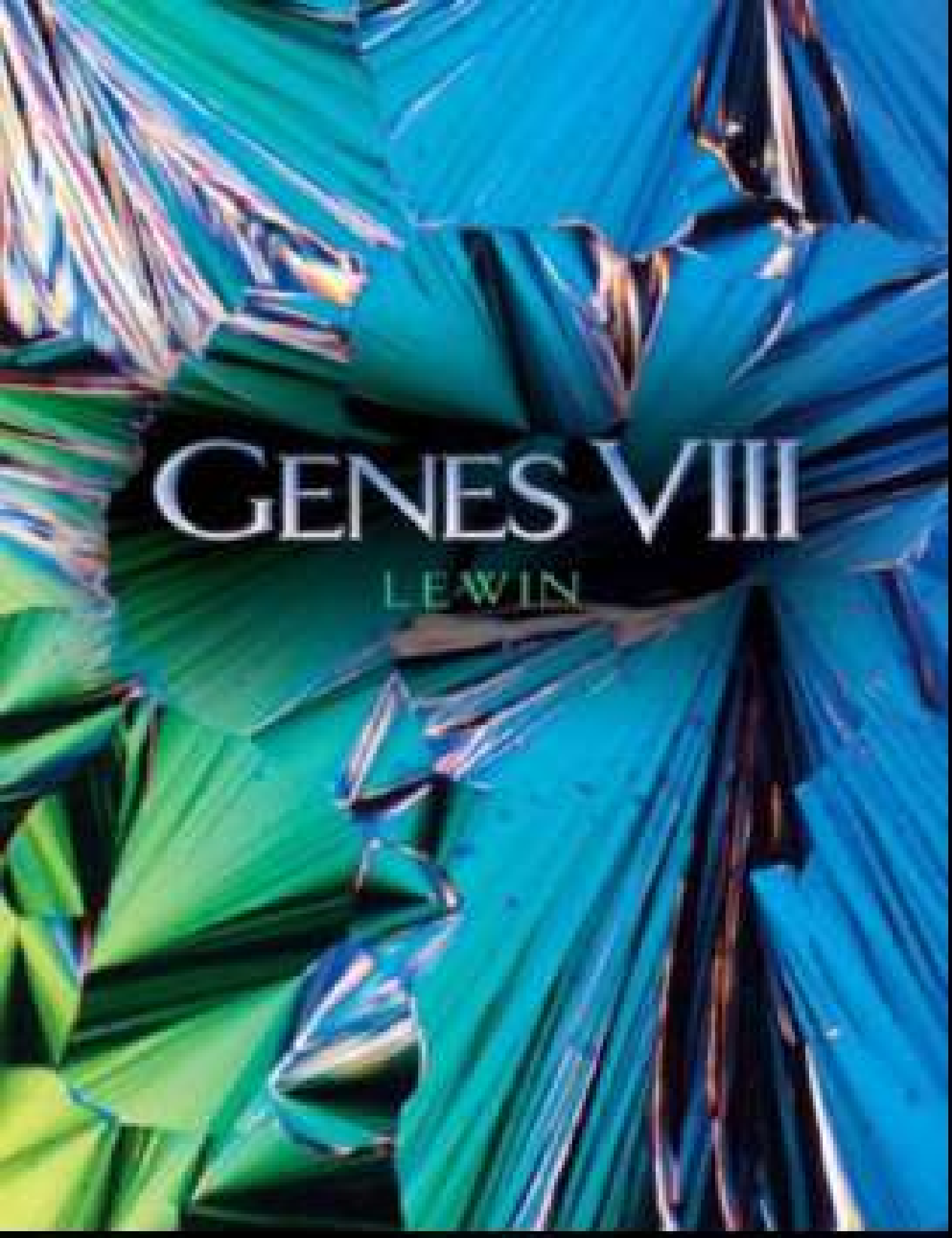

GENES VIII
LEWIN

Executive Editor: *Gary Carlson*
Editor-in-Chief: *John Challice*
President: *Paul F. Corey*
Assistant Vice President of Production and Manufacturing: *David Riccardi*
Manager of Electronic Composition: ***Jim** Sullivan*
Executive Managing Editor: *Kathleen Schiaparelli*
Editorial Assistant: *Susan Zeigler*
Assistant Managing Editor, Science Media: ***Nicole** Bush*
Media Editor: *Andrew **Stall***
Assistant Editor: *Chrissy Dudonis*
Senior Marketing Manager: *Shari Meffert*
Art Director: *John Christiana*
Book Design: *Bang Wong (Virtual Text)*
Manufacturing Buyer: *Alan Fischer*
Manufacturing Manager: *Trudy Pisciotti*
Marketing Assistant: *Juliana Tarris*
Director of Creative Services: *Paul Belfanti*
Cover Designer: *Bruce Kenselaar*
Cover Credit: *High Density Liquid Crystalline DNA by Michael W. Davidson and The Florida State University (National High Magnetic Field Laboratory)*

Published by Pearson Prentice Hall
Pearson Education, Inc.
Upper Saddle River, NT 07458

Printed in the United States of America

] 0 9 8 7 6 5 4 3 2

ISBN 0-13-123826-4

Pearson Education LTD., *London*
Pearson Education Australia PTY, Limited, *Sydney*
Pearson Education *Singapore, Pte. Ltd*
Pearson Education North Asia Ltd, *Hong Kong*
Pearson Education *Canada, Ltd., Toronto*
Pearson Educacion de *Mexico, S.A. de C.V.*
Pearson Education—*Japan, Tokyo*
Pearson Education *Malaysia, Pte. **Ltd***
Pearson Education Inc., *Upper Saddle River, New Jersey*

instant access to key research in this field. The unique user-interface allows you to view the site in three different formats, highlighting text, images or a combination of both, to best support your teaching style.

Instructor's Resource Manual (0-13-144944-3)

Test Item File (0-13-144945-1)

Transparency Package (0-13-144946-X)

For the Student:

Student Study Companion: This study tool provides students with the resources to review fundamental concepts from the text through practice questions and exercises. Additional study aids help students-to study more effectively.

Website with E-Book (*www.prenhall.com/lewin*) This powerful website contains an online version of the text, supported by weekly updates to maintain currency on key topics. Links connect the student directly to the original source material for immediate access to key articles wherever possible. The unique user-interface allows students to view the site in three different formats, highlighting text, images or a combination of both, to best support their learning style.

Outline

Contents

Part 1 **Genes**

1 Genes are DNA

2 The interrupted gene

3 The content of the genome

4 Clusters and repeats

Part 2 **Proteins**

5 Messenger RNA

6 Protein synthesis

7 Using the genetic code

8 Protein localization

Part 3 Gene expression

9 Transcription

10 The operon

11 Regulatory circuits

12 Phage strategies

Part 4 DNA

13 The replicon

14 DNA replication

15 Recombination and repair

16 Transposons

17 Retroviruses and retroposons

18 Rearrangement of DNA

Part 5 **The Nucleus**

19 Chromosomes

20 Nucleosomes

21 Promoters and enhancers

22 Activating transcription

23 Controlling chromatin structure

24 RNA splicing and processing

25 Catalytic RNA

26 Immune diversity

Part 6 **Cells**

27 Protein trafficking

28 Signal transduction

29 Cell cycle and growth regulation

30 Oncogenes and cancer

GENES is continuously updated on the web site, www.ergito.com with revisions posted weekly. This allows readers to check for revised sections and relate them to the printed book. The web site can be viewed as either sections from the book or as a slide show of the figures from the book. Some of the figures shown are animated and there are references hyperlinked to the original sources. Other features of the web site include a glossary, sophisticated searches, and ancillary material such as the essays in the Great Experiments and Structures Series. To subscribe to this site, please visit www.ergito.com.

Chapter 1

Genes are DNA

1.1 Introduction

The hereditary nature of every living organism is defined by its genome, which consists of a long sequence of nucleic acid that provides the *information* needed to construct the organism. We use the term "information" because the genome does not itself perform any active role in building the organism; rather it is the sequence of the individual subunits (bases) of the nucleic acid that determines hereditary features. By a complex series of interactions, this sequence is used to produce all the proteins of the organism in the appropriate time and place. The proteins either form part of the structure of the organism, or have the capacity to build the structures or to perform the metabolic reactions necessary for life.

The genome contains the complete set of hereditary information for any organism. Physically the genome may be divided into a number of different nucleic acid molecules. Functionally it may be divided into genes. Each gene is a sequence within the nucleic acid that represents a single protein. Each of the discrete nucleic acid molecules comprising the genome may contain a large number of genes. Genomes for living organisms may contain as few as <500 genes (for a mycoplasma, a type of bacterium) to as many as >40,000 for Man.

In this chapter, we analyze the properties of the gene in terms of its basic molecular construction. **Figure 1.1** summarizes the stages in **the** transition from the historical concept of the gene to the modern definition of the genome.

The basic behavior of the gene was defined by Mendel more than a century ago. Summarized in his two laws, the gene was recognized as a "particulate factor" that passes unchanged from parent to progeny. A gene may exist in alternative forms. These forms are called alleles.

In diploid organisms, which have two sets of chromosomes, one copy of each chromosome is inherited from each parent. This is the same behavior that is displayed by genes. One of the two copies of each gene is the paternal allele (inherited from the father), the other is the maternal allele (inherited from the mother). The equivalence led to the discovery that chromosomes in fact carry the genes.

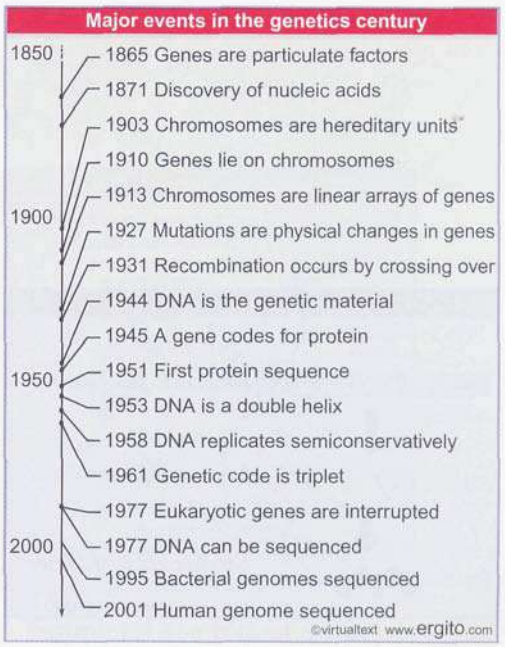


Figure 1.1 A brief history of genetics.

Each chromosome consists of a linear array of genes. Each gene resides at a particular location on the chromosome. This is more formally called a genetic locus. We can then define the alleles of this gene as the different forms that are found at this locus.

The key to understanding the organization of genes into chromosomes was the discovery of genetic linkage. This describes the observation that alleles on the same chromosome tend to remain together in the progeny instead of assorting independently as predicted by Mendel's laws. Once the unit of recombination (reassortment) was introduced as the measure of linkage, the construction of genetic maps became possible.

On the genetic maps of higher organisms established during the first half of this century, the genes are arranged like beads on a string. They occur in a fixed order, and genetic recombination involves transfer of corresponding portions of the string between homologous chromosomes. The gene is to all intents and purposes a mysterious object (the bead), whose relationship to its surroundings (the string) is unclear.

The resolution of the recombination map of a higher eukaryote is restricted by the small number of progeny that can be obtained from each mating. Recombination occurs so infrequently between nearby points that it is rarely observed between different mutations in the same gene. By moving to a microbial system in which a very large number of progeny can be obtained from each genetic cross, it became possible to demonstrate that recombination occurs within genes. It follows the same rules that were previously deduced for recombination between genes.

Mutations within a gene can be arranged into a linear order, showing that the gene itself has the same linear construction as the array of genes on a chromosome. So the genetic map is linear within as well as between loci: it consists of an unbroken sequence within which the genes reside. This conclusion leads naturally into the modern view that the genetic material of a chromosome consists of an uninterrupted length of DNA representing many genes.

A genome consists of the entire set of chromosomes for any particular organism. It therefore comprises a series of DNA molecules (one for each chromosome), each of which contains many genes. The ultimate definition of a genome is to determine the sequence of the DNA of each chromosome.

The first definition of the gene as a functional unit followed from the discovery that individual genes are responsible for the production of specific proteins. The difference in chemical nature between the DNA of the gene and its protein product led to the concept that a gene *codes* for a protein. This in turn led to the discovery of the complex apparatus that allows the DNA sequence of gene to generate the amino acid sequence of a protein.

Understanding the process by which a gene is expressed allows us to make a more rigorous definition of its nature. **Figure 1.2** shows the basic theme of this book. A gene is a sequence of DNA that produces another nucleic acid, RNA. The DNA has two strands of nucleic acid, and the RNA has only one strand. The sequence of the RNA is determined by the sequence of the DNA (in fact, it is identical to one of the DNA strands). In many, but not in all cases, the RNA is in turn used to direct production of a protein. *Thus a gene is a sequence of DNA that codes for an RNA; in protein-coding genes, the RNA in turn codes for a protein.*

From the demonstration that a gene consists of DNA, and that a chromosome consists of a long stretch of DNA representing many genes, we move to the overall organization of the genome in terms of its DNA sequence. In *2 The interrupted gene* we take up in more detail the organization of the gene and its representation in proteins. In *3 The content of the genome* we consider the total number of genes, and in *4 Clusters and repeats* we discuss other components of the genome and the maintenance of its organization.

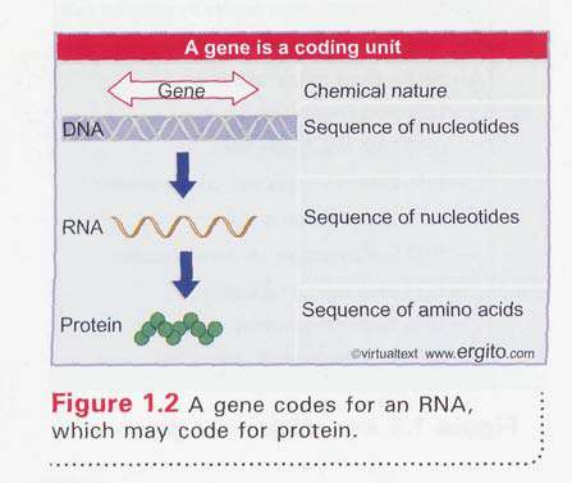


Figure 1.2 A gene codes for an RNA, which may code for protein.

1.2 DNA is the genetic material of bacteria

Key Concepts

- Bacterial transformation provided the first proof that **DNA** is **the** genetic material. Genetic properties can be transferred from one bacterial strain to another by extracting **DNA** from the first strain and adding it to the second strain.

The idea that genetic material is nucleic acid had its roots in the discovery of transformation in 1928. The bacterium *Pneumococcus* kills mice by causing pneumonia. The virulence of the bacterium is determined by its *capsular polysaccharide*. This is a component of the surface that allows the bacterium to escape destruction by the host. Several types (I, II, III) of *Pneumococcus* have different capsular polysaccharides. They have a smooth (S) appearance.

Each of the smooth *Pneumococcal* types can give rise to variants that fail to produce the capsular polysaccharide. These bacteria have a *rough* (R) surface (consisting of the material that was beneath the capsular polysaccharide). They are avirulent. They do not kill the mice, because the absence of the polysaccharide allows the animal to destroy the bacteria.

When smooth bacteria are killed by heat treatment, they lose their ability to harm the animal. But inactive heat-killed S bacteria and the ineffectual variant R bacteria together have a quite different effect from either bacterium by itself. **Figure 1.3** shows that when they are jointly injected into an animal, the mouse dies as the result of a *Pneumococcal* infection. Virulent S bacteria can be recovered from the mouse postmortem.

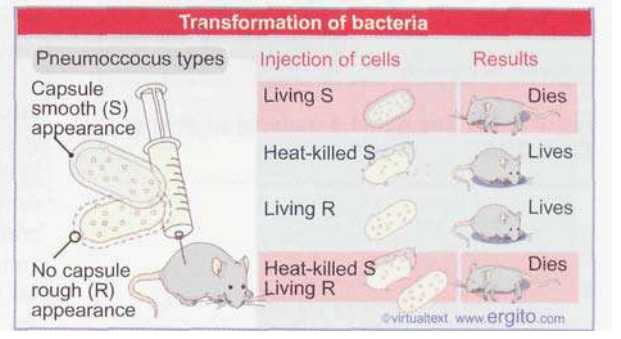


Figure 1.3 Neither heat-killed S-type nor live R-type bacteria can kill mice, but simultaneous infection of them both can kill mice just as effectively as the live S-type.

In this experiment, the dead S bacteria were of type III. The live R bacteria had been derived from type II. The virulent bacteria recovered from the mixed infection had the smooth coat of type III. So some property of the dead type III S bacteria can *transform* the live R bacteria so that they make the type III capsular polysaccharide, and as a result become virulent.

Figure 1.4 shows the identification of the component of the dead bacteria responsible for transformation. This was called the transforming principle. It was purified by developing a cell-free system, in which extracts of the dead S bacteria could be added to the live R bacteria before injection into the animal. Purification of the transforming principle in 1944 showed that it is deoxyribonucleic acid (DNA).

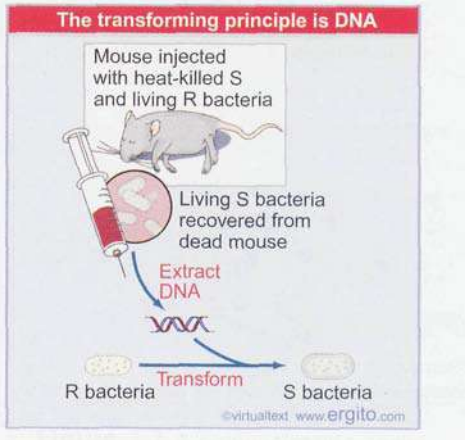


Figure 1.4 The DNA of S-type bacteria can transform R-type bacteria into the same S-type.

1.3 DNA is the genetic material of viruses

Key Concepts

- Phage infection proved that DNA is the genetic material of viruses. When the **DNA** and protein components of bacteriophages are labeled with different radioactive isotopes, only the **DNA** is transmitted to the progeny phages produced by infecting bacteria.

Having shown that DNA is the genetic material of bacteria, the next step was to demonstrate that DNA provides the genetic material in a quite different system. Phage T2 is a virus that infects the

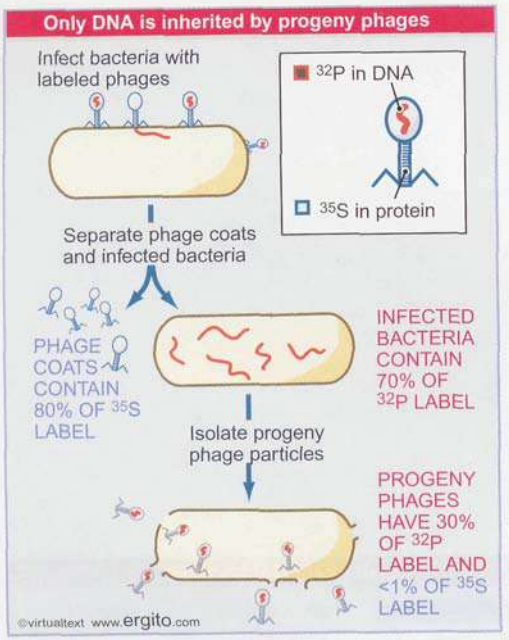


Figure 1.5 The genetic material of phage T2 is DNA.

bacterium *E. coli*. When phage particles are added to bacteria, they adsorb to the outside surface, some material enters the bacterium, and then ~20 minutes later each bacterium bursts open (lyses) to release a large number of progeny phage.

Figure 1.5 illustrates the results of an experiment in 1952 in which bacteria were infected with T2 phages that had been radioactively labeled *either* in their DNA component (with ^{32}P) *or* in their protein component (with ^{35}S). The infected bacteria were agitated in a blender, and two fractions were separated by centrifugation. One contained the empty phage coats that were released from the surface of the bacteria. The other fraction consisted of the infected bacteria themselves.

Most of the ^{32}P label was present in the infected bacteria. The progeny phage particles produced by the infection contained ~30% of the original ^{32}P label. The progeny received very little—less than 1%—of the protein contained in the original phage population. The phage coats consist of protein and therefore carried the ^{35}S radioactive label. This experiment therefore showed directly that only the DNA of the parent phages enters the bacteria and then becomes part of the progeny phages, exactly the pattern of inheritance expected of genetic material.

A phage (virus) reproduces by commandeering the machinery of an infected host cell to manufacture more copies of itself. The phage possesses genetic material whose behavior is analogous to that of cellular genomes: its traits are faithfully reproduced, and they are subject to the same rules that govern inheritance. The case of Ti reinforces the general conclusion that the genetic material is DNA, whether part of the genome of a cell or virus.

1.4 DNA is the genetic material of animal cells

Key Concepts

- DNA can be used to introduce new genetic features into animal cells or whole animals.
- In some viruses, the genetic material is RNA.

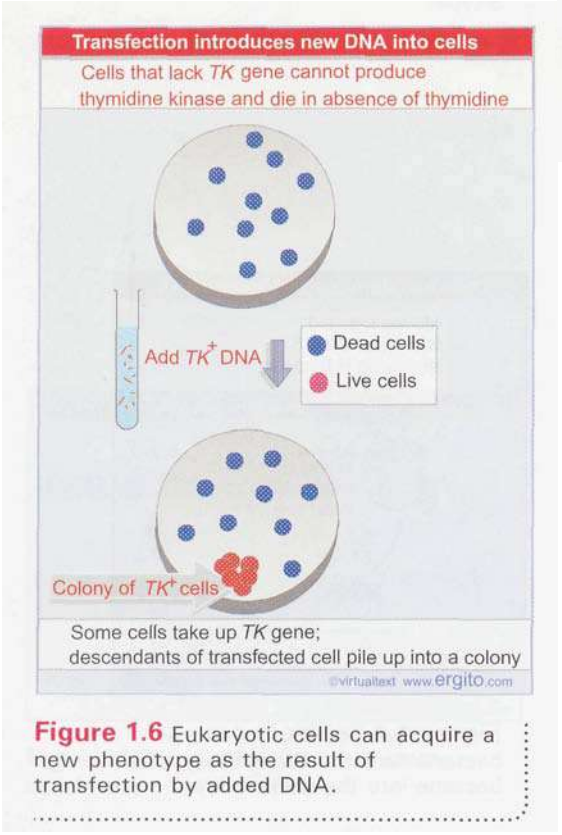


Figure 1.6 Eukaryotic cells can acquire a new phenotype as the result of transfection by added DNA.

When DNA is added to populations of single eukaryotic cells growing in culture, the nucleic acid enters the cells, and in some of them results in the production of new proteins. When a purified DNA is used, its incorporation leads to the production of a particular protein. **Figure 1.6** depicts one of the standard systems.

Although for historical reasons these experiments are described as transfection when performed with eukaryotic cells, they are a direct counterpart to bacterial transformation. The DNA that is introduced into the recipient cell becomes part of its genetic material, and is inherited in the same way as any other part. Its expression confers a new trait upon the cells (synthesis of thymidine kinase in the example of the figure). At first, these experiments were successful only with individual cells adapted to grow in a culture medium. Since then, however, DNA has been introduced into mouse eggs by microinjection; and it may become a stable part of the genetic material of the mouse (see *18.18 Genes can be injected into animal eggs).*

Such experiments show directly not only that DNA is the genetic material in eukaryotes, but also that *it can be transferred between different species and yet remain functional.*

The genetic material of all known organisms and many viruses is DNA. However, some viruses use an alternative type of nucleic acid,

ribonucleic acid (RNA), as the genetic material. The general principle of the nature of the genetic material, then, is that it is always nucleic acid; in fact, it is DNA except in the RNA viruses.

1.5 Polynucleotide chains have nitrogenous bases linked to a sugar-phosphate backbone

Key Concepts

- A nucleoside consists of a purine or pyrimidine base linked to position 1 of a pentose sugar.
- Positions on the ribose ring are described with a prime (′) to distinguish them.
- The difference between DNA and RNA is in the group at the 2′ position of the sugar. DNA has a deoxyribose sugar (2′-H); RNA has a ribose sugar (2′-OH).
- A nucleotide consists of a nucleoside linked to a phosphate group on either the 5′ or 3′ position of the (deoxy)ribose.
- Successive (deoxy)ribose residues of a polynucleotide chain are joined by a phosphate group between the 3′ position of one sugar and the 5′ position of the next sugar.
- One end of the chain (conventionally the left) has a free 5′ end and the other end has a free 3′ end.
- DNA contains the four bases adenine, guanine, cytosine, and thymine; RNA has uracil instead of thymine.

The basic building block of nucleic acids is the nucleotide. This has three components:

- a nitrogenous base;
- a sugar;
- and a phosphate.

The nitrogenous base is a purine or pyrimidine ring. The base is linked to position 1 on a pentose sugar by a glycosidic bond from N_1 of pyrimidines or N_9 of purines. To avoid ambiguity between the numbering systems of the heterocyclic rings and the sugar, positions on the pentose are given a prime (′).

Nucleic acids are named for the type of sugar; DNA has 2'-deoxyribose, whereas RNA has ribose. The difference is that the sugar in RNA has an OH group at the 2' position of the pentose ring. The sugar can be linked by its 5' or 3' position to a phosphate group.

A nucleic acid consists of a long chain of nucleotides. **Figure 1.7** shows that the backbone of the polynucleotide chain consists of an alternating series of pentose (sugar) and phosphate residues. This is constructed by linking the 5' position of one pentose ring to the 3' position of the next pentose ring via a phosphate group. So the sugar-phosphate backbone is said to consist of 5′–3′ phosphodiester linkages. The nitrogenous bases "stick out" from the backbone.

Each nucleic acid contains 4 types of base. The same two purines, adenine and guanine, are present in both DNA and RNA. The two pyrimidines in DNA are cytosine and thymine; in RNA uracil is found instead of thymine. The only difference between uracil and thymine is the presence of a methyl substituent at position C_5. The bases are usually referred to by their initial letters. DNA contains A, G, C, T, while RNA contains A, G, C, U.

The terminal nucleotide at one end of the chain has a free 5' group; the terminal nucleotide at the other end has a free 3' group. It is conventional to write nucleic acid sequences in the 5′→3′ direction—that is, from the 5' terminus at the left to the 3' terminus at the right.

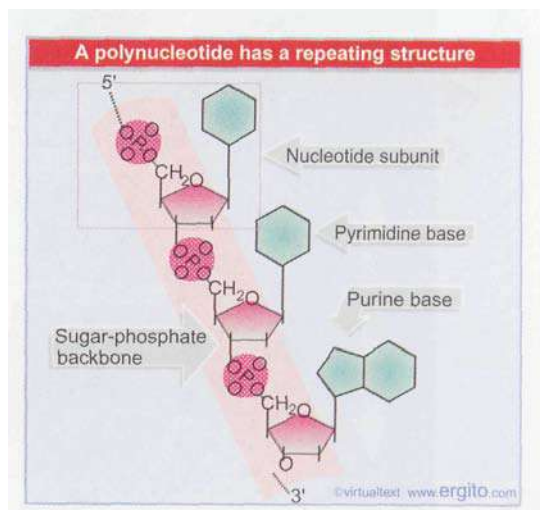


Figure 1.7 A polynucleotide chain consists of a series of 5′-3′ sugar-phosphate links that form a backbone from which the bases protrude.

1.6 DNA is a double helix

Key Concepts

- The B-form of DNA is a double helix consisting of two polynucleotide chains that run antiparallel.
- The nitrogenous bases of each chain are flat purine or pyrimidine rings that face inwards and pair with one another by hydrogen bonding to form A-T or G-C pairs only.
- The diameter of the double helix is 20 Å, and there is a complete turn every 34 Å, with 10 base pairs per turn.
- The double helix forms a major (wide) groove and a minor (narrow) groove.

The observation that the bases are present in different amounts in the DNAs of different species led to the concept that the *sequence of bases is the form in which genetic information is carried.* By the 1950s, the concept of genetic information was common: the twin problems it posed were working out the structure of the nucleic acid, and explaining how a sequence of bases in DNA could represent the sequence of amino acids in a protein.

Three notions converged in the construction of the double helix model for DNA by Watson and Crick in 1953:

- X-ray diffraction data showed that DNA has the form of a regular helix, making a complete turn every 34 A (3.4 nm), with a diameter of ~20 A (2 nm). Since the distance between adjacent nucleotides is 3.4 A, there must be 10 nucleotides per turn.
- The density of DNA suggests that the helix must contain two polynucleotide chains. The constant diameter of the helix can be explained if the bases in each chain face inward and are restricted so that a purine is always opposite a pyrimidine, avoiding partnerships of purine-purine (too wide) or pyrimidine-pyrimidine (too narrow).
- Irrespective of the absolute amounts of each base, the proportion of G is always the same as the proportion of C in DNA, and the proportion of A is always the same as that of T. So the composition of any DNA can be described by the proportion of its bases that is G + C. This ranges from 26% to 74% for different species.

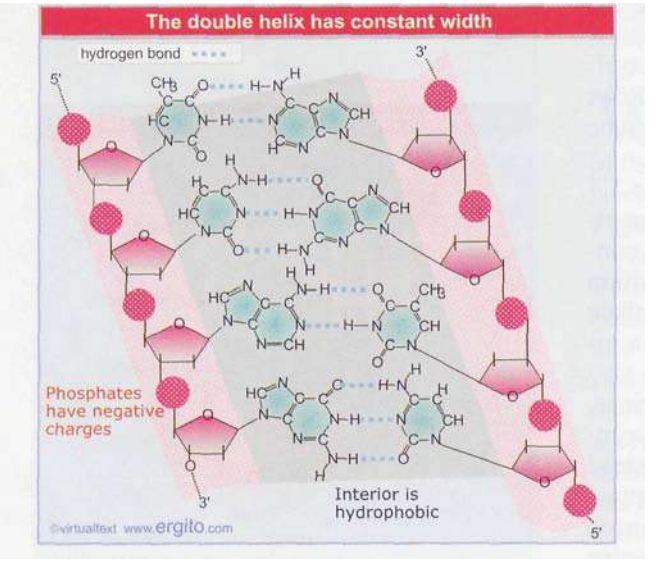


Figure 1.8 The double helix maintains a constant width because purines always face pyrimidines in the complementary A-T and G-C base pairs. The sequence in the figure is T-A, C-G, A-T, G-C.

Watson and Crick proposed that the two polynucleotide chains in the double helix associate by *hydrogen bonding between the nitrogenous bases.* G can hydrogen bond specifically only with C, while A can bond specifically only with T. These reactions are described as base pairing, and the paired bases (G with C, or A with T) are said to be complementary.

The model proposed that the two polynucleotide chains run in opposite directions (antiparallel), as illustrated in **Figure 1.8.** Looking along the helix, one strand runs in the 5′→3′ direction, while its partner runs 3′→5′.

The sugar-phosphate backbone is on the outside and carries negative charges on the phosphate groups. When DNA is in solution *in vitro,* the charges are neutralized by the binding of metal ions, typically by Na^+. In the cell, positively charged proteins provide some of the neutralizing force. These proteins play an important role in determining the organization of DNA in the cell.

The bases lie on the inside. They are flat structures, lying in pairs perpendicular to the axis of the helix. Consider the double helix in

terms of a spiral staircase: the base pairs form the treads, as illustrated schematically in **Figure 1.9.** Proceeding along the helix, bases are stacked above one another, in a sense like a pile of plates.

Each base pair is rotated ~36° around the axis of the helix relative to the next base pair. So ~10 base pairs make a complete turn of 360°. The twisting of the two strands around one another forms a double helix with a minor groove (~12 A across) and a major groove (~22 A across), as can be seen from the scale model of **Figure 1.10.** The double helix is right-handed; the turns run clockwise looking along the helical axis. These features represent the accepted model for what is known as the B-form of DNA.

It is important to realize that the B-form represents an *average*, not a precisely specified structure. DNA structure can change locally. If it has more base pairs per turn it is said to be overwound; if it has fewer base pairs per turn it is underwound. Local winding can be affected by the overall conformation of the DNA double helix in space or by the binding of proteins to specific sites.

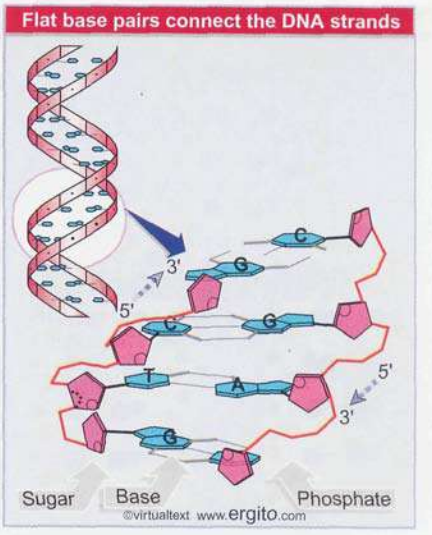


Figure 1.9 Flat base pairs lie perpendicular to the sugar-phosphate backbone.

1.7 DNA replication is semiconservative

Key Concepts

- The Meselson-Stahl experiment used density labeling to prove that the single polynucleotide strand is the unit of DNA that is conserved during replication.
- Each strand of a DNA duplex acts as a template to synthesize a daughter strand.
- The sequences of the daughter strands are determined by complementary base pairing with the separated parental strands.

It is crucial that the genetic material is reproduced accurately. Because the two polynucleotide strands are joined only by hydrogen bonds, they are able to separate without requiring breakage of covalent bonds. The specificity of base pairing suggests that each of the separated parental strands could act as a template strand for the synthesis of a complementary daughter strand. **Figure 1.11** shows the principle that a new daughter strand is assembled on each parental strand. The sequence of the daughter strand is dictated by the parental strand; an *A* in the parental strand causes a *T* to be placed in the daughter strand, a parental *G* directs incorporation of a daughter *C*, and so on.

The top part of the figure shows a parental (unreplicated) duplex that consists of the original two parental strands. The lower part shows the two daughter duplexes that are being produced by complementary base pairing. Each of the daughter duplexes is identical in sequence with the original parent, and contains one parental strand and one newly synthesized strand. *The structure of DNA carries the information needed to perpetuate its sequence.*

The consequences of this mode of replication are illustrated in Figure **1.12.** The parental duplex is replicated to form two daughter duplexes, each of which consists of one parental strand and one (newly synthesized) daughter strand. *The unit conserved from one generation to the next is one of the two individual strands comprising the parental duplex.* This behavior is called semiconservative replication.

The figure illustrates a prediction of this model. If the parental DNA "*heavy*" *density label because the organism has been grown in*

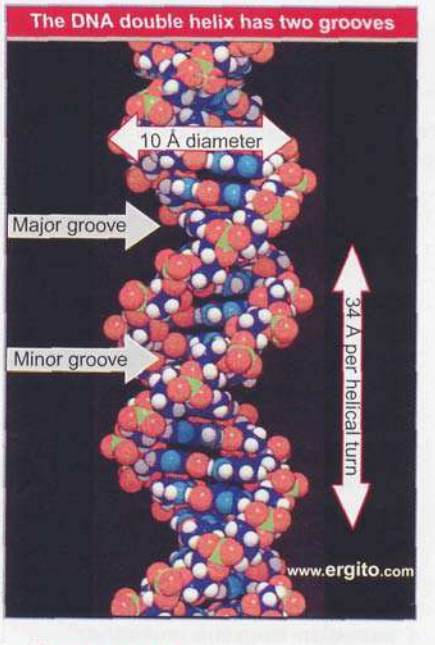


Figure 1.10 The two strands of DNA form a double helix.

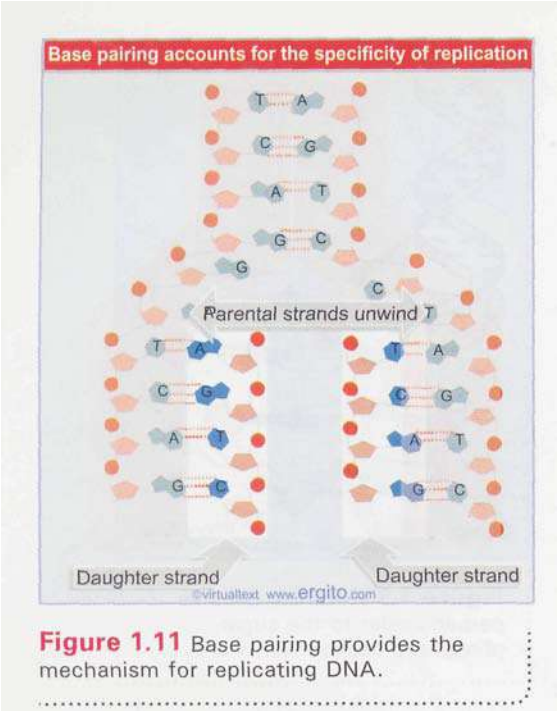


Figure 1.11 Base pairing provides the mechanism for replicating DNA.

medium containing a suitable isotope (such as ^{15}N), its strands can be distinguished from those that are synthesized when the organism is transferred to a medium containing normal "light" isotopes.

The parental DNA consists of a duplex of two heavy strands (red). After one generation of growth in light medium, the duplex DNA is "hybrid" in density—it consists of one heavy parental strand (red) and one light daughter strand (blue). After a second generation, the two strands of each hybrid duplex have separated; each gains a light partner, so that now half of the duplex DNA remains hybrid while half is entirely light (both strands are blue).

The individual strands of these duplexes are entirely heavy or entirely light. This pattern was confirmed experimentally in the Meselson-Stahl experiment of 1958, which followed the semiconservative replication of DNA through three generations of growth of *E. coli*. When DNA was. extracted from bacteria and its density measured by centrifugation, the DNA formed bands corresponding to its density—heavy for parental, hybrid for the first generation, and half hybrid and half light in the second generation.

1.8 DNA strands separate at the replication fork

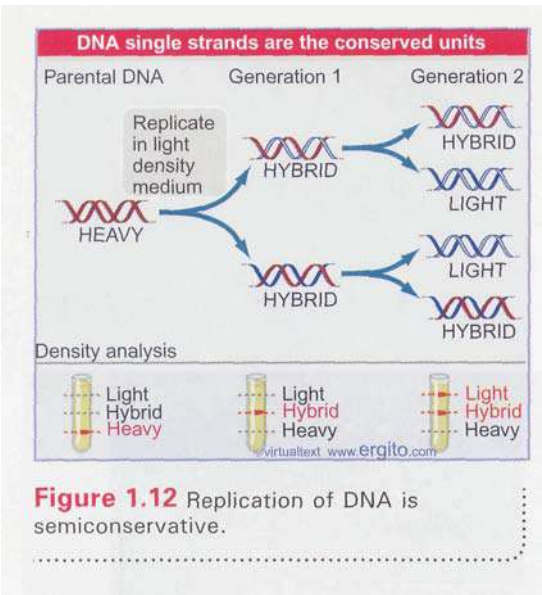


Figure 1.12 Replication of DNA is semiconservative.

Key Concepts

- Replication of DNA is undertaken by a complex of enzymes that separate the parental strands and synthesize the daughter strands.
- The replication fork is the point at which the parental strands are separated.
- The enzymes that synthesize DNA are called DNA polymerases; the enzymes that synthesize RNA are RNA polymerases.
- Nucleases are enzymes that degrade nucleic acids; they include DNAases and RNAases, and can be divided into endonucleases and exonucleases.

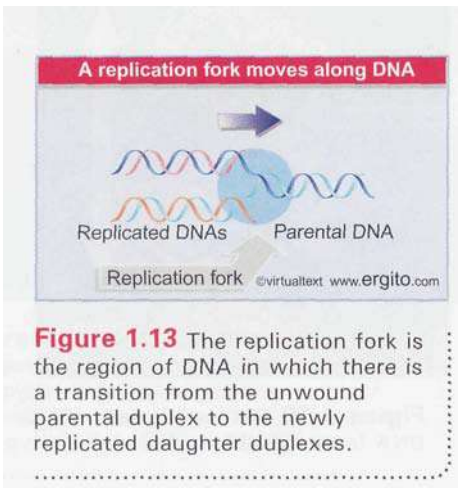


Figure 1.13 The replication fork is the region of DNA in which there is a transition from the unwound parental duplex to the newly replicated daughter duplexes.

Replication requires the two strands of the parental duplex to separate. However, the disruption of structure is only transient and is reversed as the daughter duplex is formed. Only a small stretch of the duplex DNA is separated into single strands at any moment.

The helical structure of a molecule of DNA engaged in replication is illustrated in **Figure 1.13.** The nonreplicated region consists of the parental duplex, opening into the replicated region where the two daughter duplexes have formed. The double helical structure is disrupted at the junction between the two regions, which is called the replication fork. Replication involves movement of the replication fork along the parental DNA, so there is a continuous unwinding of the parental strands and rewinding into daughter duplexes.

The synthesis of nucleic acids is catalyzed by specific enzymes, which recognize the template and undertake the task of catalyzing the addition of subunits to the polynucleotide chain that is being synthesized. The enzymes are named according to the type of chain that is synthesized: DNA polymerases synthesize DNA, and RNA polymerases synthesize RNA.

Degradation of nucleic acids also requires specific enzymes: deoxyribonucleases (DNAases) degrade DNA, and ribonucleases (RNAases) degrade RNA. The nucleases fall into the general classes of exonucleases and endonucleases:

Endonucleases cut individual bonds *within* RNA or DNA molecules, generating discrete fragments. Some DNAases cleave both strands of a duplex DNA at the target site, while others cleave only one of the two strands. Endonucleases are involved in cutting reactions, as shown in **Figure 1.14.**

Exonucleases remove residues one at a time from the end of the molecule, generating mononucleotides. They always function on a single nucleic acid strand, and each exonuclease proceeds in a specific direction, that is, starting at either a 5' or at a 3' end and proceeding toward the other end. They are involved in trimming reactions, as shown in **Figure 1.15.**

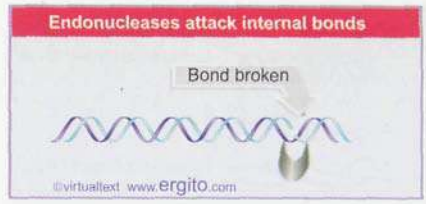


Figure 1.14 An endonuclease cleaves a bond within a nucleic acid. This example shows an enzyme that attacks one strand of a DNA duplex.

1.9 Nucleic acids hybridize by base pairing

Key Concepts

- Heating causes the two strands of a DNA duplex to separate.
- The T_m is the midpoint of the temperature range for denaturation.
- Complementary single strands can renature when the temperature is reduced.
- Denaturation and renaturation/hybridization can occur with DNA-DNA, DNA-RNA, or RNA-RNA combinations, and can be intermolecular or intramolecular.
- The ability of two single-stranded nucleic acid preparations to hybridize is a measure of their complementarity.

Figure 1.15 An exonuclease removes bases one at a time by cleaving the last bond in the polynucleotide chain. This example shows an enzyme that degrades an RNA chain.

A crucial property of the double helix is the ability to separate the two strands without disrupting covalent bonds. This makes it possible for the strands to separate and reform under physiological conditions at the (very rapid) rates needed to sustain genetic functions. The specificity of the process is determined by complementary base pairing.

The concept of base pairing is central to all processes involving nucleic acids. Disruption of the base pairs is a crucial aspect of the function of a double-stranded molecule, while the ability to form base pairs is essential for the activity of a single-stranded nucleic acid. **Figure 1.16** shows that base pairing enables complementary single-stranded nucleic acids to form a duplex structure.

- An intramolecular duplex region can form by base pairing between two complementary sequences that are part of a single-stranded molecule.
- A single-stranded molecule may base pair with an independent, complementary single-stranded molecule to form an intermolecular duplex.

Formation of duplex regions from single-stranded nucleic acids is most important for RNA, but single-stranded DNA also exists (in the form of viral genomes). Base pairing between independent complementary single strands is not restricted to DNA-DNA or RNA-RNA, but can also occur between a DNA molecule and an RNA molecule.

The lack of covalent links between complementary strands makes it possible to manipulate DNA *in vitro*. The noncovalent forces that stabilize the double helix are disrupted by heating or by exposure to low salt concentration. The two strands of a double helix separate entirely when all the hydrogen bonds between them are broken.

The process of strand separation is called denaturation or (more colloquially) *melting*. ("Denaturation" is also used to describe loss of

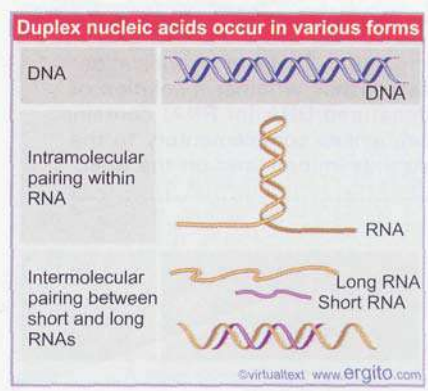


Figure 1.16 Base pairing occurs in duplex DNA and also in intra- and inter-molecular interactions in single-stranded RNA (or DNA).

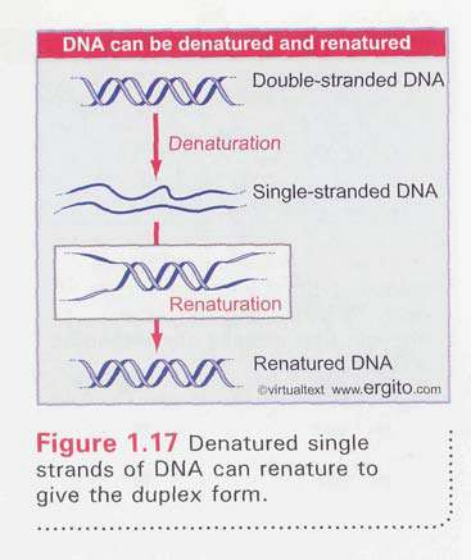


Figure 1.17 Denatured single strands of DNA can renature to give the duplex form.

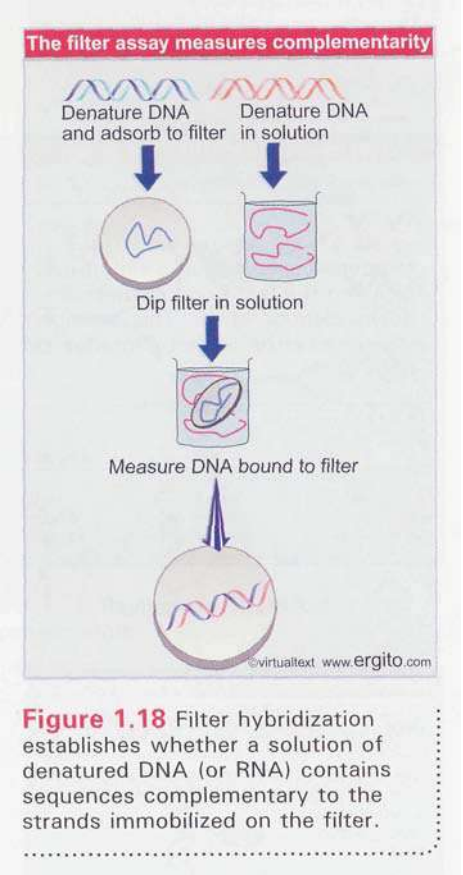


Figure 1.18 Filter hybridization establishes whether a solution of denatured DNA (or RNA) contains sequences complementary to the strands immobilized on the filter.

authentic protein structure; it is a general term implying that the natural conformation of a macromolecule has been converted to some other form.)

Denaturation of DNA occurs over a narrow temperature range and results in striking changes in many of its physical properties. The midpoint of the temperature range over which the strands of DNA separate is called the *melting temperature* (T_m). It depends on the proportion of GC base pairs. Because each G·C base pair has three hydrogen bonds, it is more stable than an A·T base pair, which has only two hydrogen bonds. The more G-C base pairs are contained in a DNA, the greater the energy that is needed to separate the two strands. In solution under physiological conditions, a DNA that is 40% G·C—a value typical of mammalian genomes—denatures with a T_m of about 87°C. So duplex DNA is stable at the temperature prevailing in the cell.

The denaturation of DNA is reversible under appropriate conditions. The ability of the two separated complementary strands to reform into a double helix is called renaturation. Renaturation depends on specific base pairing between the complementary strands. **Figure 1.17** shows that the reaction takes place in two stages. First, single strands of DNA in the solution encounter one another by chance; if their sequences are complementary, the two strands base pair to generate a short double-helical region. Then the region of base pairing extends along the molecule by a zipper-like effect to form a lengthy duplex molecule. Renaturation of the double helix restores the original properties that were lost when the DNA was denatured.

Renaturation describes the reaction between two complementary sequences that were separated by denaturation. However, the technique can be extended to allow any two complementary nucleic acid sequences to react with each other to form a duplex structure. This is sometimes called annealing, but the reaction is more generally described as hybridization whenever nucleic acids of different sources are involved, as in the case when one preparation consists of DNA and the other consists of RNA. *The ability of two nucleic acid preparations to hybridize constitutes a precise test for their complementarity since* only *complementary sequences can form a duplex structure.*

The principle of the hybridization reaction is to expose two single-stranded nucleic acid preparations to each other and then to measure the amount of double-stranded material that forms. **Figure 1.18** illustrates a procedure in which a DNA preparation is denatured and the single strands are adsorbed to a filter. Then a second denatured DNA (or RNA) preparation is added. The filter is treated so that the second preparation can adsorb to it only if it is able to base pair with the DNA that was originally adsorbed. Usually the second preparation is radioactively labeled, so that the reaction can be measured as the amount of radioactive label retained by the filter.

The extent of hybridization between two single-stranded nucleic acids is determined by their complementarity. Two sequences need not be *perfectly* complementary to hybridize. If they are closely related but not identical, an imperfect duplex is formed in which base pairing is interrupted at positions where the two single strands do not correspond.

1.10 Mutations change the sequence of DNA

Key Concepts

* All mutations consist of changes in the sequence of DNA.
* Mutations may occur spontaneously or may be induced by mutagens.

Mutations provide decisive evidence that DNA is the genetic material. When a change in the sequence of DNA causes an alteration in the sequence of a protein, we may conclude that the DNA codes for that protein. Furthermore, a change in the phenotype of the organism may allow us to identify the function of the protein. The existence of many mutations in a gene may allow many variant forms of a protein to be compared, and a detailed analysis can be used to identify regions of the protein responsible for individual enzymatic or other functions.

All organisms suffer a certain number of mutations as the result of normal cellular operations or random interactions with the environment. These are called spontaneous mutations; the rate at which they occur is characteristic for any particular organism and is sometimes called the background level. Mutations are rare events, and of course those that damage a gene are selected against during evolution. It is therefore difficult to obtain large numbers of spontaneous mutants to study from natural populations.

The occurrence of mutations can be increased by treatment with certain compounds. These are called mutagens, and the changes they cause are referred to as induced mutations. Most mutagens act directly by virtue of an ability either to modify a particular base of DNA or to become incorporated into the nucleic acid. The effectiveness of a mutagen is judged by how much it increases the rate of mutation above background. By using mutagens, it becomes possible to induce many changes in any gene.

Spontaneous mutations that inactivate gene function occur in bacteriophages and bacteria at a relatively constant rate of 3-4 x 10^{-3} per genome per generation. Given the large variation in genome sizes between bacteriophages and bacteria, this corresponds to wide differences in the mutation rate per base pair. This suggests that the overall rate of mutation has been subject to selective forces that have balanced the deleterious effects of most mutations against the advantageous effects of some mutations. This conclusion is strengthened by the observation that an archaeal microbe that lives under harsh conditions of high temperature and acidity (which are expected to damage DNA) does not show an elevated mutation rate, but in fact has an overall mutation rate just below the average range.

Figure 1.19 shows that in bacteria, the mutation rate corresponds to ~10^{-6} events per locus per generation or to an average rate of change per base pair of 10^{-9}-10^{-10} per generation. The rate at individual base pairs varies very widely, over a 10,000 fold range. We have no accurate measurement of the rate of mutation in eukaryotes, although usually it is thought to be somewhat similar to that of bacteria on a per-locus per-generation basis. We do not know what proportion of the spontaneous events results from point mutations.

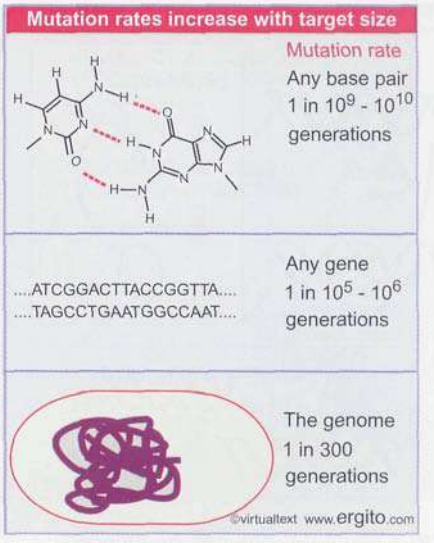


Figure 1.19 A base pair is mutated at a rate of 10^{-9} - 10^{-10} per generation, a gene of 1000 bp is mutated at ~10^{-6} per generation, and a bacterial genome is mutated at 3 × 10^{-3} per generation.

1.11 Mutations may affect single base pairs or longer sequences

Key Concepts

- A point mutation changes a single base pair.
- Point mutations can be caused by the chemical conversion of one base into another or by mistakes that occur during replication.
- A transition replaces a G·C base pair with an A·T base pair or vice-versa.
- A transversion replaces a purine with a pyrimidine, such as changing A-T to T·A.
- Insertions are the most common type of mutation, and result from the movement of transposable elements.

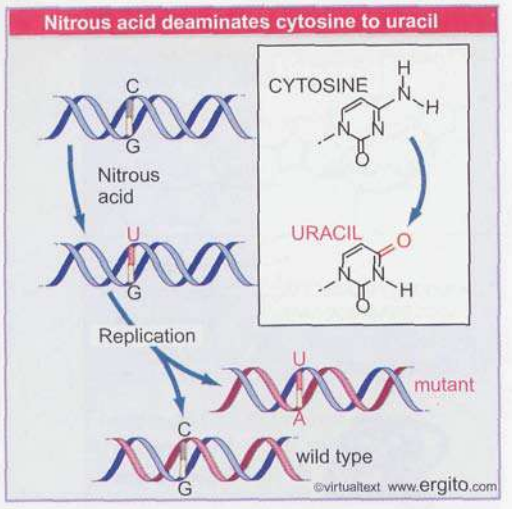


Figure 1.20 Mutations can be induced by chemical modification of a base.

Any base pair of DNA can be mutated. A point mutation changes only a single base pair, and can be caused by either of two types of event:

- Chemical modification of DNA directly changes one base into a different base.
- A malfunction during the replication of DNA causes the wrong base to be inserted into a polynucleotide chain during DNA synthesis.

Point mutations can be divided into two types, depending on the nature of the change when one base is substituted for another:

- The most common class is the transition, comprising the substitution of one pyrimidine by the other, or of one purine by the other. This replaces a GC pair with an AT pair or vice versa.
- The less common class is the transversion, in which a purine is replaced by a pyrimidine or vice versa, so that an AT pair becomes a T A or C G pair.

The effects of nitrous acid provide a classic example of a transition caused by the chemical conversion of one base into another. **Figure 1.20** shows that nitrous acid performs an oxidative deamination that converts cytosine into uracil. In the replication cycle following the transition, the U pairs with an A, instead of with the G with which the original C would have paired. So the CG pair is replaced by a T·A pair when the A pairs with the T in the next replication cycle. (Nitrous acid also deaminates adenine, causing the reverse transition from AT to G·C.)

Transitions are also caused by base mispairing, when unusual partners pair in defiance of the usual restriction to Watson-Crick pairs. Base mispairing usually occurs as an aberration resulting from the incorporation into DNA of an abnormal base that has ambiguous pairing properties. **Figure 1.21** shows the example of bromouracil (BrdU), an analog of thymine that contains a bromine atom in place of the methyl group of thymine. BrdU is incorporated into DNA in place of thymine. But it has ambiguous pairing properties, because the presence of the bromine atom allows a shift to occur in which the base changes structure from a keto (=O) form to an enol (–OH) form. The enol form can base pair with guanine, which leads to substitution of the original AT pair by a GC pair.

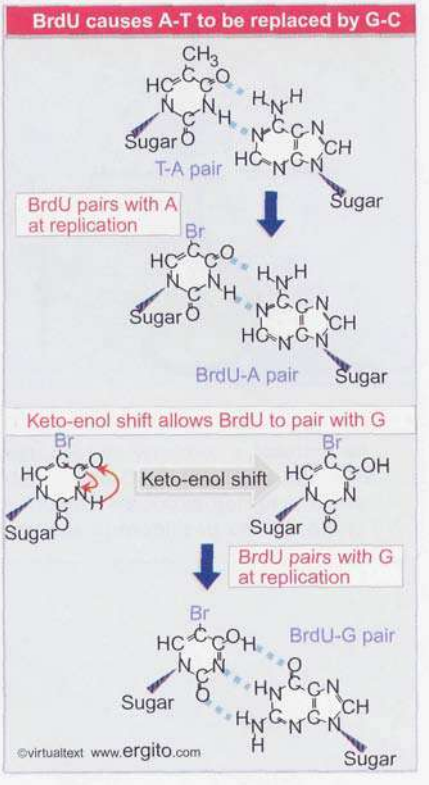


Figure 1.21 Mutations can be induced by the incorporation of base analogs into DNA.

The mistaken pairing can occur either during the original incorporation of the base or in a subsequent replication cycle. The transition is induced with a certain probability in each replication cycle, so the incorporation of BrdU has continuing effects on the sequence of DNA.

Point mutations were thought for a long time to be the principal means of change in individual genes. However, we now know that insertions of stretches of additional material are quite frequent. The source of the inserted material lies with transposable elements, sequences of DNA with the ability to move from one site to another (see *16 Transposons* and *17 Retroviruses and retroposons).* An insertion usually abolishes the activity of a gene. Where such insertions have occurred, deletions of part or all of the inserted material, and sometimes of the adjacent regions, may subsequently occur.

A significant difference between point mutations and the insertions/deletions is that the frequency of point mutation can be increased by mutagens, whereas the occurrence of changes caused by transposable elements is not affected. However, insertions and deletions can also occur by other mechanisms—for example, involving mistakes made during replication or recombination—although probably these are less common. And a class of mutagens called the acridines introduce (very small) insertions and deletions.

1.12 The effects of mutations can be reversed

Key Concepts

- Forward mutations inactivate a gene, and back mutations (or revertants) reverse their effects.
- Insertions can revert by deletion of the inserted material, but deletions cannot revert.
- Suppression occurs when a mutation in a second gene bypasses the effect of mutation in the first gene.

Figure 1.22 shows that the isolation of revertants is an important characteristic that distinguishes point mutations and insertions from deletions:

- A point mutation can revert by restoring the original sequence or by gaining a compensatory mutation elsewhere in the gene.
- An insertion of additional material can revert by deletion of the inserted material.
- A deletion of part of a gene cannot revert.

Mutations that inactivate a gene are called forward mutations. Their effects are reversed by back mutations, which are of two types.

An exact reversal of the original mutation is called true reversion. So if an AT pair has been replaced by a G·C pair, another mutation to restore the AT pair will exactly regenerate the wild-type sequence.

Alternatively, another mutation may occur elsewhere in the gene, and its effects compensate for the first mutation. This is called second-site reversion. For example, one amino acid change in a protein may abolish gene function, but a second alteration may compensate for the first and restore protein activity.

A forward mutation results from any change that inactivates a gene, whereas a back mutation must restore function to a protein damaged by a particular forward mutation. So the demands for back mutation are much more specific than those for forward mutation. The rate of back mutation is correspondingly lower than that of forward mutation, typically by a factor of ~\ 0.

Mutations can also occur in other genes to circumvent the effects of mutation in the original gene. This effect is called suppression. A locus in which a mutation suppresses the effect of a mutation in another locus is called a suppressor.

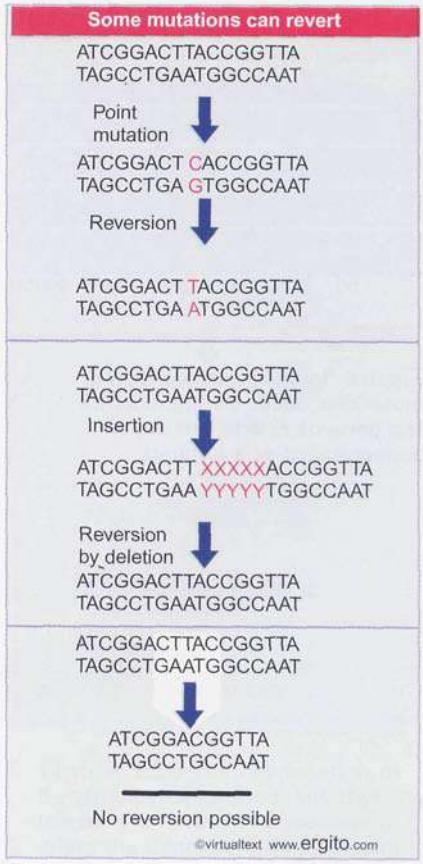


Figure 1.22 Point mutations and insertions can revert, but deletions cannot revert.

1.13 Mutations are concentrated at hotspots

Key Concepts

- The frequency of mutation at any particular base pair is determined by statistical fluctuation, except for hotspots, where the frequency is increased by at least an order of magnitude.

So far we have dealt with mutations in terms of individual changes in the sequence of DNA that influence the activity of the genetic unit in which they occur. When we consider mutations in terms of the inactivation of the gene, most genes within a species show more or less similar rates of mutation relative to their size. This suggests that the gene can be regarded as a target for mutation, and that damage to

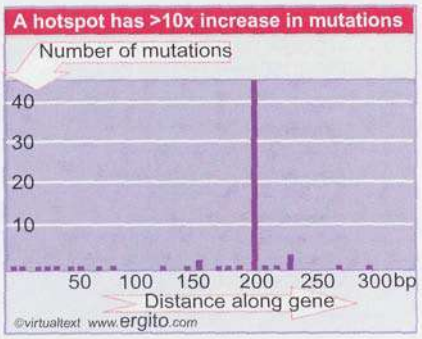


Figure 1.23 Spontaneous mutations occur throughout the *lacI* gene of *E. coli*, but are concentrated at a hotspot.

any part of it can abolish its function. As a result, susceptibility to mutation is roughly proportional to the size of the gene. But consider the sites of mutation within the sequence of DNA; are all base pairs in a gene equally susceptible or are some more likely to be mutated than others?

What happens when we isolate a large number of independent mutations in the same gene? Many mutants are obtained. Each is the result of an individual mutational event. Then the site of each mutation is determined. Most mutations will lie at different sites, but some will lie at the same position. Two independently isolated mutations at the same site may constitute exactly the same change in DNA (in which case the same mutational event has happened on more than one occasion), or they may constitute different changes (three different point mutations are possible at each base pair).

The histogram of **Figure 1.23** shows the frequency with which mutations are found at each base pair in the *lacI* gene of *E. coli*. The statistical probability that more than one mutation occurs at a particular site is given by random-hit kinetics (as seen in the Poisson distribution). So some sites will gain one, two, or three mutations, while others will not gain any. But some sites gain far more than the number of mutations expected from a random distribution; they may have 10× or even 100x more mutations than predicted by random hits. These sites are called hotspots. Spontaneous mutations may occur at hotspots; and different mutagens may have different hotspots.

1.14 Many hotspots result from modified bases

Key Concepts

- A common cause of hotspots is the modified base 5-methylcytosine, which is spontaneously deaminated to thymine.

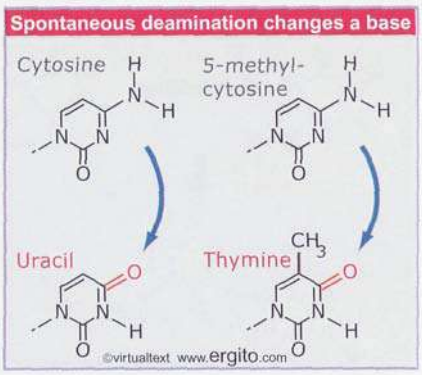


Figure 1.24 Deamination of cytosine produces uracil, whereas deamination of 5-methyl-cytosine produces thymine.

A major cause of spontaneous mutation results from the presence of an unusual base in the DNA. In addition to the four bases that are inserted into DNA when it is synthesized, modified bases are sometimes found. The name reflects their origin; they are produced by chemically modifying one of the four bases already present in DNA. The most common modified base is 5-methylcytosine, generated by a methylase enzyme that adds a methyl group to certain cytosine residues at specific sites in the DNA.

Sites containing 5-methylcytosine provide hotspots for spontaneous point mutation in *E. coli*. In each case, the mutation takes the form of a GC to AT transition. The hotspots are not found in strains of *E. coli* that cannot methylate cytosine.

The reason for the existence of the hotspots is that cytosine bases suffer spontaneous deamination at an appreciable frequency. In this reaction, the amino group is replaced by a keto group. Recall that deamination of cytosine generates uracil (see Figure 1.20). **Figure 1.24** compares this reaction with the deamination of 5-methylcytosine where deamination generates thymine. The effect in DNA is to generate the base pairs GU and GT, respectively, where there is a mismatch between the partners.

All organisms have repair systems that correct mismatched base pairs by removing and replacing one of the bases. The operation of these systems determines whether mismatched pairs such as GU and GT result in mutations.

Figure 1.25 shows that the consequences of deamination are different for 5-methylcytosine and cytosine. Deaminating the (rare) 5-methylcytosine causes a mutation, whereas deamination of the more common cytosine does not have this effect. This happens because the repair systems are much more effective in recognizing GU than G·T.

E. coli contains an enzyme, uracil-DNA-glycosidase, that removes uracil residues from DNA (see *15.22 Base flipping is used by methylases and glycosylases*). This action leaves an unpaired G residue, and a "repair system" then inserts a C base to partner it. The net result of these reactions is to restore the original sequence of the DNA. This system protects DNA against the consequences of spontaneous deamination of cytosine (although it is not active enough to prevent the effects of the increased level of deamination caused by nitrous acid; see Figure 1.20).

But the deamination of 5-methylcytosine leaves thymine. This creates" a mismatched base pair, G-T. If the mismatch is not corrected before the next replication cycle, a mutation results. At the next replication, the bases in the mispaired G-T partnership separate, and then they pair with new partners to produce one wild-type G·C pair and one mutant AT pair.

Deamination of 5-methylcytosine is the most common cause of production of G-T mismatched pairs in DNA. Repair systems that act on G-T mismatches have a bias toward replacing the T with a C (rather than the alternative of replacing the G with an A), which helps to reduce the rate of mutation (see *15.24 Controlling the direction of mismatch repair*). However, these systems are not as effective as the removal of U from GU mismatches. As a result, deamination of 5-methylcytosine leads to mutation much more often than does deamination of cytosine.

5-methylcytosine also creates hotspots in eukaryotic DNA. It is common at CpG dinucleotides that are concentrated in regions called CpG islands (see *21.19 CpG islands are regulatory targets*). Although 5-methylcytosine accounts for ~1% of the bases in human DNA, sites containing the modified base account for ~30% of all point mutations. This makes the state of 5-methylcytosine a particularly important determinant of mutation in animal cells.

The importance of repair systems in reducing the rate of mutation is emphasized by the effects of eliminating the mouse enzyme MBD4, a glycosylase that can remove T (or U) from mismatches with G. The result is to increase the mutation rate at CpG sites by a factor of 3x. (The reason the effect is not greater is that MBD4 is only one of several systems that act on G-T mismatches; we can imagine that elimination of all the systems would increase the mutation rate much more.)

The operation of these systems casts an interesting light on the use of T in DNA compared with U in RNA. Perhaps it relates to the need of DNA for stability of sequence; the use of T means that any deaminations of C are immediately recognized, because they generate a base (U) not usually present in the DNA. This greatly increases the efficiency with which repair systems can function (compared with the situation when they have to recognize G-T mismatches, which can be produced also by situations where removing the T would not be the appropriate response). Also, the phosphodiester bond of the backbone is more labile when the base is U.

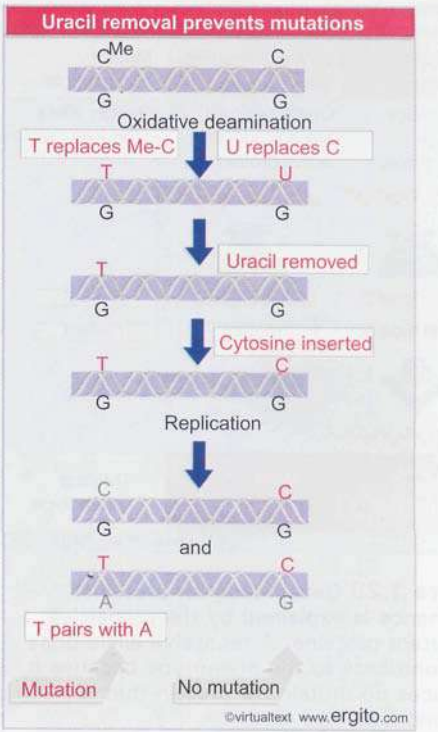


Figure 1.25 The deamination of 5-methylcytosine produces thymine (causing C·G to T·A transitions), while the deamination of cytosine produces uracil (which usually is removed and then replaced by cytosine).

1.15 A gene codes for a single polypeptide

Key Concepts

- The one gene: one enzyme hypothesis summarizes the basis of modern genetics: that a gene is a stretch of DNA coding for a single polypeptide chain.
- Most mutations damage gene function.

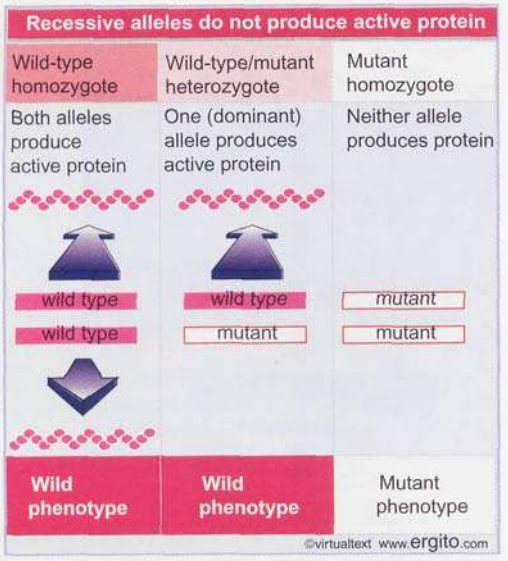


Figure 1.26 Genes code for proteins; dominance is explained by the properties of mutant proteins. A recessive allele does not contribute to the phenotype because it produces no protein (or protein that is nonfunctional).

The first systematic attempt to associate genes with enzymes showed that each stage in a metabolic pathway is catalyzed by a single enzyme and can be blocked by mutation in a different gene. This led to the *one gene: one enzyme hypothesis*. Each metabolic step is catalyzed by a particular enzyme, whose production is the responsibility of a single gene. A mutation in the gene alters the activity of the protein for which it is responsible.

A modification in the hypothesis is needed to accommodate proteins that consist of more than one subunit. If the subunits are all the same, the protein is a homomultimer, represented by a single gene. If the subunits are different, the protein is a heteromultimer. Stated as a more general rule applicable to any heteromultimeric protein, the one gene: one enzyme hypothesis becomes more precisely expressed as **one gene: one polypeptide chain.**

Identifying which protein represents a particular gene can be a protracted task. The mutation responsible for creating Mendel's wrinkled-pea mutant was identified only in 1990 as an alteration that inactivates the gene for a starch branching enzyme!

It is important to remember that a gene does not directly generate a protein. As shown previously in Figure 1.2, a gene codes for an RNA, which may in turn code for a protein. Most genes code for proteins, but some genes code for RNAs that do not give rise to proteins. These RNAs may be structural components of the apparatus responsible for synthesizing proteins or may have roles in regulating gene expression. The basic principle is that the gene is a sequence of DNA that specifies the sequence of an independent product. The process of gene expression may terminate in a product that is either RNA or protein.

A mutation is a random event with regard to the structure of the gene, so the greatest probability is that it will damage or even abolish gene function. Most mutations that affect gene function are recessive: *they represent an absence of function, because the mutant gene has been prevented from producing its usual protein.* **Figure 1.26** illustrates the relationship between recessive and wild-type alleles. When a heterozygote contains one wild-type allele and one mutant allele, the wild-type allele is able to direct production of the enzyme. The wild-type allele is therefore dominant. (This assumes that an adequate *amount* of protein is made by the single wild-type allele. When this is not true, the smaller amount made by one allele as compared to two alleles results in the intermediate phenotype of a partially dominant allele in a heterozygote.)

1.16 Mutations in the same gene cannot complement

Key Concepts

- A mutation in a gene affects only the protein coded by the mutant copy of the gene, and does not affect the protein coded by any other allele.
- Failure of two mutations to complement (produce wild-phenotype) when they are present in *trans* configuration in a heterozygote means that they are part of the same gene.

How do we determine whether two mutations that cause a similar phenotype lie in the same gene? If they map close together, they may be alleles. However, they could also represent mutations in two *dif-*

ferent genes whose proteins are involved in the same function. The complementation test is used to determine whether two mutations lie in the same gene or in different genes. The test consists of making a heterozygote for the two mutations (by mating parents homozygous for each mutation).

If the mutations lie in the same gene, the parental genotypes can be represented as

$$\frac{m_1}{m_1} \text{ and } \frac{m_2}{m_2}$$

The first parent provides an m_1 mutant allele and the second parent provides an m_2 allele, so that the heterozygote has the constitution

$$\frac{m_1}{m_2}$$

No wild-type gene is present, so the heterozygote has mutant phenotype.

If the mutations lie in different genes, the parental genotypes can be represented as

$$\frac{m_1 +}{m_1 +} \text{ and } \frac{+ m_2}{+ m_2}$$

Each chromosome has a wild-type copy of one gene (represented by the plus sign) and a mutant copy of the other. Then the heterozygote has the constitution

$$\frac{m_1 +}{+ m_2}$$

in which the two parents between them have provided a wild-type copy of each gene. The heterozygote has wild phenotype; the two genes are said to complement.

The complementation test is shown in more detail in **Figure 1.27.** The basic test consists of the comparison shown in the top part of the figure. If two mutations lie in the same gene, we see a difference in the phenotypes of the *trans* configuration and the *cis* configuration. The *trans* configuration is mutant, because each allele has a (different) mutation. But the *cis* configuration is wild-type, because one allele has two mutations but the other allele has no mutations. The lower part of the figure shows that if the two mutations lie in different genes, we always see a wild phenotype. There is always one wild-type and one mutant allele of each gene, and the configuration is irrelevant.

Failure to complement means that two mutations are part of the *same* genetic unit. Mutations that do not complement one another are said to comprise part of the same complementation group. Another term that is used to describe the unit defined by the complementation test is the cistron. This is the same as the gene. Basically these three terms all describe a stretch of DNA that functions as a unit to give rise to an RNA or protein product. The properties of the gene with regards to complementation are explained by the fact that this product is a single molecule that behaves as a functional unit.

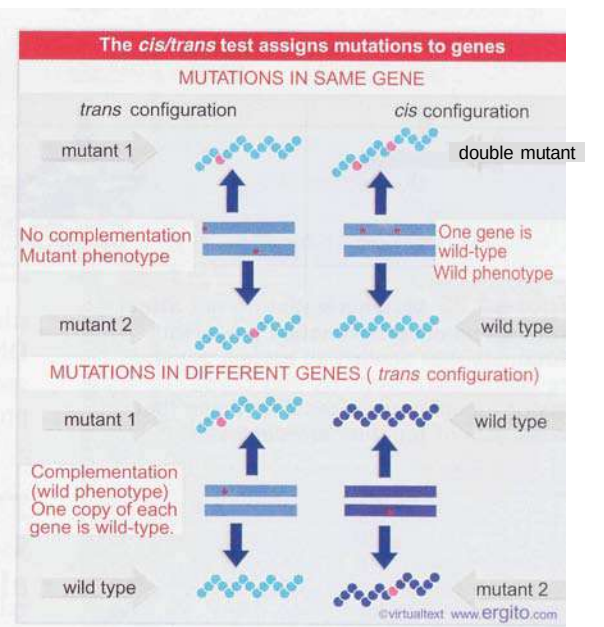


Figure 1.27 The cistron is defined by the complementation test. Genes are represented by bars; red stars identify sites of mutation.

1.17 Mutations may cause loss-of-function or gain-of-function

Key Concepts

- Recessive mutations are due to loss-of-function by the protein product.
- Dominant mutations result from a gain-of-function.
- Testing whether a gene is essential requires a null mutation (one that completely eliminates its function).
- Silent mutations have no effect, either because the base change does not change the sequence or amount of protein, or because the change in protein sequence has no effect.
- Leaky mutations do affect the function of the gene product, but are not revealed in the phenotype because sufficient activity remains.

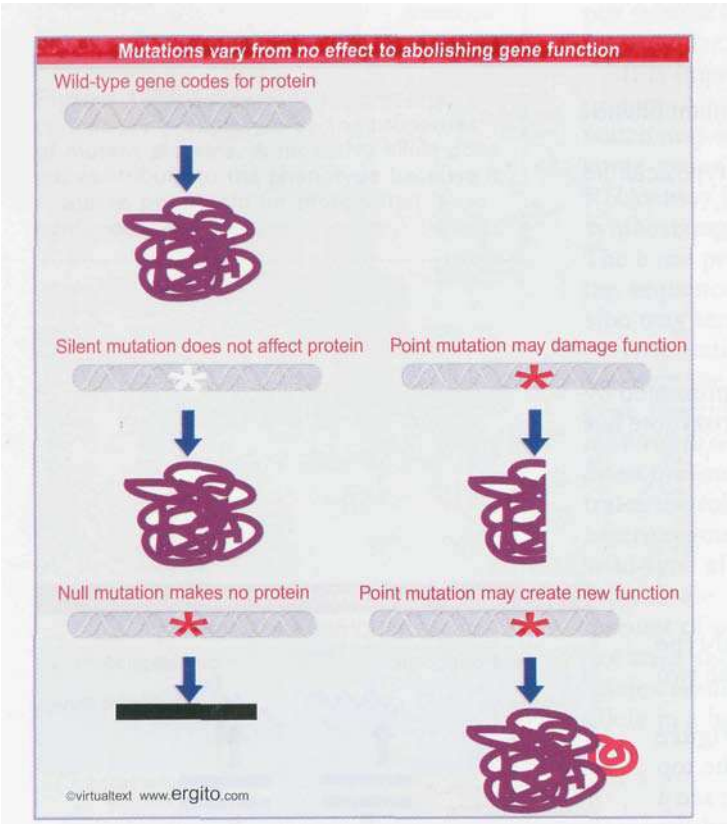


Figure 1.28 Mutations that do not affect protein sequence or function are silent. Mutations that abolish all protein activity are null. Point mutations that cause loss-of-function are recessive; those that cause gain-of-function are dominant.

The various possible effects of mutation in a gene are summarized in **Figure 1.28.**

When a gene has been identified, insight into its function in principle can be gained by generating a mutant organism that entirely lacks the gene. A mutation that completely eliminates gene function, usually because the gene has been deleted, is called a null mutation. If a gene is essential, a null mutation is lethal.

To determine what effect a gene has upon the phenotype, it is essential to characterize a null mutant. When a mutation fails to affect the phenotype, it is always possible that this is because it is a leaky mutation—enough active product is made to fulfill its function, even though the activity is quantitatively reduced or qualitatively different from the wild type. But if a null mutant fails to affect a phenotype, we may safely conclude that the gene function is not necessary.

Null mutations, or other mutations that impede gene function (but do not necessarily abolish it entirely) are called loss-of-function mutations. A loss-of-function mutation is recessive (as in the example of Figure 1.26). Sometimes a mutation has the opposite effect and causes a protein to acquire a new function; such a change is called a gain-of-function mutation. A gain-of-function mutation is dominant.

Not all mutations in DNA lead to a detectable change in the phenotype. Mutations without apparent effect are called silent mutations. They fall into two types. Some involve base changes in DNA that do not cause any change in the amino acid present in the corresponding protein. Others change the amino acid, but the replacement in the protein does not affect its activity; these are called neutral substitutions.

1.18 A locus may have many different mutant alleles

Key Concepts

- The existence of multiple alleles allows heterozygotes to occur representing any pairwise combination of alleles.

By Book_Crazy [IND]

If a recessive mutation is produced by every change in a gene that prevents the production of an active protein, there should be a large number of such mutations in any one gene. Many amino acid replacements may change the structure of the protein sufficiently to impede its function.

Different variants of the same gene are called multiple alleles, and their existence makes it possible to create a heterozygote between mutant alleles. The relationship between these multiple alleles takes various forms.

In the simplest case, a wild-type gene codes for a protein product that is functional. Mutant allele(s) code for proteins that are nonfunctional.

But there are often cases in which a series of mutant alleles have different phenotypes. For example, wild-type function of the *white* locus of *D. melanogaster* is required for development of the normal red color of the eye. The locus is named for the effect of extreme (null) mutations, which cause the fly to have a white eye in mutant homozygotes.

To describe wild-type and mutant alleles, wild genotype is indicated by a plus superscript after the name of the locus (w^+ is the wild-type allele for [red] eye color in *D. melanogaster).* Sometimes + is used by itself to describe the wild-type allele, and only the mutant alleles are indicated by the name of the locus.

An entirely defective form of the gene (or absence of phenotype) may be indicated by a minus superscript. To distinguish among a variety of mutant alleles with different effects, other superscripts may be introduced such as w' or w^a.

The w^+ allele is dominant over any other allele in heterozygotes. There are many different mutant alleles. **Figure 1.29** shows a (small) sample. Although some alleles have no eye color, many alleles produce some color. Each of these mutant alleles must therefore represent a different mutation of the gene, which does not eliminate its function entirely, but leaves a residual activity that produces a characteristic phenotype. These alleles are named for the color of the eye in a homozygote. (Most *w* alleles affect the quantity of pigment in the eye, and the examples in the figure are arranged in [roughly] declining amount of color, but others, such as w^{sp}, affect the pattern in which it is deposited.)

When multiple alleles exist, an animal may be a heterozygote that carries two different mutant alleles. The phenotype of such a heterozygote depends on the nature of the residual activity of each allele. The relationship between two mutant alleles is in principle no different from that between wild-type and mutant alleles: one allele may be dominant, there may be partial dominance, or there may be codominance.

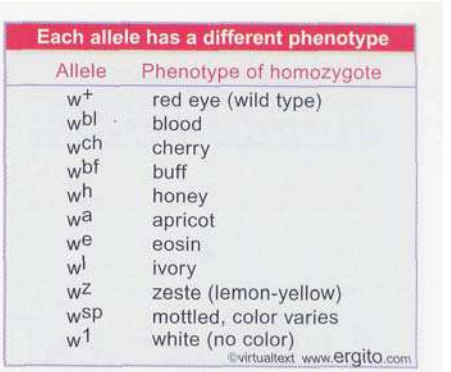


Figure 1.29 The *w* locus has an extensive series of alleles, whose phenotypes extend from wild-type (red) color to complete lack of pigment.

1.19 A locus may have more than one wild-type allele

Key Concepts

- A locus may have a polymorphic distribution of alleles, with no individual allele that can be considered to be the sole wild-type.

There is not necessarily a unique wild-type allele at any particular locus. Control of the human blood group system provides an example. Lack of function is represented by the null type, *O* group. But the functional alleles *A* and *B* provide activities that are codominant with one another and dominant over *O* group. The basis for this relationship is illustrated in **Figure 1.30.**

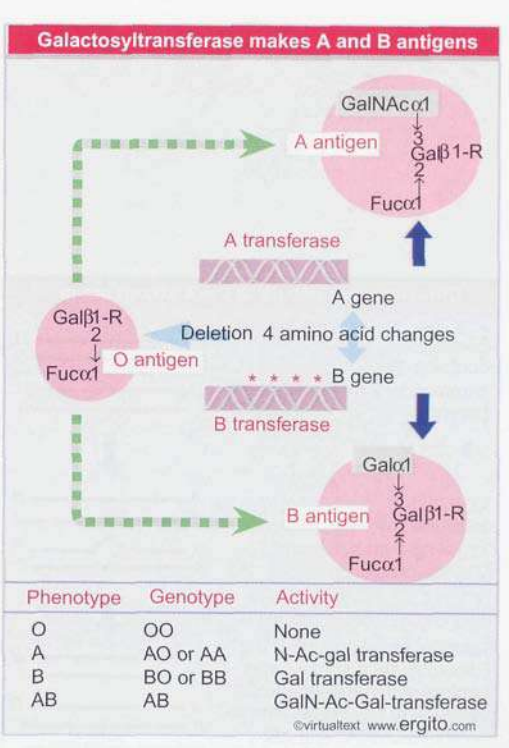


Figure 1.30 The ABO blood group locus codes for a galactosyltransferase whose specificity determines the blood group.

The O (or H) antigen is generated in all individuals, and consists of a particular carbohydrate group that is added to proteins. The *ABO* locus codes for a galactosyltransferase enzyme that adds a further sugar group to the O antigen. The specificity of this enzyme determines the blood group. The *A* allele produces an enzyme that uses the cofactor UDP-N-acetylgalactose, creating the A antigen. The *B* allele produces an enzyme that uses the cofactor UDP-galactose, creating the B antigen. The A and B versions of the transferase protein differ in 4 amino acids that presumably affect its recognition of the type of cofactor. The *O* allele has a mutation (a small deletion) that eliminates activity, so no modification of the O antigen occurs.

This explains why *A* and *B* alleles are dominant in the *AO* and *BO* heterozygotes: the corresponding transferase activity creates the A or B antigen. The *A* and *B* alleles are codominant in *AB* heterozygotes, because both transferase activities are expressed. The *OO* homozygote is a null that has neither activity, and therefore lacks both antigens.

Neither *A* nor *B* can be regarded as uniquely wild type, since they represent alternative activities rather than loss or gain of function. A situation such as this, in which there are multiple functional alleles in a population, is described as a polymorphism (see 3.5 *Individual genomes show extensive variation*).

1.20 Recombination occurs by physical exchange of DNA

Key Concepts

- Recombination is the result of crossing-over that occurs at chiasmata and involves two of the four chromatids.
- Recombination occurs by a breakage and reunion that proceeds via an intermediate of hybrid DNA.

Genetic recombination describes the generation of new combinations of alleles that occurs at each generation in diploid organisms. The two copies of each chromosome may have different alleles at some loci. By exchanging corresponding parts between the chromosomes, recombinant chromosomes can be generated that are different from the parental chromosomes.

Recombination results from a physical exchange of chromosomal material. This is visible in the form of the crossing-over that occurs during meiosis (the specialized division that produces haploid germ cells). Meiosis starts with a cell that has duplicated its chromosomes, so that it has four copies of each chromosome. Early in meiosis, all four copies are closely associated (synapsed) in a structure called a bivalent. Each individual chromosomal unit is called a chromatid at this stage. Pairwise exchanges of material occur between the chromatids.

The visible result of a crossing-over event is called a chiasma, and is illustrated diagrammatically in **Figure 1.31.** A chiasma represents a site at which two of the chromatids in a bivalent have been broken at corresponding points. The broken ends have been rejoined crosswise, generating new chromatids. Each new chromatid consists of material derived from one chromatid on one side of the junction point, with material from the other chromatid on the opposite side. The two recombinant chromatids have reciprocal structures. The event is described as a breakage and reunion. Its nature explains why a single recombination event can produce only 50% recombinants: each individual recombination event involves only two of the four associated chromatids.

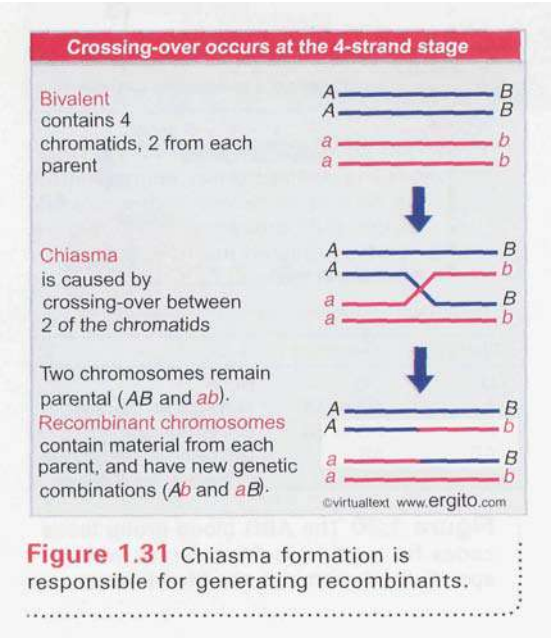


Figure 1.31 Chiasma formation is responsible for generating recombinants.

The complementarity of the two strands of DNA is essential for the recombination process. Each of the chromatids shown in Figure 1.31 consists of a very long duplex of DNA. For them to be broken and reconnected without any loss of material requires a mechanism to recognize exactly corresponding positions. This is provided by complementary base pairing.

Recombination involves a process in which the single strands in the region of the crossover exchange their partners. **Figure 1.32** shows that this creates a stretch of hybrid DNA in which the single strand of one duplex is paired with its complement from the other duplex. The mechanism of course involves other stages (strands must be broken and resealed), and we discuss this in more detail in 75 *Recombination and repair,* but the crucial feature that makes precise recombination possible is the complementarity of DNA strands. The figure shows only some stages of the reaction, but we see that a stretch of hybrid DNA forms in the recombination intermediate when a single strand crosses over from one duplex to the other. Each recombinant consists of one parental duplex DNA at the left, connected by a stretch of hybrid DNA to the other parental duplex at the right. Each duplex DNA corresponds to one of the chromatids involved in recombination in Figure 1.31.

The formation of hybrid DNA requires the sequences of the two recombining duplexes to be close enough to allow pairing between the complementary strands. If there are no differences between the two parental genomes in this region, formation of hybrid DNA will be perfect. But the reaction can be tolerated even when there are small differences. In this case, the hybrid DNA has points of mismatch, at which a base in one strand faces a base in the other strand that is not complementary to it. The correction of such mismatches is another feature of genetic recombination (see *15 Recombination and repair).*

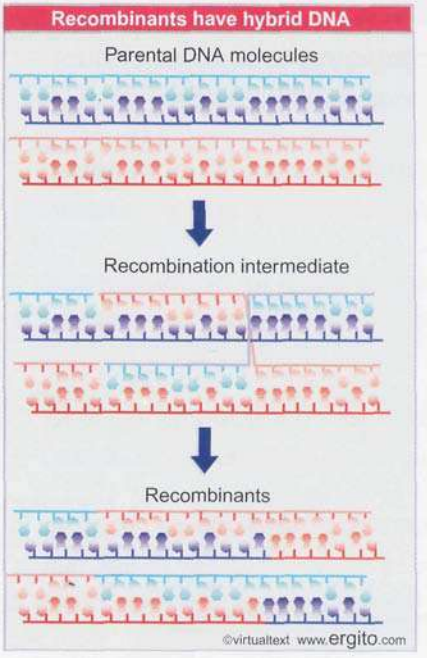


Figure 1.32 Recombination involves pairing between complementary strands of the two parental duplex DNAs.

1.21 The genetic code is triplet

Key Concepts

- The genetic code is read in triplet nucleotides called codons.
- The triplets are nonoverlapping and are read from a fixed starting point.
- Mutations that insert or delete individual bases cause a shift in the triplet sets after the site of mutation.
- Combinations of mutations that together insert or delete 3 bases (or multiples of three) insert or delete amino acids but do not change the reading of the triplets beyond the last site of mutation.

Each gene represents a particular protein chain. The concept that each protein consists of a particular series of amino acids dates from Sanger's characterization of insulin in the 1950s. The discovery that a gene consists of DNA faces us with the issue of how a sequence of nucleotides in DNA represents a sequence of amino acids in protein.

A crucial feature of the general structure of DNA is that *it is independent of the particular sequence of its component nucleotides.* The sequence of nucleotides in DNA is important not because of its structure *per se,* but because it *codes* for the sequence of amino acids that constitutes the corresponding polypeptide. The relationship between a sequence of DNA and the sequence of the corresponding protein is called the genetic code.

The structure and/or enzymatic activity of each protein follows from its primary sequence of amino acids. By determining the sequence of amino acids in each protein, the gene is able to carry all the information needed to specify an active polypeptide chain. In this way, a single type of structure—the gene—is able to represent itself in innumerable polypeptide forms.

Together the various protein products of a cell undertake the catalytic and structural activities that are responsible for establishing its phenotype. Of course, in addition to sequences that code for proteins, DNA also contains certain sequences whose function is to be recognized by regulator molecules, usually proteins. Here the function of the DNA is determined by its sequence directly, not via any intermediary code. Both types of regions, genes expressed as proteins and sequences recognized as such, constitute genetic information.

The genetic code is deciphered by a complex apparatus that interprets the nucleic acid sequence. This apparatus is essential if the information carried in DNA is to have meaning. In any given region, only one of the two strands of DNA codes for protein, so we write the genetic code as a sequence of bases (rather than base pairs).

The genetic code is read in groups of three nucleotides, each group representing one amino acid. Each trinucleotide sequence is called a codon. A gene includes a series of codons that is read sequentially from a starting point at one end to a termination point at the other end. Written in the conventional 5′→3′ direction, the nucleotide sequence of the DNA strand that codes for protein corresponds to the amino acid sequence of the protein written in the direction from N-terminus to C-terminus.

The genetic code is read in *nonoverlapping triplets from a fixed starting point:*

- Nonoverlapping implies that each codon consists of three nucleotides and that successive codons are represented by successive trinucleotides.
- The use of a *fixed starting point* means that assembly of a protein must start at one end and work to the other, so that different parts of the coding sequence cannot be read independently.

The nature of the code predicts that two types of mutations will have different effects. If a particular sequence is read sequentially, such as

```
UUU AAA GGG CCC (codons)
aa1 aa2 aa3 aa4 (amino acids)
```

then a point mutation will affect only one amino acid. For example, because only the second codon has been changed, the substitution of an A by some other base (X) causes aa2 to be replaced by aa5:

```
UUU AAX GGG CCC
aal aa5 aa3 aa4
```

But a mutation that inserts or deletes a single base will change the triplet sets for the entire subsequent sequence. A change of this sort is called a frameshift. An insertion might take the following form:

```
UUU AAX AGG GCC C
aal aa5 aa6 aa7
```

Because the new sequence of triplets is completely different from the old one, the entire amino acid sequence of the protein is altered beyond the site of mutation. So the function of the protein is likely to be lost completely.

Frameshift mutations are induced by the acridines, compounds that bind to DNA and distort the structure of the double helix, causing additional bases to be incorporated or omitted during replication. Each mutagenic event sponsored by an acridine results in the addition or removal of a single base pair.

If an acridine mutant is produced by, say, addition of a nucleotide, it should revert to wild type by deletion of the nucleotide. But reversion can also be caused by deletion of a different base, at a site close to the first. Combinations of such mutations provided revealing evidence about the nature of the genetic code.

Figure 1.33 illustrates the properties of frameshift mutations. An insertion or a deletion changes the entire protein sequence following the site of mutation. But the combination of an insertion and a deletion causes the code to be read incorrectly only between the two sites of mutation; correct reading resumes after the second site.

Genetic analysis of acridine mutations in the *rII* region of the phage T6 in 1961 showed that all the mutations could be classified into one of two sets, described as (+) and (-). Either type of mutation by itself causes a frameshift, the (+) type by virtue of a base addition, the (-) . type by virtue of a base deletion. Double mutant combinations of the types (+ +) and (—) continue to show mutant behavior. But combinations of the types (+ -) or (- +) suppress one another, giving rise to a description in which one mutation is described as a suppressor of the other. (In the context of this work, "suppressor" is used in an unusual sense, because the second mutation is in the same gene as the first.)

These results show that the genetic code must be read as a sequence that is fixed by the starting point, so additions or deletions compensate for each other, whereas double additions or double deletions remain mutant. But this does not reveal how many nucleotides make up each codon.

When triple mutants are constructed, only (+ + +) and (———) combinations show the wild phenotype, while other combinations remain mutant. If we take three additions or three deletions to correspond respectively to the addition or omission overall of a single amino acid, this implies that the code is read in triplets. An incorrect amino acid sequence is found between the two outside sites of mutation, and the sequence on either side remains wild type, as indicated in Figure 1.33.

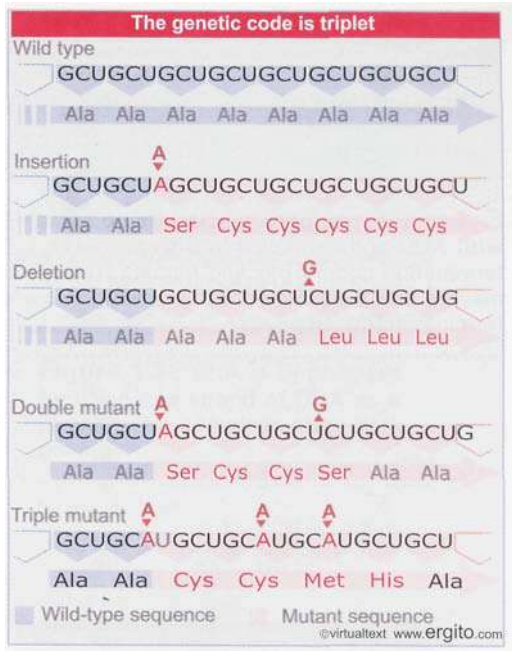


Figure 1.33 Frameshift mutations show that the genetic code is read in triplets from a fixed starting point.

1.22 Every sequence has three possible reading frames

Key Concepts

- Usually only one reading frame is translated and the other two are blocked by frequent termination signals.

If the genetic code is read in nonoverlapping triplets, there are three possible ways of translating any nucleotide sequence into protein, depending on the starting point. These called reading frames. For the sequence

A C G A C G A C G A C G A C G A C G

the three possible reading frames are

ACG ACG ACG ACG ACG ACG ACG
CGA CGA CGA CGA CGA CGA CGA
GAC GAC GAC GAC GAC GAC GAC

A reading frame that consists exclusively of triplets representing amino acids is called an open reading frame or ORF. A sequence that is translated into protein has a reading frame that starts with a special initiation codon (AUG) and that extends through a series of triplets representing amino acids until it ends at one of three types of termination codon (see *5 Messenger RNA).*

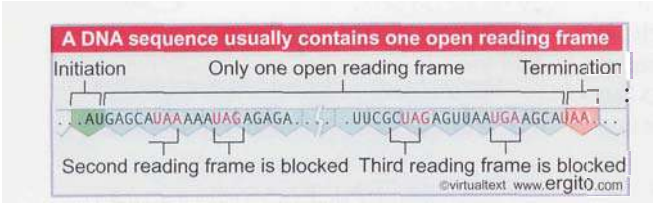


Figure 1.34 An open reading frame starts with AUG and continues in triplets to a termination codon. Blocked reading frames may be interrupted frequently by termination codons.

A reading frame that cannot be read into protein because termination codons occur frequently is said to be blocked. If a sequence is blocked in all three reading frames, it cannot have the function of coding for protein.

When the sequence of a DNA region of unknown function is obtained, each possible reading frame is analyzed to determine whether it is open or blocked. Usually no more than one of the three possible frames of reading is open in any single stretch of DNA. **Figure 1.34** shows an example of a sequence that can be read in only one reading frame, because the alternative reading frames are blocked by frequent termination codons. A long open reading frame is unlikely to exist by chance; if it were not translated into protein, there would have been no selective pressure to prevent the accumulation of termination codons. So the identification of a lengthy open reading frame is taken to be *prima facie* evidence that the sequence is translated into protein in that frame. An open reading frame (ORF) for which no protein product has been identified is sometimes called an unidentified reading frame (URF).

1.23 Prokaryotic genes are colinear with their proteins

Key Concepts

- A prokaryotic gene consists of a continuous length of 3N nucleotides that codes for N amino acids.
- The gene, mRNA, and protein are all colinear.

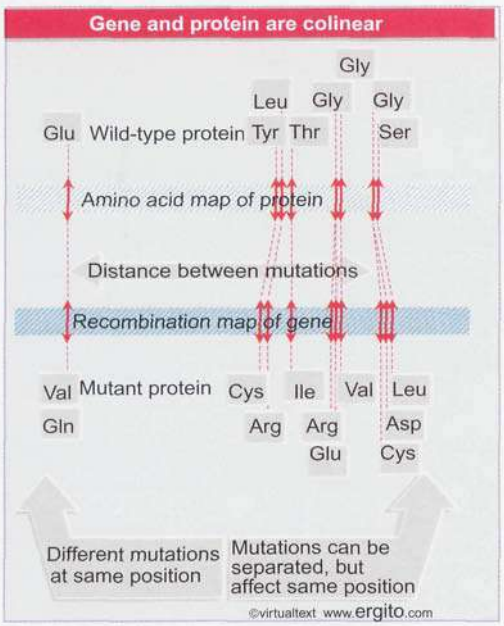


Figure 1.35 The recombination map of the tryptophan synthetase gene corresponds with the amino acid sequence of the protein.

By comparing the nucleotide sequence of a gene with the amino acid sequence of a protein, we can determine directly whether the gene and the protein are colinear: whether the sequence of nucleotides in the gene corresponds exactly with the sequence of amino acids in the protein. In bacteria and their viruses, there is an exact equivalence. Each gene contains a continuous stretch of DNA whose length is directly related to the number of amino acids in the protein that it represents. A gene of $3N$ bp is required to code for a protein of N amino acids, according to the genetic code.

The equivalence of the bacterial gene and its product means that a physical map of DNA will exactly match an amino acid map of the protein. How well do these maps fit with the recombination map?

The colinearity of gene and protein was originally investigated in the tryptophan synthetase gene of *E. coli*. Genetic distance was measured by the percent recombination between mutations; protein distance was measured by the number of amino acids separating sites of replacement. **Figure 1.35** compares the two maps. The order of seven sites of mutation is the same as the order of the corresponding sites of amino acid replacement. And the recombination distances are relatively similar to the actual distances in the protein. The recombination map expands the distances between some mutations, but otherwise there is little distortion of the recombination map relative to the physical map.

The recombination map makes two further general points about the organization of the gene. Different mutations may cause a wild-type amino acid to be replaced with different substituents. If two such mutations cannot recombine, they must involve different point mutations at the same position in DNA. If the mutations can be separated on the genetic map, but affect the same amino acid on the upper map (the con-

necting lines converge in the figure), they must involve point mutations at different positions that affect the same amino acid. This happens because the unit of genetic recombination (actually 1 bp) is smaller than the unit coding for the amino acid (actually 3 bp).

1.24 Several processes are required to express the protein product of a gene

Key Concepts

- A prokaryotic gene is expressed by transcription into mRNA and then by translation of the mRNA into protein.
- In eukaryotes, a gene may contain internal regions that are not represented in protein.
- Internal regions are removed from the RNA transcript by RNA splicing to give an mRNA that is colinear with the protein product.
- Each mRNA consists of a nontranslated 5′ leader, a coding region, and a nontranslated 3′ trailer.

In comparing gene and protein, we are restricted to dealing with the sequence of DNA stretching between the points corresponding to the ends of the protein. However, a gene is not directly translated into protein, but is expressed via the production of a messenger RNA (abbreviated to mRNA), a nucleic acid intermediate actually used to synthesize a protein (as we see in detail in *5 Messenger RNA).*

Messenger RNA is synthesized by the same process of complementary base pairing used to replicate DNA, with the important difference that it corresponds to only one strand of the DNA double helix. **Figure 1.36** shows that the sequence of messenger RNA is complementary with the sequence of one strand of DNA and is identical (apart from the replacement of T with U) with the other strand of DNA. The convention for writing DNA sequences is that the top strand runs 5′→3′, with the sequence that is the same as RNA.

The process by which a gene gives rise to a protein is called gene expression. In bacteria, it consists of two stages. The first stage is transcription, when an mRNA copy of one strand of the DNA is produced. The second stage is translation of the mRNA into protein. This is the process by which the sequence of an mRNA is read in triplets to give the series of amino acids that make the corresponding protein.

A messenger RNA includes a sequence of nucleotides that corresponds with the sequence of amino acids in the protein. This part of the nucleic acid is called the coding region. But the messenger RNA includes additional sequences on either end; these sequences do not directly represent protein. The 5' nontranslated region is called the leader, and the 3' nontranslated region is called the trailer.

The *gene* includes the entire sequence represented in messenger RNA. Sometimes mutations impeding gene function are found in the additional, noncoding regions, confirming the view that these comprise a legitimate part of the genetic unit.

Figure 1.37 illustrates this situation, in which the gene is considered to comprise a continuous stretch of DNA, needed to produce a particular protein. It includes the sequence coding for that protein, but also includes sequences on either side of the coding region.

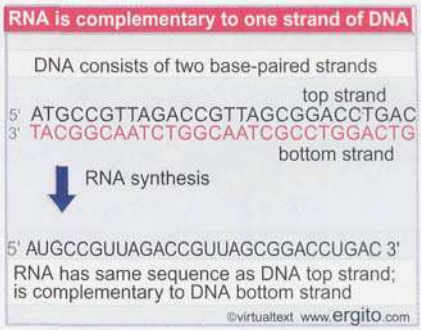


Figure 1.36 RNA is synthesized by using one strand of DNA as a template for complementary base pairing.

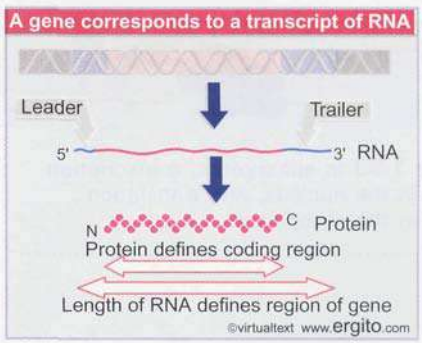


Figure 1.37 The gene may be longer than the sequence coding for protein.

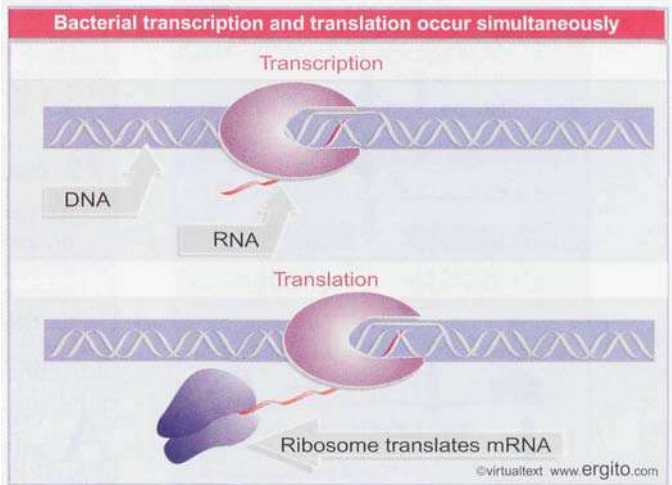


Figure 1.38 Transcription and translation take place in the same compartment in bacteria.

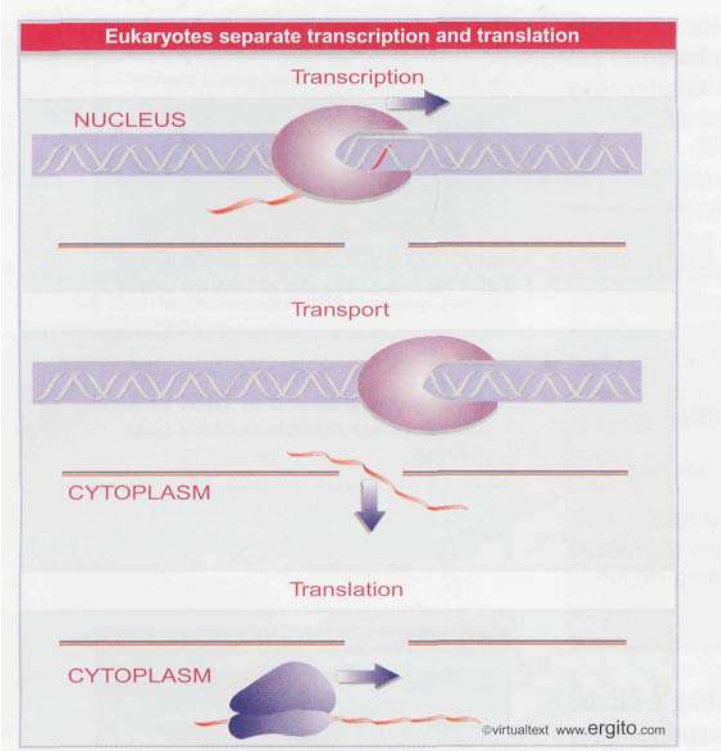


Figure 1.39 In eukaryotes, transcription occurs in the nucleus, and translation occurs in the cytoplasm.

A bacterium consists of only a single compartment, so transcription and translation occur in the same place, as illustrated in **Figure 1.38.**

In eukaryotes transcription occurs in the nucleus, but the RNA product must be transported to the cytoplasm in order to be translated, as shown in **Figure 1.39.** For the simplest eukaryotic genes (just like in bacteria) the transcript RNA is in fact the mRNA. But for more complex genes, the immediate transcript of the gene is a pre-mRNA that requires processing to generate the mature mRNA. The basic stages of gene expression in a eukaryote are outlined in **Figure 1.40.**

The most important stage in processing is RNA splicing. Many genes in eukaryotes (and a majority in higher eukaryotes) contain internal regions that do not code for protein. The process of splicing removes these regions from the pre-mRNA to generate an RNA that has a continuous open reading frame (see Figure 2.1). Other processing events that occur at this stage involve the modification of the 5' and 3' ends of the pre-mRNA (see Figure 5.16).

Translation is accomplished by a complex apparatus that includes both protein and RNA components. The actual "machine" that undertakes the process is the *ribosorne,* a large complex that includes some large RNAs *(ribosomal RNAs,* abbreviated to *rRNAs)* and many small proteins. The process of recognizing which amino acid corresponds to a particular nucleotide triplet requires an intermediate *transfer RNA* (abbreviated to *tRNA);* there is at least one tRNA species for every amino acid. Many ancillary proteins are involved. We describe translation in *5 Messenger RNA,* but note for now that the ribosomes are the large structures in Figure 1.38 and Figure 1.39 that move along the mRNA.

The important point to note at this stage is that the process of gene expression involves RNA not only as the essential substrate, but also in providing components of the apparatus. The rRNA and tRNA components are coded by genes and are generated by the process of transcription (just like mRNA, except that there is no subsequent stage of translation).

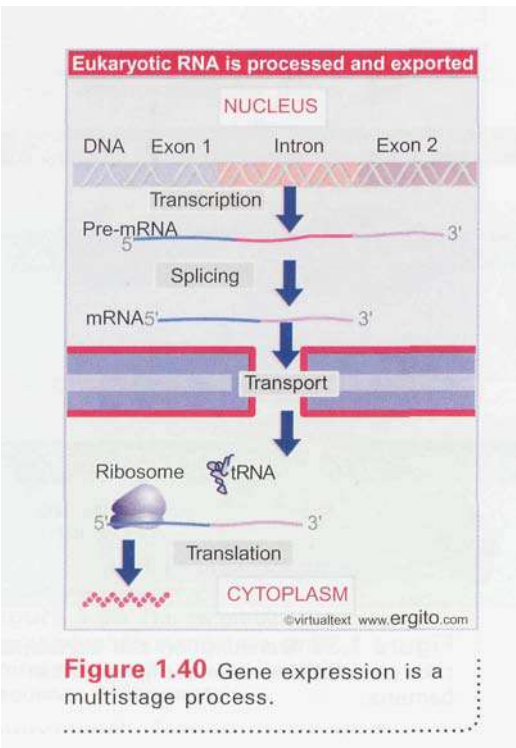


Figure 1.40 Gene expression is a multistage process.

1.25 Proteins are trans-acting but sites on DNA are *cis*-acting

Key Concepts

- All gene products (RNA or proteins) are frans-acting. They can act on any copy of a gene in the cell.
- c/s-acting mutations identify sequences of DNA that are targets for recognition by frans-acting products. They are not expressed as RNA or protein and affect only the contiguous stretch of DNA.

A crucial step in the definition of the gene was the realization that all its parts must be present on one contiguous stretch of DNA. In genetic terminology, sites that are located on the same DNA are said to be in *cis.* Sites that are located on two different molecules of DNA are described as being in *trans.* So two mutations may be in *cis* (on the same DNA) or in *trans* (on different DNAs). The complementation test uses this concept to determine whether two mutations are in the same

gene (see Figure 1.27 in *1.16 Mutations in the same gene cannot complement).* We may now extend the concept of the difference between *cis* and *trans* effects from defining the coding region of a gene to describing the interaction between regulatory elements and a gene.

Suppose that the ability of a gene to be expressed is controlled by a protein that binds to the DNA close to the coding region. In the example depicted in **Figure 1.41,** messenger RNA can be synthesized only when the protein is bound to the DNA. Now suppose that a mutation occurs in the DNA sequence to which this protein binds, so that the protein can no longer recognize the DNA. As a result, the DNA can no longer be expressed.

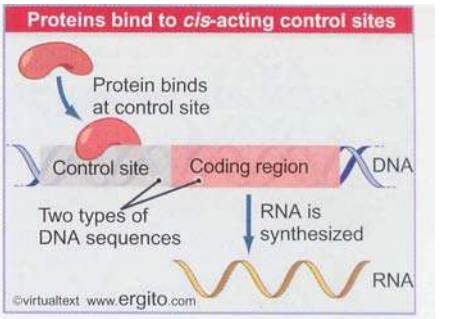


Figure 1.41 Control sites in DNA provide binding sites for proteins; coding regions are expressed via the synthesis of RNA.

So a gene can be inactivated either by a mutation in a control site or by a mutation in a coding region. The mutations cannot be distinguished genetically, because both have the property of acting only on the DNA sequence of the single allele in which they occur. They have identical properties in the complementation test, and a mutation in a control region is therefore defined as comprising part of the gene in the same way as a mutation in the coding region.

Figure 1.42 shows that a deficiency in the control site *affects only the coding region to which it is connected; it does not affect the ability of the other allele to be expressed.* A mutation that acts solely by affecting the properties of the contiguous sequence of DNA is called cis-acting.

We may contrast the behavior of the *cis-acting* mutation shown in Figure 1.42 with the result of a mutation in the gene coding for the regulator protein. **Figure 1.43** shows that the absence of regulator protein would prevent *both* alleles from being expressed. A mutation of this sort is said to be *trans-acting.*

Reversing the argument, if a mutation is *trans-acting,* we know that its effects must be exerted through some diffusible product (typically a protein) that acts on multiple targets within a cell. But if a mutation is ris-acting, it must function via affecting directly the properties of the contiguous DNA, which means that it is *not expressed in the form of RNA or protein.*

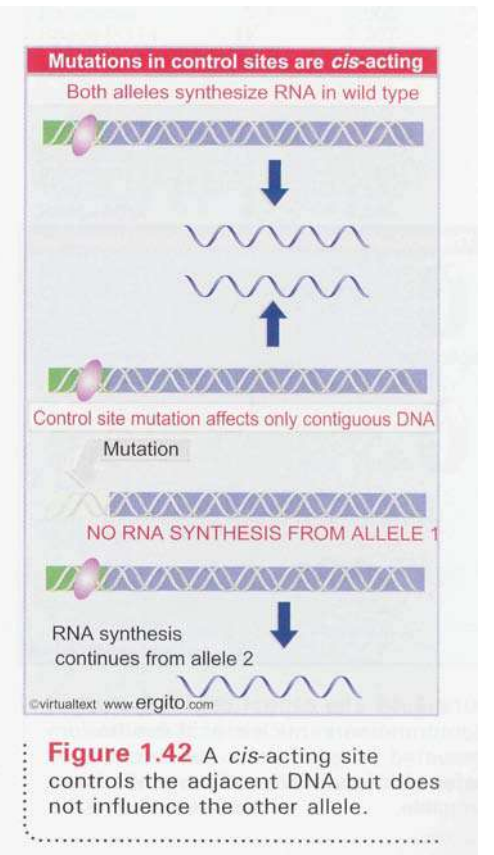
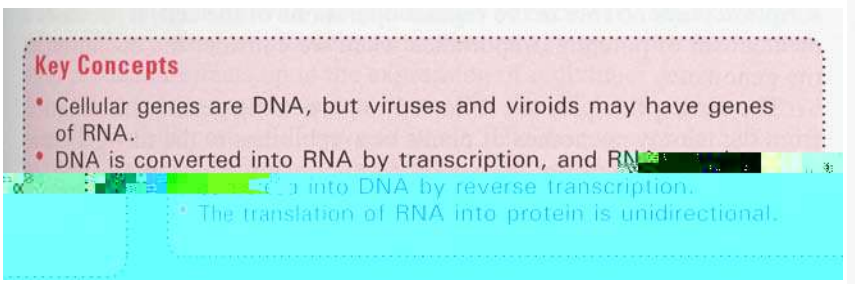


Figure 1.42 A *cis*-acting site controls the adjacent DNA but does not influence the other allele.

1.26 Genetic information can be provided by DNA or RNA



The central dogma defines the paradigm of molecular biology. Genes are perpetuated as sequences of nucleic acid, but function by being expressed in the form of proteins. Replication is responsible for the inheritance of genetic information. Transcription and translation are responsible for its conversion from one form to another.

Figure 1.44 illustrates the roles of replication, transcription, and translation, viewed from the perspective of the central dogma:

- *The perpetuation of nucleic acid may involve either DNA or RNA as the genetic material.* Cells use only DNA. Some viruses use RNA, and replication of viral RNA occurs in the infected cell.

By Book_Crazy [IND]

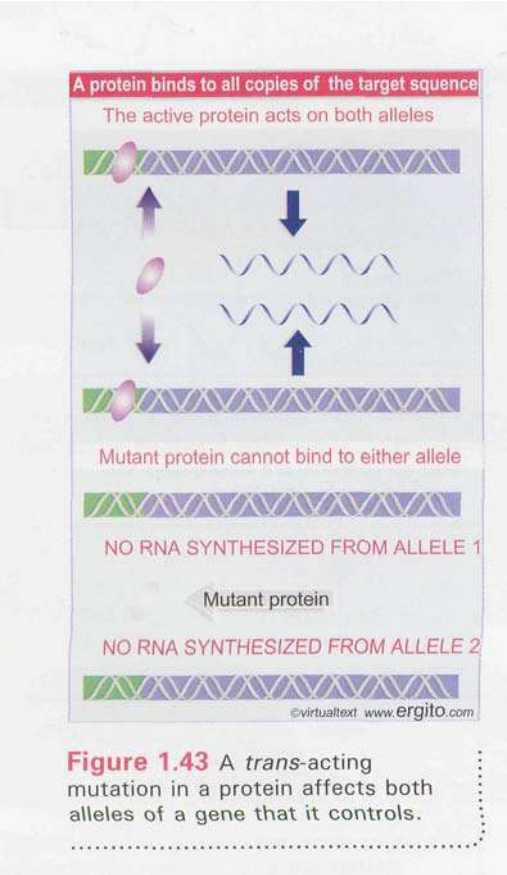


Figure 1.43 A *trans*-acting mutation in a protein affects both alleles of a gene that it controls.

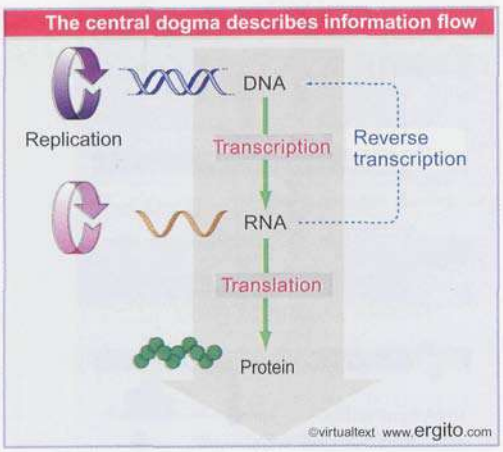


Figure 1.44 The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into protein is irreversible.

- *The expression of cellular genetic information usually is unidirectional.* Transcription of DNA generates RNA molecules that can be used further *only* to generate protein sequences; generally they cannot be retrieved for use as genetic information. Translation of RNA into protein is always irreversible.

These mechanisms are equally effective for the cellular genetic information of prokaryotes or eukaryotes, and for the information carried by viruses. The genomes of all living organisms consist of duplex DNA. Viruses have genomes that consist of DNA or RNA; and there are examples of each type that are double-stranded (ds) or single-stranded (ss). Details of the mechanism used to replicate the nucleic acid vary among the viral systems, but the principle of replication via synthesis of complementary strands remains the same, as illustrated in **Figure 1.45.**

Cellular genomes reproduce DNA by the mechanism of semi-conservative replication. Double-stranded virus genomes, whether DNA or RNA, also replicate by using the individual strands of the duplex as templates to synthesize partner strands.

Viruses with single-stranded genomes use the single strand as a template to synthesize a complementary strand; and this complementary strand in turn is used to synthesize its complement, which is, of course, identical with the original starting strand. Replication may involve the formation of stable double-stranded intermediates or may use double-stranded nucleic acid only as a transient stage.

The restriction to unidirectional transfer from DNA to RNA is not absolute. It is overcome by the retroviruses, whose genomes consist of single-stranded RNA molecules. During the infective cycle, the RNA is converted by the process of reverse transcription into a single-stranded DNA, which in turn is converted into a double-stranded DNA. This duplex DNA becomes part of the genome of the cell, and is inherited like any other gene. *So reverse transcription allows a sequence of RNA to be retrieved and used as genetic information.*

The existence of RNA replication and reverse transcription establishes the general principle that *information in the form of either type of nucleic acid sequence can be converted into the other type.* In the usual course of events, however, the cell relies on the processes of DNA replication, transcription, and translation. But on rare occasions (possibly mediated by an RNA virus), information from a cellular RNA is converted into DNA and inserted into the genome. Although reverse transcription plays no role in the regular operations of the cell, it becomes a mechanism of potential importance when we consider the evolution of the genome.

The same principles are followed to perpetuate genetic information from the massive genomes of plants or amphibians to the tiny genomes of mycoplasma and the yet smaller genetic information of DNA or RNA viruses. **Figure 1.46** summarizes some examples that illustrate the range of genome types and sizes.

Throughout the range of organisms, with genomes varying in total content over a 100,000 fold range, a common principle prevails. *The DNA codes for all the proteins that the cell(s) of the organism must synthesize; and the proteins in turn (directly or indirectly) provide the functions needed for survival.* A similar principle describes the function of the genetic information of viruses, whether DNA or RNA. *The nucleic acid codes for the protein(s) needed to package the genome and also for any functions additional to those provided by the host cell that are needed to reproduce the virus during its infective cycle.* (The smallest virus, the satellite tobacco necrosis virus [STNV], cannot replicate independently, but requires the simultaneous presence of a "helper" virus [tobacco necrosis virus, TNV], which is itself a normally infectious virus.)

1.27 Some hereditary agents are extremely small

Key Concepts

- Some very small hereditary agents do not code for protein but consist of RNA or of protein that has hereditary properties.

Viroids are infectious agents that cause diseases in higher plants. They are very small circular molecules of RNA. Unlike viruses, where the infectious agent consists of a virion, a genome encapsulated in a protein coat, *the viroid RNA is itself the infectious agent.* The viroid consists solely of the RNA, which is extensively but imperfectly base paired, forming a characteristic rod like the example shown in **Figure 1.47.** Mutations that interfere with the structure of the rod reduce infectivity.

A viroid RNA consists of a single molecular species that is replicated autonomously in infected cells. Its sequence is faithfully perpetuated in its descendants. Viroids fall into several groups. A given viroid is identified with a group by its similarity of sequence with other members of the group. For example, four viroids related to PSTV (potato spindle tuber viroid) have 70-83% similarity of sequence with it. Different isolates of a particular viroid strain vary from one another, and the change may affect the phenotype of infected cells. For example, the *mild* and *severe* strains of PSTV differ by three nucleotide substitutions.

Viroids resemble viruses in having heritable nucleic acid genomes. They fulfill the criteria for genetic information. Yet viroids differ from viruses in both structure and function. They are sometimes called subviral pathogens. Viroid RNA does not appear to be translated into protein. So it cannot itself code for the functions needed for its survival. This situation poses two questions. How does viroid RNA replicate? And how does it affect the phenotype of the infected plant cell?

Replication must be carried out by enzymes of the host cell, subverted from their normal function. The heritability of the viroid sequence indicates that viroid RNA provides the template.

Viroids are presumably pathogenic because they interfere with normal cellular processes. They might do this in a relatively random way, for example, by sequestering an essential enzyme for their own replication or by interfering with the production of necessary cellular RNAs. Alternatively, they might behave as abnormal regulatory molecules, with particular effects upon the expression of individual genes.

An even more unusual agent is scrapie, the cause of a degenerative neurological disease of sheep and goats. The disease is related to the human diseases of kuru and Creutzfeldt-Jakob syndrome, which affect brain function.

The infectious agent of scrapie does not contain nucleic acid. This extraordinary agent is called a prion (proteinaceous infectious agent). It is

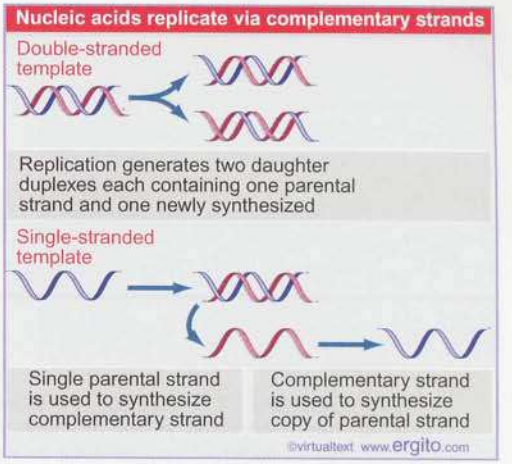


Figure 1.45 Double-stranded and single-stranded nucleic acids both replicate by synthesis of complementary strands governed by the rules of base pairing.

Genomes have nucleic acids

Genome	Gene Number	Base Pairs
Organisms		11
Plants	<50,000	$<10_8$
Mammals	30,000	$\sim 3 \times 10_9$
Worms	14,000	$\sim 10_8$
Flies	12,000	$1.6 \times 10_7$
Fungi	6,000	$1.3 \times 10_7$
Bacteria	2-4,000	$<10_6$
Mycoplasma	500	<10
dsDNA Viruses		
Vaccinia	<300	187,000
Papova (SV40)	~6	5,226
Phage T4	~200	165,000
ssDNA Viruses		
Parvovirus	5	5,000
Phage fX174	11	5,387
dsRNA Viruses		
Reovirus	22	23,000
ssRNA Viruses		
Coronavirus	7	20,000
Influenza	12	13,500
TMV	4	6,400
Phage MS2	4	3,569
STNV	1	1,300
Viroids		
PSTV RNA	0	359
Scrapie		
Prion	?	?

©virtualtext www.ergito.com

Figure 1.46 The amount of nucleic acid in the genome varies over an enormous range.

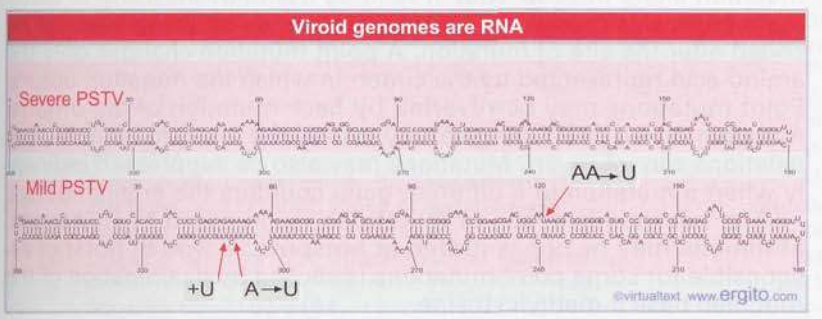


Figure 1.47 PSTV RNA is a circular molecule that forms an extensive double-stranded structure, interrupted by many interior loops. The severe and mild forms differ at three sites.

a 28 kD hydrophobic glycoprotein, PrP. PrP is coded by a cellular gene (conserved among the mammals) that is expressed in normal brain. The protein exists in two forms. The product found in normal brain is called PrP^{c}. It is entirely degraded by proteases. The protein found in infected brains is called PrP^{sc}. It is extremely resistant to degradation by proteases. PrP^{c} is converted to PrP^{sc} by a modification or conformational change that confers protease-resistance, and which has yet to be fully defined.

As the infectious agent of scrapie, PrP^{sc} must in some way modify the synthesis of its normal cellular counterpart so that it becomes infectious instead of harmless (see *23.24 Prions cause diseases in mammals).* Mice that lack a PrP gene cannot be infected to develop scrapie, which demonstrates that PrP is essential for development of the disease.

1.28 Summary

Two classic experiments proved that DNA is the genetic material. DNA isolated from one strain of *Pneumococcus* bacteria can confer properties of that strain upon another strain. And DNA is the only component that is inherited by progeny phages from the parental phages. DNA can be used to transfect new properties into eukaryotic cells.

DNA is a double helix consisting of antiparallel strands in which the nucleotide units are linked by 5'-3' phosphodiester bonds. The backbone provides the exterior; purine and pyrimidine bases are stacked in the interior in pairs in which A is complementary to T while G is complementary to C. The strands separate and use complementary base pairing to assemble daughter strands in semiconservative replication. Complementary base pairing is also used to transcribe an RNA representing one strand of a DNA duplex.

A stretch of DNA may code for protein. The genetic code describes the relationship between the sequence of DNA and the sequence of the protein. Only one of the two strands of DNA codes for protein. A codon consists of three nucleotides that represent a single amino acid. A coding sequence of DNA consists of a series of codons, read from a fixed starting point. Usually only one of the three possible reading frames can be translated into protein.

A chromosome consists of an uninterrupted length of duplex DNA that contains many genes. Each gene (or cistron) is transcribed into an RNA product, which in turn is translated into a polypeptide sequence if the gene codes for protein. An RNA or protein product of a gene is said to be *trans*-acting. A gene is defined as a unit on a single stretch of DNA by the complementation test. A site on DNA that regulates the activity of an adjacent gene is said to be *cis*-acting.

A gene may have multiple alleles. Recessive alleles are caused by a loss-of-function. A null allele has total loss-of-function. Dominant alleles are caused by gain-of-function.

A mutation consists of a change in the sequence of AT and GC base pairs in DNA. A mutation in a coding sequence may change the sequence of amino acids in the corresponding protein. A frameshift mutation alters the subsequent reading frame by inserting or deleting a base; this causes an entirely new series of amino acids to be coded after the site of mutation. A point mutation changes only the amino acid represented by the codon in which the mutation occurs. Point mutations may be reverted by back mutation of the original mutation. Insertions may revert by loss of the inserted material, but deletions cannot revert. Mutations may also be suppressed indirectly when a mutation in a different gene counters the original defect.

The natural incidence of mutations is increased by mutagens. Mutations may be concentrated at hotspots. A type of hotspot responsible for some point mutations is caused by deamination of the modified base 5-methylcytosine.

Forward mutations occur at a rate of ~10^{-6} per locus per generation; back mutations are rarer. Not all mutations have an effect on the phenotype.

Although all genetic information in cells is carried by DNA, viruses have genomes of double-stranded or single-stranded DNA or RNA. Viroids are subviral pathogens that consist solely of small circular molecules of RNA, with no protective packaging. The RNA does not code for protein and its mode of perpetuation and of pathogenesis is unknown. Scrapie consists of a proteinaceous infectious agent.

References

1.1 Introduction

rev Cairns, J., Stent, G., and Watson, J. D. (1966). Phage and the Origins of Molecular Biology. Cold Spring Harbor Symp. Quant. Biol.

Judson, H. (1978). The Eighth Day of Creation. Knopf, New York.

Olby, R. (1974). The Path to the Double Helix. MacMillan, London.

1.2 DNA is the genetic material of bacteria

ref Avery, 0. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J. Exp. Med. 98, 451-460.

Griffith, F. (1928). The significance of pneumococcal types. J. Hyg. 27, 113-159.

1.3 DNA is the genetic material of viruses

ref Hershey, A. D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. 36, 39-56.

1.4 DNA is the genetic material of animal cells

ref Pellicer, A., Wigler, M., Axel, R., and Silverstein, S. (1978). The transfer and stable integration of the HSV thymidine kinase gene into mouse cells. Cell 14, 133-141.

1.6 DNA is a double helix

ref Watson, J. D., and Crick, F. H. C. (1953). A structure for DNA. Nature 171, 737-738.

Watson, J. D., and Crick, F. H. C. (1953). Genetic implications of the structure of DNA. Nature 171, 964-967.

Wilkins, M. F. H., Stokes, A. R., and Wilson, H. R. (1953). Molecular structure of DNA. Nature 171, 738-740.

1.7 DNA replication is semiconservative

rev Holmes, F. (2001). Yale University Press. Meselson, Stahl, and the Replication of DNA: A History of The Most Beautiful Experiment in Biology.

ref Meselson, M. and Stahl, F. W. (1958). The replication of DNA in *E. coli*. Proc. Nat. Acad. Sci. USA 44, 671-682.

1.10 Mutations change the sequence of DNA

rev Drake, J. W., and Balz, R. H. (1976). The biochemistry of mutagenesis. Ann. Rev. Biochem. 45, 11-37.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. Genetics 148, 1667-1686.

ref Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. Proc. Nat. Acad. Sci. USA 88, 7160-7164.

Grogan, D. W., Carver, G. T., and Drake, J. W. (2001). Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. Proc. Nat. Acad. Sci. USA 98, 7928-7933.

1.11 Mutations may affect single base pairs or longer sequences

rev Maki, H. (2002). Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. Ann. Rev. Genet. 36, 279-303.

1.14 Many hotspots result from modified bases

ref Coulondre, C. et al. (1978). Molecular basis of base substitution hotspots in *E. coli*. Nature 274, 775-780.

Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P. D., Bishop, S. M., Clarke, A. R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science 297, 403-405.

1.21 The genetic code is triplet

rev Roth, J. R. (1974). Frameshift mutations. Ann. Rev. Genet. 8, 319-346.

ref Benzer, S., and Champe, S. P. (1961). Ambivalent rII mutants of phage T4. Proc. Nat. Acad. Sci. USA 47, 403-416.

Crick, F. H. C, Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. Nature 192, 1227-1232.

1.23 Prokaryotic genes are colinear with their proteins

exp Yanofsky, C. (2002). Gene-Protein Colinearity (www.ergito.com/lookup.jsp?expt=yanofsky)

ref Yanofsky, C. et al. (1964). On the colinearity of gene structure and protein structure. Proc. Nat. Acad. Sci. USA 51, 266-272.

Yanofsky, C, Drapeau, G. R., Guest, J. R., and Carlton, B. C. (1967). The complete amino acid sequence of the tryptophan synthetase A protein (μ subunit) and its colinear relationship with the genetic map of the A gene. Proc. Nat. Acad. Sci. USA 57, 2966-2968.

1.27 Some hereditary agents are extremely small

rev Diener, T. O. (1986). Viroid processing: a model involving the central conserved region and hairpin. Proc. Nat. Acad. Sci. USA 83, 58-62.

Diener, T. O. (1999). Viroids and the nature of viroid diseases. Arch. Virol. Suppl. 15, 203-220.

Prusiner, S. B. (1998). Prions. Proc. Nat. Acad. Sci. USA 95, 13363-13383.

ref Bueler, H. et al. (1993). Mice devoid of PrP are resistant to scrapie. Cell 73, 1339-1347.

McKinley, M. P., Bolton, D. C, and Prusiner, S. B. (1983). A protease-resistant protein is a structural component of the scrapie prion. Cell 35, 57-62.

Chapter 2

The interrupted gene

2.1 Introduction

Key Concepts

- Eukaryotic genomes contain interrupted genes that consist of an alternation of exons (represented in the final RNA product) and introns (removed from the initial transcript).
- The exon sequences occur in the same order in the gene and in the RNA, but an interrupted gene is longer than its final RNA product because of the presence of the introns.

Until eukaryotic genes were characterized by molecular mapping, we assumed that they would have the same organization as prokaryotic genes. We expected the gene to consist of a length of DNA that is colinear with the protein. But a comparison between the structure of DNA and the corresponding mRNA shows a discrepancy in many cases. The mRNA always includes a nucleotide sequence that corresponds exactly with the protein product according to the rules of the genetic code. *But the gene includes additional sequences that lie within the coding region, interrupting the sequence that represents the protein.*

The sequences of DNA comprising an interrupted gene are divided into the two categories depicted in **Figure 2.1:**

- The exons are the sequences represented in the mature RNA. By definition, a gene starts and ends with exons, corresponding to the 5' and 3' ends of the RNA.
- The introns are the intervening sequences that are removed when the primary transcript is processed to give the mature RNA.

The expression of interrupted genes requires an additional step that does not occur for uninterrupted genes. The DNA gives rise to an RNA copy (a transcript) that exactly represents the genome sequence. But this RNA is only a precursor; it cannot be used for producing protein. First the introns must be removed from the RNA to give a messenger RNA that consists only of the series of exons. This process is called RNA splicing. It involves a precise deletion of an intron from the primary transcript; the ends of the RNA on either side are joined to form a covalently intact molecule (see *24 RNA splicing and processing).*

The structural gene comprises the region in the genome between points corresponding to the 5' and 3' terminal bases of mature mRNA. We know that transcription starts at the 5' end of the mRNA, but

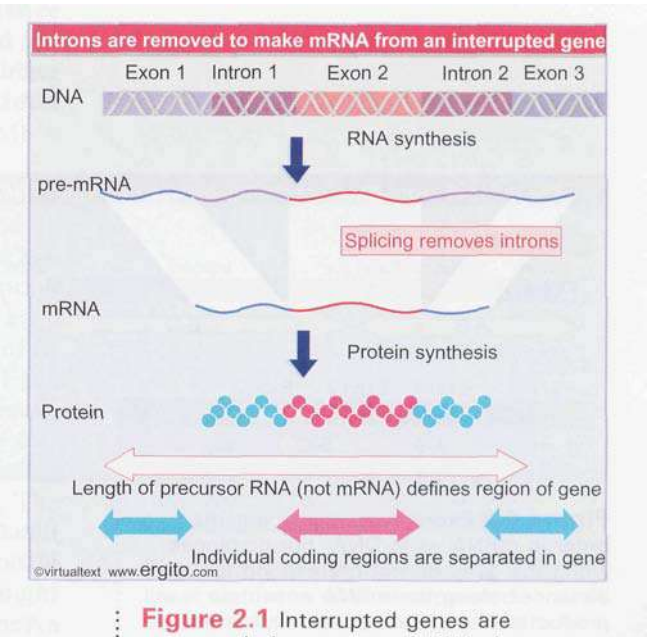


Figure 2.1 Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.

usually it extends beyond the 3′ end, which is generated by cleavage of the RNA (see *24.19 The 3′ ends of mRNAs are generated by cleavage and polyadenylation*). The gene is considered to include the regulatory regions on both sides of the gene that are required for initiating and (sometimes) terminating gene expression.

2.2 An interrupted gene consists of exons and introns

Key Concepts

- Introns are removed by the process of RNA splicing, which occurs only in *cis* on an individual RNA molecule.
- Only mutations in exons can affect protein sequence, but mutations in introns can affect processing of the RNA and therefore prevent production of protein.

How does the existence of introns change our view of the gene? Following splicing, the exons are always joined together in the same order in which they lie in DNA. So the colinearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. Figure 2.2 shows that the *order* of mutations in the gene remains the same as the order of amino acid replacements in the protein. But the *distances* in the gene do not correspond at all with the distances in the protein. Genetic distances, as seen on a recombination map, have no relationship to the distances between the corresponding points in the protein. The length of the gene is defined by the length of the initial (precursor) RNA instead of by the length of the messenger RNA.

All the exons are represented on the same molecule of RNA, and their splicing together occurs only as an *intra*molecular reaction. There is usually no joining of exons carried by *different* RNA molecules, so the mechanism excludes any splicing together of sequences representing different alleles. Mutations located in different exons of a gene cannot complement one another; thus they continue to be defined as members of the same complementation group.

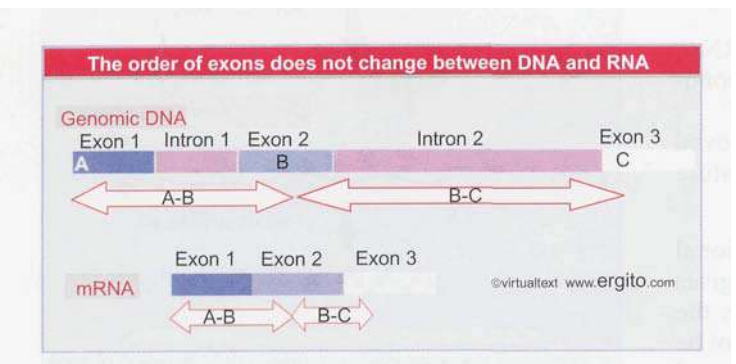


Figure 2.2 Exons remain in the same order in mRNA as in DNA, but distances along the gene do not correspond to distances along the mRNA or protein products. The distance from A-B in the gene is smaller than the distance from B-C; but the distance from A-B in the mRNA (and protein) is greater than the distance from B-C.

Mutations that directly affect the sequence of a protein must lie in exons. What are the effects of mutations in the introns? Since the introns are not part of the messenger RNA, mutations in them cannot directly affect protein structure. However, they can prevent the production of the messenger RNA—for example, by inhibiting the splicing together of exons. A mutation of this sort acts only on the allele that carries it. So it fails to complement any other mutation in that allele, and constitutes part of the same complementation group as the exons.

Mutations that affect splicing are usually deleterious. The majority are single base substitutions at the junctions between introns and exons. They may cause an exon to be left out of the product, cause an intron to be included, or make splicing occur at an aberrant site. The most common result is to introduce a termination codon that results in truncation of the protein sequence. About 15% of the point mutations that cause human diseases are caused by disruption of splicing.

Eukaryotic genes are not necessarily interrupted. Some correspond directly with the protein product in the same manner as prokaryotic

genes. In yeast, most genes are uninterrupted. In higher eukaryotes, most genes are interrupted; and the introns are usually much longer than exons, creating genes that are very much larger than their coding regions.

2.3 Restriction endonucleases are a key tool in mapping DNA

Key Concepts

- Restriction endonucleases can be used to cleave DNA into defined fragments.
- A map can be generated by using the overlaps between the fragments generated by different restriction enzymes.

The characterization of eukaryotic genes was made possible by the development of techniques for physically mapping DNA. The techniques can be extended to (single-stranded) RNA by making a (double-stranded) DNA copy of the RNA. A physical map of any DNA molecule can be obtained by breaking it at defined points whose distance apart can be accurately determined. Specific breaks are made possible by the ability of restriction endonucleases to recognize rather short sequences of double-stranded DNA as targets for cleavage.

Each restriction enzyme has a particular target in duplex DNA, usually a specific sequence of 4-6 base pairs. The enzyme cuts the DNA at every point at which its target sequence occurs. Different restriction enzymes have different target sequences, and a large range of these activities (obtained from a wide variety of bacteria) now is available.

A restriction map represents a linear sequence of the sites at which particular restriction enzymes find their targets. Distance along such maps is measured directly in base pairs (abbreviated *bp)* for short distances; longer distances are given in kb, corresponding to kilobase (10^3) pairs in DNA or to kilobases in RNA. At the level of the chromosome, a map is described in megabase pairs (1 Mb = 10^6 bp).

When a DNA molecule is cut with a suitable restriction enzyme, it is cleaved into distinct fragments. These fragments can be separated on the basis of their size by gel electrophoresis, as shown in **Figure 2.3.** The cleaved DNA is placed on top of a gel made of agarose or polyacrylamide. When an electric current is passed through the gel, each fragment moves down at a rate that is inversely related to the log of its molecular weight. This movement produces a series of bands. Each band corresponds to a fragment of particular size, decreasing down the gel.

By analyzing the restriction fragments of DNA, we can generate a map of the original molecule in the form shown in **Figure 2.4.** The map shows the positions at which particular restriction enzymes cut DNA; the distances between the sites of cutting are measured in base pairs. *So the DNA is divided into a series of regions of defined lengths that lie between sites recognized by the restriction enzymes.* An important feature is that a restriction map can be obtained for any sequence of DNA, *irrespective of whether mutations have been identified in it*, or, indeed, whether we have any knowledge of its function.

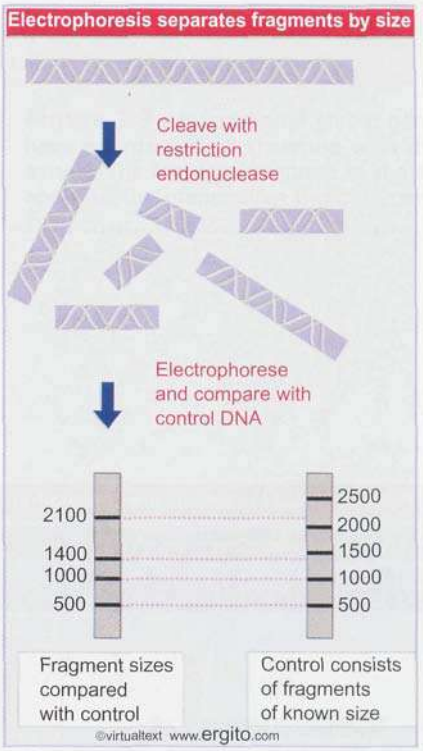


Figure 2.3 Fragments generated by cleaving DNA with a restriction endonuclease can be separated according to their sizes.

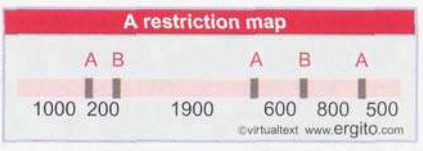


Figure 2.4 A restriction map is a linear sequence of sites separated by defined distances on DNA. The map identifies the sites cleaved by enzymes A and B, as defined by the individual fragments produced by the single and double digests.

2.4 Organization of interrupted genes may be conserved

Key Concepts

- Introns can be detected by the presence of additional regions when genes are compared with their RNA products by restriction mapping or electron microscopy, but the ultimate definition is based on comparison of sequences.
- The positions of introns are usually conserved when homologous genes are compared between different organisms, but the lengths of the corresponding introns may vary greatly.
- Introns usually do not code for proteins.

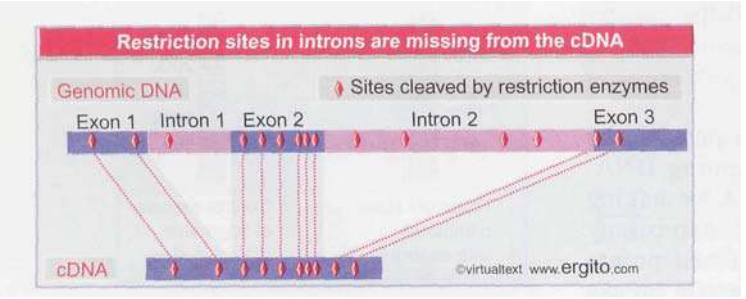


Figure 2.5 Comparison of the restriction maps of cDNA and genomic DNA for mouse β-globin shows that the gene has two introns that are not present in the cDNA. The exons can be aligned exactly between cDNA and gene.

When a gene is uninterrupted, the restriction map of its DNA corresponds exactly with the map of its mRNA.

When a gene possesses an intron, the map at each end of the gene corresponds with the map at each end of the message sequence. But within the gene, the maps diverge, because additional regions are found in the gene, but are not represented in the message. Each such region corresponds to an intron. The example of **Figure 2.5** compares the restriction maps of a β-globin gene and mRNA. There are two introns. Each intron contains a series of restriction sites that are absent from the cDNA. The pattern of restriction sites in the exons is the same in both the cDNA and the gene.

Ultimately a comparison of the nucleotide sequences of the genomic and mRNA sequences precisely defines the introns. As indicated in **Figure 2.6,** an intron usually has no open reading frame. An intact reading frame is created in the mRNA sequence by the removal of the introns.

The structures of eukaryotic genes show extensive variation. Some genes are uninterrupted, so that the genomic sequence is colinear with that of the mRNA. Most higher eukaryotic genes are interrupted, but the introns vary enormously in both number and size.

All classes of genes may be interrupted: nuclear genes coding for proteins, nucleolar genes coding for rRNA, and genes coding for tRNA. Interruptions also are found in mitochondrial genes in lower eukaryotes, and in chloroplast genes. Interrupted genes do not appear to be excluded from any class of eukaryotes, and have been found in bacteria and bacteriophages, although they are extremely rare in prokaryotic genomes.

Some interrupted genes possess only one or a few introns. The globin genes provide an extensively studied example (see *2.11 The members of a gene family have a common organization).* The two general types of globin gene, a and β, share a common type of structure. The consis-

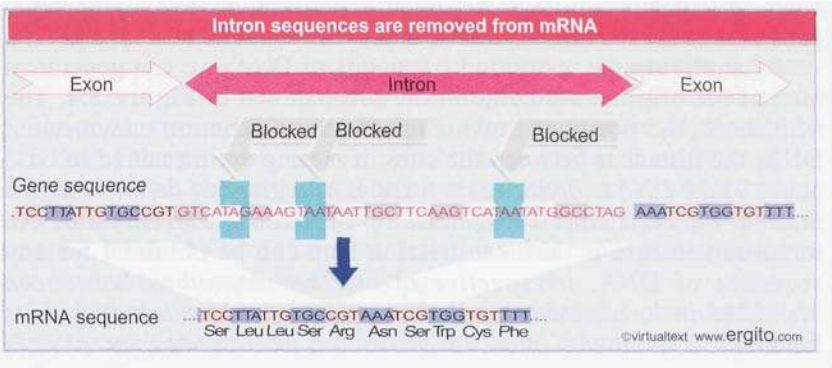


Figure 2.6 An intron is a sequence present in the gene but absent from the mRNA (here shown in terms of the cDNA sequence). The reading frame is indicated by the alternating open and shaded blocks; note that all three possible reading frames are blocked by termination codons in the intron.

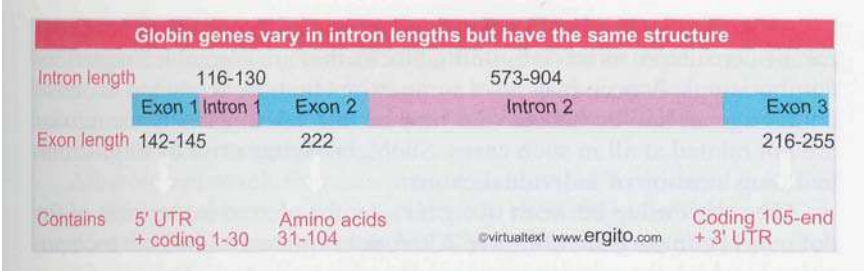


Figure 2.7 All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β-globin genes.

tency of the organization of mammalian globin genes is evident from the structure of the "generic" globin gene summarized in **Figure 2.7.**

Interruptions occur at homologous positions (relative to the coding sequence) in all known active globin genes, including those of mammals, birds, and frogs. The first intron is always fairly short, and the second usually is longer, but the actual lengths can vary. Most of the variation in overall lengths between different globin genes results from the variation in the second intron. In the mouse, the second intron in the α-globin gene is only 150 bp long, so the overall length of the gene is 850 bp, compared with the major β-globin gene where the intron length of 585 bp gives the gene a total length of 1382 bp. The variation in length of the genes is much greater than the range of lengths of the mRNAs (α-globin mRNA = 585 bases, β-globin mRNA = 620 bases).

The example of DHFR, a somewhat larger gene, is shown in **Figure 2.8.** The mammalian DHFR (dihydrofolate reductase) gene is organized into 6 exons that correspond to the 2000 base mRNA. But they extend over a much greater length of DNA because the introns are very long. In three mammals the exons remain essentially the same, and the relative positions of the introns are unaltered, but the lengths of individual introns vary extensively, resulting in a variation in the length of the gene from 25-31 kb.

The globin and DHFR genes present examples of a general phenomenon: *genes that are related by evolution have related organizations, with conservation of the positions of (at least some) of the introns. Variations in the lengths of the genes are primarily determined by the lengths of the introns.*

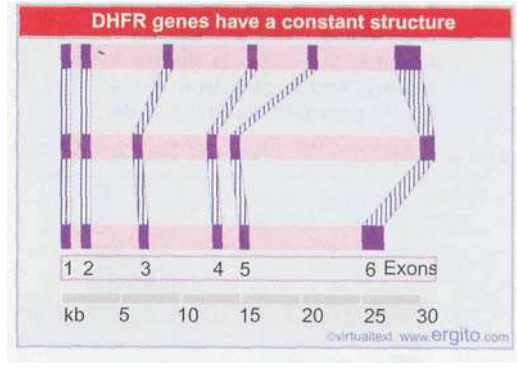


Figure 2.8 Mammalian genes for DHFR have the same relative organization of rather short exons and very long introns, but vary extensively in the lengths of corresponding introns.

2.5 Exon sequences are conserved but introns vary

Key Concepts

- Comparisons of related genes in different species show that the sequences of the corresponding exons are usually conserved but the sequences of the introns are much less well related.
- Introns evolve much more rapidly than exons because of the lack of selective pressure to produce a protein with a useful sequence.

Is a structural gene unique in its genome? The answer can be ambiguous. The entire length of the gene is unique as such, but its exons often are related to those of other genes. As a general rule, when two genes are related, the relationship between their exons is closer than the relationship between the introns. In an extreme case, the exons of two genes may code for the same protein sequence, but the introns may be different. This implies that the two genes originated by a duplication of some common ancestral gene. Then differences accumulated between the copies, but they were restricted in the exons by the need to code for protein functions.

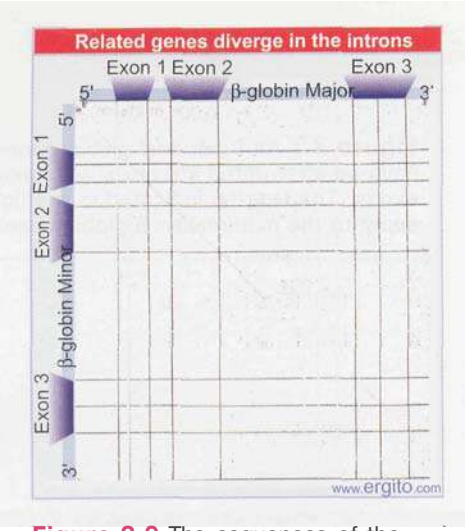


Figure 2.9 The sequences of the mouse α^{maj} and α^{min} globin genes are closely related in coding regions, but differ in the flanking regions and large intron. Data kindly provided by Philip Leder.

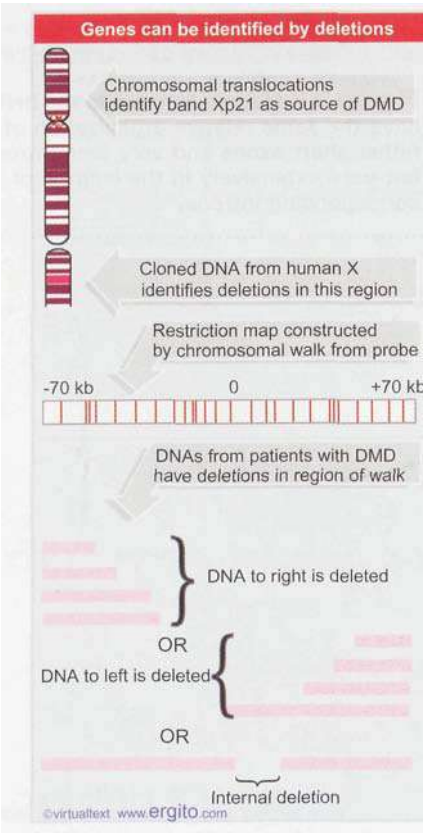


Figure 2.10 The gene involved in Duchenne muscular dystrophy was tracked down by chromosome mapping and walking to a region in which deletions can be identified with the occurrence of the disease.

As we see later when we consider the evolution of the gene, exons can be considered as basic building blocks that are assembled in various combinations. A gene may have some exons that are related to exons of another gene, but the other exons may be unrelated. Usually the introns are not related at all in such cases. Such genes may arise by duplication and translocation of individual exons.

The relationship between two genes can be plotted in the form of the dot matrix comparison of **Figure 2.9.** A dot is placed to indicate each position at which the same sequence is found in each gene. The dots form a line at an angle of 45° if two sequences are identical. The line is broken by regions that lack similarity, and it is displaced laterally or vertically by deletions or insertions in one sequence relative to the other.

When the two β-globin genes of the mouse are compared, such a line extends through the three exons and through the small intron. The line peters out in the flanking regions and in the large intron. This is a typical pattern, in which coding sequences are well related, the relationship can extend beyond the boundaries of the exons, but it is lost in longer introns and the regions on either side of the gene.

The overall degree of divergence between two exons is related to the differences between the proteins. It is caused mostly by base substitutions. In the translated regions, the exons are under the constraint of needing to code for amino acid sequences, so they are limited in their potential to change sequence. Many of the changes do not affect codon meanings, because they change one codon into another that represents the same amino acid. Changes occur more freely in nontranslated regions (corresponding to the 5' leader and 3' trailer of the mRNA).

In corresponding introns, the pattern of divergence involves both changes in size (due to deletions and insertions) and base substitutions. Introns evolve much more rapidly than exons. When a gene is compared in different species, sometimes the exons are homologous, while the introns have diverged so much that corresponding sequences cannot be recognized.

Mutations occur at the same rate in both exons and introns, but are removed more effectively from the exons by adverse selection. However, in the absence of the constraints imposed by a coding function, an intron is quite freely able to accumulate point substitutions and other changes. These changes imply that the intron does not have a sequence-specific function. Whether its presence is at all necessary for gene function is not clear.

2.6 Genes can be isolated by the conservation of exons

Key Concepts

- Conservation of exons can be used as the basis for identifying coding regions by identifying fragments whose sequences are present in multiple organisms.

Some major approaches to identifying genes are based on the contrast between the conservation of exons and the variation of introns. In a region containing a gene whose function has been conserved among a range of species, the sequence representing the protein should have two distinctive properties:

- it must have an open reading frame;
- it is likely to have a related sequence in other species.

These features can be used to isolate genes.

Suppose we know by genetic data that a particular genetic trait is located in a given chromosomal region. If we lack knowledge about the nature of the gene product, how are we to identify the gene in a region that may be (for example) > 1 Mb?

A heroic approach that has proved successful with some genes of medical importance is to screen relatively short fragments from the region for the two properties expected of a conserved gene. First we seek to identify fragments that cross-hybridize with the genomes of other species. Then we examine these fragments for open reading frames.

The first criterion is applied by performing a zoo blot. We use short fragments from the region as (radioactive) probes to test for related DNA from a variety of species by Southern blotting. If we find hybridizing fragments in several species related to that of the probe—the probe is usually human—the probe becomes a candidate for an exon of the gene.

The candidates are sequenced, and if they contain open reading frames, are used to isolate surrounding genomic regions. If these appear to be part of an exon, we may then use them to identify the entire gene, to isolate the corresponding cDNA or mRNA, and ultimately to identify the protein.

This approach is especially important when the target gene is spread out because it has many large introns. This proved to be the case with Duchenne muscular dystrophy (DMD), a degenerative disorder of muscle, which is X-linked and affects 1 in 3500 of human male births. The steps in identifying the gene are summarized in **Figure 2.10.**

Linkage analysis localized the DMD locus to chromosomal band Xp21. Patients with the disease often have chromosomal rearrangements involving this band. By comparing the ability of X-linked DNA probes to hybridize with DNA from patients and with normal DNA, cloned fragments were obtained that correspond to the region that was rearranged or deleted in patients' DNA.

Once some DNA in the general vicinity of the target gene has been obtained, it is possible to "walk" along the chromosome until the gene is reached. A chromosomal walk was used to construct a restriction map of the region on either side of the probe, covering a region of > 100 kb. Analysis of the DNA from a series of patients identified large deletions in this region, extending in either direction. The most telling deletion is one contained entirely within the region, since this delineates a segment that must be important in gene function and indicates that the gene, or at least part of it, lies in this region.

Having now come into the region of the gene, we need to identify its exons and introns. A zoo blot identified fragments that cross-hybridize with the mouse X chromosome and with other mammalian DNAs. As summarized in **Figure 2.11,** these were scrutinized for open reading frames and the sequences typical of exon-intron junctions. Fragments that met these criteria were used as probes to identify homologous sequences in a cDNA library prepared from muscle mRNA.

The cDNA corresponding to the gene identifies an unusually large mRNA, ~14 kb. Hybridization back to the genome shows that the mRNA is represented in >60 exons, which are spread over ~2000 kb of DNA. This makes DMD the longest and most complex gene identified.

The gene codes for a protein of ~500 kD, called dystrophin, which is a component of muscle, present in rather low amounts. All patients with the disease have deletions at this locus, and lack (or have defective) dystrophin.

Muscle also has the distinction of having the largest known protein, titin, with almost 27,000 amino acids. Its gene has the largest number of exons (178) and the longest single exon in the human genome (17,000 bp).

Another technique that allows genomic fragments to be scanned rapidly for the presence of exons is called exon trapping. **Figure 2.12** shows

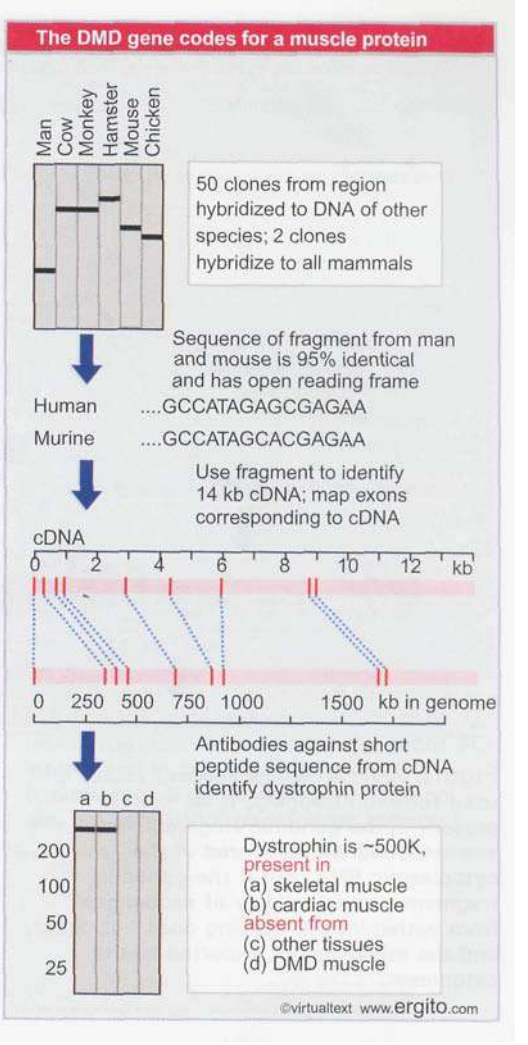


Figure 2.11 The Duchenne muscular dystrophy gene was characterized by zoo blotting, cDNA hybridization, genomic hybridization, and identification of the protein.

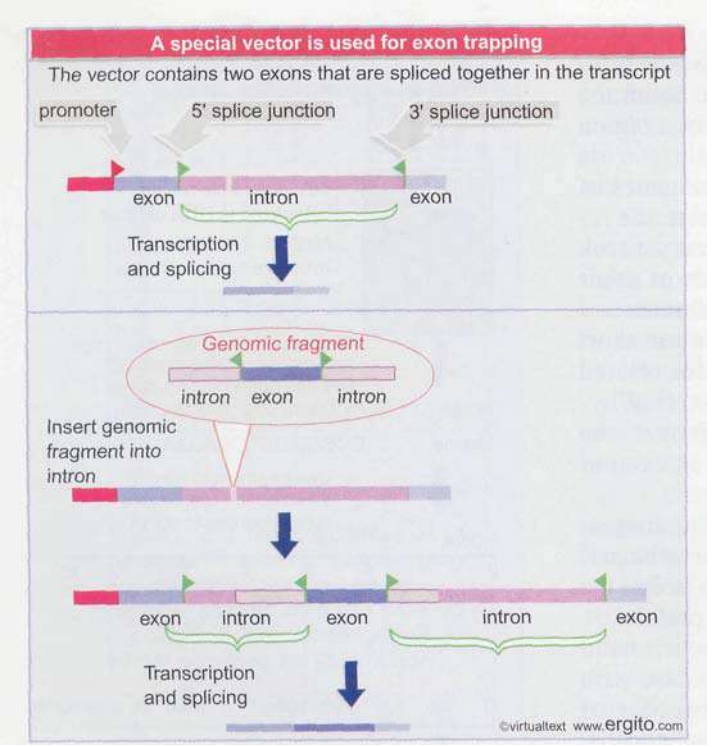


Figure 2.12 A special splicing vector is used for exon trapping. If an exon is present in the genomic fragment, its sequence will be recovered in the cytoplasmic RNA, but if the genomic fragment consists solely of sequences from within intron, splicing does not occur, and the mRNA is not exported to the cytoplasm.

that it starts with a vector that contains a strong promoter, and has a single intron between two exons. When this vector is transfected into cells, its transcription generates large amounts of an RNA containing the sequences of the two exons. A restriction cloning site lies within the intron, and is used to insert genomic fragments from a region of interest. If a fragment does not contain an exon, there is no change in the splicing pattern, and the RNA contains only the same sequences as the parental vector. But if the genomic fragment contains an exon flanked by two partial intron sequences, the splicing sites on either side of this exon are recognized, and the sequence of the exon is inserted into the RNA between the two exons of the vector. This can be detected readily by reverse transcribing the cytoplasmic RNA into cDNA, and using PCR to amplify the sequences between the two exons of the vector. So the appearance in the amplified population of sequences from the genomic fragment indicates that an exon has been trapped. Because introns are usually large and exons are small in animal cells, there is a high probability that a random piece of genomic DNA will contain the required structure of an exon surrounded by partial introns. In fact, exon trapping may mimic the events that have occurred naturally during evolution of genes (see *2.9 How did interrupted genes evolve?*).

2.7 Genes show a wide distribution of sizes

Key Concepts

- Most genes are uninterrupted in yeasts, but are interrupted in higher eukaryotes.
- Exons are usually short, typically coding for <100 amino acids.
- Introns are short in lower eukaryotes, but range up to several 10s of kb in length in higher eukaryotes.
- The overall length of a gene is determined largely by its introns.

Figure 2.13 shows the overall organization of genes in yeasts, insects, and mammals. In 5. *cerevisiae*, the great majority of genes (>96%) are not interrupted, and those that have exons usually remain reasonably compact. There are virtually no *S. cerevisiae* genes with more than 4 exons.

In insects and mammals, the situation is reversed. Only a few genes have uninterrupted coding sequences (6% in mammals). Insect genes tend to have a fairly small number of exons, typically fewer than 10. Mammalian genes are split into more pieces, and some have several 10s of exons. Approximately 50% of mammalian genes have >10 introns.

Examining the consequences of this type of organization for the overall size of the gene, we see in **Figure 2.14** that there is a striking difference between yeast and the higher eukaryotes. The average yeast gene is 1.4 kb long, and very few are longer than 5 kb. The predominance of interrupted genes in high eukaryotes, however, means that the gene can be much larger than the unit that codes for protein. Relatively few genes in flies or mammals are shorter than 2 kb, and many have lengths between 5 kb and 100 kb. The average human gene is 27 kb long (see Figure 3.22).

The switch from largely uninterrupted to largely interrupted genes occurs in the lower eukaryotes. In fungi (excepting the yeasts), the majority of genes are interrupted, but they have a relatively small number of exons (<6) and are fairly short (<5 kb). The switch to long genes occurs within the higher eukaryotes, and genes become significantly larger in the insects. With this increase in the length of the gene, the relationship between genome complexity and organism complexity is lost (see Figure 3.5).

As genome size increases, the tendency is for introns to become rather large, while exons remain quite small.

Figure 2.15 shows that the exons coding for stretches of protein tend to be fairly small. In higher eukaryotes, the average exon codes for ~50 amino acids, and the general distribution fits well with the idea that genes have evolved by the slow addition of units that code for small, individual domains of proteins (see *2.9 How did interrupted genes evolve?).* There is no very significant difference in the sizes of exons in different types of higher eukaryotes, although the distribution is more compact in vertebrates where there are few exons longer than 200 bp. In yeast, there are some longer exons that represent uninterrupted genes where the coding sequence is intact. There is a tendency for exons coding for untranslated 5' and 3' regions to be longer than those that code for proteins.

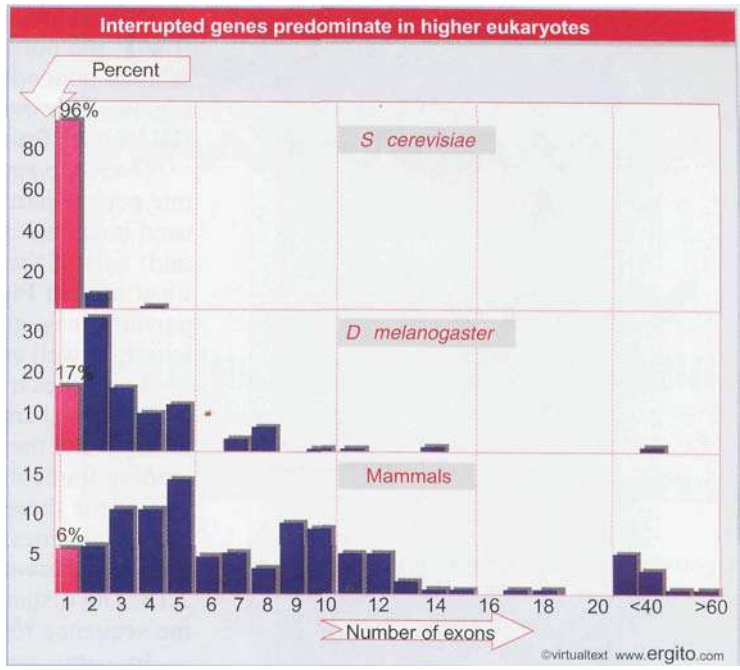


Figure 2.13 Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totaled in the leftmost column.)

Figure 2.16 shows that introns vary widely in size. In worms and flies, the average intron is not much longer than the exons. There are no very long introns in worms, but flies contain a significant proportion. In vertebrates, the size distribution is much wider, extending from approximately the same length as the exons (<200 bp) to lengths measured in 10s of kbs, and extending up to 50-60 kb in extreme cases.

Very long genes are the result of very long introns, not the result of coding for longer products. There is no correlation between gene size and mRNA size in higher eukaryotes; nor is there a good correlation between gene size and the number of exons. The size of a gene therefore depends primarily on the lengths of its individual introns. In mammals, insects, and birds, the "average" gene is approximately 5 X the length of its mRNA.

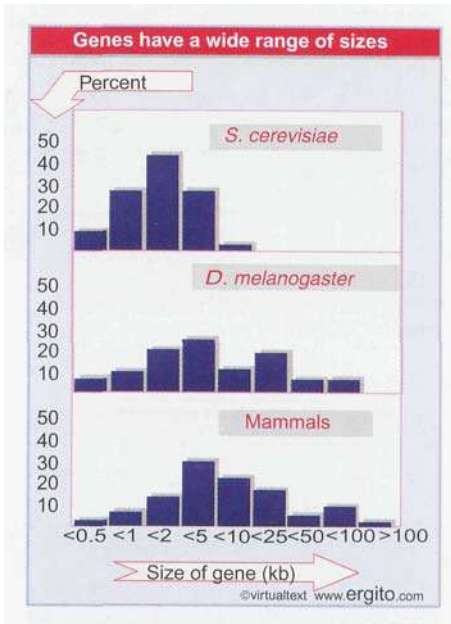


Figure 2.14 Yeast genes are small, but genes in flies and mammals have a dispersed distribution extending to very large sizes.

2.8 Some DNA sequences code for more than one protein

Key Concepts

- The use of alternative initiation or termination codons allows two proteins to be generated where one is equivalent to a fragment of the other.
- Nonhomologous protein sequences can be produced from the same sequence of DNA when it is read in different reading frames by two (overlapping) genes.
- Homologous proteins that differ by the presence or absence of certain regions can be generated by differential (alternative) splicing, when certain exons are included or excluded. This may take the form of including or excluding individual exons or of choosing between alternative exons.

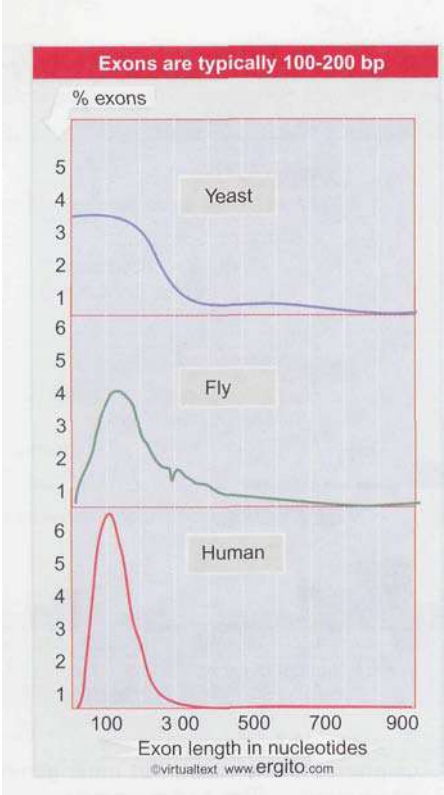


Figure 2.15 Exons coding for proteins are usually short.

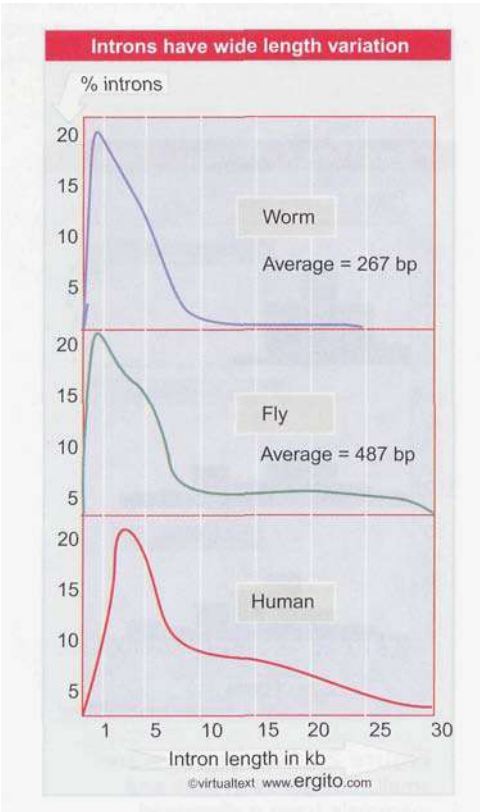


Figure 2.16 Introns range from very short to very long.

Most genes consist of a sequence of DNA that is devoted solely to the purpose of coding for one protein (although the gene may include noncoding regions at either end and introns within the coding region). However, there are some cases in which a single sequence of DNA codes for more than one protein.

Overlapping genes occur in the relatively simple situation in which one gene is part of the other. The first half (or second half) of a gene is used independently to specify a protein that represents the first (or second) half of the protein specified by the full gene. This relationship is illustrated in **Figure 2.17.** The end result is much the same as though a partial cleavage took place in the protein product to generate part-length as well as full-length forms.

Two genes overlap in a more subtle manner when the same sequence of DNA is shared between two *nonhomologous* proteins. This situation arises when the same sequence of DNA is translated in more than one reading frame. In cellular genes, a DNA sequence usually is read in only one of the three potential reading frames, but in some viral and mitochondrial genes, there is an overlap between two adjacent genes that are read in different reading frames. This situation is illustrated in **Figure 2.18.** The distance of overlap is usually relatively short, so that most of the sequence representing the protein retains a unique coding function.

In some genes, *alternative* patterns of gene expression create switches in the pathway for connecting the exons. A single gene may generate a variety of mRNA products that differ in their content of exons. The difference may be that certain exons are optional—they may be included or spliced out. Or there may be exons that are treated as mutually exclusive—one or the other is included, but not both. The alternative forms produce proteins in which one part is common while the other part is different.

In some cases, the alternative means of expression do not affect the sequence of the protein; for example, changes that affect the 5' nontranslated leader or the 3' nontranslated trailer may have regulatory consequences, but the same protein is made. In other cases, one exon is substituted for another, as indicated in **Figure 2.19.**

In this example, the proteins produced by the two mRNAs contain sequences that overlap extensively, but that are different within the alternatively spliced region. The 3' half of the troponin T gene of rat muscle contains 5 exons, but only 4 are used to construct an individual mRNA. Three exons, *WXZ,* are the same in both expression patterns. However, in one pattern the a exon is spliced between *X* and *Z;* in the other pattern, the β exon is used. The a and β forms of troponin T therefore differ in the sequence of the amino acids present between sequences W and Z, depending on which of the alternative exons, a or β, is used. Either one of the a and β exons can be used to form an individual mRNA, but both cannot be used in the same mRNA.

Figure 2.20 illustrates an example in which alternative splicing leads to the inclusion of an exon in some mRNAs, while it is left out of others. A single type of transcript is made from the gene, but it can be spliced in either of two ways. In the first pathway, two introns are spliced out, and the three exons are joined together. In the second pathway, the second exon is not recognized. As a result, a single large intron is spliced out. This intron consists of intron 1 + exon 2 + intron 2. In effect, exon 2 has been treated in this pathway as part of the single intron. The pathways produce two proteins that are the same at their ends, but one of which has an additional sequence in the middle. So the region of DNA codes for more than one protein. (Other types of combinations that are produced by alternative splicing are discussed in *24.12 Alternative splicing involves differential use of splice junctions).*

Sometimes two pathways operate simultaneously, a certain proportion of the RNA being spliced in each way; sometimes the pathways

are alternatives that are expressed under different conditions, one in one cell type and one in another cell type-

So alternative (or differential) splicing can generate proteins with overlapping sequences from a single stretch of DNA. It is curious that the higher eukaryotic genome is extremely spacious in having large genes that are often quite dispersed, but at the same time it may make multiple products from an individual locus. Alternative splicing expands the number of proteins relative to the number of genes by ~15% in flies and worms, but has much bigger effects in man, where ~60% of genes may have alternative modes of expression (see *3.11 The human genome has fewer genes than expected).* About 80% of the alternative splicing events result in a change in the protein sequence.

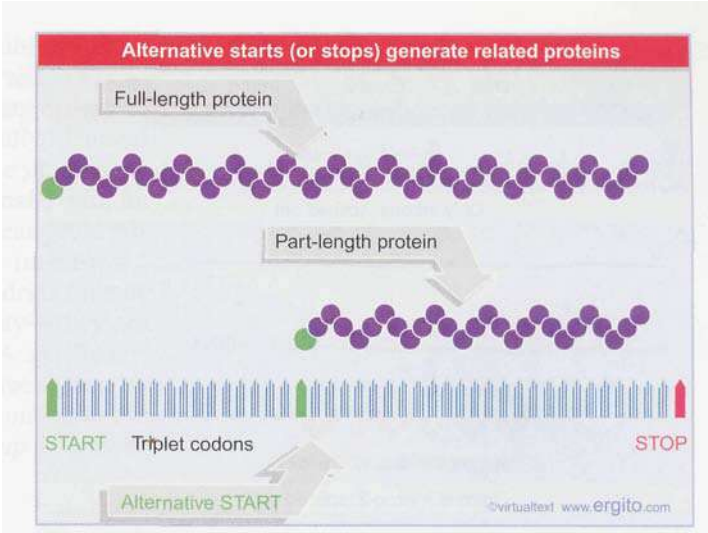


Figure 2.17 Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

2.9 How did interrupted genes evolve?

Key Concepts

- The major evolutionary question is whether genes originated as sequences interrupted by exons or whether they were originally uninterrupted.
- Most protein-coding genes probably originated in an interrupted form, but interrupted genes that code for RNA may have originally been uninterrupted.
- A special class of introns is mobile and can insert itself into genes.

The highly interrupted structure of eukaryotic genes suggests a picture of the eukaryotic genome as a sea of introns (mostly but not exclusively unique in sequence), in which islands of exons (sometimes very short) are strung out in individual archipelagoes that constitute genes.

What was the original form of genes that today are interrupted?

- The "introns early" model supposes that introns have always been an integral part of the gene. Genes originated as interrupted structures, and those without introns have lost them in the course of evolution.
- The "introns late" model supposes that the ancestral protein-coding units consisted of uninterrupted sequences of DNA. Introns were subsequently inserted into them.

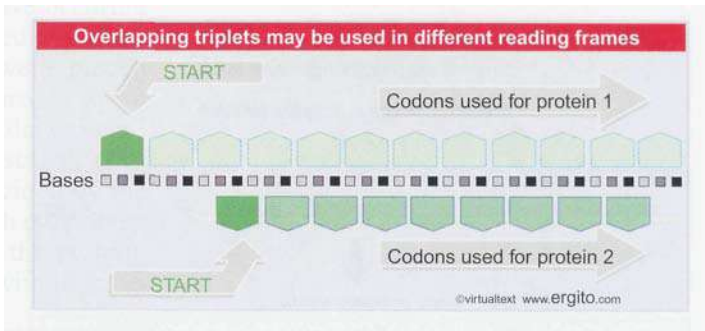


Figure 2.18 Two genes may share the same sequence by reading the DNA in different frames.

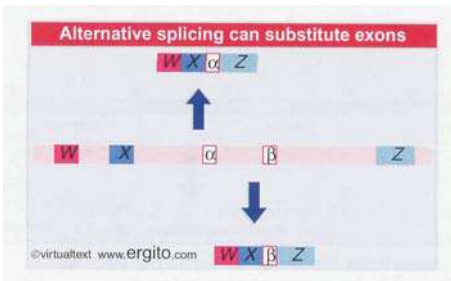


Figure 2.19 Alternative splicing generates the a and β variants of troponin T.

A test of the models is to ask whether the difference between eukaryotic and prokaryotic genes can be accounted for by the acquisition of introns in the eukaryotes or by the loss of introns from the prokaryotes.

The introns early model suggests that the mosaic structure of genes is a remnant of an ancient approach to the reconstruction of genes to make novel proteins. Suppose that an early cell had a number of separate protein-coding sequences. One aspect of its evolution is likely to have been the reorganization and juxtaposition of different polypeptide units to build up new proteins.

If the protein-coding unit must be a continuous series of codons, every such reconstruction would require a precise recombination of

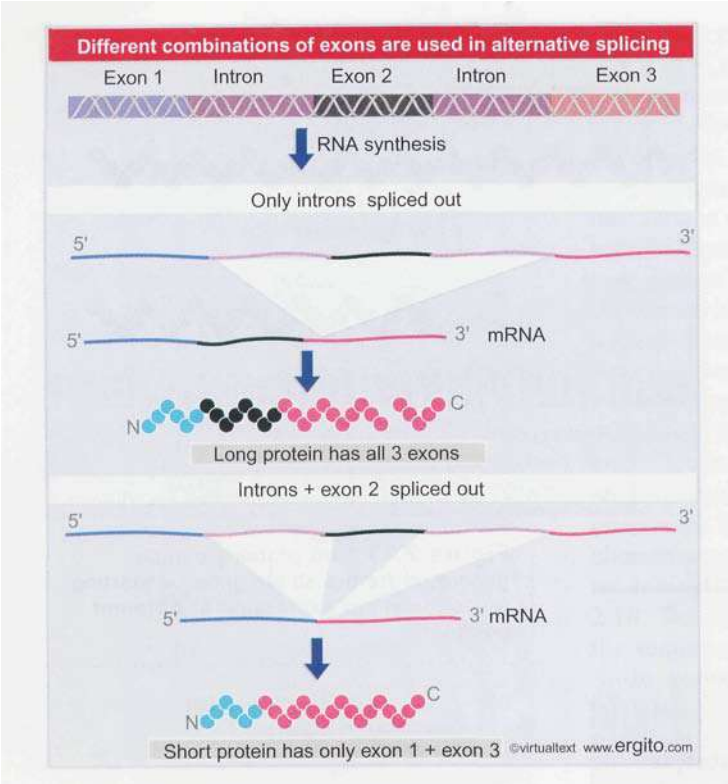


Figure 2.20 Alternative splicing uses the same pre-mRNA to generate mRNAs that have different combinations of exons.

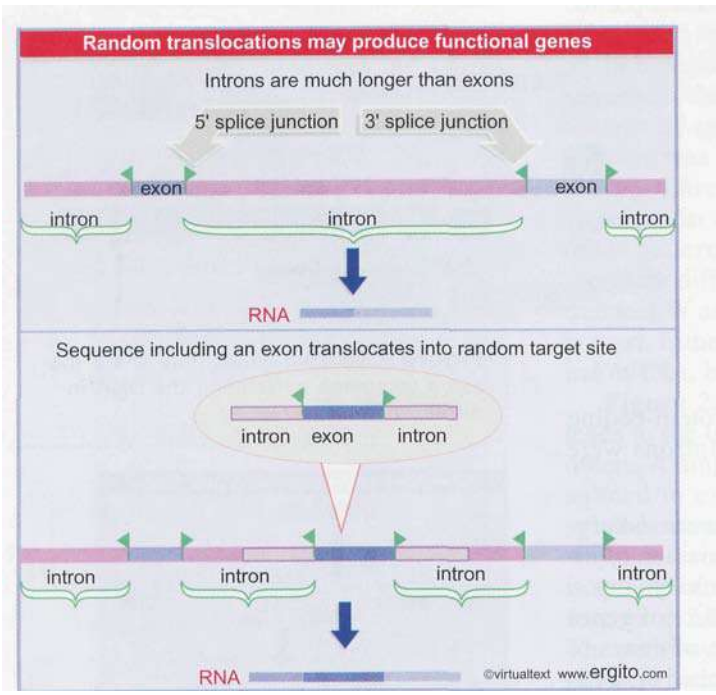


Figure 2.21 An exon surrounded by flanking sequences that is translocated into an intron may be spliced into the RNA product.

DNA to place the two protein-coding units in register, end to end in the same reading frame. Furthermore, if this combination is not successful, the cell has been damaged, because it has lost the original protein-coding units.

But if an approximate recombination of DNA could place the two protein-coding units within the same transcription unit, splicing patterns could be tried out at the level of RNA to combine the two proteins into a single polypeptide chain. And if these combinations are not successful, the original protein-coding units remain available for further trials. Such an approach essentially allows the cell to try out controlled deletions in RNA without suffering the damaging instability that could occur from applying this procedure to DNA. This argument is supported by the fact that we can find related exons in different genes, as though the gene had been assembled by mixing and matching exons (see next section).

Figure 2.21 illustrates the outcome when a random sequence that includes an exon is translocated to a new position in the genome. Exons are very small relative to introns, so it is likely that the exon will find itself within an intron. Because only the sequences at the exon-intron junctions are required for splicing, the exon is likely to be flanked by functional 3' and 5' splice junctions, respectively. Because splicing junctions are recognized in pairs, the 5' splicing junction of the original intron is likely to interact with the 3' splicing junction introduced by the new exon, instead of with its original partner. Similarly, the 5' splicing junction of the new exon will interact with the 3' splicing junction of the original intron. The result is to insert the new exon into the RNA product between the original two exons. So long as the new exon is in the same coding frame as the original exons, a new protein sequence will be produced. This type of event could have been responsible for generating new combinations of exons during evolution. Note that the principle of this type of event is mimicked by the technique of exon trapping that is used to screen for functional exons (see Figure 2.12).

Alternative forms of genes for rRNA and tRNA are sometimes found with and without introns. In the case of the tRNAs, where all the molecules conform to the same general structure, it seems unlikely that evolution brought together the two regions of the gene. After all, the different regions are involved in the base pairing that gives significance to the structure. So here it must be that the introns were inserted into continuous genes.

Organelle genomes provide some striking connections between the prokaryotic and eukaryotic worlds. Because of many general similarities between mitochondria or chloroplasts and bacteria, it seems likely that the organelles originated by an *endosymbiosis* in which an early bacterial prototype was inserted into eukaryotic cytoplasm. Yet in contrast with the resemblances with bacteria—for example, as seen in protein or RNA synthesis—some organelle genes possess introns, and therefore resemble eukaryotic nuclear genes.

Introns are found in several chloroplast genes, including some that have homologies with genes of *E. coli*. This suggests that the endosymbiotic event occurred before introns were lost from the prokaryotic line.

If a suitable gene can be found, it may therefore be possible to trace gene lineage back to the period when endosymbiosis occurred.

The mitochondrial genome presents a particularly striking case. The genes of yeast and mammalian mitochondria code for virtually identical mitochondrial proteins, in spite of a considerable difference in gene organization. Vertebrate mitochondrial genomes are very small, with an extremely compact organization of continuous genes, whereas yeast mitochondrial genomes are larger and have some complex interrupted genes. Which is the ancestral form? The yeast mitochondrial introns (and certain other introns) can have the property of mobility—they are self-contained sequences that can splice out of the RNA and insert DNA copies elsewhere—which suggests that they may have arisen by insertions into the genome (see *25.5 Some group I introns code for endonucleases that sponsor mobility* and *25.6 Some group II intronS code for reverse transcriptases).*

2.10 Some exons can be equated with protein functions

Key Concepts

- Facts suggesting that exons were the building blocks of evolution and the first genes were interrupted are:
  - Gene structure is conserved between genes in very distant species.
  - Many exons can be equated with coding for protein sequences that have particular functions.
  - Related exons are found in different genes.

If current proteins evolved by combining ancestral proteins that were originally separate, the accretion of units is likely to have occurred sequentially over some period of time, with one exon added at a time. Can the different functions from which these genes were pieced together be seen in their present structures? In other words, can we equate particular functions of current proteins with individual exons?

In some cases, there is a clear relationship between the structures of the gene and protein. The example *par excellence* is provided by the immunoglobulin proteins, which are coded by genes in which every exon corresponds exactly with a known functional domain of the protein. Figure 2.22 compares the structure of an immunoglobulin with its gene.

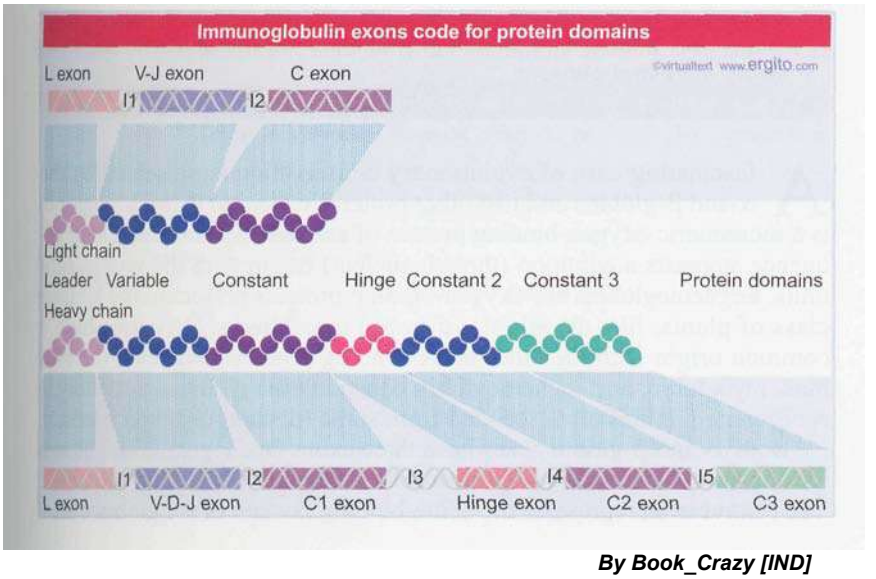


Figure 2.22 Immunoglobulin light chains and heavy chains are coded by genes whose structures (in their expressed forms) correspond with the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered 1-5.

By Book_Crazy [IND]

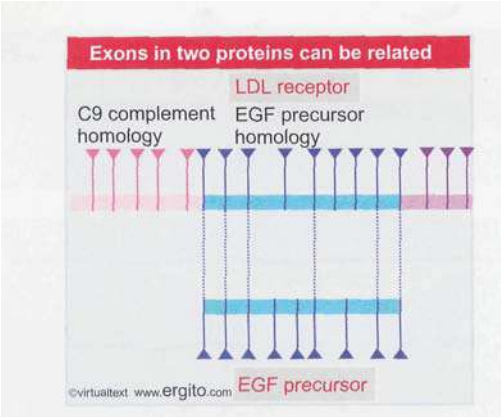


Figure 2.23 The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor and some to the C9 blood complement gene. Triangles mark the positions of introns. Only some of the introns in the region related to EGF precursor are identical in position to those in the EGF gene.

An immunoglobulin is a tetramer of two light chains and two heavy chains, which aggregate to generate a protein with several distinct domains. Light chains and heavy chains differ in structure, and there are several types of heavy chain. Each type of chain is expressed from a gene that has a series of exons corresponding with the structural domains of the protein.

In many instances, some of the exons of a gene can be identified with particular functions. In secretory proteins, the first exon, coding for the N-terminal region of the polypeptide, often specifies the signal sequence involved in membrane secretion. An example is insulin.

The view that exons are the functional building blocks of genes is supported by cases in which two genes may have some exons that are related to one another, while other exons are found only in one of the genes. **Figure 2.23** summarizes the relationship between the receptor for human LDL (plasma low density lipoprotein) and other proteins. In the center of the LDL receptor gene is a series of exons related to the exons of the gene for the precursor for EGF (epidermal growth factor). In the N-terminal part of the protein, a series of exons codes for a sequence related to the blood protein complement factor C9. So the LDL receptor gene was created by assembling *modules* for its various functions. These modules are also used in different combinations in other proteins.

Exons tend to be fairly small (see Figure 2.12), around the size of the smallest polypeptide that can assume a stable folded structure, ~20-40 residues. Perhaps proteins were originally assembled from rather small modules. Each module need not necessarily correspond to a current function; several modules could have combined to generate a function. The number of exons in a gene tends to increase with the length of its protein, which is consistent with the view that proteins acquire multiple functions by successively adding appropriate modules.

This idea might explain another feature of protein structure: it seems that the sites represented at exon-intron boundaries often are located at the surface of a protein. As modules are added to a protein, the connections, at least of the most recently added modules, could tend to lie at the surface.

2.11 The members of a gene family have a common organization

Key Concepts

- A common feature in a set of genes is assumed to identify a property that preceded their separation in evolution.
- All globin genes have a common form of organization with 3 exons and 2 introns, suggesting that they are descended from a single ancestral gene.

A fascinating case of evolutionary conservation is presented by the α-and β-globins and two other proteins related to them. Myoglobin is a monomeric oxygen-binding protein of animals, whose amino acid sequence suggests a common (though ancient) origin with the globin subunits. Leghemoglobins are oxygen-binding proteins present in the legume class of plants; like myoglobin, they are monomeric. They too share a common origin with the other heme-binding proteins. Together, the globins, myoglobin, and leghemoglobin constitute the globin superfamily, a set of gene families all descended from some (distant) common ancestor.

Both a- and β-globin genes have three exons (see Figure 2.7). The two introns are located at constant positions relative to the coding sequence. The central exon represents the heme-binding domain of the globin chain.

Myoglobin is represented by a single gene in the human genome, whose structure is essentially the same as that of the globin genes. The three-exon structure therefore predates the evolution of separate myoglobin and globin functions.

Leghemoglobin genes contain three introns, the first and last of which occur at points in the coding sequence that are homologous to the locations of the two introns in the globin genes. This remarkable similarity suggests an exceedingly ancient origin for the heme-binding proteins in the form of a split gene, as illustrated in **Figure 2.24.**

The central intron of leghemoglobin separates two exons that together code for the sequence corresponding to the single central exon in globin. Could the central exon of the globin gene have been derived by a fusion of two central exons in the ancestral gene? Or is the single central exon the ancestral form; in this case, an intron must have been inserted into it at the start of plant evolution?

Cases in which homologous genes differ in structure may provide information about their evolution. An example is insulin. Mammals and birds have only one gene for insulin, except for the rodents, which have two genes. **Figure 2.25** illustrates the structures of these genes.

The principle we use in comparing the organization of related genes in different species is that *a common feature identifies a structure that predated the evolutionary separation of the two species.* In chickens, the single insulin gene has two introns; one of the two rat genes has the same structure. The common structure implies that the ancestral insulin gene had two introns. However, the second rat gene has only one intron. It must have evolved by a gene duplication in rodents that was followed by the precise removal of one intron from one of the copies.

The organization of some genes shows extensive discrepancies between species. In these cases, there must have been extensive removal or insertion of introns during evolution.

A well characterized case is represented by the actin genes. The typical actin gene has a nontranslated leader of <100 bases, a coding region of ~1200 bases, and a trailer of ~200 bases. Most actin genes are interrupted; the positions of the introns can be aligned with regard to the coding sequence (except for a single intron sometimes found in the leader).

Figure 2.26 shows that almost every actin gene is different in its pattern of interruptions. Taking all the genes together, introns occur at 19 different sites. However, no individual gene has more than 6 introns; some genes have only one intron, and one is uninterrupted altogether. How did this situation arise? If we suppose that the primordial actin gene was interrupted, and all current actin genes are related to it by loss of introns, different introns have been lost in each evolutionary branch. Probably some introns have been lost entirely, so the primordial gene could well have had 20 or more. The alternative is to suppose that a process of intron insertion continued independently in the different lines of evolution. The

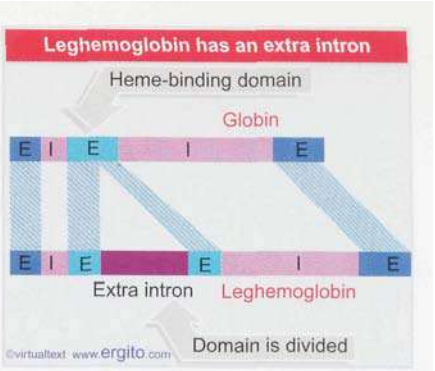


Figure 2.24 The exon structure of globin genes corresponds with protein function, but leghemoglobin has an extra intron in the central domain.

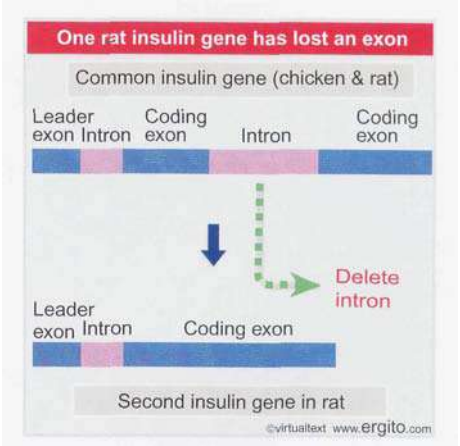


Figure 2.25 The rat insulin gene with one intron evolved by losing an intron from an ancestor with two interruptions.

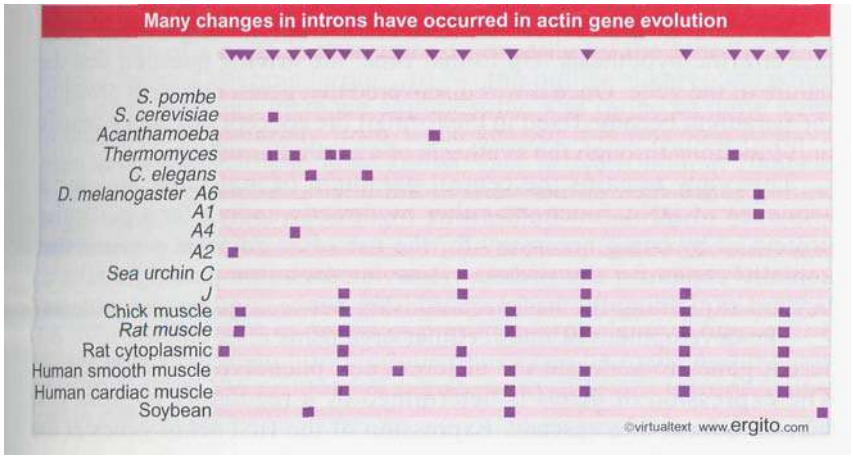


Figure 2.26 Actin genes vary widely in their organization. The sites of introns are indicated in purple.

relationships between the intron locations found in different species may be used ultimately to construct a tree for the evolution of the gene.

The relationship between exons and protein domains is somewhat erratic. In some cases there is a clear 1:1 relationship; in others no pattern is to be discerned. One possibility is that removal of introns has fused the adjacent exons. This means that the intron must have been precisely removed without changing the integrity of the coding region. An alternative is that some introns arose by insertion into a coherent domain. Together with the variations that we see in exon placement in cases such as the actin genes, this argues that intron positions can be adjusted in the course of evolution.

The equation of at least some exons with protein domains, and the appearance of related exons in different proteins, leaves no doubt that the duplication and juxtaposition of exons has played an important role in evolution. It is possible that the number of ancestral exons, from which all proteins have been derived by duplication, variation, and recombination, could be relatively small (a few thousands or tens of thousands). By taking exons as the building blocks of evolution, this view implicitly accepts the introns early model for the origin of genes coding for proteins.

2.12 Is all genetic information contained in DNA?

Key Concepts

- The definition of the gene has reversed from "one gene: one protein" to "one protein: one gene".
- Positional information is also important in development.

The concept of the gene has evolved significantly in the past few years. The question of what's in a name is especially appropriate for the gene. We can no longer say that a gene is a sequence of DNA that continuously and uniquely codes for a particular protein. In situations in which a stretch of DNA is responsible for production of one particular protein, current usage regards the entire sequence of DNA, from the first point represented in the messenger RNA to the last point corresponding to its end as comprising the "gene," exons, introns, and all.

When the sequences representing proteins overlap or have alternative forms of expression, we may reverse the usual description of the gene. Instead of saying "one gene, one polypeptide," we may describe the relationship as "one polypeptide, one gene." So we regard the sequence actually responsible for production of the polypeptide (including introns as well as exons) as constituting the gene, while recognizing that from the perspective of another protein, part of this same sequence also belongs to *its* gene. This allows the use of descriptions such as "overlapping" or "alternative" genes.

We can now see how far we have come from the original one gene: one enzyme hypothesis. Up to that time, the driving question was the nature of the gene. Once it was discovered that genes represent proteins, the paradigm became fixed in the form of the concept that every genetic unit functions through the synthesis of a particular protein.

This view remains the central paradigm of molecular biology: a sequence of DNA functions either by directly coding for a particular protein or by being necessary for the use of an adjacent segment that actually codes for the protein. How far does this paradigm take us beyond explaining the basic relationship between genes and proteins?

The development of multicellular organisms rests on the use of different genes to generate the different cell phenotypes of each tissue. The expression of genes is determined by a regulatory network that takes the form of a cascade. Expression of the first set of genes at the

start of embryonic development leads to expression of the genes involved in the next stage of development, which in turn leads to a further stage, and so on until all the tissues of the adult are functioning. The molecular nature of this regulatory network is largely unknown, but we assume that it consists of genes that code for products (probably protein, perhaps sometimes RNA) that act on other genes.

While such a series of interactions is almost certainly the means by which the developmental program is executed, we can ask whether it is entirely sufficient. One specific question concerns the nature and role of positional information. We know that all parts of a fertilized egg are not equal; one of the features responsible for development of different tissue parts from different regions of the egg is location of information (presumably specific macromolecules) within the cell.

We do not know how these particular regions are formed. But we may speculate that the existence of positional information in the egg leads to the differential expression of genes in the cells subsequently formed in these regions, which leads to the development of the adult organism, which leads to the development of an egg with the appropriate positional information . . .

This possibility prompts us to ask whether some information needed for development of the organism is contained in a form that we cannot directly attribute to a sequence of DNA (although the expression of particular sequences may be needed to perpetuate the positional information). Put in a more general way, we might ask: when we read out the entire sequence of DNA comprising the genome of some organism and interpret it in terms of proteins and regulatory regions, could we in principle construct an organism (or even a single living

2.13 Summary

All types of eukaryotic genomes contain interrupted genes. The proportion of interrupted genes is low in yeasts and increases in the lower eukaryotes; few genes are uninterrupted in higher eukaryotes.

Introns are found in all classes of eukaryotic genes. The structure of the interrupted gene is the same in all tissues, exons are joined together in RNA in the same order as their organization in DNA, and the introns usually have no coding function. Introns are removed from RNA by splicing. Some genes are expressed by alternative splicing patterns, in which a particular sequence is removed as an intron in some situations, but retained as an exon in others.

Positions of introns are often conserved when the organization of homologous genes is compared between species. Intron sequences vary, and may even be unrelated, although exon sequences remain well related. The conservation of exons can be used to isolate related genes in different species.

The size of a gene is determined primarily by the lengths of its introns. Introns become larger early in the higher eukaryotes, when gene sizes therefore increase significantly. The range of gene sizes in mammals is generally from 1-100 kb, but it is possible to have even larger genes; the longest known case is dystrophin at 2000 kb.

Some genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing individual modules of the protein. Such modules may have been incorporated into a variety of different proteins. The idea that genes have been assembled by accretion of exons implies that introns were present in genes of primitive organisms. Some of the relationships between homologous genes can be explained by loss of introns from the primordial genes, with different introns being lost in different lines of descent.

References

2.1 Introduction

exp Sharp, P. A. (2002). The Discovery of RNA Splicing (www.ergito.com/lookup.jsp?expt=sharp)

2.2 An interrupted gene consists of exons and introns

rev Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. Ann. Rev. Biochem. 50, 349-383.
Faustino, N. A. and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes Dev. 17, 419-437.

2.3 Restriction endonucleases are a key tool in mapping DNA

rev Nathans, D. and Smith, H. 0. (1975). Restriction endonucleases in the analysis and restructuring of DNA molecules. Ann. Rev. Biochem. 44, 273-293.
Wu, R. (1978). DNA sequence analysis. Ann. Rev. Biochem. 47, 607-734.

ref Danna, K. J., Sack, G. H., and Nathans, D. (1973). Studies of SV40 DNA VII A cleavage map of the SV40 genome. J. Mol. Biol. 78, 363-376.

2.4 Organization of interrupted genes may be conserved

ref Berget, S. M., Moore, C, and Sharp, P. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Nat. Acad. Sci. USA 74, 3171-3175.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 mRNA. Cell 12, 1-8.

Glover, D. M. and Hogness, D. S. (1977). A novel arrangement of the 8S and 28S sequences in a repeating unit of *D. melanogaster* rDNA. Cell 10, 167-176.

Jeffreys, A. J. and Flavell, R. A. (1977). The rabbit β-globin gene contains a large insert in the coding sequence. Cell 12, 1097-1108.

Wenskink, P. et al. (1974). A system for mapping DNA sequences in the chromosomes of *D. melanogaster.* Cell 3, 315-325.

2.6 Genes can be isolated by the conservation of exons

ref Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc. Nat. Acad. Sci. USA 88, 4005-4009.
Koenig, M., Hoffman, E. P., Bertelson, C. J., Monaco, A. P., Feener, C, and Kunkel, L. M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 50, 509-517.
Kunkel, L. M., Monaco, A. P., Middlesworth, W., Ochs, H. D., and Latt, S. A. (1985). Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. Proc. Nat. Acad. Sci. USA 82, 4778-4782.
Monaco, A. P., Bertelson, C. J., Middlesworth, W., Colletti, C. A., Aldridge, J., Fischbeck, K. H., Bartlett, R., Pericak-Vance, M. A., Roses, A. D., and Kunkel, L. M. (1985). Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment. Nature 316, 842-845.

van Ommen, G. J., Verkerk, J. M., Hofker, M. H., Monaco, A. P., Kunkel, L. M., Ray, P., Worton, R., Wieringa, B., Bakker, E., and Pearson, P. L. (1986). A physical map of 4 million bp around the Duchenne muscular dystrophy gene on the human X-chromosome. Cell 47, 499-504.

2.10 Some exons can be equated with protein functions

rev Blake, C. C. (1985). Exons and the evolution of proteins. Int. Rev. Cytol. 93, 149-185.

Chapter 3

The content of the genome

3.1 Introduction

The key question about the genome is how many genes it contains. We can think about the total number of genes at four levels, corresponding to successive stages in gene expression:

- The genome is the complete set of genes of an organism. Ultimately it is defined by the complete DNA sequence, although as a practical matter it may not be possible to identify every gene unequivocally solely on the basis of sequence.
- The transcriptome is the complete set of genes expressed under particular conditions. It is defined in terms of the set of RNA molecules that is present, and can refer to a single cell type or to any more complex assembly of cells up to the complete organism. Because some genes generate multiple mRNAs, the transcriptome is likely to be larger than the number of genes defined directly in the genome. The transcriptome includes noncoding RNAs as well as mRNAs.
- The proteome is the complete set of proteins. It should correspond to the mRNAs in the transcriptome, although there can be differences of detail reflecting changes in the relative abundance or stabilities of mRNAs and proteins. It can be used to refer to the set of proteins coded by the whole genome or produced in any particular cell or tissue.
- Proteins may function independently or as part of multiprotein assemblies. If we could identify all protein-protein interactions, we could define the total number of independent assemblies of proteins.

The number of genes in the genome can be identified directly by defining open reading frames. Large scale mapping of this nature is complicated by the fact that interrupted genes may consist of many separated open reading frames. Since we do not necessarily have information about the functions of the protein products, or indeed proof that they are expressed at all, this approach is restricted to defining the *potential* of the genome. However, a strong presumption exists that any conserved open reading frame is likely to be expressed.

Another approach is to define the number of genes directly in terms of the transcriptome (by directly identifying all the mRNAs) or

proteome (by directly identifying all the proteins). This gives an assurance that we are dealing with *bona fide* genes that are expressed under known circumstances. It allows us to ask how many genes are expressed in a particular tissue or cell type, what variation exists in the relative levels of expression, and how many of the genes expressed in one particular cell are unique to that cell or are also expressed elsewhere.

Concerning the types of genes, we may ask whether a particular gene is essential: what happens to a null mutant? If a null mutation is lethal, or the organism has a visible defect, we may conclude that the gene is essential or at least conveys a selective advantage. But some genes can be deleted without apparent effect on the phenotype. Are these genes really dispensable, or does a selective disadvantage result from the absence of the gene, perhaps in other circumstances, or over longer periods of time?

3.2 Genomes can be mapped by linkage, restriction cleavage, or DNA sequence

Defining the contents of a genome essentially means making a map. We can think about mapping genes and genomes at several levels of resolution:

- A genetic (or linkage) map identifies the distance between mutations in terms of recombination frequencies. It is limited by its reliance on the occurrence of mutations that affect the phenotype. Because recombination frequencies can be distorted relative to the physical distance between sites, it does not accurately represent physical distances along the genetic material.
- A linkage map can also be constructed by measuring recombination between sites in genomic DNA. These sites have sequence variations that generate differences in the susceptibility to cleavage by certain (restriction) enzymes. Because such variations are common, such a map can be prepared for any organism irrespective of the occurrence of mutants. It has the same disadvantage as any linkage map that the relative distances are based on recombination.
- A restriction map is constructed by cleaving DNA into fragments with restriction enzymes and measuring the distances between the sites of cleavage. This represents distances in terms of the length of DNA, so it provides a physical map of the genetic material. A restriction map does not intrinsically identify sites of genetic interest. For it to be related to the genetic map, mutations have to be characterized in terms of their effects upon the restriction sites. Large changes in the genome can be recognized because they affect the sizes or numbers of restriction fragments. Point mutations are more difficult to detect.
- The ultimate map is to determine the sequence of the DNA. From the sequence, we can identify genes and the distances between them. By analyzing the protein-coding potential of a sequence of the DNA, we can deduce whether it represents a protein. The basic assumption here is that natural selection prevents the accumulation of damaging mutations in sequences that code for proteins. Reversing the argument, we may assume that an intact coding sequence is likely to be used to generate a protein.

By comparing the sequence of a wild-type DNA with that of a mutant allele, we can determine the nature of a mutation and its exact site of occurrence. This defines the relationship between the genetic map

(based entirely on sites of mutation) and the physical map (based on or even comprising the sequence of DNA).

Similar techniques are used to identify and sequence genes and to map the genome, although there is of course a difference of scale. In each case, the principle is to obtain a series of overlapping fragments of DNA, which can be connected into a continuous map. The crucial feature is that each segment is related to the next segment on the map by characterizing the overlap between them, so that we can be sure no segments are missing. This principle is applied both at the level of ordering large fragments into a map, and in connecting the sequences that make up the fragments.

3.3 Individual genomes show extensive variation

Key Concepts

- Polymorphism may be detected at the phenotypic level when a sequence affects gene function, at the restriction fragment level when it affects a restriction enzyme target site, and at the sequence level by direct analysis of DNA.
- The alleles of a gene show extensive polymorphism at the sequence level, but many sequence changes do not affect function.

The original Mendelian view of the genome classified alleles as either wild-type or mutant. Subsequently we recognized the existence of multiple alleles, each with a different effect on the phenotype. In some cases it may not even be appropriate to define any one allele as "wild-type".

The coexistence of multiple alleles at a locus is called genetic polymorphism. Any site at which multiple alleles exist as stable components of the population is by definition polymorphic. An allele is usually defined as polymorphic if it is present at a frequency of > 1 % in the population.

What is the basis for the polymorphism among the mutant alleles? They possess different mutations that alter the protein function, thus producing changes in phenotype. If we compare the restriction maps or the DNA sequences of these alleles, they too will be polymorphic in the sense that each map or sequence will be different from the others.

Although not evident from the phenotype, the wild type may itself be polymorphic. Multiple versions of the wild-type allele may be distinguished by differences in sequence that do not affect their function, and which therefore do not produce phenotypic variants. A population may have extensive polymorphism at the level of genotype. Many different sequence variants may exist at a given locus; some of them are evident because they affect the phenotype, but others are hidden because they have no visible effect.

So there may be a continuum of changes at a locus, including those that change DNA sequence but do not change protein sequence, those that change protein sequence without changing function, those that create proteins with different activities, and those that create mutant proteins that are nonfunctional.

A change in a single nucleotide when alleles are compared is called a single nucleotide polymorphism (SNP). One occurs every ~1330 bases in the human genome. Defined by their SNPs, every human being

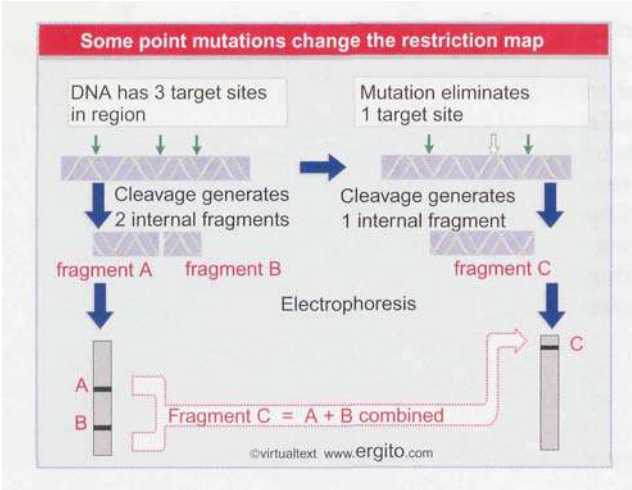


Figure 3.1 A point mutation that affects a restriction site is detected by a difference in restriction fragments.

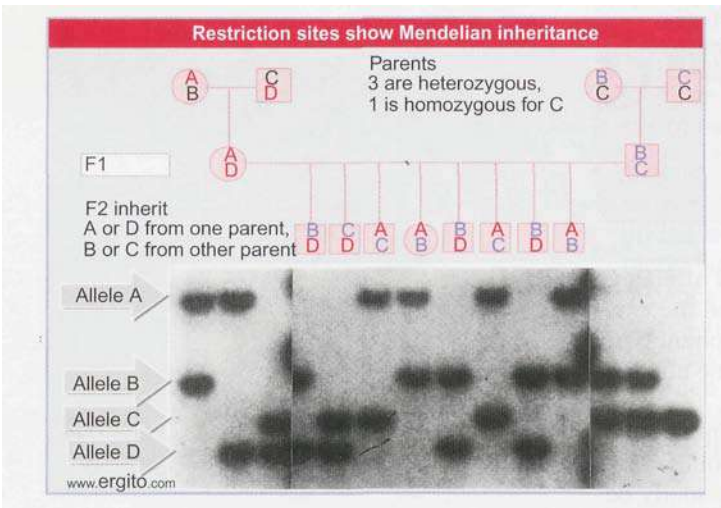


Figure 3.2 Restriction site polymorphisms are inherited according to Mendelian rules. Four alleles for a restriction marker are found in all possible pairwise combinations, and segregate independently at each generation. Photograph kindly provided by Ray White.

is unique. SNPs can be detected by various means, ranging from direct comparisons of sequence to mass spectroscopy or biochemical methods that produce differences based on sequence variations in a defined region.

One aim of genetic mapping is to obtain a catalog of common variants. The observed frequency of SNPs per genome predicts that, over the human population as a whole (taking the sum of all human genomes of all living individuals), there should be >10 million SNPs that occur at a frequency of >1%. Already > 1 million have been identified.

Some polymorphisms in the genome can be detected by comparing the restriction maps of different individuals. The criterion is a change in the pattern of fragments produced by cleavage with a restriction enzyme. **Figure 3.1** shows that when a target site is present in the genome of one individual and absent from another, the extra cleavage in the first genome will generate two fragments corresponding to the single fragment in the second genome.

Because the restriction map is independent of gene function, a polymorphism at this level can be detected *irrespective of whether the sequence change affects the phenotype*. Probably very few of the restriction site polymorphisms in a genome actually affect the phenotype. Most involve sequence changes that have no effect on the production of proteins (for example, because they lie between genes).

A difference in restriction maps between two individuals is called a restriction fragment length polymorphism (RFLP). Basically a RFLP is a SNP that is located in the target site for a restriction enzyme. It can be used as a genetic marker in exactly the same way as any other marker. Instead of examining some feature of the phenotype, we directly assess the genotype, as revealed by the restriction map. **Figure 3.2** shows a pedigree of a restriction polymorphism followed through three generations. It displays Mendelian segregation at the level of DNA marker fragments.

3.4 RFLPs and SNPs can be used for genetic mapping

Key Concepts

- RFLPs and SNPs can be the basis for linkage maps and are useful for establishing parent-progeny relationships.

Recombination frequency can be measured between a restriction marker and a visible phenotypic marker as illustrated in **Figure 3.3.** So a genetic map can include both genotypic and phenotypic markers.

Because restriction markers are not restricted to those genome changes that affect the phenotype, they provide the basis for an extremely powerful technique for identifying genetic loci at the molecular level. A typical problem concerns a mutation with known effects on the phenotype, where the relevant genetic locus can be placed on a genetic map, but for which we have no knowledge about the corresponding gene or protein. Many damaging or fatal human diseases fall into this category. For example cystic fibrosis shows Mendelian inheritance, but

the molecular nature of the mutant function was unknown until it could be identified as a result of characterizing the gene.

If restriction polymorphisms occur at random in the genome, some should occur near any particular target gene. We can identify such restriction markers by virtue of their tight linkage to the mutant phenotype. If we compare the restriction map of DNA from patients suffering from a disease with the DNA of normal people, we may find that a particular restriction site is always present (or always absent) from the patients.

A hypothetical example is shown in **Figure 3.4.** This situation corresponds to finding 100% linkage between the restriction marker and the phenotype. It would imply that the restriction marker lies so close to the mutant gene that it is never separated from it by recombination.

The identification of such a marker has two important consequences:

- It may offer a diagnostic procedure for detecting the disease. Some of the human diseases that are genetically well characterized but ill defined in molecular terms cannot be easily diagnosed. If a restriction marker is reliably linked to the phenotype, then its presence can be used to diagnose the disease.
- It may lead to isolation of the gene. The restriction marker must lie relatively near the gene on the genetic map if the two loci rarely or never recombine. Although "relatively near" in genetic terms can be a substantial distance in terms of base pairs of DNA, nonetheless it provides a starting point from which we can proceed along the DNA to the gene itself.

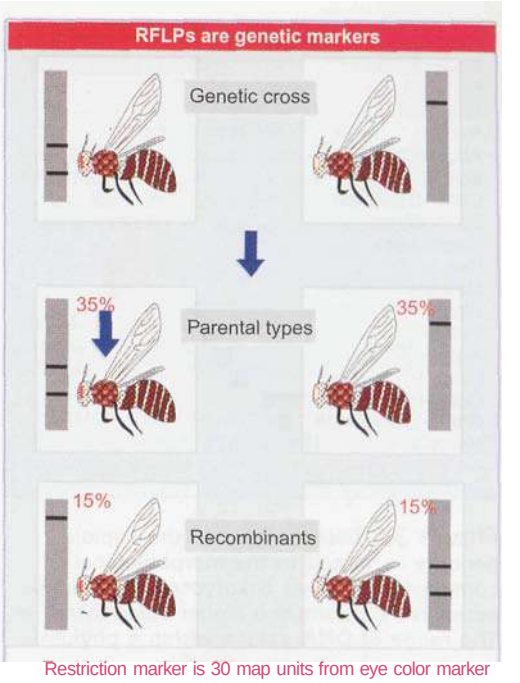


Figure 3.3 A restriction polymorphism can be used as a genetic marker to measure recombination distance from a phenotypic marker (such as eye color). The figure simplifies the situation by showing only the DNA bands corresponding to the allele of one genome in a diploid.

The frequent occurrence of SNPs in the human genome makes them useful for genetic mapping. From the 1.4×10^6 SNPs that have already been identified, there is on average an SNP every 1-2 kb. This should allow rapid localization of new disease genes by locating them between the nearest SNPs.

On the same principle, RFLP mapping has been in use for some time. Once an RFLP has been assigned to a linkage group, it can be placed on the genetic map. RFLP mapping in man and mouse has led to the construction of linkage maps for both genomes. Any unknown site can be tested for linkage to these sites and by this means rapidly placed on to the map. Because there are fewer RFLPs than SNPs, the resolution of the RFLP map is in principle more limited.

The frequency of polymorphism means that every individual has a unique constellation of SNPs or RFLPs. The particular combination of sites found in a specific region is called a haplotype, a genotype in miniature. Haplotype was originally introduced as a concept to describe the genetic constitution of the major histocompatibility locus, a region specifying proteins of importance in the immune system (see *26 Immune diversity).* The concept now has been extended to describe the particular combination of alleles or restriction sites (or any other genetic marker) present in some defined area of the genome.

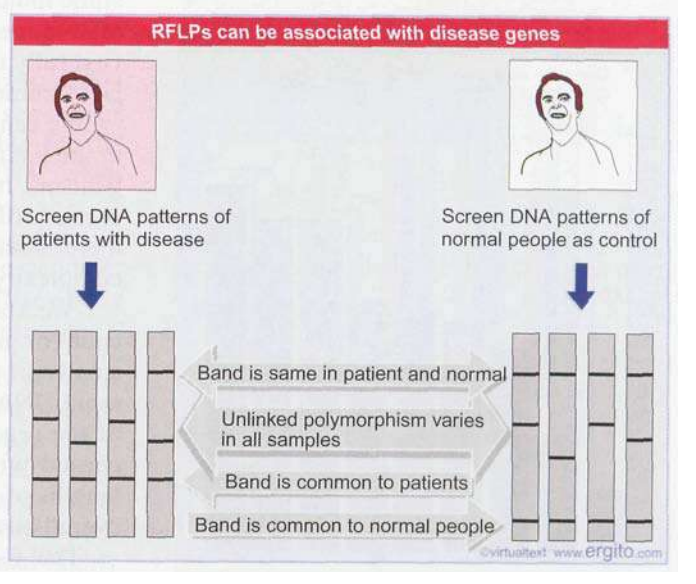


Figure 3.4 If a restriction marker is associated with a phenotypic characteristic, the restriction site must be located near the gene responsible for the phenotype. The mutation changing the band that is common in normal people into the band that is common in patients is very closely linked to the disease gene.

The existence of RFLPs provides the basis for a technique to establish unequivocal parent-progeny relationships. In cases where parentage is in doubt, a comparison of the RFLP map in a suitable chromosome region between potential parents and child allows absolute assignment of the relationship. The use of DNA restriction analysis to identify individuals has been called DNA fingerprinting. Analysis of especially variable "minisatellite" sequences is used mapping in the human genome (see *4.14 Minisatellites are useful for genetic mapping).*

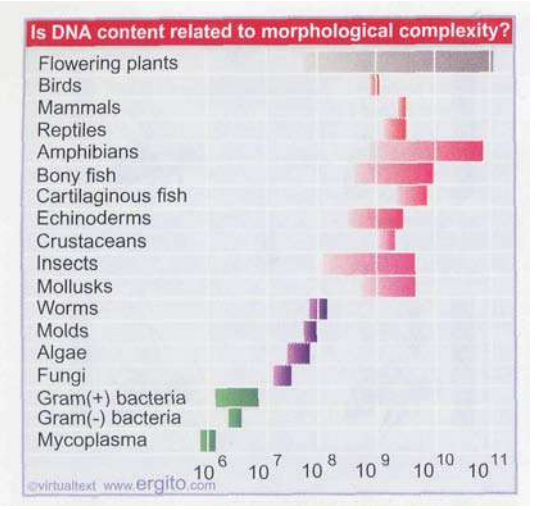


Figure 3.5 DNA content of the haploid genome is related to the morphological complexity of lower eukaryotes, but varies extensively among the higher eukaryotes. The range of DNA values within a phylum is indicated by the shaded area.

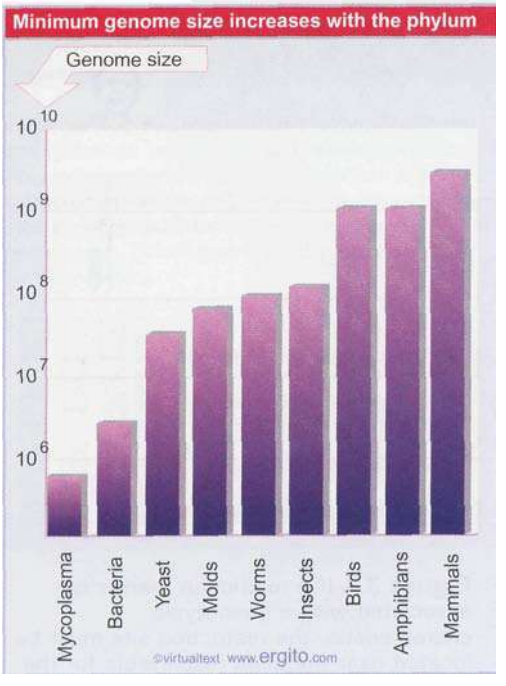


Figure 3.6 The minimum genome size found in each phylum increases from prokaryotes to mammals.

3.5 Why are genomes so large?

Key Concepts

- There is no good correlation between genome size and genetic complexity.
- There is an increase in the minimum genome size required to make organisms of increasing complexity.
- There are wide variations in the genome sizes of organisms within many phyla.

The total amount of DNA in the (haploid) genome is a characteristic of each living species known as its C-value. There is enormous variation in the range of C-values, from $<10^6$ bp for a mycoplasma to $>10^{11}$ bp for some plants and amphibians.

Figure 3.5 summarizes the range of C-values found in different evolutionary phyla. There is an increase in the minimum genome size found in each group as the complexity increases. But as absolute amounts of DNA increase in the higher eukaryotes, we see some wide variations in the genome sizes within some phyla.

Plotting the *minimum* amount of DNA required for a member of each group suggests in **Figure 3.6** that an increase in genome size is required to make more complex prokaryotes and lower eukaryotes.

Mycoplasma are the smallest prokaryotes, and have genomes only ~3× the size of a large bacteriophage. Bacteria start at ~2 X 10^6 bp. Unicellular eukaryotes (whose life-styles may resemble the prokaryotic) get by with genomes that are also small, although larger than those of the bacteria. Being eukaryotic *per se* does not imply a vast increase in genome size; a yeast may have a genome size of ~ 1.3 X 10^7 bp, not much above the size of the largest bacterial genomes.

A further twofold increase in genome size is adequate to support the slime mold *D. discoideum*, able to live in either unicellular or multicellular modes. Another increase in complexity is necessary to produce the first fully multicellular organisms; the nematode worm C. *elegans* has a DNA content of 8 X 10^7 bp.

We can also see the steady increase in genome size with complexity in the listing in **Figure 3.7** of some of the most commonly analyzed organisms. It is necessary to increase the genome size in order to make insects, birds or amphibians, and mammals. However, after this point there is no good relationship between genome size and morphological complexity of the organism.

We know that genes are much larger than the sequences needed to code for proteins, because exons (coding regions) may comprise only a small part of the total length of a gene). This explains why there is much more DNA than is needed to provide reading frames for all the proteins of the organism. Large parts of an interrupted gene may not be concerned with coding for protein. And there may also be significant lengths of DNA between genes. So it is not possible to deduce from the overall size of the genome anything about the number of genes.

The C-value paradox refers to the lack of correlation between genome size and genetic complexity. There are some extremely curious variations in relative genome size. The toad *Xenopus* and man have genomes of essentially the same size. But we assume that man is more complex in terms of genetic development! And in some phyla there are extremely large variations in DNA content between organisms that do not vary much in complexity (see Figure 3.5). (This is especially marked in insects, amphibians, and plants, but does not occur in birds, reptiles, and mammals, which all show little variation within the group,

with an ~2X range of genome sizes.) A cricket has a genome 11× the size of a fruit fly. In amphibians, the smallest genomes are $<10^9$ bp, while the largest are $\sim10^{11}$ bp. There is unlikely to be a large difference in the number of genes needed to specify these amphibians. We do not understand why natural selection allows this variation and whether it has evolutionary consequences.

Useful genome sizes

Phylum	Species	Genome (bp)
Algae	*Pyrenomas salina*	6.6×10^5
Mycoplasma	*M. pneumoniae*	1.0×10^6
Bacterium	*E. coli*	4.2×10^6
Yeast	*S. cerevisiae*	1.3×10^7
Slime mold	*D. discoideum*	5.4×10^7
Nematode	*C. elegans*	8.0×10^7
Insect	*D. melanogaster*	1.4×10^8
Bird	*G. domesticus*	1.2×10^9
Amphibian	*X. laevis*	3.1×10^9
Mammal	*H. sapiens*	3.3×10^9

Figure 3.7 The genome sizes of some common experimental animals.

3.6 Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences

Key Concepts

- The kinetics of DNA reassociation after a genome has been denatured distinguish sequences by their frequency of repetition in the genome.
- Genes are generally coded by sequences in nonrepetitive DNA.
- Larger genomes within a phylum do not contain more genes, but have large amounts of repetitive DNA.
- A large part of repetitive DNA may be made up of transposons.

The general nature of the eukaryotic genome can be assessed by the kinetics of reassociation of denatured DNA. This technique was used extensively before large scale DNA sequencing became possible.

Reassociation kinetics identify two general types of genomic sequences:

- Nonrepetitive DNA consists of sequences that are unique: there is only one copy in a haploid genome.
- Repetitive DNA describes sequences that are present in more than one copy in each genome.

Repetitive DNA is often divided into two general types:

- Moderately repetitive DNA consists of relatively short sequences that are repeated typically 10-1000× in the genome. The sequences are dispersed throughout the genome, and are responsible for the high degree of secondary structure formation in pre-mRNA, when (inverted) repeats in the introns pair to form duplex regions.
- Highly repetitive DNA consists of very short sequences (typically < 100 bp) that are present many thousands of times in the genome, often organized as long tandem repeats (see *4.11 Satellite DNAs often lie in heterochromatin*). Neither class represents protein.

The proportion of the genome occupied by nonrepetitive DNA varies widely. **Figure 3.8** summarizes the genome organization of some representative organisms. Prokaryotes contain only nonrepetitive DNA. For lower eukaryotes, most of the DNA is nonrepetitive; <20% falls into one or more moderately repetitive components. In animal cells, up to half of the DNA often is occupied by moderately and highly repetitive components. In plants and amphibians, the moderately and highly repetitive components may account for up to 80% of the genome, so that the nonrepetitive DNA is reduced to a minority component.

A significant part of the moderately repetitive DNA consists of transposons, short sequences of DNA (~1 kb) that have the ability to move to new locations in the genome and/or to make additional copies of themselves (see *16 Transposons* and *17 Retroviruses and retroposons).* In some higher eukaryotic genomes they may even occupy more than half of the genome (see *3.11 The human genome has fewer genes than expected).*

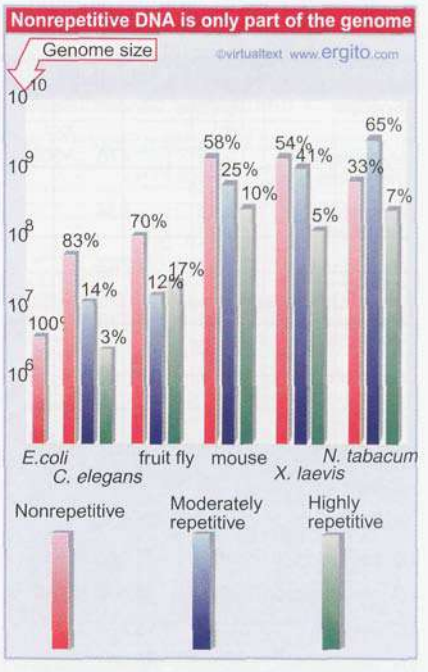


Figure 3.8 The proportions of different sequence components vary in eukaryotic genomes. The absolute content of nonrepetitive DNA increases with genome size, but reaches a plateau at ~2 x 10^9 bp.

Transposons are sometimes viewed as fitting the concept of selfish DNA, defined as sequences that propagate themselves within a genome, without contributing to the development of the organism. Transposons may sponsor genome rearrangements, and these could confer selective advantages, but it is fair to say that we do not really understand why selective forces do not act against transposons becoming such a large proportion of the genome. Another term that is sometimes used to describe the apparent excess of DNA is *junk DNA*, meaning genomic sequences without any apparent function. Of course, it is likely that there is a balance in the genome between the generation of new sequences and the elimination of unwanted sequences, and some proportion of DNA that apparently lacks function may be in the process of being eliminated.

The length of the nonrepetitive DNA component tends to increase with overall genome size, as we proceed up to a total genome size ~3 X 10^9 (characteristic of mammals). Further increase in genome size, however, generally reflects an increase in the amount and proportion of the repetitive components, so that it is rare for an organism to have a nonrepetitive DNA component >2 X 10^9. The nonrepetitive DNA content of genomes therefore accords better with our sense of the relative complexity of the organism. *E. coli* has 4.2 X 10^6 bp, *C. elegans* increases an order of magnitude to 6.6 X 10^7 bp, *D. melanogaster* increases further to ~10^8 bp, and mammals increase another order of magnitude to ~2 X 10^9 bp.

What type of DNA corresponds to protein-coding genes? Reassociation kinetics typically show that mRNA is derived from nonrepetitive DNA. The amount of nonrepetitive DNA is therefore a better indication that the total DNA of the coding potential. (However, more detailed analysis based on genomic sequences shows that many exons have related sequences in other exons [see *2.5 Exon sequences are conserved but introns vary]*. Such exons evolve by a duplication to give copies that initially are identical, but which then diverge in sequence during evolution.)

Sequenced genomes vary from 470-40,000 genes

Species	Genome (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1,743	
Methanococcus jannaschi	1.66	1,738	
B. subtilis	4.2	4,100	
E. coli	4.6	4,288	1,800
S. cerevisiae	13.5	6,034	1,090
S. pombe	12.5	4,929	
A. thaliana	119	25,498	
O. sativa (rice)	466	~40,000	
D. melanogaster	165	13,601	3,100
C. elegans	97	18,424	
H. sapiens	3,300	<40,000	

Figure 3.9 Genome sizes and gene numbers are known from complete sequences for several organisms. Lethal loci are estimated from genetic data.

3.7 Bacterial gene numbers range over an order of magnitude

Key Concepts

- Genome sequences show that there are 500-1200 genes in parasitic bacteria, 1500-7500 genes in free-living bacteria, and 1500-2700 genes in archaea.

Large-scale efforts have now led to the sequencing of many genomes. A range is summarized in **Figure 3.9.** They extend from the 0.6 X 10^6 bp of a mycoplasma to the 3.3 X 10^9 bp of the human genome, and include several important experimental animals, including yeasts, the fruit fly, and a nematode worm.

Figure 3.10 summarizes the minimum number of genes found in each class of organism; of course, many species may have more than the minimum number required for their type.

The sequences of the genomes of bacteria and archaea show that virtually all of the DNA (typically 85-90%) codes for RNA or protein. **Figure 3.11** shows that the range of genome sizes is about an order of magnitude, and that the genome size is proportional to the number of genes. The typical gene is about 1000 bp in length.

All of the bacteria with genome sizes below 1.5 Mb are obligate intracellular parasites—they live within a eukaryotic host that provides them with small molecules. Their genomes identify the minimum number of functions required to construct a cell. All classes of genes are reduced in number compared with bacteria with larger genomes, but the most significant reduction is in loci coding for enzymes concerned with metabolic functions (which are largely provided by the host cell) and with regulation of gene expression. *Mycoplasma genitalium* has the smallest genome, ~470 genes.

The archaea have biological properties that are intermediate between the prokaryotes and eukaryotes, but their genome sizes and gene numbers fall in the same range as bacteria. Their genome sizes vary from 1.5-3 Mb, corresponding to 1500 - 2700 genes. *M. jannaschii* is a methane-producing species that lives under high pressure and temperature. Its total gene number is similar to that of *H. influenzae*, but fewer of its genes can be identified on the basis of comparison with genes known in other organisms. Its apparatus for gene expression resembles eukaryotes more than prokaryotes, but its apparatus for cell division better resembles prokaryotes.

The archaea and the smallest free-living bacteria identify the minimum number of genes required to make a cell able to function independently in the environment. The smallest archaeal genome has ~1500 genes. The free-living bacterium with the smallest known genome is the thermophile *Aquifex aeolicus*, with 1.5 Mb and 1512 genes. A "typical" gram-negative bacterium, *H. influenzae*, has 1,743 genes each of ~900 bp. So we can conclude that ~1500 genes are required to make a free-living organism.

Bacterial genome sizes extend over almost an order of magnitude to <8 Mb. The larger genomes have more genes. The bacteria with the largest genomes, *S. meliloti* and *M. loti*, are nitrogen-fixing bacteria that live on plant roots. Their genome sizes (~7 Mb) and total gene numbers (>6000) are similar to those of yeasts.

The size of the genome of *E. coli* is in the middle of the range. The common laboratory strain has 4,288 genes, with an average length ~950 bp, and an average separation between genes of 118 bp. But there can be quite significant differences between strains. The known extremes of *E. coli* are from the smallest strain that has 4.6 Mb with 4249 genes to the largest strain that has 5.5 Mb bp with 5361 genes.

We still do not know the functions of all the genes. In most of these genomes, ~60% of the genes can be identified on the basis of homology with known genes in other species. These genes fall approximately equally into classes whose products are concerned with metabolism, cell structure or transport of components, and gene expression and its regulation. In virtually every genome, >25% of the genes cannot be ascribed any function. Many of these genes can be found in related organisms, which implies that they have a conserved function.

There has been some emphasis on sequencing the genomes of pathogenic bacteria, given their medical importance. An important insight into the nature of pathogenicity has been provided by the demonstration that "pathogenicity islands" are a characteristic feature of their genomes. These are large regions, ~10-200 kb, that are present in the genome of a pathogenic species, but absent from the genomes of nonpathogenic variants of the same or related species. Their G-C content often differs from that of the rest of the genome, and it is likely that they migrate between bacteria by a process of horizontal transfer. For example, the bacterium that causes anthrax *(B. anthracis)* has two large plasmids (extrachromosomal DNA), one of which has a pathogenicity island that includes the gene coding for the anthrax toxin.

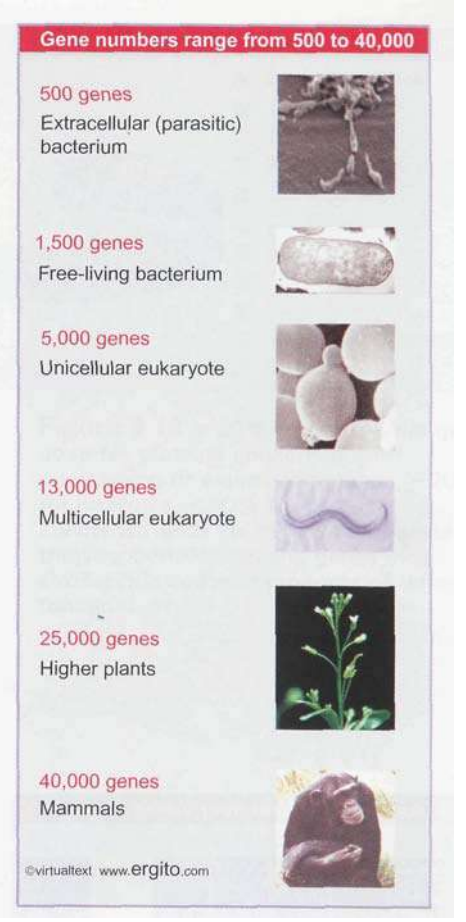


Figure 3.10 The minimum gene number required for any type of organism increases with its complexity. Photograph of mycoplasma kindly provided by A. Albay, K. Frantz, and K. Bott.

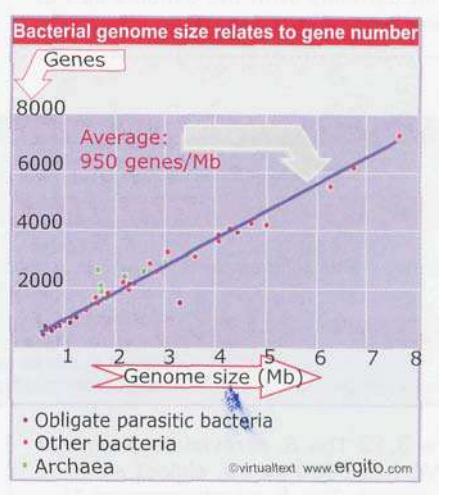


Figure 3.11 The number of genes in bacterial and archaeal genomes is proportional to genome size.

3.8 Total gene number is known for several eukaryotes

Key Concepts

- There are 6000 genes in yeast, 18,500 in worm, 13,600 in fly, 25,000 in the small plant *Arabidopsis,* and probably 30,000 in mouse and <40,000 in Man.

As soon as we look at eukaryotic genomes, the relationship between genome size and gene number is lost. The genomes of unicellular eukaryotes fall in the same size range as the largest bacterial genomes. Higher eukaryotes have more genes, but the number does not correlate with genome size, as can be seen from **Figure 3.12.**

The most extensive data for lower eukaryotes are available from the sequences of the genomes of the yeasts *S. cerevisiae* and *S. pombe.* **Figure 3.13** summarizes the most important features. The yeast genomes of 13.5 Mb and 12.5 Mb have ~6000 and ~5000 genes, respectively. The average open reading frame is ~1.4 kb, so that ~70% of the genome is occupied by coding regions. The major difference between them is that only 5% of *S. cerevisiae* genes have introns, compared to 43% in 5. *pombe.* The density of genes is high; organization is generally similar, although the spaces between genes are a bit shorter in *S. cerevisiae.* About half of the genes identified by sequence were either known previously or related to known genes. The remainder are new, which gives some indication of the number of new types of genes that may be discovered.

The identification of long reading frames on the basis of sequence is quite accurate. However, ORFs coding for <100 amino acids cannot be identified solely by sequence because of the high occurrence of false positives. Analysis of gene expression suggests that ~300 of 600 such ORFs in *S. cerevisiae* are likely to be genuine genes.

The genome of *C. elegans* DNA varies between regions rich in genes and regions in which genes are more sparsely organized. The total sequence contains ~18,500 genes. Only ~42% of the genes have putative counterparts outside the Nematoda.

Although the fly genome is larger than the worm genome, there are fewer genes (13,600) in *D. melanogaster.* The number of different transcripts is slightly larger (14,100) as the result of alternative splicing. We do not understand why the fly—a much more complex organism—has only 70% of the number of genes in the worm. This emphasizes forcefully the lack of an exact relationship between gene number and complexity of the organism.

The plant *Arabidopsis thaliana* has a genome size intermediate between the worm and the fly, but has a larger gene number (25,000) than either. This again shows the lack of a clear relationship, and also emphasizes the special quality of plants, which may have more genes (due to ancestral duplications) than animal cells. A majority of the *Arabidopsis* genome is found in duplicated segments, suggesting that there was an ancient doubling of the genome (to give a tetraploid). Only 35% of *Arabidopsis* genes are present as single copies.

The genome of rice (*Oryza sativa)* is ~4× larger than *Arabidopsis,* but the number of genes is only ~50% larger, probably ~40,000. Repetitive DNA occupies 42-45% of the genome. More than 80% of the genes found in *Arabidopsis* are represented in rice. Of these common genes, ~8000 are found in *Arabidopsis* and rice but not in any of the

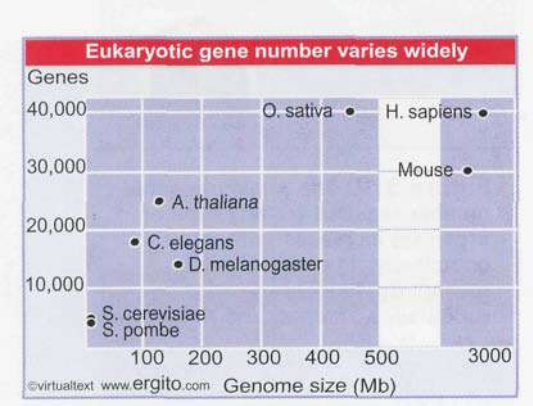


Figure 3.12 The number of genes in a eukaryote varies from 6000 to 40,000 but does not correlate with the genome size or the complexity of the organism.

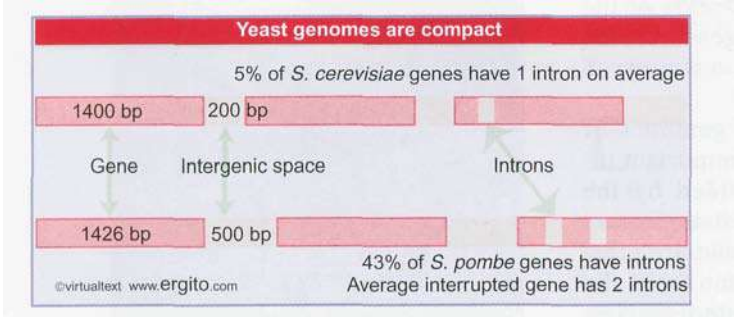


Figure 3.13 The S. *cerevisiae* genome of 13.5 Mb has 6000 genes, almost all uninterrupted. The S. *pombe* genome of 12.5 Mb has 5000 genes, almost half having introns. Gene sizes and spacing are fairly similar.

bacterial or animal genomes that have been sequenced. These are probably the set of genes that code for plant-specific functions, such as photosynthesis.

From the fly genome, we can form an impression of how many genes are devoted to each type of function. Figure **3.14** breaks down the functions into different categories. Among the genes that are identified, we find 2500 enzymes, ~750 transcription factors, ~700 transporters and ion channels, and ~700 proteins involved with signal transduction. But just over the half genes code for products of unknown function. Approximately 20% of the proteins reside in membranes.

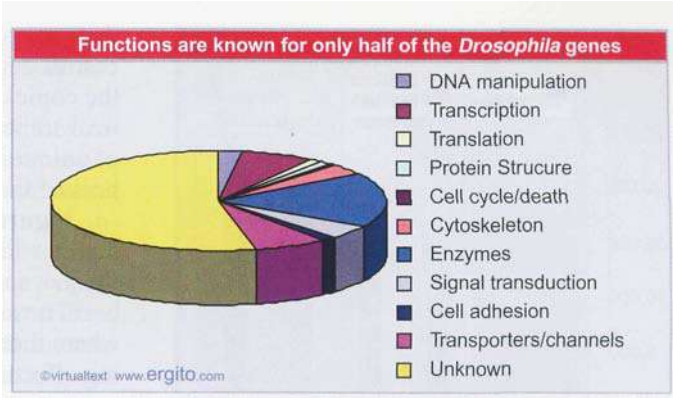


Figure 3.14 ~20% of *Drosophila* genes code for proteins concerned with maintaining or expressing genes, ~20% for enzymes, <10% for proteins concerned with the cell cycle or signal transduction. Half of the genes of *Drosophila* code for products of unknown function.

Protein size increases from prokaryotes and archaea to eukaryotes. The archaea *M. jannaschi* and bacterium *E. coli* have average protein lengths of 287 and 317 amino acids, respectively; whereas *S. cerevisiae* and *C. elegans* have average lengths of 484 and 442 amino acids, respectively. Large proteins (>500 amino acids) are rare in bacteria, but comprise a significant component (~l/3) in eukaryotes. The increase in length is due to the addition of extra domains, with each domain typically constituting 100-300 amino acids. But the increase in protein size is responsible for only a very small part of the increase in genome size.

Another insight into gene number is obtained by counting the number of expressed genes. If we rely upon the estimates of the number of different mRNA species that can be counted in a cell, we would conclude that the average vertebrate cell expresses ~10,000-20,000 genes. The existence of significant overlaps between the messenger populations in different cell types would suggest that the total expressed gene number for the organism should be within a few fold of this. The estimate for the total human genome number of 30,000-40,000 (see *3.11 The human genome has fewer genes than expected)* would imply that a significant proportion of the total gene number is actually expressed in any given cell.

Eukaryotic genes are transcribed individually, each gene producing a monocistronic messenger. There is only one general exception to this rule; in the genome of *C. elegans*, ~15% of the genes are organized into polycistronic units (which is associated with the use of frzms-splicing to allow expression of the downstream genes in these units; see *24.13 trans-splicing reactions use small RNAs).*

3.9 How many different types of genes are there?

Key Concepts

- Only some genes are unique; others belong to families where the other members are related (but not usually identical).
- The proportion of unique genes declines with genome size, and the proportion of genes in families increases.
- The minimum number of gene families required to code a bacterium is >1000, a yeast is >4000, and a higher eukaryote 11,000-14,000.

Because some genes are present in more than one copy or are related to one another, the number of different types of genes is less than the total number of genes. We can divide the total number of genes into sets that have related members, as defined by comparing

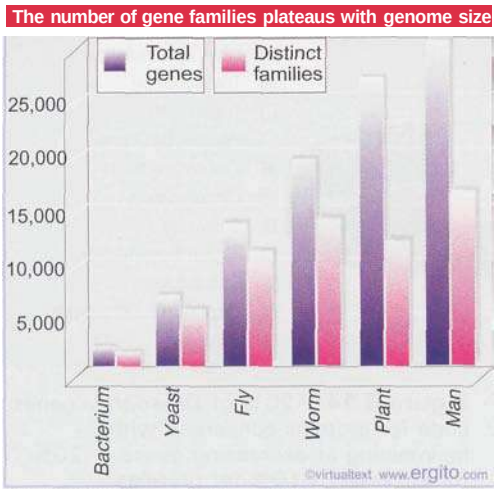


Figure 3.15 Because many genes are duplicated, the number of different gene families is much less than the total number of genes. The histogram compares the total number of genes with the number of distinct gene families.

Family size increases with genome size

Organism	Unique genes	Families with 2-4 members	Families with >4 members
H. influenzae	89%	10%	1%
S. cerevisiae	72%	19%	9%
D. melanogaster	72%	14%	14%
C. elegans	55%	20%	26%
A. thaliana	35%	24%	41%

Figure 3.16 The proportion of genes that are present in multiple copies increases with genome size in higher eukaryotes.

their exons. (A family of related genes arises by duplication of an ancestral gene followed by accumulation of changes in sequence between the copies. Most often the members of a family are related but not identical.) The number of types of genes is calculated by adding the number of unique genes (where there is no other related gene at all) to the numbers of families that have 2 or more members.

Figure 3.15 compares the total number of genes with the number of distinct families in each of six genomes. In bacteria, most genes are unique, so the number of distinct families is close to the total gene number. The situation is different even in the lower eukaryote *S. cerevisiae*, where there is a significant proportion of repeated genes. The most striking effect is that the number of genes increases quite sharply in the higher eukaryotes, but the number of gene families does not change much.

Figure 3.16 shows that the proportion of unique genes drops sharply with genome size. When genes are present in families, the number of members in a family is small in bacteria and lower eukaryotes, but is large in higher eukaryotes. Much of the extra genome size of *Arabidopsis* is accounted for by families with >4 members.

If every gene is expressed, the total number of genes will account for the total number of proteins required to make the organism (the proteome). However, two effects mean that the proteome is different from the total gene number. Because genes are duplicated, some of them code for the same protein (although it may be expressed in a different time or place) and others may code for related proteins that again play the same role in different times or places. And because some genes can produce more than one protein by means of alternative splicing, the proteome can be larger than the number of genes.

What is the core proteome—the basic number of the different types of proteins in the organism? A minimum estimate is given by the number of gene families, ranging from 1400 in the bacterium, >4000 in the yeast, and a range of 11,000-14,000 for the fly and worm.

What is the distribution of the proteome among types of proteins? The 6000 proteins of the yeast proteome include 5000 soluble proteins and 1000 transmembrane proteins. About half of the proteins are cytoplasmic, a quarter are in the nucleolus, and the remainder are split between the mitochondrion and the ER/Golgi system.

How many genes are common to all organisms (or to groups such as bacteria or higher eukaryotes) and how many are specific for the individual type of organism? **Figure 3.17** summarizes the comparison between yeast, worm, and fly. Genes that code for corresponding proteins in different organisms are called orthologs. Operationally, we usually reckon that two genes in different organisms can be considered to provide corresponding functions if their sequences are similar over >80% of the length. By this criterion, ~20% of the fly genes have orthologs in both yeast and the worm. These genes are probably required by all eukaryotes. The proportion increases to 30% when fly and worm are compared, probably representing the addition of gene functions that are common to multicellular eukaryotes. This still leaves a major proportion of genes as coding for proteins that are required specifically by either flies or worms, respectively

The proteome can be deduced from the number and structures o genes, and can also be directly measured by analyzing the total proteir content of a cell or organism. By such approaches, some proteins have been identified that were not suspected on the basis of genome analysis and that have therefore led to the identification of new genes. Severa methods are used for large scale analysis of proteins. Mass spectrometry can be used for separating and identifying proteins in a mixture obtainec directly from cells or tissues. Hybrid proteins bearing tags can be ob tained by expression of cDNAs made by linking the sequences of oper reading frames to appropriate expression vectors that incorporate the se quences for affinity tags. This allows array analysis to be used to analyz

the products. These methods also can be effective in comparing the proteins of two tissues, for example, a tissue from a normal individual and one from a patient with disease, to pinpoint the differences.

Once we know the total number of proteins, we can ask how they interact. By definition, proteins in structural multiprotein assemblies must form stable interactions with one another. Proteins in signaling pathways interact with one another transiently. In both cases, such interactions can be detected in test systems where essentially a readout system magnifies the effect of the interaction. One popular such system is the two hybrid assay discussed in *22.3 Independent domains bind DNA and activate transcription.* Such assays cannot detect all interactions: for example, if one enzyme in a metabolic pathway releases a soluble metabolite that then interacts with the next enzyme, the proteins may not interact directly.

As a practical matter, assays of pairwise interactions can give us an indication of the minimum number of independent structures or pathways. An analysis of the ability of all 6000 (predicted) yeast proteins to interact in pairwise combinations shows that ~1000 proteins can bind to at least one other protein. Direct analyses of complex formation have identified 1440 different proteins in 232 multiprotein complexes. This is the beginning of an analysis that will lead to definition of the number of functional assemblies or pathways.

In addition to functional genes, there are also copies of genes that have become nonfunctional (identified as such by various inactivating mutations). These are called pseudogenes (see *4.6 Pseudogenes are dead ends of evolution).* The number of pseudogenes can be large. In the mouse and human genomes, the number of pseudogenes is ~10% of the number of (potentially) active genes (see next section).

Besides needing to know the density of genes to estimate the total gene number, we must also ask: is it important in itself? Are there structural constraints that make it necessary for genes to have a certain spacing, and does this contribute to the large size of eukaryotic genomes?

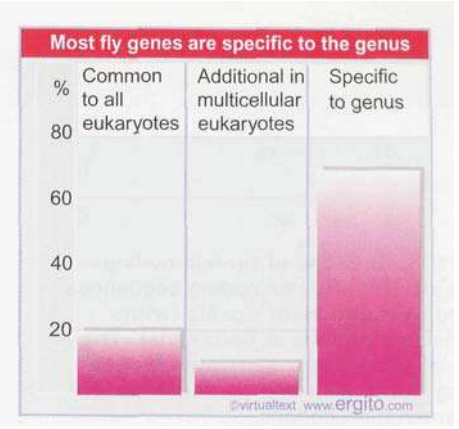


Figure 3.17 The fly genome can be divided into genes that are (probably) present in all eukaryotes, additional genes that are (probably) present in all multicellular eukaryotes, and genes that are more specific to subgroups of species that include flies.

3.10 The conservation of genome organization helps to identify genes

Key Concepts

- Algorithms for identifying genes are not perfect and many corrections must be made to the initial data set.
- Pseudogenes must be distinguished from active genes.
- Syntenic relationships are extensive between mouse and human genomes, and most active genes are in a syntenic region.

Once we have assembled the sequence of a genome, we still have to identify the genes within it. Coding sequences represent a very small fraction. Exons can be identified as uninterrupted open reading frames flanked by appropriate sequences. What criteria need to be satisfied to identify an active gene from a series of exons?

Figure 3.18 shows that an active gene should consist of a series of exons where the first exon immediately follows a promoter, the internal exons are flanked by appropriate splicing junctions, the last exon is followed by 3' processing signals, and a single open reading frame starting with an initiation codon and ending with a termination codon can be deduced by joining the exons together. Internal exons can be identified as open reading frames flanked by splicing junctions. In the simplest cases, the first and last exons contain the start and end of the coding

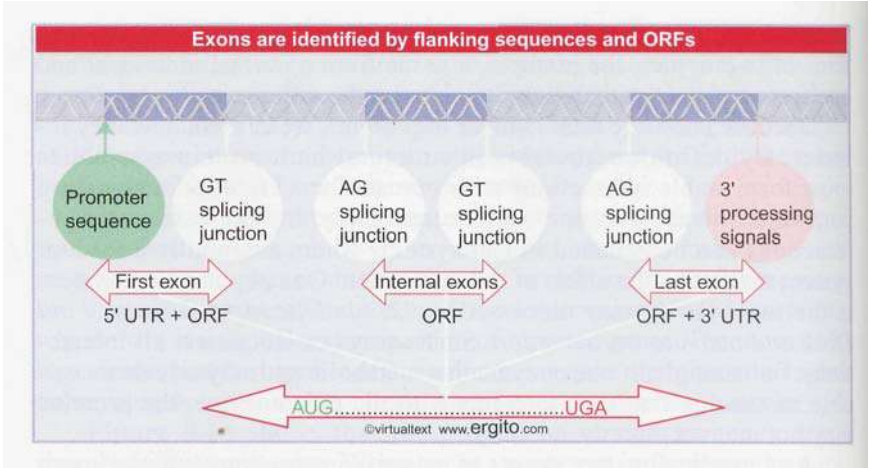


Figure 3.18 Exons of protein-coding genes are identified as coding sequences flanked by appropriate signals (with untranslated regions at both ends). The series of exons must generate an open reading frame with appropriate initiation and termination codons.

region, respectively, (as well as the 5' and 3' untranslated regions), but in more complex cases the first or last exons may have only untranslated regions, and may therefore be more difficult to identify.

The algorithms that are used to connect exons are not completely effective when the genome isVery large and the exons may be separated by very large distances. For example, the initial analysis of the human genome mapped 170,000 exons into 32,000 genes. This is unlikely to be correct, because it gives an average of 5.3 exons per gene, whereas the average of individual genes that have been fully characterized is 10.2. Either we have missed many exons, or they should be connected differently into a smaller number of genes in the whole genome sequence.

Even when the organization of a gene is correctly identified, there is the problem of distinguishing active genes from pseudogenes. Many pseudogenes can be recognized by obvious defects in the form of multiple mutations that create an inactive coding sequence. However, pseudogenes that have arisen more recently, and which have not accumulated so many mutations, may be more difficult to recognize. In an extreme example, the mouse has only one active *Gapdh* gene (coding for glyceraldehyde phosphate dehydrogenase), but has ~400 pseudogenes. However, >100 of these pseudogenes initially appeared to be active in the mouse genome sequence. Individual examination was necessary to exclude them from the list of active genes.

Confidence that a gene is active can be increased by comparing regions of the genomes of different species. There has been extensive overall reorganization of sequences between the mouse and human genomes, as seen in the simple fact that there are 23 chromosomes in the human haploid genome and 20 chromosomes in the mouse haploid genome. However, at the local level, the order of genes is generally the same: when pairs of human and mouse homologues are compared, the genes located on either side also tend to be homologues. This relationship is called synteny.

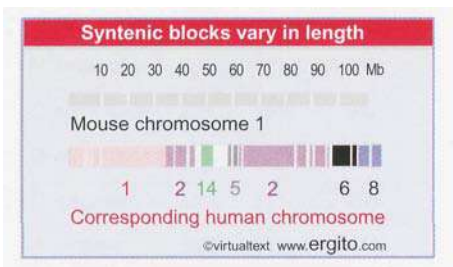


Figure 3.19 Mouse chromosome 1 has 21 segments of 1 - 25 Mb that are syntenic with regions corresponding to parts of 6 human chromosomes.

Figure 3.19 shows the relationship between mouse chromosome 1 and the human chromosomal set. We can recognize 21 segments in this mouse chromosome that have syntenic counterparts in human chromosomes. The extent of reshuffling that has occurred between the genomes is shown by the fact that the segments are spread among 6 different human chromosome. The same types of relationships are found in all mouse chromosomes, except for the X chromosome, which is syntenic only with the human X chromosome. This is explained by the fact that the X is a special case, subject to dosage compensation to adjust for the difference between males (one copy) and females (two copies) (see *23.17 X chromosomes undergo global changes)*. This may apply selective pressure against the translocation of genes to and from the X chromosome.

Comparison of the mouse and human genome sequences shows that >90% of each genome lies in syntenic blocks that range widely in size (from 300 kb to 65 Mb). There is a total of 342 syntenic segments, with

an average length of 7 Mb (0.3% of the genome). 99% of mouse genes have a homologue in the human genome; and for 96% that homologue is in a syntenic region.

Comparing the genomes provides interesting information about the evolution of species. The number of gene families in the mouse and human genomes is the same, and a major difference between the species is the differential expansion of particular families in one of the genomes. This is especially noticeable in genes that affect phenotypic features that are unique to the species. Of 25 families where the size has been expanded in mouse, 14 contain genes specifically involved in rodent reproduction, and 5 contain genes specific to the immune system.

A validation of the importance of syntenic blocks comes from pairwise comparisons of the genes within them. Looking for likely pseudogenes on the basis of sequence comparisons, a gene that is not in a syntenic location (that is, its context is different in the two species) is twice as likely to be a pseudogene. Put another way, translocation away from the original locus tends to be associated with the creation of pseudogenes. The lack of a related gene in a syntenic position is therefore grounds for suspecting that an apparent gene may really be a pseudogene. Overall, >10% of the genes that are initially identified by analysis of the genome are likely to turn out to be pseudogenes.

As a general rule, comparisons between genomes add significantly to the effectiveness of gene prediction. When sequence features indicating active genes are conserved, for example, between man and mouse, there is an increased probability that they identify active homologues.

Identifying genes coding for RNA is more difficult, because we cannot use the criterion of the open reading frame. It is true here also that comparative genome analysis increased the rigor of the analysis. For example, analysis of either the human or mouse genome alone identifies ~500 genes coding for tRNA in each case, but comparison of features suggests that <350 of these genes are in fact active in each genome.

3.11 The human genome has fewer genes than expected

Key Concepts

- Only 1% of the human genome consists of coding regions.
- The exons comprise ~5% of each gene, so genes (exons plus introns) comprise ~25% of the genome.
- The human genome has 30,000-40,000 genes.
- ~60% of human genes are alternatively spliced.
- Up to 80% of the alternative splices change protein sequence, so the proteome has ~50,000-60,000 members.

The human genome was the first vertebrate genome to be sequenced. This massive task has revealed a wealth of information about the genetic makeup of our species, and about the evolution of the genome in general. Our understanding is deepened further by the ability to compare the human genome sequence with the more recently sequenced mouse genome.

Mammal and rodent genomes generally fall into a narrow size range, ~3 X 10^9 bp (see 5.5 *Why are genomes so large?).* The mouse genome is ~14% smaller than the human genome, probably because it has had a higher rate of deletion. The genomes contain similar gene families and genes, with most genes having an ortholog in the other genome, but with differences in the number of members of a family, especially in those

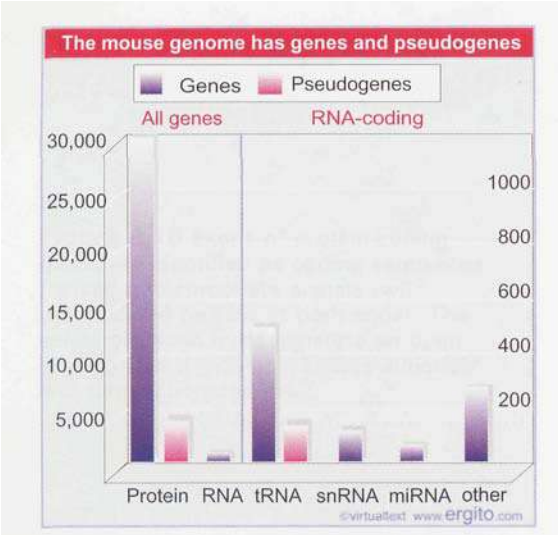


Figure 3.20 The mouse genome has ~ 30,000 protein-coding genes, which have ~4000 pseudogenes. There are ~800 RNA-coding genes. The data for RNA-coding genes are replotted on the right, at an expanded scale to show that there are ~350 tRNA genes and 150 pseudogenes, and ~450 other noncoding RNA genes, including snRNAs and miRNAs.

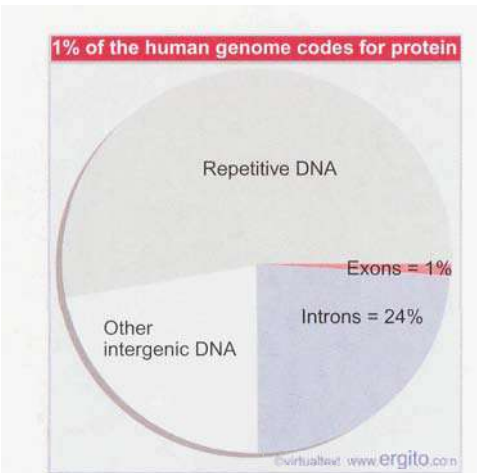


Figure 3.21 Genes occupy 25% of the human genome, but protein-coding sequences are only a tiny part of this fraction.

cases where the functions are specific to the species (see previous section). The estimate of 30,000 genes for the mouse genome is at the lower end of the range of estimates for the human genome. **Figure 3.20** plots the distribution of the mouse genes. The 30,000 protein-coding genes are accompanied by ~4000 pseudogenes. There are ~800 genes representing RNAs that do not code for proteins; these are generally small (aside from the rRNAs). Almost half of these genes code for tRNAs, for which a large number of pseudogenes also have been identified.

The human (haploid) genome contains 22 autosomes plus the X or Y. The chromosomes range in size from 45-279 Mb of DNA, making a total genome content of 3,286 Mb (~3.3 X 10^9 bp). On the basis of chromosome structure, the overall genome can be divided into regions of euchromatin (potentially containing active genes) and heterochromatin (see *19.7 Chromatin is divided into euchromatin and heterochromatin*). The euchromatin comprises the majority of the genome, ~2.9 X 10^9 bp. The identified genome sequence represents ~90% of the euchromatin. In addition to providing information on the genetic content of the genome, the sequence also identifies features that may be of structural importance (see *19.8Chromosomes have banding patterns).*

Figure 3.21 shows that a tiny proportion (~1%) of the human genome is accounted for by the exons that actually code for proteins. The introns that constitute the remaining sequences in the genes bring the total of DNA concerned with producing proteins to ~25%. As shown in **Figure 3.22,** the average human gene is 27 kb long, with 9 exons that include a total coding sequence of 1,340 bp. The average coding sequence is therefore only 5% of the length of the gene.

Based on comparisons with other species and with known protein-coding genes, there are ~24,000 clearly identifiable genes. Sequence analysis identifies ~12,000 more potential genes. Two independent analyses have produced estimates of ~30,000 and ~40,000 genes, respectively. One measure of the accuracy of the analyses is whether they identify the same genes. The surprising answer is that the overlap between the two sets of genes is only ~50%, as summarized in **Figure 3.23.** An earlier analysis of the human gene set based on RNA transcripts had identified ~11,000 genes, almost all of which are present in both the large human gene sets, and which account for the major part of the overlap between them. So there is no question about the authenticity of half of each human gene set, but we have yet to establish the relationship between the other half of each set. It is possible that they identify the same genes, but that the equivalencies have not been evident. In the extreme (but unlikely) case in which they corresponded to different but authentic genes, the total human gene number would inflate to ~50,000. The discrepancies illustrate the pitfalls of large scale sequence analysis!

But by any measure, the total human gene number is much less than we had expected—most previous estimates had been ~100,000. It shows a relatively small increase over flies and worms (13,600 and 18,500, respectively), not to mention the plant *Arabidopsis* (25,000) (see Figure 3.9). However, we should not be particularly surprised by the notion that it does not take a great number of additional genes to make a more complex organism. The difference in DNA sequences be-

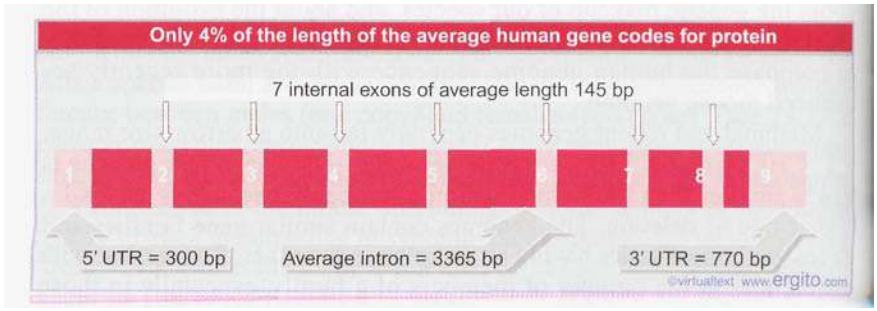


Figure 3.22 The average human gene is 27 kb long and has 9 exons, usually comprising two longer exons at each end and 7 internal exons. The UTRs in the terminal exons are the untranslated (noncoding) regions at each end of the gene. (This is based on the average. Because some genes are *extremely long,* the median length is 14 kb with 7 exons.)

tween Man and chimpanzee is extremely small (there is >99% similarity), so it is clear that the functions and interactions between a similar set of genes can produce very different results.

The number of genes is less than the number of potential proteins because of alternative splicing. The extent of alternative splicing is greater in man than in flies or worms; it may affect as many as 60% of the genes, so the increase in size of the human proteome relative to the other eukaryotes may be larger than the increase in the number of genes. A sample of genes from two chromosomes suggests that the proportion of the alternative splices that actually result in changes in the protein sequence may be as high as 80%. This could increase the size of the proteome to 50,000-60,000 members.

In terms of the diversity of the number of gene families, however, the discrepancy between Man and the other eukaryotes may not be so great. Many of the human genes belong to families. An analysis of ~25,000 genes identified 3500 unique genes and 10,300 gene pairs. As can be seen from Figure 3.15, this extrapolates to a number of gene families only slightly larger than worm or fly.

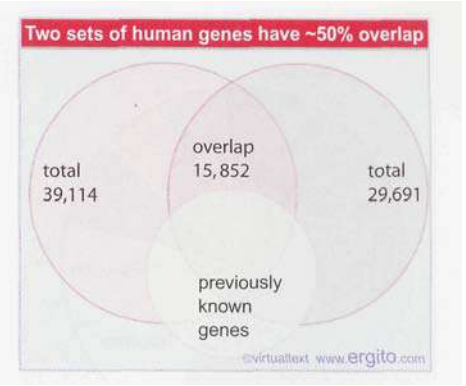


Figure 3.23 The two sets of genes identified in the human genome overlap only partially, as shown in the two large upper circles. However, they include almost all previously known genes, as shown by the overlap with the smaller, lower circle.

3.12 How are genes and other sequences distributed in the genome?

Key Concepts

- Repeated sequences (present in more than one copy) account for >50% of the human genome.
- The great bulk of repeated sequences consist of copies of nonfunctional transposons.
- There are many duplications of large chromosome regions.

Are genes uniformly distributed in the genome? Some chromosomes are relatively poor in genes, and have >25% of their sequences as "deserts"—regions longer than 500 kb where there are no genes. Even the most gene-rich chromosomes have >10% of their sequences as deserts. So overall ~20% of the human genome consists of deserts that have no genes.

Repetitive sequences account for >50% of the human genome, as seen in Figure 3.24. The repetitive sequences fall into five classes:

- Transposons (either active or inactive) account for the vast majority (45% of the genome). All transposons are found in multiple copies.
- Processed pseudogenes (~3000 in all, account for ~0.1% of total DNA). (These are sequences that arise by insertion of a copy of an mRNA sequence into the genome; see *4.6 Pseudogenes are dead ends of evolution).*
- Simple sequence repeats (highly repetitive DNA such as $(CA)_n$ account for ~3%).
- Segmental duplications (blocks of 10-300 kb that have been duplicated into a new region) account for ~5%. Only a minority of these duplications are found on the same chromosome; in the other cases, the duplicates are on different chromosomes.
- Tandem repeats form blocks of one type of sequence (especially found at centromeres and telomeres).

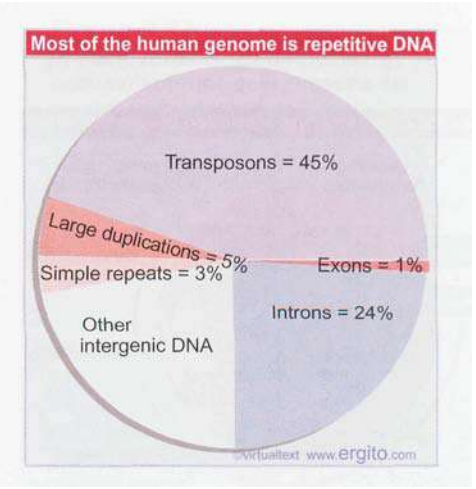


Figure 3.24 The largest component of the human genome consists of transposons. Other repetitive sequences include large duplications and simple repeats.

The sequence of the human genome emphasizes the importance of transposons. (Transposons have the capacity to replicate themselves and insert into new locations. They may function exclusively as DNA elements [see *16 Transposons]* or may have an active form that is RNA [see *17 Retroviruses and retroposons]*. Their distribution in the human

genome is summarized in Figure 17.18.) Most of the transposons in the human genome are nonfunctional; very few are currently active. However, the high proportion of the genome occupied by these elements indicates that they have played an active role in shaping the genome. One interesting feature is that some present genes originated as transposons, and evolved into their present condition after losing the ability to transpose. Almost 50 genes appear to have originated like this.

Segmental duplication at its simplest involves the tandem duplication of some region within a chromosome (typically because of an aberrant recombination event at meiosis; see *4.7 Unequal crossing-over rearranges gene clusters).* In many cases, however, the duplicated regions are on different chromosomes, implying that either there was originally a tandem duplication followed by a translocation of one copy to a new site, or that the duplication arose by some different mechanism altogether. The extreme case of a segmental duplication is when a whole genome is duplicated, in which case the diploid genome initially becomes tetraploid. As the duplicated copies develop differences from one another, the genome may gradually become effectively a diploid again, although homologies between the diverged copies leave evidence of the event. This is especially common in plant genomes. The present state of analysis of the human genome identifies many individual duplicated regions, but does not indicate whether there was a whole genome duplication in the vertebrate lineage.

3.13 More complex species evolve by adding new gene functions

Key Concepts

* Comparisons of different genomes show a steady increase in gene number as additional genes are added to make eukaryotes, make multicellular organisms, make animals, and make vertebrates.
* Most of the genes that are unique to vertebrates are concerned with the immune or nervous systems.

Comparison of the human genome sequence with sequences found in other species is revealing about the process of evolution. **Figure 3.25** analyzes human genes according to the breadth of their distribution in nature. Starting with the most generally distributed (top right corner of the figure), 21% of genes are common to eukaryotes and prokaryotes. These tend to code for proteins that are essential for all living forms—typically basic metabolism, replication, transcription, and translation. Moving clockwise, another 32% of genes are added in eukaryotes in general—for example, they may be found in yeast. These tend to code for proteins involved in functions that are general to eukaryotic cells but not to bacteria—for example, they may be concerned with specifying organelles or cytoskeletal components. Another 24% of genes are needed to specify animals. These include genes necessary for multicellularity and for development of different tissue types. And 22% of genes are unique to vertebrates. These mostly code for proteins of the immune and nervous systems; they code for very few enzymes, consistent with the idea that enzymes have ancient origins, and that metabolic pathways originated early in evolution. We see, therefore, that the progression from bacteria to vertebrates requires addition of groups of genes representing the necessary new functions at each stage.

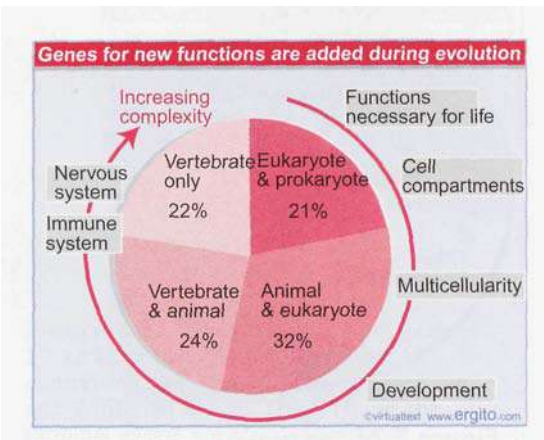


Figure 3.25 Human genes can be classified according to how widely their homologues are distributed in other species.

One way to define commonly needed proteins is to identify the proteins present in all proteomes. Comparing the human proteome in more

detail with the proteomes of other organisms, 46% of the yeast proteome, 43% of the worm proteome, and 61% of the fly proteome is represented in the human proteome. A key group of ~1300 proteins is present in all four proteomes. The common proteins are basic housekeeping proteins required for essential functions, falling into the types summarized in **Figure 3.26.** The main functions are concerned with transcription and translation (35%), metabolism (22%), transport (12%), DNA replication and modification (10%), protein folding and degradation (8%), and cellular processes (6%).

A small number of human genes (223) have homologues in bacteria, but not in the other eukaryotic genomes that have been sequenced. This suggests that they have not become part of the human genome by evolutionary descent from the (very ancient) last common ancestor shared with bacteria. These genes appear to be present in other vertebrates, but not in other classes of eukaryotes. One possible explanation is that they have become part of the vertebrate genome by direct (horizontal) transfer from bacteria to a vertebrate ancestor. A more complex alternative is that they have a more ancient origin, but have been lost separately in the other eukaryotic lineages.

One of the striking features of the human proteome is that it has many new proteins compared with other eukaryotes, but it has relatively few new protein domains. Most protein domains appear to be common to the animal kingdom. However, there are many new protein architectures, defined as new combinations of domains. **Figure 3.27** shows that the greatest increase occurs in transmembrane and extracellular proteins. In yeast, the vast majority of architectures are concerned with intracellular proteins. About twice as many intracellular architectures are found in fly (or worm), but there is a very striking increase in transmembrane and extracellular proteins, as might be expected from the addition of functions required for the interactions between the cells of a multicellular organism. The increase in intracellular architectures required to make a vertebrate (Man) is relatively small, but there is again a large increase in transmembrane and extracellular architectures.

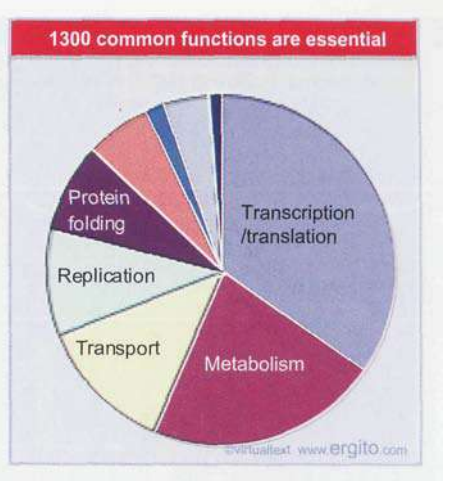


Figure 3.26 Common eukaryotic proteins are concerned with essential cellular functions.

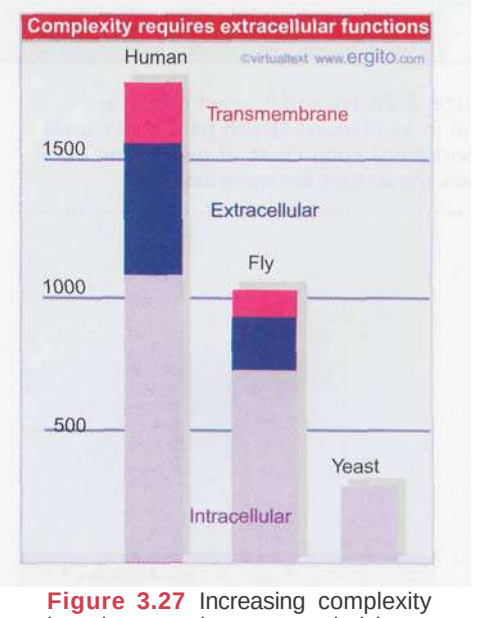


Figure 3.27 Increasing complexity in eukaryotes is accompanied by accumulation of new proteins for transmembrane and extracellular functions.

3.14 How many genes are essential?

Key Concepts

- Not all genes are essential. In yeast and fly, deletions of <50% of the genes have detectable effects.
- When two or more genes are redundant, a mutation in any one of them may not have detectable effects.
- We do not fully understand the survival in the genome of genes that are apparently dispensable.

Natural selection is the force that ensures that useful genes are retained in the genome. Mutations occur at random, and their most common effect in an open reading frame will be to damage the protein product. An organism with a damaging mutation will be at a disadvantage in evolution, and ultimately the mutation will be eliminated by the competitive failure of organisms carrying it. The frequency of a disadvantageous allele in the population is balanced between the generation of new mutations and the elimination of old mutations. Reversing this argument, whenever we see an intact open reading frame in the genome, we assume that its product plays a useful role in the organism. Natural selection must have prevented mutations from accumulating in the gene. The ultimate fate of a gene that ceases to be useful is to accumulate mutations until it is no longer recognizable.

By Book_Crazy [IND]

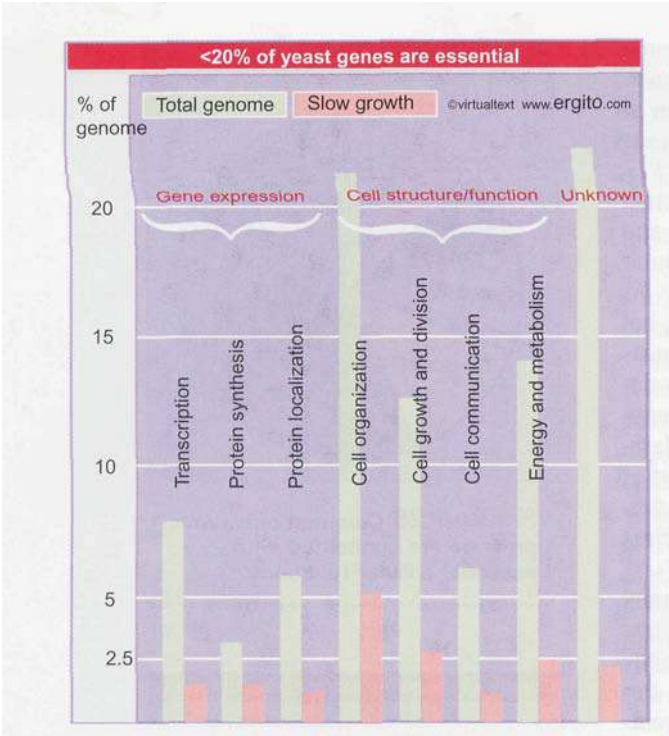


Figure 3.28 Essential yeast genes are found in all classes. Green bars show total proportion of each class of genes, red bars shows those that are essential.

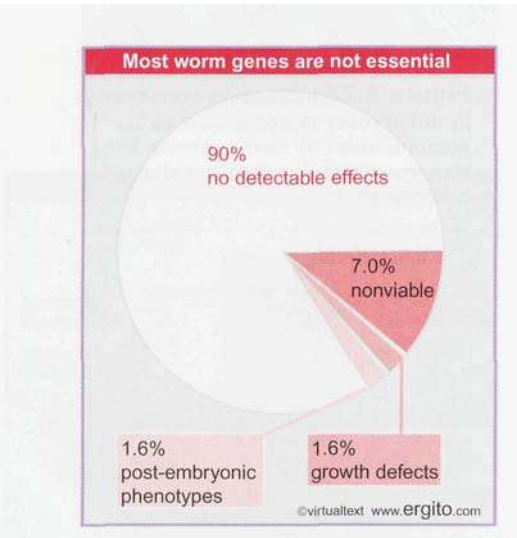


Figure 3.29 A systematic analysis of loss of function for 86% of worm genes shows that only 10% have detectable effects on the phenotype.

The maintenance of a gene implies that it confers a selective advantage on the organism. But in the course of evolution, even a small relative advantage may be the subject of natural selection, and a phenotypic defect may not necessarily be immediately detectable as the result of a mutation. However, we should like to know how many genes are actually *essential*. This means that their absence is lethal to the organism. In the case of diploid organisms, it means of course that the homozygous null mutation is lethal.

We might assume that the proportion of essential genes will decline with increase in genome size, given that larger genomes may have multiple, related copies of particular gene functions. So far this expectation has not been borne out by the data (see Figure 3.9).

One approach to the issue of gene number is to determine the number of essential genes by mutational analysis. If we saturate some specified region of the chromosome with mutations that are lethal, the mutations should map into a number of complementation groups that corresponds to the number of lethal loci in that region. By extrapolating to the genome as a whole, we may calculate the total essential gene number.

In the organism with the smallest known genome (*Mycoplasma genitalium*), random insertions have detectable effects only in about two thirds of the genes. Similarly, fewer than half of the genes of *E. coli* appear to be essential. The proportion is even lower in the yeast *S. cerevisiae*. When insertions were introduced at random into the genome in one early analysis, only 12% were lethal, and another 14% impeded growth. The majority (70%) of the insertions had no effect. A more systematic survey based on completely deleting each of 5,916 genes (>96% of the identified genes) shows that only 18.7% are essential for growth on a rich medium (that is, when nutrients are fully provided). **Figure 3.28** shows that these include genes in all categories. The only notable concentration of defects is in genes coding for products involved in protein synthesis, where ~50% are essential. Of course, this approach underestimates the number of genes that are essential for the yeast to live in the wild, when it is not so well provided with nutrients.

Figure 3.29 summarizes the results of a systematic analysis of the effects of loss of gene function in the worm *C. elegans*. The sequences of individual genes were predicted from the genome sequence, and by targeting an inhibitory RNA against these sequences (see *11.22 RNA interference is related to gene silencing*), a large set of worms were made in which one (predicted) gene was prevented from functioning in each worm. Detectable effects on the phenotype were only observed for 10% of these knockouts, suggesting that most genes do not play essential roles.

There is a greater proportion of essential genes (21%) among those worm genes that have counterparts in other eukaryotes, suggesting that widely conserved genes tend to play more basic functions. There is also an increased proportion of essential genes among those that are present in only one copy per haploid genome, compared with those where there are multiple copies of related or identical genes. This suggests that many of the multiple genes might be relatively recent duplications that can substitute for one another's functions.

Extensive analyses of essential gene number in a higher eukaryote have been made in *Drosophila* through attempts to correlate visible aspects of chromosome structure with the number of functional genetic units. The notion that this might be possible arose originally from the presence of bands in the polytene chromosomes of *D. melanogaster*. (These chromosomes are found at certain developmental stages and represent an unusually extended physical form, in which a series of bands

[more formally called chromomeres] are evident; see *19.10 Polytene chromosomes form bands.)* From the early concept that the bands might represent a linear order of genes, we have come to the attempt to correlate the organization of genes with the organization of bands. There are ~5000 bands in the *D. melanogaster* haploid set; they vary in size over an order of magnitude, but on average there is ~20 kb of DNA per band.

The basic approach is to saturate a chromosomal region with mutations. Usually the mutations are simply collected as lethals, without analyzing the cause of the lethality. Any mutation that is lethal is taken to identify a locus that is essential for the organism. Sometimes mutations cause visible deleterious effects short of lethality, in which case we also count them as identifying an essential locus. When the mutations are placed into complementation groups, the number can be compared with the number of bands in the region, or individual complementation groups may even be assigned to individual bands. The purpose of these experiments has been to determine whether there is a consistent relationship between bands and genes; for example, does every band contain a single gene?

Totaling the analyses that have been carried out over the past 30 years, the number of lethal complementation groups is ~70% of the number of bands. It is an open question whether there is any functional significance to this relationship. But irrespective of the cause, the equivalence gives us a reasonable estimate for the lethal gene number of ~3600. By any measure, the number of lethal loci in *Drosophila* is significantly less than the total number of genes.

If the proportion of essential human genes is similar to other eukaryotes, we would predict a range of 4000-8000 genes in which mutations would be lethal or produce evidently damaging effects. At the present, 1300 genes have been identified in which mutations cause evident defects. This is a substantial proportion of the expected total, especially in view of the fact that many lethal genes may act so early that we never see their effects. This sort of bias may also explain the results in Figure 3.30, which show that the majority of known genetic defects are due to point mutations (where there is more likely to be at least some residual function of the gene).

How do we explain the survival of genes whose deletion appears to have no effect? One possibility is that there is redundancy, that such genes are present in multiple copies. This is certainly true in some cases, in which multiple (related) genes must be knocked out in order to produce an effect. It is clear that there are cases in which a genome has more than one gene capable of providing a protein to fulfill a certain function, and all of them must be deleted to produce a lethal effect.

The idea that some genes are not essential (or at least cannot be shown to have serious effects upon the phenotype) raises some important questions. Does the genome contain genuinely dispensable genes, or do these genes actually have effects upon the phenotype that are significant at least during the long march of evolution? The theory of natural selection would suggest that the loss of individual genes in such circumstances produces a small disadvantage, which although not evident to us, is sufficient for the gene to be retained during the course of evolution.

Key questions that remain to be answered systematically are: What proportion of the total number of genes is essential, in how many do mutations produce at least detectable effects, and are there genes that are genuinely dispensable? Subsidiary questions about the genome as a whole are: What are the functions (if any) of DNA that does not reside in genes? What effect does a large change in total size have on the operation of the genome, as in the case of the related amphibians?

Most human mutations are small

Missense/nonsense	58%
Splicing	10%
Regulatory	<1%
Small deletions	16%
Small insertions	6%
Large deletions	5%
Large rearrangements	2%

Figure 3.30 Most known genetic defects in human genes are due to point mutations. The majority directly affect the protein sequence. The remainder are due to insertion, deletions, or rearrangements of varying sizes.

3.15 Genes are expressed at widely differing levels

Key Concepts

- In any given cell, most genes are expressed at a low level.
- Only a small number of genes, whose products are specialized for the cell type, are highly expressed.

The proportion of DNA represented in an mRNA population can be determined by the amount of the DNA that can hybridize with the RNA. Such a saturation analysis typically identifies ~1% of the DNA as providing a template for mRNA. From this we can calculate the number of genes so long as we know the average length of an mRNA. For a lower eukaryote such as yeast, the total number of expressed genes is ~4000. For somatic tissues of higher eukaryotes, the number usually is 10,000-15,000. The value is similar for plants and for vertebrates. (The only consistent exception to this type of value is presented by mammalian brain, where much larger numbers of genes appear to be expressed, although the exact quantitation is not certain.)

Kinetic analysis of the reassociation of an RNA population can be used to determine its sequence complexity. This type of analysis typically identifies three components in a eukaryotic cell. Just as with a DNA reassociation curve, a single component hybridizes over about two decades of Rot (RNA concentration X time) values, and a reaction extending over a greater range must be resolved by computer curve-fitting into individual components. Again this represents what is really a continuous spectrum of sequences.

An example of an excess mRNA X cDNA reaction that generates three components is given in **Figure 3.31:**

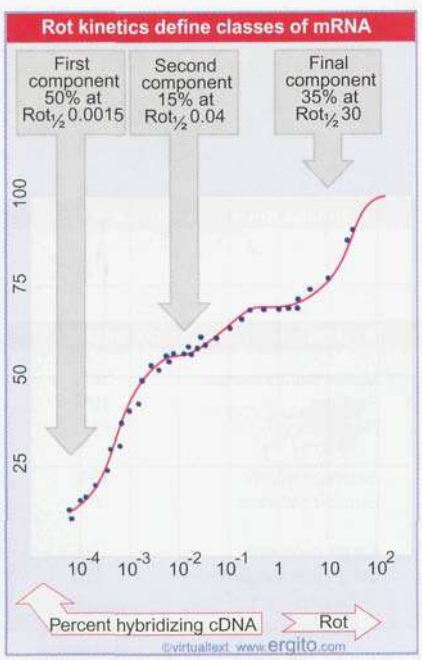


Figure 3.31 Hybridization between excess mRNA and cDNA identifies several components in chick oviduct cells, each characterized by the $Rot_{1/2}$ of reaction.

- The first component has the same characteristics as a control reaction of ovalbumin mRNA with its DNA copy. This suggests that the first component is in fact just ovalbumin mRNA (which indeed occupies about half of the messenger mass in oviduct tissue).
- The next component provides 15% of the reaction, with a total complexity of 15 kb. This corresponds to 7-8 mRNA species of average length 2000 bases.
- The last component provides 35% of the reaction, which corresponds to a complexity of 26 Mb. This corresponds to ~13,000 mRNA species of average length 2000 bases.

From this analysis, we can see that about half of the mass of mRNA in the cell represents a single mRNA, ~15% of the mass is provided by a mere 7-8 mRNAs, and ~35% of the mass is divided into the large number of 13,000 mRNA species. It is therefore obvious that the mRNAs comprising each component must be present in very different amounts.

The average number of molecules of each mRNA per cell is called its abundance. It can be calculated quite simply if the total mass of RNA in the cell is known. In the example shown in Figure 3.31, the total mRNA can be accounted for as 100,000 copies of the first component (ovalbumin mRNA), 4000 copies of each of the 7-8 mRNAs in the second component, but only ~5 copies of each of the 13,000 mRNAs that constitute the last component.

We can divide the mRNA population into two general classes, according to their abundance:

- The oviduct is an extreme case, with so much of the mRNA represented in only one species, but most cells do contain a small number of RNAs present in many copies each. This abundant mRNA component typically consists of <100 different mRNAs present in 1000-10,000 copies per cell. It often corresponds to a major part of the mass, approaching *50%* of *the total* mRNA.
- About half of the mass of the mRNA consists of a large number of sequences, of the order of 10,000, each represented by only a small number of copies in the mRNA—say, < 10. This is the scarce mRNA or complex mRNA class. It is this class that drives a saturation reaction.

3.16 How many genes are expressed?

Key Concepts

- mRNAs expressed at low levels overlap extensively when different cell types are compared.
- The abundantly expressed mRNAs are usually specific for the cell type.
- ~ 10,000 expressed genes may be common to most cell types of a higher eukaryote.

Many somatic tissues of higher eukaryotes have an expressed gene number in the range of 10,000-20,000. How much overlap is there between the genes expressed in different tissues? For example, the expressed gene number of chick liver is ~11,000-17,000, compared with the value for oviduct of ~13,000-15,000. How many of these two sets of genes are identical? How many are specific for each tissue? These questions are usually addressed by analyzing the transcriptome—the set of sequences represented in RNA.

We see immediately that there are likely to be substantial differences among the genes expressed in the abundant class. Ovalbumin, for example, is synthesized only in the oviduct, not at all in the liver. This means that 50% of the mass of mRNA in the oviduct is specific to that tissue.

But the abundant mRNAs represent only a small proportion of the number of expressed genes. In terms of the total number of genes of the organism, and of the number of changes in transcription that must be made between different cell types, we need to know the extent of overlap between the genes represented in the scarce mRNA classes of different cell phenotypes.

Comparisons between different tissues show that, for example, ~75% of the sequences expressed in liver and oviduct are the same. In other words, ~12,000 genes are expressed in both liver and oviduct, ~5000 additional genes are expressed only in liver, and ~3000 additional genes are expressed only in oviduct.

The scarce mRNAs overlap extensively. Between mouse liver and kidney, ~90% of the scarce mRNAs are identical, leaving a difference between the tissues of only 1000-2000 in terms of the number of expressed genes. The general result obtained in several comparisons of this sort is that only ~10% of the mRNA sequences of a cell are unique to it. The majority of sequences are common to many, perhaps even all, cell types.

This suggests that the common set of expressed gene functions, numbering perhaps ~10,000 in a mammal, comprise functions that are needed in all cell types. Sometimes this type of function is referred to

as a housekeeping gene or constitutive gene. It contrasts with the activities represented by specialized functions (such as ovalbumin or globin) needed only for particular cell phenotypes. These are sometimes called luxury genes.

3.17 Expressed gene number can be measured *en masse*

Key Concepts

- "Chip" technology allows a snapshot to be taken of the expression of the entire genome in a yeast cell.
- ~75% (~4500 genes) of the yeast genome is expressed under normal growth conditions.
- Chip technology allows detailed comparisons of related animal cells to determine (for example) the differences in expression between a normal cell and a cancer cell.

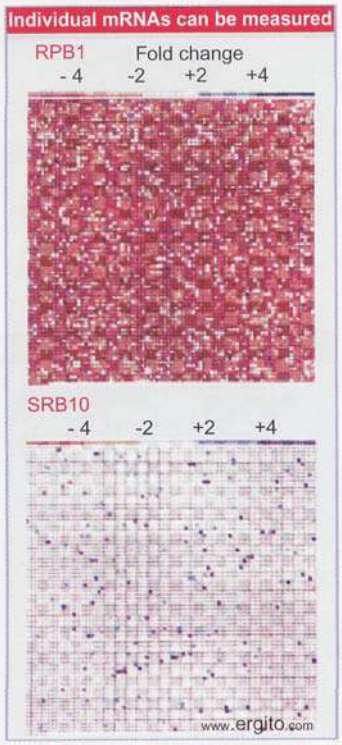


Figure 3.32 HDA analysis allows change in expression of each gene to be measured. Each square represents one gene (top left is first gene on chromosome I, bottom right is last gene on chromosome XVI). Change in expression relative to wild type is indicated by red (reduction), white (no change) or blue (increase). Photograph kindly provided by Rick Young.

Recent technology allows more systematic and accurate estimates of the number of expressed genes. One approach (SAGE, serial analysis of gene expression) allows a unique sequence tag to be used to identify each mRNA. The technology then allows the abundance of each tag to be measured. This approach identifies 4,665 expressed genes in *S. cerevisiae* growing under normal conditions, with abundances varying from 0.3 to >200 transcripts/cell. This means that ~75% of the total gene number (~6000) is expressed under these conditions.

The most powerful new technology uses chips that contain high-density oligonucleotide arrays (HDAs). Their construction is made possibly by knowledge of the sequence of the entire genome. In the case of *S. cerevisiae*, each of 6181 ORFs is represented on the HDA by 20 25-mer oligonucleotides that perfectly match the sequence of the message and 20 mismatch oligonucleotides that differ at one base position. The expression level of any gene is calculated by subtracting the average signal of a mismatch from its perfect match partner. The entire yeast genome can be represented on 4 chips. This technology is sensitive enough to detect transcripts of 5460 genes (~90% of the genome), and shows that 80% of genes are expressed at low levels, with abundances of 0.1-2 transcripts/cell. An abundance of <1 transcript/cell means that not all cells have a copy of the transcript at any given moment.

The technology allows not only measurement of levels of gene expression, but also detection of differences in expression in mutant cells compared with wild-type, cells growing under different growth conditions, and so on. The results of comparing two states are expressed in the form of a grid, in which each square represents a particular gene, and the relative change in expression is indicated by color. The upper part of **Figure 3.32** shows the effect of a mutation in RNA polymerase II, the enzyme that produces mRNA, which as might be expected causes the expression of most genes to be heavily reduced. By contrast, the lower part shows that a mutation in an ancillary component of the transcription apparatus (*SRB10*) has much more restricted effects, causing increases in expression of some genes.

The extension of this technology to animal cells will allow the general descriptions based on RNA hybridization analysis to be replaced by exact descriptions of the genes that are expressed, and the abundances of their products, in any given cell type.

3.18 Organelles have DNA

Key Concepts

- Mitochondria and chloroplasts have genomes that show nonMendelian inheritance. Typically they are maternally inherited.
- Organelle genomes may undergo somatic segregation in plants.
- Comparisons of mitochondrial DNA suggest that humans are descended from a single female who lived 200,000 years ago in Africa.

The first evidence for the presence of genes outside the nucleus was provided by non-Mendelian inheritance in plants (observed in the early years of this century, just after the rediscovery of Mendelian inheritance). Non-Mendelian inheritance is sometimes associated with the phenomenon of somatic segregation. They have a similar cause:

- Non-Mendelian inheritance is defined by the failure of the progeny of a mating to display Mendelian segregation for parental characters. It reflects lack of association between the segregating character and the meiotic spindle.
- Somatic segregation describes a phenomenon in which parental characters segregate in somatic cells, and therefore display heterogeneity in the organism. This is a notable feature of plant development. It reflects lack of association between the segregating character and the mitotic spindle.

Non-Mendelian inheritance and somatic segregation are therefore taken to indicate the presence of genes that reside outside the nucleus and do not utilize segregation on the meiotic and mitotic spindles to distribute replicas to gametes or to daughter cells, respectively. **Figure 3.33** shows that this happens when the mitochondria inherited from the male and female parents have different alleles, and by chance a daughter cell receives an unbalanced distribution of mitochondria that represents only one parent (see *13.24 How do mitochondria replicate and segregate?).*

The extreme form of non-Mendelian inheritance is uniparental inheritance, when the genotype of only one parent is inherited and that of the other parent is permanently lost. In less extreme examples, the progeny of one parental genotype exceed those of the other genotype. Usually it is the mother whose genotype is preferentially (or solely) inherited. This effect is sometimes described as maternal inheritance. The important point is that the genotype contributed by the parent of one particular sex predominates, as seen in abnormal segregation ratios when a cross is made between mutant and wild type. This contrasts with the behavior of Mendelian genetics when reciprocal crosses show the contributions of both parents to be equally inherited.

The bias in parental genotypes is established at or soon after the formation of a zygote. There are various possible causes. The contribution of maternal or paternal information to the organelles of the zygote may be unequal; in the most extreme case, only one parent contributes. In other cases, the contributions are equal, but the information provided by one parent does not survive. Combinations of both effects are possible. Whatever the cause, the unequal representation of the information from the two parents contrasts with nuclear genetic information, which derives equally from each parent.

Non-Mendelian inheritance results from the presence in mitochondria and chloroplasts of DNA genomes that are inherited independently of nuclear genes. In effect, the organelle genome comprises a length of DNA that has been physically sequestered in a defined part of the cell,

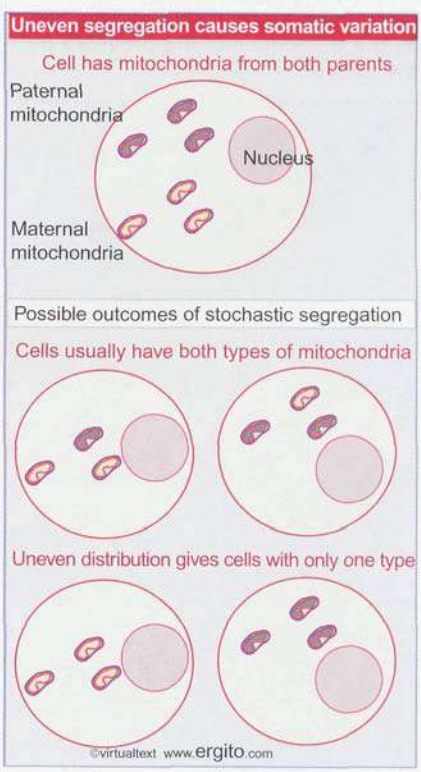


Figure 3.33 When paternal and maternal mitochondrial alleles differ, a cell has two sets of mitochondrial DNAs. Mitosis usually generates daughter cells with both sets. Somatic variation may result if unequal segregation generates daughter cells with only one set.

Animal mtDNA is inherited from the mother

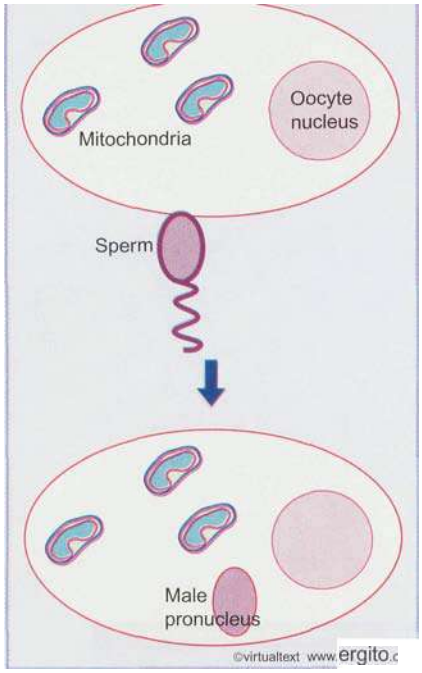


Figure 3.34 DNA from the sperm enters the oocyte to form the male pronucleus in the fertilized egg, but all the mitochondria are provided by the oocyte.

and is subject to its own form of expression and regulation. An organelle genome can code for some or all of the RNAs, but codes for only some of the proteins needed to perpetuate the organelle. The other proteins are coded in the nucleus, expressed via the cytoplasmic protein synthetic apparatus, and imported into the organelle.

Genes not residing within the nucleus are generally described as extranuclear genes; they are transcribed and translated in the *same* organelle compartment (mitochondrion or chloroplast) in which they reside. By contrast, *nuclear* genes are expressed by means of *cytoplasmic* protein synthesis. (The term cytoplasmic inheritance is sometimes used to describe the behavior of genes in organelles. However, we shall not use this description, since it is important to be able to distinguish between events in the general cytosol and those in specific organelles.)

Higher animals show maternal inheritance, which can be explained if the mitochondria are contributed entirely by the ovum and not at all by the sperm. **Figure 3.34** shows that the sperm contributes only a copy of the nuclear DNA. So the mitochondrial genes are derived exclusively from the mother; and in males they are discarded each generation.

Conditions in the organelle are different from those in the nucleus, and organelle DNA therefore evolves at its own distinct rate. If inheritance is uniparental, there can be no recombination between parental genomes; and usually recombination does not occur in those cases where organelle genomes are inherited from both parents. Since organelle DNA has a different replication system from that of the nucleus, the error rate during replication may be different. Mitochondrial DNA accumulates mutations more rapidly than nuclear DNA in mammals, but in plants the accumulation in the mitochondrion is slower than in the nucleus (the chloroplast is intermediate).

One consequence of maternal inheritance is that the sequence of mitochondrial DNA is more sensitive than nuclear DNA to reductions in the size of the breeding population. Comparisons of mitochondrial DNA sequences in a range of human populations allow an evolutionary tree to be constructed. The divergence among human mitochondrial DNAs spans 0.57%. A tree can be constructed in which the mitochondrial variants diverged from a common (African) ancestor. The rate at which mammalian mitochondrial DNA accumulates mutations is 2-4% per million years, >10× faster than the rate for globin. Such a rate would generate the observed divergence over an evolutionary period of 140,000-280,000 years. This implies that the human race is descended from a single female, who lived in Africa ~200,000 years ago.

3.19 Organelle genomes are circular DNAs that code for organelle proteins

Key Concepts

- Organelle genomes are usually (but not always) circular molecules of DNA.
- Organelle genomes code for some but not all of the proteins found in the organelle.

Most organelle genomes take the form of a single circular molecule of DNA of unique sequence (denoted mtDNA in the mitochondrion and ctDNA in the chloroplast). There are a few exceptions where mitochondrial DNA is a linear molecule, generally in lower eukaryotes.

Usually there are several copies of the genome in the individual organelle. Since there are multiple organelles per cell, there are many

organelle genomes per cell. Although the organelle genome itself is unique, it constitutes a repetitive sequence relative to any nonrepetitive nuclear sequence.

Chloroplast genomes are relatively large, usually ~140 kb in higher plants, and <200 kb in lower eukaryotes. This is comparable to the size of a large bacteriophage, for example, T4 at ~165 kb. There are multiple copies of the genome per organelle, typically 20-40 in a higher plant, and multiple copies of the organelle per cell, typically 20-40.

Mitochondrial genomes vary in total size by more than an order of magnitude. Animal cells have small mitochondrial genomes, ~16.5 kb in mammals. There are several hundred mitochondria per cell. Each mitochondrion has multiple copies of the DNA. The total amount of mitochondrial DNA relative to nuclear DNA is small, <1%.

In yeast, the mitochondrial genome is much larger. In *S. cerevisiae*, the * exact size varies among different strains, but is ~80 kb. There are ~22 mitochondria per cell, which corresponds to ~4 genomes per organelle. In growing cells, the proportion of mitochondrial DNA can be as high as 18%.

Plants show an extremely wide range of variation in mitochondrial DNA size, with a minimum of ~ 100 kb. The size of the genome makes it difficult to isolate intact, but restriction mapping in several plants suggests that the mitochondrial genome is usually a single sequence, organized as a circle. Within this circle, there are short homologous sequences. Recombination between these elements generates smaller, subgenomic circular molecules that coexist with the complete, "master" genome, explaining the apparent complexity of plant mitochondrial DNAs.

With mitochondrial genomes sequenced from many organisms, we can now see some general patterns in the representation of functions in mitochondrial DNA. **Figure 3.35** summarizes the distribution of genes in mitochondrial genomes. The total number of protein-coding genes is rather small, but does not correlate with the size of the genome. Mammalian mitochondria use their 16 kb genomes to code for 13 proteins, whereas yeast mitochondria use their 60-80 kb genomes to code for as few as 8 proteins. Plants, with much larger mitochondrial genomes, code for more proteins. Introns are found in most mitochondrial genomes, although not in the very small mammalian genomes.

The two major rRNAs are always coded by the mitochondrial genome. The number of tRNAs coded by the mitochondrial genome varies from none to the full complement (25-26 in mitochondria). This accounts for the variation in Figure 3.35.

The major part of the protein-coding activity is devoted to the components of the multisubunit assemblies of respiration complexes I-IV. Many ribosomal proteins are coded in protist and plant mitochondrial genomes, but there are few or none in fungi and animal genomes. There are genes coding for proteins involved in import in many protist mitochondrial genomes.

Mitochondria code for RNAs and proteins

Species	Size (kb)	Protein-coding genes	RNA-coding genes
Fungi	19-100	8-14	10-28
Protists	6-100	3-62	2-29
Plants	186-366	27-34	21-30
Animals	16-17	13	4-24

Figure 3.35 Mitochondrial genomes have genes coding for (mostly complex I-IV) proteins, rRNAs, and tRNAs.

3.20 Mitochondrial DNA organization is variable

Key Concepts

- Animal cell mitochondrial DNA is extremely compact and typically codes for 13 proteins, 2 rRNAs, and 22 tRNAs.
- Yeast mitochondrial DNA is 5x longer than animal cell mtDNA because of the presence of long introns.

Animal mitochondrial DNA is extremely compact. There are extensive differences in the detailed gene organization found in

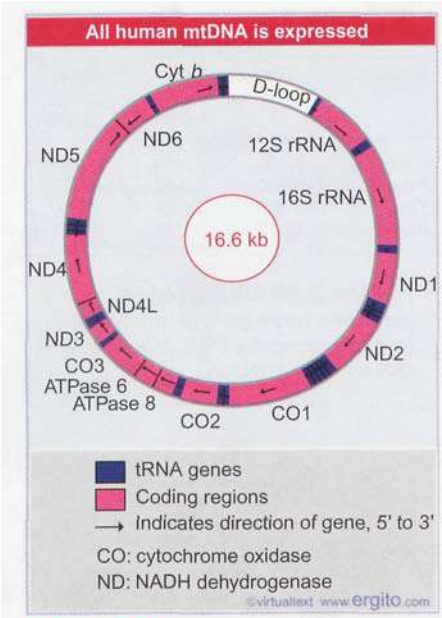


Figure 3.36 Human mitochondrial DNA has 22 tRNA genes, 2 rRNA genes, and 13 protein-coding regions. 14 of the 15 protein-coding or rRNA-coding regions are transcribed in the same direction. 14 of the tRNA genes are expressed in the clockwise direction and 8 are read counter clockwise.

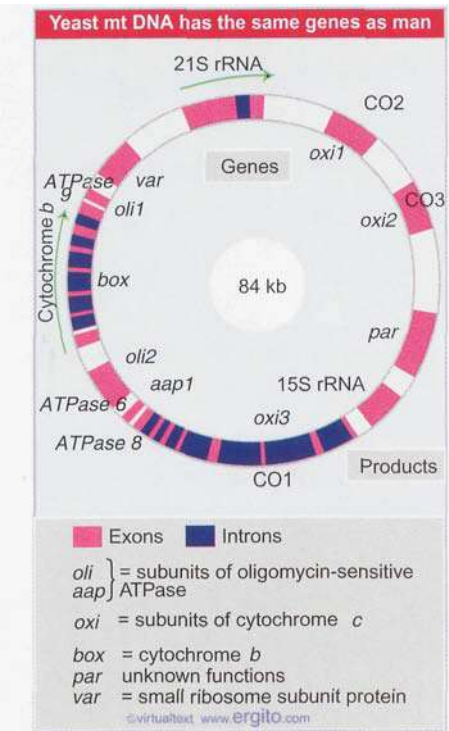


Figure 3.37 The mitochondrial genome of *S. cerevisiae* contains both interrupted and uninterrupted protein-coding genes, rRNA genes, and tRNA genes (positions not indicated). Arrows indicate direction of transcription.

different animal phyla, but the general principle is maintained of a small genome coding for a restricted number of functions. In mammalian mitochondrial genomes, the organization is extremely compact. There are no introns, some genes actually overlap, and almost every single base pair can be assigned to a gene. With the exception of the D loop, a region concerned with the initiation of DNA replication, no more than 87 of the 16,569 bp of the human mitochondrial genome can be regarded as lying in intercistronic regions.

The complete nucleotide sequences of mitochondrial genomes in animal cells show extensive homology in organization. The map of the human mitochondrial genome is summarized in **Figure 3.36.** There are 13 protein-coding regions. All of the proteins are components of the apparatus concerned with respiration. These include cytochrome *b,* 3 subunits of cytochrome oxidase, one of the subunits of ATPase, and 7 subunits (or associated proteins) of NADH dehydrogenase.

The five-fold discrepancy in size between the *S. cerevisiae* (84 kb) and mammalian (16 kb) mitochondrial genomes alone alerts us to the fact that there must be a great difference in their genetic organization in spite of their common function. The number of endogenously synthesized products concerned with mitochondrial enzymatic functions appears to be similar. Does the additional genetic material in yeast mitochondria represent other proteins, perhaps concerned with regulation, or is it unexpressed?

The map shown in **Figure 3.37** accounts for the major RNA and protein products of the yeast mitochondrion. The most notable feature is the dispersion of loci on the map.

The two most prominent loci are the interrupted genes *box* (coding for cytochrome *b)* and *oxi3* (coding for subunit 1 of cytochrome oxidase). Together these two genes are almost as long as the entire mitochondrial genome in mammals! Many of the long introns in these genes have open reading frames in register with the preceding exon (see *25.5 Some group I introns code for endonucleases that sponsor mobility*). This adds several proteins, all synthesized in low amounts, to the complement of the yeast mitochondrion.

The remaining genes are uninterrupted. They correspond to the other two subunits of cytochrome oxidase coded by the mitochondrion, to the subunit(s) of the ATPase, and (in the case of *var1*)to a mitochondrial ribosomal protein. The total number of yeast mitochondrial genes is unlikely to exceed ~25.

3.21 Mitochondria evolved by endosymbiosis

How did a situation evolve in which an organelle contains genetic information for some of its functions, while others are coded in the nucleus? **Figure 3.38** shows the endosymbiosis model for mitochondrial evolution,4n which primitive cells captured bacteria that provided the functions that evolved into mitochondria and chloroplasts. At this point, the proto-organelle must have contained all of the genes needed to specify its functions.

Sequence homologies suggest that mitochondria and chloroplasts evolved separately, from lineages that are common with eubacteria, with mitochondria sharing an origin with α-purple bacteria, and chloroplasts sharing an origin with cyanobacteria. The closest known relative of mitochondria among the bacteria is *Rickettsia* (the causative agent of typhus), which is an obligate intracellular parasite that is probably descended from free-living bacteria. This reinforces the idea that mitochondria originated in an endosymbiotic event involving an ancestor that is also common to *Rickettsia*.

Two changes must have occurred as the bacterium became integrated into the recipient cell and evolved into the mitochondrion (or chloroplast). The organelles have far fewer genes than an independent bacterium, and have lost many of the gene functions that are necessary for independent life (such as metabolic pathways). And since the majority of genes coding for organelle functions are in fact now located in the nucleus, these genes must have been transferred there from the organelle.

Transfer of DNA between organelle and nucleus has occurred over evolutionary time periods, and still continues. The rate of transfer can be measured directly by introducing into an organelle a gene that can function only in the nucleus, for example, because it contains a nuclear intron, or because the protein must function in the cytosol. In terms of providing the material for evolution, the transfer rates from organelle to nucleus are roughly equivalent to the rate of single gene mutation. DNA introduced into mitochondria is transferred to the nucleus at a rate of 2 X 10^{-5} per generation. Experiments to measure transfer in the reverse direction, from nucleus to mitochondrion, suggest that it is much lower, $<10^{-10}$. When a nuclear-specific antibiotic resistance gene is introduced into chloroplasts, its transfer to the nucleus and successful expression can be followed by screening seedlings for resistance to the antibiotic. This shows that transfer occurs at a rate of 1 in 16,000 seedlings, or 6 X 10^{-5}.

Transfer of a gene from an organelle to the nucleus requires physical movement of the DNA, of course, but successful expression also requires changes in the coding sequence. Organelle proteins that are coded by nuclear genes have special sequences that allow them to be imported into the organelle after they have been synthesized in the cytoplasm (see *8.17 Post-translational membrane insertion depends on leader sequences).* These sequences are not required by proteins that are synthesized within the organelle. Perhaps the process of effective gene transfer occurred at a period when compartments were less rigidly defined, so that it was easier both for the DNA to be relocated, and for the proteins to be incorporated into the organelle irrespective of the site of synthesis.

Phylogenetic maps show that gene transfers have occurred independently in many different lineages. It appears that transfers of mitochondrial genes to the nucleus occurred only early in animal cell evolution, but it is possible that the process is still continuing in plant cells. The number of transfers can be large; there are >800 nuclear genes in *Arabidopsis* whose sequences are related to genes in the chloroplasts of other plants. These genes are candidates for evolution from genes that originated in the chloroplast.

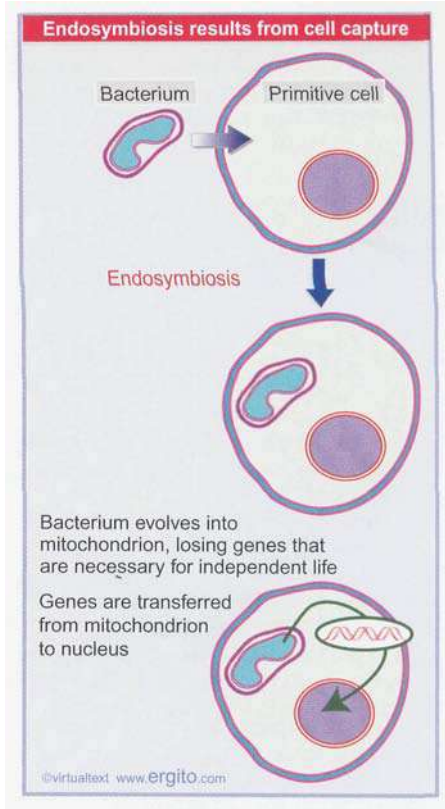


Figure 3.38 Mitochondria originated by a endosymbiotic event when a bacterium was captured by a eukaryotic cell.

3.22 The chloroplast genome codes for many proteins and RNAs

Key Concepts

- Chloroplast genomes vary in size, but are large enough to code for 50-100 proteins as well as the rRNAs and tRNAs.

What genes are carried by chloroplasts? Chloroplast DNAs vary in length from 120-190 kb. The sequenced chloroplast genomes (>10 in total) have from 87-183 genes. **Figure 3.39** summarizes the functions coded by the chloroplast genome in land plants. There is more variation in the chloroplast genomes of algae.

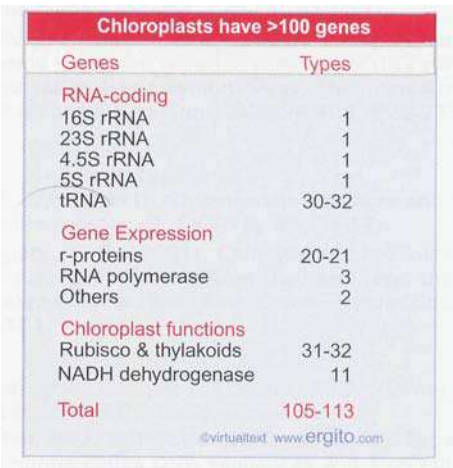


Figure 3.39 The chloroplast genome in land plants codes for 4 rRNAs, 30 tRNAs, and ~60 proteins.

By Book_Crazy [IND]

The situation is generally similar to that of mitochondria, except that more genes are involved. The chloroplast genome codes for all the rRNA and tRNA species needed for protein synthesis. The ribosome includes two small rRNAs in addition to the major species. The tRNA set may include all of the necessary genes. The chloroplast genome codes for ~50 proteins, including RNA polymerase and ribosomal proteins. Again the rule is that organelle genes are transcribed and translated by the apparatus of the organelle.

About half of the chloroplast genes code for proteins involved in protein synthesis. The endosymbiotic origin of the chloroplast is emphasized by the relationships between these genes and their counterparts in bacteria. The organization of the rRNA genes in particular is closely related to that of a cyanobacterium, which pins down more precisely the last common ancestor between chloroplasts and bacteria.

Introns in chloroplasts fall into two general classes. Those in tRNA genes are usually (although not inevitably) located in the anticodon loop, like the introns found in yeast nuclear tRNA genes (see *24.14 Yeast tRNA splicing involves cutting and rejoining).* Those in protein-coding genes resemble the introns of mitochondrial genes (see *25 Catalytic RNA).* This places the endosymbiotic event at a time in evolution before the separation of prokaryotes with uninterrupted genes.

The role of the chloroplast is to undertake photosynthesis. Many of its genes code for proteins of complexes located in the thylakoid membranes. The constitution of these complexes shows a different balance from that of mitochondrial complexes. Although some complexes are like mitochondrial complexes in having some subunits coded by the organelle genome and some by the nuclear genome, other chloroplast complexes are coded entirely by one genome.

3.23 Summary

Genomes that have been sequenced include many bacteria and archaea, yeasts, and a worm, fly, mouse, and man. The minimum number of genes required to make a living cell (an obligatory intracellular parasite) is ~470. The minimum number required to make a free-living cell is ~1700. A typical gram-negative bacterium has ~1500 genes. Strains of *E. coli* vary from 4300 to 5400 genes. The average bacterial gene is ~1000 bp long and is separated from the next gene by a space of ~100 bp. The yeasts *S. pombe* and *S. cerevisiae* have 5000 and 6000 genes, respectively.

Although the fly *D. melanogaster\s* a more complex organism and has a larger genome than the worm *C. elegans,* the fly has fewer genes (13,600) than the worm (18,500). The plant *Arabidopsis* has 25,000 genes, and the lack of a clear relationship between genome size and gene number is shown by the fact that the rice genome is 4x larger, but contains only a 50% increase in gene number, to ~40,000. Mouse has ~30,000 genes. Man has <40,000 genes, which is much less than had been expected. The complexity of development of an organism may depend on the nature of the interactions between genes as well as their total number.

About 8000 genes are common to prokaryotes and eukaryotes and are likely to be involved in basic functions. A further 12,000 genes are found in multicellular organisms. Another 8000 genes are added to make an animal, and a further 8000 (largely involved with the immune and nervous systems) are found in vertebrates. In each organism that has been sequenced, only ~50% of the genes have defined functions. Analysis of lethal genes suggests that only a minority of genes are essential in each organism.

The sequences comprising a eukaryotic genome can be classified in three groups: nonrepetitive sequences are unique; moderately repetitive sequences are dispersed and repeated a small number of

times in the form of related but not identical copies; and highly repetitive sequences are short and usually repeated as tandem arrays. The proportions of the types of sequence are characteristic for each genome, although larger genomes tend to have a smaller proportion of nonrepetitive DNA. Almost 50% of the human genome consists of repetitive sequences, the vast majority corresponding to transposon sequences. Most structural genes are located in nonrepetitive DNA. The complexity of nonrepetitive DNA is a better reflection of the complexity of the organism than the total genome complexity; nonrepetitive DNA reaches a maximum complexity of ~2 X 10^9 bp.

Genes are expressed at widely varying levels. There may be 10^5 copies of mRNA for an abundant gene whose protein is the principal product of the cell, 10^3 copies of each mRNA for <10 moderately abundant messages, and <10 copies of each mRNA for >10,000 scarcely expressed genes. Overlaps between the mRNA populations of ceils of different phenotypes are extensive; the majority of mRNAs are present in most cells.

Non-Mendelian inheritance is explained by the presence of DNA in organelles in the cytoplasm. Mitochondria and chloroplasts both represent membrane-bounded systems in which some proteins are synthesized within the organelle, while others are imported. The organelle genome is usually a circular DNA that codes for all of the RNAs and for some of the proteins that are required.

Mitochondrial genomes vary greatly in size from the 16 kb minimalist mammalian genome to the 570 kb genome of higher plants. It is assumed that the larger genomes code for additional functions. Chloroplast genomes range from 120-200 kb. Those that have been sequenced have a similar organization and coding functions. In both mitochondria and chloroplasts, many of the major proteins contain some subunits synthesized in the organelle and some subunits imported from the cytosol.

Mammalian mtDNAs are transcribed into a single transcript from the major coding strand, and individual products are generated by RNA processing. Rearrangements occur in mitochondrial DNA rather frequently in yeast; and recombination between mitochondrial or between chloroplast genomes has been found. Transfers of DNA have occurred from chloroplasts or mitochondria to nuclear genomes.

References

3.3 Individual genomes show extensive variation

ref Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., and Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513-516.

Mullikin, J. C, Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., Burton, J., Matthews, L. H., Pavitt, R., Plumb, R. W., Sims, S. K., Ainscough, R. M., Attwood, J., Bailey, J. M., Barlow, K., Bruskiewich, R. M., Butcher, P. N., Carter, N. P., Chen, Y., and Clee, C. M. (2000). An SNP map of human chromosome 22. Nature 407, 516-520.

3.4 RFLPs and SNPs can be used for genetic mapping

rev Gusella, J. F. (1986). DNA polymorphism and human disease. Ann. Rev. Biochem. 55, 831-854.

White, R. et al. (1985). Construction of linkage maps with DNA markers for human chromosomes. Nature 313, 101-105.

ref Dib, C. et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380, 152-154.

Dietrich, W. F. et al. (1996). A comprehensive genetic map of the mouse genome. Nature 380, 149-152.

Donis-Keller, J. et al. (1987). A genetic linkage map of the human genome. Cell 51, 319-337.

Sachidanandam, R. et al. (2001). A map of human genome sequence variation containing 1 .42 million single nucleotide polymorphisms. The International SNP Map Working Group. Nature 409, 928-933.

3.5 Why are genomes so large?

Gall, J. G. (1981). Chromosome structure and the C-value paradox. J. Cell Biol. 91, 3s-14s.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. Camb. Philos. Soc. 76, 65-101.

3.6 Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences

rev Britten, R. J. and Davidson, E. H. (1971). Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. 46, 111-133.

Davidson, E. H. and Britten, R. J. (1973). Organization, transcription, and regulation in the animal genome. Q. Rev. Biol. 48, 565-613.

3.7 Bacterial gene numbers range over an order of magnitude

rev Hacker, J. and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. Ann. Rev. Microbiol. 54, 641-679.

ref Blattner, F. et al. (1997). The complete genome sequence of *E. coli* K12. Science 277, 1453-1462.
Deckert, G. et al. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392, 353-358.
Galibert, F. et al. (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti.* Science 293, 668-672.

3.8 Total gene number is known for several eukaryotes

ref The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408, 796-815.
Adams, M. D. et al. (2000). The genome sequence of *D. melanogaster.* Science 287, 2185-2195.
C. elegans sequencing consortium. (1998). Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282, 2012-2022.
Dujon, B. et al. (1994). Complete DNA sequence of yeast chromosome XI. Nature 369, 371-378.
Goff, S. A. et al. (2002). A draft sequence of the rice genome *(Oryza sativa L. ssp. japonica).* Science 296, 92-114.
Johnston, M. et al. (1994). Complete nucleotide sequence of *S. cerevisiae* chromosome VIII. Science 265, 2077-2082.
Oliver, S. G. et al. (1992). The complete DNA sequence of yeast chromosome III. Nature 357, 38-46.
Wilson, R. et al. (1994). 22 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans.* Nature 368, 32-38.
Wood, V. et al. (2002). The genome sequence of *S. pombe.* Nature 415, 871-880.
Yu, J. et al. (2002). A draft sequence of the rice genome *(Oryza sativa L. ssp. indica).* Science 296, 79-92.

3.9 How many different types of genes are there?

rev Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M., and Fields, S. (2003). Protein analysis on a proteomic scale. Nature 422, 208-215.

ref The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408, 796-815.
Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. Nature 422, 198-207.
Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Miller, P., Gerstein, M., Roeder, G. S., and Snyder, M. (2002). Subcellular localization of the yeast proteome. Genes Dev. 1 6, 707-719.
Gavin, A. C, et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141-147.
Hanash, S. (2003). Disease proteomics. Nature 422, 226-232.
Ho, Y. et al. (2002). Systematic identification of protein complexes in *S. cerevisiae* by mass spectrometry. Nature 415, 180-183.
Rubin, G. M. et al. (2000). Comparative genomics of the eukaryotes. Science 287, 2204-2215.
Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. Nature 422, 216-225.
Uetz, P. et al. (2000). A comprehensive analysis of protein-protein interactions in *S. cerevisiae.* Nature 403, 623-630.
Venter, J. C. et al. (2001). The sequence of the human genome. Science 291, 1304-1350.

3.10 The conservation of genome organization helps to identify genes

ref Waterston et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

3.11 The human genome has fewer genes than expected

ref Waterston et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.
International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.
Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 106, 413-415.
Venter, J. C. et al. (2001). The sequence of the human genome. Science 291, 1304-1350.

3.14 How many genes are essential?

ref Giaever et al. (2002). Functional profiling of the *S. cerevisiae* genome. Nature 418, 387-391.
Goebl, M. G. and Petes, T. D. (1986). Most of the yeast genomic sequences are not essential for cell growth and division. Cell 46, 983-992.
Hutchison, C. A. et al. (1999). Global transposon mutagenesis and a minimal mycoplasma genome. Science 286, 2165-2169.
Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P., and Ahringer, J. (2003). Systematic functional analysis of the *C. elegans* genome using RNAi. Nature 421, 231-237.

3.15 Genes are expressed at widely differing levels

ref Hastie, N. B. and Bishop, J. 0. (1976). The expression of three abundance classes of mRNA in mouse tissues. Cell 9, 761-774.

3.17 Expressed gene number can be measured *en masse*

rev Young, R. A. (2000). Biomedical discovery with DNA arrays. Cell 102, 9-15.

ref Holstege, F. C. P. et al. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95, 717-728.
Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D. et al. (2000). Functional discovery via a compendium of expression profiles. Cell 102, 109-126.
Mikos, G. L. G. and Rubin, G. M. (1996). The role of the genome project in determining gene function: insights from model organisms. Cell 86, 521-529.
Velculescu, V. E. et al. (1997). Characterization of the yeast transcriptosome. Cell 88, 243-251.

3.18 Organelles have DNA

ref Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. Nature 325, 31-36.

3.19 Organelle genomes are circular DNAs that code for organelle proteins

rev Lang, B. F., Gray, M. W., and Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. Ann. Rev. Genet. 33, 351-397.

3.20 Mitochondrial DNA organization is variable

rev Attardi, G. (1985). Animal mitochondrial DNA: an extreme example of economy. Int. Rev. Cytol. 93, 93-146.

Boore, J. L (1999). Animal mitochondrial genomes. Nuc. Acids Res. 27, 1767-1780.

Clayton, D. A. (1984). Transcription of the mammalian mitochondrial genome. Ann. Rev. Biochem. 53, 573-594.

Gray, M. W. (1989). Origin and evolution of mitochondrial DNA. Ann. Rev. Cell Biol. 5, 25-50.

ref Anderson, S. et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457-465.

3.21 Mitochondria evolved by endosymbiosis

rev Lang, B. F., Gray, M. W., and Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. Ann. Rev. Genet. 33, 351-397.

ref The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408, 796-815.

Adams, K. L., Daley, D. 0., Qiu, Y. L, Whelan, J., and Palmer, J. D. (2000). Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408, 354-357.

Huang, C. Y., Ayliffe, M. A., and Timmis, J. N. (2003). Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422, 72-76.

Thorsness, P. E. and Fox, T. D. (1990). Escape of DNA from mitochondria to the nucleus in *S. cerevisiae.* Nature 346, 376-379.

3.22 The chloroplast genome codes for many proteins and RNAs

rev Shimada H, et al. (1991). Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nuc. Acids Res. 11, 983-995.
Palmer, J. D. (1985). Comparative organization of chloroplast genomes. Ann. Rev. Genet. 19, 325-354.
Sugiura, M., Hirose, T., and Sugita, M. (1998). Evolution and mechanism of translation in chloroplasts. Ann. Rev. Genet. 32, 437-459.

Chapter 4

Clusters and repeats

4.1 Introduction

A set of genes descended by duplication and variation from some ancestral gene is called a gene family. Its members may be clustered together or dispersed on different chromosomes (or a combination of both). Genome analysis shows that many genes belong to families; the 40,000 genes identified in the human genome fall into ~ 15,000 families, so the average gene has a couple of relatives in the genome (see Figure 3.15). Gene families vary enormously in the degree of relatedness between members, from those consisting of multiple identical members to those where the relationship is quite distant. Genes are usually related only by their exons, with introns having diverged (see *2.5 Exon sequences are conserved but introns vary).* Genes may also be related by only some of their exons, while others are unique (see *2.10 Some exons can be equated with protein functions).*

The initial event that allows related exons or genes to develop is a duplication, when a copy is generated of some sequence within the genome. Tandem duplication (when the duplicates remain together) may arise through errors in replication or recombination. Separation of the duplicates can occur by a translocation that transfers material from one chromosome to another. A duplicate at a new location may also be produced directly by a transposition event that is associated with copying a region of DNA from the vicinity of the transposon. Duplications may apply either to intact genes or to collections of exons or even individual exons. When an intact gene is involved, the act of duplication generates two copies of a gene whose activities are indistinguishable, but then usually the copies diverge as each accumulates different mutations.

The members of a well-related structural gene family usually have related or even identical functions, although they may be expressed at different times or in different cell types. So different globin proteins are expressed in embryonic and adult red blood cells, while different actins are utilized in muscle and nonmuscle cells. When genes have diverged significantly, or when only some exons are related, the proteins may have different functions.

Some gene families consist of identical members. Clustering is a prerequisite for maintaining identity between genes, although clustered genes are not necessarily identical. Gene clusters range from extremes where a duplication has generated two adjacent related genes to cases where hundreds of identical genes lie in a tandem array. Extensive tandem repetition of a gene may occur when the product is needed in unusually large amounts. Examples are the genes for rRNA or histone proteins. This creates a special situation with regards to the maintenance of identity and the effects of selective pressure.

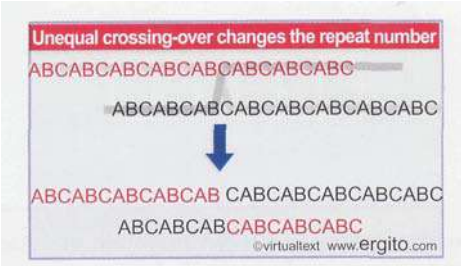


Figure 4.1 Unequal crossing-over results from pairing between non-equivalent repeats in regions of DNA consisting of repeating units. Here the repeating unit is the sequence ABC, and the third repeat of the red allele has aligned with the first repeat of the blue allele. Throughout the region of pairing, ABC units of one allele are aligned with ABC units of the other allele. Crossing-over generates chromosomes with 10 and 6 repeats each, instead of the 8 repeats of each parent.

Gene clusters offer us an opportunity to examine the forces involved in evolution of the genome over larger regions than single genes. Duplicated sequences, especially those that remain in the same vicinity, provide the substrate for further evolution by recombination. A population evolves by the classical recombination illustrated in Figure **1.31** and Figure 1.32, in which an exact crossing-over occurs. The recombinant chromosomes have the same organization as the parental chromosome. They contain precisely the same loci in the same order, but contain different combinations of alleles, providing the raw material for natural selection. However, the existence of duplicated sequences allows aberrant events to occur occasionally, changing the content of genes and not just the combination of alleles.

Unequal crossing-over describes a recombination event occurring between two sites that are not homologous. The feature that makes such events possible is the existence of repeated sequences. **Figure 4.1** shows that this allows one copy of a repeat in one chromosome to misalign for recombination with a different copy of the repeat in the homologous chromosome, instead of with the corresponding copy. When recombination occurs, this increases the number of repeats in one chromosome and decreases it in the other. In effect, one recombinant chromosome has a deletion and the other has an insertion. This mechanism is responsible for the evolution of clusters of related sequences. We can trace its operation in expanding or contracting the size of an array in both gene clusters and regions of highly repeated DNA.

The highly repetitive fraction of the genome consists of multiple tandem copies of very short repeating units. These often have unusual properties. One is that they may be identified as a separate peak on a density gradient analysis of DNA, which gave rise to the name satellite DNA. They are often associated with inert regions of the chromosomes, and in particular with centromeres (which contain the points of attachment for segregation on a mitotic or meiotic spindle). Because of their repetitive organization, they show some of the same behavior with regard to evolution as the tandem gene clusters. In addition to the satellite sequences, there are shorter stretches of DNA that show similar behavior, called minisatellites. They are useful in showing a high degree of divergence between individual genomes that can be used for mapping purposes.

All of these events that change the constitution of the genome are rare, but they are significant over the course of evolution.

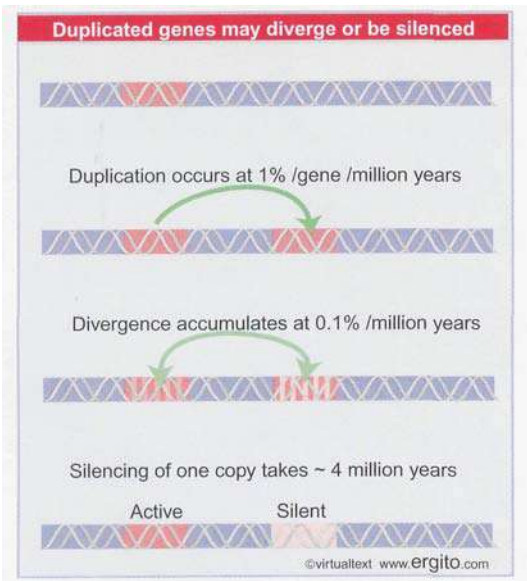


Figure 4.2 After a gene has been duplicated, differences may accumulate between the copies. The genes may acquire different functions or one of the copies may become inactive.

4.2 Gene duplication is a major force in evolution

Key Concepts

- Duplicated genes may diverge to generate different genes or one copy may become inactive.

Exons behave like modules for building genes that are tried out in the course of evolution in various combinations. At one extreme, an individual exon from one gene may be copied and used in another gene. At the other extreme, an entire gene, including both exons and introns, may be duplicated. In such a case, mutations can accumulate in one copy without attracting the adverse attention of natural selection. This copy may then evolve to a new function; it may become expressed in a different time or place from the first copy, or it may acquire different activities.

Figure 4.2 summarizes our present view of the rates at which these processes occur. There is ~1% probability that a given gene will be

included in a duplication in a period of 1 million years. After the gene has duplicated, differences develop as the result of the occurrence of different mutations in each copy. These accumulate at a rate of ~0.1% per million years (see *4.4 Sequence divergence is the basis for the evolutionary clock)*.

The organism is not likely to need to retain two identical copies of the gene. As differences develop between the duplicated genes, one of two types of event is likely to occur.

- Both of the genes become necessary. This can happen either because the differences between them generate proteins with different functions, or because they are expressed specifically in different times or places.
- If this does not happen, one of the genes is likely to be eliminated, because it will by chance gain a deleterious mutation, and there will be no adverse selection to eliminate this copy. Typically this takes ~4 million years. In such a situation, it is purely a matter of chance which of the two copies becomes inactive. (This can contribute to incompatibility between different individuals, and ultimately to speciation, if different copies become inactive in different populations.)

Analysis of the human genome sequence shows that ~5% comprises duplications of identifiable segments ranging in length from 10-300 kb. These have arisen relatively recently, that is, there has not been sufficient time for divergence between them to eliminate their relationship. They include a proportional share (~6%) of the expressed exons, which shows that the duplications are occurring more or less irrespective of genetic content. The genes in these duplications may be especially interesting because of the implication that they have evolved recently, and therefore could be important for recent evolutionary developments (such as the separation of man from the monkeys).

4.3 Globin clusters are formed by duplication and divergence

Key Concepts

- All globin genes are descended by duplication and mutation from an ancestral gene that had three exons.
- The ancestral gene gave rise to myoglobin, leghemoglobin, and α- and β-globins.
- The α- and β-globin genes separated in the period of early vertebrate evolution, after which duplications generated the individual clusters of separate α-like and β-like genes.
- Once a gene has been inactivated by mutation, it may accumulate further mutations and become a pseudogene, which is homologous to the active gene(s) but has no functional role.

The most common type of duplication generates a second copy of the gene close to the first copy. In some cases, the copies remain associated, and further duplication may generate a cluster of related genes. The best characterized example of a gene cluster is presented by the globin genes, which constitute an ancient gene family, concerned with a function that is central to the animal kingdom: the transport of oxygen through the bloodstream.

The major constituent of the red blood cell is the globin tetramer, associated with its heme (iron-binding) group in the form of hemoglobin. Functional globin genes in all species have the same general structure, divided into three exons as shown previously in Figure 2.7. We conclude that all globin genes are derived from a single ancestral gene; so by tracing the development of individual globin genes within and

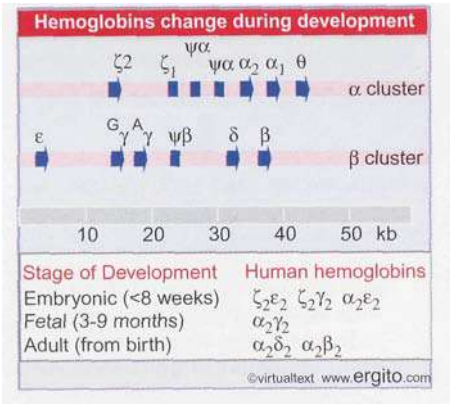


Figure 4.3 Each of the α-like **and** β-like globin gene families is organized into **a** single cluster that includes functional genes and pseudogenes (ψ).

between species, we may learn about the mechanisms involved in the evolution of gene families.

In adult cells, the globin tetramer consists of two identical a chains and two identical β chains. Embryonic blood cells contain hemoglobin tetramers that are different from the adult form. Each tetramer contains two identical α-like chains and two identical β-like chains, each of which is related to the adult polypeptide and is later replaced by it. This is an example of developmental control, in which different genes are successively switched on and off to provide alternative products that fulfill the same function at different times.

The details of the relationship between embryonic and adult hemoglobins vary with the organism. The human pathway has three stages: embryonic, fetal, and adult. The distinction between embryonic and adult is common to mammals, but the number of pre-adult stages varies. In man, zeta and alpha are the two α-like chains. Epsilon, gamma, delta, and beta are the β-like chains. The chains are expressed at different stages of development.

In the human pathway, ζ is the first α-like chain to be expressed, but is soon replaced by a. In the β-pathway, e and γ are expressed first, with 5 and β replacing them later. In adults, the $\alpha_2\beta_2$ form provides 97% of the hemoglobin, $\alpha_2\delta_2$ is ~2%, and ~1% is provided by persistence of the fetal form $\alpha_2\gamma_2$.

What is the significance of the differences between embryonic and adult globins? The embryonic and fetal forms have a higher affinity for oxygen. This is necessary in order to obtain oxygen from the mother's blood. This explains why there is no equivalent in (for example) chicken, where the embryonic stages occur outside the body (that is, within the egg).

The division of globin chains into α-like and β-like reflects the organization of the genes. Each type of globin is coded by genes organized into a single cluster. The structures of the two clusters in the higher primate genome are illustrated in **Figure 4.3.**

Stretching over 50 kb, the β cluster contains five functional genes (e, two 7, o. and β) and one nonfunctional gene (ψβ). The two 7 genes differ in their coding sequence in only one amino acid; the G variant has glycine at position 136, where the A variant has alanine.

The more compact a cluster extends over 28 kb and includes one active ζ gene, one ζ nonfunctional gene, two a genes, two a nonfunctional genes, and the 0 gene of unknown function. The two a genes code for the same protein. Two (or more) identical genes present on the same chromosome are described as nonallelic copies.

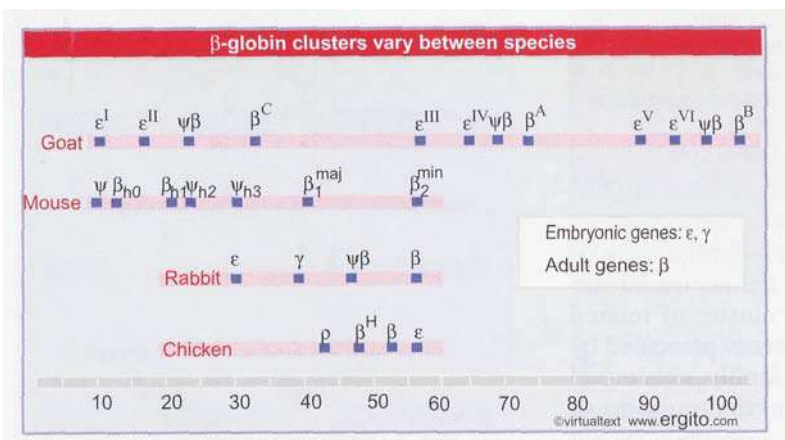


Figure 4.4 Clusters of β-globin genes **and** pseudogenes are found in vertebrates. Seven mouse genes include 2 early embryonic, 1 late embryonic, 2 adult genes, and 2 pseudogenes. Rabbit and chicken each have four genes.

Functional genes are defined by their expression in RNA, and ultimately by the proteins for which they code. Nonfunctional genes are defined as such by their inability to code for proteins; the reasons for inactivity vary, and the deficiencies may be in transcription or translation (or both). They are called pseudogenes and given the symbol ψ.

A similar general organization is found in other vertebrate globin gene clusters, but details of the types, numbers, and order of genes all vary, as illustrated in **Figure 4.4.** Each cluster contains both embryonic and adult genes. The total lengths of the clusters vary widely. The longest is found in the goat, where a basic cluster of 4 genes has been duplicated twice. The distribution of active genes and pseudogenes differs in each case, illustrating the random nature oi the conversion of one copy of a duplicated gene into the inactive state,

The characterization of these gene clusters makes an important general point. *There may be more members of a gene family, both functional and nonfunctional, than we would suspect on the basis of protein analysis.* Thi

extra functional genes may represent duplicates that code for identical polypeptides; or they may be related to known proteins, although different from them (and presumably expressed only briefly or in low amounts).

With regard to the question of how much DNA is needed to code for a particular function, we see that coding for the β-like globins requires a range of 20-120 kb in different mammals. This is much greater than we would expect just from scrutinizing the known β-globin proteins or even considering the individual genes. However, clusters of this type are not common; most genes are found as individual loci.

From the organization of globin genes in a variety of species, we should be able to trace the evolution of present globin gene clusters from a single ancestral globin gene. Our present view of the evolutionary descent is pictured in **Figure 4.5.**

The leghemoglobin gene of plants, which is related to the globin genes, may represent the ancestral form. The furthest back that we can trace a globin gene in modern form is provided by the sequence of the single chain of mammalian myoglobin, which diverged from the globin line of descent ~800 million years ago. The myoglobin gene has the same organization as globin genes, so we may take the three-exon structure to represent their common ancestor.

Some "primitive fish" have only a single type of globin chain, so they must have diverged from the line of evolution before the ancestral globin gene was duplicated to give rise to the a and β variants. This appears to have occurred ~500 million years ago, during the evolution of the bony fish.

The next stage of evolution is represented by the state of the globin genes in the frog *X. laevis,* which has two globin clusters. However, each cluster contains *both a* and β genes, of both larval and adult types. The cluster must therefore have evolved by duplication of a linked (X-β pair, followed by divergence between the individual copies. Later the entire cluster was duplicated.

The amphibians separated from the mammalian/avian line ~350 million years ago, so the separation of the a- and β-globin genes must have resulted from a transposition in the mammalian/avian forerunner after this time. This probably occurred in the period of early vertebrate evolution. Since there are separate clusters for a and β globins in both birds and mammals, the α and β genes must have been physically separated before the mammals and birds diverged from their common ancestor, an event that occurred probably ~270 million years ago.

Changes have occurred within the separate a and β clusters in more recent times, as we see from the description of the divergence of the individual genes in the next section.

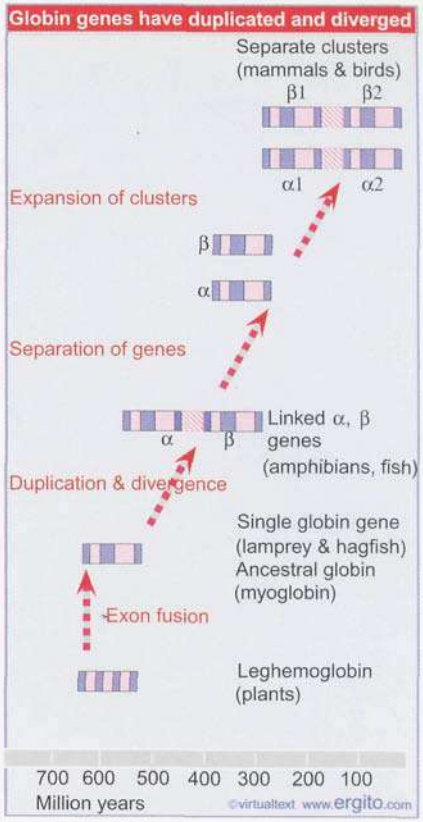


Figure 4.5 All globin genes have evolved by a series of duplications, transpositions, and mutations from a single ancestral gene.

4.4 Sequence divergence is the basis for the evolutionary clock

Key Concepts

- The sequences of homologous genes in different species vary at replacement sites (where mutation causes amino acid substitutions) and silent sites (where mutation does not affect the protein sequence).
- Mutations accumulate at silent sites ~10× faster than at replacement sites.
- The evolutionary divergence between two proteins is measured by the per cent of positions at which the corresponding amino acids are different.
- Mutations accumulate at a more or less even speed after genes separate, so that the divergence between any pair of globin sequences is proportional to the time since their genes separated.

Most changes in protein sequences occur by small mutations that accumulate slowly with time. Point mutations and small insertions and deletions occur by chance, probably with more or less equal probability in all regions of the genome, except for hotspots at which mutations occur much more frequently. Most mutations that change the amino acid sequence are deleterious and will be eliminated by natural selection.

Few mutations are advantageous, but when a rare one occurs, it is likely to spread through the population, eventually replacing the former sequence. When a new variant replaces the previous version of the gene, it is said to have become fixed in the population.

A contentious issue is what proportion of mutational changes in an amino acid sequence are neutral, that is, without any effect on the function of the protein, and able therefore to accrue as the result of random drift and fixation.

The rate at which mutational changes accumulate is a characteristic of each protein, presumably depending at least in part on its flexibility with regard to change. Within a species, a protein evolves by mutational substitution, followed by elimination or fixation within the single breeding pool. Remember that when we scrutinize the gene pool of a species, we see only the variants that have survived. When multiple variants are present, they may be stable (because neither has any selective advantage) or one may in fact be transient because it is in process of being displaced.

When a species separates into two new species, each now constitutes an independent pool for evolution. By comparing the corresponding proteins in two species, we see the differences that have accumulated between them *since the time when their ancestors ceased to interbreed*. Some proteins are highly conserved, showing little or no change from species to species. This indicates that almost any change is deleterious and therefore selected against.

The difference between two proteins is expressed as their divergence, the percent of positions at which the amino acids are different. The divergence between proteins can be different from the divergence between the corresponding nucleic acid sequences. The source of this difference is the representation of each amino acid in a three-base codon, in which often the third base has no effect on the meaning.

We may divide the nucleotide sequence of a coding region into potential replacement sites and silent sites:

- At replacement sites, a mutation alters the amino acid that is coded. The effect of the mutation (deleterious, neutral, or advantageous) depends on the result of the amino acid replacement.
- At silent sites, mutation only substitutes one synonym codon for another, so there is no change in the protein. Usually the replacement sites account for 75% of a coding sequence and the silent sites provide 25%.

In addition to the coding sequence, a gene contains nontranslated regions. Here again, mutations are potentially neutral, apart from their effects on either secondary structure or (usually rather short) regulatory signals.

Although silent mutations are neutral with regard to the protein, j they could affect gene expression via the sequence change in RNA. For example, a change in secondary structure might influence transcription, processing, or translation. Another possibility is that a change in synonym codons calls for a different tRNA to respond, influencing the efficiency of translation.

The mutations in replacement sites should correspond with the amino acid divergence (determined by the percent of changes in the protein sequence). A nucleic acid divergence of 0.45% at replacement

sites corresponds to an amino acid divergence of 1 % (assuming that the average number of replacement sites per codon is 2.25). Actually, the measured divergence underestimates the differences that have occurred during evolution, because of the occurrence of multiple events at one codon. Usually a correction is made for this.

To take the example of the human β- and δ-globin chains, there are 10 differences in 146 residues, a divergence of 6.9%. The DNA sequence has 31 changes in 441 residues. However, these changes are distributed very differently in the replacement and silent sites. There are 11 changes in the 330 replacement sites, but 20 changes in only 111 silent sites. This gives (corrected) rates of divergence of 3.7% in the replacement sites and 32% in the silent sites, almost an order of magnitude in difference.

The striking difference in the divergence of replacement and silent sites demonstrates the existence of much greater constraints on nucleotide positions that influence protein constitution relative to those that do not. So probably very few of the amino acid changes are neutral.

Suppose we take the rate of mutation at silent sites to indicate the underlying rate of mutational fixation (this assumes that there is no selection at all at the silent sites). Then over the period since the β and 8 genes diverged, there should have been changes at 32% of the 330 replacement sites, a total of 105. All but 11 of them have been eliminated, which means that ~90% of the mutations did not survive.

The divergence between any pair of globin sequences is (more or less) proportional to the time since they separated. This provides an evolutionary clock that measures the accumulation of mutations at an apparently even rate during the evolution of a given protein.

The rate of divergence can be measured as the percent difference per million years, or as its reciprocal, the unit evolutionary period (UEP), the time in millions of years that it takes for 1% divergence to develop. Once the clock has been established by pairwise comparisons between species (remembering the practical difficulties in establishing the actual time of speciation), it can be applied to related genes *within* a species. From their divergence, we can calculate how much time has passed since the duplication that generated them.

By comparing the sequences of homologous genes in different species, the rate of divergence at both replacement and silent sites can be determined, as plotted in **Figure 4.6.**

In pairwise comparisons, there is an average divergence of 10% in the replacement sites of either the a- or β-globin genes of mammals that have been separated since the mammalian radiation occurred ~85 million years ago. This corresponds to a replacement divergence rate of 0.12% per million years.

The rate is steady when the comparison is extended to genes that diverged in the more distant past. For example, the average replacement divergence between corresponding mammalian and chicken globin genes is 23%. Relative to a separation ~270 million years ago, this gives a rate of 0.09% per million years.

Going further back, we can compare the a- with the β-globin genes within a species. They have been diverging since the individual gene types separated ≥500 million years ago (see Figure 4.5). They have an average replacement divergence of ~50%, which gives a rate of 0.1% per million years.

The summary of these data in Figure 4.6 shows that replacement divergence in the globin genes has an average rate of ~0.096% per million years (or a UEP of 10.4). Considering the uncertainties in estimating the times at which the species diverged, the results lend good support to the idea that there is a linear clock.

The data on silent site divergence are much less clear. In every case, it is evident that the silent site divergence is much greater than the

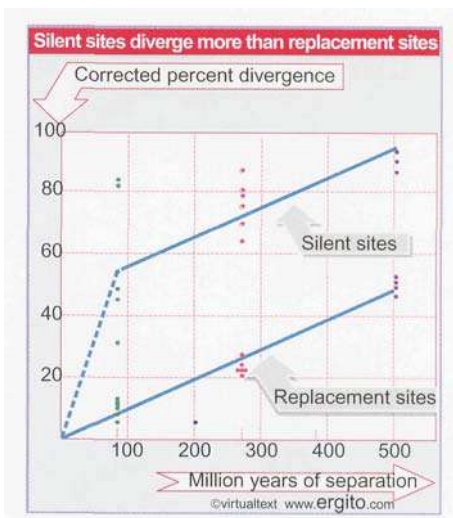


Figure 4.6 Divergence of DNA sequences depends on evolutionary separation. Each point on the graph represents a pairwise comparison.

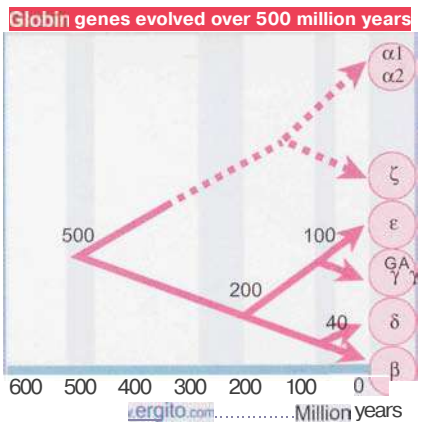


Figure 4.7 Replacement site divergences between pairs of β-globin genes allow the history of the human cluster to be reconstructed. This tree accounts for the separation of classes of globin genes.

replacement site divergence, by a factor that varies from 2 to 10. But the spread of silent site divergences in pairwise comparisons is too great to show whether a clock is applicable (so we must base temporal comparisons on the replacement sites).

From Figure 4.6, it is clear that the rate at silent sites is not linear with regard to time. *If we assume that there must be zero divergence at zero years of separation,* we see that the rate of silent site divergence is much greater for the first ~100 million years of separation. One interpretation is that a fraction of roughly half of the silent sites is rapidly (within 100 million years) saturated by mutations; this fraction behaves as neutral sites. The other fraction accumulates mutations more slowly, at a rate approximately the same as that of the replacement sites; this fraction identifies sites that are silent with regard to the protein, but that come under selective pressure for some other reason.

Now we can reverse the calculation of divergence rates to estimate the times since genes within a species have been apart. The difference between the human β and δ genes is 3.7% for replacement sites. At a UEP of 10.4, these genes must have diverged 10.4 X 3.7 = 40 million years ago—about the time of the separation of the lines leading to New World monkeys, Old World monkeys, great apes, and man. All of these higher primates have both β and δ genes, which suggests that the gene divergence commenced just before this point in evolution.

Proceeding further back, the divergence between the replacement sites of γ and e genes is 10%, which corresponds to a time of separation ~100 million years ago. The separation between embryonic and fetal globin genes therefore may have just preceded or accompanied the mammalian radiation.

An evolutionary tree for the human globin genes is constructed in **Figure 4.7.** Features that evolved before the mammalian radiation—such as the separation of β/δ from γ—should be found in all mammals. Features that evolved afterward—such as the separation of β- and δ-globin genes—should be found in individual lines of mammals.

In each species, there have been comparatively recent changes in the structures of the clusters, since we see differences in gene number (one adult β-globin gene in man, two in mouse) or in type (most often concerning whether there are separate embryonic and fetal genes).

When sufficient data have been collected on the sequences of a particular gene, the arguments can be reversed, and comparisons between genes in different species can be used to assess taxonomic relationships.

4.5 The rate of neutral substitution can be measured from divergence of repeated sequences

Key Concepts

- The rate of substitution per year at neutral sites is greater in the mouse than in the human genome.

We can make the best estimate of the rate of substitution at neutral sites by examining sequences that do not code for protein. (We use the term neutral here rather than silent, because there is no coding potential). An informative comparison can be made by comparing the members of common repetitive family in the human and mouse genomes.

The principle of the analysis is summarized in **Figure 4.8.** We start with a family of related sequences that have evolved by duplication and substitution from an original family member. We assume that the common ancestral sequence can be deduced by taking the base that is most common at each position. Then we can calculate the divergence of each individual family member as the proportion of bases that differ from the deduced ancestral sequence. In this example, individual members vary from 0.13-0.18 divergence, and the average is 0.16.

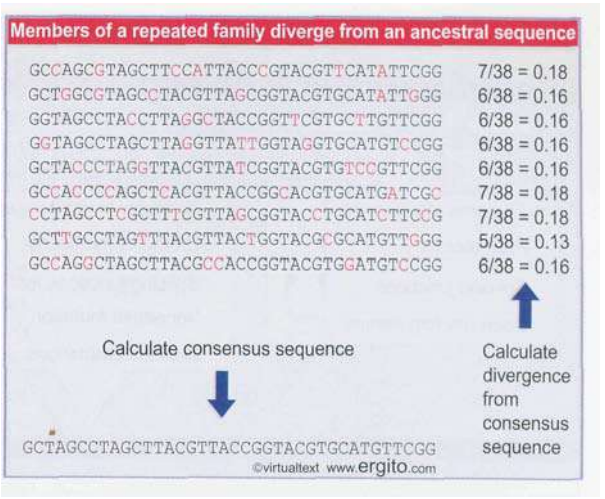


Figure 4.8 An ancestral consensus sequence for a family is calculated by taking the most common base at each position. The divergence of each existing current member of the family is calculated as the proportion of bases at which it differs from the ancestral sequence.

One family used for this analysis in the human and mouse genomes derives from a sequence that is thought to have ceased to be active at about the time of the divergence between Man and rodents (the LINES family; see *17.9 Retroposons fall into three classes).* This means that it has been diverging without any selective pressure for the same length of time in both species. Its average divergence in Man is ~0.17 substitutions per site, corresponding to a rate of 2.2 X 10^{-9} substitutions per base per year over the 75 million years since the separation. In the mouse genome, however, neutral substitutions have occurred at twice this rate, corresponding to 0.34 substitutions per site in the family, or a rate of 4.5 X 10^{-9}. However, note that if we calculated the rate per generation instead of per year, it would be greater in man than in mouse (~2.2 X 10^{-8} as opposed to ~10^{-9}).

These figures probably underestimate the rate of substitution in the mouse, because at the time of divergence the rates in both species would have been the same, and the difference must have evolved since then. The current rate of neutral substitution per year in the mouse is probably 2-3 X greater than the historical average. These rates reflect the balance between the occurrence of mutations and the ability of the genetic system of the organism to correct them. The difference between the species demonstrates that each species has systems that operate with a characteristic efficiency.

Comparing the mouse and human genomes allows us to assess whether syntenic (corresponding) sequences show signs of conservation or have differed at the rate expected from accumulation of neutral substitutions. The proportion of sites that show signs of selection is ~5%. This is much higher than the proportion that codes for protein or RNA (~1%). It implies that the genome includes many more stretches whose sequence is important for non-coding functions than for coding functions. Known regulatory elements are likely to comprise only a small part of this proportion. This number also suggests that most (i.e., the rest) of the genome sequences do not have any function that depends on the exact sequence.

4.6 Pseudogenes are dead ends of evolution

Key Concepts

* Pseudogenes have no coding function, but they can be recognized by sequence similarities with existing functional genes. They arise by the accumulation of mutations in (formerly) functional genes.

Pseudogenes (ψ) are defined by their possession of sequences that are related to those of the functional genes, but that cannot be translated into a functional protein.

Some pseudogenes have the same general structure as functional genes, with sequences corresponding to exons and introns in the usual locations. They may have been rendered inactive by mutations that

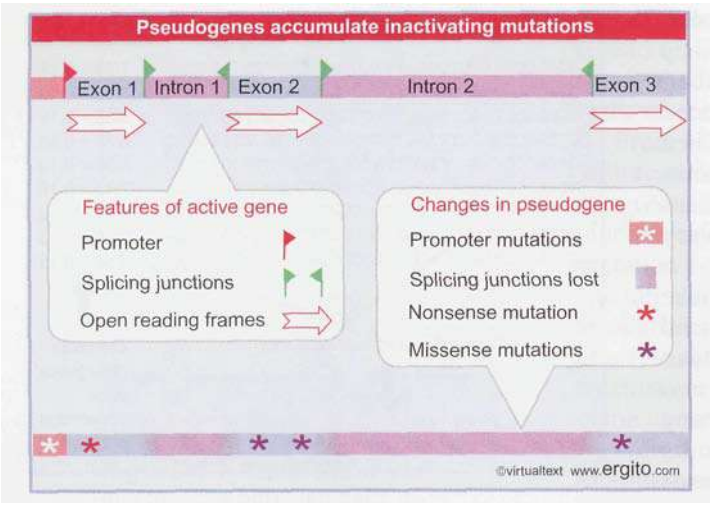


Figure 4.9 Many changes have occurred in a *ft* globin gene since it became a pseudogene.

prevent any or all of the stages of gene expression. The changes can take the form of abolishing the signals for initiating transcription, preventing splicing at the exon-intron junctions, or prematurely terminating translation.

Usually a pseudogene has several deleterious mutations. Presumably once it ceased to be active, there was no impediment to the accumulation of further mutations. Pseudogenes that represent inactive versions of currently active genes have been found in many systems, including globin, immunoglobulins, and histocompatibility antigens, where they are located in the vicinity of the gene cluster, often interspersed with the active genes.

A typical example is the rabbit pseudogene, *ty$2,* which has the usual organization of exons and introns, and is related most closely to the functional globin gene β1. But it is not functional. **Figure 4.9** summarizes the many changes that have occurred in the pseudogene. The deletion of a base pair at codon 20 of ψβ2 has caused a frameshift that would lead to termination shortly after. Several point mutations have changed later codons representing amino acids that are highly conserved in the β globins. Neither of the two introns any longer possesses recognizable boundaries with the exons, so probably the introns could not be spliced out even if the gene were transcribed. However, there are no transcripts corresponding to the gene, possibly because there have been changes in the 5' flanking region.

Since this list of defects includes mutations potentially preventing each stage of gene expression, we have no means of telling which event originally inactivated this gene. However, from the divergence between the pseudogene and the functional gene, we can estimate when the pseudogene originated and when its mutations started to accumulate.

If the pseudogene had become inactive as soon as it was generated by duplication from β1, we should expect both replacement site and silent site divergence rates to be the same. (They will be different only if the gene is translated to create selective pressure on the replacement sites.) But actually there are fewer replacement site substitutions than silent site substitutions. This suggests that at first (while the gene was expressed) there was selection against replacement site substitution. From the relative extents of substitution in the two types of site, we can calculate that ψβ2 diverged from β1 ~55 million years ago, remained a functional gene for 22 million years, but has been a pseudogene for the last 33 million years.

Similar calculations can be made for other pseudogenes. Some appear to have been active for some time before becoming pseudogenes, but others appear to have been inactive from the very time of their original generation. The general point made by the structures of these pseudogenes is that each has evolved independently during the development of the globin gene cluster in each species. This reinforces the conclusion that the creation of new genes, followed by their acceptance as functional duplicates, variation to become new functional genes, or inactivation as pseudogenes, is a continuing process in the gene cluster. Most gene families have members that are pseudogenes. Usually the pseudogenes represent a small minority of the total gene number.

The mouse ψα3 globin gene has an interesting property: it precisely lacks both introns. Its sequence can be aligned (allowing for accumulated mutations) with the α-globin mRNA. The apparent time of inactivation coincides with the original duplication, which suggests that the original inactivating event was associated with the loss of introns.

Inactive genomic sequences that resemble the RNA transcript are called processed pseudogenes. They originate by insertion at some ran-

dom site of a product derived from the RNA, following a retrotransposition event, as discussed in *17 Retroviruses and retroposons*. Their characteristic features are summarized in Figure 17.19.

If pseudogenes are evolutionary dead ends, simply an unwanted accompaniment to the rearrangement of functional genes, why are they still present in the genome? Do they fulfill any function or are they entirely without purpose, in which case there should be no selective pressure for their retention?

We should remember that we see those genes that have survived in present populations. In past times, any number of other pseudogenes may have been eliminated. This elimination could occur by deletion of the sequence as a sudden event or by the accretion of mutations to the point where the pseudogene can no longer be recognized as a member of its original sequence family (probably the ultimate fate of any pseudogene that is not suddenly eliminated).

Even relics of evolution can be duplicated. In the β-globin genes of the goat, there are three adult species, β^A, β^B, and β^C (see Figure 4.4). Each of these has a pseudogene a few kb upstream of it. The pseudogenes are better related to each other than to the adult β-globin genes; in particular, they share several inactivating mutations. Also, the adult β-globin genes are better related to each other than to the pseudogenes. This implies that an original ψβ-β structure was itself duplicated, giving functional β genes (which diverged further) and two nonfunctional genes (which diverged into the current pseudogenes).

The mechanisms responsible for gene duplication, deletion, and rearrangement act on all sequences that are recognized as members of the cluster, whether or not they are functional. It is left to selection to discriminate among the products.

4.7 Unequal crossing-over rearranges gene clusters

Key Concepts

- When a genome contains a cluster of genes with related sequences, mispairing between nonallelic genes can cause unequal crossing-over. This produces a deletion in one recombinant chromosome and a corresponding duplication in the other.
- Different thalassemias are caused by various deletions that eliminate a- or β-globin genes. The severity of the disease depends on the individual deletion.

There are frequent opportunities for rearrangement in a cluster of related or identical genes. We can see the results by comparing the mammalian β clusters included in Figure 4.4. Although the clusters serve the same function, and all have the same general organization, each is different in size, there is variation in the total number and types of β-globin genes, and the numbers and structures of pseudogenes are different. All of these changes must have occurred since the mammalian radiation, ~85 million years ago (the last point in evolution common to all the mammals).

The comparison makes the general point that gene duplication, rearrangement, and variation is as important a factor in evolution as the slow accumulation of point mutations in individual genes. What types of mechanisms are responsible for gene reorganization?

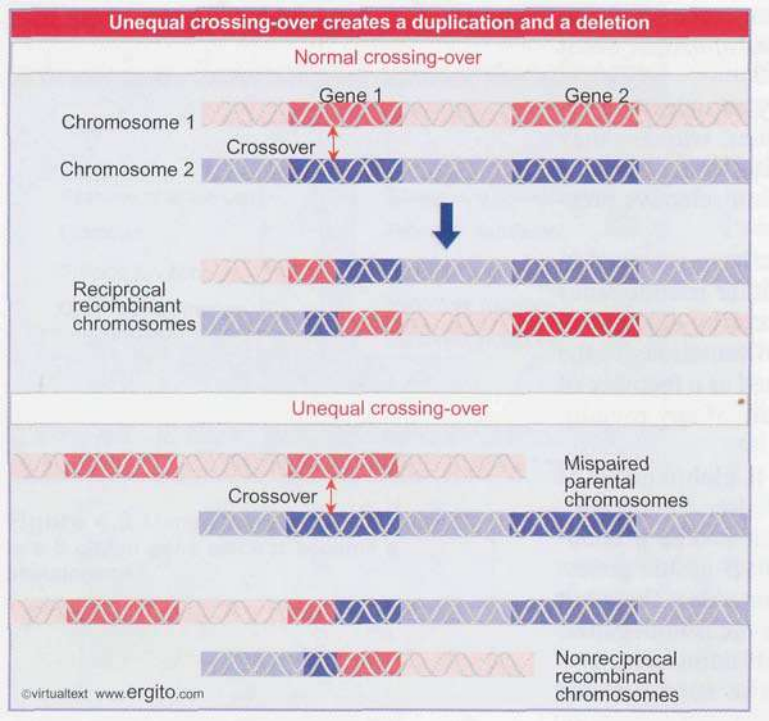


Figure 4.10 Gene number can be changed by unequal crossing-over. If gene 1 of one chromosome pairs with gene 2 of the other chromosome, the other gene copies are excluded from pairing. Recombination between the mispaired genes produces one chromosome with a single (recombinant) copy of the gene and one chromosome with three copies of the gene (one from each parent and one recombinant).

Unequal crossing-over (also known as nonreciprocal recombination) can occur as the result of pairing between two sites that are *not* homologous. Usually, recombination involves corresponding sequences of DNA held in exact alignment between the two homologous chromosomes. However, when there are two copies of a gene on each chromosome, an occasional misalignment allows pairing between them. (This requires some of the adjacent regions to go unpaired.) This can happen in a region of short repeats (see Figure 4.1) or in a gene cluster. **Figure 4.10** shows that unequal crossing-over in a gene cluster can have two consequences, quantitative and qualitative:

- The number of repeats increases in one chromosome and decreases in the other. In effect, one recombinant chromosome has a deletion and the other has an insertion. This happens irrespective of the exact location of the crossover. In the figure, the first recombinant has an increase in the number of gene copies from 2 to 3, while the second has a decrease from 2 to 1.
- If the recombination event occurs within a gene (as opposed to between genes), the result depends on whether the recombining genes are identical or only related. If the noncorresponding gene copies 1 and 2 are entirely homologous, there is no change in the sequence of either gene. However, unequal crossing-over also can occur when the adjacent genes are well related (although the probability is less than when they are identical). In this case, each of the recombinant genes has a sequence that is different from either parent.

Whether the chromosome has a selective advantage or disadvantage will depend on the consequence of any change in the sequence of the gene product as well as on the change in the number of gene copies.

An obstacle to unequal crossing-over is presented by the interrupted structure of the genes. In a case such as the globins, the corresponding exons of adjacent gene copies are likely to be well enough related to support pairing; but the sequences of the introns have diverged appreciably. The restriction of pairing to the exons considerably reduces the continuous length of DNA that can be involved. This lowers the chance of unequal crossing-over. So divergence between introns could enhance the stability of gene clusters by hindering the occurrence of unequal crossing-over.

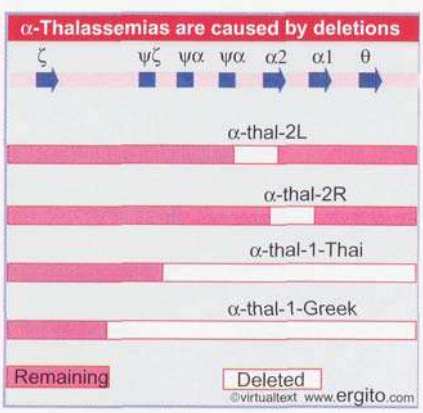


Figure 4.11 Thalassemias result from various deletions in the α-globin gene cluster.

Thalassemias result from mutations that reduce or prevent synthesis of either a or β globin. The occurrence of unequal crossing-over in the human globin gene clusters is revealed by the nature of certain thalassemias.

Many of the most severe thalassemias result from deletions of part of a cluster. In at least some cases, the ends of the deletion lie in regions that are homologous, which is exactly what would be expected if it had been generated by unequal crossing-over.

Figure 4.11 summarizes the deletions that cause the α-thalassemias. α-thal-1 deletions are long, varying in the location of the left end, with the positions of the right ends located beyond the known genes. They j eliminate both the a genes. The α-thal-2 deletions are short and eliminate only one of the two a genes. The L deletion removes 4.2 kb of DNA, including the a2 gene. It probably results from unequal crossing-over, because the ends of the deletion lie in homologous regions, just to the right of the ψα and a2 genes, respectively. The R deletion results from the removal of exactly 3.7 kb of DNA, the precise distance

between the *a*\ and α2 genes. It appears to have been generated by unequal crossing-over between the α1 and α2 genes themselves. This is precisely the situation depicted in Figure 4.10.

Depending on the diploid combination of thalassemic chromosomes, an affected individual may have any number of a chains from zero to three. There are few differences from the wild type (four a genes) in individuals with three or two a genes. But with only one a gene, the excess β chains form the unusual tetramer β_4, which causes HbH disease. The complete absence of a genes results in hydrops fetalis, which is fatal at or before birth.

The same unequal crossing-over that generated the thalassemic chromosome should also have generated a chromosome with three a genes. Individuals with such chromosomes have been identified in several populations. In some populations, the frequency of the triple a locus is about the same as that of the single α locus; in others, the triple α genes are much *less* common than single a genes. This suggests that (unknown) selective factors operate in different populations to adjust the gene levels.

Variations in the number of α genes are found relatively frequently, which argues that unequal crossing-over in the cluster must be fairly common. It occurs more often in the a cluster than in the β cluster, possibly because the introns in a genes are much shorter, and therefore present less impediment to mispairing between nonhomologous genes.

The deletions that cause β-thalassemias are summarized in **Figure 4.12.** In some (rare) cases, only the *ft* gene is affected. These have a deletion of 600 bp, extending from the second intron through the 3' flanking regions. In the other cases, more than one gene of the cluster is affected. Many of the deletions are very long, extending from the 5' end indicated on the map for >50 kb toward the right.

The Hb Lepore type provided the classic evidence that deletion can result from unequal crossing-over between linked genes. The β and 8 genes differ only ~7% in sequence. Unequal recombination deletes the material between the genes, thus fusing them together (see Figure 4.10). The fused gene produces a single β-like chain that consists of the N-terminal sequence of 8 joined to the C-terminal sequence of *ft*.

Several types of Hb Lepore now are known, the difference between them lying in the point of transition from 8 to β sequences. So when the 8 and β genes pair for unequal crossing-over, the exact point of recombination determines the position at which the switch from 8 to *ft* sequence occurs in the amino acid chain.

The reciprocal of this event has been found in the form of Hb anti-Lepore, which is produced by a gene that has the N-terminal part of β and the C-terminal part of 8. The fusion gene lies between normal 8 and β genes.

Evidence that unequal crossing-over can occur between more distantly related genes is provided by the identification of Hb Kenya, another fused hemoglobin. This contains the N-terminal sequence of the $^{A}\gamma$ gene and the C-terminal sequence of the β gene. The fusion must have resulted from unequal crossing-over between $^{A}\gamma$ and *ft,* which differ ~20% in sequence.

From the differences between the globin gene clusters of various mammals, we see that duplication followed (sometimes) by variation has been an important feature in the evolution of each cluster. The human thalassemic deletions demonstrate that unequal crossing-over continues to occur in both globin gene clusters. Each such event generates a duplication as well as the deletion, and we must account for the fate of both recombinant loci in the population. Deletions can also occur (in principle) by recombination between homologous sequences lying on the *same* chromosome. This does not generate a corresponding duplication.

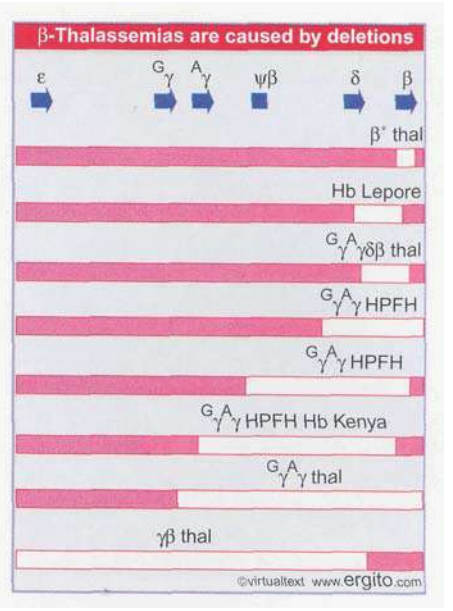


Figure 4.12 Deletions in the β-globin gene cluster cause several types of thalassemia.

It is difficult to estimate the natural frequency of these events, because selective forces rapidly adjust the levels of the variant clusters in the population. Generally a contraction in gene number is likely to be deleterious and selected against. However, in some populations, there may be a balancing advantage that maintains the deleted form at a low frequency.

The structures of the present human clusters show several duplications that attest to the importance of such mechanisms. The *functional* sequences include two a genes coding the same protein, fairly well related β and 8 genes, and two almost identical γ genes. These comparatively recent independent duplications have survived in the population, not to mention the more distant duplications that originally generated the various types of globin genes. Other duplications may have given rise to pseudogenes or have been lost. We expect continual duplication and deletion to be a feature of all gene clusters.

4.8 Genes for rRNA form tandem repeats

Key Concepts

- Ribosomal RNA is coded by a large number of identical genes that are tandemly repeated to form one or more clusters.
- Each rDNA cluster is organized so that transcription units giving a joint precursor to the major rRNAs alternate with nontranscribed spacers.

In the cases we have discussed so far, there are differences between the individual members of a gene cluster that allow selective pressure to act independently upon each gene. A contrast is provided by two cases of large gene clusters that contain many identical copies of the same gene or genes. Most organisms contain multiple copies of the genes for the histone proteins that are a major component of the chromosomes; and there are almost always multiple copies of the genes that code for the ribosomal RNAs. These situations pose some interesting evolutionary questions.

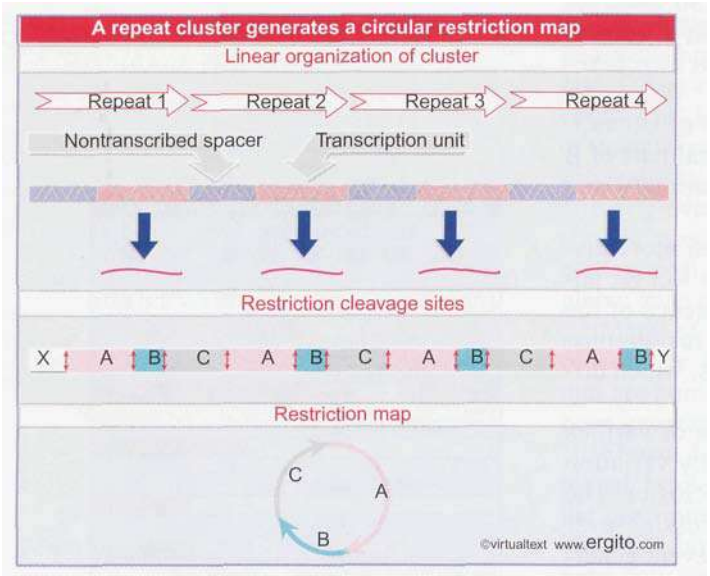


Figure 4.13 A tandem gene cluster has an alternation of transcription unit and nontranscribed spacer and generates a circular restriction map.

Ribosomal RNA is the predominant product of transcription, constituting some 80-90% of the total mass of cellular RNA in both eukaryotes and prokaryotes. The number of major rRNA genes varies from 7 in *E. coli* 100-200 in lower eukaryotes, to several hundred in higher eukaryotes. The genes for the large and small rRNA (found respectively in the large and small subunits of the ribosome) usually form a tandem pair. (The sole exception is the yeast mitochondrion.)

The lack of any detectable variation in the sequences of the rRNA molecules implies that all the copies of each gene must be identical, or at least must have differences below the level of detection in rRNA (~1%). A point of major interest is what mechanism(s) are used to prevent variations from accruing in the individual sequences.

In bacteria, the multiple rRNA gene pairs are dispersed. In most eukaryotic nuclei, the rRNA genes are contained in a tandem cluster or clusters. Sometimes these regions are called rDNA. (In some cases, the proportion of rDNA in the total DNA, together with its atypical base composition, is great enough to allow its isolation as a separate fraction directly from sheared genomic DNA.) An important diagnostic feature of a tandem cluster is that it generates a circular restriction map, as shown in Figure 4.13.

Suppose that each repeat unit has 3 restriction sites. In the example shown in the figure, fragments A and B are contained entirely within a repeat unit, and fragment C contains the end of one repeat and the beginning of the next. When we map these fragments by conventional means, we find that A is next to B, which is next to C, which is next to A, generating the circular map. If the cluster is large, the internal fragments (A, B, C) will be present in much greater quantities than the terminal fragments (X, Y) which connect the cluster to adjacent DNA. In a cluster of 100 repeats, X and Y would be present at 1% of the level of A, B, C. This can make it difficult to obtain the ends of a gene cluster for mapping purposes.

The region of the nucleus where rRNA synthesis occurs has a characteristic appearance, with a core of fibrillar nature surrounded by a granular cortex. The fibrillar core is where the rRNA is transcribed from the DNA template; and the granular cortex is formed by the ribonucleoprotein particles into which the rRNA is assembled. The whole area is called the nucleolus. Its characteristic morphology is evident in **Figure 4.14.**

The particular chromosomal regions associated with a nucleolus are called nucleolar organizers. Each nucleolar organizer corresponds to a cluster of tandemly repeated rRNA genes on one chromosome. The concentration of the tandemly repeated rRNA genes, together with their very intensive transcription, is responsible for creating the characteristic morphology of the nucleoli.

The pair of major rRNAs is transcribed as a single precursor in both bacteria and eukaryotic nuclei. Following transcription, the precursor is cleaved to release the individual rRNA molecules. The transcription unit is shortest in bacteria and is longest in mammals (where it is known as 45S RNA, according to its rate of sedimentation). An rDNA cluster contains many transcription units, each separated from the next by a nontranscribed spacer. The alternation of transcription unit and nontranscribed spacer can be seen directly in electron micrographs. The example shown in **Figure 4.15** is taken from the newt *N. viridescens,* in which each transcription unit is intensively expressed, so that many RNA polymerases are simultaneously engaged in transcription on one repeating unit. The polymerases are so closely packed that the RNA transcripts form a characteristic matrix displaying increasing length along the transcription unit.

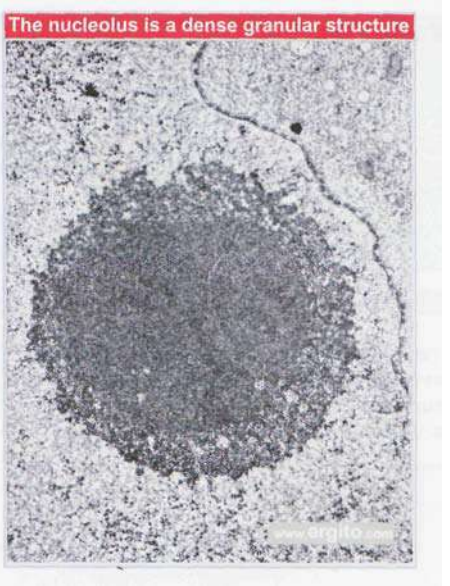


Figure 4.14 The nucleolar core identifies rDNA under transcription, and the surrounding granular cortex consists of assembling ribosomal subunits. This thin section shows the nucleolus of the newt *Notopthalmus viridescens.* Photograph kindly provided by Oscar Miller.

4.9 The repeated genes for rRNA maintain constant sequence

Key Concepts

- The genes in an rDNA cluster all have an identical sequence.
- The nontranscribed spacers consist of shorter repeating units whose number varies so that the lengths of individual spacers are different.

The nontranscribed spacer varies widely in length between and (sometimes) within species. In yeast there is a short nontranscribed spacer, relatively constant in length. In *D. melanogaster,* there is almost a twofold variation in the length of the nontranscribed spacer between different copies of the repeating unit. A similar situation is seen in *X. laevis.* In each of these cases, all of the repeating units are present as a single tandem cluster on one particular chromosome. (In the example of *D. melanogaster,* this happens to be the sex chromosome. The cluster on the X chromosome is larger than that on the Y chromosome, so female flies have more copies of the rRNA genes than male flies.)

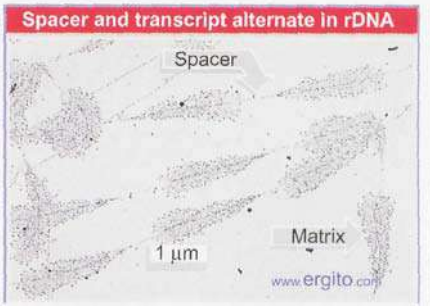


Figure 4.15 Transcription of rDNA clusters generates a series of matrices, each corresponding to one transcription unit and separated from the next by the nontranscribed spacer. Photograph kindly provided by Oscar Miller.

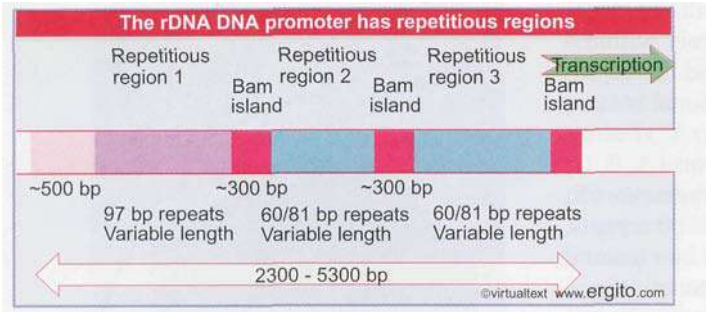


Figure 4.16 The nontranscribed spacer of *X. laevis* rDNA has an internally repetitious structure that is responsible for its variation in length.

In mammals the repeating unit is very much larger, comprising the transcription unit of ~13 kb and a nontranscribed spacer of ~30 kb. Usually, the genes lie in several dispersed clusters—in the case of man and mouse residing on five and six chromosomes, respectively. One interesting (but unanswered) question is how the corrective mechanisms that presumably function within a single cluster to ensure constancy of rRNA sequence are able to work when there are several clusters.

The variation in length of the nontranscribed spacer in a single gene cluster contrasts with the conservation of sequence of the transcription unit. In spite of this variation, the sequences of longer nontranscribed spacers remain homologous with those of the shorter nontranscribed spacers. This implies that each nontranscribed spacer is *internally repetitious*, so that the variation in length results from changes in the number of repeats of some subunit.

The general nature of the nontranscribed spacer is illustrated by the example of *X. laevis*. **Figure 4.16** illustrates the situation. Regions that are fixed in length alternate with regions that vary. Each of the three repetitious regions comprises a variable number of repeats of a rather short sequence. One type of repetitious region has repeats of a 97 bp sequence; the other, which occurs in two locations, has a repeating unit found in two forms, 60 bp and 81 bp long. The variation in the number of repeating units in the repetitious regions accounts for the overall variation in spacer length. The repetitious regions are separated by shorter constant sequences called Bam islands. (This description takes its name from their isolation via the use of the BamHI restriction enzyme.) From this type of organization, we see that the cluster has evolved by duplications involving the promoter region.

We need to explain the lack of variation in the expressed copies of the repeated genes. One model would suppose that there is a quantitative demand for a certain number of "good" sequences. But this would enable mutated sequences to accumulate up to a point at which their proportion of the cluster is great enough for selective pressure to be exerted. We can exclude such models because of the lack of such variation in the cluster.

The lack of variation implies the existence of selective pressure in some form that is sensitive to individual variations. One model would suppose that the entire cluster is regenerated periodically from one or from a very few members. As a practical matter any mechanism would need to involve regeneration every generation. We can exclude such models because a regenerated cluster would not show variation in the nontranscribed regions of the individual repeats.

We are left with a dilemma. Variation in the nontranscribed regions suggests that there is frequent unequal crossing over. This will change the size of the cluster, but will not otherwise change the properties of the individual repeats. So how are mutations prevented from accumulating? We see in the next section that continuous contraction and expansion of a cluster may provide a mechanism for homogenizing its copies.

4.10 Crossover fixation could maintain identical repeats

Key Concepts

- Unequal crossing-over changes the size of a cluster of tandem repeats.
- Individual repeating units can be eliminated or can spread through the cluster.

The same problem is encountered whenever a gene has been duplicated. How can selection be imposed to prevent the accumulation of deleterious mutations?

The duplication of a gene is likely to result in an immediate relaxation of the evolutionary pressure on its sequence. Now that there are two identical copies, a change in the sequence of either one will not deprive the organism of a functional protein, since the original amino acid sequence continues to be coded by the other copy. Then the selective pressure on the two genes is diffused, until one of them mutates sufficiently away from its original function to refocus all the selective pressure on the other.

Immediately following a gene duplication, changes might accumulate more rapidly in one of the copies, leading eventually to a new function (or to its disuse in the form of a pseudogene). If a new function" develops, the gene then evolves at the same, slower rate characteristic of the original function. Probably this is the sort of mechanism responsible for the separation of functions between embryonic and adult globin genes.

Yet there are instances where duplicated genes retain the same function, coding for the identical or nearly identical proteins. Identical proteins are coded by the two human α-globin genes, and there is only a single amino acid difference between the two γ-globin proteins. How is selective pressure exerted to maintain their sequence identity?

The most obvious possibility is that the two genes do not actually have identical functions, but differ in some (undetected) property, such as time or place of expression. Another possibility is that the need for two copies is quantitative, because neither by itself produces a sufficient amount of protein.

In more extreme cases of repetition, however, it is impossible to avoid the conclusion that no single copy of the gene is essential. When there are many copies of a gene, the immediate effects of mutation in any one copy must be very slight. The consequences of an individual mutation are diluted by the large number of copies of the gene that retain the wild-type sequence. Many mutant copies could accumulate before a lethal effect is generated.

Lethality becomes quantitative, a conclusion reinforced by the observation that half of the units of the rDNA cluster of *X. laevis* or *D. melanogaster* can be deleted without ill effect. So how are these units prevented from gradually accumulating deleterious mutations? And what chance is there for the rare favorable mutation to display its advantages in the cluster?

The basic principle of models to explain the maintenance of identity among repeated copies is to suppose that nonallelic genes are not independently inherited, but must be continually regenerated from *one* of the copies of a preceding generation. In the simplest case of two identical genes, when a mutation occurs in one copy, either it is by chance eliminated (because the sequence of the other copy takes over), or it is spread to both duplicates (because the mutant copy becomes the dominant version). Spreading exposes a mutation to selection. The result is that the two genes evolve together as though only a single locus existed. This is called coincidental evolution or concerted evolution (occasionally coevolution). It can be applied to a pair of identical genes or (with further assumptions) to a cluster containing many genes.

One mechanism supposes that the sequences of the nonallelic genes are directly compared with one another and homogenized by enzymes that recognize any differences. This can be done by exchanging single strands between them, to form genes one of whose strands derives from one copy, one from the other copy. Any differences show as improperly paired bases, which attract attention from enzymes able to excise and

replace a base, so that only A·T and G·C pairs survive. This type of event is called gene conversion and is associated with genetic recombination as described in *15 Recombination and repair*.

We should be able to ascertain the scope of such events by comparing the sequences of duplicate genes. If they are subject to concerted evolution, we should not see the accumulation of silent site substitutions between them (because the homogenization process applies to these as well as to the replacement sites). We know that the extent of the maintenance mechanism need not extend beyond the gene itself, since there are cases of duplicate genes whose flanking sequences are entirely different. Indeed, we may see abrupt boundaries that mark the ends of the sequences that were homogenized.

We must remember that the existence of such mechanisms can invalidate the determination of the history of such genes via their divergence, *because the divergence reflects only the time since the last homogenization/regeneration event, not the original duplication.*

The crossover fixation model supposes that an entire cluster is subject to continual rearrangement by the mechanism of unequal crossing-over. Such events can explain the concerted evolution of multiple genes if unequal crossing-over causes all the copies to be regenerated physically from one copy.

Following the sort of event depicted in Figure 4.10, for example, the chromosome carrying a triple locus could suffer deletion of one of the genes. Of the two remaining genes, 1 ½ represent the sequence of one of the original copies; only ½ of the sequence of the other original copy has survived. Any mutation in the first region now exists in both genes and is subject to selective pressure.

Tandem clustering provides frequent opportunities for "mispairing" of genes whose sequences are the same, but that lie in different positions in their clusters. By continually expanding and contracting the number of units via unequal crossing-over, it is possible for all the units in one cluster to be derived from rather a small proportion of those in an ancestral cluster. The variable lengths of the spacers are consistent with the idea that unequal crossing-over events take place in spacers that are internally mispaired. This can explain the homogeneity of the genes compared with the variability of the spacers. The genes are exposed to selection when individual repeating units are amplifiec within the cluster; but the spacers are irrelevant and can accumulate changes.

In a region of nonrepetitive DNA, recombination occurs between precisely matching points on the two homologous chromosomes, generating reciprocal recombinants. The basis for this precision is the ability of two duplex DNA sequences to align exactly. We know that unequal recombination can occur when there are multiple copies of genes whose exons are related, even though their flanking and intervening sequences may differ. This happens because of the mispairing between corresponding exons in *nonallelic* genes.

Imagine how much more frequently misalignment must occur in a tandem cluster of identical or nearly identical repeats. Except at the very ends of the cluster, the close relationship between successive repeats makes it impossible even to define the exactly corresponding repeats! This has two consequences: there is continual adjustment of the size of the cluster; and there is homogenization of the repeating unit.

Consider a sequence consisting of a repeating unit "ab" with ends "x" and "y." If we represent one chromosome in black and the other in color, the exact alignment between "allelic" sequences would be

xabababababababababababababababy
xabababababababababababababababy

But probably *any* sequence *ab* in one chromosome could pair with *any*sequence *abin* the other chromosome. In a misalignment such as

xabababababababababababababababy
xababababababababababababababababy

the region of pairing is no less stable than in the perfectly aligned pair, although it is shorter. We do not know very much about how pairing is initiated prior to recombination, but very likely it starts between short corresponding regions and then spreads. If it starts within satellite DNA, it is more likely than not to involve repeating units that do not have exactly corresponding locations in their clusters.

Now suppose that a recombination event occurs within the unevenly paired region. The recombinants will have different numbers of repeating units. In one case, the cluster has become longer; in the other, it has become shorter,

xabababababababababababababababy
x
xababababababababababababababababy
↓
xababababababababababababababababy
+
xabababababababababababababy

where "X" indicates the site of the crossover.

If this type of event is common, clusters of tandem repeats will undergo continual expansion and contraction. This can cause a particular repeating unit to spread through the cluster, as illustrated in **Figure 4.17.** Suppose that the cluster consists initially of a sequence *abcde,* where each letter represents a repeating unit. The different repeating units are closely enough related to one another to mispair for recombination. Then by a series of unequal recombination events, the size of the repetitive region increases or decreases, and also one unit spreads to replace all the others.

The crossover fixation model predicts that *any sequence of DNA that is not under selective pressure will be taken over by a series of identical tandem repeats generated in this way.* The critical assumption is that the process of crossover fixation is fairly rapid relative to mutation, so that new mutations either are eliminated (their repeats are lost) or come to take over the entire cluster. In the case of the rDNA cluster, of course, a further factor is imposed by selection for an effective transcribed sequence.

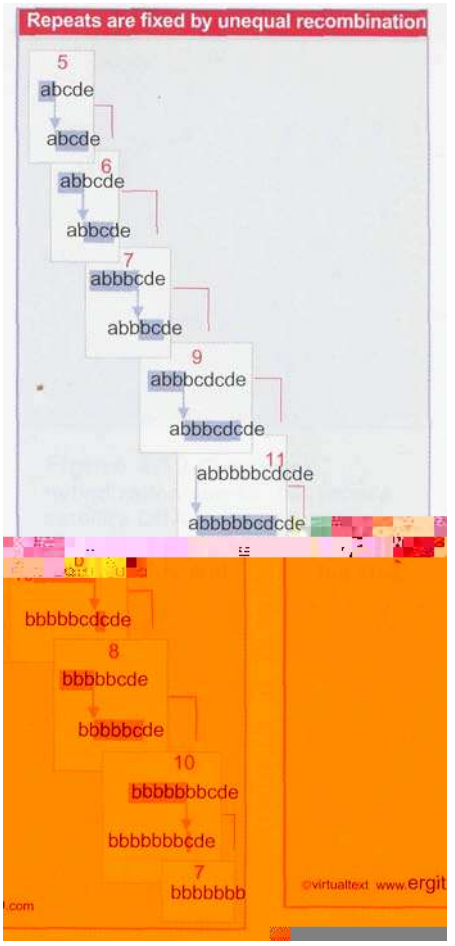


Figure 4.17 Unequal recombination allows one particular repeating unit to occupy the entire cluster. The numbers indicate the length of the repeating unit at each stage.

4.11 Satellite DNAs often lie in heterochromatin

Key Concepts

- Highly repetitive DNA has a very short repeating sequence and no coding function.
- It occurs in large blocks that can have distinct physical properties.
- It is often the major constituent of centromeric heterochromatin.

Repetitive DNA is defined by its (relatively) rapid rate of renaturation. The component that renatures most rapidly in a eukaryotic genome is called *highly repetitive* DNA, and consists of very short sequences repeated many times in tandem in large clusters. Because of its short repeating unit, it is sometimes described as simple sequence

DNA. This type of component is present in almost all higher eukaryotic genomes, but its overall amount is extremely variable. In mammalian genomes it is typically <10%, but in (for example) *Drosophila virilis*, it amounts to ~50%. In addition to the large clusters in which this type of sequence was originally discovered, there are smaller clusters interspersed with nonrepetitive DNA. It typically consists of short sequences that are repeated in identical or related copies in the genome.

The tandem repetition of a short sequence often creates a fraction with distinctive physical properties that can be used to isolate it. In some cases, the repetitive sequence has a base composition distinct from the genome average, which allows it to form a separate fraction by virtue of its distinct buoyant density. A fraction of this sort is called satellite DNA. The term satellite DNA is essentially synonymous with simple sequence DNA. Consistent with its simple sequence, this DNA is not transcribed or translated.

Tandemly repeated sequences are especially liable to undergo misalignments during chromosome pairing, and thus the sizes of tandem clusters tend to be highly polymorphic, with wide variations between individuals. In fact, the smaller clusters of such sequences can be used to characterize individual genomes in the technique of "DNA fingerprinting" (see *4.14 Minisatellites are useful for genetic mapping*).

The buoyant density of a duplex DNA depends on its G-C content according to the empirical formula

$$\rho = 1.660 + 0.00098\ (\%G{\cdot}C)\ \text{g-cm}^{-3}$$

Buoyant density usually is determined by centrifuging DNA through a density gradient of CsCl. The DNA forms a band at the position corresponding to its own density. Fractions of DNA differing in G-C content by >5% can usually be separated on a density gradient.

When eukaryotic DNA is centrifuged on a density gradient, two types of material may be distinguished:

- Most of the genome forms a continuum of fragments that appear as a rather broad peak centered on the buoyant density corresponding to the average G·C content of the genome. This is called the main band.
- Sometimes an additional, smaller peak (or peaks) is seen at a different value. This material is the satellite DNA.

Satellites are present in many eukaryotic genomes. They may be either heavier or lighter than the main band; but it is uncommon for them to represent >5% of the total DNA. A clear example is provided by mouse DNA, shown in **Figure 4.18.** The graph is a quantitative scan of the bands formed when mouse DNA is centrifuged through a CsCl density gradient. The main band contains 92% of the genome and is centered on a buoyant density of 1.701 g-cm^{-3} (corresponding to its average G-C of 42%, typical for a mammal). The smaller peak represents 8% of the genome and has a distinct buoyant density of 1.690 g-cnf^{3}. It contains the mouse satellite DNA, whose G-C content (30%) is much lower than any other part of the genome.

The behavior of satellite DNA on density gradients is often anomalous. When the actual base composition of a satellite is determined, it is different from the prediction based on its buoyant density. The reason is that ρ is a function not just of base composition, but of the constitution in terms of nearest neighbor pairs. For simple sequences, these are likely to deviate from the random pairwise relationships needed to obey the equation for buoyant density. Also, satellite DNA may be methylated, which changes its density.

Often most of the highly repetitive DNA of a genome can be isolated in the form of satellites. When a highly repetitive DNA component does not separate as a satellite, on isolation its properties often prove to be similar to those of satellite DNA. That is to say that it consists of multiple tandem repeats with anomalous centrifugation. Material isolated in this manner is sometimes referred to as a cryptic satellite. Together the

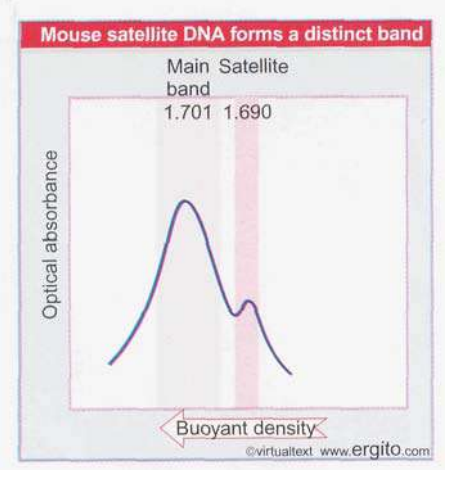


Figure 4.18 Mouse DNA is separated into a main band and a satellite by centrifugation through a density gradient of CsCl.

cryptic and apparent satellites usually account for all the large tandemly repeated blocks of highly repetitive DNA. When a genome has more than one type of highly repetitive DNA, each exists in its own satellite block (although sometimes different blocks are adjacent).

Where in the genome are the blocks of highly repetitive DNA located? An extension of nucleic acid hybridization techniques allows the location of satellite sequences to be determined directly in the chromosome complement. In the technique of *in situ* hybridization, the chromosomal DNA is denatured by treating cells that have been squashed on a cover slip. Then a solution containing a radioactively labeled DNA or RNA probe is added. The probe hybridizes with its complements in the denatured genome. The location of the sites of hybridization can be determined by autoradiography (see Figure 19.19).

Satellite DNAs are found in regions of heterochromatin. Heterochromatin is the term used to describe regions of chromosomes that are permanently tightly coiled up and inert, in contrast with the euchromatin that represents most of the genome (see *19.7 Chromatin is divided into euchromatin and heterochromatin).* Heterochromatin is commonly found at centromeres (the regions where the kinetochores are formed at mitosis and meiosis for controlling chromosome movement). The centromeric location of satellite DNA suggests that it has some structural function in the chromosome. This function could be connected with the process of chromosome segregation.

An example of the localization of satellite DNA for the mouse chromosomal complement is shown in **Figure 4.19.** In this case, one end of each chromosome is labeled, because this is where the centromeres are located in *M. musculus* chromosomes.

Figure 4.19 Cytological hybridization shows that mouse satellite DNA is located at the centromeres. Photograph kindly provided by Mary Lou Pardue and Joe Gall.

4.12 Arthropod satellites have very short identical repeats

Key Concepts

- The repeating units of arthropod satellite DNAs are only a few nucleotides long. Most of the copies of the sequence are identical.

In the arthropods, as typified by insects and crabs, each satellite DNA appears to be rather homogeneous. Usually, a single very short repeating unit accounts for >90% of the satellite. This makes it relatively straightforward to determine the sequence.

Drosophila virilis has three major satellites and also a cryptic satellite, together representing >40% of the genome. The sequences of the satellites are summarized in **Figure 4.20.** The three major satellites have closely related sequences. A single base substitution is sufficient to generate either satellite II or III from the sequence of satellite I.

The satellite I sequence is present in other species of *Drosophila* related to *virilis*, and so may have preceded speciation. The sequences of satellites II and III seem to be specific to *D. virilis*, and so may have evolved from satellite I after speciation.

The main feature of these satellites is their very short repeating unit: only 7 bp. Similar satellites are found in other species. *D. melanogaster* has a variety of satellites, several of which have very short repeating units (5, 7, 10, or 12 bp). Comparable satellites are found in the crabs.

The close sequence relationship found among the *D. virilis* satellites is not necessarily a feature of other genomes, where the satellites may have unrelated sequences. *Each satellite has arisen by a lateral amplification of*

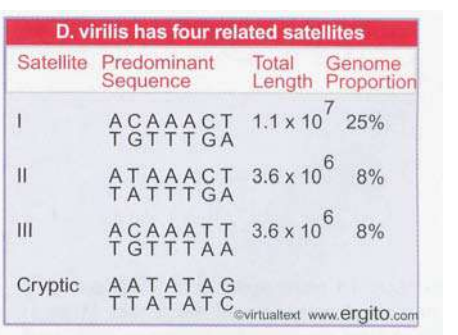


Figure 4.20 Satellite DNAs of *D. virilis* are related. More than 95% of each satellite consists of a tandem repetition of the predominant sequence.

a very short sequence. This sequence may represent a variant of a previously existing satellite (as in *D. virilis*), or could have some other origin.

Satellites are continually generated and lost from genomes. This makes it difficult to ascertain evolutionary relationships, since a current satellite could have evolved from some previous satellite that has since been lost. The important feature of these satellites is that *they represent very long stretches of DNA of very low sequence complexity, within which constancy of sequence can be maintained.*

One feature of many of these satellites is a pronounced asymmetry in the orientation of base pairs on the two strands. In the example of the *D. virilis* satellites shown in Figure 4.19, in each of the major satellites one of the strands is much richer in T and G bases. This increases its buoyant density, so that upon denaturation this heavy strand (H) can be separated from the complementary light strand (L). This can be useful in sequencing the satellite.

4.13 Mammalian satellites consist of hierarchical repeats

Key Concepts

- Mouse satellite DNA has evolved by duplication and mutation of a short repeating unit to give a basic repeating unit of 234 bp in which the original half, quarter, and eighth repeats can be recognized.

In the mammals, as typified by various rodents, the sequences comprising each satellite show appreciable divergence between tandem repeats. Common short sequences can be recognized by their preponderance among the oligonucleotide fragments released by chemical or enzymatic treatment. However, the predominant short sequence usually accounts for only a small minority of the copies. The other short sequences are related to the predominant sequence by a variety of substitutions, deletions, and insertions.

But a series of these variants of the short unit can constitute a longer repeating unit that is itself repeated in tandem with some variation. So mammalian satellite DNAs are constructed from a hierarchy of repeating units. These longer repeating units constitute the sequences that renature in reassociation analysis. They can also be recognized by digestion with restriction enzymes.

When any satellite DNA is digested with an enzyme that has a recognition site in its repeating unit, one fragment will be obtained for every repeating unit in which the site occurs. In fact, when the DNA of a eukaryotic genome is digested with a restriction enzyme, most of it gives a general smear, due to the random distribution of cleavage sites. But satellite DNA generates sharp bands, because a large number of fragments of identical or almost identical size are created by cleavage at restriction sites that lie a regular distance apart.

Determining the sequence of satellite DNA can be difficult. Using the discrete bands generated by restriction cleavage, we can attempt to obtain a sequence directly. However, if there is appreciable divergence between individual repeating units, different nucleotides will be present at the same position in different repeats, so the sequencing gels will be obscure. If the divergence is not too great—say, within ~2%—it may be possible to determine an average repeating sequence.

Individual segments of the satellite can be inserted into plasmids for cloning. A difficulty is that the satellite sequences tend to be excised

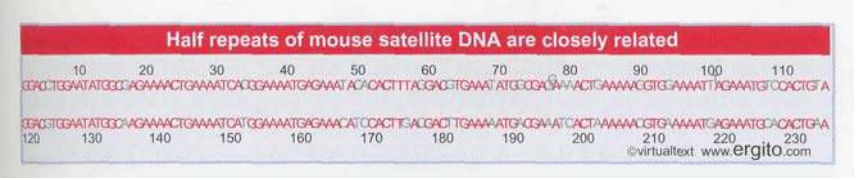


Figure 4.21 The repeating unit of mouse satellite DNA contains two half-repeats, which are aligned to show the identities (in red).

from the chimeric plasmid by recombination in the bacterial host. However, when the cloning succeeds, it is possible to determine the sequence of the cloned segment unambiguously. While this gives the actual sequence of a repeating unit or units, we should need to have many individual such sequences to reconstruct the type of divergence typical of the satellite as a whole.

By either sequencing approach, the information we can gain is limited to the distance that can be analyzed on one set of sequence gels. The repetition of divergent tandem copies makes it impossible to reconstruct longer sequences by obtaining overlaps between individual restriction fragments. The satellite DNA of the mouse *M. musculus* is cleaved by the enzyme EcoRII into a series of bands, including a predominant monomeric fragment of 234 bp. This sequence must be repeated with few variations throughout the 60-70% of the satellite that is cleaved into the monomeric band. We may analyze this sequence in terms of its successively smaller constituent repeating units.

Figure **4.21** depicts the sequence in terms of two half-repeats. By writing the 234 bp sequence so that the first 117 bp are aligned with the second 117 bp, we see that the two halves are quite well related. They differ at 22 positions, corresponding to 19% divergence. This means that the current 234 bp repeating unit must have been generated at some time in the past by duplicating a 117 bp repeating unit, after which differences accumulated between the duplicates.

Within the 117 bp unit, we can recognize two further subunits. Each of these is a quarter-repeat relative to the whole satellite. The four quarter-repeats are aligned in **Figure 4.22.** The upper two lines represent the first half-repeat of Figure 4.21; the lower two lines represent the second half-repeat. We see that the divergence between the four quarter-repeats has increased to 23 out of 58 positions, or 40%. The first three quarter-repeats are somewhat better related, and a large proportion of the divergence is due to changes in the fourth quarter-repeat.

Looking within the quarter-repeats, we find that each consists of two related subunits (one-eighth-repeats), shown as the a and β sequences in **Figure 4.23.** The a sequences all have an insertion of a C, and the *ft* sequences all have an insertion of a trinucleotide, relative to a common consensus sequence. This suggests that the quarter-repeat originated by the duplication of a sequence like the consensus sequence, after which changes occurred to generate the components we now see as a and β. Further changes then took place between tandemly repeated *aft* sequences to generate the individual quarter- and half-repeats that exist today. Among the one-eighth-repeats, the present divergence is 19/31 = 61%.

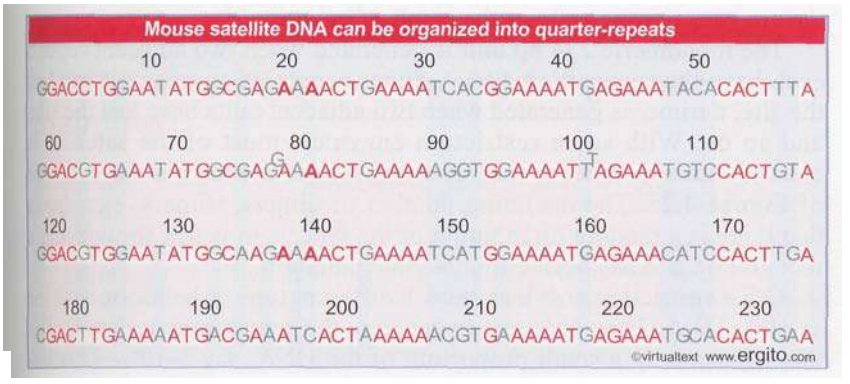


Figure 4.22 The alignment of quarter-repeats identifies homologies between the first and second half of each half-repeat. Positions that are the same in all 4 quarter-repeats are shown in color; identities that extend only through 3 quarter-repeats are indicated by grey letters in the pink area.

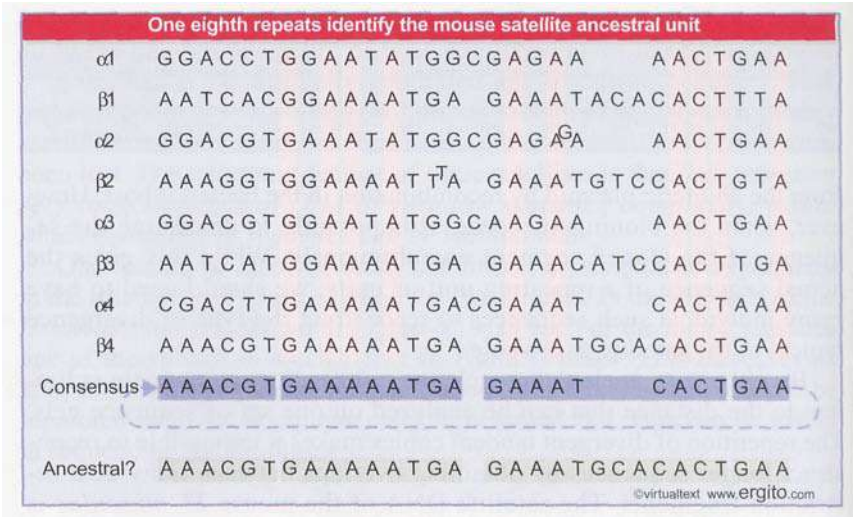


Figure 4.23 The alignment of eighth-repeats shows that each quarter-repeat consists of an a and a β half. The consensus sequence gives the most common base at each position. The "ancestral" sequence shows a sequence very closely related to the consensus sequence, which could have been the predecessor to the a and β units. (The satellite sequence is continuous, so that for the purposes of deducing the consensus sequence, we can treat it as a circular permutation, as indicated by joining the last GAA triplet to the first 6 bp.)

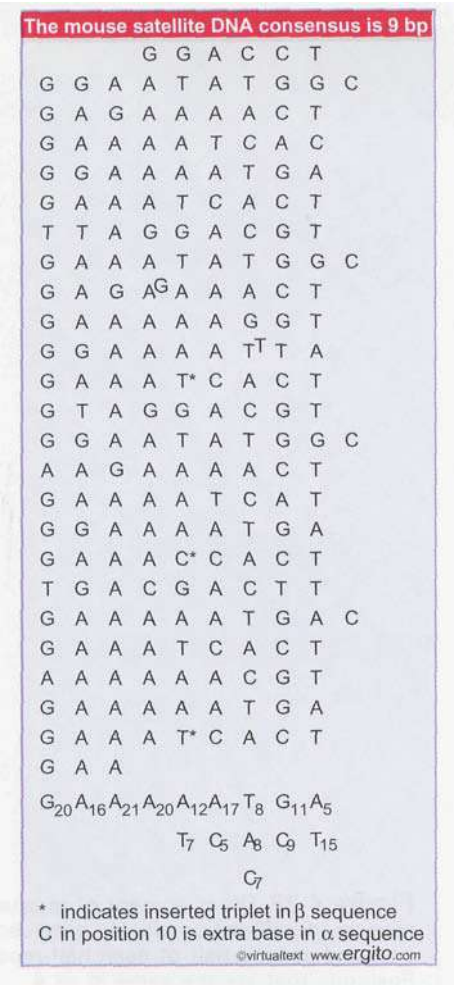


Figure 4.24 The existence of an overall consensus sequence is shown by writing the satellite sequence in terms of a 9 bp repeat.

The consensus sequence is analyzed directly in **Figure 4.24,** which demonstrates that the current satellite sequence can be treated as derivatives of a 9 bp sequence. We can recognize three variants of this sequence in the satellite, as indicated at the bottom of Figure 4.24. If in one of the repeats we take the next most frequent base at two positions instead of the most frequent, we obtain three well-related 9 bp sequences.

G A A A A A C G T
G A A A A A T G A
G A A A A A A C T

The origin of the satellite could well lie in an amplification of one of these three nonamers. The overall consensus sequence of the present satellite is GAAAAA(A/T)(T/C)T, which is effectively an amalgam of the three 9 bp repeats.

The average sequence of the monomeric fragment of the mouse satellite DNA explains its properties. The longest repeating unit of 234 bp is identified by the restriction cleavage. The unit of reassociation between single strands of denatured satellite DNA is probably the 117 bp half-repeat, because the 234 bp fragments can anneal both in register and in half-register (in the latter case, the first half-repeat of one strand renatures with the second half-repeat of the other).

So far, we have treated the present satellite as though it consisted of identical copies of the 234 bp repeating unit. Although this unit accounts for the majority of the satellite, variants of it also are present. Some of them are scattered at random throughout the satellite; others are clustered.

The existence of variants is implied by our description of the starting material for the sequence analysis as the "monomeric" fragment. When the satellite is digested by an enzyme that has one cleavage site in the 234 bp sequence, it also generates dimers, trimers, and tetramers relative to the 234 bp length. They arise when a repeating unit has lost the enzyme cleavage site as the result of mutation.

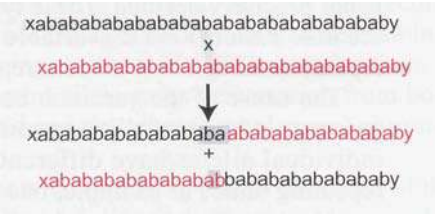
The monomeric 234 bp unit is generated when two adjacent repeats each have the recognition site. A dimer occurs when one unit has lost the site, a trimer is generated when two adjacent units have lost the site, and so on. With some restriction enzymes, most of the satellite is cleaved into a member of this repeating series, as shown in the example of **Figure 4.25.** The declining number of dimers, trimers, etc. shows that there is a random distribution of the repeats in which the enzyme's recognition site has been eliminated by mutation.

Other restriction enzymes show a different type of behavior with the satellite DNA. They continue to generate the same series of bands. But they cleave only a small proportion of the DNA, say 5-10%. This im-

plies that a certain region of the satellite contains a concentration of the repeating units with this particular restriction site. Presumably the series of repeats in this domain all are derived from an ancestral variant that possessed this recognition site (although in the usual way, some members since have lost it by mutation).

A satellite DNA suffers unequal recombination. This has additional consequences when there is internal repetition in the repeating unit. Let us return to our cluster consisting of "ab" repeats. Suppose that the "a" and "b" components of the repeating unit are themselves sufficiently well related to pair. Then the two clusters can align in *half-register*, with the "a" sequence of one aligned with the "b" sequence of the other. How frequently this occurs will depend on the closeness of the relationship between the two halves of the repeating unit. In mouse satellite DNA, reassociation between the denatured satellite DNA strands *in vitro* commonly occurs in the half-register.

When a recombination event occurs out of register, it changes the length of the repeating units that are involved in the reaction:



In the upper recombinant cluster, an "ab" unit has been replaced by an "aab" unit. In the lower cluster, the "ab" unit has been replaced by a "b" unit.

This type of event explains a feature of the restriction digest of mouse satellite DNA. Figure 4.24 shows a fainter series of bands at lengths of 'A, 1½, and 3½ repeating units, in addition to the stronger integral length repeats. Suppose that in the preceding example, "ab" represents the 234 bp repeat of mouse satellite DNA, generated by cleavage at a site in the "b" segment. The "a" and "b" segments correspond to the 117 bp half-repeats.

Then in the upper recombinant cluster, the "aab" unit generates a fragment of 1 'A times the usual repeating length. And in the lower recombinant cluster, the "b" unit generates a fragment of half of the usual length. (The multiple fragments in the half-repeat series are generated in the same way as longer fragments in the integral series, when some repeating units have lost the restriction site by mutation.)

Turning the argument the other way around, the identification of the half-repeat series on the gel shows that the 234 bp repeating unit consists of two half-repeats well enough related to pair sometimes for recombination. Also visible in Figure 4.25 are some rather faint bands corresponding to ¼- and ¾-spacings. These will be generated in the same way as the ½-spacings, when recombination occurs between clusters aligned in a quarter-register. The decreased relationship between quarter-repeats compared with half-repeats explains the reduction in frequency of the ¼- and ¾-bands compared with the ½-bands.

Mouse satellite DNA has repeats and half-repeats

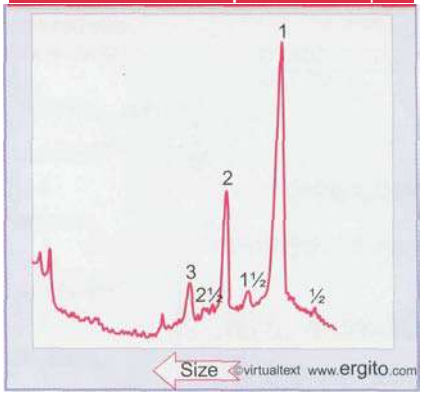


Figure 4.25 Digestion of mouse satellite DNA with the restriction enzyme EcoRII identifies a series of repeating units (1, 2, 3) that are multimers of 234 bp and also a minor series (½, 1½, 2½) that includes half-repeats (see text later). The band at the far left is **a** fraction resistant to digestion.

4.14 Minisatellites are useful for genetic mapping

Key Concepts

- The variation between microsatellites or minisatellites in individual genomes can be used to identify heredity unequivocally by showing that 50% of the bands in an individual are derived from a particular parent.

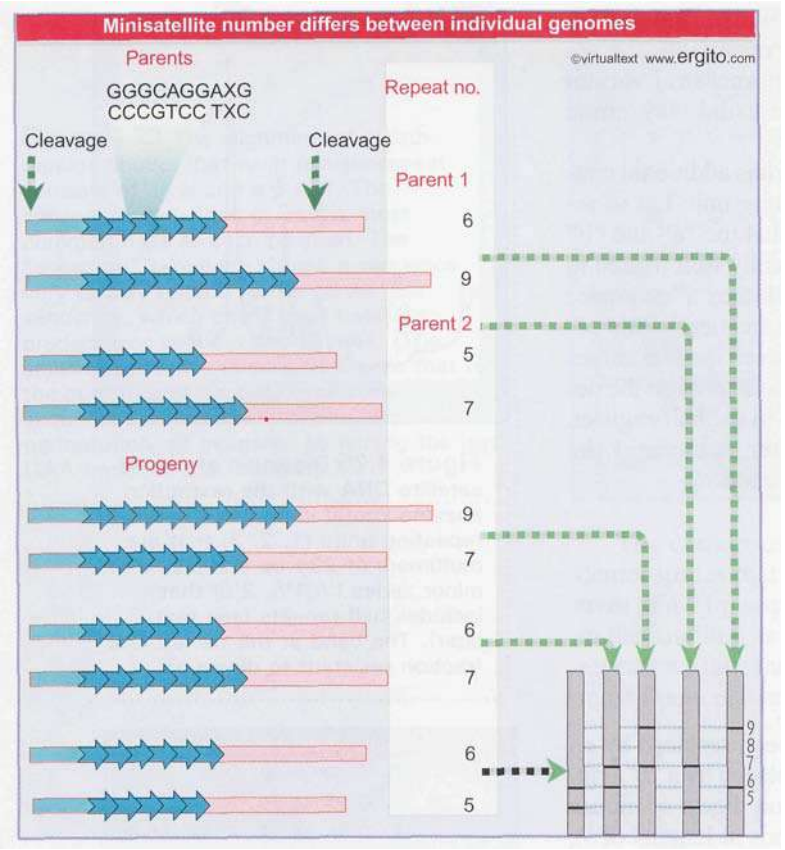


Figure 4.26 Alleles may differ in the number of repeats at a minisatellite locus, so that cleavage on either side generates restriction fragments that differ in length. By using a minisatellite with alleles that differ between parents, the pattern of inheritance can be followed.

Sequences that resemble satellites in consisting of tandem repeats of a short unit, but that overall are much shorter, consisting of (for example) from 5-50 repeats, are common in mammalian genomes. They were discovered by chance as fragments whose size is extremely variable in genomic libraries of human DNA. The variability is seen when a population contains fragments of many different sizes that represent the same genomic region; when individuals are examined, it turns out that there is extensive polymorphism, and that many different alleles can be found.

The name microsatellite is usually used when the length of the repeating unit is < 10 bp, and the name minisatellite is used when the length of the repeating unit is ~10-100 bp, but the terminology is not precisely defined. These types of sequences are also called VNTR (variable number tandem repeat) or STR (short tandem repeat).

The cause of the variation between individual genomes at microsatellites or minisatellites is that individual alleles have different numbers of the repeating unit. For example, one minisatellite has a repeat length of 64 bp, and is found in the population with the following distribution:

7% 18 repeats
11% 16 repeats
43% 14 repeats
36% 13 repeats
4% 10 repeats

The rate of genetic exchange at minisatellite sequences is high, $\sim 10^{-4}$ per kb of DNA. (The frequency of exchanges per actual locus is assumed to be proportional to the length of the minisatellite.) This rate is ~ 10 X greater than the rate of homologous recombination at meiosis, that is, in any random DNA sequence.

The high variability of minisatellites makes them especially useful for genomic mapping, because there is a high probability that individuals will vary in their alleles at such a locus. An example of mapping by minisatellites is illustrated in **Figure 4.26.** This shows an extreme case in which two individuals both are heterozygous at a minisatellite locus, and in fact all four alleles are different. All progeny gain one allele from each parent in the usual way, and it is possible unambiguously to determine the source of every allele in the progeny. In the terminology of human genetics, the meiosis described in this figure are highly informative, because of the variation between alleles.

One family of minisatellites in the human genome share a common "core" sequence. The core is a G·C-rich sequence of 10-15 bp, showing an asymmetry of purine/pyrimidine distribution on the two strands. Each individual minisatellite has a variant of the core sequence, but ~1000 minisatellites can be detected on Southern blot by a probe consisting of the core sequence.

Consider the situation shown in Figure 4.26, but multiplied 1000×. The effect of the variation at individual loci is to create a unique pattern for every individual. This makes it possible to assign heredity unambiguously between parents and progeny, by showing that 50% of the bands in any individual are derived from a particular parent. This is the basis of the technique known as DNA fingerprinting.

By Book_Crazy [IND]

Both microsatellites and minisatellites are unstable, although for different reasons. Microsatellites undergo intrastrand mispairing, when slippage during replication leads to expansion of the repeat, as shown in **Figure 4.27.** Systems that repair damage to DNA, in particular those that recognize mismatched base pairs, are important in reversing such changes, as shown by a large increase in frequency when repair genes are inactivated. Because mutations in repair systems are an important contributory factor in the development of cancer, tumor cells often display variations in microsatellite sequences (see *30.29 Defects in repair systems cause mutations to accumulate in tumors).*

Minisatellites undergo the same sort of unequal crossing-over between repeats that we have discussed for satellites (see Figure 4.1). One telling case is that increased variation is associated with a meiotic hotspot. The recombination event is not usually associated with recombination between flanking markers, but has a complex form in which the new mutant allele gains information from both the sister chromatid and the other (homologous) chromosome.

It is not clear at what repeating length the cause of the variation shifts from replication slippage to recombination.

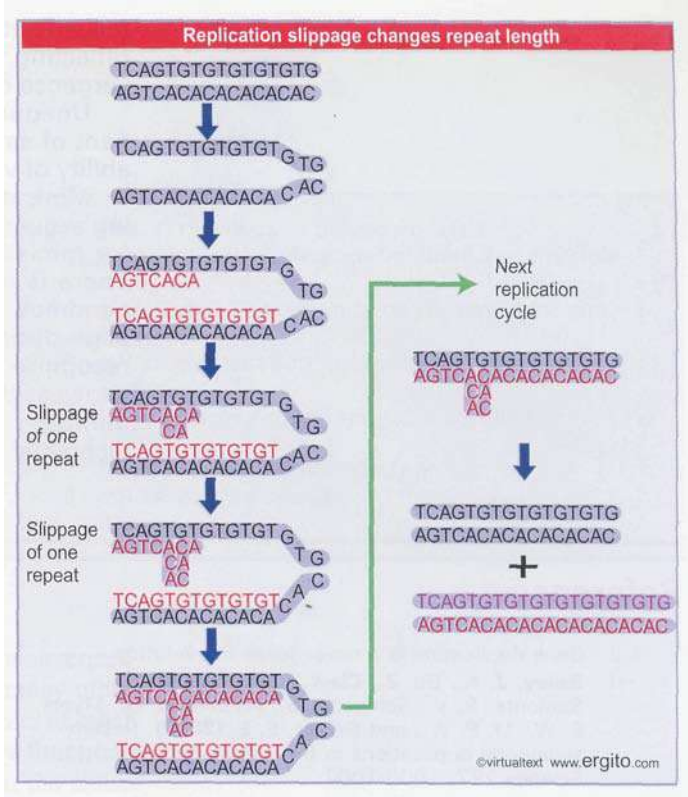


Figure 4.27 Replication slippage occurs when the daughter strand slips back one repeating unit in pairing with the template strand. Each slippage event adds one repeating unit to the daughter strand. The extra repeats are extruded as a single strand loop. Replication of this daughter strand in the next cycle generates a duplex DNA with an increased number of repeats.

4.15 Summary

Almost all genes belong to families, defined by the possession of related sequences in the exons of individual members. Families evolve by the duplication of a gene (or genes), followed by divergence between the copies. Some copies suffer inactivating mutations and become pseudogenes that no longer have any function. Pseudogenes also may be generated as DNA copies of the mRNA sequences.

An evolving set of genes may remain together in a cluster or may be dispersed to new locations by chromosomal rearrangement. The organization of existing clusters can sometimes be used to infer the series of events that has occurred. These events act with regard to sequence rather than function, and therefore include pseudogenes as well as active genes.

Mutations accumulate more rapidly in silent sites than in replacement sites (which affect the amino acid sequence). The rate of divergence at replacement sites can be used to establish a clock, calibrated in percent divergence per million years. The clock can then be used to calculate the time of divergence between any two members of the family.

A tandem cluster consists of many copies of a repeating unit that includes the transcribed sequence(s) and a nontranscribed spacer(s). rRNA gene clusters code only for a single rRNA precursor. Maintenance of active genes in clusters depends on mechanisms such as gene conversion or unequal crossing-over that cause mutations to spread through the cluster, so that they become exposed to evolutionary pressure.

Satellite DNA consists of very short sequences repeated many times in tandem. Its distinct centrifugation properties reflect its biased base composition. Satellite DNA is concentrated in centromeric heterochromatin, but its function (if any) is unknown. The individual repeating units of arthropod satellites are identical. Those of mam-

malian satellites are related, and can be organized into a hierarchy reflecting the evolution of the satellite by the amplification and divergence of randomly chosen sequences.

Unequal crossing-over appears to have been a major determinant of satellite DNA organization. Crossover fixation explains the ability of variants to spread through a cluster.

Minisatellites and microsatellites consist of even shorter repeating sequences than satellites, <10 bp for microsatellites and 10-50 bp for minisatellites. The number of repeating units is usually 5-50. There is high variation in the repeat number between individual genomes. Microsatellite repeat number varies as the result of slippage during replication; the frequency is affected by systems that recognize and repair damage in DNA. Minisatellite repeat number varies as the result of recombination-like events. Variations in repeat number can be used to determine hereditary relationships by the technique known as DNA fingerprinting.

References

4.2 Gene duplication is a major force in evolution

ref Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). Recent segmental duplications in the human genome. Science 297, 1003-1007.

4.3 Globin clusters are formed by duplication and divergence

Hardison, R. (1998). Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J. Exp. Biol. 201, 1099-1117.

4.5 The rate of neutral substitution can be measured from divergence of repeated sequences

ref Waterston et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

4.10 Crossover fixation could maintain identical repeats

rev Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215-220.

4.14 Minisatellites are useful for genetic mapping

ref Jeffreys, A. J., Tamaki, K., McCleod, A., Monckton, D. G., Neil, D. L and Armour, J. A. L. Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L., and Armour, J. A. (1994). Complex gene conversion events in germline mutation at human minisatellites. Nat. Genet. 6, 136-145.

Jeffreys, A. J., Jeffreys, A. J., Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332, 278-281.

Jeffreys, A. J., Murray, J., and Neumann, R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol. Cell 2, 267-273.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable minisatellite regions in human DNA. Nature 314, 67-73.

Strand, M., Prolla, T. A., Liskay, and Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365, 274-276.

Chapter 5

Messenger RNA

5.1 Introduction

RNA is a central player in gene expression. It was first characterized as an intermediate in protein synthesis, but since then many other RNAs have been discovered that play structural or functional roles at other stages of gene expression. The involvement of RNA in many functions concerned with gene expression supports the general view that the entire process may have evolved in an "RNA world" in which RNA was originally the active component in maintaining and expressing genetic information. Many of these functions were subsequently assisted or taken over by proteins, with a consequent increase in versatility and probably efficiency.

As summarized in **Figure 5.1,** three major classes of RNA are directly involved in the production of proteins:

- Messenger RNA (mRNA) provides an intermediate that carries the copy of a DNA sequence that represents protein.
- Transfer RNAs (tRNA) are small RNAs that are used to provide amino acids corresponding to each particular codon in mRNA.
- Ribosomal RNAs (rRNA) are components of the ribosome, a large ribonucleoprotein complex that contains many proteins as well as its RNA components, and which provides the apparatus for actually polymerizing amino acids into a polypeptide chain.

The type of role that RNA plays in each of these cases is distinct. For messenger RNA, its sequence is the important feature: each nucleotide triplet within the coding region of the mRNA represents an amino acid in the corresponding protein. However, the structure of the mRNA, in particular the sequences on either side of the coding region, can play an important role in controlling its activity, and therefore the amount of protein that is produced from it.

In tRNA, we see two of the common themes governing the use of RNA: its three dimensional structure is important; and it has the ability to base pair with another RNA (mRNA). The three dimensional structure is recognized first by an enzyme as providing a target that is appropriate for linkage to a specific amino acid. The linkage creates an aminoacyl-tRNA, which is recognized as the structure that is used for protein synthesis. The specificity with which an aminoacyl-tRNA is used is controlled by base pairing, when a short triplet sequence (the anticodon) pairs with the nucleotide triplet representing its amino acid.

With rRNA, we see another type of activity. One role of RNA is structural, in providing a framework to which ribosomal proteins attach. But it also participates directly in the activities of the ribosome. One of

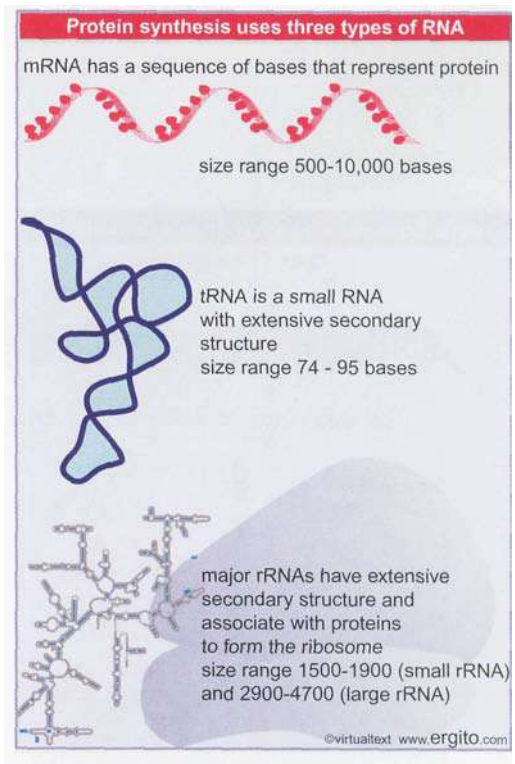


Figure 5.1 The three types of RNA universally required for gene expression are mRNA (carries the coding sequence), tRNA (provides the amino acid corresponding to each codon), and rRNA (a major component of the ribosome that provides the environment for protein synthesis).

the crucial activities of the ribosome is the ability to catalyze the formation of a peptide bond by which an amino acid is incorporated into protein. This activity resides in one of the rRNAs.

The important thing about this background is that, as we consider the role of RNA in protein synthesis, we have to view it as a component that plays an active role and that can be a target for regulation by either proteins or by other RNAs, and we should remember that the RNAs may have been the basis for the original apparatus. The theme that runs through all of the activities of RNA, in both protein synthesis and elsewhere, is that its functions depend critically upon base pairing, both to form its secondary structure, and to interact specifically with other RNA molecules. The coding function of mRNA is unique, but tRNA and rRNA are examples of a much broader class of noncoding RNAs with a variety of functions in gene expression.

5.2 mRNA is produced by transcription and is translated

Key Concepts

- Only one of the two strands of DNA is transcribed into RNA.

Gene expression occurs by a two-stage process.

- Transcription generates a single-stranded RNA identical in sequence with one of the strands of the duplex DNA.
- Translation converts the nucleotide sequence of mRNA into the sequence of amino acids comprising a protein. The entire length of an mRNA is not translated, but each mRNA contains at least one coding region that is related to a protein sequence by the genetic code: each nucleotide triplet (codon) of the coding region represents one amino acid.

Only one strand of a DNA duplex is transcribed into a messenger RNA. We distinguish the two strands of DNA as depicted in **Figure 5.2**:

- The strand of DNA that directs synthesis of the mRNA via complementary base pairing is called the template strand or antisense strand. (*Antisense* is used as a general term to describe a sequence of DNA or RNA that is complementary to mRNA.)
- The other DNA strand bears the *same* sequence as the mRNA (except for possessing T instead of U), and is called the coding strand or sense strand.

In this chapter we discuss mRNA and its use as a template for protein synthesis. In *6 Protein synthesis* we discuss the process by which a protein is synthesized. In *7 Using the genetic code* we discuss the way the genetic code is used to interpret the meaning of a sequence of mRNA. And in *8 Protein localization* we turn to the question of how a protein finds its proper location in the cell when or after it is synthesized.

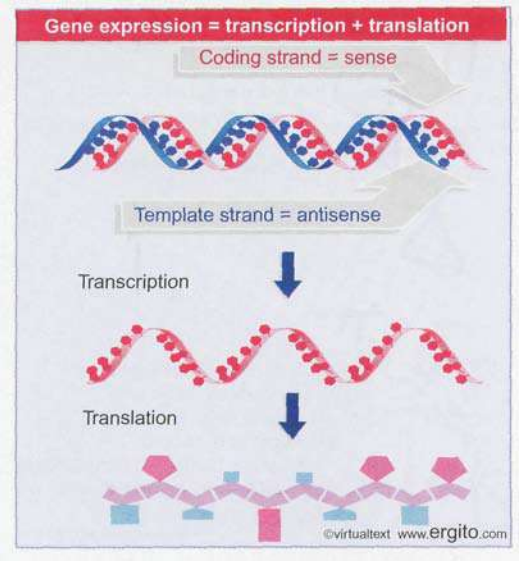


Figure 5.2 Transcription generates an RNA which is complementary to the DNA template strand and has the same sequence as the DNA coding strand. Translation reads each triplet of bases into one amino acid. Three turns of the DNA double helix contain 30 bp, which code for 10 amino acids.

Key Concepts

- A tRNA has a sequence of 74-95 bases that folds into a cloverleaf secondary structure with four constant arms (and an additional arm in the longer tRNAs).
- tRNA is charged to form aminoacyl-tRNA by forming an ester link from the 2' or 3' OH group of the adenylic acid at the end of the acceptor arm to the COOH group of the amino acid.

Messenger RNA can be distinguished from the apparatus responsible for its translation by the use of *in vitro* cell-free systems to synthesize proteins. *A protein-synthesizing system from one cell type can translate the mRNA from another, demonstrating that both the genetic code and the translation apparatus are universal.*

Each nucleotide triplet in the mRNA represents an amino acid. The incongruity of structure between trinucleotide and amino acid immediately raises the question of how each codon is matched to its particular amino acid. The "adapter" is transfer RNA (tRNA). A tRNA has two crucial properties:

- It represents a single amino acid to which it is *covalently linked.*
- It contains a trinucleotide sequence, the anticodon, which is *complementary to the codon representing its amino acid.* The anticodon enables the tRNA to recognize the codon via complementary base pairing.

All tRNAs have common secondary and tertiary structures. The tRNA secondary structure can be written in the form of a cloverleaf, illustrated in **Figure 5.3,** in which complementary base pairing forms stems for single-stranded loops. The stem-loop structures are called the arms of tRNA. Their sequences include "unusual" bases that are generated by modification of the 4 standard bases after synthesis of the polynucleotide chain.

The construction of the cloverleaf is illustrated in more detail in Figure 5.4. The four major arms are named for their structure or function:

- The acceptor arm consists of a base-paired stem that ends in an unpaired sequence whose free 2'- or 3′-OH group can be linked to an amino acid.
- The *TψC arm* is named for the presence of this triplet sequence. (ψ stands for pseudouridine, a modified base.)
- The anticodon arm always contains the anticodon triplet in the center of the loop.
- The D arm is named for its content of the base dihydrouridine (another of the modified bases in tRNA).
- The extra arm lies between the TψC and anticodon arms and varies from 3-21 bases.

The numbering system for tRNA illustrates the constancy of the structure. Positions are numbered from 5' to 3' according to the most common tRNA structure, which has 76 residues. The overall range of tRNA lengths is 74-95 bases. The variation in length is caused by differences in the D arm and extra arm.

The base pairing that maintains the secondary structure is shown in Figure 5.4. Within a given tRNA, most of the base pairings are conventional partnerships of A·U and G·C, but occasional G·U, G·ψ, or A·ψ pairs are found. The additional types of base pairs are less stable than the regular pairs, but still allow a double-helical structure to form in RNA.

When the sequences of tRNAs are compared, the bases found at some positions are invariant (or conserved); almost always a particular base is found at the position. Some positions are described as semi-invariant (or semiconserved) because they are restricted to one type of base (purine versus pyrimidine), but either base of that type may be present.

When a tRNA is *charged* with the amino acid corresponding to its anticodon, it is called aminoacyl-tRNA. The amino acid is linked by an ester bond from its carboxyl group to the 2' or 3' hydroxyl group of the ribose of the 3' terminal base of the tRNA (which is always adenine). The process of charging a tRNA is catalyzed by a specific enzyme,

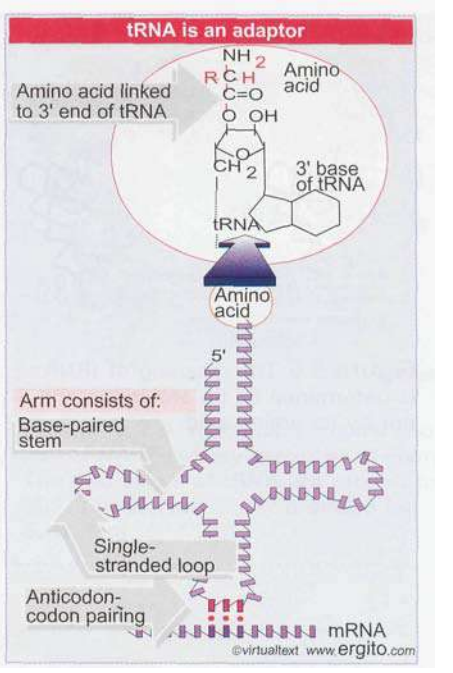


Figure 5.3 A tRNA has the dual properties of an adaptor that recognizes both the amino acid and codon. The 3' adenosine is covalently linked to an amino acid. The anticodon base pairs with the codon on mRNA.

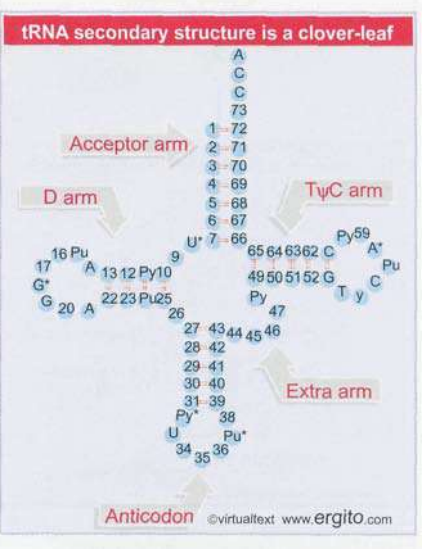


Figure 5.4 The tRNA cloverleaf has invariant and semi-invariant bases, and a conserved set of base pairing interactions.

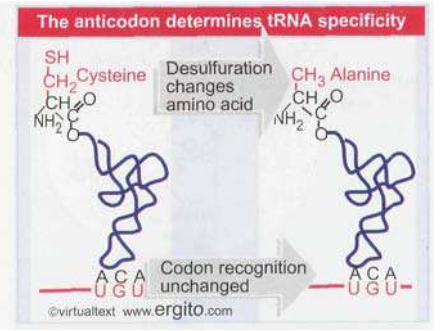


Figure 5.5 **The** meaning of tRNA is determined by its anticodon and not by its amino acid.

aminoacyl-tRNA synthetase. There are (at least) 20 aminoacyl-tRNA synthetases. Each recognizes a single amino acid and all the tRNAs on to which it can legitimately be placed.

There is at least one tRNA (but usually more) for each amino acid. A tRNA is named by using the three letter abbreviation for the amino acid as a superscript. If there is more than one tRNA for the same amino acid, subscript numerals are used to distinguish them. So two tRNAs for tyrosine would be described as $tRNA_1^{Tyr}$ and $tRNA_2^{Tyr}$. A tRNA carrying an amino acid—that is, an aminoacyl-tRNA—is indicated by a prefix that identifies the amino acid. Ala-tRNA describes $tRNA^{Ala}$ carrying its amino acid.

Does the anticodon sequence alone allow aminoacyl-tRNA to recognize the correct codon? A classic experiment to test this question is illustrated in **Figure 5.5.** Reductive desulfuration converts the amino acid of cysteinyl-tRNA into alanine, generating alanyl-$tRNA^{Cys}$. The tRNA has an anticodon that responds to the codon UGU. Modification of the amino acid does not influence the specificity of the anticodon-codon interaction, so the alanine residue is incorporated into protein in place of cysteine. *Once a tRNA has been charged, the amino acid plays no further role in its specificity, which is determined exclusively by the anticodon.*

5.4 The acceptor stem and anticodon are at ends of the tertiary structure

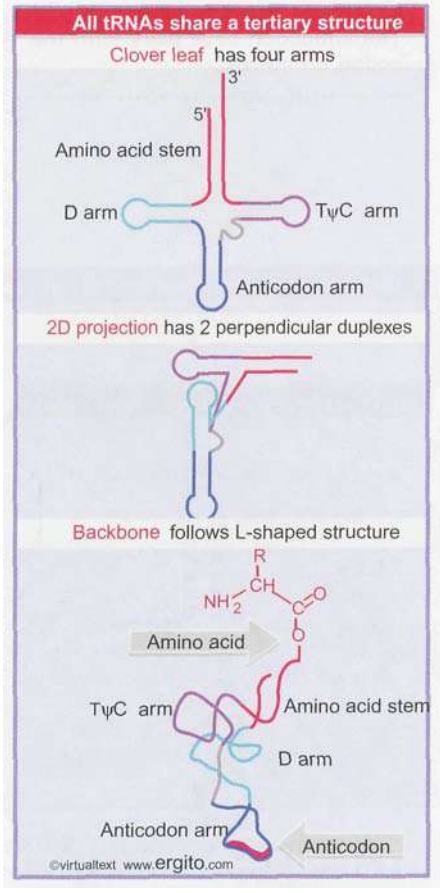


Figure 5.6 Transfer RNA folds into a compact L-shaped tertiary structure with the amino acid at one end and the anticodon at the other end.

Key Concepts

- The clover-leaf forms an L-shaped tertiary structure with the acceptor arm at one end and the anticodon arm at the other end.
- The sequence of the anticodon is solely responsible for the specificity of the aminoacyl-tRNA.

The secondary structure of each tRNA folds into a compact L-shaped tertiary structure in which the 3′ end that binds the amino acid is distant from the anticodon that binds the mRNA. All tRNAs have the same general tertiary structure, although they are distinguished by individual variations.

The base paired double-helical stems of the secondary structure are maintained in the tertiary structure, but their arrangement in three dimensions essentially creates two double helices at right angles to each other, as illustrated in **Figure 5.6.** The acceptor stem and the TψC stem form one continuous double helix with a single gap; the D stem and anticodon stem form another continuous double helix, also with a gap. The region between the double helices, where the turn in the L-shape is made, contains the TψC loop and the D loop. So the amino acid resides at the extremity of one arm of the L-shape, and the anticodon loop forms the other end.

The tertiary structure is created by hydrogen bonding, mostly involving bases that are unpaired in the secondary structure. Many of the invariant and semiinvariant bases are involved in these H-bonds, which explains their conservation. Not every one of these interactions is universal, but probably they identify the *general* pattern for establishing tRNA structure.

A molecular model of the structure of yeast $tRNA^{Phe}$ is shown in **Figure 5.7.** The left view corresponds with the bottom panel in Figure 5.6. Differences in the structure are found in other tRNAs, thus accommodating the dilemma that all tRNAs must have a similar shape, yet it

must be possible to recognize differences between them. For example, in tRNA ^{Phe}, the angle between the two axes is slightly greater, so the molecule has a slightly more open conformation.

The structure suggests a general conclusion about the function of tRNA. *Its sites for exercising particular functions are maximally separated.* The amino acid is as far distant from the anticodon as possible, *which is* consistent with their roles in protein synthesis.

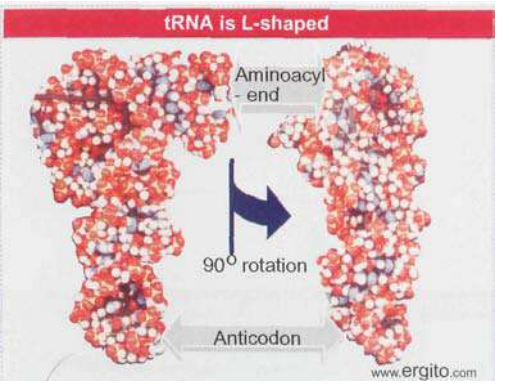


Figure 5.7 A space-filling model shows that tRNA^{Phe} tertiary structure is compact. The two views of tRNA are rotated by 90°. Photograph kindly provided by S. H. Kim.

5.5 Messenger RNA is translated by ribosomes

Key Concepts

- Ribosomes are characterized by their rate of sedimentation (70S for bacterial ribosomes and 80S for eukaryotic ribosomes).
- A ribosome consists of a large subunit (50S or 60S for bacteria and eukaryotes) and a small subunit (30S or 40S).
- The ribosome provides the environment in which aminoacyl-tRNAs add amino acids to the growing polypeptide chain in response to the corresponding triplet codons.
- A ribosome moves along an mRNA from 5' to 3'.

Translation of an mRNA into a polypeptide chain is catalyzed by the ribosome. Ribosomes are traditionally described in terms of their (approximate) rate of sedimentation (measured in Svedbergs, in which a higher S value indicates a greater rate of sedimentation and a larger mass). Bacterial ribosomes generally sediment at ~70S. The ribosomes of the cytoplasm of higher eukaryotic cells are larger, usually sedimenting at ~80S.

The ribosome is a compact ribonucleoprotein particle consisting of two subunits. Each subunit has an RNA component, including one very large RNA molecule, and many proteins. The relationship between a ribosome and its subunits is depicted in **Figure 5.8.** The two subunits dissociate *in vitro* when the concentration of Mg^{2+} ions is reduced. In each case, the large subunit is about twice the mass of the small subunit. Bacterial (70S) ribosomes have subunits that sediment at 50S and 30S. The subunits of eukaryotic cytoplasmic (80S) ribosomes sediment at 60S and 40S. The two subunits work together as part of the complete ribosome, but each undertakes distinct reactions in protein synthesis.

All the ribosomes of a given cell compartment are identical. *They undertake the synthesis of different proteins by associating with the different mRNAs that provide the actual coding sequences.*

The ribosome provides the environment that controls the recognition between a codon of mRNA and the anticodon of tRNA. Reading the genetic code as a series of adjacent triplets, protein synthesis proceeds from the start of a coding region to the end. *A protein is assembled by the sequential addition of amino acids in the direction from the N-terminus to the C-terminus as a ribosome moves along the mRNA.*

A ribosome begins translation at the 5' end of a coding region; it translates each triplet codon into an amino acid as it proceeds towards the 3' end. At each codon, the appropriate aminoacyl-tRNA associates with the ribosome, donating its amino acid to the polypeptide chain. At any given moment, the ribosome can accommodate the two aminoacyl-tRNAs corresponding to successive codons, making it possible for a peptide bond to form between the two corresponding amino acids. At each step, *the* growing polypeptide chain becomes longer by one amino acid.

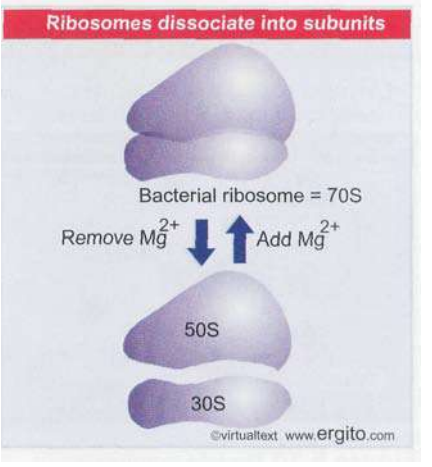


Figure 5.8 A ribosome consists of two subunits.

5.6 Many ribosomes bind to one mRNA

Key Concepts

- An mRNA is simultaneously translated by several ribosomes. Each ribosome is at a different stage of progression along the mRNA.

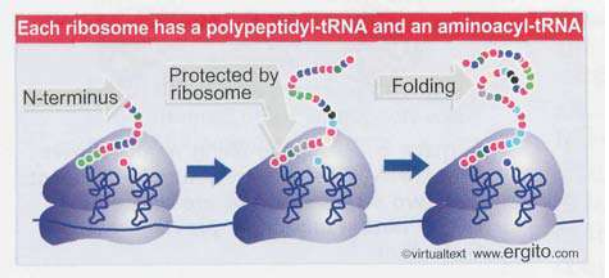


Figure 5.9 A polyribosome consists of an mRNA being translated simultaneously by several ribosomes moving in the direction from 5'-3'. Each ribosome has two tRNA molecules, one carrying the nascent protein, the second carrying the next amino acid to be added.

When active ribosomes are isolated in the form of the fraction associated with newly synthesized proteins, they are found in the form of a complex consisting of an mRNA associated with several ribosomes. This is the polyribosome or polysome. The 30S subunit of each ribosome is associated with the mRNA, and the 50S subunit carries the newly synthesized protein. The tRNA spans both subunits.

Each ribosome in the polysome independently synthesizes a single polypeptide during its traverse of the messenger sequence. Essentially the mRNA is pulled through the ribosome, and each triplet nucleotide is translated into an amino acid. So the mRNA has a series of ribosomes that carry increasing lengths of the protein product, moving from the 5' to the 3' end, as illustrated in **Figure 5.9.** A polypeptide chain in the process of synthesis is sometimes called a nascent protein.

Roughly the most recent 30-35 amino acids added to a growing polypeptide chain are protected from the environment by the structure of the ribosome. Probably all of the preceding part of the polypeptide protrudes and is free to start folding into its proper conformation. So proteins can display parts of the mature conformation even before synthesis has been completed.

A classic characterization of polysomes is shown in the electron micrograph of **Figure 5.10.** Globin protein is synthesized by a set of 5 ribosomes attached to each mRNA (pentasomes). The ribosomes appear as squashed spherical objects of ~7 nm (70 A) in diameter, connected by a thread of mRNA. The ribosomes are located at various positions along the messenger. Those at one end have just started protein synthesis; those at the other end are about to complete production of a polypeptide chain.

The size of the polysome depends on several variables. In bacteria, it is very large, with tens of ribosomes simultaneously engaged in translation. Partly the size is due to the length of the mRNA (which usually codes for several proteins); partly it is due to the high efficiency with which the ribosomes attach to the mRNA.

Polysomes in the cytoplasm of a eukaryotic cell are likely to be smaller than those in bacteria; again, their size is a function both of the length of the mRNA (usually representing only a single protein in eukaryotes) and of the characteristic frequency with which ribosomes attach. An average eukaryotic mRNA probably has ~8 ribosomes attached at any one time.

Figure 5.11 illustrates the life cycle of the ribosome. Ribosomes are drawn from a pool (actually the pool consists of ribosomal subunits), used to translate an mRNA, and then return to the pool for further cycles. The number of ribosomes on each mRNA molecule synthesizing a particular protein is not precisely determined, in either bacteria or eukaryotes, but is a matter of statistical fluctuation, determined by the variables of mRNA size and efficiency.

An overall view of the attention devoted to protein synthesis in the intact bacterium is given in **Figure 5.12.** The 20,000 or so ribosomes account for a quarter of the cell mass. There are >3000 copies of each tRNA, and altogether, the tRNA molecules outnumber the ribosomes by

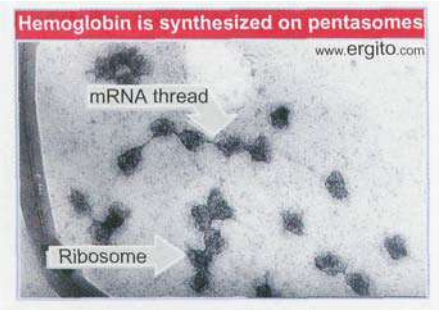


Figure 5.10 Protein synthesis occurs on polysomes. Photograph kindly provided by Alex Rich.

almost tenfold; most of them are present as aminoacyl-tRNAs, that is, ready to be used at once in protein synthesis. Because of their instability, it is difficult to calculate the number of mRNA molecules, but a reasonable guess would be ~1500, in varying states of synthesis and decomposition. There are ~600 different types of mRNA in a bacterium. This suggests that there are usually only 2-3 copies of each mRNA per bacterium. On average, each probably codes for ~3 proteins. If there are 1850 different soluble proteins, there must on average be ≥1000 copies of each protein in a bacterium.

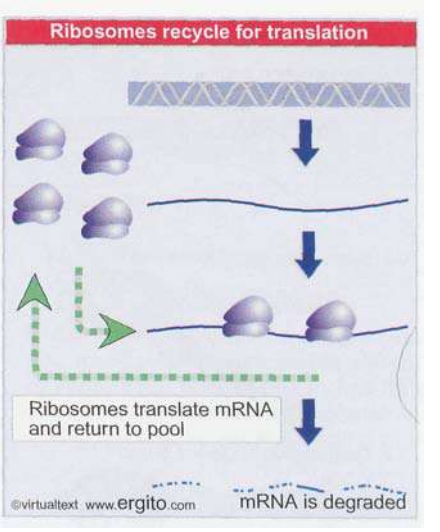


Figure 5.11 Messenger RNA is translated by ribosomes that cycle through a pool.

5.7 The life cycle of bacterial messenger RNA

Key Concepts

- Transcription and translation occur simultaneously in bacteria, as ribosomes begin translating an mRNA before its synthesis has been completed.
- Bacterial mRNA is unstable and has a half-life of only a few minutes.
- A bacterial mRNA may be polycistronic in having several coding regions that represent different genes.

Messenger RNA has the same function in all cells, but there are important differences in the details of the synthesis and structure of prokaryotic and eukaryotic mRNA.

A major difference in the production of mRNA depends on the locations where transcription and translation occur:

- In bacteria, mRNA is transcribed and translated in the single cellular compartment; and the two processes are so closely linked that they occur simultaneously. Since ribosomes attach to bacterial mRNA even before its transcription has been completed, the polysome is likely still to be attached to DNA. Bacterial mRNA usually is unstable, and is therefore translated into proteins for only a few minutes.
- In a eukaryotic cell, synthesis and maturation of mRNA occur exclusively in the nucleus. Only after these events are completed is the mRNA exported to the cytoplasm, where it is translated by ribosomes. Eukaryotic mRNA is relatively stable and continues to be translated for several hours.

Figure 5.13 shows that transcription and translation are intimately related in bacteria. Transcription begins when the enzyme RNA polymerase binds to DNA and then moves along making a copy of one strand. As soon as transcription begins, ribosomes attach to the 5′ end of the mRNA and start translation, even before the rest of the message has been synthesized. A bunch of ribosomes moves along the mRNA while it is being synthesized. The 3′ end of the mRNA is generated when transcription terminates. Ribosomes continue to translate the mRNA while it survives, but it is degraded in the overall 5′→3′ direction quite rapidly. The mRNA is synthesized, translated by the ribosomes, and degraded, all in rapid succession. An individual molecule of mRNA survives for only a matter of minutes or even less.

Bacterial transcription and translation take place at similar rates. At 37°C, transcription of mRNA occurs at ~40 nucleotides/second. This is very close to the rate of protein synthesis, roughly 15 amino acids/second. It therefore takes ~2 minutes to transcribe and translate an mRNA of 5000 bp, corresponding to 180 kD

25% of bacterial dry mass is concerned with gene expression

Component	Dry Cell Mass (%)	Molecules /cell	Different types	Copies of each type
Wall	10	1	1	1
Membrane	10	2	2	1
DNA	1.5	1	1	1
mRNA	1	1,500	600	2-3
tRNA	3	200,000	60	>3,000
rRNA	16	38,000	2	19,000
Ribosomal proteins	9	10^6	52	19,000
Soluble proteins	46	2.0×10^6	1,850	>1,000
Small molecules	3	7.5×10^6	800	

Figure 5.12 Considering *E. coli* in terms of its macromolecular components.

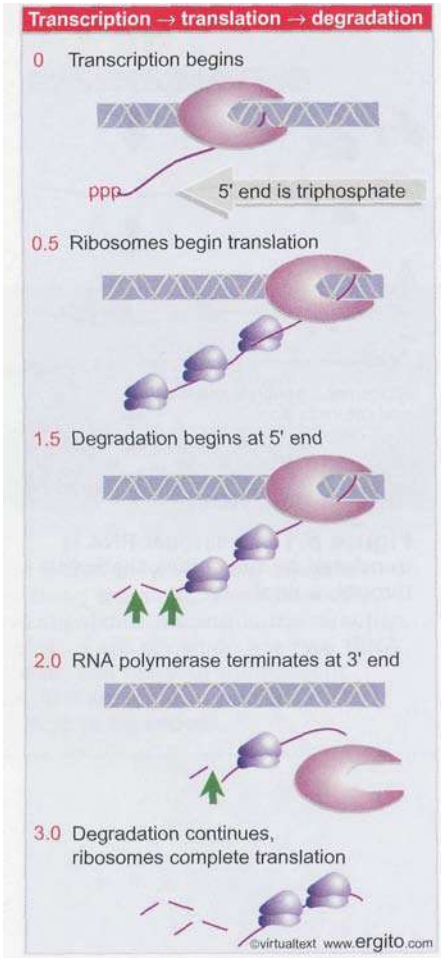


Figure 5.13 Overview: mRNA is transcribed, translated, and degraded simultaneously in bacteria.

of protein. When expression of a new gene is initiated, its mRNA typically will appear in the cell within ~2.5 minutes. The corresponding protein will appear within perhaps another 0.5 minute.

Bacterial translation is very efficient, and most mRNAs are translated by a large number of tightly packed ribosomes. In one example (*trp* mRNA), about 15 initiations of transcription occur every minute, and each of the 15 mRNAs probably is translated by ~30 ribosomes in the interval between its transcription and degradation.

The instability of most bacterial mRNAs is striking. Degradation of mRNA closely follows its translation. Probably it begins within 1 minute of the start of transcription. The 5' end of the mRNA starts to decay before the 3′ end has been synthesized or translated. Degradation seems to follow the last ribosome of the convoy along the mRNA. But degradation proceeds more slowly, probably at about half the speed of transcription or translation.

The stability of mRNA has a major influence on the amount of protein that is produced. It is usually expressed in terms of the half-life. The mRNA representing any particular gene has a characteristic half-life, but the average is ~2 minutes in bacteria.

This series of events is only possible, of course, because transcription, translation, and degradation all occur in the same direction. The dynamics of gene expression have been caught *in flagrante delicto* in the electron micrograph of **Figure 5.14.** In these (unknown) transcription units, several mRNAs are under synthesis simultaneously; and each carries many ribosomes engaged in translation. (This corresponds to the stage shown in the second panel in Figure 5.13.) An RNA whose synthesis has not yet been completed is often called a nascent RNA.

Bacterial mRNAs vary greatly in the number of proteins for which they code. Some mRNAs represent only a single gene: they are monocistronic. Others (the majority) carry sequences coding for several proteins: they are polycistronic. In these cases, a single mRNA is transcribed from a group of adjacent genes. (Such a cluster of genes constitutes an operon that is controlled as a single genetic unit; see *10 The operon.)*

All mRNAs contain two types of region. The coding region consists of a series of codons representing the amino acid sequence of the protein, starting (usually) with AUG and ending with a termination codon. But the mRNA is always longer than the coding region, extra regions are present at both ends. An additional sequence at the 5' end, preceding the start of the coding region, is described as the leader or 5' UTR (untranslated region). An additional sequence following the termination signal, forming the 3' end, is called the trailer or 3' UTR. Although part of the transcription unit, these sequences are not used to code for protein.

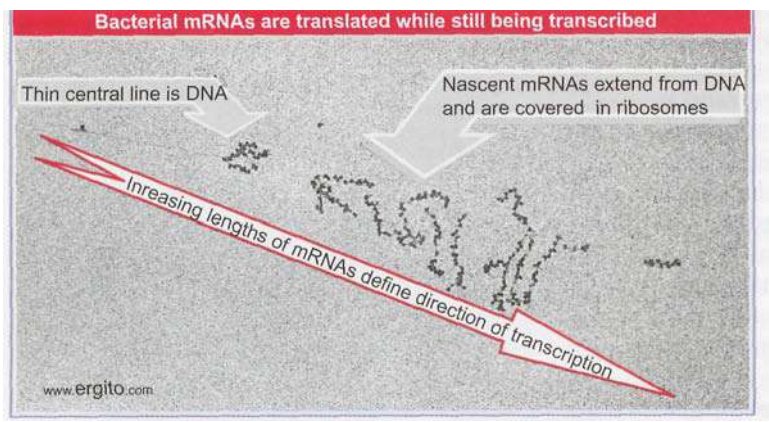


Figure 5.14 Transcription units can be visualized in bacteria. Photograph kindly provided by Oscar Miller.

A polycistronic mRNA also contains intercistronic regions, as illustrated in **Figure 5.15.** They vary greatly in size. They may be as long as 30 nucleotides in bacterial mRNAs (and even longer in phage RNAs), but they can also be very short, with as few as 1 or 2 nucleotides separating the termination codon for one protein from the initiation codon for the next. In an extreme case, two genes actually overlap, so that the last base of one coding region is also the first base of the next coding region.

The number of ribosomes engaged in translating a particular cistron depends on the efficiency of its initiation site. The initiation site for the first cistron becomes available as soon as the 5′ end of the mRNA is synthesized. How are subsequent cistrons translated? Are the several coding

regions in a polycistronic mRNA translated independently or is their expression connected? Is the mechanism of initiation the same for all cistrons, or is it different for the first cistron and the internal cistrons?

Translation of a bacterial mRNA proceeds sequentially through its cistrons. At the time when ribosomes attach to the first coding region, the subsequent coding regions have not yet even been transcribed. By the time the second ribosome site is available, translation is well under way through the first cistron. Usually ribosomes terminate translation at the end of the first cistron (and dissociate into subunits), and a new ribosome assembles independently at the start of the next coding region. (We discuss the processes of initiation and termination in *6 Protein Synthesis.*)

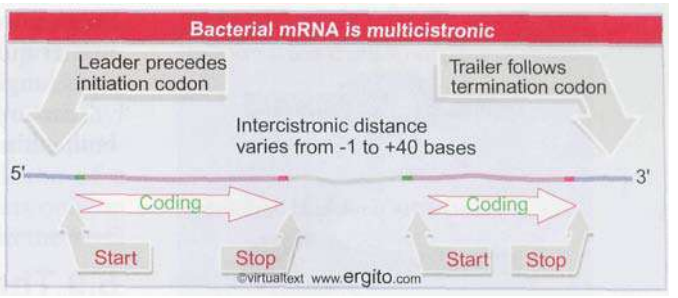


Figure 5.15 Bacterial mRNA includes non translated as well as translated regions. Each coding region has its own initiation and termination signals. A typical mRNA may have several coding regions.

5.8 Eukaryotic mRNA is modified during or after its transcription

Key Concepts

- A eukaryotic mRNA transcript is modified in the nucleus during or shortly after transcription.
- The modifications include the addition of a methylated cap at the 5' end and a sequence of poly(A) at the 3' end,
- The mRNA is exported from the nucleus to the cytoplasm only after all modifications have been completed.

The production of eukaryotic mRNA involves additional stages after transcription. Transcription occurs in the usual way, initiating a transcript with a 5′ triphosphate end. However, the 3′ end is generated by cleaving the transcript, rather than by terminating transcription at a fixed site. Those RNAs that are derived from interrupted genes require splicing to remove the introns, generating a smaller mRNA that contains an intact coding sequence.

Figure 5.16 shows that both ends of the transcript are modified by additions of further nucleotides (involving additional enzyme systems). The 5' end of the RNA *is modified by addition of a "cap" virtually as* soon as it appears. This replaces the triphosphate of the initial transcript with a nucleotide in reverse (3′→5′) orientation, thus "sealing" the end. The 3' end is modified by addition of a series of adenylic acid nucleotides [polyadenylic acid or poly(A)] immediately after its cleavage. Only after the completion of all modification and processing events can the mRNA be exported from the nucleus to the cytoplasm. The average delay in leaving for the cytoplasm is ~20 minutes. Once the mRNA has entered the cytoplasm, it is recognized by ribosomes and translated.

Figure 5.17 shows that the life cycle of eukaryotic mRNA is more protracted than that of bacterial mRNA. Transcription in animal cells occurs at about the same speed as in bacteria (~40 nucleotides per second). Many eukaryotic genes are large; a gene of 10,000 bp takes ~5 minutes to transcribe. Transcription of mRNA is not terminated by the release of enzyme from the DNA; instead the enzyme continues past the end of the gene. A coordinated series of events generates the 3′ end of the mRNA by cleavage, and adds a length of poly(A) to the newly generated 3′ end.

Eukaryotic mRNA constitutes only a small proportion of the total cellular RNA (~3% of the mass). Half-lives are relatively short in yeast,

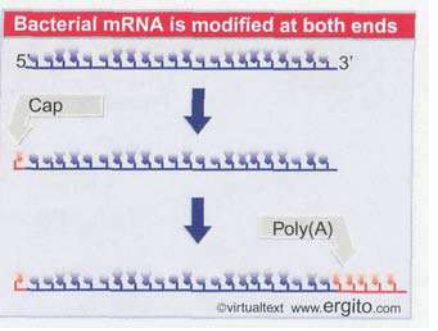


Figure 5.16 Eukaryotic mRNA is modified by addition of a cap to the 5' end and poly(A) to the 3' end.

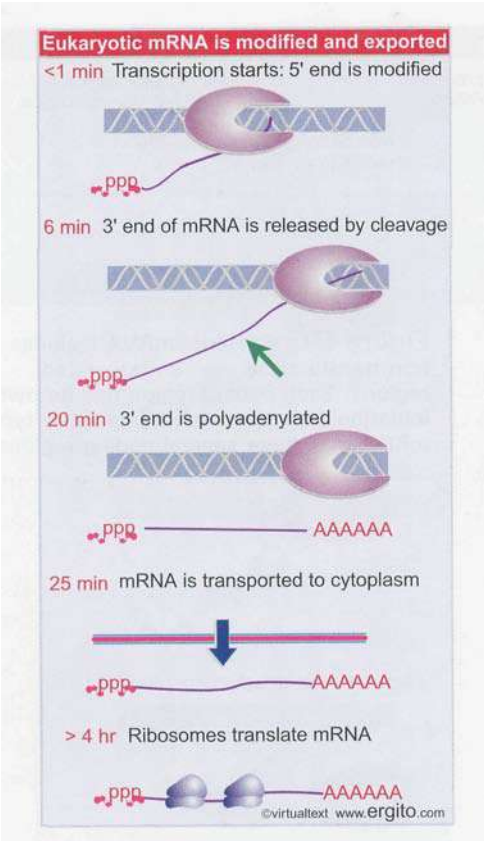


Figure 5.17 Overview: expression of mRNA in animal cells requires transcription, modification, processing, nucleocytoplasmic transport, and translation.

ranging from 1-60 minutes. There is a substantial increase in stability in higher eukaryotes; animal cell mRNA is relatively stable, with half-lives ranging from 1-24 hours.

Eukaryotic polysomes are reasonably stable. The modifications at both ends of the mRNA contribute to the stability.

5.9 The 5′ end of eukaryotic mRNA is capped

Key Concepts

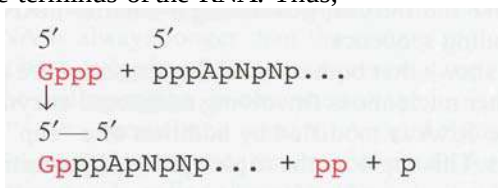
- A 5' cap is formed by adding a G to the terminal base of the transcript via a 5'-5' link. 1-3 methyl groups are added to the base or ribose of the new terminal guanosine.

Transcription starts with a nucleoside triphosphate (usually a purine, A or G). The first nucleotide retains its 5' triphosphate group and makes the usual phosphodiester bond from its 3' position to the 5' position of the next nucleotide. The initial sequence of the transcript can be represented as

$$5'\text{ppp}^{A}/_{G}\text{pNpNpNp...}$$

But when the mature mRNA is treated *in vitro* with enzymes that should degrade it into individual nucleotides, the 5' end does not give rise to the expected nucleoside triphosphate. Instead it contains two nucleotides, connected by a 5′–5′ triphosphate linkage and also bearing methyl groups. The terminal base is always a guanine that is added to the original RNA molecule after transcription.

Addition of the 5' terminal G is catalyzed by a nuclear enzyme, guanylyl transferase. The reaction occurs so soon after transcription has started that it is not possible to detect more than trace amounts of the original 5′ triphosphate end in the nuclear RNA. The overall reaction can be represented as a condensation between GTP and the original 5' triphosphate terminus of the RNA. Thus,



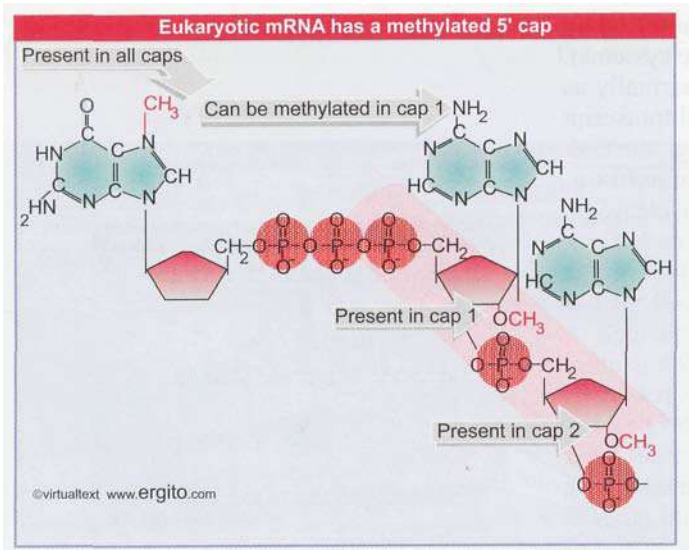


Figure 5.18 The cap blocks the 5' end of mRNA and may be methylated at several positions.

The new G residue added to the end of the RNA is in the reverse orientation from all the other nucleotides.

This structure is called a cap. It is a substrate for several methylation events. **Figure 5.18** shows the full structure of a cap after all possible methyl groups have been added. Types of caps are distinguished by how many of these methylations have occurred:

- The first methylation occurs in all eukaryotes, and consists of the addition of a methyl group to the 7 position of the terminal guanine. A cap that possesses this single methyl group is known as a cap 0. This is as far as the reaction proceeds in unicellular eukaryotes. The enzyme responsible for this modification is called guanine-7-methyltransferase.
- The next step is to add another methyl group to the 2'-O position of the penultimate base (which was actually the original first base of the transcript before any modifications were made). This

reaction is catalyzed by another enzyme (2′-O-methyl-transferase). A cap with the two methyl groups is called cap 1. This is the predominant type of cap in all eukaryotes except unicellular organisms.

- In a small minority of cases in higher eukaryotes, another methyl group is added to the second base. This happens only when the position is occupied by adenine; the reaction involves addition of a methyl group at the N^6 position. The enzyme responsible acts only on an adenosine substrate that already has the methyl group in the 2 -O position.
- In some species, a methyl group is added to the third base of the capped mRNA. The substrate for this reaction is the cap 1 mRNA that already possesses two methyl groups. The third-base modification is always a 2'-0 ribose methylation. This creates the cap 2 type. This cap usually represents less than 10-15% of the total capped population.

In a population of eukaryotic mRNAs, every molecule is capped. The proportions of the different types of cap are characteristic for a particular organism. We do not know whether the structure of a particular mRNA is invariant or can have more than one type of cap.

In addition to the methylation involved in capping, a low frequency of internal methylation occurs in the mRNA only of higher eukaryotes. This is accomplished by the generation of N^6 methyladenine residues at a frequency of about one modification per 1000 bases. There are 1-2 methyladenines in a typical higher eukaryotic mRNA, although their presence is not obligatory, since some mRNAs do not have any.

5.10 The 3′ terminus is polyadenylated

Key Concepts

- A length of poly(A) ~200 nucleotides long is added to a nuclear transcript after transcription.
- The poly(A) is bound by a specific protein (PABP).
- The poly(A) stabilizes the mRNA against degradation.

The 3' terminal stretch of A residues is often described as the poly(A) tail; and mRNA with this feature is denoted poly(A)$^+$.

The poly(A) sequence is not coded in the DNA, but is added to the RNA in the nucleus after transcription. The addition of poly(A) is catalyzed by the enzyme poly(A) polymerase, which adds ~200 A residues to the free 3'-OH end of the mRNA. The poly(A) tract of both nuclear RNA and mRNA is associated with a protein called the poly(A)-binding protein (PABP). Related forms of this protein are found in many eukaryotes. One PABP monomer of ~70 kD is bound every 10-20 bases of the poly(A) tail. So a common feature in many or most eukaryotes is that the 3' end of the mRNA consists of a stretch of poly(A) bound to a large mass of protein. Addition of poly(A) occurs as part of a reaction in which the 3' end of the mRNA is generated and modified by a complex of enzymes (see *24.19 The 3' ends of mRNAs are generated by cleavage and polyadenylation).*

Binding of the PABP to the initiation factor eIF4G generates a closed loop, in which the 5' and 3' ends of the mRNA find themselves held in the same protein complex (see Figure 6.20 in *6.9 Eukaryotes use a complex of many initiation factors).* The formation of this complex may be responsible for some of the effects of poly(A) on the properties of mRNA. Poly(A) usually stabilizes mRNA. The ability of the poly(A) to protect mRNA against degradation requires binding of the PABP.

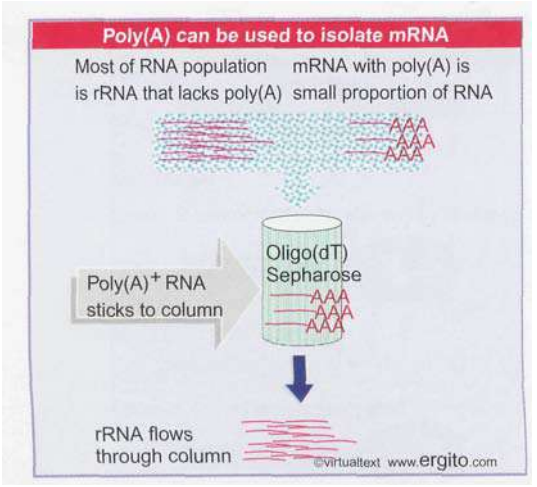


Figure 5.19 Poly(A)+ RNA can be separated from other RNAs by fractionation on Sepharose-oligo(dT).

Removal of poly(A) inhibits the initiation of translation *in vitro*, and depletion of PABP has the same effect in yeast *in vivo*. These effects could depend on the binding of PABP to the initiation complex at the 5' end of m RNA. There are many examples in early embryonic development where polyadenylation of a particular mRNA is correlated with its translation. In some cases, mRNAs are stored in a nonpolyadenylated form, and poly(A) is added when their translation is required; in other cases, poly(A)$^+$ mRNAs are de-adenylated, and their translation is reduced.

The presence of poly(A) has an important practical consequence. The poly(A) region of mRNA can base pair with oligo(U) or oligo(dT); and this reaction can be used to isolate poly(A)$^+$ mRNA. The most convenient technique is to immobilize the oligo(U or dT) on a solid support material. Then when an RNA population is applied to the column, as illustrated in Figure 5.19, only the poly(A)$^+$ RNA is retained. It can be retrieved by treating the column with a solution that breaks the bonding to release the RNA.

The only drawback to this procedure is that it isolates all the RNA that contains poly(A). If RNA of the whole cell is used, for example, both nuclear and cytoplasmic poly(A)$^+$ RNA will be retained. If preparations of polysomes are used (a common procedure), most of the isolated poly(A)$^+$ RNA will be active mRNA. However, in addition to mRNA in polysomes, there are also ribonucleoprotein particles in the cytosol that contain poly(A)$^+$ mRNA, but which are not translated. This RNA may be "stored" for use at some other time. Isolation of total poly(A)$^+$ mRNA therefore does not correspond exactly with the active mRNA population.

The "cloning" approach for purifying mRNA uses a procedure in which the mRNA is copied to make a complementary DNA strand (known as cDNA). Then the cDNA can be used as a template to synthesize a DNA strand that is identical with the original mRNA sequence. The product of these reactions is a double-stranded DNA corresponding to the sequence of the mRNA. This DNA can be reproduced in large amounts.

The availability of a cloned DNA makes it easy to isolate the corresponding mRNA by hybridization techniques. Even mRNAs that are present in only very few copies per cell can be isolated by this approach. Indeed, only mRNAs that are present in relatively large amounts can be isolated directly without using a cloning step.

Almost all cellular mRNAs possess poly(A). A significant exception is provided by the mRNAs that code for the histone proteins (a major structural component of chromosomal material). These mRNAs comprise most or all of the poly(A)" fraction. The significance of the absence of poly(A) from histone mRNAs is not clear, and there is no particular aspect of their function for which this appears to be necessary.

5.11 Bacterial mRNA degradation involves multiple enzymes

Key Concepts

- **The overall** direction of degradation of bacterial **mRNA** is 5'-3'.
- Degradation results **from the** combination of exonucleolytic cleavages followed by endonucleolytic degradation of the fragment from 3'-5'.

Bacterial mRNA is constantly degraded by a combination of endonucleases and exonucleases. Endonucleases cleave an RNA at an internal site. Exonucleases are involved in trimming reactions in which the extra residues are whittled away, base by base from the end.

Bacterial exonucleases that act on single-stranded RNA proceed along the nucleic acid chain from the 3′ end.

The way the two types of enzymes work together to degrade an *mRNA is* shown in **Figure 5.20.** Degradation of a bacterial mRNA is initiated by an endonucleolytic attack. Several 3′ ends may be generated by endonucleolytic cleavages within the mRNA. The overall direction of degradation (as measured by loss of ability to synthesize proteins) is from 5′ to 3′. This probably results from a succession of endonucleolytic cleavages following the last ribosome. Degradation of the released fragments of mRNA into nucleotides then proceeds by exonucleolytic attack from the free 3′–OH end toward the 5′ terminus (that is, in the opposite direction from transcription). Endonucleolytic attack releases fragments that may have different susceptibilities to exonucleases. A region of secondary structure within the mRNA may provide an obstacle to the exonuclease, thus protecting the regions on its 5' side. The stability of each mRNA is therefore determined by the susceptibility of its particular sequence to both endo- and exonucleolytic cleavages.

There are ~12 ribonucleases in *E. coli*. Mutants in the endoribonucleases (except ribonuclease I, which is without effect) accumulate unprocessed precursors to rRNA and tRNA, but are viable. Mutants in the exonucleases often have apparently unaltered phenotypes, which suggests that one enzyme can substitute for the absence of another. Mutants lacking multiple enzymes sometimes are inviable.

RNAase E is the key enzyme in initiating cleavage of mRNA. It may be the enzyme that makes the first cleavage for many mRNAs. Bacterial mutants that have a defective ribonuclease E have increased stability (2-3 fold) of mRNA. However, this is not its only function. RNAase E was originally discovered as the enzyme that is responsible for processing 5′ rRNA from the primary transcript by a specific endonucleolytic processing event.

The process of degradation may be catalyzed by a multienzyme complex (sometimes called the degradosome) that includes ribonuclease E, PNPase, and a helicase. RNAase E plays dual roles. Its N-terminal domain provides an endonuclease activity. The C-terminal domain provides a scaffold that holds together the other components. The helicase unwinds the substrate RNA to make it available to PNPase. According to this model, RNAase E makes the initial cut and then passes the fragments to the other components of the complex for processing.

Polyadenylation may play a role in initiating degradation of some mRNAs in bacteria. Poly(A) polymerase is associated with ribosomes in *E. coli*, and short (10-40 nucleotide) stretches of poly(A) are added to at least some mRNAs. Triple mutations that remove poly(A) polymerase, ribonuclease E, and polynucleotide phosphorylase (PNPase is a 3′–5′ exonuclease) have a strong effect on stability. (Mutations in individual genes or pairs of genes have only a weak effect.) Poly(A) polymerase may create a poly(A) tail that acts as a binding site for the nucleases. The role of poly(A) in bacteria would therefore be different from that in eukaryotic cells.

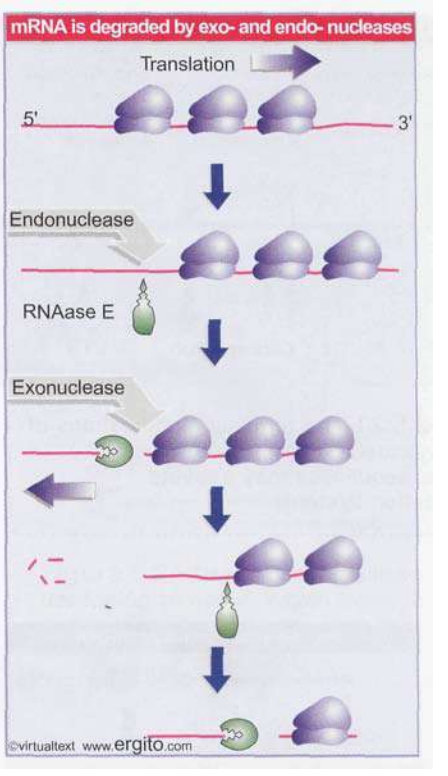


Figure 5.20 Degradation of bacterial mRNA is a two stage process. Endonucleolytic cleavages proceed 5'-3' behind the ribosomes. The released fragments are degraded by exonucleases that move 3'-5'.

5.12 mRNA stability depends on its structure and sequence

Key Concepts

- The modifications at both ends of mRNA protect it against degradation by exonucleases.
- Specific sequences within an mRNA may have stabilizing or destabilizing effects.
- Destabilization may be triggered by loss of poly(A).

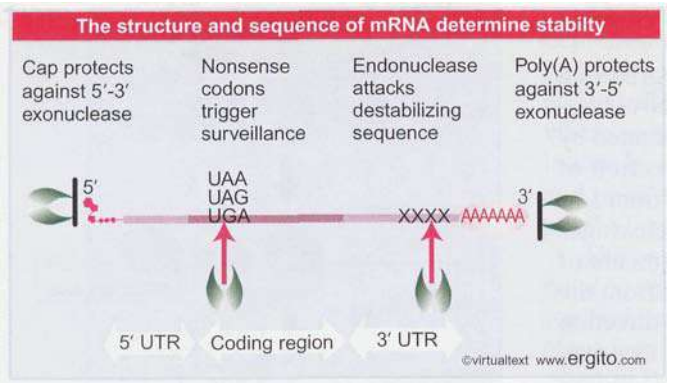


Figure 5.21 The terminal modifications of mRNA protect it against degradation. Internal sequences may activate degradation systems.

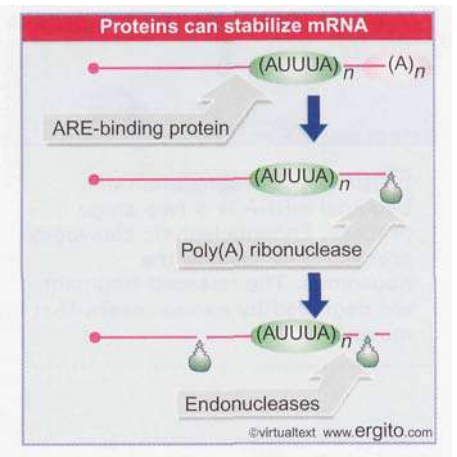


Figure 5.22 An ARE in a 3' nontranslated region initiates degradation of mRNA.

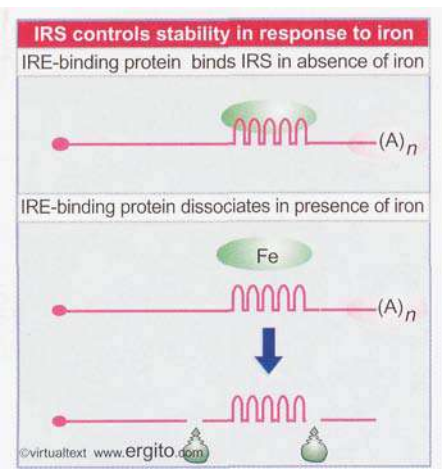


Figure 5.23 An IRE in a 3' nontranslated region controls mRNA stability.

The major features of mRNA that affect its stability are summarized in **Figure 5.21.** Both structure and sequence are important. The 5' and 3' terminal structures protect against degradation, and specific sequences within the mRNA may either serve as targets to trigger degradation or may protect against degradation:

- The modifications at the 5′ and 3′ ends of mRNA play an important role in preventing exonuclease attack. The cap prevents 5′-3′ exonucleases from attacking the 5' end, and the poly(A) prevents 3′-5′ exonucleases from attacking the 3′ end.
- Specific sequence elements within the mRNA may stabilize or destabilize it. The most common location for destabilizing elements is within the 3' untranslated region. The presence of such an element shortens the lifetime of the mRNA.
- Within the coding region, mutations that create termination codons trigger a surveillance system that degrades the mRNA (see *5.14 Nonsense mutations trigger a surveillance system).*

Destabilizing elements have been found in several yeast mRNAs, although as yet we do not see any common sequences or know how they destabilize the mRNA. They do not necessarily act directly (by providing targets for endonucleases), but may function indirectly, perhaps by encouraging deadenylation. The criterion for defining a destabilizing sequence element is that its introduction into a new mRNA may cause it to be degraded. The removal of an element from an mRNA does not necessarily stabilize it, suggesting that an individual mRNA can have more than one destabilizing element.

A common feature in some unstable mRNAs is the presence of an AU-rich sequence of ~50 bases (called the ARE) that is found in the 3' trailer region. The consensus sequence in the ARE is the pentanucleotide AUUUA, repeated several times. **Figure 5.22** shows that the ARE triggers destabilization by a two stage process: first the mRNA is deadenylated; then it decays. The deadenylation is probably needed because it causes loss of the poly(A)-binding protein, whose presence stabilizes the 3' region (see next section).

In some cases, an mRNA can be stabilized by specifically inhibiting the function of a destabilizing element. Transferrin mRNA contains a sequence called the IRE, which controls the response of the mRNA to changes in iron concentration. The IRE is located in the 3' nontranslated region, and contains stem-loop structures that bind a protein whose affinity for the mRNA is controlled by iron. **Figure 5.23** shows that binding of the protein to the IRE stabilizes the mRNA by inhibiting the function of (unidentified) destabilizing sequences in the vicinity. This is a general model for the stabilization of mRNA, that is, stability is conferred by inhibiting the function of destabilizing sequences.

5.13 mRNA degradation involves multiple activities

Key Concepts

- Degradation of yeast mRNA requires removal of the 5' cap and the 3' poly(A).
- One yeast pathway involves exonucleolytic degradation from 5′-3′.
- Another yeast pathway uses a complex of several exonucleases that work in the 3'-5' direction.
- The deadenylase of animal cells may bind directly to the 5' cap.

We know most about the degradation of mRNA in yeast. There are basically two pathways. Both start with removal of the poly(A) tail. This is catalyzed by a specific deadenylase which probably functions as part of a large protein complex. (The catalytic subunit is the exonuclease Ccr4 in yeast, and is the exonuclease PARN in vertebrates, which is related to RNAase D.) The enzyme action is processive—once it has started to degrade a particular mRNA substrate, it continues to whittle away that mRNA, base by base.

The major degradation pathway is summarized in **Figure 5.24.** Deadenylation at the 3′ end triggers decapping at the 5′ end. The basis for this relationship is that the presence of the PABP (poly(A)-binding protein) on the poly(A) prevents the decapping enzyme from binding to the 5' end. PABP is released when the length of poly(A) falls below 10-15 residues. The decapping reaction occurs by cleavage 1-2 bases from the 5′ end.

Each end of the mRNA influences events that occur at the other end. This is explained by the fact that the two ends of the mRNA are held together by the factors involved in protein synthesis (see *6.9 Eukaryotes use a complex of many initiation factors).* The effect of PABP on decapping allows the 3' end to have an effect in stabilizing the 5' end. There is also a connection between the structure at the 5′ end and degradation at the 3' end. The deadenylase directly binds to the 5' cap, and this interaction is in fact needed for its exonucleolytic attack on the poly(A).

What is the rationale for the connection between events occurring at both ends of an mRNA? Perhaps it is necessary to ensure that the mRNA is not left in a state (having the structure of one end but not the other) that might compete with active mRNA for the proteins that bind to the ends.

Removal of the cap triggers the 5′-3′ degradation pathway in which the mRNA is degraded rapidly from the 5' end, by the 5′–3′ exonuclease XRN1.

In the second pathway, deadenylated yeast mRNAs can be degraded by the 3′–5′ exonuclease activity of the exosome, a complex of >9 exonucleases. The exosome is also involved in processing precursors for rRNAs. The aggregation of the individual exonucleases into the exosome complex may enable 3′–5′ exonucleolytic activities to be coordinately controlled. The exosome may also degrade fragments of mRNA released by endonucleolytic cleavage. **Figure 5.25** shows that the 3′–5′ degradation pathway may actually involve combinations of endonucleolytic and exonucleolytic action. The exosome is also found in the nucleus, where it degrades unspliced precursors to mRNA.

Yeast mutants lacking either exonucleolytic pathway degrade their mRNAs more slowly, but the loss of both pathways is lethal.

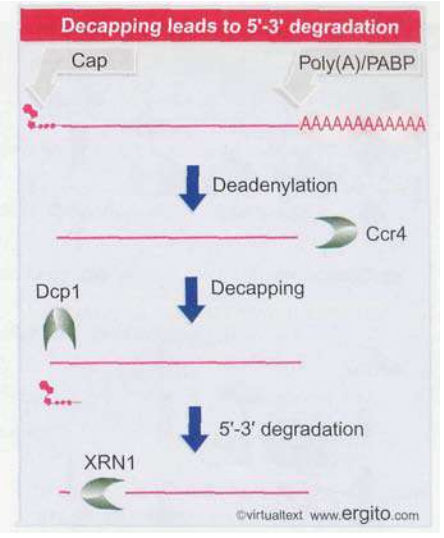


Figure 5.24 Deadenylation allows decapping to occur, which leads to endonucleolytic cleavage from the 5' end.

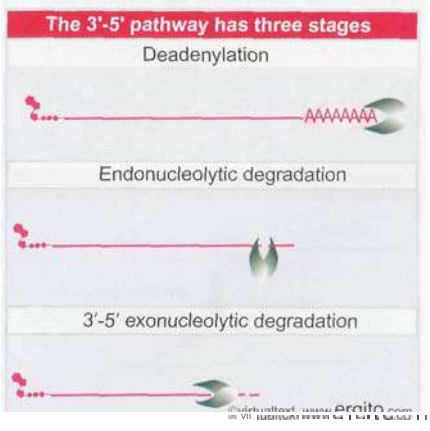


Figure 5.25 Deadenylation may lead directly to exonucleolytic cleavage and endonucleolytic cleavage from 3' end(s).

5.14 Nonsense mutations trigger a surveillance system

Key Concepts

- Nonsense mutations cause mRNA to be degraded.
- Genes coding for the degradation system have been found in yeast and worm.

Another pathway for degradation is identified by nonsense-mediated mRNA decay. **Figure 5.26** shows that the introduction of a nonsense mutation often leads to increased degradation of the mRNA. As may be expected from dependence on a termination codon, the degradation occurs in the cytoplasm. It may represent a

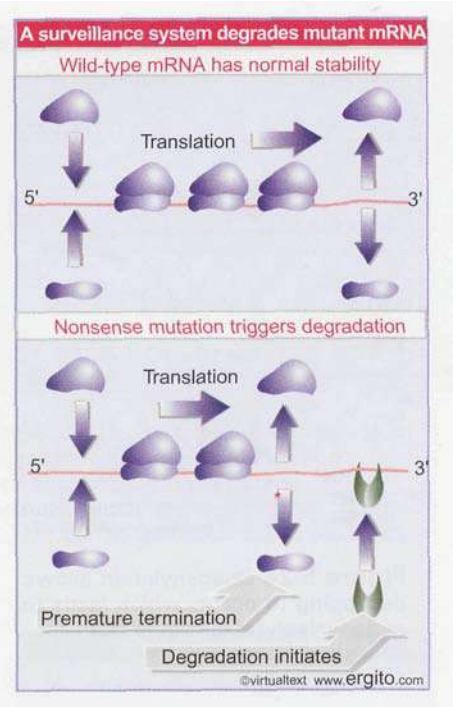


Figure 5.26 Nonsense mutations may cause mRNA to be degraded.

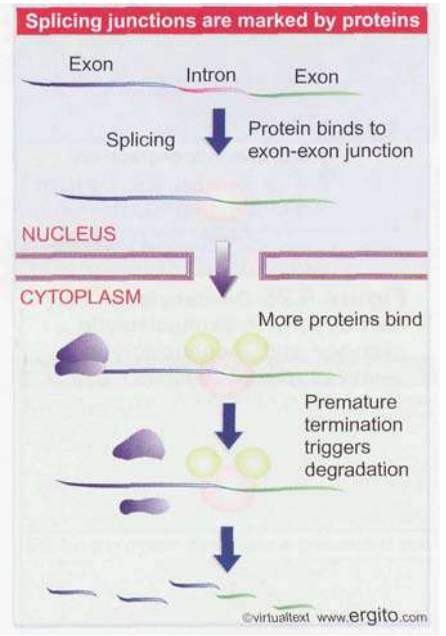


Figure 5.27 A surveillance system could have two types of components. Protein(s) must bind in the nucleus to mark the result of a splicing event. Other proteins could bind to the mark either in the nucleus or cytoplasm. They are triggered to act to degrade the mRNA when ribosomes terminate prematurely.

quality control or surveillance system for removing nonfunctional mRNAs.

The surveillance system has been studied best in yeast and *C. elegans*, but may also be important in animal cells. For example, during the formation of immunoglobulins and T cell receptors in cells of the immune system, genes are modified by somatic recombination and mutation (see *26 Immune diversity*). This generates a significant number of nonfunctional genes, whose RNA products are disposed of by a surveillance system.

In yeast, the degradation requires sequence elements (called *DSE*) that are downstream of the nonsense mutation. The simplest possibility would be that these are destabilizing elements, and that translation suppresses their use. However, when translation is blocked, the mRNA is stabilized. This suggests that the process of degradation is linked to translation of the mRNA or to the termination event in some direct way.

Genes that are required for the process have been identified in *S. cerevisiae* (*upf* loci) and *C. elegans* (*smg* loci) by identifying suppressors of nonsense-mediated degradation. Mutations in these genes stabilize aberrant mRNAs, but do not affect the stability of most wild-type transcripts. One of these genes is conserved in eukaryotes (*upf1/smg2*). It codes for an ATP-dependent helicase (an enzyme that unwinds double-stranded nucleic acids into single strands). This implies that recognition of the mRNA as an appropriate target for degradation requires a change in its structure.

Upf1 interacts with the release factors (eRF1 and eRF3) that catalyze termination, which is probably how it recognizes the termination event. It may then "scan" the mRNA by moving toward the 3' end to look for the downstream sequence elements.

In mammalian cells, the surveillance system appears to work only on mutations located prior to the last exon—in other words, there must be an intron after the site of mutation. This suggests that the system requires some event to occur in the nucleus, before the introns are removed by splicing. One possibility is that proteins attach to the mRNA in the nucleus at the exon-exon boundary when a splicing event occurs. **Figure 5.27** shows a general model for the operation of such a system. This is similar to the way in which an mRNA may be marked for export from the nucleus (see *24.10 Splicing is connected to export of mRNA*). Attachment of a protein to the exon-exon junction creates a mark of the event that persists into the cytoplasm. Human homologues of the yeast Upf 2,3 proteins may be involved in such a system. They bind specifically to mRNA that has been spliced.

5.15 Eukaryotic RNAs are transported

Key Concepts

- RNA is transported through a membrane as a ribonucleoprotein particle.
- All eukaryotic RNAs that function in the cytoplasm must be exported from the nucleus.
- tRNAs and the RNA component of a ribonuclease are imported into mitochondria.
- mRNAs can travel long distances between plant cells.

A bacterium consists of only a single compartment, so all the RNAs function in the same environment in which they are

synthesized. This is most striking in the case of mRNA, where translation occurs simultaneously with transcription *(see J. 7 The life cycle of bacterial messenger RNA).*

RNA is transported through membranes in the variety of instances summarized in **Figure 5.28.** It poses a significant thermodynamic problem to transport a highly negative RNA through a hydrophobic membrane, and the solution is to transport the RNA packaged with proteins.

In eukaryotic cells, RNAs are transcribed in the nucleus, but translation occurs in the cytoplasm. Each type of RNA must be transported into the cytoplasm to assemble the apparatus for translation. The rRNA assembles with ribosomal proteins into immature ribosome subunits that are the substrates for the transport system. tRNA is transported by a specific protein system (see *8.28 Transport receptors carry cargo proteins through the pore).* mRNA is transported as a ribonucleoprotein, which forms on the RNA transcript in the nucleus (see *24 RNA splicing and processing).* These processes are common to all eukaryotic cells. Many mRNAs are translated in the cytosol, but some are localized within the cell, by means of attachment to a cytoskeletal element. One situation in which localization occurs is when it is important for a protein product to be produced near to the site of its incorporation into some macromolecular structure.

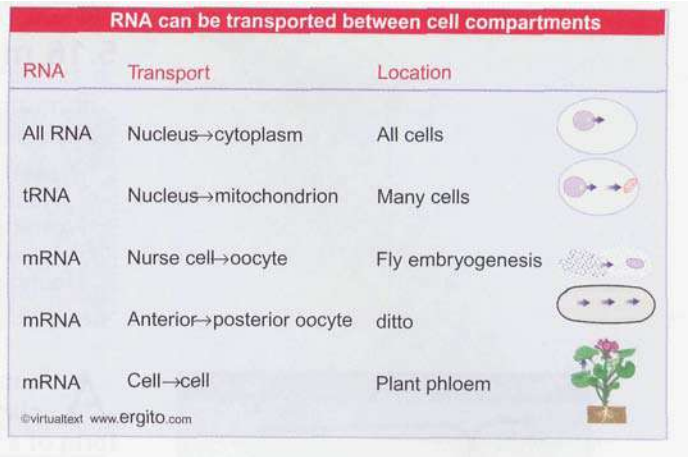


Figure 5.28 RNAs are transported through membranes in a variety of systems.

Some RNAs are made in the nucleus, exported to the cytosol, and then imported into mitochondria. The mitochondria of some organisms do not code for all of the tRNAs that are required for protein synthesis (see *3.19 Organelle genomes are circular DNAs that code for organelle proteins).* In these cases, the additional tRNAs must be imported from the cytosol. The enzyme ribonuclease P, which contains both RNA and protein subunits, is coded by nuclear genes, but is found in mitochondria as well as the nucleus. This means that the RNA must be imported into the mitochondria.

We know of some situations in which mRNA is even transported between cells. During development of the oocyte in Drosophila, certain mRNAs are transported into the egg from the nurse cells that surround it. The nurse cells have specialized junctions with the oocyte that allow passage of material needed for early development. This material includes certain mRNAs. Once in the egg, these mRNAs take up specific locations. Some simply diffuse from the anterior end where they enter, but others are transported the full length of the egg to the posterior end by a motor attached to microtubules (see *31.7 How are mRNAs and proteins transported and localized?).*

The most striking case of transport of mRNA has been found in plants. Movement of individual nucleic acids over long distances was first discovered in plants, where viral movement proteins help propagate the viral infection by transporting an RNA virus genome through the plasmodesmata (connections between cells). Plants also have a defense system, that causes cells to silence an infecting virus, and this too may involve the spread of components including RNA over long distance between cells. Now it has turned out that similar systems may transport mRNAs between plant cells. Although the existence of the systems has been known for some time, it is only recently that their functional importance has been demonstrated. This was shown by grafting wild-type tomato plants onto plants that had the dominant mutation *Me* (which causes a change in the shape of the leaf). mRNA from the mutant stock was transported into the leaves of the wild-type graft, where it changed their shape.

5.16 mRNA can be specifically localized

Key Concepts

- Yeast Ash1 mRNA forms a ribonucleoprotein that binds to a myosin motor.
- A motor transports it along actin filaments into the daughter bud.
- It is anchored and translated in the bud, so that the protein is found only in the bud.

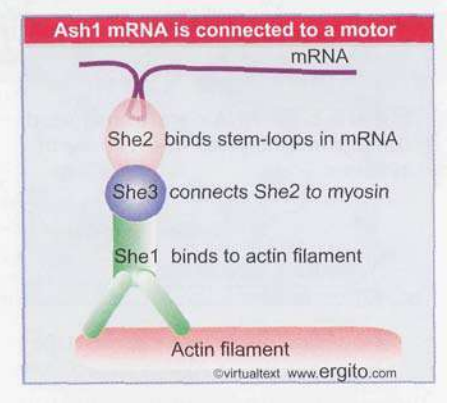


Figure 5.29 Ash1 mRNA forms a ribonucleoprotein containing a myosin motor that moves it along an actin filament.

An mRNA is synthesized in the nucleus but translated in the cytoplasm of a eukaryotic cell. It passes into the cytoplasm in the form of a ribonucleoprotein particle that is transported through the nuclear pore. Once in the cytosol, it may associate with ribosomes and be translated. The cytosol is a crowded place, occupied by a high concentration of proteins. It is not clear how freely a polysome can diffuse within the cytosol, and most mRNAs are probably translated in random locations, determined by their point of entry into the cytosol, and the distance that they may have moved away from it. However, some mRNAs are translated at specific sites. This may be accomplished by several mechanisms:

- An mRNA may be specifically transported to a site where it is translated.
- It may be universally distributed but degraded at all sites except the site of translation.
- It may be freely diffusible but become trapped at the site of translation.

One of the best characterized cases of localization within a cell is that of Ash 1 in yeast. Ash1 represses expression of the HO endonuclease in the budding daughter cell, with the result that HO is expressed only in the mother cell. The consequence is that mating type is changed only in the mother cell (see *18.9 Regulation of HO expression controls switching).* The cause of the restriction to the daughter cell is that all the Ashl mRNA is transported from the mother cell, where it is made, into the budding daughter cell.

Mutations in any one of 5 genes, called *SHE1-5,* prevent the specific localization and cause Ashl mRNA to be symmetrically distributed in both mother and daughter compartments. The proteins She 1,2,3 bind Ashl mRNA into a ribonucleoprotein particle that transports the mRNA into the daughter cell. **Figure 5.29** shows the functions of the proteins. She1p is a myosin (previously identified as Myo4), and She3 and She2 are proteins that connect the myosin to the mRNA. The myosin is a motor that moves the mRNA along actin filaments.

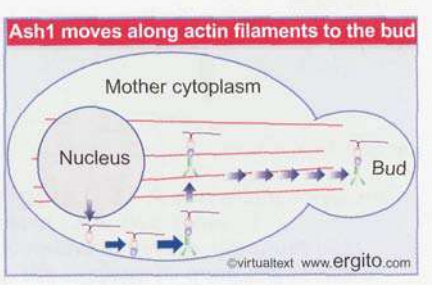


Figure 5.30 Ash1 mRNA is exported from the nucleus into the cytoplasm where it is assembled into a complex with the She proteins. The complex transports it along actin filaments to the bud.

Figure 5.30 summarizes the overall process. Ash1 mRNA is exportec from the nucleus in the form of a ribonucleoprotein. In the cytoplasm it is first bound by She2, which recognizes some stem-loop secondary structures within the mRNA. Then She3 binds to She2, after which the myosin She1 binds. Then the particle hooks on to an actin filament anc moves to the bud. When Ash1 mRNA reaches the bud, it is anchorec there, probably by proteins that bind specifically to the mRNA.

Similar principles govern other cases where mRNAs are transportec to specific sites. The mRNA is recognized by means of *cis*-acting se quences, which usually are regions of secondary structure in the 3' un translated region. (Ashl mRNA is unusual in that the *cis*-acting region are in the coding frame.) The mRNA is packaged into a ribonucleopro tein particle. In some cases, the transported mRNA can be visualized i very large particles, called mRNA granules. The particles are larg

enough (several times the size of a ribosome) to contain many protein and RNA components.

A transported mRNP must be connected to a motor that moves it along a system of tracks. The tracks can be either actin filaments or microtubules. Whereas Ash1 uses a myosin motor on actin tracks, *oscar* mRNA in the *Drosophila* egg uses a kinesin motor to move along microtubules (see *31.7 How are mRNAs and proteins transported and localized?*). Once the mRNA reaches its destination, it needs to be anchored in order to prevent it from diffusing away. Less is known about this, but the process appears to be independent of transport. An mRNA that is transported along microtubules may anchored to actin filaments at its destination.

5.17 Summary

Genetic information carried by DNA is expressed in two stages: transcription of DNA into mRNA; and translation of the mRNA into protein. Messenger RNA is transcribed from one strand of DNA and is complementary to this (noncoding) strand and identical with the other (coding) strand. The sequence of mRNA, in triplet codons 5'-3', is related to the amino acid sequence of protein, N- to C-terminal.

The adaptor that interprets the meaning of a codon is transfer RNA, which has a compact L-shaped tertiary structure; one end of the tRNA has an anticodon that is complementary to the codon, and the other end can be covalently linked to the specific amino acid that corresponds to the target codon. A tRNA carrying an amino acid is called an aminoacyl-tRNA.

The ribosome provides the apparatus that allows aminoacyl-tRNAs to bind to their codons on mRNA. The small subunit of the ribosome is bound to mRNA; the large subunit carries the nascent polypeptide. A ribosome moves along mRNA from an initiation site in the 5' region to a termination site in the 3' region, and the appropriate aminoacyl-tRNAs respond to their codons, unloading their amino acids, so that the growing polypeptide chain extends by one residue for each codon traversed.

The translational apparatus is not specific for tissue or organism; an mRNA from one source can be translated by the ribosomes and tRNAs from another source. The number of times any mRNA is translated is a function of the affinity of its initiation site(s) for ribosomes and its stability. There are some cases in which translation of groups of mRNA or individual mRNAs is specifically prevented: this is called translational control.

A typical mRNA contains both a nontranslated 5' leader and 3' trailer as well as coding region(s). Bacterial mRNA is usually polycistronic, with nontranslated regions between the cistrons. Each cistron is represented by a coding region that starts with a specific initiation site and ends with a termination site. Ribosome subunits associate at the initiation site and dissociate at the termination site of each coding region.

A growing *E. coli* bacterium has ~20,000 ribosomes and ~200,000 tRNAs, mostly in the form of aminoacyl-tRNA. There are ~1500 mRNA molecules, representing 2-3 copies of each of 600 different messengers.

A single mRNA can be translated by many ribosomes simultaneously, generating a polyribosome (or polysome). Bacterial polysomes are large, typically with tens of ribosomes bound to a single mRNA. Eukaryotic polysomes are smaller, typically with fewer than 10 ribosomes; each mRNA carries only a single coding sequence.

Bacterial mRNA has an extremely short half-life, only a few minutes. The 5' end starts translation even while the downstream

sequences are being transcribed. Degradation is initiated by endonucleases that cut at discrete sites, following the ribosomes in the 5'-3' direction, after which exonucleases reduce the fragments to nucleotides by degrading them from the released *3'* end toward the 5' end. Individual sequences may promote or retard degradation in bacterial mRNAs.

Eukaryotic mRNA must be processed in the nucleus before it is transported to the cytoplasm for translation. A methylated cap is added to the 5' end. It consists of a nucleotide added to the original end by a 5'-5' bond, after which methyl groups are added. Most eukaryotic mRNA has an ~200 base sequence of poly(A) added to its 3' terminus in the nucleus after transcription, but poly(A)$^-$ mRNAs appear to be translated and degraded with the same kinetics as poly(A)$^+$ mRNAs. Eukaryotic mRNA exists as a ribonucleoprotein particle; in some cases mRNPs are stored that fail to be translated. Eukaryotic mRNAs are usually stable for several hours. They may have multiple sequences that initiate degradation; examples are known in which the process is regulated.

Yeast mRNA is degraded by (at least) two pathways. Both start with removal of poly(A) from the 3' end, causing loss of poly(A)-binding protein, which in turn leads to removal of the methylated cap from the 5' end. One pathway degrades the mRNA from the 5' end by an exonuclease. Another pathway degrades from the 3' end by the exosome, a complex containing several exonucleases.

Nonsense-mediated degradation leads to the destruction of mRNAs that have a termination (nonsense) codon prior to the last exon. The *upf* loci in yeast and the *smg* loci in worms are required for the process. They includes a helicase activity to unwind mRNA and a protein that interacts with the factors that terminate protein synthesis. The features of the process in mammalian cells suggest that some of the proteins attach to the mRNA in the nucleus when RNA splicing occurs to remove introns.

mRNAs can be transported to specific locations within a cell (especially in embryonic development). In the Ash1 system in yeast, mRNA is transported from the mother cell into the daughter cell by a myosin motor that moves on actin filaments. In plants, mRNAs can be transported long distances between cells.

References

5.3 Transfer RNA forms a cloverleaf

rev Soll, D. and RajBhandary, U. L. (1995). tRNA Structure, Biosynthesis, and Function. American Society for Microbiology, Washington DC.

ref Chapeville, F. et al. (1962). On the role of soluble RNA in coding for amino acids. Proc. Nat. Acad. Sci. USA 48, 1086-1092.

Hoagland, M. B. et al. (1958). A soluble RNA intermediate in protein synthesis. J. Biol. Chem. 231, 241-257.

Holley, R. W. et al. (1965). Structure of an RNA. Science 147, 1462-1465.

5.5 Messenger RNA is translated by ribosomes

ref Dintzis, H. M. (1961). Assembly of the peptide chain of hemoglobin. Proc. Nat. Acad. Sci. USA 47, 247-261.

5.6 Many ribosomes bind to one mRNA

ref Slayter, H. S. et al. (1963). The visualization of polyribosome structure. J. Mol. Biol. 7, 652-657.

5.7 The life cycle of bacterial messenger RNA

ref Brenner, S. Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature 190, 576-581.

5.9 The 5' end of eukaryotic mRNA is capped

rev Bannerjee, A. K. (1980). 5'-terminal cap structure in eukaryotic mRNAs. Microbiol. Rev. 44, 175-205.

5.10 The 3' terminus is polyadenylated

rev Jackson, R. J. and Standart, N. (1990). Do the poly(A) tail and 3' untranslated region control mRNA translation? Cell 62, 15-24.

ref Darnell, J. et al. (1971). Poly(A) sequences: role in conversion of nuclear RNA into mRNA. Science 174 507-510.

5.11 Bacterial mRNA degradation involves multiple enzym

rev Caponigro, G. and Parker, R. (1996). Mechanisms and control of mRNA turnover in S. *cerevlsiae.* Microbiol. Rev. 60, 233-249.

Grunberg-Manago, M. (1999). mRNA stability and it role in control of gene expression in bacteria and phages. Ann. Rev. Genet. 33, 193-227.

ref Miczak, A., Kaberdin, V. R., Wei, C.-L., and Lin-chao, S. (1996). Proteins associated with RNAase i in a multicomponent ribonucleolytic complex. Proc. Nat. Acad. Sci. USA 93, 3865-3869.

O'Hara, E. B. et al. (1995). Polyadenylation helps regulate mRNA decay in *E. coli.* Proc. Nat. Acad. S USA 92, 1807-1811.

By Book_Crazy [IND]

Vanzo, N. F. et al. (1998). RNAase E organizes the protein interactions in the *E. coli* RNA degradaosome. Genes Dev. 12, 2770-2781.

5.12 mRNA stability depends on its structure and sequence

rev Ross, J. (1995). mRNA stability in mammalian cells. Microbiol. Rev. 59, 423-450.

Sachs, A. (1993). Messenger RNA degradation in eukaryotes. Cell 74, 413-421.

5.13 mRNA degradation involves multiple activities

rev Jacobson, A. and Peltz, S. W. (1996). Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Ann. Rev. Biochem. 65, 693-739.

ref Allmang, C, Petfalski, E., Podtelejnikov, A., Mann, M., Tollervey, D., and Mitchell, P. (1999). The yeast exosome and human PM-Scl are related complexes of 3'-5' exonucleases. Genes Dev. 13, 2148-2158.

Beelman, C. A. and Parker, R. (1995). Degradation of mRNA in eukaryotes. Cell 81, 179-183.

Bousquet-Antonelli, C, Presutti, C, and Tollervey, D. (2000). Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell 102, 765-775.

Gao, M., Fritz, D. T., Ford, L. P., and Wilusz, J. (2000). Interaction between a poly(A)-specific ribonuclease and the 5' cap influences mRNA deadenylation rates *in vitro.* Mol. Cell 5, 479-488.

Mitchell, P. et al. (1997). The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-5' exoribonuclease activities. Cell 91, 457-466.

Muhlrad, D., Decker, C. J., and Parker, R. (1994). Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'-3' digestion of the transcript. Genes Dev. 8, 855-866.

Tucker, M., Valencia-Sanchez, M. A., Staples, R. R., Chen, J., Denis, C. L., and Parker, R. (2001). The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in *S. cerevisiae.* Cell 104, 377-386.

5.14 Nonsense mutations trigger a surveillance system

rev Hilleren, P. and Parker, R. (1999). Mechanisms of mRNA surveillance in eukaryotes. Ann. Rev. Genet. 33, 229-260.

ref Cui, Y., Hagan, K. W., Zhang, S., and Peltz, S. W. (1995). Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. Genes Dev. 9, 423-436.

Czaplinski, K., Ruiz-Echevarria, M. J., Paushkin, S. V., Han, X., Weng, Y., Perlick, H. A., Dietz, H. C, Ter-Avanesyan, M. D., and Peltz, S. W. (1998). The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs. Genes Dev. 12, 1665-1677.

Le Hir, H., Moore, M. J., and Maquat, L. E. (2000). Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev. 14, 1098-1108.

Lykke-Andersen, J., Shu, M. D., and Steitz, J. A. (2000). Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. Cell 103, 1121-1131.

Peltz, S. W., Brown, A. H., and Jacobson, A. (1993). mRNA destabilization triggered by premature translational termination depends on at least three *cis*-acting sequence elements and one frans-acting factor. Genes Dev. 7, 1737-1754.

Pulak, R. and Anderson, P. (1993). mRNA surveillance by the *C. elegans smg* genes. Genes Dev. 7, 1885-1897.

Ruiz-Echevarria, M. J. et al. (1998). Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA. EMBO J. 15, 2810-2819.

Weng, Y., Czaplinski, K., and Peltz, S. (1996). Genetic and biochemical characterization of mutants in the ATPase and helicase regions of the Upf1 protein. Mol. Cell Biol. 16, 5477-5490.

Weng, Y., Czaplinski, K., and Peltz, S. (1996). Identification and characterization of mutations in the *upf1* gene that affect the Upf protein complex, nonsense suppression, but not mRNA turnover. Mol. Cell Biol. 16, 5491-5506.

5.15 Eukaryotic RNAs are transported

rev Ghoshroy, S., Lartey, R., Sheng, J., and Citovsky, V. (1997). Transport of proteins and nucleic acids through plasmodesmata. Ann. Rev. Plant. Physiol. Plant. Mol. Biol. 48, 27-50.

Lucas, W. J. and Gilbertson, R. L. (1994). Plasmodesmata in relation to viral movement within leaf tissues. Ann. Rev. Phytopathol. 32, 387-411.

ref Jansen, R. P. (2001). mRNA localization: message on the move. Nat. Rev. Mol. Cell Biol. 2, 247-256.

Kim, M., Canio, W., Kessler, S., and Sinha, N. (2001). Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato. Science 293, 287-289.

Puranam, R. S. and Attardi, G. (2001). The RNase P associated with HeLa cell mitochondria contains an essential RNA component identical in sequence to that of the nuclear RNase P. Mol. Cell Biol. 21, 548-561.

Vance, V. and Vaucheret, H. (2001). RNA silencing in plants—defense and counterdefense. Science 292, 2277-2280.

5.16 mRNA can be specifically localized

rev Chartrand, P., Singer, R. H., and Long, R. M. (2001). RNP localization and transport in yeast. Ann. Rev. Cell Dev. Biol. 17, 297-310.

Kloc, M., Zearfoss, N. R., and Etkin, L. D. (2002). Mechanisms of subcellular mRNA localization. Cell 108, 533-544.

Palacios, I. M. and Johnston, D. S. (2001). Getting the message across: the intracellular localization of mRNAs in higher eukaryotes. Ann. Rev. Cell Dev. Biol. 17, 569-614.

ref Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., and Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. Mol. Cell 2, 437-445.

Jansen, R. P. (2001). mRNA localization: message on the move. Nat. Rev. Mol. Cell Biol. 2, 247-256.

Long, R. M., Singer, R. H., Meng, X., Gonzalez, I., Nasmyth, K., and Jansen, R. P. (1997). Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. Science 277, 383-387.

Chapter 6

Protein synthesis

6.1 Introduction

An mRNA contains a series of codons that interact with the anticodons of aminoacyl-tRNAs so that a corresponding series of amino acids is incorporated into a polypeptide chain. The ribosome provides the environment for controlling the interaction between mRNA and aminoacyl-tRNA. The ribosome behaves like a small migrating factory that travels along the template engaging in rapid cycles of peptide bond synthesis. Aminoacyl-tRNAs shoot in and out of the particle at a fearsome rate, depositing amino acids; and elongation factors cyclically associate with and dissociate from the ribosome. Together with its accessory factors, the ribosome provides the full range of activities required for all the steps of protein synthesis.

Figure 6.1 shows the relative dimensions of the components of the protein synthetic apparatus. The ribosome consists of two subunits that have specific roles in protein synthesis. Messenger RNA is associated with the small subunit; ~30 bases of the mRNA are bound at any time. The mRNA threads its way along the surface close to the junction of the subunits. Two tRNA molecules are active in protein synthesis at any moment; so polypeptide elongation involves reactions taking place at just two of the (roughly) 10 codons covered by the ribosome. The two tRNAs are inserted into internal sites that stretch across the subunits. A third tRNA may remain present on the ribosome after it has been used in protein synthesis, before being recycled.

The basic form of the ribosome has been conserved in evolution, but there are appreciable variations in the overall size and proportions of RNA and protein in the ribosomes of bacteria, eukaryotic cytoplasm, and organelles. **Figure 6.2** compares the components of bacterial and mammalian ribosomes. Both are ribonucleoprotein particles that contain more RNA than protein. The ribosomal proteins are known as *r-proteins*.

Each of the ribosome subunits contains a major rRNA and a number of small proteins. The large subunit may also contain smaller RNA(s). In *E. coli*, the small (30S) subunit consists of the 16S rRNA and 21 r-proteins. The large (50S) subunit contains 23S rRNA, the

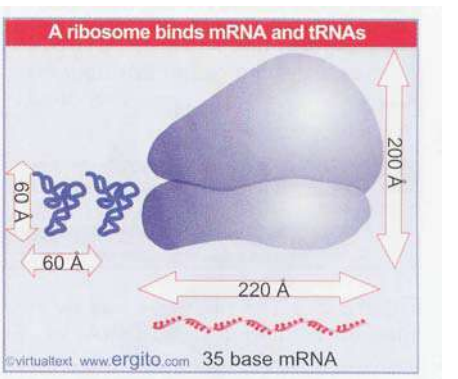


Figure 6.1 Size comparisons show that the ribosome is large enough to bind tRNAs and mRNA.

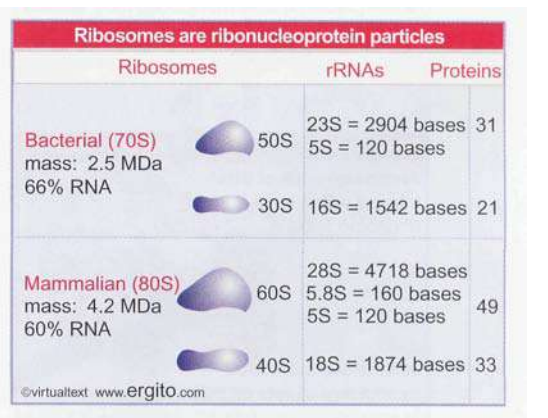


Figure 6.2 Ribosomes are large ribonucleoprotein particles that contain more RNA than protein and dissociate into large and small subunits.

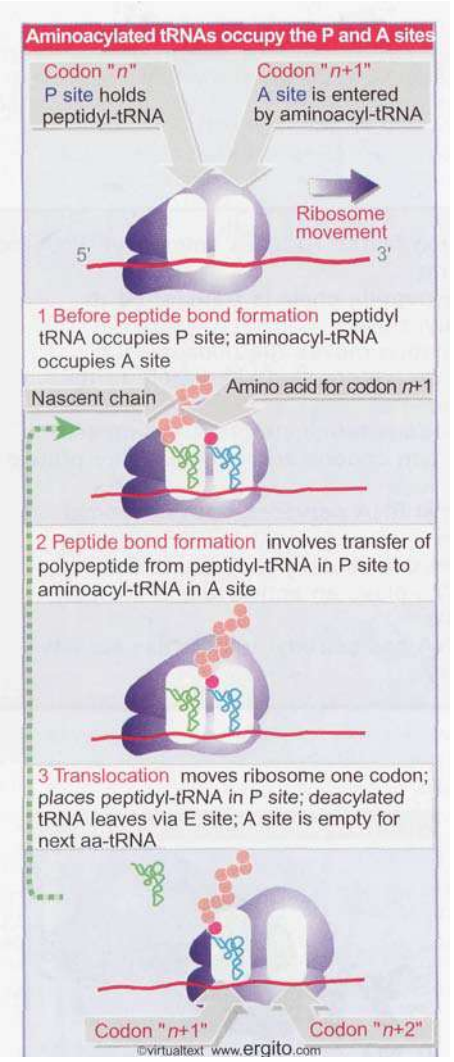


Figure 6.3 The ribosome has two sites for binding charged tRNA.

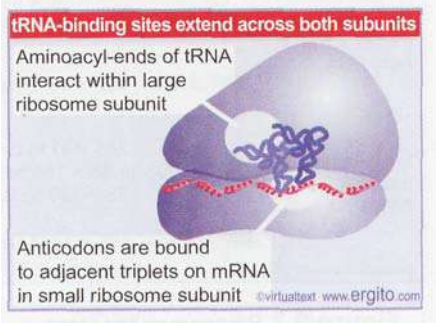


Figure 6.4 The P and A sites position the two interacting tRNAs across both ribosome subunits.

small 5S RNA, and 31 proteins. With the exception of one protein present at four copies per ribosome, there is one copy of each protein. The major RNAs constitute the major part of the mass of the bacterial ribosome. Their presence is pervasive, and probably most or all of the ribosomal proteins actually contact rRNA. So the major rRNAs form what is sometimes thought of as the backbone of each subunit, a continuous thread whose presence dominates the structure, and which determines the positions of the ribosomal proteins.

The ribosomes of higher eukaryotic cytoplasm are larger than those of bacteria. The total content of both RNA and protein is greater; the major RNA molecules are longer (called 18S and 28S rRNAs), and there are more proteins. Probably most or all of the proteins are present in stoichiometric amounts. RNA is still the predominant component by mass.

Organelle ribosomes are distinct from the ribosomes of the cytosol, and take varied forms. In some cases, they are almost the size of bacterial ribosomes and have 70% RNA; in other cases, they are only 60S and have <30% RNA.

The ribosome possesses several active centers, each of which is constructed from a group of proteins associated with a region of ribosomal RNA. The active centers require the direct participation of rRNA in a structural or even catalytic role. Some catalytic functions require individual proteins, but none of the activities can be reproduced by isolated proteins or groups of proteins; they function only in the context of the ribosome.

Two types of information are important in analyzing the ribosome. Mutations implicate particular ribosomal proteins or bases in rRNA in participating in particular reactions. Structural analysis, including direct modification of components of the ribosome and comparisons to identify conserved features in rRNA, identifies the physical locations of components involved in particular functions.

6.2 Protein synthesis occurs by initiation, elongation, and termination

Key Concepts

- The ribosome has 3 tRNA-binding sites.
- An aminoacyl-tRNA enters the A site.
- Peptidyl-tRNA is bound in the P site.
- Deacylated tRNA exits via the E site.
- An amino acid is added to the polypeptide chain by transferring the polypeptide from peptidyl-tRNA in the P site to aminoacyl-tRNA in the A site.

An amino acid is brought to the ribosome by an aminoacyl-tRNA. Its addition to the growing protein chain occurs by an interaction with the tRNA that brought the previous amino acid. Each of these tRNA lies in a distinct site on the ribosome. **Figure 6.3** shows that the two sites have different features:

- An incoming aminoacyl-tRNA binds to the A site. Prior to the entry of aminoacyl-tRNA, the site exposes the codon representing the next amino acid due to be added to the chain.
- The codon representing the most recent amino acid to have been added to the nascent polypeptide chain lies in the P site. This site is

occupied by peptidyl-tRNA, a tRNA carrying the nascent polypeptide chain.

Figure 6.4 shows that the aminoacyl end of the tRNA is located on the large subunit, while the anticodon at the other end interacts with the mRNA bound by the small subunit. So the P and A sites each extend across both ribosomal subunits.

For a ribosome to synthesize a peptide bond, it must be in the state shown in step 1 in Figure 6.3, when peptidyl-tRNA is in the P site and aminoacyl-tRNA is in the A site. Then peptide bond formation occurs when the polypeptide carried by the peptidyl-tRNA is transferred to the amino acid carried by the aminoacyl-tRNA. This reaction is catalyzed by the large subunit of the ribosome.

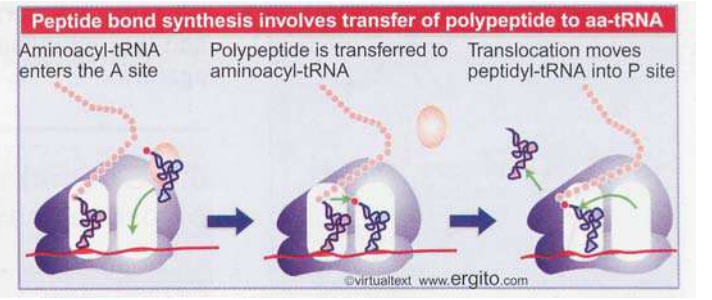


Figure 6.5 Aminoacyl-tRNA enters the A site, receives the polypeptide chain from peptidyl-tRNA, and is transferred into the P site for the next cycle of elongation.

Transfer of the polypeptide generates the ribosome shown in step 2, in which the deacylated tRNA, lacking any amino acid, lies in the P site, while a new peptidyl-tRNA has been created in the A site. This peptidyl-tRNA is one amino acid residue longer than the peptidyl-tRNA that had been in the P site in step 1.

Then the ribosome moves one triplet along the messenger. This stage is called translocation. The movement transfers the deacylated tRNA out of the P site, and moves the peptidyl-tRNA into the P site (see step 3). The next codon to be translated now lies in the A site, ready for a new aminoacyl-tRNA to enter, when the cycle will be repeated. **Figure 6.5** summarizes the interaction between tRNAs and the ribosome.

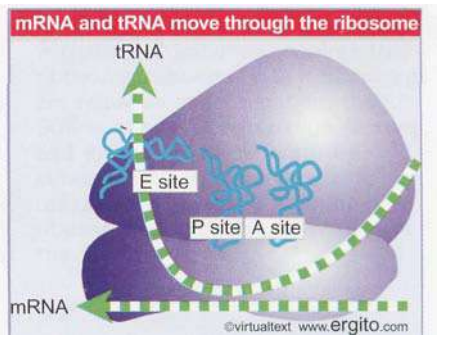


Figure 6.6 tRNA and mRNA move through the ribosome in the same direction.

The deacylated tRNA leaves the ribosome via another tRNA-binding site, the E site. This site is transiently occupied by the tRNA en route between leaving the P site and being released from the ribosome into the cytosol. So the flow of tRNA is into the A site, through the P site, and out through the E site (see also Figure 6.28). **Figure 6.6** compares the movement of tRNA and mRNA, which may be thought of as a sort of ratchet in which the reaction is driven by the codon-anticodon interaction.

Protein synthesis falls into the three stages shown in **Figure 6.7:**

Initiation involves the reactions that precede formation of the peptide bond between the first two amino acids of the protein. It requires the ribosome to bind to the mRNA, forming an initiation complex that contains the first aminoacyl-tRNA. This is a relatively slow step in protein synthesis, and usually determines the rate at which an mRNA is translated.

Elongation includes all the reactions from synthesis of the first peptide bond to addition of the last amino acid. Amino acids are added to the chain one at a time; the addition of an amino acid is the most rapid step in protein synthesis.

Termination encompasses the steps that are needed to release the completed polypeptide chain; at the same time, the ribosome dissociates from the mRNA.

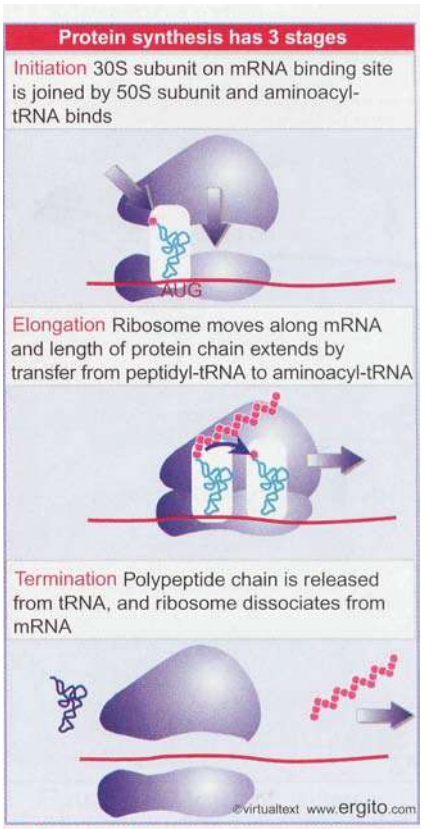


Figure 6.7 Protein synthesis falls into three stages.

Different sets of accessory factors assist the ribosome at each stage. Energy is provided at various stages by the hydrolysis of GTP.

During initiation, the small ribosomal subunit binds to mRNA and then is joined by the 50S subunit. During elongation, the mRNA moves through the ribosome and is translated in triplets. (Although we usually talk about the ribosome moving along mRNA, it is more realistic to think in terms of the mRNA being pulled through the

ribosome.) At termination, the protein is released, mRNA is released, and the individual ribosomal subunits dissociate in order to be used again.

6.3 Special mechanisms control the accuracy of protein synthesis

Key Concepts

- The accuracy of protein synthesis is controlled by specific mechanisms at each stage.

We know that protein synthesis is generally accurate, because of the consistency that is found when we determine the sequence of a protein. There are few detailed measurements of the error rate *in vivo*, but it is generally thought to lie in the range of 1 error for every 10^4 - 10^5 amino acids incorporated. Considering that most proteins are produced in large quantities, this means that the error rate is too low to have any effect on the phenotype of the cell.

It is not immediately obvious how such a low error rate is achieved. In fact, the nature of discriminatory events is a general issue raised by several steps in gene expression. How do synthetases recognize just the corresponding tRNAs and amino acids? How does a ribosome recognize only the tRNA corresponding to the codon in the A site? How do the enzymes that synthesize DNA or RNA recognize only the base complementary to the template? Each case poses a similar problem: how to distinguish one particular member from the entire set, all of which share the same general features.

Probably any member initially can contact the active center by a random-hit process, but then the wrong members are rejected and only the appropriate one is accepted. The appropriate member is always in a minority (1 of 20 amino acids, 1 of ~40 tRNAs, 1 of 4 bases), so the criteria for discrimination must be strict. The point is that the enzyme must have some mechanism for increasing discrimination from the level that would be achieved merely by making contacts with the available surfaces of the substrates.

Figure 6.8 summarizes the error rates at the steps that can affect the accuracy of protein synthesis.

Errors in transcribing mRNA are rare—probably $<10^{-6}$. This is an important stage to control, because a single mRNA molecule is translated into many protein copies. We do not know very much about the mechanisms.

The ribosome can make two types of errors in protein synthesis. It may cause a frameshift by skipping a base when it reads the mRNA (or in the reverse direction by reading a base twice, once as the last base of one codon and then again as the first base of the next codon). These errors are rare, $\sim 10^{-5}$. Or it may allow an incorrect aminoacyl-tRNA to (mis)pair with a codon, so that the wrong amino acid is incorporated. This is probably the most common error in protein synthesis, ~ 5 X 10^{-4}. It is controlled by ribosome structure and velocity (see *7.14 The ribosome influences the accuracy of translation).*

A tRNA synthetase can make two types of error. It can place the wrong amino acid on its tRNA; or it can charge its amino acid with the wrong tRNA. The incorporation of the wrong amino acid is more common, probably because the tRNA offers a larger surface with which the enzyme can make many more contacts to ensure specificity. Aminoacyl-tRNA synthetases have specific mechanisms to correct errors before a mischarged tRNA is released (see *7.10 Synthetases use proofreading to improve accuracy).*

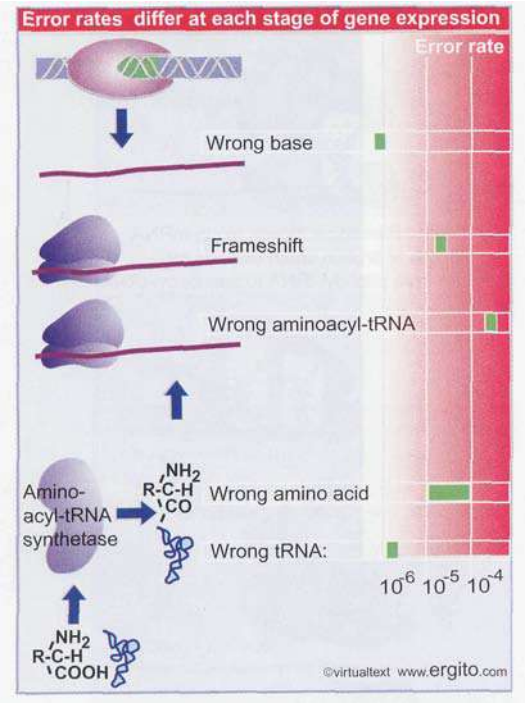


Figure 6.8 Errors occur at rates from 10 [illegible] to 5 X 10^{-4} at different stages of protein synthesis.

6.4 Initiation in bacteria needs 30S subunits and accessory factors

Key Concepts

- Initiation of protein synthesis requires separate 30S and 50S ribosome subunits.
- Initiation factors (IF-1,2,3), which bind to 30S subunits, are also required.
- A 30S subunit carrying initiation factors binds to an initiation site on mRNA to form an initiation complex.
- IF-3 must be released to allow 50S subunits to join the 30S-mRNA complex.

Bacterial ribosomes engaged in elongating a polypeptide chain exist as 70S particles. At termination, they are released from the mRNA as free ribosomes. In growing bacteria, the majority of ribosomes are synthesizing proteins; the free pool is likely to contain ~20% of the ribosomes.

Ribosomes in the free pool can dissociate into separate subunits; so 70S ribosomes are in dynamic equilibrium with 30S and 50S subunits. *Initiation of protein synthesis is not a function of intact ribosomes, but is undertaken by the separate subunits,* which reassociate during the initiation reaction. **Figure 6.9** summarizes the ribosomal subunit cycle during protein synthesis in bacteria.

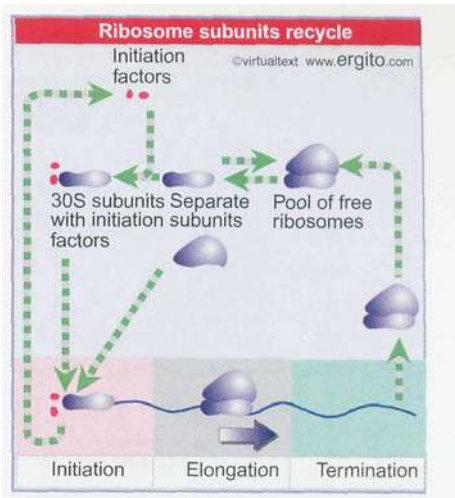


Figure 6.9 Initiation requires free ribosome subunits. When ribosomes are released at termination, the 30S subunits bind initiation factors, and dissociate to generate free subunits. When subunits reassociate to give a functional ribosome at initiation, they release the factors.

Initiation occurs at a special sequence on mRNA called the ribosome-binding site. This is a short sequence of bases that precedes the coding region (see Figure 6.16). The ribosome-binding site is a sequence at which the small and large subunits associate on mRNA to form an intact ribosome. The reaction occurs in two steps:

- Recognition of mRNA occurs when a small subunit binds to form an initiation complex at the ribosome-binding site.
- Then a large subunit joins the complex to generate a complete ribosome.

Although the 30S subunit is involved in initiation, it is not by itself competent to undertake the reactions of binding mRNA and tRNA. It requires additional proteins called initiation factors (IF). These factors are found only on 30S subunits, and they are released when the 30S subunits associate with 50S subunits to generate 70S ribosomes. This behavior distinguishes initiation factors from the structural proteins of the ribosome. The initiation factors are concerned solely with formation of the initiation complex, they are absent from 70S ribosomes, and they play no part in the stages of elongation. **Figure 6.10** summarizes the stages of initiation.

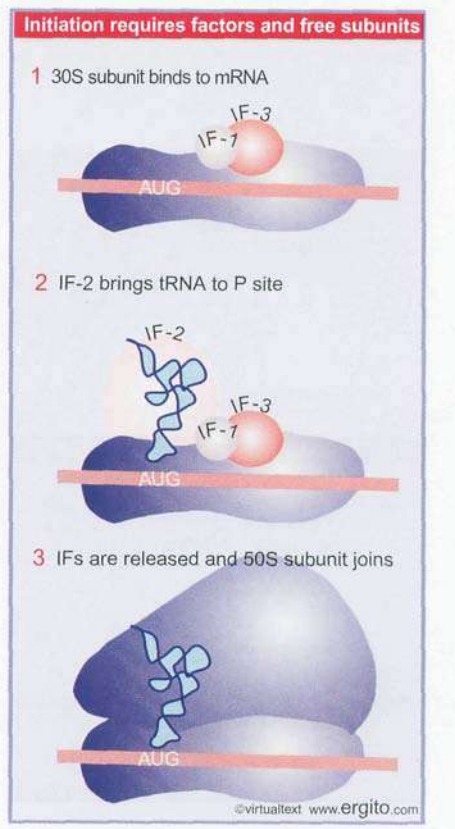


Figure 6.10 Initiation factors stabilize free 30S subunits and bind initiator tRNA to the 30S-mRNA complex.

Bacteria use three initiation factors, numbered IF-1, IF-2, and IF-3. They are needed for both mRNA and tRNA to enter the initiation complex:

- IF-3 is needed for 30S subunits to bind specifically to initiation sites in mRNA.
- IF-2 binds a special initiator tRNA and controls its entry into the ribosome.
- IF-1 binds to 30S subunits only as a part of the complete initiation complex. It binds to the A site and prevents aminoacyl-tRNA from entering. Its location also may impede the 30S subunit from binding to the 50S subunit.

The role of IF-3 is illustrated in **Figure 6.11.** The factor has dual functions: it is needed first to stabilize (free) 30S subunits; and then it

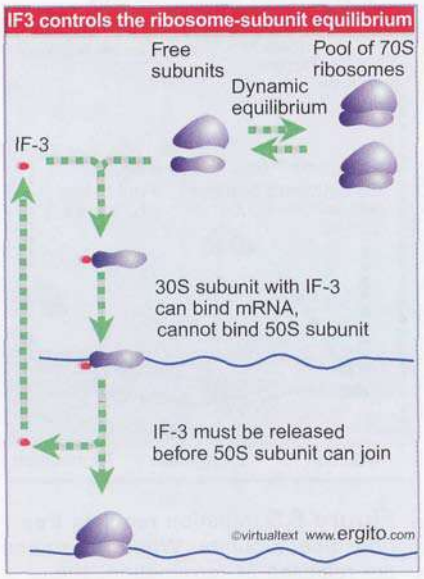


Figure 6.11 Initiation requires 30S subunits that carry IF-3.

enables them to bind to mRNA. IF-3 essentially controls the freedom of 30S subunits, which lasts from their dissociation from the pool of ribosomes to their reassociation with a 50S subunit at initiation.

The first function of IF-3 controls the equilibrium between ribosomal states. IF-3 binds to free 30S subunits that are released from the pool of 70S ribosomes. The presence of IF-3 prevents the 30S subunit from reassociating with a 50S subunit. The reaction between IF-3 and the 30S subunit is stoichiometric: one molecule of IF-3 binds per subunit. There is a relatively small amount of IF-3, so its availability determines the number of free 30S subunits.

IF-3 binds to the surface of the 30S subunit in the vicinity of the A site. There is significant overlap between the bases in 16S rRNA protected by IF-3 and those protected by binding of the 50S subunit, suggesting that it physically prevents junction of the subunits. IF-3 therefore behaves as an anti-association factor that causes a 30S subunit to remain in the pool of free subunits.

The second function of IF-3 controls the ability of 30S subunits to bind to mRNA. Small subunits must have IF-3 in order to form initiation complexes with mRNA. IF-3 must be released from the 30S·mRNA complex in order to enable the 50S subunit to join. On its release, IF-3 immediately recycles by finding another 30S subunit.

IF-2 has a ribosome-dependent GTPase activity: it sponsors the hydrolysis of GTP in the presence of ribosomes, releasing the energy stored in the high-energy bond. The GTP is hydrolyzed when the 50S subunit joins to generate a complete ribosome. The GTP cleavage could be involved in changing the conformation of the ribosome, so that the joined subunits are converted into an active 70S ribosome.

6.5 A special initiator tRNA starts the polypeptide chain

Key Concepts

- Protein synthesis starts with a methionine amino acid usually coded by AUG.
- Different methionine tRNAs are involved in initiation and elongation.
- The initiator tRNA has unique structural features that distinguish it from all other tRNAs.
- The NH_2 group of the methionine bound to bacterial initiator tRNA is formylated.

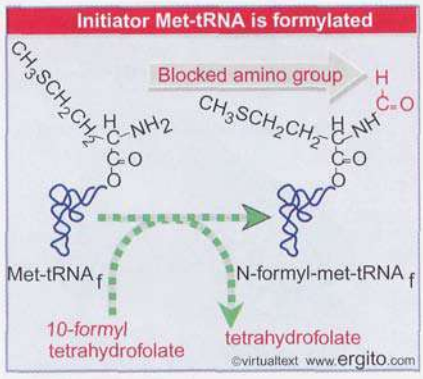


Figure 6.12 The initiator N-formyl-methionyl-tRNA (fMet-tRNA$_f$) is generated by formylation of methionyl-tRNA, using formyl-tetrahydrofolate as cofactor.

Synthesis of all proteins starts with the same amino acid: methionine. The signal for initiating a polypeptide chain is a special initiation codon that marks the start of the reading frame. Usually the initiation codon is the triplet AUG, but in bacteria, GUG or UUG are also used.

The AUG codon represents methionine, and two types of tRNA can carry this amino acid. One is used for initiation, the other for recognizing AUG codons during elongation.

In bacteria and in eukaryotic organelles, the initiator tRNA carries a methionine residue that has been formylated on its amino group, forming a molecule of N-formyl-methionyl-tRNA. The tRNA is known as $tRNA_f^{Met}$. The name of the aminoacyl-tRNA is usually abbreviated to fMet-tRNA$_f$.

The initiator tRNA gains its modified amino acid in a two stage reaction. First, it is charged with the amino acid to generate Met-tRNAf; then the formylation reaction shown in **Figure 6.12** blocks the free NH_2 group. Although the blocked amino acid group would prevent the initia-

tor from participating in chain elongation, it does not interfere with the ability to initiate a protein.

This tRNA is used only for initiation. It recognizes the codons AUG or *GUG (occasionally UUG).* The codons are not recognized equally well: the extent of initiation declines about half when AUG is replaced by GUG, and declines by about half again when UUG is employed.

The species responsible for recognizing AUG codons in internal locations is $tRNA_m^{Met}$. This tRNA responds only to internal AUG codons. Its methionine cannot be formylated.

What features distinguish the fMet-tRNAf initiator and the Met-$tRNA_m$ elongator? Some characteristic features of the tRNA sequence are important, as summarized in **Figure 6.13.** Some of these features are needed to prevent the initiator from being used in elongation, others are necessary for it to function in initiation:

- Formylation is not strictly necessary, because nonformylated Met-tRNAf can function as an initiator, but it improves the efficiency with which the Met-$tRNA_f$ is used, because it is one of the features recognized by the factor IF-2 that binds the initiator tRNA.
- The bases that face one another at the last position of the stem to which the amino acid is connected are paired in all tRNAs except $tRNA_f^{Met}$. Mutations that create a base pair in this position of $tRNA_f^{Met}$ allow it to function in elongation. The absence of this pair is therefore important in preventing $tRNA_f^{Met}$ from being used in elongation. It is also needed for the formylation reaction.
- A series of 3 G·C pairs in the stem that precedes the loop containing the anticodon is unique to $tRNA_f^{Met}$. These base pairs are required to allow the fMet-$tRNA_f$ to be inserted directly into the P site.

In bacteria and mitochondria, the formyl residue on the initiator methionine is removed by a specific deformylase enzyme to generate a normal NH_2 terminus. If methionine is to be the N-terminal amino acid of the protein, this is the only necessary step. In about half the proteins, the methionine at the terminus is removed by an aminopeptidase, creating a new terminus from R_2 (originally the second amino acid incorporated into the chain). When both steps are necessary, they occur sequentially. The removal reaction(s) occur rather rapidly, probably when the nascent polypeptide chain has reached a length of 15 amino acids.

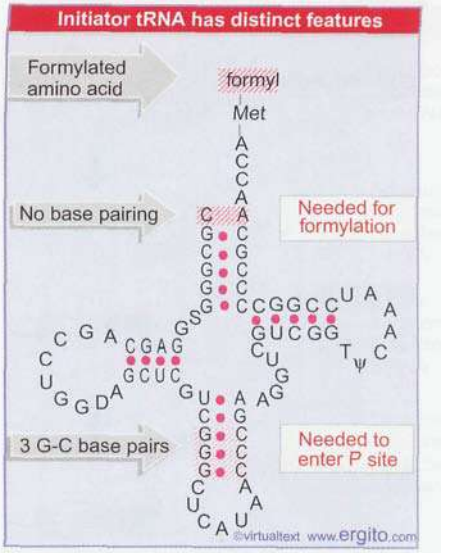


Figure 6.13 fMet-$tRNA_f$ has unique features that distinguish it as the initiator tRNA.

6.6 Use of fMet-$tRNA_f$ is controlled by IF-2 and the ribosome

Key Concepts

IF-2 binds the initiator fMet-tRNAf and allows it to enter the partial P site on the 30S subunit.

The meaning of the AUG and GUG codons depends on their context. When the AUG codon is used for initiation, it is read as formyl-methionine; when used within the coding region, it represents methionine. The meaning of the GUG codon is even more dependent on its location. When present as the first codon, it is read via the initiation reaction as formyl-methionine. Yet when present within a gene, it is read by Val-tRNA, one of the regular members of the tRNA set, to provide valine as required by the genetic code.

How is the context of AUG and GUG codons interpreted? **Figure 6.14** illustrates the decisive role of the ribosome, in conjunction with accessory factors.

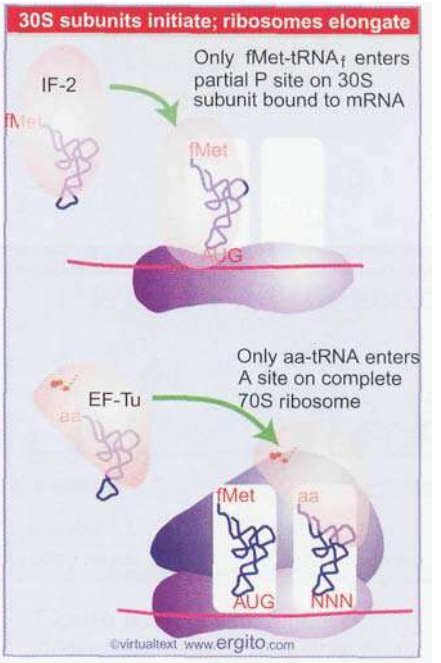


Figure 6.14 Only fMet-$tRNA_f$ can be used for initiation by 30S subunits; only other aminoacyl-tRNAs (αα-tRNA) can be used for elongation by 70S ribosomes.

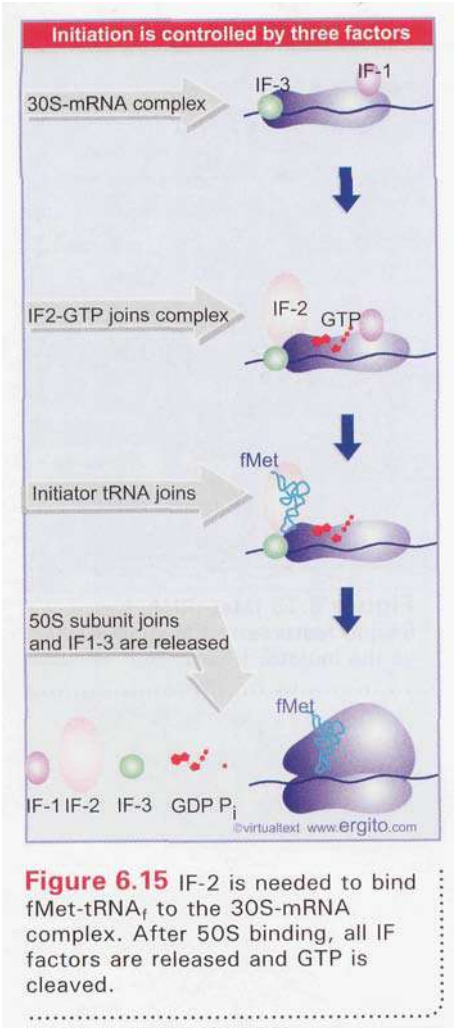


Figure 6.15 IF-2 is needed to bind fMet-tRNA$_f$ to the 30S-mRNA complex. After 50S binding, all IF factors are released and GTP is cleaved.

In an initiation complex, the small subunit alone is bound to mRNA. The initiation codon lies within the part of the P site carried by the small subunit. The only aminoacyl-tRNA that can become part of the initiation complex is the initiator, which has the unique property of being able to enter directly into the partial P site to recognize its codon.

When the large subunit joins the complex, the partial tRNA-binding sites are converted into the intact P and A sites. The initiator fMet-tRNA$_f$ occupies the P site, and the A site is available for entry of the aminoacyl-tRNA complementary to the second codon of the gene. The first peptide bond forms between the initiator and the next aminoacyl-tRNA.

Initiation prevails when an AUG (or GUG) codon lies within a ribosome-binding site, because only the initiator tRNA can enter the partial P site generated when the 30S subunit binds *de novo* to the mRNA. Internal reading prevails subsequently, when the codons are encountered by a ribosome that is continuing to translate an mRNA, because only the regular aminoacyl-tRNAs can enter the (complete) A site.

Accessory factors are critical in controlling the usage of aminoacyl-tRNAs. All aminoacyl-tRNAs associate with the ribosome by binding to an accessory factor. The factor used in initiation is IF-2 (see *6.4 Initiation in bacteria needs 30S subunits and accessory factors)*, and the corresponding factor used at elongation is EF-Tu (see *6.10 Elongation factor Tu loads aminoacyl-tRNA into the A site).*

The initiation factor IF-2 places the initiator tRNA into the P site. By forming a complex specifically with fMet-tRNA$_f$, IF-2 ensures that only the initiator tRNA, and none of the regular aminoacyl-tRNAs, participates in the initiation reaction. Conversely, EF-Tu, which places aminoacyl-tRNAs in the A site cannot bind fMet-tRNA$_f$, which is therefore excluded from use during elongation.

Figure 6.15 details the series of events by which IF-2 places the fMet-tRNA$_f$ initiator in the P site. IF-2, bound to GTP, associates with the P site of the 30S subunit. At this point, the 30S subunit carries all the initiation factors. fMet-tRNA$_f$ then binds to the IF-2 on the 30S subunit. IF-2 then transfers the tRNA into the partial P site.

6.7 Initiation involves base pairing between mRNA and rRNA

Key Concepts

- An initiation site on bacterial mRNA consists of the AUG initiation codon preceded with a gap of ~10 bases by the Shine-Dalgarno polypurine hexamer.
- The rRNA of the 30S bacterial ribosomal subunit has a complementary sequence that base pairs with the Shine-Dalgarno sequence during initiation.

An mRNA contains many AUG triplets: how is the initiation codon recognized as providing the starting point for translation? The sites on mRNA where protein synthesis is initiated can be identified by binding the ribosome to mRNA under conditions that block elongation. Then the ribosome remains at the initiation site. When ribonuclease is added to the blocked initiation complex, all the regions of mRNA outside the ribosome are degraded, but those actually bound to it are protected, as illustrated in **Figure 6.16.** The protected fragments can be recovered and characterized.

The initiation sequences protected by bacterial ribosomes are ~30 bases long. The ribosome-binding sites of different bacterial mRNAs display two common features:

- The AUG (or less often, GUG or UUG) initiation codon is always included within the protected sequence.
- Within 10 bases upstream of the AUG is a sequence that corresponds to part or all of the hexamer.

5′... A G G A G G... 3′

This polypurine stretch is known as the Shine-Dalgarno sequence. It is complementary to a highly conserved sequence close to the 3' end of 16S rRNA. (The extent of complementarity differs with individual mRNAs, and may extend from a 4-base core sequence GAGG to a 9-base sequence extending beyond each end of the hexamer.) Written in reverse direction, the rRNA sequence is the hexamer:

3′... U C C U C C... 5′

Does the Shine-Dalgarno sequence pair with its complement in rRNA during mRNA-ribosome binding? Mutations of both partners in this reaction demonstrate its importance in initiation. Point mutations in the Shine-Dalgarno sequence can prevent an mRNA from being translated. And the introduction of mutations into the complementary sequence in rRNA is deleterious to the cell and changes the pattern of protein synthesis. The decisive confirmation of the base pairing reaction is that a mutation in the Shine-Dalgarno sequence of an mRNA can be suppressed by a mutation in the rRNA that restores base pairing.

The sequence at the 3′ end of rRNA is conserved between prokaryotes and eukaryotes except that in all eukaryotes there is a deletion of the five-base sequence CCUCC that is the principal complement to the Shine-Dalgarno sequence. There does not appear to be base pairing between eukaryotic mRNA and 18S rRNA. This is a significant difference in the mechanism of initiation.

In bacteria, a 30S subunit binds directly to a ribosome-binding site. As a result, the initiation complex forms at a sequence surrounding the AUG initiation codon. When the mRNA is polycistronic, each coding region starts with a ribosome-binding site.

The nature of bacterial gene expression means that translation of a bacterial mRNA proceeds sequentially through its cistrons. At the time when ribosomes attach to the first coding region, the subsequent coding regions have not yet even been transcribed. By the time the second ribosome site is available, translation is well under way through the first cistron.

What happens between the coding regions depends on the individual mRNA. Probably in most cases the ribosomes bind independently at the beginning of each cistron. The most common series of events is illustrated in Figure 6.17. When synthesis of the first protein terminates, the ribosomes leave the mRNA and dissociate into subunits. Then a new ribosome must assemble at the next coding region, and set out to translate the next cistron.

In some bacterial mRNAs, translation between adjacent cistrons is directly linked, because ribosomes gain access to the initiation codon of the second cistron as they complete translation of the first cistron. This effect requires the space between the two coding regions to be small. It may depend on the high local density of ribosomes; or the juxtaposition of termination and initiation sites could allow some of the usual intercistronic events to be bypassed.

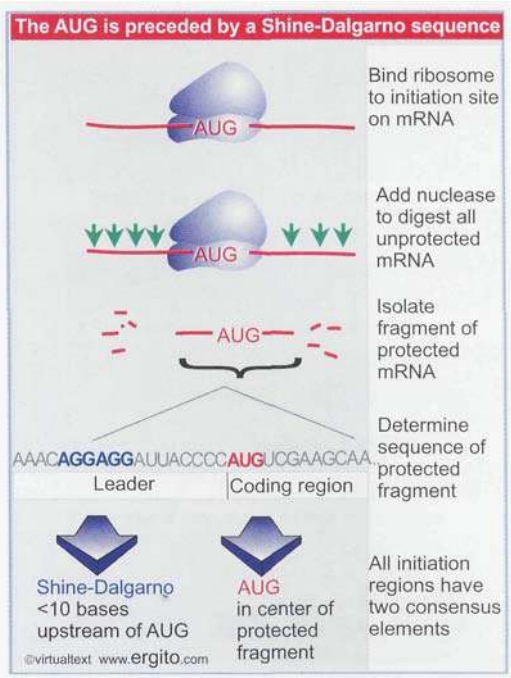


Figure 6.16 Ribosome-binding sites on mRNA can be recovered from initiation complexes. They include the upstream Shine-Dalgarno sequence and the initiation codon.

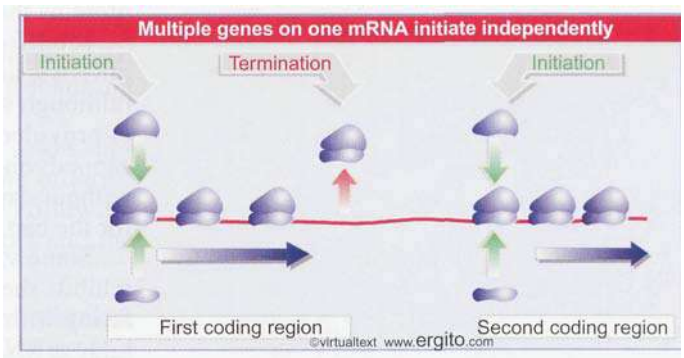


Figure 6.17 Initiation occurs independently at each cistron in a polycistronic mRNA. When the intercistronic region is longer than the span of the ribosome, dissociation at the termination site is followed by independent reinitiation at the next cistron.

A ribosome physically spans ~30 bases of mRNA, so that it could simultaneously contact a termination codon and the next initiation site if they are separated by only a few bases.

6.8 Small subunits scan for initiation sites on eukaryotic mRNA

Key Concepts

- Eukaryotic 40S ribosomal subunits bind to the 5' end of mRNA and scan the mRNA until they reach an initiation site.
- A eukaryotic initiation site consists of a 10 nucleotide sequence that includes an AUG codon.
- 60S ribosomal subunits join the complex at the initiation site.

Initiation of protein synthesis in eukaryotic cytoplasm resembles the process in bacteria, but the order of events is different, and the number of accessory factors is greater. Some of the differences in initiation are related to a difference in the way that bacterial 30S and eukaryotic 40S subunits find their binding sites for initiating protein synthesis on mRNA. In eukaryotes, small subunits first recognize the 5' end of the mRNA, and then move to the initiation site, where they are joined by large subunits. (In prokaryotes, small subunits bind directly to the initiation site.)

Virtually all eukaryotic mRNAs are monocistronic, but each mRNA usually is substantially longer than necessary just to code for its protein. The average mRNA in eukaryotic cytoplasm is 1000-2000 bases long, has a methylated cap at the 5' terminus, and carries 100-200 bases of poly(A) at the 3' terminus.

The nontranslated 5' leader is relatively short, usually <100 bases. The length of the coding region is determined by the size of the protein. The nontranslated 3' trailer is often rather long, sometimes ~1000 bases.

The first feature to be recognized during translation of a eukaryotic mRNA is the methylated cap that marks the 5' end. Messengers whose caps have been removed are not translated efficiently *in vitro*. Binding of 40S subunits to mRNA requires several initiation factors, including proteins that recognize the structure of the cap.

Modification at the 5' end occurs to almost all cellular or viral mRNAs, and is essential for their translation in eukaryotic cytoplasm (although it is not needed in organelles). The sole exception to this rule is provided by a few viral mRNAs (such as poliovirus) that are not capped; only these exceptional viral mRNAs can be translated *in vitro* without caps. They use an alternative pathway that bypasses the need for the cap.

Some viruses take advantage of this difference. Poliovirus infection inhibits the translation of host mRNAs. This is accomplished by interfering with the cap binding proteins that are needed for initiation of cellular mRNAs, but that are superfluous for the noncapped poliovirus mRNA.

We have dealt with the process of initiation as though the ribosome-binding site is always freely available. However, its availability may be impeded by secondary structure. The recognition of mRNA requires several additional factors; an important part of their function is to remove any secondary structure in the mRNA (see Figure 6.20).

Sometimes the AUG initiation codon lies within 40 bases of the 5' terminus of the mRNA, so that both the cap and AUG lie within the

span of ribosome binding. But in many mRNAs the cap and AUG are farther apart, in extreme cases ~1000 bases distant. Yet the presence of the cap still is necessary for a stable complex to be formed at the initiation codon. How can the ribosome rely on two sites so far apart?

Figure 6.18 illustrates the "scanning" model, which supposes that the 40S subunit initially recognizes the 5' cap and then "migrates" along the mRNA. Scanning from the 5' end is a linear process. When 40S subunits scan the leader region, they can melt secondary structure hairpins with stabilities <–30 kcal, but hairpins of greater stability impede or prevent migration.

Migration stops when the 40S subunit encounters the AUG initiation codon. Usually, although not always, the first AUG triplet sequence to be encountered will be the initiation codon. However, the AUG triplet by itself is not sufficient to halt migration; it is recognized efficiently as an initiation codon only when it is in the right context. The most important determinants of context are the bases in positions —4 and +1. An initiation codon may be recognized in the sequence NNNPuNN*AUGG*. The purine (A or G) 3 bases before the AUG codon, and the G immediately following it, can influence the efficiency of translation by 10×. When the leader sequence is long, further 40S subunits can recognize the 5' end before the first has left the initiation site, creating a queue of subunits proceeding along the leader to the initiation site.

It is probably true that the initiation codon is the first AUG to be encountered in the most efficiently translated mRNAs. But what happens when there is an AUG triplet in the 5′ nontranslated region? There are two possible escape mechanisms for a ribosome that starts scanning at the 5' end. The most common is that scanning is leaky, that is, a ribosome may continue past a non-initiation AUG because it is not in the right context. In the rare case that it does recognize the AUG, it may initiate translation but terminate before the proper initiation codon, after which it resumes scanning.

The vast majority of eukaryotic initiation events involve scanning from the 5' cap, but there is an alternative means of initiation, used especially by certain viral RNAs, in which a 40S subunit associates directly with an internal site called an IRES. (This entirely bypasses any AUG codons that may be in the 5' nontranslated region.) There are few sequence homologies between known IRES elements. We can distinguish three types on the basis of their interaction with the 40S subunit:

- One type of IRES includes the AUG initiation codon at its upstream boundary. The 40S subunit binds directly to it, using a subset of the same factors that are required for initiation at 5' ends.
- Another is located as much as *100 nucleotides upstream of the* AUG, requiring a 40S subunit to migrate, again probably by a scanning mechanism.
- An exceptional type of IRES in hepatitis C virus can bind a 40S subunit directly, without requiring any initiation factors. The order of events is different from all other eukaryotic initiation. Following 40S-mRNA binding, a complex containing initiator factors and the initiator tRNA binds.

Use of the IRES is especially important in picornavirus infection, where it was first discovered, because the virus inhibits host protein synthesis by destroying cap structures and inhibiting the initiation factors that bind them (see next section).

Binding is stabilized at the initiation site. When the 40S subunit is joined by a 60S subunit, the intact ribosome is located at the site identified by the protection assay. A 40S subunit protects a region of up to 60 bases; when the 60S subunits join the complex, the protected region contracts to about the same length of 30-40 bases seen in prokaryotes.

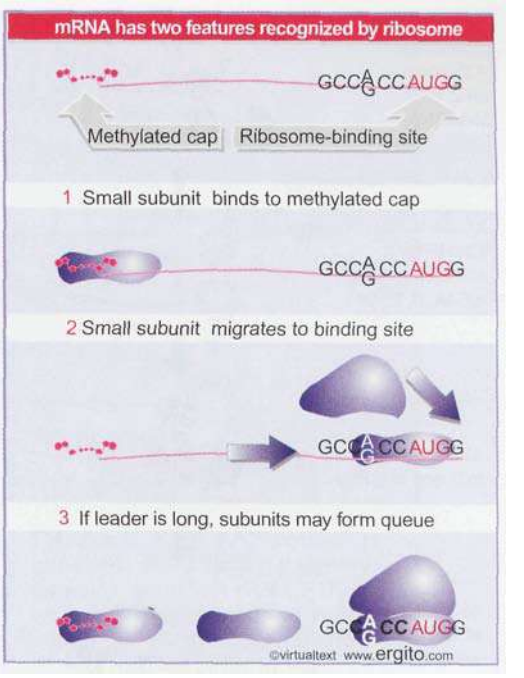


Figure 6.18 Eukaryotic ribosomes migrate from the 5' end of mRNA to the ribosome binding site, which includes an AUG initiation codon.

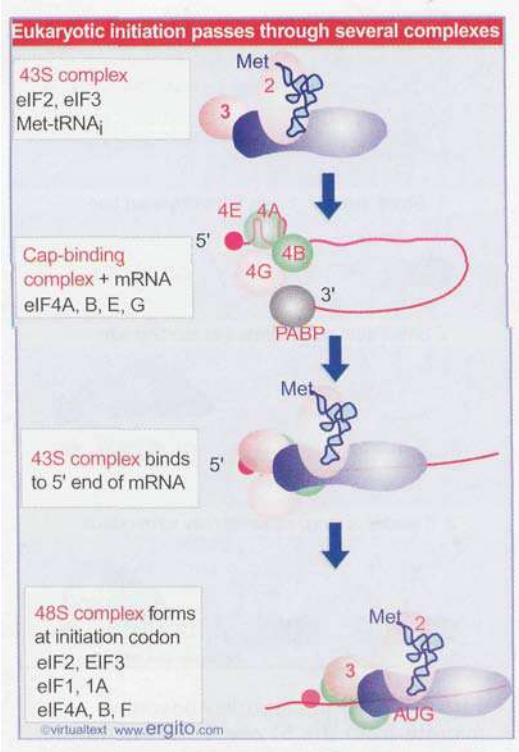


Figure 6.19 Some initiation factors bind to the 40S ribosome subunit to form the 43S complex; others bind to mRNA. When the 43S complex binds to mRNA, it scans for the initiation codon and can be isolated as the 48S complex.

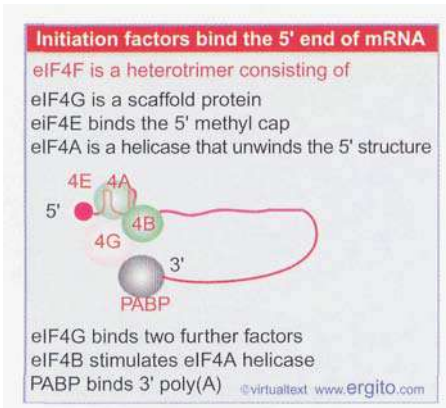


Figure 6.20 The heterotrimer eIF4F binds the 5′ end of mRNA and also binds further factors.

6.9 Eukaryotes use a complex of many initiation factors

Key Concepts

- Initiation factors are required for all stages of initiation, including binding the initiator tRNA, 40S subunit attachment to mRNA, movement along the mRNA, and joining of the 60S subunit.
- Eukaryotic initiator tRNA is a Met-tRNA that is different from the Met-tRNA used Vn elongation, but the methionine is not formylated.
- eIF2 binds the initiator Met-tRNA$_i$ and GTP, and the complex binds to the 40S subunit before it associates with mRNA.

Initiation in eukaryotes has the same general features as in bacteria in using a specific initiation codon and initiator tRNA. Initiation in eukaryotic cytoplasm uses AUG as the initiator. The initiator tRNA is a distinct species, but its methionine does not become formylated. It is called tRNA$_i^{Met}$. So the difference between the initiating and elongating Met-tRNAs lies solely in the tRNA moiety, with Met-tRNA$_i$ used for initiation and Met-tRNA$_m$ used for elongation.

At least two features are unique to the initiator tRNA$_i^{Met}$ in yeast; it has an unusual tertiary structure; and it is modified by phosphorylation of the 2' ribose position on base 64 (if this modification is prevented, the initiator can be used in elongation). So the principle of a distinction between initiator and elongator Met-tRNAs is maintained in eukaryotes, but its structural basis is different from that in bacteria (for comparison see Figure 6.13).

Eukaryotic cells have more initiation factors than bacteria—the current list includes 12 factors that are directly or indirectly required for initiation. The factors are named similarly to those in bacteria, sometimes by analogy with the bacterial factors, and are given the prefix "e" to indicate their eukaryotic origin. They act at all stages of the process, including

- forming an initiation complex with the 5' end of mRNA
- forming a complex with Met-tRNA$_i$
- binding the mRNA-factor complex to the Met-tRNA$_i$-factor complex
- enabling the ribosome to scan mRNA from the 5′ end to the first AUG
- detecting binding of initiator tRNA to AUG at the start site
- mediating joining of the 60S subunit.

Figure 6.19 summarizes the stages of initiation, and shows which initiation factors are involved at each stage. Factors eIF2 and eIF3 bind to the 40S ribosome subunit. Factors eIF4A, eIF4B, eIF4F bind to the mRNA. Factors eIF1 and eIF1A bind to the ribosome subunit-mRNA complex.

Figure 6.20 shows the group of factors that bind to the 5' end of mRNA. The factor eIF4F is a protein complex that contains three of the initiation factors. It is not clear whether it preassembles as a complex before binding to mRNA or whether the individual subunits are added individually to form the complex on mRNA. It includes the cap-binding subunit eIF4E, the helicase eIF4A, and the "scaffolding" subunit eIF4G. After eIF4E binds the cap, eIF4A unwinds any secondary structure that exists in the first 15 bases of the mRNA. Energy for the unwinding is provided by hydrolysis of ATR Unwinding of structure farther along the mRNA is accomplished by eIF4A together with another factor, eIF4B. The main role of eIF4G is to link other components of the initiation complex.

Factor eIF4E is a focus for regulation. Its activity is increased by phosphorylation, which is triggered by stimuli that increase protein synthesis, and reversed by stimuli that repress protein synthesis. Factor eIF4F has a kinase activity that phosphorylates eIF4E. The availability of eIF4E is also controlled by proteins that bind to it (called 4E-BP1,2,3), to prevent it from functioning in initiation. eIF4G is also a target for degradation during picornavirus infection, as part of the destruction of the capacity to initiate at 5' cap structures (see previous section).

The presence of poly(A) on the 3' tail of an mRNA stimulates the formation of an initiation complex at the 5' end. The poly(A)-binding protein (Pab1p in yeast) is required for this effect. Pab1p binds to the eIF4G scaffolding protein. This implies that the mRNA will have a circular organization so long as eIFG is bound, with both the 5' and 3' ends held in this complex (see Figure 6.20). The significance of the formation of this closed loop is not clear, although it could have several effects, such as:

- stimulating initiation of translation;
- promoting reinitiation of ribosomes, so that when they terminate at the 3' end, the released subunits are already in the vicinity of the 5' end;
- stabilizing the mRNA against degradation;
- allowing factors that bind to the 3' end to regulate the initiation of translation.

Factor eIF2 is the key factor in binding Met-tRNA$_i$. It is a typical monomeric GTP-binding protein that is active when bound to GTP, and inactive when bound to GDP. **Figure 6.21** shows that the eIF2-GTP binds to Met-tRNA$_i$. The product is sometimes called the ternary complex (after its three components, eIF2, GTP, Met-tRNAj).

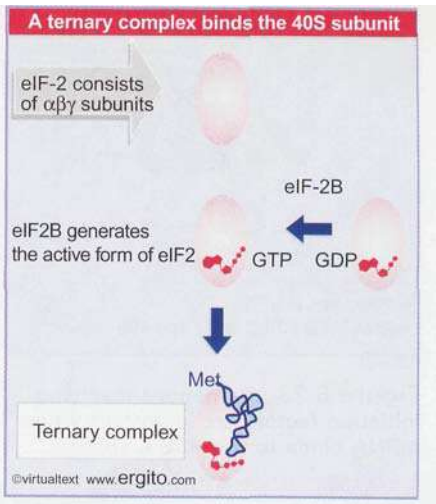


Figure 6.21 In eukaryotic initiation, eIF-2 forms a ternary complex with Met-tRNA$_i$. The ternary complex binds to free 40S subunits, which attach to the 5′ end of mRNA. Later in the reaction, GTP is hydrolyzed when eIF-2 is released in the form of eIF2-GDP. Factor eIF-2B regenerates the active form.

Figure 6.22 shows that the ternary complex places Met-tRNAj onto the 40S subunit. This generates the 43S initiation complex. The reaction is independent of the presence of mRNA. In fact, the Met-tRNA; initiator must be present in order for the 40S subunit to bind to mRNA. One of the factors in this complex is eIF3, which is required to maintain 40S subunits in their dissociated state. eIF3 is a very large factor, with 8-10 subunits.

The next step is for the 43S complex to bind to the 5' end of the mRNA. **Figure 6.23** shows that the interactions involved at this stage are not completely defined, but probably involve eIF4G and eIF3 as well as the mRNA and 40S subunit. Factor eIF4G binds to eIF3. This provides the means by which the 40S ribosomal subunit binds to eIF4F, and thus is recruited to the complex. In effect, eIF4F functions to get eIF4G in place so that it can attract the small ribosomal subunit.

When the small subunit has bound mRNA, it migrates to (usually) the first AUG codon. This requires expenditure of energy in the form of ATP. It is assisted by the factors eIF1 and eIF1A. **Figure 6.24** shows that the small subunit stops when it reaches the initiation site, forming a 48S complex.

Junction of the 60S subunits with the initiation complex cannot occur until eIF2 and eIF3 have been released from the initiation complex. This is mediated by eIF5, and causes eIF2 to hydrolyze its GTP. The reaction occurs on the small ribosome subunit, and requires the initiator tRNA to be base paired with the AUG initiation codon. Probably all of the remaining factors are released when the complete 80S ribosome is formed.

Finally the factor eIF5B enables the 60S subunit to join the complex, forming an intact ribosome that is ready to start elongation. eIF5B has a similar sequence to the prokaryotic factor IF2, which has a similar role in hydrolyzing GTP (in addition to its role in binding the initiator tRNA).

Once the factors have been released, they can associate with the initiator tRNA and ribosomal subunits in another initiation cycle. Because eIF2 has hydrolyzed its GTP, the active form must be regenerated. This

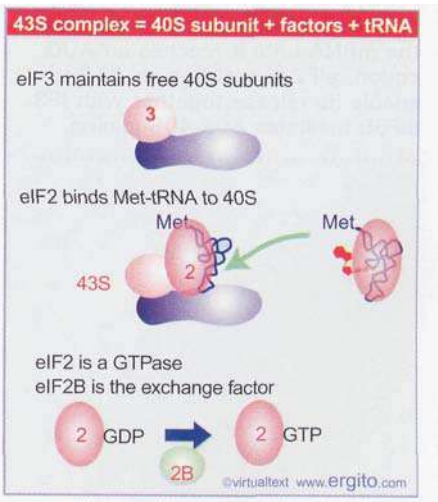


Figure 6.22 Initiation factors bind the initiator Met-tRNA to the 40S subunit to form a 43S complex. Later in the reaction, GTP is hydrolyzed when eIF-2 is released in the form of eIF2-GDP. Factor eIF-2B regenerates the active form.

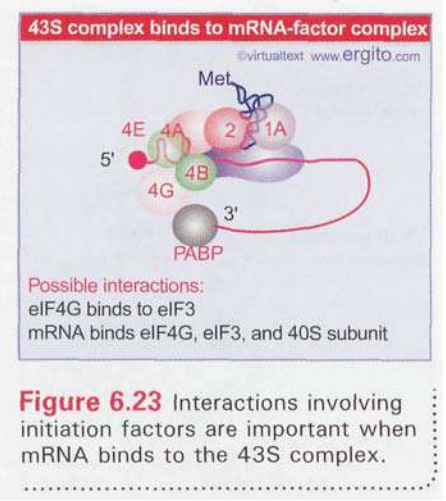


Figure 6.23 Interactions involving initiation factors are important when mRNA binds to the 43S complex.

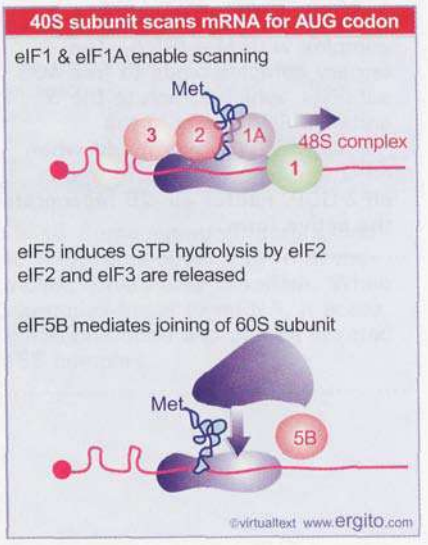


Figure 6.24 eIF1 and eIF1A help the 43S initiation complex to scan the mRNA until it reaches an AUG codon. eIF2 hydrolyzes its GTP to enable its release together with IF3. eIF5B mediates 60S-40S joining.

is accomplished by another factor, eIF2B, which displaces the GDP so that it can be replaced by GTP.

Factor eIF2 is a target for regulation. Several regulatory kinases act on the a subunit of eIF2. Phosphorylation prevents eIF2B from regenerating the active form. This limits the action of eIF2B to one cycle of initiation, and thereby inhibits protein synthesis.

6.10 Elongation factor Tu loads aminoacyl-tRNA into the A site

Key Concepts

- EF-Tu is a monomeric G protein whose active form (bound to GTP) binds aminoacyl-tRNA.
- The EF-Tu·GTP·aminoacyl-tRNA complex binds to the ribosome A site.

Once the complete ribosome is formed at the initiation codon, the stage is set for a cycle in which aminoacyl-tRNA enters the A site of a ribosome whose P site is occupied by peptidyl-tRNA. Any aminoacyl-tRNA except the initiator can enter the A site. Its entry is mediated by an elongation factor (*EF-Tu* in bacteria). The process is similar in eukaryotes. EF-Tu is a highly conserved protein throughout bacteria and mitochondria, and is homologous to its eukaryotic counterpart.

Just like its counterpart in initiation (IF-2), EF-Tu is associated with the ribosome only during the process of aminoacyl-tRNA entry. Once the aminoacyl-tRNA is in place, EF-Tu leaves the ribosome, to work again with another aminoacyl-tRNA. So it displays the cyclic association with, and dissociation from, the ribosome that is the hallmark of the accessory factors.

The pathway for aminoacyl-tRNA entry to the A site is illustrated in **Figure 6.25.** EF-Tu carries a guanine nucleotide. The factor is a monomeric G protein whose activity is controlled by the state of the guanine nucleotide:

- When GTP is present, the factor is in its active state.
- When the GTP is hydrolyzed to GDP, the factor becomes inactive.
- Activity is restored when the GDP is replaced by GTP.

The binary complex of EF-Tu·GTP binds aminoacyl-tRNA to form a ternary complex of aminoacyl-tRNA·EF-Tu·GTP. The ternary complex binds only to the A site of ribosomes whose P site is already occupied by peptidyl-tRNA. This is the critical reaction in ensuring that the aminoacyl-tRNA and peptidyl-tRNA are correctly positioned for peptide bond formation.

Aminoacyl-tRNA is loaded into the A site in two stages. First the anticodon end binds to the A site of the 30S subunit. Then codon-anticodon recognition triggers a change in the conformation of the ribosome. This stabilizes tRNA binding and causes EF-Tu to hydrolyze its GTP. The CCA end of the tRNA now moves into the A site on the 50S subunit. The binary complex EF-Tu·GDP is released. This form of EF-Tu is inactive and does not bind aminoacyl-tRNA effectively.

Another factor, EF-Ts, mediates the regeneration of the used form, EF-Tu·GDP, into the active form, EF-Tu-GTP. First, EF-Ts displaces the GDP from EF-Tu, forming the combined factor EF-Tu·EF-Ts. Then the EF-Ts is in turn displaced by GTP, reforming EF-Tu-GTP. The active binary complex binds aminoacyl-tRNA; and the released EF-Ts can recycle.

There are ~70,000 molecules of EF-Tu per bacterium (~5% of the total bacterial protein), which approaches the number of aminoacyl-tRNA n

ecules. This implies that most aminoacyl-tRNAs are likely to be present in ternary complexes. There are only ~10,000 molecules of EF-Ts per cell (about the same as the number of ribosomes). The kinetics of the interaction between EF-Tu and EF-Ts suggest that the EF-Tu·EF-Ts complex exists only transiently, so that the EF-Tu is very rapidly converted to the GTP-bound form, and then to a ternary complex.

The role of GTP in the ternary complex has been studied by substituting an analog that cannot be hydrolyzed. The compound GMP-PCP has a methylene bridge in place of the oxygen that links the β and γ phosphates in GTP. In the presence of GMP-PCP, a ternary complex can be formed that binds aminoacyl-tRNA to the ribosome. But the peptide bond cannot be formed. So the presence of GTP is needed for aminoacyl-tRNA to be bound at the A site; but the hydrolysis is not required until later.

Kirromycin is an antibiotic that inhibits the function of EF-Tu. When EF-Tu is bound by kirromycin, it remains able to bind aminoacyl-tRNA to the A site. But the EF-Tu·GDP complex cannot be released from the ribosome. Its continued presence prevents formation of the peptide bond between the peptidyl-tRNA and the aminoacyl-tRNA. As a result, the ribosome becomes "stalled" on mRNA, bringing protein synthesis to a halt.

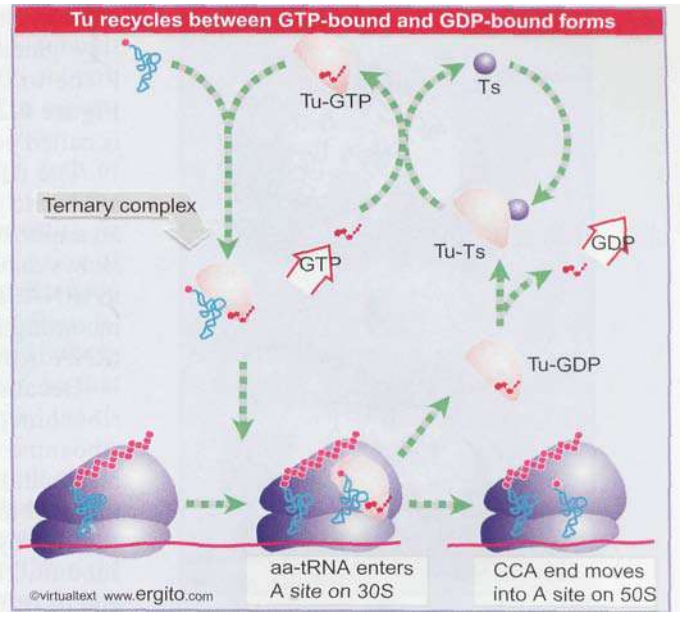


Figure 6.25 EF-Tu-GTP places aminoacyl-tRNA on the ribosome and then is released as EF-Tu-GDP. EF-Ts is required to mediate the replacement of GDP by GTP. The reaction consumes GTP and releases GDP. The only aminoacyl-tRNA that cannot be recognized by EF-Tu-GTP is $fMet\text{-}tRNA_f$, whose failure to bind prevents it from responding to internal AUG or GUG codons.

This effect of kirromycin demonstrates that inhibiting one step in protein synthesis blocks the next step. The reason is that the continued presence of EF-Tu prevents the aminoacyl end of aminoacyl-tRNA from entering the A site on the 50S subunit (see Figure 6.30). So the release of EF-Tu-GDP is needed for the ribosome to undertake peptide bond formation. The same principle is seen at other stages of protein synthesis: one reaction must be completed properly before the next can occur.

The interaction with EF-Tu also plays a role in quality control. Aminoacyl-tRNAs are brought into the A site without knowing whether their anticodons will fit the codon. The hydrolysis of EF-Tu·GTP is relatively slow: because it takes longer than the time required for an incorrect aminoacyl-tRNA to dissociate from the A site, most incorrect species are removed at this stage. The release of EF-Tu-GDP after hydrolysis also is slow, so any surviving incorrect aminoacyl-tRNAs may dissociate at this stage. The basic principle is that the reactions involving EF-Tu occur slowly enough to allow incorrect aminoacyl-tRNAs to dissociate before they become trapped in protein synthesis.

In eukaryotes, the factor *eEF1α* is responsible for bringing aminoacyl-tRNA to the ribosome, again in a reaction that involves cleavage of a high-energy bond in GTP. Like its prokaryotic homologue (EF-Tu), it is an abundant protein. After hydrolysis of GTP, the active form is regenerated by the factor eEF1βγ, a counterpart to EF-Ts.

6.11 The polypeptide chain is transferred to aminoacyl-tRNA

Key Concepts

- The 50S subunit has peptidyl transferase activity.
- The nascent polypeptide chain is transferred from peptidyl-tRNA in the P site to aminoacyl-tRNA in the A site.
- Peptide bond synthesis generates deacylated tRNA in the P site and peptidyl-tRNA in the A site.

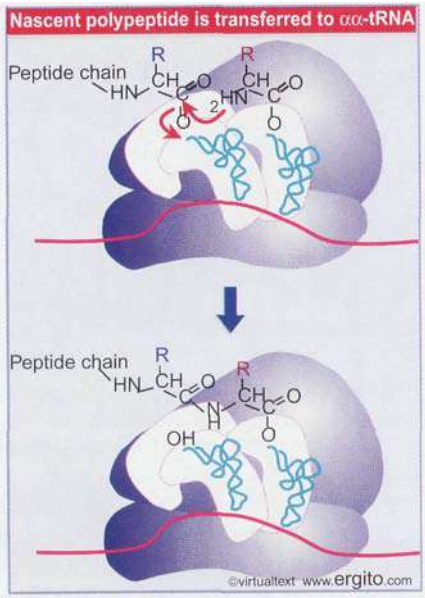


Figure 6.26 Peptide bond formation takes place by reaction between the polypeptide of peptidyl-tRNA in the P site and the amino acid of aminoacyl-tRNA in the A site.

The ribosome remains in place while the polypeptide chain is elongated by transferring the polypeptide attached to the tRNA in the P site to the aminoacyl-tRNA in the A site. The reaction is shown in **Figure 6.26.** The activity responsible for synthesis of the peptide bond is called peptidyl transferase.

The nature of the transfer reaction is revealed by the ability of the antibiotic puromycin to inhibit protein synthesis. Puromycin resembles an amino acid attached to the terminal adenosine of tRNA. **Figure 6.27** shows that puromycin has an N instead of the O that joins an amino acid to tRNA. The antibiotic is treated by the ribosome as though it were an incoming aminoacyl-tRNA. Then the polypeptide attached to peptidyl-tRNA is transferred to the NH_2 group of the puromycin.

Because the puromycin moiety is not anchored to the A site of the ribosome, the polypeptidyl-puromycin adduct is released from the ribosome in the form of polypeptidyl-puromycin. This premature termination of protein synthesis is responsible for the lethal action of the antibiotic.

Peptidyl transferase is a function of the large (50S or 60S) ribosomal subunit. The reaction is triggered when EF-Tu releases the aminoacyl end of its tRNA. The aminoacyl end then swings into a location close to the end of the peptidyl-tRNA. This site has a peptidyl transferase activity that essentially ensures a rapid transfer of the peptide chain to the aminoacyl-tRNA. Both rRNA and 50S subunit proteins are necessary for this activity, but the actual act of catalysis is a property of the ribosomal RNA of the 50S subunit (see *6.19 23S rRNA has peptidyl transferase activity).*

6.12 Translocation moves the ribosome

Key Concepts

- Ribosomal translocation moves the mRNA through the ribosome by 3 bases.
- Translocation moves deacylated tRNA into the E site, peptidyl-tRNA into the P site, and empties the A site.
- The hybrid state model proposes that translocation occurs in two stages, in which the 50S moves relative to the 30S, and then the 30S moves along mRNA to restore the original conformation.

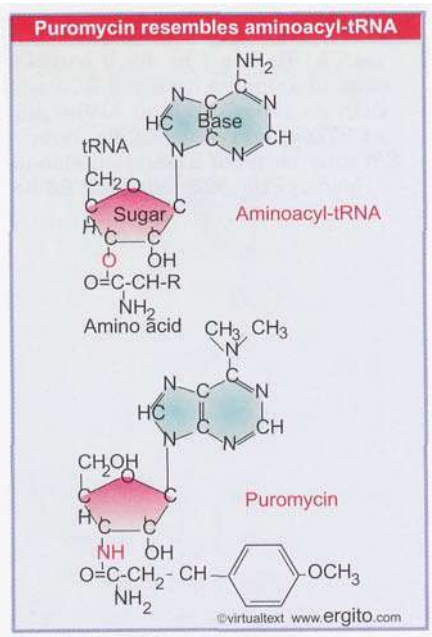


Figure 6.27 Puromycin mimics aminoacyl-tRNA because it resembles an aromatic amino acid linked to a sugar-base moiety.

The cycle of addition of amino acids to the growing polypeptide chain is completed by translocation, when the ribosome advances three nucleotides along the mRNA. **Figure 6.28** shows that translocation expels the uncharged tRNA from the P site, so that the new peptidyl-tRNA can enter. The ribosome then has an empty A site ready for entry of the aminoacyl-tRNA corresponding to the next codon. As the figure shows, in bacteria the discharged tRNA is transferred from the P site to the E site (from which it is then expelled into the cytoplasm). In eukaryotes it is expelled directly into the cytosol. The A and P sites straddle both the large and small subunits; the E site (in bacteria) is located largely on the 50S subunit, but has some contacts in the 30S subunit.

Most thinking about translocation follows the hybrid state model, which proposes that translocation occurs in two stages. **Figure** 6.29 shows that first there is a shift of the 50S subunit relative to the 30S subunit; then a second shift occurs when the 30S subunit moves along mRNA to restore the original conformation. The basis for this model was the observation that the pattern of contacts that tRNA makes with the ribosome (measured by chemical footprinting) changes in two

stages. When puromycin is added to a ribosome that has an aminoacylated tRNA in the P site, the contacts of tRNA on the 50S subunit change from the P site to the E site, but the contacts on the 30S subunit do not change. This suggests that the 50S subunit has moved to a post-transfer state, but the 30S subunit has not changed.

The interpretation of these results is that first the aminoacyl ends of the tRNAs (located in the 50S subunit) move into the new sites (while the anticodon ends remain bound to their anticodons in the 30S subunit). At this stage, the tRNAs are effectively bound in hybrid sites, consisting of the 50S E/ 30S P and the 50S P/ 30S A sites. Then movement is extended to the 30S subunits, so that the anticodon-codon pairing region finds itself in the right site. The most likely means of creating the hybrid state is by a movement of one ribosomal subunit relative to the other, so that translocation in effect involves two stages, the normal structure of the ribosome being restored by the second stage.

The ribosome faces an interesting dilemma at translocation. It needs to break many of its contacts with tRNA in order to allow movement. But at the same time it must maintain pairing between tRNA and the anticodon (breaking the pairing of the deacylated tRNA only at the right moment). One possibility is that the ribosome switches between alternative, discrete conformations. The switch could consist of changes in rRNA base pairing. The accuracy of translation is influenced by certain mutations that influence alternative base pairing arrangements. The most likely interpretation is that the effect is mediated by the tightness of binding to tRNA of the alternative conformations.

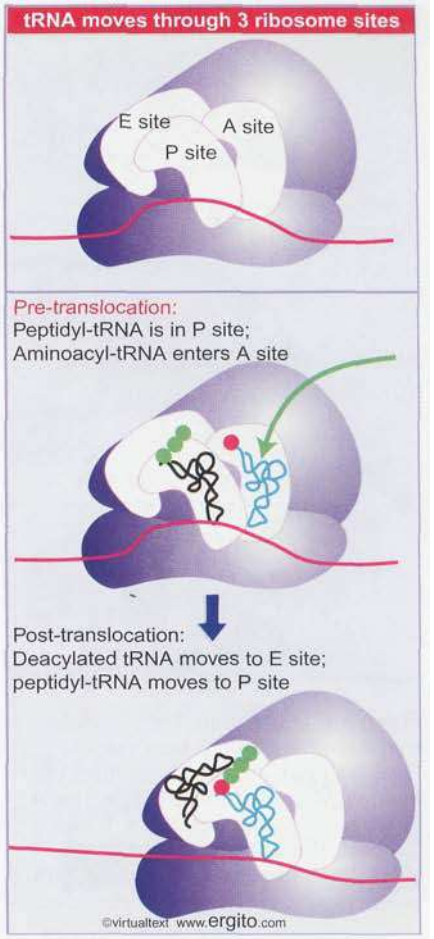


Figure 6.28 A bacterial ribosome has 3 tRNA-binding sites. Aminoacyl-tRNA enters the A site of a ribosome that has peptidyl-tRNA in the P site. Peptide bond synthesis deacylates the P site tRNA and generates peptidyl-tRNA in the A site. Translocation moves the deacylated tRNA into the E site and moves peptidyl-tRNA into the P site.

6.13 Elongation factors bind alternately to the ribosome

Key Concepts

- Translocation requires EF-G, whose structure resembles the aminoacyl-tRNA·EF-Tu·GTP complex.
- Binding of EF-Tu and EF-G to the ribosome is mutually exclusive.
- Translocation requires GTP hydrolysis, which triggers a change in *EF-G,* which in turn triggers a change in ribosome structure.

Translocation requires GTP and another elongation factor, EF-G. This factor is a major constituent of the cell; it is present at a level of ~1 copy per ribosome (20,000 molecules per cell).

Ribosomes cannot bind EF-Tu and EF-G simultaneously, so protein synthesis follows the cycle illustrated in **Figure 6.30** in which the factors are alternately bound to, and released from, the ribosome. So EF-Tu·GDP must be released before EF-G can bind; and then EF-G must be released before aminoacyl-tRNA-EF-Tu-GTP can bind.

Does the ability of each elongation factor to exclude the other rely on an allosteric effect on the overall conformation of the ribosome or on direct competition for overlapping binding sites? **Figure 6.31** shows an extraordinary similarity between the structures of the ternary complex of aminoacyl-tRNA·EF-Tu·GDP and EF-G. The structure of EF-G mimics the overall structure of EF-Tu bound to the amino acceptor stem of aminoacyl-tRNA. This creates the immediate assumption that they compete for the same binding site (presumably in the vicinity of the A site). The need for each factor to be released before the other can bind ensures that the events of protein synthesis proceed in an orderly manner.

Both elongation factors are monomeric GTP-binding proteins that are active when bound to GTP but inactive when bound to GDP. The

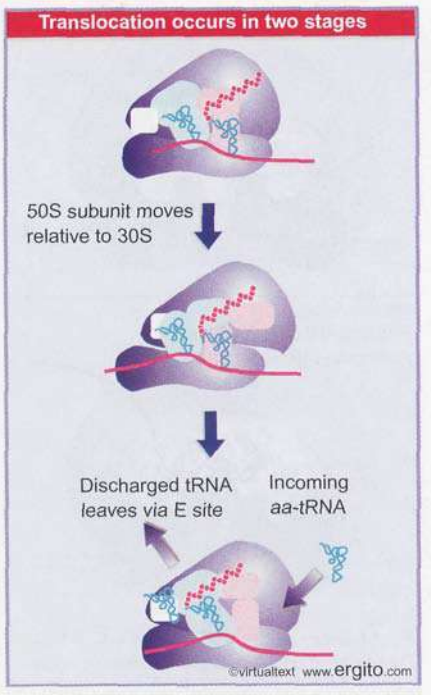


Figure 6.29 Models for translocation involve two stages. First, at peptide bond formation the aminoacyl end of the tRNA in the A site becomes located in the P site. Second, the anticodon end of the tRNA becomes located in the P site.

triphosphate form is required for binding to the ribosome, which ensures that each factor obtains access to the ribosome only in the company of the GTP that it needs to fulfill its function.

EF-G binds to the ribosome to sponsor translocation; and then is released following ribosome movement. EF-G can still bind to the ribosome when GMP-PCP is substituted for GTP; thus the presence of a guanine nucleotide is needed for binding, but its hydrolysis is not absolutely essential for translocation (although translocation is much slower in the absence of GTP hydrolysis). The hydrolysis of GTP is needed to release EF-G.

The need for EF-G release was discovered by the effects of the steroid antibiotic fusidic acid, which "jams" the ribosome in its post-translocation state (see Figure 6.30). In the presence of fusidic acid, one round of translocation occurs: EF-G binds to the ribosome, GTP is hydrolyzed, and the ribosome moves three nucleotides. But fusidic acid stabilizes the ribosome·EF-G·GDP complex, so that EF-G and GDP remain on the ribosome instead of being released. Because the ribosome then cannot bind aminoacyl-tRNA, no further amino acids can be added to the chain.

EF-G drives the ribosome's ability to translocate. It is an important part of the mechanism for translocation. The hydrolysis of GTP occurs before translocation and accelerates the ribosome movement. The most likely mechanism is that GTP hydrolysis causes a change in the structure of EF-G, which in turn forces a change in the ribosome structure. An extensive reorientation of EF-G occurs at translocation. Before translocation, it is bound across the two ribosomal subunits. Most of its contacts with the 30S subunit are made by a region called domain 4, which is inserted into the A site. This domain could be responsible for displacing the tRNA. After translocation, domain 4 is instead oriented toward the 50S subunit.

The eukaryotic counterpart to EF-G is the protein eEF2, which functions in a similar manner, as a translocase dependent on GTP hydrolysis. Its action also is inhibited by fusidic acid. A stable complex of eEF2 with GTP can be isolated; and the complex can bind to ribosomes with consequent hydrolysis of its GTP.

A unique reaction of eEF2 is its susceptibility to diphtheria toxin. The toxin uses NAD (nicotinamide adenine dinucleotide) as a cofactor to transfer an ADPR moiety (adenosine diphosphate ribosyl) on to the eEF2. The ADPR-eEF2 conjugate is inactive in protein synthesis. The substrate for the attachment is an unusual amino acid, produced by modifying a histidine; it is common to the eEF2 of many species.

The ADP-ribosylation is responsible for the lethal effects of diphtheria toxin. The reaction is extraordinarily effective: a single molecule of toxin can modify sufficient eEF2 molecules to kill a cell.

6.14 Three codons terminate protein synthesis

Key Concepts

- The codons UAA (ochre), UAG (amber) and UGA (opal) terminate protein synthesis.
- In bacteria they are used most often with relative frequencies UAA>UGA>UAG.

Only 61 triplets are assigned to amino acids. The other three triplet; are termination codons (or stop codons) that end protein synthesis. They have casual names from the history of their discovery. Th(

UAG triplet is called the amber codon; UAA is the ochre codon; and UGA is sometimes called the opal codon.

The nature of these triplets was originally shown by a genetic test that distinguished two types of point mutation:

- A point mutation that changes a codon to represent a different amino acid is called a missense mutation. One amino acid replaces the other in the protein; the effect on protein function depends on the site of mutation and the nature of the amino acid replacement.
- When a point mutation creates one of the three termination codons, it causes premature termination of protein synthesis at the mutant codon. Only the first part of the protein is made in the mutant cell. This is likely to abolish protein function (depending, of course, on how far along the protein the mutant site is located). A change of this sort is called a nonsense mutation.

(Sometimes the term *nonsense codon* is used to describe the termination triplets. "Nonsense" is really a misnomer, since the codons do have meaning, albeit a disruptive one in a mutant gene. A better term is stop codon.)

In every gene that has been sequenced, one of the termination codons lies immediately after the codon representing the C-terminal amino acid of the wild-type sequence. Nonsense mutations show that any one of the three codons is sufficient to terminate protein synthesis within a gene. The UAG, UAA, and UGA triplet sequences are therefore necessary and sufficient to end protein synthesis, whether occurring naturally at the end of a gene or created by mutation within a coding sequence.

In bacterial genes, UAA is the most commonly used termination codon. UGA is used more heavily than UAG, although there appear to be more errors reading UGA. (An error in reading a termination codon, when an aminoacyl-tRNA improperly responds to it, results in the continuation of protein synthesis until another termination codon is encountered.)

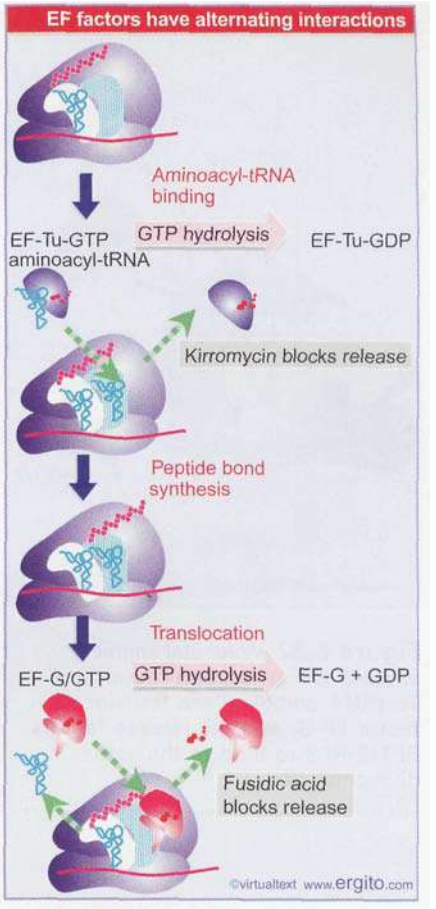


Figure 6.30 Binding of factors EF-Tu and EF-G alternates as ribosomes accept new aminoacyl-tRNA, form peptide bonds, and translocate.

6.15 Termination codons are recognized by protein factors

Key Concepts

- Termination codons are recognized by protein release factors not by aminoacyl-tRNAs.
- The structures of the class 1 release factors resemble aminoacyl-tRNA·EF-Tu and EF-G.
- The class 1 release factors respond to specific termination codons and hydrolyze the polypeptide-tRNA linkage.
- The class 1 release factors are assisted by class 2 release factors that depend on GTP.
- The mechanism is similar in bacteria (which have two types of class 1 release factors) and eukaryotes (which have only one class 1 release factor).

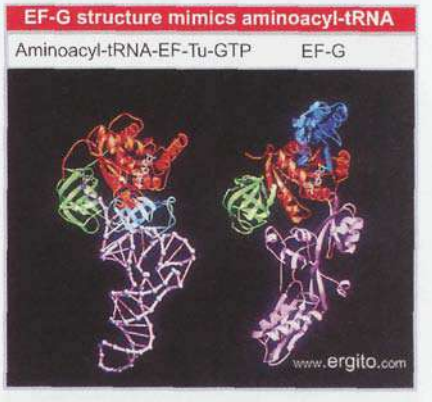


Figure 6.31 The structure of the ternary complex of aminoacyl-tRNA·EF-Tu·GTP (left) resembles the structure of EF-G (right). Structurally conserved domains of EF-Tu and EF-G are in red and green; the tRNA and the domain resembling it in EF-G are in purple. Photograph kindly provided by Poul Nissen.

Two stages are involved in ending translation. The *termination reaction* itself involves release of the protein chain from the last tRNA. The *post-termination reaction* involves release of the tRNA and mRNA, and dissociation of the ribosome into its subunits.

None of the termination codons is represented by a tRNA. They function in an entirely different manner from other codons, and are recognized directly by protein factors. (Since the reaction does not depend

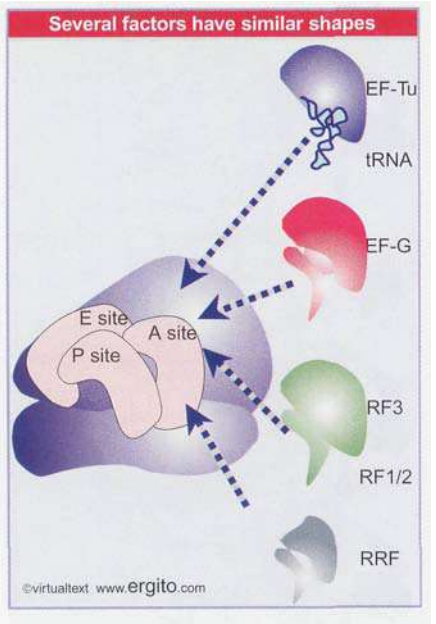


Figure 6.32 Molecular mimicry enables the elongation factor Tu-tRNA complex, the translocation factor EF-G, and the release factors RF1/2-RF3 to bind to the same ribosomal site.

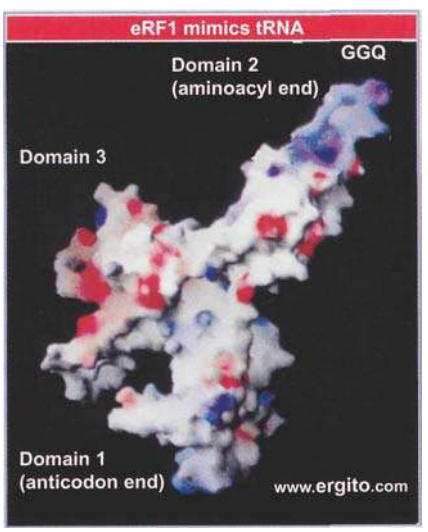


Figure 6.33 The eukaryotic termination factor eRF1 has a structure that mimics tRNA. The motif GGQ at the tip of domain 2 is essential for hydrolyzing the polypeptide chain from tRNA. Photograph kindly provided by David Barford.

on codon-anticodon recognition, there seems to be no particular reason why it should require a triplet sequence. Presumably this reflects the evolution of the genetic code.)

Termination codons are recognized by class 1 release factors (RF). In *E. coli* two class 1 release factors are specific for different sequences. RF1 recognizes UAA and UAG; RF2 recognizes UGA and UAA. The factors act at the ribosomal A site and require polypeptidyl-tRNA in the P site. The release factors are present at much lower levels than initiation or elongation factors; there are ~600 molecules of each per cell, equivalent to 1 RF per 10 ribosomes. Probably at one time there was only a single release factor, recognizing all termination codons, and later it evolved into two factors with specificities for particular codons. In eukaryotes, there is only a single class 1 release factor, called eRF. The efficiency with which the bacterial factors recognize their target codons is influenced by the bases on the 3' side.

The class 1 release factors are assisted by class 2 release factors, which are not codon-specific. The class 2 factors are GTP-binding proteins. In *E. coli*, the role of the class 2 factor is to release the class 1 factor from the ribosome.

Although the general mechanism of termination is similar in prokaryotes and eukaryotes, the interactions between the class 1 and class 2 factors have some differences.

The class 1 factors RF1 and RF2 recognize the termination codons and activate the ribosome to hydrolyze the peptidyl tRNA. Cleavage of polypeptide from tRNA takes place by a reaction analogous to the usual peptidyl transfer, except that the acceptor is H_2O instead of aminoacyl-tRNA (see Figure 6.34).

Then RF1 or RF2 is released from the ribosome by the class 2 factor RF3, which is related to EF-G. RF3 resembles the GTP-binding domains of EF-Tu and EF-G, and RF1/2 resemble the C-terminal domain of EF-G, which mimics tRNA. This suggests that the release factors utilize the same site that is used by the elongation factors. **Figure 6.32** illustrates the basic idea that these factors all have the same general shape and bind to the ribosome successively at the same site (basically the A site or a region extensively overlapping with it).

The eukaryotic class 1 release factor, eRF1, is a single protein that recognizes all three termination codons. Its sequence is unrelated to the bacterial factors. It can terminate protein synthesis *in vitro* without the class 3 factor, eRF3, although eRF3 is essential in yeast *in vivo*. The structure of eRF1 follows a familiar theme: **Figure 6.33** shows that it consists of three domains that mimic the structure of tRNA.

An essential motif of three amino acids, GGQ, is exposed at the top of domain 2. Its position in the A site corresponds to the usual location of an amino acid on an aminoacyl-tRNA. This positions it to use the glutamine (Q) to position a water molecule to substitute for the amino acid of aminoacyl-tRNA in the peptidyl transfer reaction. **Figure 6.34** compares the termination reaction with the usual peptide transfer reaction. Termination transfers a hydroxyl group from the water, thus effectively hydrolyzing the peptide-tRNA bond (and see Figure 6.48 for discussion of how the peptidyl transferase center works).

Mutations in the RF genes reduce the efficiency of termination, as seen by an increased ability to continue protein synthesis past the termination codon. Overexpression of RF1 or RF2 increases the efficiency of termination at the codons on which it acts. This suggests that codon recognition by RF 1 or RF2 competes with aminoacyl-tRNAs that erroneously recognize the termination codons. The release factors recognize their target sequences very efficiently.

The termination reaction involves release of the completed polypeptide, but leaves a deacylated tRNA and the mRNA still associated with

the ribosome. **Figure 6.35** shows that the dissociation of the remaining components (tRNA, mRNA, 30S and 50S subunits) requires the factor RRF, ribosome recycling factor. This acts together with EF-G in a reaction that uses hydrolysis of GTP. Like the other factors involved in release, RRF has a structure that mimics tRNA, except that it lacks an equivalent for the 3' amino acid-binding region. IF-3 is also required, which brings the wheel full circle to its original discovery, when it was proposed to be a dissociation factor! RRF acts on the 50S subunit, and IF-3 acts to remove deacylated tRNA from the 30S subunit. Once the subunits have separated, IF-3 remains necessary, of course, to prevent their reassociation.

6.16 Ribosomal RNA pervades both ribosomal subunits

Key Concepts

- Each rRNA has several distinct domains that fold independently.
- Virtually all ribosomal proteins are in contact with rRNA.
- Most of the contacts between ribosomal subunits are made between the 16S and 23S rRNAs.

Two thirds of the mass of the bacterial ribosome is made up of rRNA. The most penetrating approach to analyzing secondary structure of large RNAs is to compare the sequences of corresponding rRNAs in related organisms. Those regions that are important in the secondary structure retain the ability to interact by base pairing. So if a base pair is required, it can form at the same relative position in each rRNA. This approach has enabled detailed models to be constructed for both 16S and 23S rRNA.

Each of the major rRNAs can be drawn in a secondary structure with several discrete domains. 16S rRNA forms four general domains, *in which just under half of the sequence is base* paired (see Figure 6.45). 23S rRNA forms six general domains. The individual double-helical regions tend to be short (<8 bp). Often the duplex regions are not perfect, but contain bulges of unpaired bases. Comparable models have been drawn for mitochondrial rRNAs (which are shorter and have fewer domains) and for eukaryotic cytosolic rRNAs (which are longer and have more domains). The increase in length in eukaryotic rRNAs is due largely to the acquisition of sequences representing additional domains. The crystal structure of the ribosome shows that in each subunit the domains of the major rRNA fold independently and have a discrete location in the subunit.

Differences in the ability of 16S rRNA to react with chemical agents are found when 30S subunits are compared with 70S ribosomes; also there are differences between free ribosomes and those engaged in protein synthesis. Changes in the reactivity of the rRNA occur when mRNA is bound, when the subunits associate, or when tRNA is bound. Some changes reflect a direct interaction of the rRNA with mRNA or tRNA, while others are caused indirectly by other changes in ribosome structure. The main point is that ribosome conformation is flexible during protein synthesis.

A feature of the primary structure of rRNA is the presence of methylated residues. There are ~10 methyl groups in 16S rRNA (located mostly toward the 3' end of the molecule) and ~20 in 23S rRNA. In

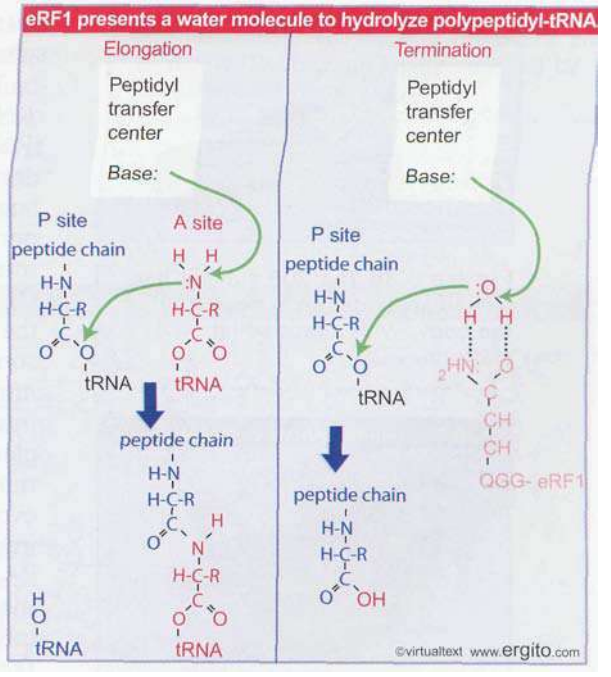


Figure 6.34 Peptide transfer and termination are similar reactions in which a base in the peptidyl transfer center triggers a transesterification reaction by attacking an N-H or O-H bond, releasing the N or O to attack the link to tRNA.

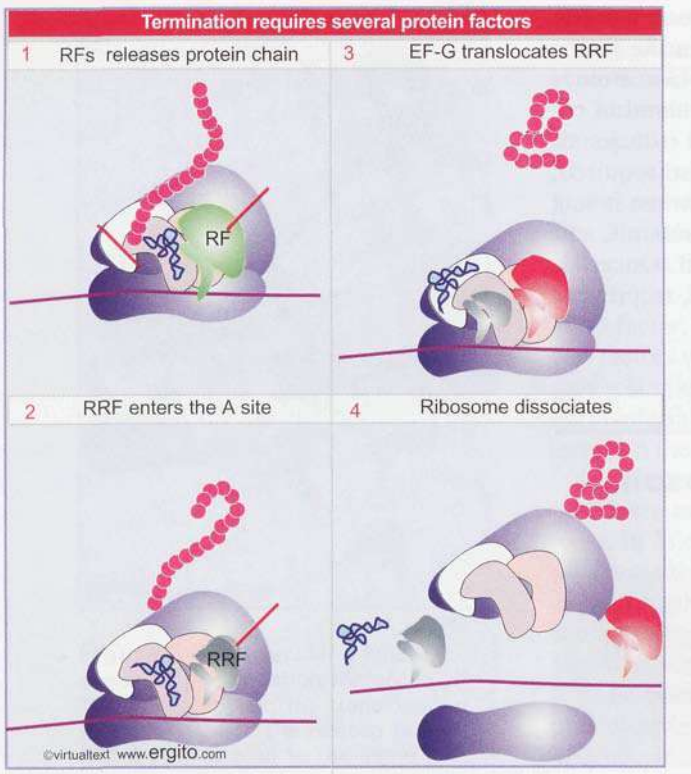


Figure 6.35 The RF (release factor) terminates protein synthesis by releasing the protein chain. The RRF (ribosome recycling factor) releases the last tRNA, and EF-G releases RRF, causing the ribosome to dissociate.

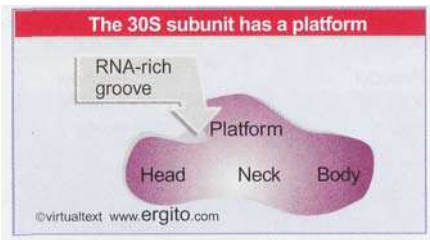


Figure 6.36 The 30S subunit has a head separated by a neck from the body, with a protruding platform.

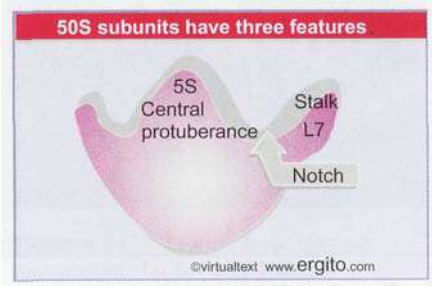


Figure 6.37 The 50S subunit has a central protuberance where 5S rRNA is located, separated by a notch from a stalk made of copies of the protein L7.

mammalian cells, the 18S and 28S rRNAs carry 43 and 74 methyl groups, respectively, so ~2% of the nucleotides are methylated (about three times the proportion methylated in bacteria).

The large ribosomal subunit also contains a molecule of a 120 base 5S RNA (in all ribosomes except those of mitochondria). The sequence of 5S RNA is less well conserved than those of the major rRNAs. All 5S RNA molecules display a highly base-paired structure.

In eukaryotic cytosolic ribosomes, another small RNA is present in the large subunit. This is the 5.8S RNA. Its sequence corresponds to the 5' end of the prokaryotic 23S rRNA.

Some ribosomal proteins bind strongly to isolated rRNA. Some do not bind to free rRNA, but can bind after other proteins have bound. This suggests that the conformation of the rRNA is important in determining whether binding sites exist for some proteins. As each protein binds, it induces conformational changes in the rRNA that make it possible for other proteins to bind. In *E. coli*, virtually all the 30S ribosomal proteins interact (albeit to varying degrees) with 16S rRNA. The binding sites on the proteins show a wide variety of structural features, suggesting that protein-RNA recognition mechanisms may be diverse.

The 70S ribosome has an asymmetric construction. **Figure** 6.36 shows a schematic of the structure of the 30S subunit, which is divided into four regions: the head, neck, body, and platform. **Figure 6.37** shows a similar representation of the 50S subunit, where two prominent features are the central protuberance (where 5S rRNA is located) and the stalk (made of multiple copies of protein L7). **Figure** 6.38 shows that the platform of the small subunit fits into the notch of the large subunit. There is a cavity between the subunits which contains some of the important sites.

The structure of the 30S subunit follows the organization of 16S rRNA, with each structural feature corresponding to a domain of the rRNA. The body is based on the 5' domain, the platform on the central domain, and the head on the 3' region. **Figure 6.39** shows that the 30S subunit has an asymmetrical distribution of RNA and protein. One important feature is that the platform of the 30S subunit that provides the interface with the 50S subunit is composed almost entirely of RNA. Only two proteins (a small part of S7 and possibly part of S12) lie near the interface. This means that the association and dissociation of ribosomal subunits must depend on interactions with the 16S rRNA. Subunit association is affected by a mutation in a loop of 16S rRNA (at position 791) that is located at the subunit interface, and other nucleotides in 16S rRNA have been shown to be involved by modification/interference experiments. This behavior supports the idea that the evolutionary origin of the ribosome may have been as a particle consisting of RNA rather than protein.

The 50S subunit has a more even distribution of components than the 30S, with long rods of double-stranded RNA crisscrossing the structure. The RNA forms a mass of tightly packed helices. The exterior surface largely consists of protein, except for the peptidyl transferase center (see *6.19 23S rRNA has peptidyl transferase activity*). Almost all segments of the 23 S rRNA interact with protein, but many of the proteins are relatively unstructured.

The junction of subunits in the 70S ribosome involves contacts between 16S rRNA (many in the platform region) with 23S rRNA. There are also some interactions between rRNA of each subunit with proteins

in the other, and a few protein-protein contacts. **Figure 6.40** identifies the contact points on the rRNA structures. **Figure 6.41** opens out the structure (imagine the 50S subunit rotated counterclockwise and the 30S subunit rotated clockwise around the axis shown in the figure) to show the locations of the contact points on the face of each subunit.

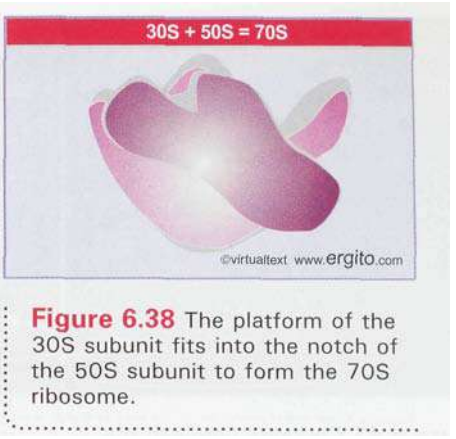


Figure 6.38 The platform of the 30S subunit fits into the notch of the 50S subunit to form the 70S ribosome.

6.17 Ribosomes have several active centers

Key Concepts

- Interactions involving rRNA are a key part of ribosome function.
- The environment of the tRNA-binding sites is largely determined by rRNA.

The basic message to remember about the ribosome is that it is a cooperative structure that depends on changes in the relationships among its active sites during protein synthesis. The active sites are not small, discrete regions like the active centers of enzymes. They are large regions whose construction and activities may depend just as much on the rRNA as on the ribosomal proteins. The crystal structures of the individual subunits and bacterial ribosomes give us a good impression of the overall organization and emphasize the role of the rRNA. The most recent structure, at 5.5 A resolution, clearly identifies the locations of the tRNAs and the functional sites. We can now account for many ribosomal functions in terms of its structure.

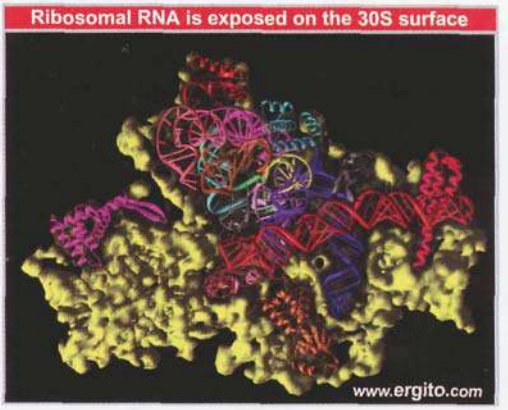


Figure 6.39 The 30S ribosomal subunit is a ribonucleoprotein particle. Proteins are in yellow. Photograph kindly provided by Venkitaraman Ramakrishnan.

Ribosomal functions are centered around the interaction with tRNAs. **Figure 6.42** shows the 70S ribosome with the positions of tRNAs in the three binding sites. The tRNAs in the A and P sites are nearly parallel to one another. All three tRNAs are aligned with their anticodon loops bound to the RNA-groove on the 30S subunit. The rest of each tRNA is bound to the 50S subunit. The environment surrounding each tRNA is mostly provided by rRNA. In each site, the rRNA contacts the tRNA at parts of the structure that are universally conserved.

It has always been a big puzzle to understand how two bulky tRNAs can fit next to one another in reading adjacent codons. The crystal structure shows a 45° kink in the mRNA between the P and A sites, which allows the tRNAs to fit as shown in the expansion of **Figure 6.43.** The tRNAs in the P and A sites are angled at 26° relative to each other at their anticodons. The closest approach between the backbones of the tRNAs occurs at the 3' ends, where they converge to within 5 A (perpendicular to the plane of the view). This allows the peptide chain to be transferred from the peptidyl-tRNA in the A site to the aminoacyl-tRNA in the A site.

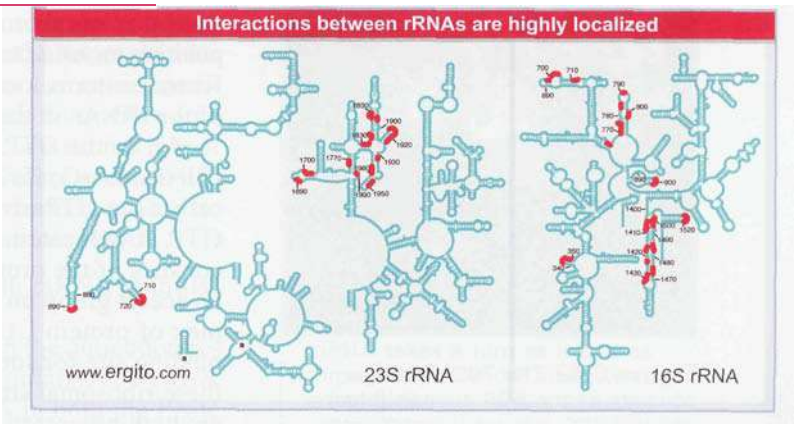


Figure 6.40 Contact points between the rRNAs are located in two domains of 16S rRNA and one domain of 23S rRNA. Photograph kindly provided by Harry Noller.

Aminoacyl-tRNA is inserted into the A site by EF-Tu, and its pairing with the codon is necessary for EF-Tu to hydrolyze GTP and be released from the ribosome (see *6.10 Elongation factor Tu loads aminoacyl-tRNA into the A site).* EF-Tu initially places the aminoacyl-tRNA into the small subunit, where the anticodon pairs with the codon. Movement of the tRNA is required to bring it fully into the A site, when its 3' end enters the peptidyl transferase center on the large subunit. There are different models for how this process may occur. One calls for the entire tRNA to

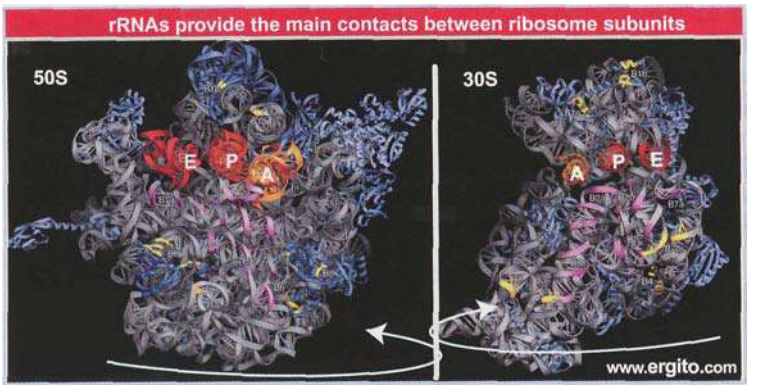


Figure 6.41 Contacts between the ribosomal subunits are mostly made by RNA (shown in purple). Contacts involving proteins are shown in yellow. The two subunits are rotated away from one another to show the faces where contacts are made; from a plane of contact perpendicular to the screen, the 50S subunit is rotated 90° counter-clockwise, and the 30S is rotated 90° clockwise (this shows it in the reverse of the usual orientation). Photograph kindly provided by Harry Noller.

swivel, so that the elbow in the L-shaped structure made by the D and TψC arms moves into the ribosome, enabling the TψC arm to pair with rRNA. Another calls for the internal structure of the tRNA to change, using the anticodon loop as a hinge, with the rest of the tRNA rotating from a position in which it is stacked on the 3' side of the anticodon loop to one in which it is stacked on the 5' side. Following the transition, EF-Tu hydrolyzes GTP, allowing peptide synthesis to proceed.

Translocation involves large movements in the positions of the tRNAs within the ribosome. The anticodon end of tRNA moves ~28 A from the A site to the P site, and then a further 20 A from the P site to the E site. Because of the angle of each tRNA relative to the anticodon, the bulk of the tRNA moves much larger distances, 40 A from A to P site, and 55 A from P site to E site. This suggests that translocation requires a major reorganization of structure.

The hybrid states model suggests that translocation may take place in two stages, with one ribosomal subunit moving relative to the other to create an intermediate stage in which they are hybrid tRNA-binding sites (50S E/30S P and 50S P/30S A) (see Figure 6.29). Comparisons of the ribosome structure between pre- and post-translocation states, and comparisons in 16S rRNA conformation between free 30S subunits and 70S ribosomes, suggest that mobility of structure is especially marked in the head and platform regions of the 30S subunit. An interesting insight on the hybrid states model is cast by the fact that many bases in rRNA involved in subunit association are close to bases involved in interacting with tRNA. This suggests that tRNA-binding sites are close to the interface between subunits, and carries the implication that changes in subunit interaction could be connected with movement of tRNA.

Much of the structure of the ribosome is occupied by its active centers. The schematic view of the ribosomal sites in **Figure 6.44** shows they comprise about two thirds of the ribosomal structure. A tRNA enters the A site, is transferred by translocation into the P site, and then leaves the (bacterial) ribosome by the E site. The A and P sites extend across both ribosome subunits; tRNA is paired with mRNA in the 30S subunit, but peptide transfer takes place in the 50S subunit. The A and P sites are adjacent, enabling translocation to move the tRNA from one site into the other. The E site is located near the P site (representing a position *en route* to the surface of the 50S subunit). The peptidyl transferase center is located on the 50S subunit, close to the aminoacyl ends of the tRNAs in the A and P sites (see next section).

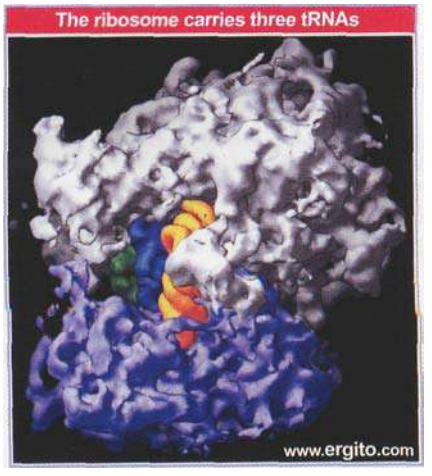


Figure 6.42 The 70S ribosome consists of the 50S subunit (blue) and the 30S subunit (purple) with three tRNAs located superficially: yellow in the A site, blue in the P site, and green in the E site. Photograph kindly provided by Harry Noller.

All of the GTP-binding proteins that function in protein synthesis (EF-Tu, EF-G, IF-2, RF 1,2,3) bind to a factor-binding site (sometimes called the GTPase center), which probably triggers their hydrolysis of GTP. It is located at the base of the stalk of the large subunit, which consists of the proteins L7/L12. (L7 is a modification of L12, and has an acetyl group on the N-terminus.) In addition to this region, the complex of protein L11 with a 58 base stretch of 23S rRNA provides the binding site for some antibiotics that affect GTPase activity. Neither of these ribosomal structures actually possesses GTPase activity, but they are both necessary for it. The role of the ribosome is to trigger GTP hydrolysis by factors bound in the factor-binding site.

Initial binding of 30S subunits to mRNA requires protein S1, which has a strong affinity for single-stranded nucleic acid. It is responsible for maintaining the single-stranded state in mRNA that is bound to th 30S subunit. This action is necessary to prevent the mRNA from takin

up a base-paired conformation that would be unsuitable for translation. S1 has an extremely elongated structure and associates with S18 and S21. The three proteins constitute a domain that is involved in the initial binding of mRNA and in binding initiator tRNA. This locates the mRNA-binding site in the vicinity of the cleft of the small subunit (see Figure 6.38). The 3' end of rRNA, which pairs with the mRNA initiation site, is located in this region.

The initiation factors bind in the same region of the ribosome. IF-3 can be crosslinked to the 3' end of the rRNA, as well as to several ribosomal proteins, including those probably involved in binding mRNA. The role of IF-3 could be to stabilize mRNA·30S subunit binding; then it would be displaced when the 50S subunit joins.

The incorporation of 5S RNA into 50S subunits that are assembled *in vitro* depends on the ability of three proteins, L5, L8, and L25, to form a stoichiometric complex with it. The complex can bind to 23S rRNA, although none of the isolated components can do so. It lies in the vicinity of the P and A sites.

A nascent protein debouches through the ribosome, away from the active sites, into the region in which ribosomes may be attached to membranes (see *8 Protein localization).* A polypeptide chain emerges from the ribosome through an exit channel, which leads from the peptidyl transferase site to the surface of the 50S subunit. The tunnel is composed mostly of rRNA. It is quite narrow, only 1-2 nm wide, and ~10 nm long. The nascent polypeptide emerges from the ribosome ~15 A away from the peptidyl transferase site. The tunnel can hold ~50 amino acids, and probably constrains the polypeptide chain so that it cannot fold until it leaves the exit domain.

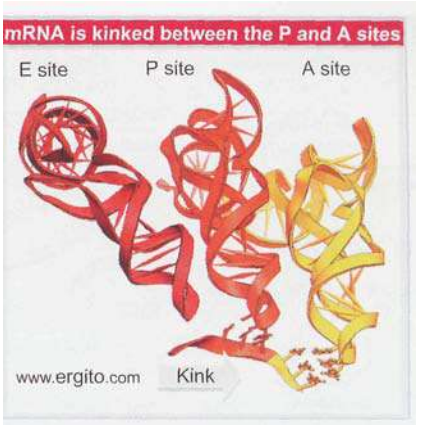


Figure 6.43 Three tRNAs have different orientations on the ribosome. mRNA turns between the P and A sites to allow aminoacyl-tRNAs to bind adjacent codons. Photograph kindly provided by Harry Noller.

6.18 16S rRNA plays an active role in protein synthesis

Key Concepts

16S rRNA plays an active role in the functions of the 30S subunit. It interacts directly with mRNA, with the 50S subunit, and with the anticodons of tRNAs in the P and A sites.

The ribosome was originally viewed as a collection of proteins with various catalytic activities, held together by protein-protein interactions and by binding to rRNA. But the discovery of RNA molecules with catalytic activities (see *24 RNA splicing and processing)* immediately suggests that rRNA might play a more active role in ribosome function. There is now evidence that rRNA interacts with mRNA or tRNA at each stage of translation, and that the proteins are necessary to maintain the rRNA in a structure in which it can perform the catalytic functions. Several interactions involve specific regions of rRNA:

- The 3' terminus of the rRNA interacts directly with mRNA at initiation.
- Specific regions of 16S rRNA interact directly with the anticodon regions of tRNAs in both the A site and the P site. Similarly, 23 S rRNA interacts with the CCA terminus of peptidyl-tRNA in both the P site and A site.
- Subunit interaction involves interactions between 16S and 23S rRNAs (see *6.16 Ribosomal RNA pervades both ribosomal subunits).*

Much information about the individual steps of bacterial protein synthesis has been obtained by using antibiotics that inhibit the process

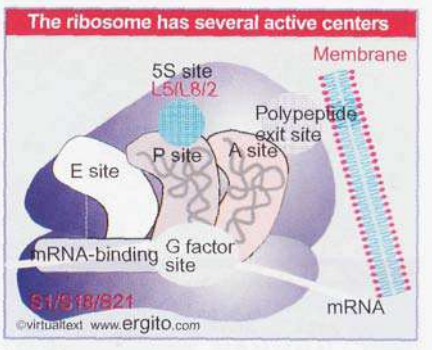


Figure 6.44 The ribosome has several active centers. It may be associated with a membrane. mRNA takes a turn as it passes through the A and P sites, which are angled with regard to each other. The E site lies beyond the P site. The peptidyl transferase site (not shown) stretches across the tops of the A and P sites. Part of the site bound by EF-Tu/G lies at the base of the A and P sites.

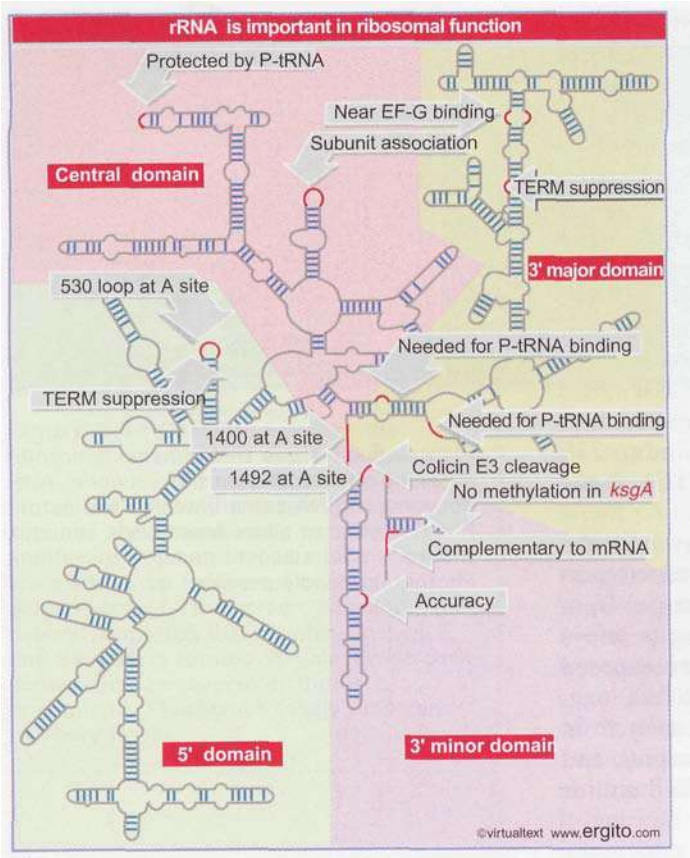


Figure 6.45 Some sites in 16S rRNA are protected from chemical probes when 50S subunits join 30S subunits or when aminoacyl-tRNA binds to the A site. Others are the sites of mutations that affect protein synthesis. TERM suppression sites may affect termination at some or several termination codons. The large colored blocks indicate the four domains of the rRNA.

at particular stages. The target for the antibiotic can be identified by the component in which resistant mutations occur. Some antibiotics act on individual ribosomal proteins, but several act on rRNA, which suggests that the rRNA is involved with many or even all of the functions of the ribosome.

The functions of rRNA have been investigated by two types of approach. Structural studies show that particular regions of rRNA are located in important sites of the ribosome, and that chemical modifications of these bases impede particular ribosomal functions. And mutations identify bases in rRNA that are required for particular ribosomal functions. **Figure 6.45** summarizes the sites in 16S rRNA that have been identified by these means.

An indication of the importance of the 3' end of 16S rRNA is given by its susceptibility to the lethal agent colicin E3. Produced by some bacteria, the colicin cleaves ~50 nucleotides from the 3' end of the 16S rRNA of *E. coli*. The cleavage entirely abolishes initiation of protein synthesis. Several important functions require the region that is cleaved: binding the factor IF-3; recognition of mRNA; and binding of tRNA.

The 3' end of the 16S rRNA is directly involved in the initiation reaction by pairing with the Shine-Dalgarno sequence in the ribosome-binding site of mRNA. Another direct role for the 3' end of 16S rRNA in protein synthesis is shown by the properties of kasugamycin-resistant mutants, which lack certain modifications in 16S rRNA. Kasugamycin blocks initiation of protein synthesis. Resistant mutants of the type *ksgA* lack a methylase enzyme that introduces four methyl groups into two adjacent adenines at a site near the 3' terminus of the 16S rRNA. The methylation generates the highly conserved sequence $G–m_2^6A–m_2^6A$, found in both prokaryotic and eukaryotic small rRNA. The methylated sequence is involved in the joining of the 30S and 50S subunits, which in turn is connected also with the retention of initiator tRNA in the complete ribosome. Kasugamycin causes $fMet\text{-}tRNA_f$ to be released from the sensitive (methylated) ribosomes, but the resistant ribosomes are able to retain the initiator.

Changes in the structure of 16S rRNA occur when ribosomes are engaged in protein synthesis, as seen by protection of particular bases against chemical attack. The individual sites fall into a few groups, concentrated in the 3′ minor and central domains. Although the locations are dispersed in the linear sequence of 16S rRNA, it seems likely that base positions involved in the same function are actually close together in the tertiary structure.

Some of the changes in 16S rRNA are triggered by joining with 50S subunits, binding of mRNA, or binding of tRNA. They indicate that these events are associated with changes in ribosome conformation that affect the exposure of rRNA. They do not necessarily indicate direct participation of rRNA in these functions. One change that occurs during protein synthesis is shown in **Figure 6.46;** it involves a local movement to change the nature of a short duplex sequence.

The 16S rRNA is involved in both A site and P site function, and significant changes in its structure occur when these sites are occupied. Certain distinct regions are protected by tRNA bound in the A site (see Figure 6.45). One is the 530 loop (which is also is the site of a mutation that prevents termination at the UAA, UAG, and UGA codons). The other is the 1400-1500 region (so-called because bases 1399-1492 and the adenines at 1492-1493 are two single-stranded stretches that are

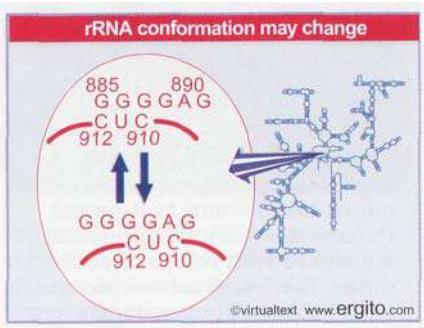


Figure 6.46 A change in conformation of 16S rRNA may occur during protein synthesis.

connected by a long hairpin). All of the effects that tRNA binding has on 16S rRNA can be produced by the isolated oligonucleotide of the anticodon stem-loop, so that tRNA-30S subunit binding must involve this region.

The adenines at 1492-1493 provide a mechanism for detecting properly paired codon-anticodon complexes. The principle of the interaction is that the structure of the 16S rRNA responds to the structure of the first two bases pairs in the minor groove of the duplex formed by the codon-anticodon interaction. Modification of the Nl position of either base 1492 or 1493 in rRNA prevents tRNA from binding in the A site. However, mutations at 1492 or 1493 can be suppressed by the introduction of fluorine at the 2' position of the corresponding bases in mRNA (which restores the interaction). **Figure 6.47** shows that codon-anticodon pairing allows the Nl of each adenine to interact with the 2′-OH in the mRNA backbone. When an incorrect tRNA enters the A site, the structure of the codon-anticodon complex is distorted, and this interaction cannot occur. The interaction stabilizes the association of tRNA with the A site.

A variety of bases in different positions of 16S rRNA are protected by tRNA in the P site; probably the bases lie near one another in the tertiary structure. In fact, there are more contacts with tRNA when it is in the P site than when it is in the A site. This may be responsible for the increased stability of peptidyl-tRNA compared with aminoacyl-tRNA. This makes sense, because once the tRNA has reached the P site, the ribosome has decided that it is correctly bound, whereas in the A site, the assessment of binding is being made. The 1400 region can be directly cross-linked to peptidyl-tRNA, which suggests that this region is a structural component of the P site.

The basic conclusion to be drawn from these results is that rRNA has many interactions with both tRNA and mRNA, and that these interactions recur in each cycle of peptide bond formation.

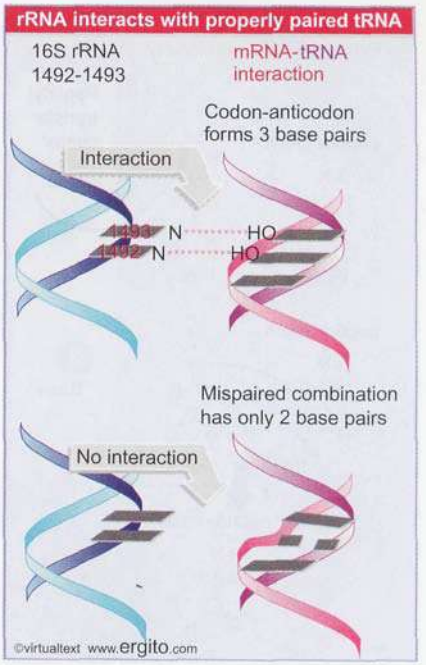


Figure 6.47 Codon-anticodon pairing supports interaction with adenines 1492-1493 of 16S rRNA, but mispaired tRNA-mRNA cannot interact.

6.19 23S rRNA has peptidyl transferase activity

Key Concepts

- Peptidyl transferase activity resides exclusively in the 23S rRNA.

The sites involved in the functions of 23S rRNA are less well identified than those of 16S rRNA, but the same general pattern is observed: bases at certain positions affect specific functions. Bases at some positions in 23S rRNA are affected by the conformation of the A site or P site. In particular, oligonucleotides derived from the 3' CCA terminus of tRNA protect a set of bases in 23 S rRNA which essentially are the same as those protected by peptidyl-tRNA. This suggests that the major interaction of 23S rRNA with peptidyl-tRNA in the P site involves the 3' end of the tRNA.

The tRNA makes contacts with the 23S rRNA in both the P and A sites. At the P site, G2552 of 23S rRNA base pairs with C74 of the peptidyl tRNA. A mutation in the G in the rRNA prevents interaction with tRNA, but interaction is restored by a compensating mutation in the C of the amino acceptor end of the tRNA. At the A site, G2553 of the 23S rRNA base pairs with C75 of the aminoacyl-tRNA. So there is a close role for rRNA in both the tRNA-binding sites. Indeed, when we have a clearer structural view of the region, we should be able to understand the movements of tRNA between the A and P sites in terms of making and breaking contacts with rRNA.

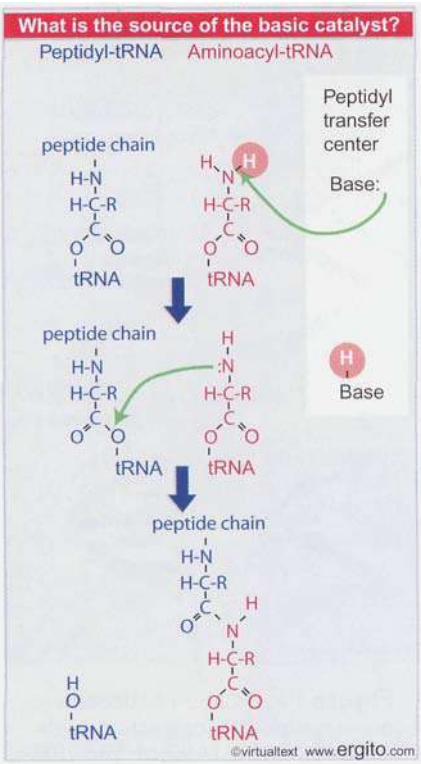


Figure 6.48 Peptide bond formation requires acid-base catalysis in which an H atom is transferred to a basic residue.

Another site that binds tRNA is the E site, which is localized almost exclusively on the 5 OS subunit. Bases affected by its conformation can be identified in 23S rRNA.

What is the nature of the site on the 50S subunit that provides peptidyl transferase function? The involvement of rRNA was first indicated because a region of the 23 S rRNA is the site of mutations that confer resistance to antibiotics that inhibit peptidyl transferase.

A long search for ribosomal proteins that might possess the catalytic activity has been unsuccessful. More recent results suggest that the ribosomal RNA of the large subunit has the catalytic activity. Extraction of almost all the protein content of 50S subunits leaves the 23S rRNA associated largely with fragments of proteins, amounting to <5% of the mass of the ribosomal proteins. This preparation retains peptidyl transferase activity. Treatments that damage the RNA abolish the catalytic activity.

Following from these results, 23S rRNA prepared by transcription *in vitro* can catalyze the formation of a peptide bond between Ac-Phe-tRNA and Phe-tRNA. The yield of Ac-Phe-Phe is very low, suggesting that the 23S rRNA requires proteins in order to function at a high efficiency. But since the rRNA has the basic catalytic activity, the role of the proteins must be indirect, serving to fold the rRNA properly or to present the substrates to it. The reaction also works, although less effectively, if the domains of 23 S rRNA are synthesized separately and then combined. In fact, some activity is shown by domain V alone, which has the catalytic center. Activity is abolished by mutations in position 2252 of domain V that lies in the P site.

The crystal structure of an archaeal 50S subunit shows that the peptidyl transferase site basically consists of 23 S rRNA. There is no protein within 18 A of the active site where the transfer reaction occurs between peptidyl-tRNA and aminoacyl-tRNA!

Peptide bond synthesis requires an attack by the amino group of one amino acid on the carboxyl group of another amino acid. Catalysis requires a basic residue to accept the hydrogen atom that is released from the amino group, as shown in **Figure 6.48.** If rRNA is the catalyst, it must provide this residue, but we do not know how this happens. The purine and pyrimidine bases are not basic at physiological pH. A highly conserved base (at position 2451 in *E. coli*) had been implicated, but appears now neither to have the right properties nor to be crucial for peptidyl transferase activity.

Proteins that are bound to the 23S rRNA outside of the peptidyl transfer region are almost certainly required to enable the rRNA to form the proper structure *in vivo.* The idea that rRNA is the catalytic component is consistent with the results discussed in *24 RNA splicing and processing* that identify catalytic properties in RNA that are involved with several RNA processing reactions. It fits with the notion that the ribosome evolved from functions originally possessed by RNA.

6.20 Summary

Ribosomes are ribonucleoprotein particles in which a majority of the mass is provided by rRNA. The shapes of all ribosomes are generally similar, but only those of bacteria (70S) have been characterized in detail. The small (30S) subunit has a squashed shape, with a "body" containing about two-thirds of the mass divided from the "head" by a cleft. The large (50S) subunit is more spherical, with a j prominent "stalk" on the right and a "central protuberance." Locations of all proteins are known approximately in the small subunit.

Each subunit contains a single major rRNA, 16S and 23S in prokaryotes, 18S and 28S in eukaryotic cytosol. There are also minor rRNAs, most notably 5S rRNA in the large subunit. Both major

rRNAs have extensive base pairing, mostly in the form of short, imperfectly paired duplex stems with single-stranded loops. Conserved features in the rRNA can be identified by comparing sequences and the secondary structures that can be drawn for rRNA of a variety of organisms. The 16S rRNA has four distinct domains; the 23S rRNA has six distinct domains. Eukaryotic rRNAs have additional domains.

The crystal structure shows that the 30S subunit has an asymmetrical distribution of RNA and protein. RNA is concentrated at the interface with the 50S subunit. The 50S subunit has a surface of protein, with long rods of double-stranded RNA crisscrossing the structure. 30S-50S joining involves contacts between 16S rRNA and 23S rRNA.

Each subunit has several active centers, concentrated in the translational domain of the ribosome where proteins are synthesized. Proteins leave the ribosome through the exit domain, which can associate with a membrane. The major active sites are the P and A sites, the E site, the EF-Tu and EF-G binding sites, peptidyl transferase, and mRNA-binding site. Ribosome conformation may change at stages during protein synthesis; differences in the accessibility of particular regions of the major rRNAs have been detected.

The tRNAs in the A and P sites are parallel to one another. The anticodon loops are bound to mRNA in a groove on the 30S subunit. The rest of each tRNA is bound to the 50S subunit. A conformational shift of tRNA within the A site is required to bring its aminoacyl end into juxtaposition with the end of the peptidyl-tRNA in the P site. The peptidyl transferase site that links the P and A binding sites is made of 23S rRNA, which has the peptidyl transferase catalytic activity, although proteins are probably needed to acquire the right structure.

An active role for the rRNAs in protein synthesis is indicated by mutations that affect ribosomal function, interactions with mRNA or tRNA that can be detected by chemical crosslinking, and the requirement to maintain individual base pairing interactions with the tRNA or mRNA. The 3' terminal region of the rRNA base pairs with mRNA at initiation. Internal regions make individual contacts with the tRNAs in both the P and A sites. Ribosomal RNA is the target for some antibiotics or other agents that inhibit protein synthesis.

A codon in mRNA is recognized by an aminoacyl-tRNA, which has an anticodon complementary to the codon and carries the amino acid corresponding to the codon. A special initiator tRNA (fMet-tRNAf in prokaryotes or Met-tRNAj in eukaryotes) recognizes the AUG codon, which is used to start all coding sequences. In prokaryotes, GUG is also used. Only the termination (nonsense) codons UAA, UAG, and UGA are not recognized by aminoacyl-tRNAs.

Ribosomes are released from protein synthesis to enter a pool of free ribosomes that are in equilibrium with separate small and large subunits. Small subunits bind to mRNA and then are joined by large subunits to generate an intact ribosome that undertakes protein synthesis. Recognition of a prokaryotic initiation site involves binding of a sequence at the 3' end of rRNA to the Shine-Dalgarno motif which precedes the AUG (or GUG) codon in the mRNA. Recognition of a eukaryotic mRNA involves binding to the 5' cap; the small subunit then migrates to the initiation site by scanning for AUG codons. When it recognizes an appropriate AUG codon (usually but not always the first it encounters), it is joined by a large subunit.

A ribosome can carry two aminoacyl-tRNAs simultaneously: its P site is occupied by a polypeptidyl-tRNA, which carries the polypeptide chain synthesized so far, while the A site is used for entry by an aminoacyl-tRNA carrying the next amino acid to be added to the chain. Bacterial ribosomes also have an E site, through which deacylated tRNA passes before it is released after being used in protein synthesis. The polypeptide chain in the P site is transferred to the aminoacyl-tRNA in the A site, creating a deacylated tRNA in the P site and a peptidyl-tRNA in the A site.

Following peptide bond synthesis, the ribosome translocates one codon along the mRNA, moving deacylated tRNA into the E site, and

peptidyl tRNA from the A site into the P site. Translocation is catalyzed by the elongation factor EF-G, and like several other stages of ribosome function, requires hydrolysis of GTP. During translocation, the ribosome passes through a hybrid stage in which the 50S subunit moves relative to the 30S subunit.

Protein synthesis is an expensive process. ATP is used to provide energy at several stages, including the charging of tRNA with its amino acid, and the unwinding of mRNA. It has been estimated that up to 90% of all the ATP molecules synthesized in a rapidly growing bacterium are consumed in assembling amino acids into protein!

Additional factors are required at each stage of protein synthesis. They are defined by their cyclic association with, and dissociation from, the ribosome. IF factors are involved in prokaryotic initiation. IF-3 is needed for 30S subunits to bind to mRNA and also is responsible for maintaining the 30S subunit in a free form. IF-2 is needed for fMet-tRNAf to bind to the 30S subunit and is responsible for excluding other aminoacyl-tRNAs from the initiation reaction. GTP is hydrolyzed after the initiator tRNA has been bound to the initiation complex. The initiation factors must be released in order to allow a large subunit to join the initiation complex.

Eukaryotic initiation involves a greater number of factors. Some of them are involved in the initial binding of the 40S subunit to the capped 5' end of the mRNA. Then the initiator tRNA is bound by another group of factors. After this initial binding, the small subunit scans the mRNA until it recognizes the correct AUG codon. At this point, initiation factors are released and the 60S subunit joins the complex.

Prokaryotic EF factors are involved in elongation. EF-Tu binds aminoacyl-tRNA to the 70S ribosome. GTP is hydrolyzed when EF-Tu is released, and EF-Ts is required to regenerate the active form of EF-Tu. EF-G is required for translocation. Binding of the EF-Tu and EF-G factors to ribosomes is mutually exclusive, which ensures that each step must be completed before the next can be started.

Termination occurs at any one of the three special codons, UAA UAG, UGA. Class 1 RF factors that specifically recognize the termination codons activate the ribosome to hydrolyze the peptidyl-tRNA. A class 2 RF factor is required to release the class 1 RF factor from the ribosome. The GTP-binding factors IF-2, EF-Tu, EF-G, RF3 all have similar structures, with the latter two mimicking the RNA-proteir structure of the first two when they are bound to tRNA; they all bind to the same ribosomal site, the G-factor binding site.

References

6.4 Initiation in bacteria needs 30S subunits and accessory factors

rev Maitra, U. et al. (1982). Initiation factors in protein biosynthesis. Ann. Rev. Biochem. 51, 869-900.

ref Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Hartsch, T., Wimberly, B. T., and Ramakrishnan, V. (2001). Crystal structure of an initiation factor bound to the 30S ribosomal subunit. Science 291, 498-501.

Dallas, A. and Noller, H. F. (2001). Interaction of translation initiation factor 3 with the 30S ribosomal subunit. Mol. Cell 8, 855-864.

Moazed, D., Samaha, R. R., Gualerzi, C, and Noller, H. F. (1995). Specific protection of 16S rRNA by translational initiation factors. J. Mol. Biol. 248, 207-210.

6.5 A special initiator tRNA starts the polypeptide chain

ref Lee, C. P., Seong, B. L., and RajBhandary, U. L. (1991). Structural and sequence elements important for recognition of *E. coli* formylmethionine tRNA by methionyl-tRNA transformylase are clustered in the acceptor stem. J. Biol. Chem. 266, 18012-18017.

Marcker, K. and Sanger, F. (1964). N-Formyl-methionyl-S-RNA. J. Mol. Biol. 8, 835-840.

Sundari, R. M., Stringer, E. A., Schulman, L. H., and Maitra, U. (1976). Interaction of bacterial initiation factor 2 with initiator tRNA. J. Biol. Chem. 251, 3338-3345.

6.7 Initiation involves base pairing between mRNA and rRNA

exp Steitz, J. (2002). rRNA-mRNA Base Pairing Selects Translational Initiator Regions in Bacteria (www.ergito.com/lookup.jsp?expt = steitz2)

6.8 Small subunits scan for initiation sites on eukaryoti mRNA

rev Kozak, M. (1978). How do eukaryotic ribosomes select initiation regions in MRNA? Cell 15, 1109-1 123.

Kozak, M. (1983). Comparison of initiation of proteir synthesis in prokaryotes, eukaryotes, and organelles Microbiol. Rev. 47, 1-45.

ref Hellen, C. U. and Sarnow, P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. 15, 1593-1612.

Kaminski, A., Howell, M. T., and Jackson, R. J. (1990). Initiation of encephalomyocarditis virus RNA translation: the authentic initiation site is not selected by a scanning mechanism. EMBO J. 9, 3753-3759.

Pelletier, J. and Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334, 320-325.

Pestova, T. V., Shatsky, I. N., Fletcher, S. P., Jackson, R. J., and Hellen, C. U. (1998). A prokaryotic-like mode of cytoplasmic eukaryotic ribosome binding to the initiation codon during internal translation initiation of hepatitis C and classical swine fever virus RNAs. Genes Dev. 12, 67-83.

Pestova, T. V., Hellen, C. U., and Shatsky, I. N. (1996). Canonical eukaryotic initiation factors determine initiation of translation by internal ribosomal entry. Mol. Cell Biol. 16, 6859-6869.

6.9 Eukaryotes use a complex of many initiation factors

rev Dever, T. E. (2002). Gene-specific regulation by general translation factors. Cell 108, 545-556.

Gingras, A. C, Raught, B., and Sonenberg, N. (1999). eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. Ann. Rev. Biochem. 68, 913-963.

Hershey, J. W. B. (1991). Translational control in mammalian cells. Ann. Rev. Biochem. 60, 717-755.

Merrick, W. C. (1992). Mechanism and regulation of eukaryotic protein synthesis. Microbiol. Rev. 56, 291-315.

Pestova, T. V., Kolupaeva, V. G., Lomakin, I. B., Pilipenko, E. V., Shatsky, I. N., Agol, V. I., and Hellen, C. U. (2001). Molecular mechanisms of translation initiation in eukaryotes. Proc. Nat. Acad. Sci. USA 98, 7029-7036.

Sachs, A., Sarnow, P., and Hentze, M. W. (1997). Starting at the beginning, middle, and end: translation initiation in eukaryotes. Cell 89, 831-838.

ref Asano, K., Clayton, J., Shalev, A., and Hinnebusch, A. G. (2000). A multifactor complex of eukaryotic initiation factors, eIF1, eIF2, eIF3, eIF5, and initiator tRNA(Met) is an important translation initiation intermediate in vivo. Genes Dev. 14, 2534-2546.

Huang, H. K., Yoon, H., Hannig, E. M., and Donahue, T. F. (1997). GTP hydrolysis controls stringent selection of the AUG start codon during translation initiation in *S. cerevisiae*. Genes Dev. 11, 2396-2413.

Pestova, T. V., Lomakin, I. B., Lee, J. H., Choi, S. K., Dever, T. E., and Hellen, C. U. (2000). The joining of ribosomal subunits in eukaryotes requires eIF5B. Nature 403, 332-335.

Tarun, S. Z. and Sachs, A. B. (1996). Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. EMBO J. 15, 7168-7177.

6.12 Translocation moves the ribosome

rev Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. Cell 108, 557-572.

ref Moazed, D. and Noller, H. F. (1989). Intermediate states in the movement of tRNA in the ribosome. Nature 342, 142-148.

Moazed, D. and Noller, H. F. (1986). Transfer RNA shields specific nucleotides in 16S rRNA from attack by chemical probes. Cell 47, 985-994.

Wilson, K. S. and Noller, H. F. (1998). Molecular movement inside the translational engine. Cell 92, 337-349.

6.13 Elongation factors bind alternately to the ribosome

ref Nissen, P. et al. (1995). Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. Science 270, 1464-1472.

Stark, H. et al. (2000). Large-scale movement of EF-G and extensive conformational change of the ribosome during translocation. Cell 100, 301-309.

6.15 Termination codons are recognized by protein factors

rev Eggertsson, G. and Soll, D. (1988). Transfer RNA-mediated suppression of termination codons in *E. coli.* Microbiol. Rev. 52, 354-374.

Frolova, L., et al. (1994). A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. Nature 372, 701-703.

Nissen, P., Kjeldgaard, M., and Nyborg, J. (2000). Macromolecular mimicry. EMBO J. 19, 489-495.

ref Freistroffer, D. V., Kwiatkowski, M., Buckingham, R. H., and Ehrenberg, M. (2000). The accuracy of codon recognition by polypeptide release factors. Proc. Nat. Acad. Sci. USA 97, 2046-2051.

Ito, K., Ebihara, K., Uno, M., and Nakamura, Y. (1996). Conserved motifs in prokaryotic and eukaryotic polypeptide release factors: tRNA-protein mimicry hypothesis. Proc. Nat. Acad. Sci. USA 93, 5443-5448.

Mikuni, 0., Ito, K., Moffat, J., Matsumura, K., McCaughan, K., Nobukuni, T., Tate, W., and Nakamura, Y. (1994). Identification of the *prfC* gene, which encodes peptide-chain-release factor 3 of *E. coli*. Proc. Nat. Acad. Sci. USA 91, 5798-5802.

Milman, G., Goldstein, J., Scolnick, E., and Caskey, T. (1969). Peptide chain termination. 3. Stimulation of *in vitro* termination. Proc. Nat. Acad. Sci. USA 63, 183-190.

Scolnick, E. et al. (1968). Release factors differing in specificity for terminator codons. Proc. Nat. Acad. Sci. USA 61, 768-774.

Selmer, M. et al. (1999). Crystal structure of *Thermotogo maritima* ribosome recycling factor: a tRNA mimic. Science 286, 2349-2352.

Song, H., Mugnier, P., Das, A. K., Webb, H. M., Evans, D. R., Tuite, M. F., Hemmings, B. A., and Barford, D. (2000). The crystal structure of human eukaryotic release factor eRF1 —mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. Cell 100, 311-321.

6.16 Ribosomal RNA pervades both ribosomal subunits

rev Hill, W. E. et al. (1990). The Ribosome. American Society for Microbiology, Washington DC.

Noller, H. F. (1984). Structure of ribosomal RNA. Ann. Rev. Biochem. 53, 119-162.

Noller, H. F. and Nomura, M. (1987). Ribosomes. In E. coli and S. typhimurium, Ed. F. C. Neidhardt, American Society for Microbiology, Washington DC, -.

Wittman, H. G. (1983). Architecture of prokaryotic ribosomes. Ann. Rev. Biochem. 52, 35-65.

ref Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289, 905-920.

Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B., and Steitz, T. A. (1999). Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. Nature 400, 841-847.

Clemons, W. M. et al. (1999). Structure of a bacterial 30S ribosomal subunit at 5.5 A resolution. Nature 400, 833-840.

Wimberly, B. T., Brodersen, D. E., Clemons WM, Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C, Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. Nature 407, 327-339.

Yusupov, M. M., Yusupova, G. 2., Baucom, A., Lieberman, A., Earnest, T. N., Cate, J. H. D., and Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. Science 292, 883-896.

6.17 Ribosomes have several active centers

rev Lafontaine, D. L. and Tollervey, D. (2001). The function and synthesis of ribosomes. Nat. Rev. Mol. Cell Biol. 2, 514-520.

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. Cell 108, 557-572.

ref Cate, J. H., Yusupov, M. M., Yusupova, G. Z., Earnest, T. N., and Noller, H. F. (1999). X-ray crystal structures of 70S ribosome functional complexes. Science 285, 2095-2104.

Sengupta, J., Agrawal, R. K., and Frank, J. (2001). Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. Proc. Nat. Acad. Sci. USA 98, 11991-11996.

Simonson, A. B. and Simonson, J. A. (2002). The transorientation hypothesis for codon recognition during protein synthesis. Nature 416, 281-285.

Valle, M., Sengupta, J., Swami, N. K., Grassucci, R. A., Burkhardt, N., Nierhaus, K. H., Agrawal, R. K., and Frank, J. (2002). Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. EMBO J. 21, 3557-3567.

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, A., Earnest, T. N., Cate, J. H. D., and Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. Science 292, 883-896.

6.18 1 6S rRNA plays an active role in protein synthesis

exp Steitz, J. (2002). rRNA-mRNA Base Pairing Selects Translational Initiator Regions in Bacteria (www.ergito.com/lookup.jsp?expt = steitz2)

rev Noller, H. F. (1991). Ribosomal RNA and translation. Ann. Rev. Biochem. 60, 191-227.

ref Lodmell, J. S. and Dahlberg, A. E. (1997). A conformational switch in *E. coli* 16S rRNA during decoding of mRNA. Science 277, 1262-1267.

Moazed, D. and Noller, H. F. (1986). Transfer RNA shields specific nucleotides in 1 6S rRNA from attack by chemical probes. Cell 47, 985-994.

Yoshizawa, S., Fourmy, D., and Puglisi, J. D. (1999). Recognition of the codon-anticodon helix by rRNA. Science 285, 1722-1725.

6.19 23S rRNA has peptidyl transferase activity

ref Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289, 905-920.

Bayfield, M. A., Dahlberg, A. E., Schulmeister, U., Dorner, S., and Barta, A. (2001). A conformational change in the ribosomal peptidyl transferase center upon active/inactive transition. Proc. Nat. Acad. Sci. USA 98, 10096-10101.

Noller, H. F., Hoffarth, V., and Zimniak, L. (1992). Unusual resistance of peptidyl transferase to protein extraction procedures. Science 256, 1416-1419.

Samaha, R. R., Green, R., and Noller, H. F. (1995). A base pair between tRNA and 23S rRNA in the peptidyl transferase center of the ribosome. Nature 377, 309-314.

Thompson, J., Thompson, D. F., O'Connor, M., Lieberman, K. R., Bayfield, M. A., Gregory, S. T., Green, R., Noller, H. F., and Dahlberg, A. E. (2001). Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit. Proc. Nat. Acad. Sci. USA 98, 9002-9007.

Using the genetic code

7.1 Introduction

Key Concepts

- 61 of the 64 possible triplets code for 20 amino acids.
- 3 codons do not represent amino acids and cause termination.
- Most amino acids are represented by more than one codon.
- The multiple codons for an amino acid are usually related.
- Related amino acids often have related codons, minimizing the effects of mutation.

The sequence of a coding strand of DNA, read in the direction from 5' to 3', consists of nucleotide triplets (codons) corresponding to the amino acid sequence of a protein read from N-terminus to C-terminus. Sequencing of DNA and proteins makes it possible to compare corresponding nucleotide and amino acid sequences directly. There are 64 codons (each of 4 possible nucleotides can occupy each of the three positions of the codon, making $4^3 = 64$ possible trinucleotide sequences). Each of these codons has a specific meaning in protein synthesis: 61 codons represent amino acids; 3 codons cause the termination of protein synthesis.

The meaning of a codon that represents an amino acid is determined by the tRNA that corresponds to it; the meaning of the termination codons is determined directly by protein factors.

The breaking of the genetic code originally showed that genetic information is stored in the form of nucleotide triplets, but did not reveal how each codon specifies its corresponding amino acid. Before the advent of sequencing, codon assignments were deduced on the basis of two types of *in vitro* studies. A system involving the translation of synthetic polynucleotides was introduced in 1961, when Nirenberg showed that polyuridylic acid [poly(U)] directs the assembly of phenylalanine into polyphenylalanine. This result means that UUU must be a codon for phenylalanine. A second system was later introduced in which a trinucleotide was used to mimic a codon, thus causing the corresponding aminoacyl-tRNA to bind to a ribosome. By identifying the amino acid component of the aminoacyl-tRNA, the meaning of the codon can be found. The two techniques together assigned meaning to all of the codons that represent amino acids.

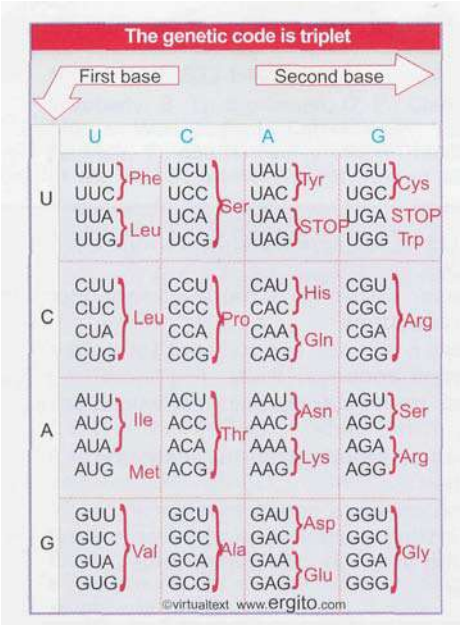


Figure 7.1 All the triplet codons have meaning: 61 represent amino acids, and 3 cause termination (STOP).

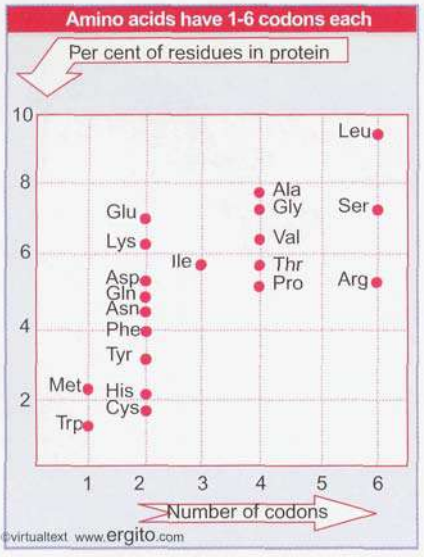


Figure 7.2 The number of codons for each amino acid does not correlate closely with its frequency of use in proteins.

The code is summarized in **Figure 7.1.** Because there are more codons (61) than there are amino acids (20), almost all amino acids are represented by more than one codon. The only exceptions are methionine and tryptophan. Codons that have the same meaning are called synonyms. Because the genetic code is actually read on the mRNA, usually it is described in terms of the four bases present in RNA: U, C, A, and G.

Codons representing the same or related amino acids tend to be similar in sequence. Often the base in the third position of a codon is not significant, because the four codons differing only in the third base represent the same amino acid. Sometimes a distinction is made only between a purine versus a pyrimidine in this position. The reduced specificity at the last position is known as third base degeneracy.

The interpretation of a codon requires base pairing with the anticodon of the corresponding aminoacyl-tRNA. The reaction occurs within the ribosome: complementary trinucleotides in isolation would usually be too short to pair in a stable manner, but the interaction is stabilized by the environment of the ribosomal A site. Also, base pairing between codon and anticodon is not solely a matter of A·U and GC base pairing. The ribosome controls the environment in such a way that conventional pairing occurs at the first two positions of the codon, but additional reactions are permitted at the third base. As a result, a single aminoacyl-tRNA may recognize more than one codon, corresponding with the pattern of degeneracy. Furthermore, pairing interactions may also be influenced by the introduction of special bases into tRNA, especially by modification in or close to the anticodon.

The tendency for similar amino acids to be represented by related codons minimizes the effects of mutations. It increases the probability that a single random base change will result in no amino acid substitution or in one involving amino acids of similar character. For example, a mutation of CUC to CUG has no effect, since both codons represent leucine; and a mutation of CUU to AUU results in replacement of leucine with isoleucine, a closely related amino acid.

Figure 7.2 plots the number of codons representing each amino acid against the frequency with which the amino acid is used in proteins (in *E. coli*). There is only a slight tendency for amino acids that are more common to be represented by more codons, and therefore it does not seem that the genetic code has been optimized with regard to the utilization of amino acids.

The three codons (UAA, UAG, and UGA) that do not represent amino acids are used specifically to terminate protein synthesis. One of these stop codons marks the end of every gene.

Is the genetic code the same in all living organisms?

Comparisons of DNA sequences with the corresponding protein sequences reveal that the identical set of codon assignments is used in bacteria and in eukaryotic cytoplasm. As a result, mRNA from one species usually can be translated correctly *in vitro* or *in vivo* by the protein synthetic apparatus of another species. So the codons used in the mRNA of one species have the same meaning for the ribosomes and tRNAs of other species.

The universality of the code argues that it must have been established very early in evolution. Perhaps the code started in a primitive form in which a small number of codons were used to represent comparatively few amino acids, possibly even with one codon corresponding to any member of a group of amino acids. More precise codon meanings and additional amino acids could have been introduced later. One possibility is that at first only two of the three bases in each codon were used; discrimination at the third position could have evolved later. (Originally there might have been a stereochemical relationship between amino acids and the codons representing them. Then a more complex system evolved.)

Evolution of the code could have become "frozen" at a point at which the system had become so complex that any changes in codon meaning would disrupt existing proteins by substituting unacceptable amino acids. Its universality implies that this must have happened at such an early stage that all living organisms are descended from a single pool of primitive cells in which this occurred.

Exceptions to the universal genetic code are rare. Changes in meaning in the principal genome of a species usually concern the termination codons. For example, in a mycoplasma, UGA codes for tryptophan; and in certain species of the ciliates Tetrahymena and Paramecium, UAA and UAG code for glutamine. Systematic alterations of the code have occurred only in mitochondrial DNA (see 7.6 *There are sporadic alterations of the universal code).*

7.2 Codon-anticodon recognition involves wobbling

Key Concepts

- Multiple codons that represent the same amino acid most often differ at the third base position.
- The wobble in pairing between the first base of the anticodon and the third base of the codon results from the structure of the anticodon loop.

The function of tRNA in protein synthesis is fulfilled when it recognizes the codon in the ribosomal A site. The interaction between anticodon and codon takes place by base pairing, but under rules that extend pairing beyond the usual G·C and A·U partnerships.

We can deduce the rules governing the interaction from the sequences of the anticodons that correspond to particular codons. The ability of any tRNA to respond to a given codon can be measured directly by the trinucleotide binding assay or by its use in an *in vitro* protein synthetic system.

The genetic code itself yields some important clues about the process of codon recognition. The pattern of third-base degeneracy is drawn in **Figure 7.3,** which shows that in almost all cases either the third base is irrelevant or a distinction is made only between purines and pyrimidines.

There are eight codon families in which all four codons sharing the same first two bases have the same meaning, so that the third base has no role at all in specifying the amino acid. There are seven codon pairs in which the meaning is the same whichever pyrimidine is present at the third position; and there are five codon pairs in which either purine may be present without changing the amino acid that is coded.

There are only three cases in which a unique meaning is conferred by the presence of a particular base at the third position: AUG (for methionine), UGG (for tryptophan), and UGA (termination). So C and U never have a unique meaning in the third position, and A never signifies a unique amino acid.

Because the anticodon is complementary to the codon, it is the first base in the anticodon sequence written conventionally in the direction from 5′ to 3′ that pairs with the third base in the codon sequence written by the same convention. So the combination

Codon	5′	A	C	G	3′
Anticodon	3′	U	G	C	5′

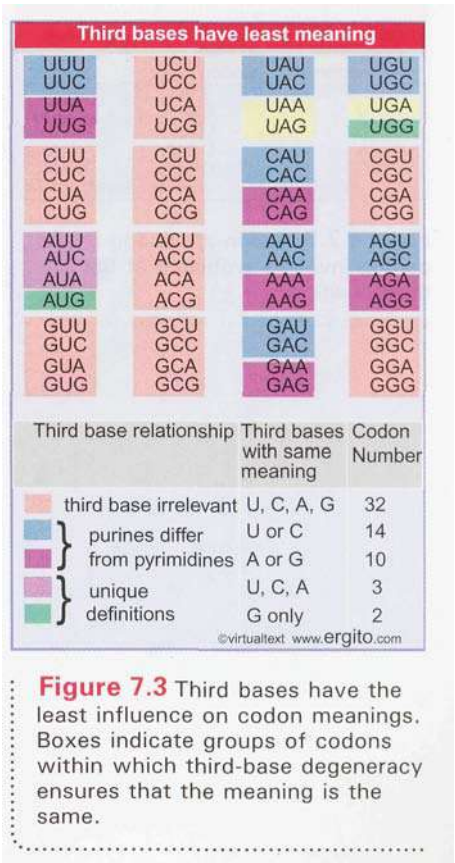


Figure 7.3 Third bases have the least influence on codon meanings. Boxes indicate groups of codons within which third-base degeneracy ensures that the meaning is the same.

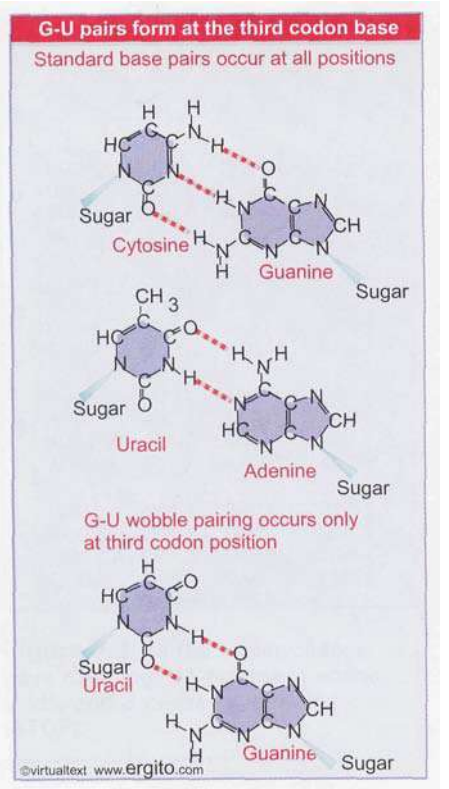


Figure 7.4 Wobble in base pairing allows G-U pairs to form between the third base of the codon and the first base of the anticodon.

The third codon base wobbles

Base in First Position of Anticodon	Base(s) Recognized in Third Position of Codon
U	A or G
C	G only
A	U only
G	C or U

©virtualtext www.ergito.com

Figure 7.5 Codon-anticodon pairing involves wobbling at the third position.

is usually written as codon ACG/anticodon CGU, where the anticodon sequence must be read backward for complementarity with the codon.

To avoid confusion, we shall retain the usual convention in which all sequences are written 5′–3′, but indicate anticodon sequences with a backward arrow as a reminder of the relationship with the codon. So the codon/anticodon pair shown above will be written as ACG and CGU←, respectively.

Does each triplet codon demand its own tRNA with a complementary anticodon? Or can a single tRNA respond to both members of a codon pair and to all (or at least some) of the four members of a codon family?

Often one tRNA can recognize more than one codon. This means that the base in the first position of the anticodon must be able to partner alternative bases in the corresponding third position of the codon. Base pairing at this position cannot be limited to the usual G·C and A·U partnerships.

The rules governing the recognition patterns are summarized in the wobble hypothesis, which states that the pairing between codon and anticodon at the first two codon positions always follows the usual rules, but that exceptional wobbles occur at the third position. Wobbling occurs because the conformation of the tRNA anticodon loop permits flexibility at the first base of the anticodon. **Figure 7.4** shows that G·U pairs can form in addition to the usual pairs.

This single change creates a pattern of base pairing in which A can no longer have a unique meaning in the codon (because the U that recognizes it must also recognize G). Similarly, C also no longer has a unique meaning (because the G that recognizes it also must recognize U). **Figure 7.5** summarizes the pattern of recognition.

It is therefore possible to recognize unique codons only when the third bases are G or U; this option is not used often, since UGG and AUG are the only examples of the first type, and there is none of the second type.

(G-U pairs are common in RNA duplex structures. But the formation of stable contacts between codon and anticodon, when only 3 base pairs can be formed, is more constrained, and thus G-U pairs can contribute only in the last position of the codon.)

7.3 tRNAs are processed from longer precursors

Key Concepts

* A mature tRNA is generated by processing a precursor.
* The 5' end *is generated by cleavage by* the endonuclease RNAase P.
* The 3' end is generated by cleavage followed by trimming of the last few bases, followed by addition of the common terminal trinucleotide sequence CCA.

tRNAs are commonly synthesized as precursor chains with additional material at one or both ends. **Figure 7.6** shows that the extra sequences are removed by combinations of endonucleolytic and exonucleolytic activities. One feature that is common to most tRNAs is that the three nucleotides at the 3' terminus, always the triplet sequence

CCA, are not coded in the genome, but are added as part of tRNA processing.

The 5' end of tRNA is generated by a cleavage action catalyzed by the enzyme ribonuclease P.

The enzymes that process the 3' end are best characterized in *E. coli*, where an endonuclease triggers the reaction by cleaving the precursor downstream, and several exonucleases then trim the end by degradation in the 3′ -5′ direction. The reaction also involves several enzymes in eukaryotes. It generates a tRNA that needs the CCA trinucleotide sequence to be added to the 3' end.

The addition of CCA is the result solely of an enzymatic process, that is, the enzymatic activity carries the specificity for the sequence of the trinucleotide, which is not determined by a template. There are several models for the process, which may be different in different organisms.

In some organisms, the process is catalyzed by a single enzyme. One model for its action proposes that a single enzyme binds to the 3' end, and sequentially adds C, C, and A, the specificity at each stage being determined by the structure of the 3′ end. Other models propose that the enzyme has different active sites for CTP and ATP.

In other organisms, different enzymes are responsible for adding the C and A residues, and they function sequentially.

When a tRNA is not properly processed, it attracts the attention of a quality control system that degrades it. This ensures that the protein synthesis apparatus does not become blocked by nonfunctional tRNAs.

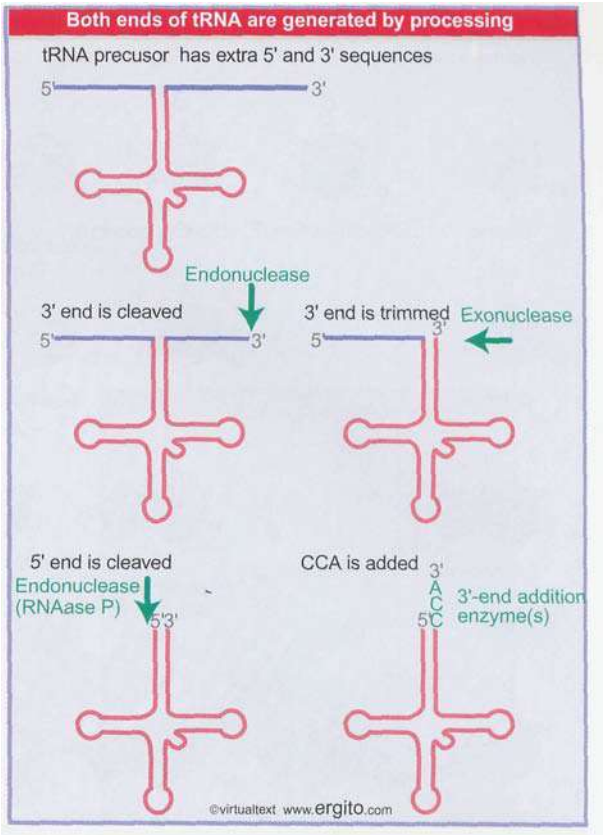


Figure 7.6 The tRNA 3′ end is generated by cutting and trimming followed by addition of CCA; the 5' end is generated by cutting.

7.4 tRNA contains modified bases

Key Concepts

- tRNAs contain >50 modified bases.
- Modification usually involves direct alteration of the primary bases in tRNA, but there are some exceptions in which a base is removed and replaced by another base.

Transfer RNA is unique among nucleic acids in its content of "unusual" bases. An unusual base is any purine or pyrimidine ring except the usual A, G, C, and U from which all RNAs are synthesized. All other bases are produced by modification of one of the four bases after it has been incorporated into the polyribonucleotide chain.

All classes of RNA display some degree of modification, but in all cases except tRNA this is confined to rather simple events, such as the addition of methyl groups. In tRNA, there is a vast range of modifications, ranging from simple methylation to wholesale restructuring of the purine ring. Modifications occur in all parts of the tRNA molecule. There are >50 different types of modified bases in tRNA.

Figure 7.7 shows some of the more common modified bases. Modifications of pyrimidines (C and U) are less complex than those of purines (A and G). In addition to the modifications of the bases themselves, methylation at the 2'-0 position of the ribose ring also occurs.

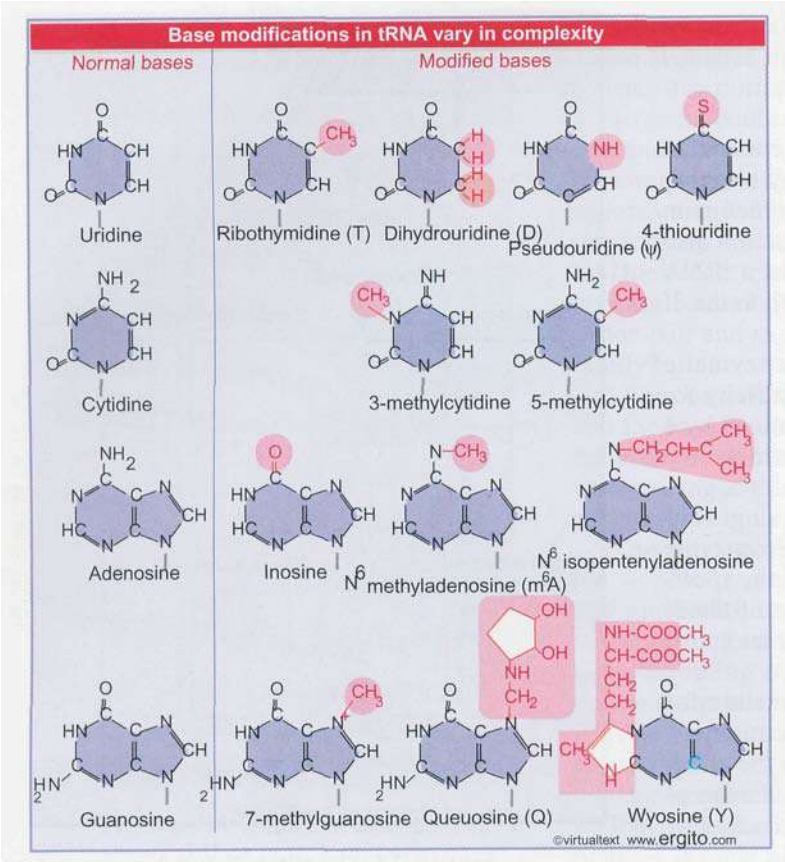


Figure 7.7 All of the four bases in tRNA can be modified.

The most common modifications of uridine are straightforward. Methylation at position 5 creates ribothymidine (T). The base is the same commonly found in DNA; but here it is attached to ribose, not deoxyribose. In RNA, thymine constitutes an unusual base, originating by modification of U.

Dihydrouridine (D) is generated by the saturation of a double bond, changing the ring structure. Pseudouridine (ψ) interchanges the positions of N and C atoms (see Figure 24.40). And 4-thiouridine has sulfur substituted for oxygen.

The nucleoside inosine is found normally in the cell as an intermediate in the purine biosynthetic pathway. However, it is not incorporated directly into RNA, where instead its existence depends on modification of A to create I. Other modifications of A include the addition of complex groups.

Two complex series of nucleotides depend on modification of G. The Q bases, such as queuosine, have an additional pentenyl ring added via an NH linkage to the methyl group of 7-methylguanosine. The pentenyl ring may carry various further groups. The Y bases, such as wyosine, have an additional ring fused with the purine ring itself; the extra ring carries a long carbon chain, again to which further groups are added in different cases.

The modification reaction usually involves the alteration of, or addition to, existing bases in the tRNA. An exception is the synthesis of Q bases, where a special enzyme exchanges free queuosine with a guanosine residue in the tRNA. The reaction involves breaking and remaking bonds on either side of the nucleoside.

The modified nucleosides are synthesized by specific tRNA-modifying enzymes. The original nucleoside present at each position can be determined either by comparing the sequence of tRNA with that of its gene or (less efficiently) by isolating precursor molecules that lack some or all of the modifications. The sequences of precursors show that different modifications are introduced at different stages during the maturation of tRNA.

Some modifications are constant features of all tRNA molecules—for example, the D residues that give rise to the name of the D arm, and the ψ found in the TψC sequence. On the 3′ side of the anticodon there is always a modified purine, although the modification varies widely.

Other modifications are specific for particular tRNAs or groups of tRNAs. For example, wyosine bases are characteristic of $tRNA^{Phe}$ in bacteria, yeast, and mammals. There are also some species-specific patterns.

The many tRNA-modifying enzymes (~60 in yeast) vary greatly in specificity. In some cases, a single enzyme acts to make a particular modification at a single position. In other cases, an enzyme can modify bases at several different target positions. Some enzymes undertake single reactions with individual tRNAs; others have a range of substrate molecules. The features recognized by the tRNA-modifying enzymes are unknown, but probably involve recognition of structural features surrounding the site of modification. Some modifications require the successive actions of more than one enzyme.

By Book_Crazy [IND]

7.5 Modified bases affect anticodon-codon pairing

Key Concepts

- Modifications in the anticodon affect the pattern of wobble pairing and therefore are important in determining tRNA specificity.

The most direct effect of modification is seen in the anticodon, where change of sequence influences the ability to pair with the codon, thus determining the meaning of the tRNA. Modifications elsewhere in the vicinity of the anticodon also influence its pairing.

When bases in the anticodon are modified, further pairing patterns become possible in addition to those predicted by the regular and wobble pairing involving A, C, U, and G. **Figure 7.8** shows the use of inosine (I), which is often present at the first position of the anticodon. Inosine can pair with any one of three bases, U, C, and A.

This ability is especially important in the isoleucine codons, where AUA codes for isoleucine, while AUG codes for methionine. Because with the usual bases it is not possible to recognize A alone in the third position, any tRNA with U starting its anticodon would have to recognize AUG as well as AUA. So AUA must be read together with AUU and AUC, a problem that is solved by the existence of tRNA with I in the anticodon.

Actually, some of the predicted regular combinations do not occur, because some bases are always modified. There seems to be an absolute ban on the employment of A; usually it is converted to I. And U at the first position of the anticodon is usually converted to a modified form that has altered pairing properties.

Some modifications create preferential readings of some codons with respect to others. Anticodons with uridine-5-oxyacetic acid and 5-methoxyuridine in the first position recognize A and G efficiently as third bases of the codon, but recognize U less efficiently. Another case in which multiple pairings can occur, but with some preferred to others, is provided by the series of queuosine and its derivatives. These modified G bases continue to recognize both C and U, but pair with U more readily.

A restriction not allowed by the usual rules can be achieved by the employment of 2-thiouridine in the anticodon. This modification allows the base to continue to pair with A, but prevents it from indulging in wobble pairing with G. **Figure 7.9** shows why 2-thiouracil pairs only with A.

These and other pairing relationships make the general point that there are multiple ways to construct a set of tRNAs able to recognize all the 61 codons representing amino acids. No particular pattern predominates in any given organism, although the absence of a certain pathway for modification can prevent the use of some recognition patterns. So a particular codon family is read by tRNAs with different anticodons in different organisms.

Often the tRNAs will have overlapping responses, so that a particular codon is read by more than one tRNA. In such cases there may be differences in the efficiencies of the alternative recognition reactions. (As a general rule, codons that are commonly used tend to be more efficiently read.) And in addition to the construction of a set of tRNAs able to recognize all the codons, there may be multiple tRNAs that respond to the same codons.

The predictions of wobble pairing accord very well with the observed abilities of almost all tRNAs. But there are exceptions in which

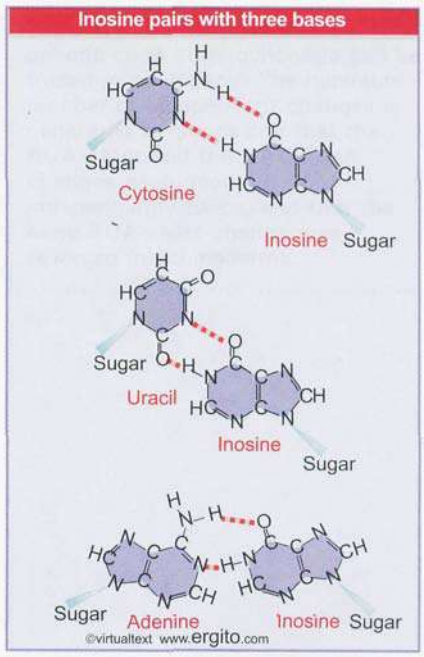


Figure 7.8 Inosine can pair with any of U, C, and A.

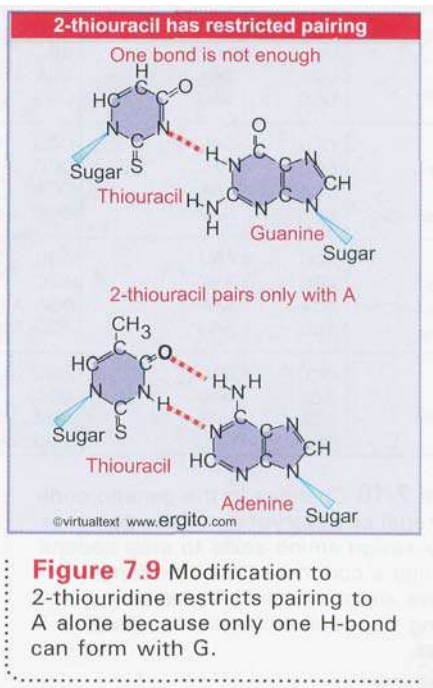


Figure 7.9 Modification to 2-thiouridine restricts pairing to A alone because only one H-bond can form with G.

the codons recognized by a tRNA differ from those predicted by the wobble rules. Such effects probably result from the influence of neighboring bases and/or the conformation of the anticodon loop in the overall tertiary structure of the tRNA. Indeed, the importance of the structure of the anticodon loop is inherent in the idea of the wobble hypothesis itself. Further support for the influence of the surrounding structure is provided by the isolation of occasional mutants in which a change in a base in some other region of the molecule alters the ability of the anticodon to recognize codons.

Another unexpected pairing reaction is presented by the ability of the bacterial initiator, fMet-tRNAf, to recognize both AUG and GUG. This misbehavior involves the third base of the anticodon.

7.6 There are sporadic alterations of the universal code

Key Concepts

- Changes in the universal genetic code have occurred in some species.
- They are more common in mitochondrial genomes, where a phylogenetic tree can be constructed for the changes.
- In nuclear genomes, they are sporadic and usually affect only termination codons.

The universality of the genetic code is striking, but some exceptions exist. They tend to affect the codons involved in initiation or termination and result from the production (or absence) of tRNAs representing certain codons. The changes found in principal (bacterial or nuclear) genomes are summarized in **Figure 7.10.**

Almost all of the changes that allow a codon to represent an amino acid affect termination codons:

- In the prokaryote *Mycoplasma capricolum*, UGA is not used for termination, but instead codes for tryptophan. In fact, it is the predominant Trp codon, and UGG is used only rarely. Two Trp-tRNA species exist, with the anticodons UCA← (reads UGA and UGG) and CCA← (reads only UGG).
- Some ciliates (unicellular protozoa) read UAA and UAG as glutamine instead of termination signals. *Tetrahymena thermophila*, one of the ciliates, contains three $tRNA^{Glu}$ species. One recognizes the usual codons CAA and CAG for glutamine, one recognizes both UAA and UAG (in accordance with the wobble hypothesis), and the last recognizes only UAG. We assume that a further change is that the release factor eRF has a restricted specificity, compared with that of other eukaryotes.
- In another ciliate *(Euplotes octacarinatus)*, UGA codes for cysteine. Only UAA is used as a termination codon, and UAG is not found. The change in meaning of UGA might be accomplished by a modification in the anticodon of $tRNA^{Cys}$ to allow it to read UGA with the usual codons UGU and UGC.

The only substitution in coding for amino acids occurs in a yeast *(Candida)*, where CUG means serine instead of leucine (and UAG is used as a sense codon).

Changes in the genetic code usually involve Stop/None signals

UUU Phe	UCU	UAU Tyr	UGU Cys
UUC	UCC Ser	UAC	UGC
UUA	UCA	UAA STOP→Gln	UGA STOP→Trp, Cys, Sel
UUG Leu	UCG	UAG	UGG Trp
CUU	CCU	CAU His	CGU
CUC Leu	CCC Pro	CAC	CGC
CUA	CCA	CAA Gln	CGA Arg
CUG Leu→Ser	CCG	CAG	CGG Arg→NONE
AUU	ACU	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC	AGC
AUA Ile→NONE	ACA	AAA Lys	AGA Arg→NONE
AUG Met	ACG	AAG	AGG Arg
GUU	GCU	GAU Asp	GGU
GUC	GCC Ala	GAC	GGC
GUA Val	GCA	GAA Glu	GGA Gly
GUG	GCG	GAG	GGG

Figure 7.10 Changes in the genetic code in bacterial or eukaryotic nuclear genomes usually assign amino acids to stop codons or change a codon so that it no longer specifies an amino acid. A change in meaning from one amino acid to another is unusual.

Acquisition of a coding function by a termination codon requires two types of change: a tRNA must be mutated so as to recognize the codon; and the class 1 release factor must be mutated so that it does not terminate at this codon.

The other common type of change is loss of the tRNA that responds to a codon, so that the codon no longer specifies any amino acid. What happens at such a codon will depend on whether the termination factor evolves to recognize it.

All of these changes are sporadic, which is to say that they appear to have occurred independently in specific lines of evolution. They may be concentrated on termination codons, because these changes do not involve substitution of one amino acid for another. Once the genetic code was established, early in evolution, any general change in the meaning of a codon would cause a substitution in all the proteins that contain that amino acid. It seems likely that the change would be deleterious in at least some of these proteins, with the result that it would be strongly selected against. The divergent uses of the termination codons could represent their "capture" for normal coding purposes. If some termination codons were used only rarely, they could be recruited to coding purposes by changes that allowed tRNAs to recognize them.

Exceptions to the universal genetic code also occur in the mitochondria from several species. **Figure 7.11** constructs a phylogeny for the changes. It suggests that there was a universal code that was changed at various points in mitochondrial evolution. The earliest change was the employment of UGA to code for tryptophan, which is common to all (non-plant) mitochondria.

Some of these changes make the code simpler, by replacing two codons that had different meanings with a pair that has a single meaning. Pairs treated like this include UGG and UGA (both Trp instead of one Trp and one termination) and AUG and AUA (both Met instead of one Met and the other Ile).

Why have changes been able to evolve in the mitochondrial code? Because the mitochondrion synthesizes only a small number of proteins (~10), the problem of disruption by changes in meaning is much less severe. Probably the codons that are altered were not used extensively in locations where amino acid substitutions would have been deleterious. The variety of changes found in mitochondria of different species suggests that they have evolved separately, and not by common descent from an ancestral mitochondrial code.

According to the wobble hypothesis, a minimum of 31 tRNAs (excluding the initiator) are required to recognize all 61 codons (at least 2 tRNAs are required for each codon family and 1 tRNA is needed per codon pair or single codon). But an unusual situation exists in (at least) mammalian mitochondria in which there are only 22 different tRNAs. How does this limited set of tRNAs accommodate all the codons?

The critical feature lies in a simplification of codon-anticodon pairing, in which one tRNA recognizes all four members of a codon family. This reduces to 23 the minimum number of tRNAs required to respond to all usual codons. The use of AG$^{A}_{G}$ for termination reduces the requirement by one further tRNA, to 22.

In all eight codon families, the sequence of the tRNA contains an unmodified U at the first position of the anticodon. The remaining codons are grouped into pairs in which all the codons ending in pyrimidines are read by G in the anticodon, and all the codons ending in purines are read by a modified U in the anticodon, as predicted by the wobble hypothesis. The complication of the single UGG codon is avoided by the change in the code to read UGA with UGG as tryptophan; and in mammals, AUA ceases to represent isoleucine and instead is read with AUG as methionine. This allows all the nonfamily codons to be read as 14 pairs.

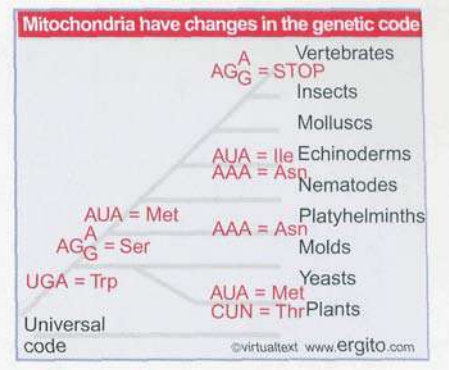


Figure 7.11 Changes in the genetic code in mitochondria can be traced in phylogeny. The minimum number of independent changes is generated by supposing that the AUA = Met and the AAA = Asn changes each occurred independently twice, and that the early AUA = Met change was reversed in echinoderms.

The 22 identified tRNA genes therefore code for 14 tRNAs representing pairs, and 8 tRNAs representing families. This leaves the two usual termination codons UAG and UAA unrecognized by tRNA, together with the codon pair AG^{A}_{G}. Similar rules are followed in the mitochondria of fungi.

7.7 Novel amino acids can be inserted at certain stop codons

Key Concepts

- Changes in the reading of specific codons can occur in individual genes.
- The insertion of seleno-Cys-tRNA at certain UGA codons requires several proteins to modify the Cys-tRNA and insert it into the ribosome.
- Pyrrolysine can be inserted at certain UAG codons.

Specific changes in reading the code occur in individual genes. The specificity of such changes implies that the reading of the particular codon must be influenced by the surrounding bases.

A striking example is the incorporation of the modified amino acid seleno-cysteine at certain UGA codons within the genes that code for selenoproteins in both prokaryotes and eukaryotes. Usually these proteins catalyze oxidation-reduction reactions, and contain a single seleno-cysteine residue, which forms part of the active site. The most is known about the use of the UGA codons in three *E. coli* genes coding for formate dehydrogenase isozymes. The internal UGA codon is read by a seleno-Cys-tRNA. This unusual reaction is determined by the local secondary structure of mRNA, in particular by the presence of a hairpin loop downstream of the UGA.

Mutations in 4 *sel* genes create a deficiency in selenoprotein synthesis. *selC* codes for tRNA (with the anticodon ACU$^{\leftarrow}$) that is charged with serine. *selA* and *selD* are required to modify the serine to seleno-cysteine. SelB is an alternative elongation factor. It is a guanine nucleotide-binding protein that acts as a specific translation factor for entry of seleno-Cys-tRNA into the A site; it thus provides (for this single tRNA) a replacement for factor EF-Tu. The sequence of SelB is related to both EF-Tu and IF-2.

Why is seleno-Cys-tRNA inserted only at certain UGA codons? These codons are followed by a stem-loop structure in the mRNA. **Figure 7.12** shows that the stem of this structure is recognized by an additional domain in SelB (one that is not present in EF-Tu or IF-2). A similar mechanism interprets some UGA codons in mammalian cells, except that two proteins are required to identify the appropriate UGA codons. One protein (SBP2) binds a stem-loop structure far downstream from the UGA codon, while the counterpart of SelB (called SECIS) binds to SBP2 and simultaneously binds the tRNA to the UGA codon.

Another example of the insertion of a special amino acid is the placement of pyrrolysine at a UAG codon. This happens in both an archaea and a bacterium. The mechanism is probably similar to the insertion of seleno-cysteine. An unusual tRNA is charged with lysine, which is presumably then modified. The tRNA has a CUA anticodon, which responds to UAG. There must be other components of the system that restricts its response to the appropriate UAG codons.

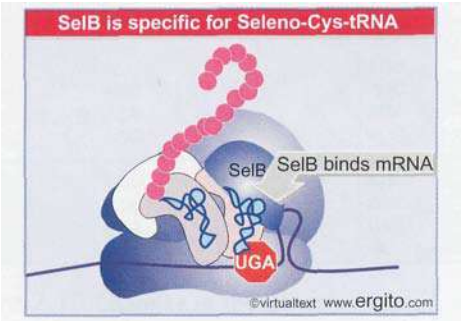


Figure 7.12 SelB is an elongation factor that specifically binds Seleno-Cys-tRNA to a UGA codon that is followed by a stem-loop structure in mRNA.

By Book_Crazy [IND]

7.8 tRNAs are charged with amino acids by synthetases

Key Concepts

- Aminoacyl-tRNA synthetases are enzymes that charge tRNA with an amino acid to generate aminoacyl-tRNA in a two-stage reaction that uses energy from ATP.
- There are 20 aminoacyl-tRNA synthetases in each cell. Each charges all the tRNAs that represent a particular amino acid.
- Recognition of a tRNA is based on a small number of points of contact in the tRNA sequence.

It is necessary for tRNAs to have certain characteristics in common, yet be distinguished by others. The crucial feature that confers this capacity is the ability of tRNA to fold into a specific tertiary structure. Changes in the details of this structure, such as the angle of the two arms of the "L" or the protrusion of individual bases, may distinguish the individual tRNAs.

All tRNAs can fit in the P and A sites of the ribosome, where at one end they are associated with mRNA via codon-anticodon pairing, while at the other end the polypeptide is being transferred. Similarly, all tRNAs (except the initiator) share the ability to be recognized by the translation factors (EF-Tu or eEF1) for binding to the ribosome. The initiator tRNA is recognized instead by IF-2 or eIF2. So the tRNA set must possess common features for interaction with elongation factors, but the initiator tRNA can be distinguished.

Amino acids enter the protein synthesis pathway through the aminoacyl-tRNA synthetases, which provide the interface for connection with nucleic acid. All synthetases function by the two-step mechanism depicted in **Figure 7.13:**

- First, the amino acid reacts with ATP to form aminoacyl~adenylate, releasing pyrophosphate. Energy for the reaction is provided by cleaving the high energy bond of the ATP.
- Then the activated amino acid is transferred to the tRNA, releasing AMP.

The synthetases sort the tRNAs and amino acids into corresponding sets. Each synthetase recognizes a single amino acid and all the tRNAs that should be charged with it. Usually, each amino acid is represented by more than one tRNA. Several tRNAs may be needed to respond to synonym codons, and sometimes there are multiple species of tRNA reacting with the same codon. Multiple tRNAs representing the same amino acid are called isoaccepting tRNAs; because they are all recognized by the same synthetase, they are also described as its cognate tRNAs.

Many attempts to deduce similarities in sequence between cognate tRNAs, or to induce chemical alterations that affect their charging, have shown that the basis for recognition is different for different tRNAs, and does not necessarily lie in some feature of primary or secondary structure alone. We know from the crystal structure that the acceptor stem and the anticodon stem make tight contacts with the synthetase, and mutations that alter recognition of a tRNA are found in these two regions. (The anticodon itself is not necessarily recognized as such; for example, the "suppressor" mutations discussed later in this chapter change a base in the anticodon, and therefore the codons to which a tRNA responds, without altering its charging with amino acids.)

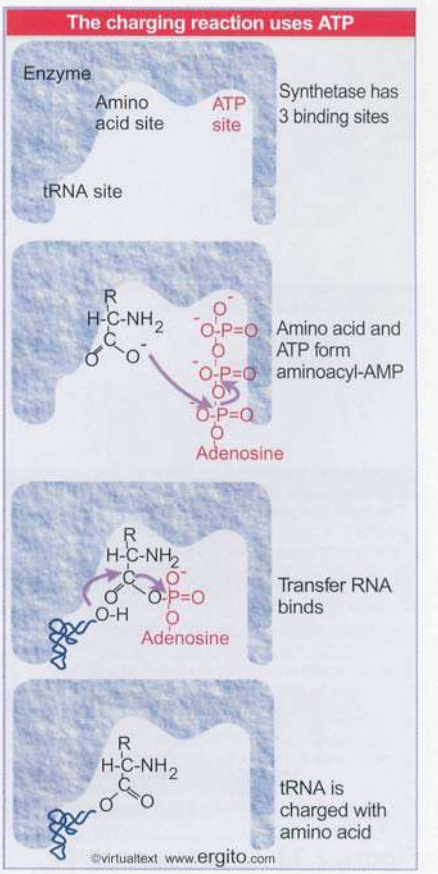


Figure 7.13 An aminoacyl-tRNA synthetase charges tRNA with an amino acid.

A group of isoaccepting tRNAs must be charged only by the single aminoacyl-tRNA synthetase specific for their amino acid. So isoaccepting tRNAs must share some common feature(s) enabling the enzyme to distinguish them from the other tRNAs. The entire complement of tRNAs is divided into 20 isoaccepting groups; each group is able to identify itself to its particular synthetase.

tRNAs are identified by their synthetases by contacts that recognize a small number of bases, typically from 1-5. Three types of feature commonly are used:

- Usually (but not always), at least one base of the anticodon is recognized. Sometimes all the positions of the anticodon are important.
- Often one of the last three base pairs in the acceptor stem is recognized. An extreme case is represented by alanine tRNA, which is identified by a single unique base pair in the acceptor stem.
- The so-called discriminator base, which lies between the acceptor stem and the CCA terminus, is always invariant among isoacceptor tRNAs.

No one of these features constitutes a unique means of distinguishing 20 sets of tRNAs, or provides sufficient specificity, so it appears that recognition of tRNAs is idiosyncratic, each following its own rules.

Several synthetases can specifically charge a "minihelix" consisting only of the acceptor and TψC arms (equivalent to one arm of the L-shaped molecule) with the correct amino acid. For certain tRNAs, specificity depends exclusively upon the acceptor stem. However, it is clear that there are significant variations between tRNAs, and in some cases the anticodon region is important. Mutations in the anticodon can affect recognition by the class II Phe-tRNA synthetase. Multiple features may be involved; minihelices from the $tRNA^{Val}$ and tRNA (where we know that the anticodon is important *in vivo)* can react specifically with their synthetases.

So recognition depends on an interaction between a few points of contact in the tRNA, concentrated at the extremities, and a few amino acids constituting the active site in the protein. The relative importance of the roles played by the acceptor stem and anticodon is different for each tRNA·synthetase interaction.

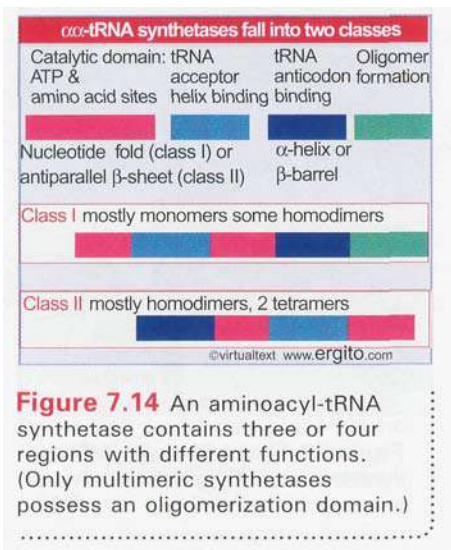


Figure 7.14 An aminoacyl-tRNA synthetase contains three or four regions with different functions. (Only multimeric synthetases possess an oligomerization domain.)

7.9 Aminoacyl-tRNA synthetases fall into two groups

Key Concepts

- Aminoacyl-tRNA synthetases are divided into the class I and class II groups by sequence and structural similarities.

In spite of their common function, synthetases are a rather diverse group of proteins. The individual subunits vary from 40-110 kD, and the enzymes may be monomeric, dimeric, or tetrameric. Homologies between them are rare. Of course, the active site that recognizes tRNA comprises a rather small part of the molecule. It is interesting to compare the active sites of different synthetases.

Synthetases have been divided into two general groups, each containing 10 enzymes, on the basis of the structure of the domain that contains the active site. A general type of organization that applies to

both groups is represented in **Figure 7.14.** The catalytic domain includes the binding sites for ATP and amino acid. It can be recognized as a large region that is interrupted by an insertion of the domain that binds the acceptor helix of the tRNA. This places the terminus of the tRNA in proximity to the catalytic site. A separate domain binds the anticodon region of tRNA. Those synthetases that are multimeric also possess an oligomerization domain.

Class I synthetases have an N-terminal catalytic domain that is identified by the presence of two short, partly conserved sequences of amino acids, sometimes called signature sequences. The catalytic domain takes the form of a motif called a nucleotide-binding fold (which is also found in other classes of enzymes that bind nucleotides). The nucleotide fold consists of alternating parallel β-strands and α-helices; the signature sequence forms part of the ATP-binding site. The insertion that contacts the acceptor helix of tRNA differs widely between different class I enzymes. The C-terminal domains of the class I synthetases, which include the tRNA anticodon-binding domain and any oligomerization domain, also are quite different from one another.

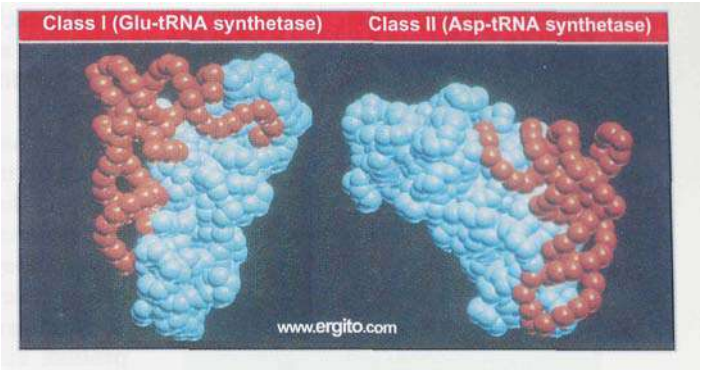


Figure 7.15 Crystal structures show that class I and class II aminoacyl-tRNA synthetases bind the opposite faces of their tRNA substrates. The tRNA is shown in red, and the protein in blue. Photographs kindly provided by Dino Moras.

Class II enzymes share three rather general similarities of sequence in their catalytic domains. The active site contains a large antiparallel β-sheet surrounded by α-helices. Again, the acceptor helix-binding domain that interrupts the catalytic domain has a structure that depends on the individual enzyme. The anticodon-binding domain tends to be N-terminal. The location of any oligomerization domain is widely variable.

The lack of any apparent relationship between the two groups of synthetases is a puzzle. Perhaps they evolved independently of one another. This makes it seem possible even that an early form of life could have existed with proteins that were made up of just the 10 amino acids coded by one type or the other.

A general model for synthetase·tRNA binding suggests that the protein binds the tRNA along the "side" of the L-shaped molecule. The same general principle applies for all synthetase·tRNA binding: the tRNA is bound principally at its two extremities, and most of the tRNA sequence is not involved in recognition by a synthetase. However, the detailed nature of the interaction is different between class I and class II enzymes, as can be seen from the models of **Figure 7.15,** which are based on crystal structures. The two types of enzyme approach the tRNA from opposite sides, with the result that the tRNA-protein models look almost like mirror images of one another.

A class I enzyme (Gln-tRNA synthetase) approaches the D-loop side of the tRNA. It recognizes the minor groove of the acceptor stem at one end of the binding site, and interacts with the anticodon loop at the other end. **Figure 7.16** is a diagrammatic representation of the crystal structure of the $tRNA^{Gln}$·synthetase complex. A revealing feature of the structure is that contacts with the enzyme change the structure of the tRNA at two important points. These can be seen by comparing the dotted and solid lines in the anticodon loop and acceptor stem:

- Bases U35 and U36 in the anticodon loop are pulled farther out of the tRNA into the protein.
- The end of the acceptor stem is seriously distorted, with the result that base pairing between U1 and A72 is disrupted. The single-stranded end of the stem pokes into a deep pocket in the synthetase protein, which also contains the binding site for ATP.

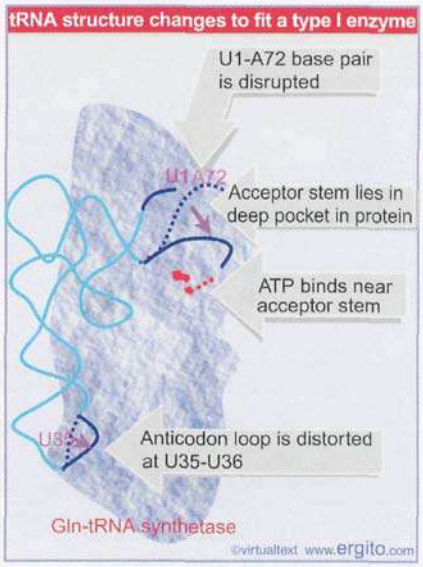


Figure 7.16 A class I tRNA synthetase contacts tRNA at the minor groove of the acceptor stem and at the anticodon.

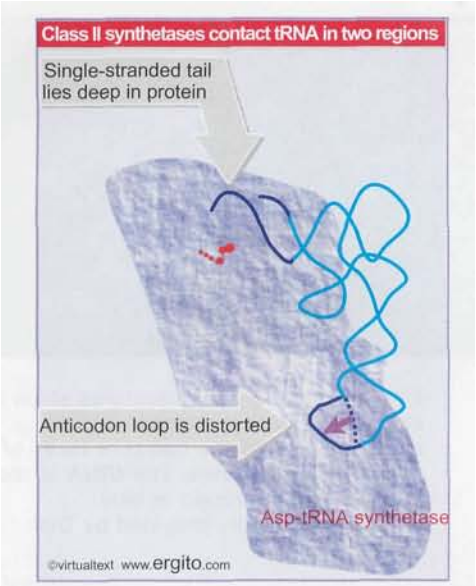


Figure 7.17 A class II aminoacyl-tRNA synthetase contacts tRNA at the major groove of the acceptor helix and at the anticodon loop.

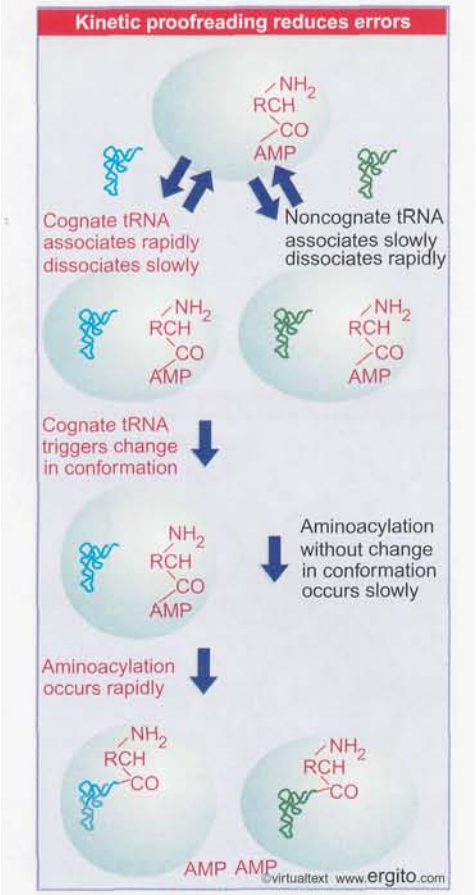


Figure 7.18 Recognition of the correct tRNA by synthetase is controlled at two steps. First, the enzyme has a greater affinity for its cognate tRNA. Second, the aminoacylation of the incorrect tRNA is very slow.

This structure explains why changes in U35, G73, or the U1-A72 base pair affect the recognition of the tRNA by its synthetase. At all of these positions, hydrogen bonding occurs between the protein and tRNA.

A class II enzyme (Asp-tRNA synthetase) approaches the tRNA from the other side, and recognizes the variable loop, and the major groove of the acceptor stem, as drawn in **Figure 7.17.** The acceptor stem remains in its regular helical conformation. ATP is probably bound near to the terminal adenine. At the other end of the binding site, there is a tight contact with the anticodon loop, which has a change in conformation that allows the anticodon to be in close contact with the protein.

7.10 Synthetases use proofreading to improve accuracy

Key Concepts

- Specificity of recognition of both amino acid and tRNA is controlled by aminoacyl-tRNA synthetases by proofreading reactions that reverse the catalytic reaction if the wrong component has been incorporated.

Aminoacyl-tRNA synthetases have a difficult job. Each synthetase must distinguish 1 out of 20 amino acids, and cognate tRNAs (typically 1-3) from the total set (perhaps 100 in all).

Many amino acids are closely related to one another, and all amino acids are related to the metabolic intermediates in their particular synthetic pathway. It is especially difficult to distinguish between two amino acids that differ only in the length of the carbon backbone (that is, by one CH_2 group). Intrinsic discrimination based on relative energies of binding two such amino acids would be only ~1/5. The synthetase enzymes improve this ratio ~1000 fold.

Intrinsic discrimination between tRNAs is better, because the tRNA offers a larger surface with which to make more contacts, but it is still true that all tRNAs conform to the same general structure, and there may be a quite limited set of features that distinguish the cognate tRNAs from the noncognate tRNAs.

We can imagine two general ways in which the enzyme might select its substrate:

- The cycle of admittance, scrutiny, rejection/acceptance could represent a single binding step that precedes all other stages of whatever reaction is involved. This is tantamount to saying that the affinity of the binding site is sufficient to control the entry of substrate. In the case of synthetases, this would mean that only the correct amino acids and cognate tRNAs could form a stable attachment at the site.
- Alternatively, the reaction proceeds through some of its stages, after which a decision is reached on whether the correct species is present. If it is not present, the reaction is reversed, or a bypass route is taken, and the wrong member is expelled. This sort of postbinding scrutiny is generally described as proofreading. In the example of synthetases, it would require that the charging reaction proceeds through certain stages even if the wrong tRNA or amino acid is present.

Synthetases use proofreading mechanisms to control the recognition of both types of substrates. They improve significantly on the intrinsic

differences among amino acids or among tRNAs, but, consistent with the intrinsic differences in each group, make more mistakes in selecting amino acids (error rates are 10^{-4} - 10^{-5}) than in selecting tRNAs (error rates are ~10^{-6}) (see Figure 6.8).

Transfer RNA binds to synthetase by the two stage reaction depicted in **Figure 7.18.** Cognate tRNAs have a greater intrinsic affinity for the binding site, so they are bound more rapidly and dissociate more slowly. Following binding, the enzyme scrutinizes the tRNA that has been bound. If the correct tRNA is present, binding is stabilized by a conformational change in the enzyme. This allows aminoacylation to occur rapidly. If the wrong tRNA is present, the conformational change does not occur. As a result, the reaction proceeds much more slowly; this increases the chance that the tRNA will dissociate from the enzyme before it is charged. This type of control is called kinetic proofreading.

Specificity for amino acids varies among the synthetases. Some are highly specific for initially binding a single amino acid, but others can also activate amino acids closely related to the proper substrate. Although the analog amino acid can sometimes be converted to the adenylate form, in none of these cases is an incorrectly activated amino acid actually used to form a stable aminoacyl-tRNA.

The presence of the cognate tRNA usually is needed to trigger proofreading, even if the reaction occurs at the stage before formation of aminoacyl-adenylate. (An exception is provided by Met-tRNA synthetase, which can reject noncognate aminoacyl-adenylate complexes even in the absence of tRNA.)

There are two stages at which proofreading of an incorrect aminoacyl-adenylate may occur during formation of aminoacyl-tRNA. **Figure 7.19** shows that both use chemical proofreading, in which the catalytic reaction is reversed. The extent to which one pathway or the other predominates varies with the individual synthetase:

- The noncognate aminoacyl-adenylate may be hydrolyzed when the cognate tRNA binds. This mechanism is used predominantly by several synthetases, including those for methionine, isoleucine, and valine. (Usually, the reaction cannot be seen *in vivo*, but it can be followed for Met-tRNA synthetase when the incorrectly activated amino acid is homocysteine, which lacks the methyl group of methionine). Proofreading releases the amino acid in an altered form, as homocysteine thiolactone. In fact, homocysteine thiolactone is produced in *E. coli* as a by-product of the charging reaction of Met-tRNA synthetase. This shows that continuous proofreading is part of the process of charging a tRNA with its amino acid.
- Some synthetases use chemical proofreading at a later stage. The wrong amino acid is actually transferred to tRNA, is then recognized as incorrect by its structure in the tRNA binding site, and so is hydrolyzed and released. The process requires a continual cycle of linkage and hydrolysis until the correct amino acid is transferred to the tRNA.

A classic example in which discrimination between amino acids depends on the presence of tRNA is provided by the Ile-tRNA synthetase of *E. coli*. The enzyme can charge valine with AMP, but hydrolyzes the valyl-adenylate when $tRNA^{Ile}$ is added. The overall error rate depends on the specificities of the individual steps, as summarized in **Figure 7.20.** The overall error rate of 1.5 X 10^{-5} is less than the measured rate at which valine is substituted for isoleucine (in rabbit globin), which is 2-5 X 10^{-4}. So mischarging probably provides only a small fraction of the errors that actually occur in protein synthesis.

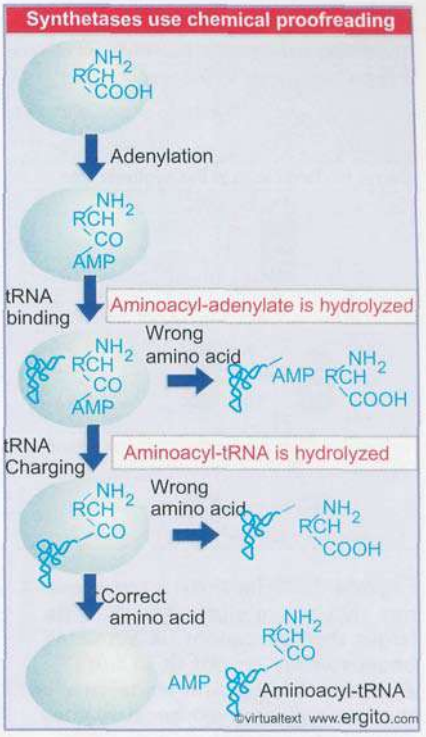


Figure 7.19 When a synthetase binds the incorrect amino acid, proofreading requires binding of the cognate tRNA. It may take place either by a conformation change that causes hydrolysis of the incorrect aminoacyl-adenylate, or by transfer of the amino acid to tRNA, followed by hydrolysis.

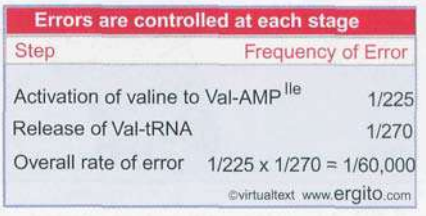


Figure 7.20 The accuracy of charging $tRNA^{Ile}$ by its synthetase depends on error control at two stages.

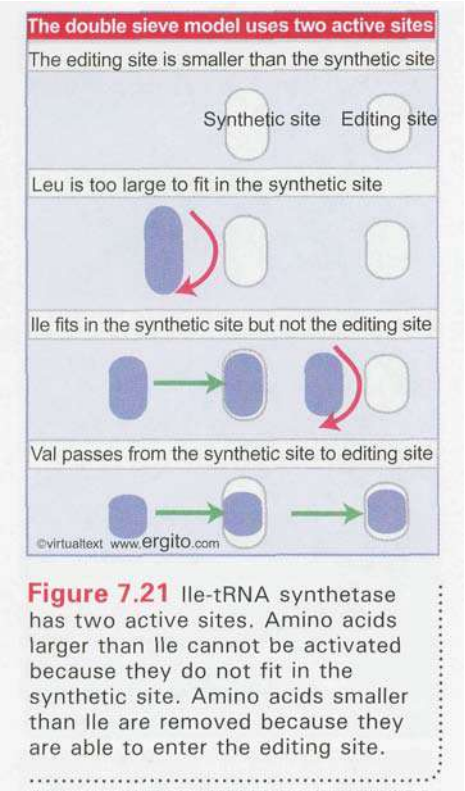


Figure 7.21 Ile-tRNA synthetase has two active sites. Amino acids larger than Ile cannot be activated because they do not fit in the synthetic site. Amino acids smaller than Ile are removed because they are able to enter the editing site.

Ile-tRNA synthetase uses size as a basis for discrimination among amino acids. **Figure 7.21** shows that it has two active sites: the synthetic (or activation) site and the editing (or hydrolytic) site. The crystal structure of the enzyme shows that the synthetic site is too small to allow leucine (a close analog of isoleucine) to enter. All amino acids large than isoleucine are excluded from activation because they cannot enter the synthetic site. An amino acid that can enter the synthetic site is placed on tRNA. Then the enzyme tries to transfer it to the editing site. Isoleucine is safe from editing because it is too large to enter the editing site. However, valine can enter this site, and as a result an incorrect Val-$tRNA^{Ile}$ is hydrolyzed. Essentially the enzyme provides a double molecular sieve, in which size of the amino acid is used to discriminate between closely related species.

One interesting feature of Ile-tRNA synthetase is that the synthetic and editing sites are a considerable distance apart, ~34 A. A crystal structure of the enzyme complexed with an edited analog of isoleucine shows that the amino acid is transported from the synthetic site to the editing site. **Figure 7.22** shows that this involves a change in the conformation of the tRNA. The amino acid acceptor stem of $tRNA^{Ile}$ can exist in alternative conformations. It adopts an unusual hairpin in order to be aminoacylated by an amino acid in the synthetic site. Then it returns to the more common helical structure in order to move the amino acid to the editing site. The translocation between sites is the rate-limiting step in proofreading. Ile-tRNA synthetase is a class I synthetase, but the double sieve mechanism is used also by class II synthetases.

7.11 Suppressor tRNAs have mutated anticodons that read new codons

Key Concepts

- A suppressor tRNA typically has a mutation in the anticodon that changes the codons to which it responds.
- When the new anticodon corresponds to a termination codon, an amino acid is inserted and the polypeptide chain is extended beyond the termination codon. This results in nonsense suppression at a site of nonsense mutation or in readthrough at a natural termination codon.
- Missense suppression occurs when the tRNA recognizes a different codon from usual, so that one amino acid is substituted for another.

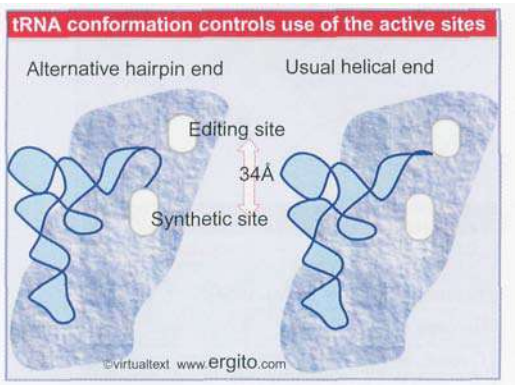


Figure 7.22 An amino acid is transported from the synthetic site to the editing site of Ile-tRNA synthetase by a change in the conformation of the amino acceptor stem of tRNA.

Isolation of mutant tRNAs has been one of the most potent tools for analyzing the ability of a tRNA to respond to its codon(s) in mRNA and for determining the effects that different parts of the tRNA molecule have on codon-anticodon recognition.

Mutant tRNAs are isolated by virtue of their ability to overcome the effects of mutations in genes coding for proteins. In general genetic terminology, a mutation that is able to overcome the effects of another mutation is called a suppressor.

In tRNA suppressor systems, the primary mutation changes a codon in an mRNA so that the protein product is no longer functional. The secondary, suppressor mutation changes the anticodon of a tRNA, so that it recognizes the mutant codon instead of (or as well as) its original target codon. The amino acid that is now inserted restores protein function. The suppressors are described as nonsense suppressors or missense suppressors, depending on the nature of the original mutation.

By Book_Crazy [IND]

In a wild-type cell, a nonsense mutation is recognized only by a release factor, terminating protein synthesis. The suppressor mutation creates an aminoacyl-tRNA that can recognize the termination codon; by inserting an amino acid, it allows protein synthesis to continue beyond the site of nonsense mutation. This new capacity of the translation system allows a full-length protein to be synthesized, as illustrated in **Figure 7.23.** If the amino acid inserted by suppression is different from the amino acid that was originally present at this site in the wild-type protein, the activity of the protein may be altered.

Missense mutations change a codon representing one amino acid into a codon representing another amino acid, one that cannot function in the protein in place of the original residue. (Formally, any substitution of amino acids constitutes a missense mutation, but in practice it is detected only if it changes the activity of the protein.) The mutation can be suppressed by the insertion either of the original amino acid or of some other amino acid that is acceptable to the protein.

Figure 7.24 demonstrates that missense suppression can be accomplished in the same way as nonsense suppression, by mutating the anticodon of a tRNA carrying an acceptable amino acid so that it responds to the mutant codon. So missense suppression involves a change in the meaning of the codon from one amino acid to another.

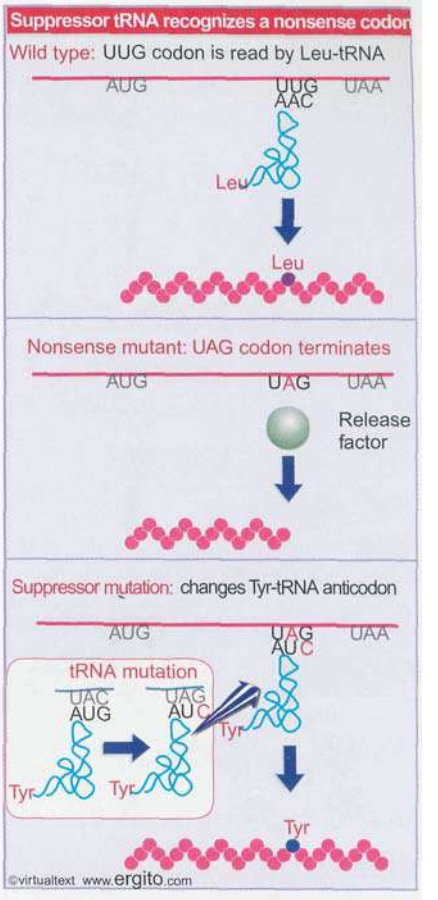


Figure 7.23 Nonsense mutations can be suppressed by a tRNA with a mutant anticodon, which inserts an amino acid at the mutant codon, producing a full length protein in which the original Leu residue has been replaced by Tyr.

7.12 There are nonsense suppressors for each termination codon

Key Concepts

- Each type of nonsense codon is suppressed by tRNAs with mutant anticodons.
- Some rare suppressor tRNAs have mutations in other parts of the molecule.

Nonsense suppressors fall into three classes, one for each type of termination codon. **Figure 7.25** describes the properties of some of the best characterized suppressors.

The easiest to characterize have been amber suppressors. In *E. coli*, at least 6 tRNAs have been mutated to recognize UAG codons. All of the amber suppressor tRNAs have the anticodon $CUA^{\leftarrow}$, in each case derived from wild type by a single base change. The site of mutation can be any one of the three bases of the anticodon, as seen from *supD*, *supE*, and *supF*. Each suppressor tRNA recognizes only the UAG codon, instead of its former codon(s). The amino acids inserted are serine, glutamine, or tyrosine, the same as those carried by the corresponding wild-type tRNAs.

Ochre suppressors also arise by mutations in the anticodon. The best known are *supC* and *supG*, which insert tyrosine or lysine in response to both ochre (UAA) and amber (UAG) codons. This conforms with the prediction of the wobble hypothesis that UAA cannot be recognized alone.

A UGA suppressor has an unexpected property. It is derived from $tRNA^{Trp}$, but its only mutation is the substitution of A in place of G at position 24. This change replaces a G·U pair in the D stem with an A·U pair, increasing the stability of the helix. The sequence of the anticodon remains the same as the wild type, $CCA^{\leftarrow}$. So the mutation in the D stem must in some way alter the conformation of the anticodon loop,

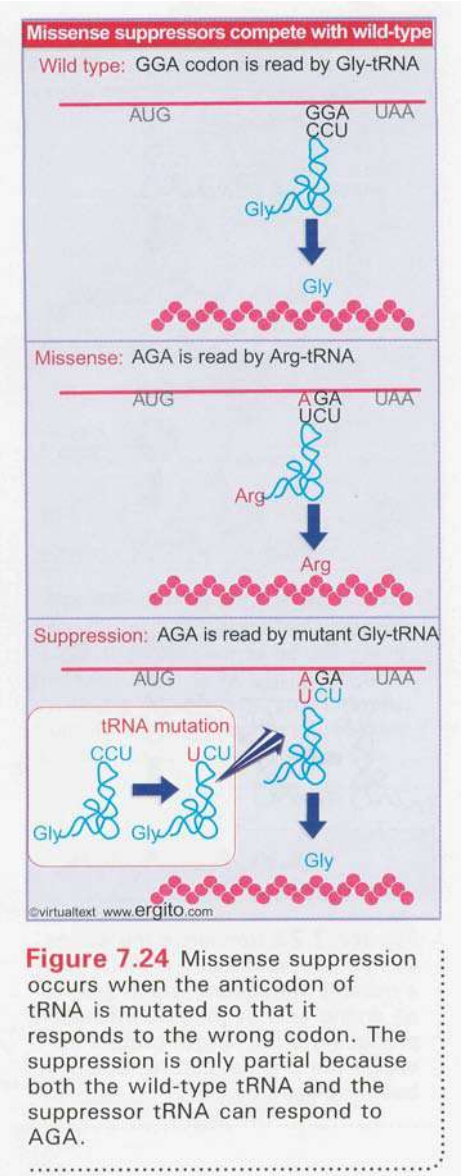


Figure 7.24 Missense suppression occurs when the anticodon of tRNA is mutated so that it responds to the wrong codon. The suppression is only partial because both the wild-type tRNA and the suppressor tRNA can respond to AGA.

Suppressors have anticodon mutations

Locus	tRNA	Wild Type Codon/Anti		Suppressor Anti/Codon	
supD (su1)	Ser	UCG	CGA	CUA	UAG
supE (su2)	Gln	CAG	CUG	CUA	UAG
supF (su3)	Tyr	UA C/U	GUA	CUA	UAG
supC (su4)	Tyr	UA C/U	GUA	UUA	UA A/G
supG (su5)	Lys	AA A/G	UUU	UUA	UA A/G
supU (su7)	Trp	UGG	CCA	UCA	UG A/G

©virtualtext www.ergito.com

Figure 7.25 Nonsense suppressor tRNAs are generated by mutations in the anticodon.

allowing CCA← to pair with UGA in an unusual wobble pairing of C with A. The suppressor tRNA continues to recognize its usual codon, UGG.

A related response is seen with a eukaryotic tRNA. Bovine liver contains a $tRNA^{Ser}$ with the anticodon $^{m}CCA←$. The wobble rules predict that this tRNA should respond to the tryptophan codon UGG; but in fact it responds to the termination codon UGA. So it is possible that UGA is suppressed naturally in this situation.

The general importance of these observations lies in the demonstration that codon-anticodon recognition of either wild-type or mutant tRNA cannot be predicted entirely from the relevant triplet sequences, but is influenced by other features of the molecule.

7.13 Suppressors may compete with wild-type reading of the code

Key Concepts

- Suppressor tRNAs compete with wild-type tRNAs that have the same anticodon to read the corresponding codon(s).
- Efficient suppression is deleterious because it results in readthrough past normal termination codons.
- The UGA codon is leaky and is misread by Trp-tRNA at 1-3% frequency.

There is an interesting difference between the usual recognition of a codon by its proper aminoacyl-tRNA and the situation in which mutation allows a suppressor tRNA to recognize a new codon. In the wild-type cell, only one meaning can be attributed to a given codon, which represents either a particular amino acid or a signal for termination. But in a cell carrying a suppressor mutation, the mutant codon has the alternatives of being recognized by the suppressor tRNA or of being read with its usual meaning.

A nonsense suppressor tRNA must compete with the release factors that recognize the termination codon(s). A missense suppressor tRNA must compete with the tRNAs that respond properly to its new codon. The extent of competition influences the efficiency of suppression; so the effectiveness of a particular suppressor depends not only on the affinity between its anticodon and the target codon, but also on its concentration in the cell, and on the parameters governing the competing termination or insertion reactions.

The efficiency with which any particular codon is read is influenced by its location. So the extent of nonsense suppression by a given tRNA can vary quite widely, depending on the context of the codon. We do not understand the effect that neighboring bases in mRNA have on codon-anticodon recognition, but the context can change the frequency with which a codon is recognized by a particular tRNA by more than an order of magnitude. The base on the 3' side of a codon appears to have a particularly strong effect.

A nonsense suppressor is isolated by its ability to respond to a mutant nonsense codon. But the same triplet sequence constitutes one of the normal termination signals of the cell! The mutant tRNA that suppresses the nonsense mutation must in principle be able to suppress natural termination at the end of any gene that uses this codon. **Figure 7.26** shows that this readthrough results in the synthesis of a longer protein, with additional C-terminal material. The extended protein will end at the next termination triplet sequence found in the phase of the reading frame. Any

extensive suppression of termination is likely to be deleterious to the cell by producing extended proteins whose functions are thereby altered.

Amber suppressors tend to be relatively efficient, usually in the range of 10-50%, depending on the system. This efficiency is possible because amber codons are used relatively infrequently to terminate protein synthesis in *E. coli*.

Ochre suppressors are difficult to isolate. They are always much less efficient, usually with activities below 10%. All ochre suppressors grow rather poorly, which indicates that suppression of both UAA and UAG is damaging to *E. coli,* probably because the ochre codon is used most frequently as a natural termination signal.

UGA is the least efficient of the termination codons in its natural function; it is misread by Trp-tRNA as frequently as 1-3% in wild-type situations. In spite of this deficiency, however, it is used more commonly than the amber triplet to terminate bacterial genes.

One gene's missense suppressor is likely to be another gene's mutator. A suppressor corrects a mutation by substituting one amino acid for another at the mutant site. But in other locations, the same substitution will replace the wild-type amino acid with a new amino acid. The change may inhibit normal protein function.

This poses a dilemma for the cell: it must suppress what is a mutant codon at one location, while failing to change too extensively its normal meaning at other locations. The absence of any strong missense suppressors is therefore explained by the damaging effects that would be caused by a general and efficient substitution of amino acids.

A mutation that creates a suppressor tRNA can have two consequences. First, it allows the tRNA to recognize a new codon. Second, sometimes it prevents the tRNA from recognizing the codons to which it previously responded. It is significant that all the high-efficiency amber suppressors are derived by mutation of one copy of a redundant tRNA set. In these cases, the cell has several tRNAs able to respond to the codon originally recognized by the wild-type tRNA. So the mutation does not abolish recognition of the old codons, which continue to be served adequately by the tRNAs of the set. In the unusual situation in which there is only a single tRNA that responds to a particular codon, any mutation that prevents the response is lethal.

Suppression is most often considered in the context of a mutation that changes the reading of a codon. However, there are some situations in which a stop codon is read as an amino acid at a low frequency in the wild-type situation. The first example to be discovered was the coat protein gene of the RNA phage Qβ. The formation of infective Qβ particles requires that the stop codon at the end of this gene is suppressed at a low frequency to generate a small proportion of coat proteins with a C-terminal extension. In effect, this stop codon is leaky. The reason is that Trp-tRNA recognizes the codon at a low frequency.

Readthrough past stop codons occurs also in eukaryotes, where it is employed most often by RNA viruses. This may involve the suppression of UAG/UAA by Tyr-tRNA, Gln-tRNA, or Leu-tRNA, or the suppression of UGA by Trp-tRNA or Arg-tRNA. The extent of partial suppression is dictated by the context surrounding the codon.

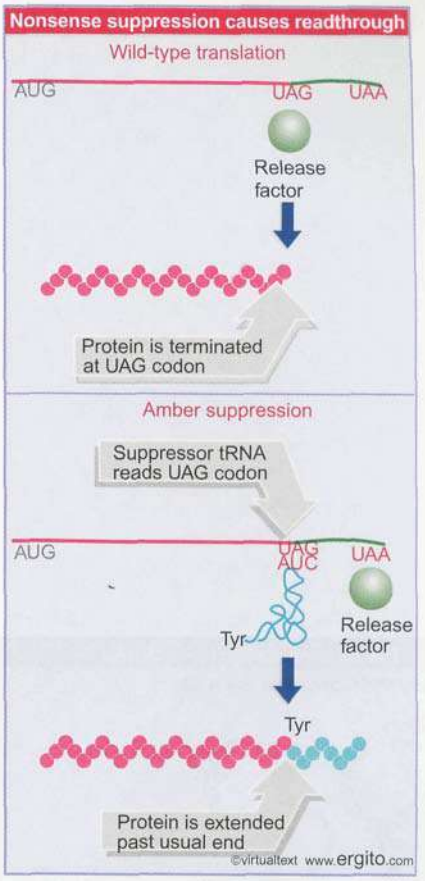


Figure 7.26 Nonsense suppressors also read through natural termination codons, synthesizing proteins that are longer than wild-type.

7.14 The ribosome influences the accuracy of translation

Key Concepts

- The structure of the 16S rRNA at the P and A sites of the ribosome influences the accuracy of translation.

The lack of detectable variation when the sequence of a protein is analyzed demonstrates that protein synthesis must be extremely accurate. Very few mistakes are apparent in the form of substitutions of one amino acid for another. There are two general stages in protein synthesis at which errors might be made (see Figure 6.8 in *6.3 Special mechanisms control the accuracy of protein synthesis):*

- Charging a tRNA only with its correct amino acid clearly is critical. This is a function of the aminoacyl-tRNA synthetase. Probably the error rate varies with the particular enzyme, but generally mistakes occur in $<1/10^5$ aminoacylations.
- The specificity of codon-anticodon recognition is crucial, but puzzling. Although binding constants vary with the individual codon-anticodon reaction, the specificity is always much too low to provide an error rate of $<10^{-5}$. When free in solution, tRNAs bind to their trinucleotide codon sequences only rather weakly. Related, but erroneous, triplets (with two correct bases out of three) are recognized 10^{-1}–10^{-2} times as efficiently as the correct triplets.

Codon-anticodon base pairing therefore seems to be a weak point in the accuracy of translation. The ribosome has an important role in controlling the specificity of this interaction, functioning directly or indirectly as a "proofreader," to distinguish correct and incorrect codon-anticodon pairs, and thus amplifying the rather modest intrinsic difference by ~1000×. And in addition to the role of the ribosome itself, the factors that place initiator- and aminoacyl-tRNAs in the ribosome also may influence the pairing reaction.

So there must be some mechanism for stabilizing the correct aminoacyl-tRNA, allowing its amino acid to be accepted as a substrate for receipt of the polypeptide chain; contacts with an incorrect aminoacyl-tRNA must be rapidly broken, so that the complex leaves without reacting. Suppose that there is no specificity in the initial collision between the aminoacyl-tRNA·EF-Tu·GTP complex and the ribosome. If any complex, irrespective of its tRNA, can enter the A site, the number of incorrect entries must far exceed the number of correct entries.

There are two basic models for how the ribosome might discriminate between correctly and incorrectly paired aminoacyl-tRNAs. The actual situation incorporates elements of both models.

- The direct recognition model supposes that the structure of the ribosome is designed to recognize aminoacyl-tRNAs that are correctly paired. This would mean that the correct pairing results in some small change in the conformation of the aminoacyl-tRNA that the ribosome can recognize. Discrimination occurs before any further reaction occurs.
- The kinetic proofreading model proposes that there are two (or more) stages in the process, so that the aminoacyl-tRNA has multiple opportunities to disengage. An incorrectly paired aminoacyl-tRNA may pass through some stages of the reaction before it is rejected. Overall selectivity can in principle be the product of the selectivities at each stage.

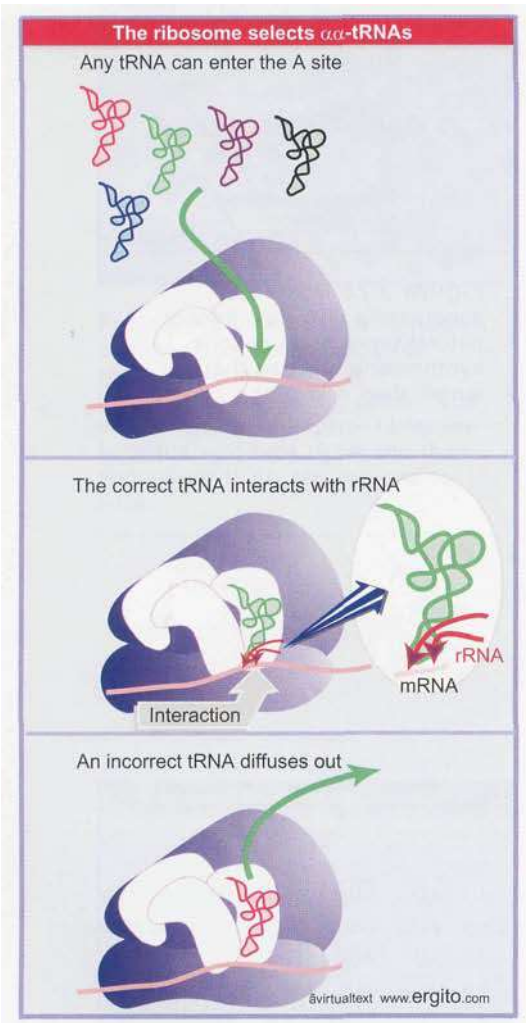


Figure 7.27 Any aminoacyl-tRNA can be placed in the A site (by EF-Tu), but only one that pairs with the anticodon ca

Figure 7.27 illustrates diagrammatically what happens to correctly and incorrectly paired aminoacyl-tRNAs. A correctly paired aminoacyl-tRNA is able to make stabilizing contacts with rRNA. An incorrectly paired aminoacyl-tRNA does not make these contacts, and therefore is able to diffuse out of the A site.

The path to discovering these interactions started with investigations of the effects of the antibiotic streptomycin in the 1960s. Streptomycin

By Book_Crazy [IND]

inhibits protein synthesis by binding to 16S rRNA and inhibiting the ability of EF-G to catalyze translocation. It also increases the level of misreading of the pyrimidines U and C (usually one is mistaken for the other, occasionally for A). The site at which streptomycin acts is influenced by the S12 protein; the sequence of this protein is altered in resistant mutants. Ribosomes with an S12 protein derived from resistant bacteria show a reduction in the level of misreading compared with wild-type ribosomes. In effect, S12 controls the level of misreading. When it is mutated to decrease misreading, it suppresses the effect of streptomycin.

S12 stabilizes the structure of 16S rRNA in the region that is bound by streptomycin. *The important point to note here is that the P/A site region influences the accuracy of translation: translation can be made more or less accurate by changing the structure of 16S rRNA.* The combination of the effects of the S12 protein and streptomycin on the rRNA structure explains the behavior of different mutants in S12, some of which even make the ribosome *dependent* on the presence of streptomycin for correct translation.

We now know from the crystal structure of the ribosome that 16S rRNA is in a position to make contacts with aminoacyl-tRNA. Two bases of 16S rRNA can contact the minor groove of the helix formed by pairing between the anticodon in tRNA with the first two bases of the codon in mRNA. This directly stabilizes the structure when the correct codon-anticodon contacts are made at the first two codon positions, but it does not monitor contacts at the third position.

The stabilization of correctly paired aminoacyl-tRNA may have two effects. By holding the aminoacyl-tRNA in the A site, it prevents it from escaping before the next stage of protein synthesis. And the conformational change in the rRNA may help to trigger the next stage of the reaction, which is the hydrolysis of GTP by EF-Tu.

Part of the proofreading effect is determined by timing. An aminoacyl-tRNA in the A site may in effect be trapped if the next stage of protein synthesis occurs while it is there. So a delay between entry into the A site and peptidyl transfer may give more opportunity for a mismatched aminoacyl-tRNA to dissociate. Mismatched aminoacyl-tRNA dissociates more rapidly than correctly matched aminoacyl-tRNA, probably by a factor of ~5×. Its chance of escaping is therefore increased when the peptide transfer step is slowed.

The specificity of decoding has been assumed to reside with the ribosome itself, but some recent results suggest that translation factors influence the process at both the P site and A site. An indication that EF-Tu is involved in maintaining the reading frame is provided by mutants of the factor that suppress frameshifting. This implies that EF-Tu does not merely bring aminoacyl-tRNA to the A site, but also is involved in positioning the incoming aminoacyl-tRNA relative to the peptidyl-tRNA in the P site.

A striking case where factors influence meaning is found at initiation. Mutation of the AUG initiation codon to UUG in the yeast gene *HIS4* prevents initiation. Extragenic suppressor mutations can be found that allow protein synthesis to be initiated at the mutant UUG codon. Two of these suppressors prove to be in genes coding for the a and β subunits of eIF2, the factor that binds Met-tRNAj to the P site. The mutation in eIFβ2 resides in a part of the protein that is almost certainly involved in binding nucleic acid. It seems likely that its target is either the initiation sequence of mRNA as such or the base-paired association between the mRNA codon and $tRNA_i^{Met}$ anticodon. This suggests that eIF2 participates in the discrimination of initiation codons as well as bringing the initiator tRNA to the P site.

The cost of protein synthesis in terms of high-energy bonds may be increased by proofreading processes. An important question in

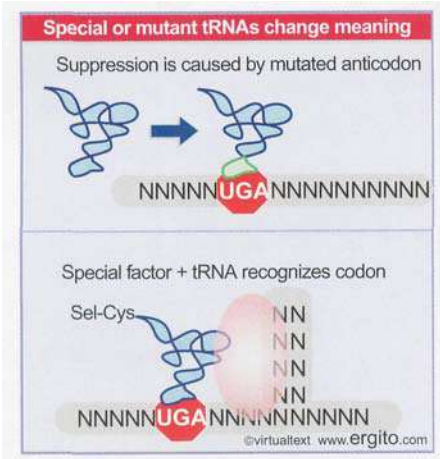


Figure 7.28 A mutation in an individual tRNA (usually in the anticodon) can suppress the usual meaning of that codon. In a special case, a specific tRNA is bound by an unusual elongation factor to recognize a termination codon adjacent to a hairpin loop.

calculating the cost of protein synthesis is the stage at which the decision is taken on whether to accept a tRNA. If a decision occurs immediately to release an aminoacyl-tRNA·EF-Tu·GTP complex, there is little extra cost for rejecting the large number of incorrect tRNAs that are likely (statistically) to enter the A site before the correct tRNA is recognized. But if GTP is hydrolyzed before the mismatched aminoacyl-tRNA dissociates, the cost will be greater. A mismatched aminoacyl-tRNA can be rejected either before or after the cleavage of GTP, although we do not know yet where on average it is rejected. There is some evidence that the use of GTP *in vivo* is greater than the three high-energy bonds that are used in adding every (correct) amino acid to the chain.

7.15 Recoding changes codon meanings

Key Concepts

- Changes in codon meaning can be caused by mutant tRNAs or by tRNAs with special properties.
- The reading frame can be changed by frameshifting or bypassing, both of which depend on properties of the mRNA.

The reading frame of a messenger usually is invariant. Translation starts at an AUG codon and continues in triplets to a termination codon. Reading takes no notice of sense: insertion or deletion of a base causes a frameshift mutation, in which the reading frame is changed beyond the site of mutation. Ribosomes and tRNAs continue inexorably in triplets, synthesizing an entirely different series of amino acids.

There are some exceptions to the usual pattern of translation that enable a reading frame with an interruption of some sort—such as a nonsense codon or frameshift—to be translated into a full-length protein. Recoding events are responsible for making exceptions to the usual rules, and can involve several types of events.

Changing the meaning of a single codon allows one amino acid to be substituted in place of another, or for an amino acid to be inserted at a termination codon. **Figure 7.28** shows that these changes rely on the properties of an individual tRNA that responds to the codon:

- Suppression involves recognition of a codon by a (mutant) tRNA that usually would respond to a different codon (see 7.77 *Suppressor tRNAs have mutated anticodons that read new codons).*
- Redefinition of the meaning of a codon occurs when an aminoacyl-tRNA is modified (see *7.7 Novel amino acids can be inserted at certain stop codons).*

Changing the reading frame occurs in two types of situation:

- Frameshifting typically involves changing the reading frame when aminoacyl-tRNA slips by one base (+1 forward or -1 backward) (see next section). The result shown in **Figure 7.29** is that translation continues past a termination codon.
- Bypassing involves a movement of the ribosome to change the codon that is paired with the peptidyl-tRNA in the P site. The sequence between the two codons fails to be represented in protein. As shown in **Figure 7.30,** this allows translation to continue past any termination codons in the intervening region.

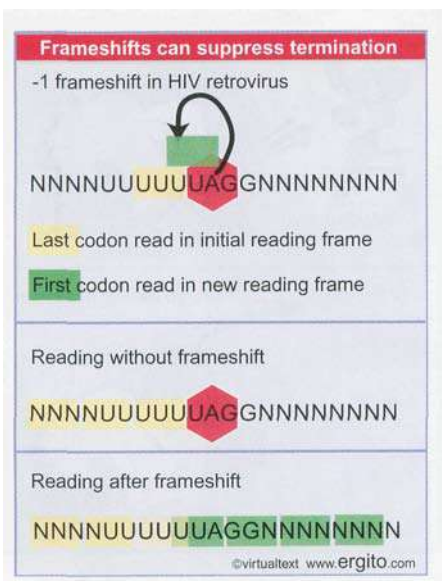


Figure 7.29 A tRNA that slips one base in pairing with a codon causes a frameshift that can suppress termination. The efficiency is usually ~5%.

7.16 Frameshifting occurs at slippery sequences

Key Concepts

- The reading frame may be influenced by the sequence of mRNA and the ribosomal environment.
- Slippery sequences allow a tRNA to shift by 1 base after it has paired with its anticodon, thereby changing the reading frame.
- Translation of some genes depends upon the regular occurrence of programmed frameshifting.

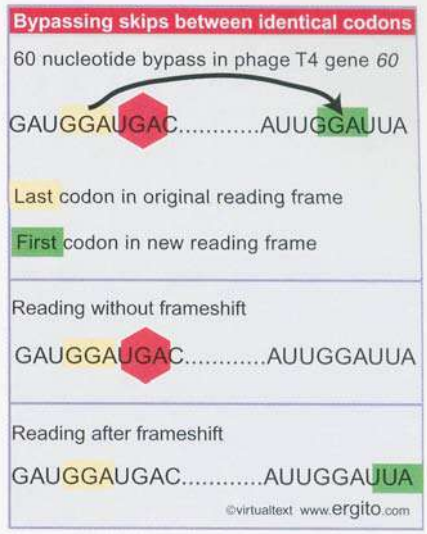


Figure 7.30 Bypassing occurs when the ribosome moves along mRNA so that the peptidyl-tRNA in the P site is released from pairing with its codon and then repairs with another codon farther along.

Frameshifting is associated with specific tRNAs in two circumstances:

' Some mutant tRNA suppressors recognize a "codon" for 4 bases instead of the usual 3 bases.
• Certain "slippery" sequences allow a tRNA to move a base up or down mRNA in the A site.

Frameshift mutants result from the insertion or deletion of a base. They can be suppressed by restoring the original reading frame. This can be achieved by compensating base deletions and insertions within a gene (see *1.21 The genetic code is triplet).* However, extragenic frameshift suppressors also can be found in the form of tRNAs with aberrant properties.

The simplest type of external frameshift suppressor corrects the reading frame when a mutation has been caused by inserting an additional base within a stretch of identical residues. For example, a G may be inserted in a run of several contiguous G bases. The frameshift suppressor is a $tRNA^{Gly}$ that has an extra base inserted in its anticodon loop, converting the anticodon from the usual triplet sequence $CCC^{\leftarrow}$ to the quadruplet sequence $CCCC^{\leftarrow}$. The suppressor tRNA recognizes a 4-base "codon".

Some frameshift suppressors can recognize more than one 4-base "codon". For example, a bacterial $tRNA^{Lys}$ suppressor can respond to either AAAA or AAAU, instead of the usual codon AAA. Another suppressor can read any 4-base "codon" with ACC in the first three positions; the next base is irrelevant. In these cases, the alternative bases that are acceptable in the fourth position of the longer "codon" are not related by the usual wobble rules. The suppressor tRNA probably recognizes a 3-base codon, but for some other reason—most likely steric hindrance—the adjacent base is blocked. This forces one base to be skipped before the next tRNA can find a codon.

Situations in which frameshifting is a normal event are presented by phages and viruses. Such events may affect the continuation or termination of protein synthesis, and result from the intrinsic properties of the mRNA.

In retroviruses, translation of the first gene is terminated by a nonsense codon in phase with the reading frame. The second gene lies in a different reading frame, and (in some viruses) is translated by a frameshift that changes into the second reading frame and therefore bypasses the termination codon (see Figure 7.29) (see *17.3 Retroviral genes codes for polyproteins).* The efficiency of the frameshift is low, typically ~5%. In fact, this is important in the biology of the virus; an increase in efficiency can be damaging. **Figure 7.31** illustrates the similar situation of the yeast Ty element, in which the termination codon of *tya* must be bypassed by a frameshift in order to read the subsequent *tyb* gene.

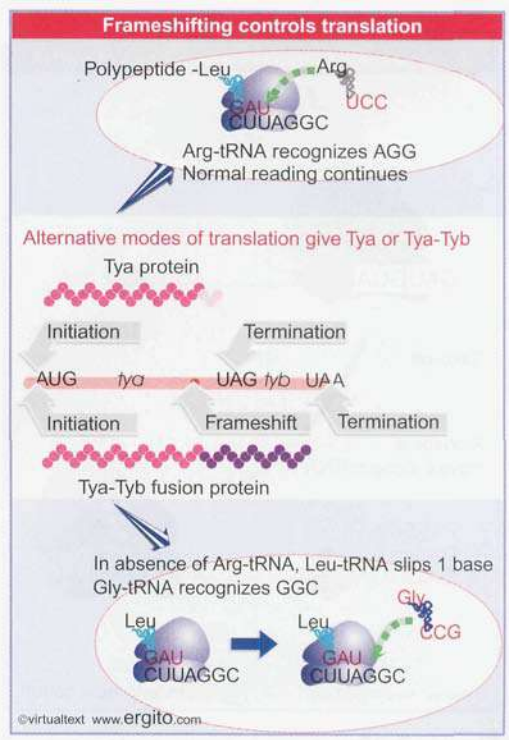


Figure 7.31 A +1 frameshift is required for expression of the *tyb* gene of the yeast Ty element. The shift occurs at a 7-base sequence at which two Leu codon(s) are followed by a scarce Arg codon.

Such situations makes the important point that the rare (but predictable) occurrence of "misreading" events can be relied on as a necessary step in natural translation. This is called programmed frameshifting. It occurs at particular sites at frequencies that are 100-1000× greater than the rate at which errors are made at nonprogrammed sites (~3×10^{-5} per codon).

There are two common features in this type of frameshifting:

- A "slippery" sequence allows an aminoacyl-tRNA to pair with its codon and then to move +1 (rare) or -1 base (more common) to pair with an overlapping triplet sequence that can also pair with its anticodon.
- The ribosome is delayed at the frameshifting site to allow time for the aminoacyl-tRNA to rearrange its pairing. The cause of the delay can be an adjacent codon that requires a scarce aminoacyl-tRNA, a termination codon that is recognized slowly by its release factor, or a structural impediment in mRNA (for example, a "pseudoknot," a particular conformation of RNA) that impedes the ribosome.

Slippery events can involve movement in either direction; a -1 frameshift is caused when the tRNA moves backwards, and a +1 frameshift is caused when it moves forwards. In either case, the result is to expose an out-of-phase triplet in the A site for the next aminoacyl-tRNA. The frameshifting event occurs before peptide bond synthesis. In the most common type of case, when it is triggered by a slippery sequence in conjunction with a downstream hairpin in mRNA, the surrounding sequences influence its efficiency.

The frameshifting in Figure 7.31 shows the behavior of a typical slippery sequence. The 7-nucleotide sequence CUUAGGC is usually recognized by Leu-tRNA at CUU followed by Arg-tRNA at AGC. However, the Arg-tRNA is scarce, and when its scarcity results in a delay, the Leu-tRNA slips from the CUU codon to the overlapping UUA triplet. This causes a frameshift, because the next triplet in phase with the new pairing (GGC) is read by Gly-tRNA. Slippage usually occurs in the P site (when the Leu-tRNA actually has become peptidyl-tRNA, carrying the nascent chain).

Frameshifting at a stop codon causes readthrough of the protein. The base on the 3′ side of the stop codon influences the relative frequencies of termination and frameshifting, and thus affects the efficiency of the termination signal. This helps to explain the significance of context on termination.

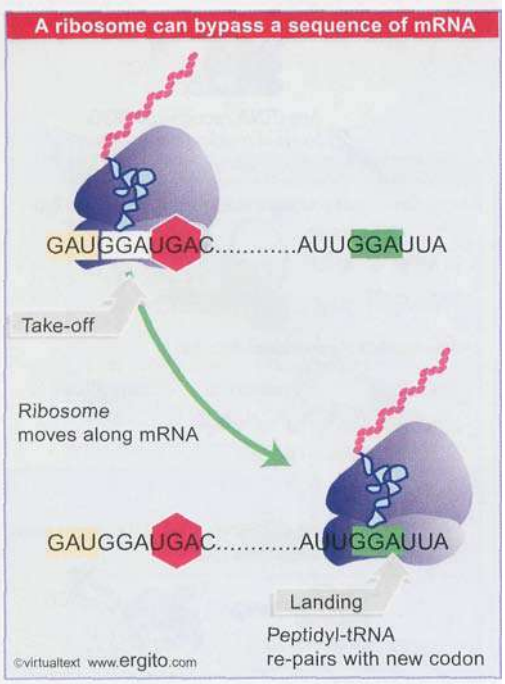


Figure 7.32 In bypass mode, a ribosome with its P site occupied can stop translation. It slides along mRNA to a site where peptidyl-tRNA pairs with a new codon in the P site. Then protein synthesis is resumed.

7.17 Bypassing involves ribosome movement

Certain sequences trigger a bypass event, when a ribosome stops translation, slides along mRNA with peptidyl-tRNA remaining in the P site, and then resumes translation (see Figure 7.30). This is a rather rare phenomenon, with only ~3 authenticated examples. The most dramatic example of bypassing is in gene *60* of phage T4, where the ribosome moves 60 nucleotides along the mRNA.

The key to the bypass system is that there are identical (or synonymous) codons at either end of the sequence that is skipped. They are sometimes referred to as the "take-off" and "landing" sites. Before bypass, the ribosome is positioned with a peptidyl-tRNA paired with the take-off codon in the P site, with an empty A site waiting for an aminoacyl-tRNA to enter. **Figure 7.32** shows that the ribosome slides along mRNA in this condition until the peptidyl-tRNA can become paired with the codon in the landing site. A remarkable feature of the system is its high efficiency, ~50%.

The sequence of the mRNA triggers the bypass. The important features are the two GGA codons for take-off and landing, the spacing between them, a stem-loop structure that includes the take-off codon, and the stop codon adjacent to the take-off codon. The protein under synthesis is also involved.

The take-off stage requires the peptidyl-tRNA to unpair from its codon. This is followed by a movement of the mRNA that prevents it from re-pairing. Then the ribosome scans the mRNA until the peptidyl-tRNA can repair with the codon in the landing reaction. This is followed by the resumption of protein synthesis when aminoacyl-tRNA enters the A site in the usual way.

Like frameshifting, the bypass reaction depends on a pause by the ribosome. The probability that peptidyl-tRNA will dissociate from its codon in the P site is increased by delays in the entry of aminoacyl-tRNA into the A site. Starvation for an amino acid can trigger bypassing in bacterial genes because of the delay that occurs when there is no aminoacyl-tRNA available to enter the A site. In phage T4 gene *60,* one role of mRNA structure may be to reduce the efficiency of termination, thus creating the delay that is needed for the take-off reaction.

7.18 Summary

The sequence of mRNA read in triplets 5' → 3' is related by the genetic code to the amino acid sequence of protein read from N-to C-terminus. Of the 64 triplets, 61 code for amino acids and 3 provide termination signals. Synonym codons that represent the same amino acids are related, often by a change in the third base of the codon. This third-base degeneracy, coupled with a pattern in which related amino acids tend to be coded by related codons, minimizes the effects of mutations. The genetic code is universal, and must have been established very early in evolution. Changes in nuclear genomes are rare, but some changes have occurred during mitochondrial evolution.

Multiple tRNAs may respond to a particular codon. The set of tRNAs responding to the various codons for each amino acid is distinctive for each organism. Codon-anticodon recognition involves wobbling at the first position of the anticodon (third position of the codon), which allows some tRNAs to recognize multiple codons. All tRNAs have modified bases, introduced by enzymes that recognize target bases in the tRNA structure. Codon-anticodon pairing is influenced by modifications of the anticodon itself and also by the context of adjacent bases, especially on the 3' side of the anticodon. Taking advantage of codon-anticodon wobble allows vertebrate mitochondria to use only 22 tRNAs to recognize all codons, compared with the usual minimum of 31 tRNAs; this is assisted by the changes in the mitochondrial code.

Each amino acid is recognized by a particular aminoacyl-tRNA synthetase, which also recognizes all of the tRNAs coding for that amino acid. Aminoacyl-tRNA synthetases have a proofreading function that scrutinizes the aminoacyl-tRNA products and hydrolyzes incorrectly joined aminoacyl-tRNAs.

Aminoacyl-tRNA synthetases vary widely, but fall into two general groups according to the structure of the catalytic domain. Synthetases of each group bind the tRNA from the side, making contacts principally with the extremities of the acceptor stem and the anticodon stem-loop; the two types of synthetases bind tRNA from opposite sides. The relative importance attached to the acceptor stem and the anticodon region for specific recognition varies with the individual tRNA.

Mutations may allow a tRNA to read different codons; the most common form of such mutations occurs in the anticodon itself. Alteration of its specificity may allow a tRNA to suppress a mutation in a gene coding for protein. A tRNA that recognizes a termination codon provides a nonsense suppressor; one that changes the amino acid responding to a codon is a missense suppressor. Suppressors of UAG and UGA codons are more efficient than those of UAA codons, which is explained by the fact that UAA is the most commonly used natural termination codon. But the efficiency of all suppressors depends on the context of the individual target codon.

Frameshifts of the +1 type may be caused by aberrant tRNAs that read "codons" of 4 bases. Frameshifts of either +1 or -1 may be caused by slippery sequences in mRNA that allow a peptidyl-tRNA to slip from its codon to an overlapping sequence that can also pair with its anticodon. This frameshifting also requires another sequence that causes the ribosome to delay. Frameshifts determined by the mRNA sequence may be required for expression of natural genes. Bypassing occurs when a ribosome stops translation and moves along mRNA with its peptidyl-tRNA in the P site until the peptidyl-tRNA pairs with an appropriate codon; then translation resumes.

References

7.1 Introduction

ref Nirenberg, M. W. and Leder, P. (1964). The effect of trinucleotides upon the binding of sRNA to ribosomes. Science 145, 1399-1407.

Nirenberg, M. W. and Matthaei, H. J. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Nat. Acad. Sci. USA 47, 1588-1602.

7.2 Codon-anticodon recognition involves wobbling

ref Crick, F. H. C. (1966). Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548-555.

7.3 tRNAs are processed from longer precursors

rev Hopper, A. K. and Phizicky, E. M. (2003). tRNA transfers to the limelight. Genes Dev. 17, 162-180.

7.4 tRNA contains modified bases

rev Hopper, A. K. and Phizicky, E. M. (2003). tRNA transfers to the limelight. Genes Dev. 17, 162-180.

7.5 Modified bases affect anticodon-codon pairing

rev Bjork, G. R (1987). Transfer RNA modification. Ann. Rev. Biochem. 56, 263-287.

7.6 There are sporadic alterations of the universal code

rev Fox, T. D. (1987). Natural variation in the genetic code. Ann. Rev. Genet. 21, 67-91.

Osawa, S. et al. (1992). Recent evidence for evolution of the genetic code. Microbiol. Rev. 56, 229-264.

7.7 Novel amino acids can be inserted at certain stop codons

rev Bock, A. (1991). Selenoprotein synthesis: an expansion of the genetic code. Trends Biochem. Sci. 16, 463-467.

ref Fagegaltier, D., Hubert, N., Yamada, K., Mizutani, T., Carbon, P., and Krol, A. (2000). Characterization of mSelB, a novel mammalian elongation factor for selenoprotein translation. EMBO J. 19, 4796-4805.

Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., and Chan, M. K. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. Science 296, 1462-1466.

Srinivasan, G., James, C. M., and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 296, 1459-1462.

7.8 tRNAs are charged with amino acids by synthetases

rev Schimmel, P. (1989). Parameters for the molecular recognition of tRNAs. Biochemistry 28, 2747-2759.

7.9 Aminoacyl-tRNA synthetases fall into two groups

rev Schimmel, P. (1987). Aminoacyl-tRNA synthetases: general scheme of structure-function relationships on the polypeptides and recognition of tRNAs. Ann. Rev. Biochem. 56, 125-158.

ref Rould, M. A. et al. (1989). Structure of *E. coli* glutaminyl-tRNA synthetase complexed with $tRNA^{Gln}$ and ATP at 28A resolution. Science 246, 1135-1142.

Ruff, M. et al. (1991). Class II aminoacyl tRNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexes with $tRNA^{Asp}$. Science 252, 1682-1689.

7.10 Synthetases use proofreading to improve accuracy

rev Jakubowski, H. and Goldman, E. (1992). Editing of errors in selection of amino acids for protein synthesis. Microbiol. Rev. 56, 412-429.

ref Dock-Bregeon, A., Sankaranarayanan, R., Romby, P., Caillet, J., Springer, M., Rees, B., Francklyn, C. S., Ehresmann, C, and Moras, D. (2000). Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem. Cell 103, 877-884.

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc, Nat. Acad. Sci. USA 71, 4135-4139.

Jakubowski, H. (1990). Proofreading in vivo: editing of homocysteine by methionyl-tRNA synthetase in *E. coli*. Proc. Nat. Acad. Sci. USA 87, 4504-4508.

Nomanbhoy, T. K., Hendrickson, T. L., and Schimmel, P. (1999). Transfer RNA-dependent • translocation of misactivated amino acids to prevent errors in protein synthesis. Mol. Cell 4, 519-528.

Nureki, 0. et al. (1998). Enzyme structure with two catalytic sites for double sieve selection of substrate. Science 280, 578-581.

Silvian, L. F., Wang, J., and Steitz, T. A. (1999). Insights into editing from an ile-tRNA synthetase structure with $tRNA^{Ile}$ and mupirocin. Science 285, 1074-1077.

7.13 Suppressors may compete with wild-type reading of the code

rev Atkins, J. F (1991). Towards a genetic dissection of the basis of triplet decoding, and its natural subversion: programmed reading frameshifts and hops. Ann. Rev. Genet. 25, 201-228.

Beier, H. and Grimm, M. (2001). Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res 29, 4767-4782.

Eggertsson, G. and Soll, D. (1988). Transfer RNA-mediated suppression of termination codons in *E. coli*. Microbiol. Rev. 52, 354-374.

Murgola, E. J. (1985). tRNA, suppression, and the code. Ann. Rev. Genet. 19, 57-80.

Normanly, J. and Abelson, J. (1989). Transfer RNA identity. Ann. Rev. Biochem. 58, 1029-1049.

ref Hirsh, D. (1971). Tryptophan transfer RNA as the UGA suppressor. J. Mol. Biol. 58, 439-458.

Weiner, A. M. and Weber, K. (1973). A single UGA codon functions as a natural termination signal in the coliphage q beta coat protein cistron. J. Mol. Biol. 80, 837-855.

7.14 The ribosome influences the accuracy of translation

rev Kurland, C. G. (1992). Translational accuracy and the fitness of bacteria. Ann. Rev. Genet. 26, 29-50.

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. Cell 108, 557-572.

ref Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature 407, 340-348.

Ogle, J. M., Brodersen, D. E., Clemons , W. M., Tarry, M. J., Carter, A. P., and Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292, 897-902.

7.16 Frameshifting occurs at slippery sequences

rev Farabaugh, P. J. (1995). Programmed translational frameshifting. Microbiol. Rev. 60, 103-134.

Farabaugh, P. J. and Bjorkk, G. R. (1999). How translational accuracy influences reading frame maintenance. EMBO J. 18, 1427-1434.

Gesteland, R. F. and Atkins, J. F. (1996). Recoding: dynamic reprogramming of translation. Ann. Rev. Biochem. 65, 741-768.

ref Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature 331, 280-283.

7.17 Bypassing involves ribosome movement

rev Herr, A. J., Atkins, J. F., and Gesteland, R. F. (2000). Coupling of open reading frames by translational bypassing. Ann. Rev. Biochem. 69, 343-372.

ref Gallant, J. A. and Lindsley, D. (1998). Ribosomes can slide over and beyond "hungry" codons, resuming protein chain elongation many nucleotides downstream. Proc. Nat. Acad. Sci. USA 95, 13771-13776.

Huang, W. M., Ao, S. Z., Casjens, S., Orlandi, R., Zeikus, R., Weiss, R., Winge, D., and Fang, M. (1988). A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science 239, 1005-1012.

Protein localization

8.1 Introduction

Proteins are synthesized in two types of location:

- The vast majority of proteins are synthesized by ribosomes in the cytosol.
- A small minority are synthesized by ribosomes within organelles (mitochondria or chloroplasts).

Proteins synthesized in the cytosol can be divided into two general classes with regard to localization: those that are not associated with membranes; and those that are associated with membranes. Each class can be subdivided further, depending on whether the protein associates with a particular structure in the cytosol or with a particular membrane. Figure 8.1 maps the cell in terms of the possible ultimate destinations for a newly synthesized protein and the systems that transport it:

- Cytosolic (or "soluble") proteins are not localized in any particular organelle. They are synthesized in the cytosol, and remain there, where they function as individual catalytic centers, acting on metabolites that are in solution in the cytosol.
- Macromolecular structures may be located at particular sites in the cytoplasm; for example, centrioles are associated with the regions that become the poles of the mitotic spindle.
- Nuclear proteins must be transported from their site of synthesis in the cytosol through the nuclear envelope into the nucleus.
- Most of the proteins in cytoplasmic organelles are synthesized in the cytosol and transported specifically to (and through) the organelle membrane, for example, to the mitochondrion or peroxisome or (in

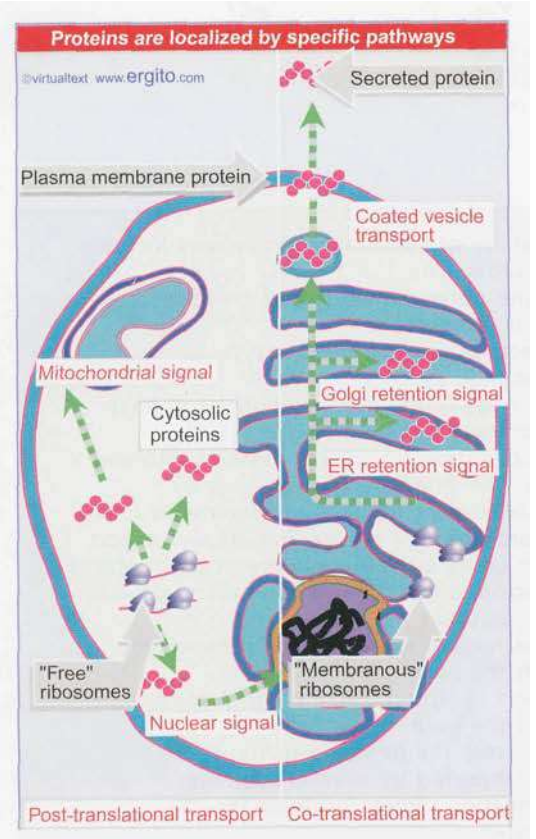


Figure 8.1 Overview: proteins that are localized post-translationally are released into the cytosol after synthesis on free ribosomes. Some have signals for targeting to organelles such as the nucleus or mitochondria. Proteins that are localized cotranslationally associate with the ER membrane during synthesis, so their ribosomes are "membrane-bound". The proteins pass into the endoplasmic reticulum, along to the Golgi, and then through the plasma membrane, unless they have signals that cause retention at one of the steps on the pathway. They may also be directed to other organelles, such as endosomes or lysosomes.

plant cells) to the chloroplast. (Those proteins that are synthesized within the organelle remain within it.)

The cytoplasm contains a series of membranous bodies, including endoplasmic reticulum (ER), Golgi apparatus, endosomes, and lysosomes. This is sometimes referred to as the "reticuloendothelial system." Proteins that reside within these compartments are inserted into ER membranes, and then are directed to their particular locations by the transport system of the Golgi apparatus.

Proteins that are secreted from the cell are transported to the plasma membrane and then must pass through it to the exterior. They start their synthesis in the same way as proteins associated with the reticuloendothelial system, but pass entirely through the system instead of halting at some particular point within it.

8.2 Passage across a membrane requires a special apparatus

Key Concepts

- Proteins pass across membranes through specialized protein structures embedded in the membrane.
- Substrate proteins interact directly with the transport apparatus of the ER or mitochondria or chloroplasts, but require carrier proteins to interact with peroxisomes.
- A much larger and complex apparatus is required for transport into the nucleus.

The process of inserting into or passing through a membrane is called protein translocation. The same dilemma must be solved for every situation in which a protein passes through a membrane. The protein presents a hydrophilic surface, but the membrane is hydrophobic. Like oil and water, the two would prefer not to mix. The solution is to create a special structure in the membrane through which the protein can pass. There are three different types of arrangement for such structures.

The endoplasmic reticulum, mitochondria, and chloroplasts contain proteinaceous structures embedded in their membranes that allow proteins to pass through without contacting the surrounding hydrophobic lipids. **Figure 8.2** shows that a substrate protein binds directly to the structure, is transported by it to the other side, and then released.

Peroxisomes also have such structures in their membranes, but the substrate proteins do not bind directly to them. **Figure 8.3** shows that instead they bind to carrier proteins in the cytosol, the carrier protein is transported through the channel into the peroxisome, and then the substrate protein is released.

For transport into the nucleus, a much larger and more complex structure is employed. This is the nuclear pore. **Figure 8.4** shows that, although the pore provides the environment that allows a substrate to enter (or to leave) the nucleus, it does not actually provide the apparatus that binds to the substrate proteins and moves them through. Included in this apparatus are carrier proteins that bind to the substrates and transport them through the pore to the other side.

8.3 Protein translocation may be post-translational or co-translational

Key Concepts

- Proteins that are imported into cytoplasmic organelles are synthesized on free ribosomes in the cytosol.
- Proteins that are imported into the ER-Golgi system are synthesized on ribosomes that are associated with the ER.
- Proteins associate with membranes by means of specific amino acid sequences called signal sequences.
- Signal sequences are most often leaders that are located at the N-terminus.
- N-terminal signal sequences are usually cleaved off the protein during the insertion process.

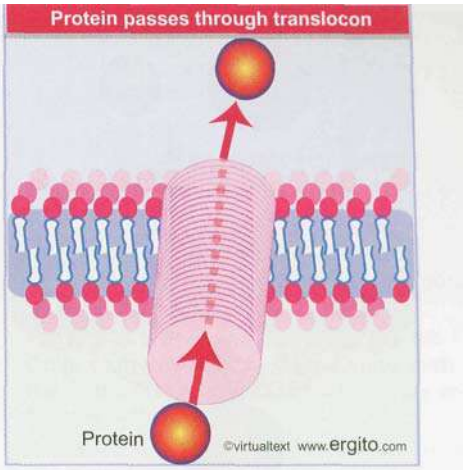


Figure 8.2 Proteins enter the ER or a mitochondrion by binding to a translocon that transports them across the membrane.

There are two ways for a protein to make its initial contact with a membrane:

- The nascent protein may associate with the translocation apparatus while it is still being synthesized on the ribosome. This called co-translational translocation.
- The protein may be released from a ribosome after translation has been completed. Then the completed protein diffuses to the appropriate membrane and associates with the translocation apparatus. This is called post-translational translocation.

The location of a ribosome depends on whether the protein under synthesis is associating with a membrane co-translationally:

- Co-translational translocation is used for proteins that enter the endoplasmic reticulum. The consequence of this association is that the ribosome is localized to the surface of the endoplasmic reticulum. Because the ribosomes are associated with the ER membranes during synthesis of these proteins, and are therefore found in membrane fractions of the cell, they are sometimes described as *membrane-bound*.
- All other ribosomes are located in the cytosol; because they are not associated with any organelle, and fractionate separately from membranes, they are sometimes called "free ribosomes". The free ribosomes synthesize all proteins except those that are translocated co-translationally. The proteins are released into the cytosol when their synthesis is completed. Some of these proteins remain free in the cytosol in quasi-soluble form; others associate with macromolecular cytosolic structures, such as filaments, microtubules, centrioles, etc., or are transported to the nucleus, or associate with membrane-bound organelles by post-translational translocation.

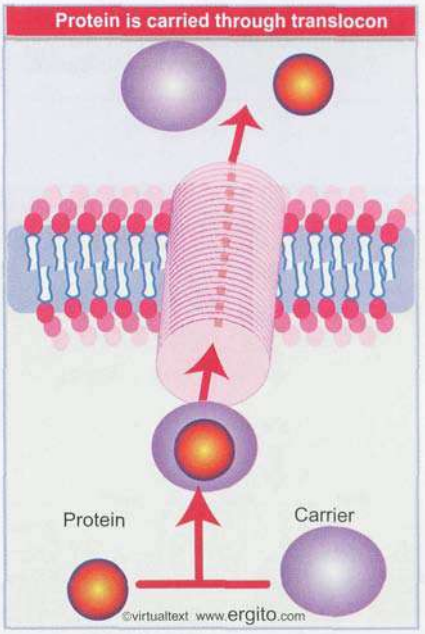


Figure 8.3 Proteins are transported into peroxisomes by a carrier protein that binds them in the cytosol, passes with them through the membrane channel, and releases them on the other side.

To associate with a membrane (or any other type of structure), a protein requires an appropriate signal, typically a sequence motif that causes it to be recognized by a translocation system (or to be assembled into a macromolecular structure).

Figure 8.5 summarizes some signals used by proteins released from cytosolic ribosomes. Import into the nucleus results from the presence of a variety of rather short sequences within proteins. These "nuclear localization signals" enable the proteins to pass through nuclear pores. One type of signal that determines transport to the peroxisome is a very short C-terminal sequence. Mitochondrial and chloroplast proteins are synthesized on "free" ribosomes; after their release into the cytosol they associate with the organelle membranes by means of N-terminal

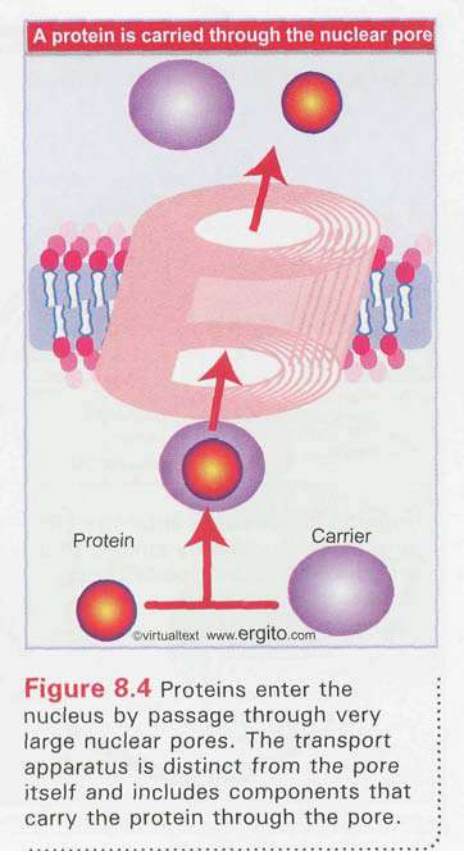


Figure 8.4 Proteins enter the nucleus by passage through very large nuclear pores. The transport apparatus is distinct from the pore itself and includes components that carry the protein through the pore.

sequences of ~25 amino acids in length that are recognized by receptors on the organelle envelope.

Proteins that reside within the reticuloendothelial system enter the endoplasmic reticulum while they are being synthesized. The principle of co-translational translocation is summarized in **Figure 8.6.** An important feature of this system is that the nascent protein is responsible for recognizing the translocation apparatus. This requires the signal for co-translational translocation to be part of the protein that is first synthesized, and, in fact, it is usually located at the N-terminus.

A common feature is found in proteins that use N-terminal sequences to be transported co-translationally to the ER or post-translationally to mitochondria or chloroplasts. The N-terminal sequence comprises a leader that is not part of the mature protein. The protein carrying this leader is called a preprotein, and is a transient precursor to the mature protein. The leader is cleaved from the protein during protein translocation.

8.4 Chaperones may be required for protein folding

Key Concepts

* Proteins that can acquire their conformation spontaneously are said to self-assemble.
* Proteins can often assemble into alternative structures.
* A chaperone directs a protein into one particular pathway by excluding alternative pathways.
* Chaperones prevent the formation of incorrect structures by interacting with unfolded proteins to prevent them from folding incorrectly.

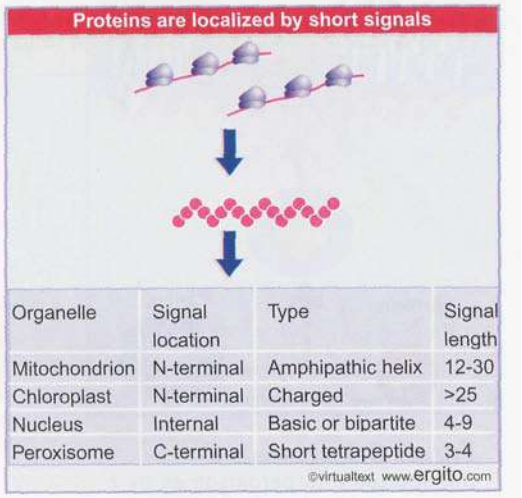


Figure 8.5 Proteins synthesized on free ribosomes in the cytosol are directed after their release to specific destinations by short signal motifs.

Some proteins are able to acquire their mature conformation spontaneously. A test for this ability is to denature the protein and determine whether it can then renature into the active form. This capacity is called self-assembly. A protein that can self-assemble can fold or refold into the active state from other conformations, including the condition in which it is initially synthesized. This implies that the internal interactions are intrinsically directed toward the right conformation. The classic case is that of ribonuclease; it was shown in the 1970s that, when the enzyme is denatured, it can renature *in vitro* into the correct conformation. More recently the process of intrinsic folding has been described in detail for some small proteins.

When correct folding does not happen, and alternative sets of interactions can occur, a protein may become trapped in a stable conformation that is not the intended final form. Proteins in this category cannot self-assemble. Their acquisition of proper structure requires the assistance of a chaperone.

Protein folding takes place by interactions between reactive surfaces. Typically these surfaces consist of exposed hydrophobic side chains. Their interactions form a hydrophobic core. The intrinsic reactivity of these surfaces means that incorrect interactions may occur unless the process is controlled. **Figure 8.7** illustrates what would happen. As a newly synthesized protein emerges from the ribosome, any hydrophobic patch in the sequence is likely to aggregate with another hydrophobic patch. Such associations are likely to occur at random and therefore will probably not represent the desired conformation of the protein.

Chaperones are proteins that mediate correct assembly by causing a target protein to acquire one possible conformation instead of others. This is accomplished by binding to reactive surfaces in the target protein that are exposed during the assembly process, and preventing those surfaces from interacting with other regions of the protein to form an incorrect conformation. Chaperones function by preventing formation of incorrect structures rather than by promoting formation of correct structures. **Figure 8.8** shows an example in which a chaperone in effect sequesters a hydrophobic patch, allowing interactions to occur that would not have been possible in its presence, as can be seen by comparing the result with Figure 8.7.

An incorrect structure may be formed either by misfolding of a single protein or by interactions with another protein. The density of proteins in the cytosol is high, and "macromolecular crowding" can increase the efficiencies of many reactions compared to the rates observed *in vitro*. Crowding can cause folding proteins to aggregate, but chaperones can counteract this effect. So one role of chaperones may be to protect a protein so that it can fold without being adversely affected by the crowded conditions in the cytosol.

We do not know what proportion of proteins can self-assemble as opposed to those that require assistance from a chaperone. (It is not axiomatic that a protein capable of self-assembly *in vitro* actually self-assembles *in vivo*, because there may be rate differences in the two conditions, and chaperones still could be involved *in vivo*. However, there is a distinction to be drawn between proteins that can in principle self-assemble and those that in principle must have a chaperone to assist acquisition of the correct structure.)

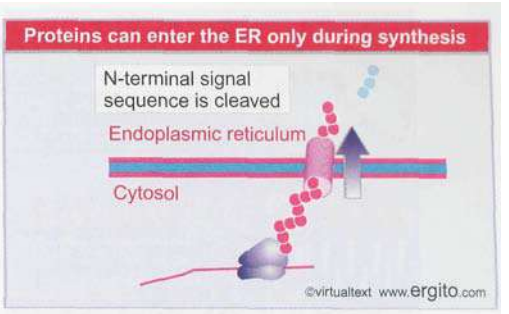


Figure 8.6 Proteins can enter the ER-Golgi pathway only by associating with the endoplasmic reticulum while they are being synthesized.

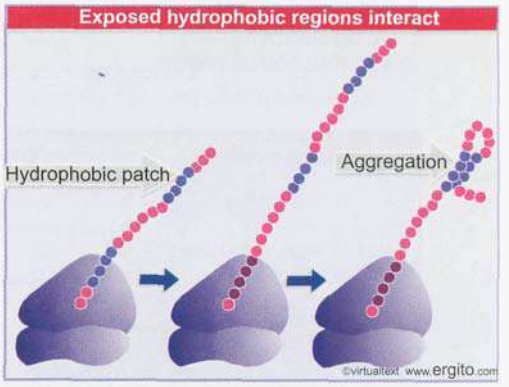


Figure 8.7 Hydrophobic regions of proteins are intrinsically interactive, and will aggregate with one another when a protein is synthesized (or denatured) unless prevented.

8.5 Chaperones are needed by newly synthesized and by denatured proteins

Key Concepts

- Chaperones act on newly synthesized proteins, proteins that are passing through membranes, or proteins that have been denatured.
- Hsp70 and some associated proteins form a major class of chaperones that act on many target proteins.
- Group I and group II chaperonins are large oligomeric assemblies that act on target proteins they sequester in internal cavities.
- Hsp90 is a specialized chaperone that acts on proteins of signal transduction pathways.

The ability of chaperones to recognize incorrect protein conformations allows them to play two related roles concerned with protein structure:

- When a protein is initially synthesized, that is to say, as it exits the ribosome to enter the cytosol, it appears in an unfolded form. Spontaneous folding then occurs as the emerging sequence interacts with regions of the protein that were synthesized previously. Chaperones influence the folding process by controlling the accessibility of the reactive surfaces. This process is involved in initial acquisition of the correct conformation.
- When a protein is denatured, new regions are exposed and become able to interact. These interactions are similar to those that occur when a protein (transiently) misfolds as it is initially synthesized.

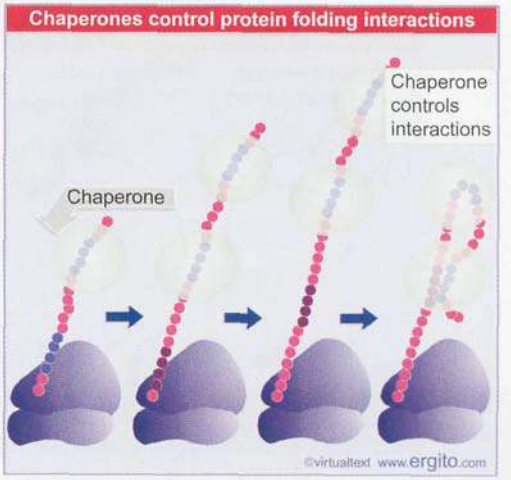


Figure 8.8 Chaperones bind to interactive regions of proteins as they are synthesized to prevent random aggregation. Regions of the protein are released to interact in an orderly manner to give the proper conformation.

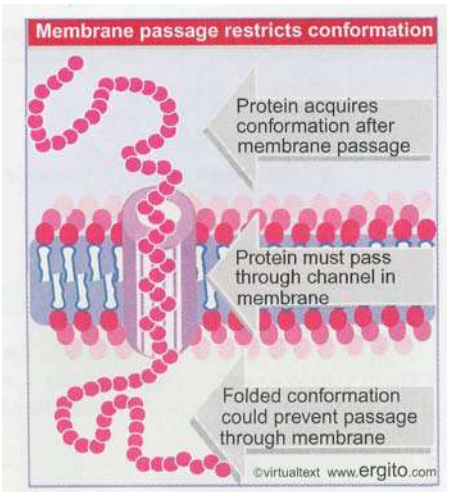


Figure 8.9 A protein is constrained to a narrow passage as it crosses a membrane.

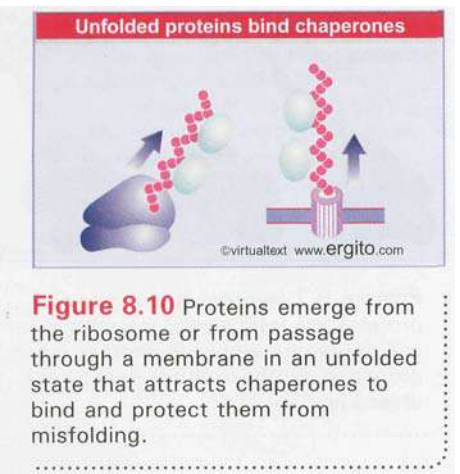


Figure 8.10 Proteins emerge from the ribosome or from passage through a membrane in an unfolded state that attracts chaperones to bind and protect them from misfolding.

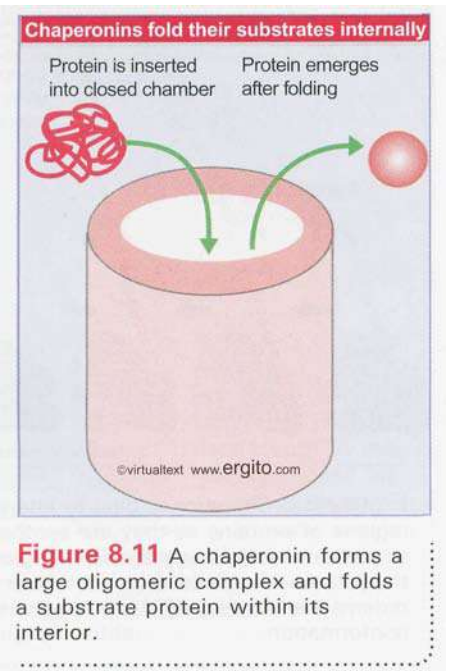


Figure 8.11 A chaperonin forms a large oligomeric complex and folds a substrate protein within its interior.

They are recognized by chaperones as comprising incorrect folds. This process is involved in recognizing a protein that has been denatured, and either assisting renaturation or leading to its removal by degradation.

Chaperones may also be required to assist the formation of oligomeric structures and for the transport of proteins through membranes. A persistent theme in membrane passage is that control (or delay) of protein folding is an important feature. **Figure 8.9** shows that it may be necessary to maintain a protein in an unfolded state before it enters the membrane because of the geometry of passage: the mature protein could simply be too large to fit into the available channel. Chaperones may prevent a protein from acquiring a conformation that would prevent passage through the membrane; in this capacity, their role is basically to maintain the protein in an unfolded, flexible state. Once the protein has passed through the membrane, it may require another chaperone to assist with folding to its mature conformation in much the same way that a cytosolic protein requires assistance from a chaperone as it emerges from the ribosome. The state of the protein as it emerges from a membrane is probably similar to that as it emerges from the ribosome—basically extended in a more or less linear condition.

Two major types of chaperones have been well characterized. They affect folding through two different types of mechanism:

- **Figure 8.10** shows that the *Hsp70 system consists of individual proteins that bind to, and act on, the substrates whose folding is to be controlled.* It recognizes proteins as they are synthesized or emerge from membranes (and also when they are denatured by stress). Basically it controls the interactions between exposed reactive regions of the protein, enabling it to fold into the correct conformation *in situ.* The components of the system are Hsp70, Hsp40, and GrpE. The name of the system reflects the original identification of Hsp70 as a protein induced by heat shock. The Hsp70 and Hsp40 proteins bind individually to the substrate proteins. They use hydrolysis of ATP to provide the energy for changing the structure of the substrate protein, and work in conjunction with an exchange factor that regenerates ATP from ADP.
- **Figure 8.11** shows that *a chaperonin system consists of a large oligomeric assembly (represented as a cylinder).* This assembly forms a structure into which unfolded proteins are inserted. The protected environment directs their folding. There are two types of chaperonin system. GroEL/GroES is found in all classes of organism. TRiC is found in eukaryotic cytosol.

The components of the systems are summarized in **Figure 8.12.** The Hsp70 system and the chaperonin systems both act on many different substrate proteins. Another system, the Hsp90 protein, functions in conjunction with Hsp70, but is directed against specific classes of proteins that are involved in signal transduction, especially the steroid hormone receptors and signaling kinases. Its basic function is to maintain its targets in an appropriate conformation until they are stabilized by interacting with other components of the pathway.

(The reason many of these proteins are named "hsp", which stands for "heat shock protein" is that increase in temperature causes production of heat shock proteins whose function is to minimize the damage caused to proteins by heat denaturation. Many of the heat shock proteins are chaperones and were first discovered, and named, as part of the heat shock response.)

By Book_Crazy [IND]

8.6 The Hsp70 family is ubiquitous

Key Concepts

- Hsp70 is a chaperone that functions on target proteins in conjunction with DnaJ and GrpE.
- Members of the Hsp70 family are found in the cytosol, in the ER, and in mitochondria and chloroplasts.

The Hsp70 family is found in bacteria, eukaryotic cytosol, in the endoplasmic reticulum, and in chloroplasts and mitochondria. A typical Hsp70 has two domains: the N-terminal domain is an ATPase; and the C-terminal domain binds the substrate polypeptide. When bound to ATP, Hsp70 binds and releases substrates rapidly; when bound to ADP, the reactions are slow. Recycling between these states is regulated by two other proteins, Hsp40 (DnaJ) and GrpE.

Figure 8.13 shows that Hsp40 (DnaJ) binds first to a nascent protein as it emerges from the ribosome. Hsp40 contains a region called the J domain (named for DnaJ), which interacts with Hsp70. Hsp70 (DnaK) binds to both Hsp40 and to the unfolded protein. In effect, two interacting chaperones bind to the protein. The J domain accounts for the specificity of the pairwise interaction, and drives a particular Hsp40 to select the appropriate partner from the Hsp70 family.

The interaction of Hsp70 (DnaK) with Hsp40 (DnaJ) stimulates the ATPase activity of Hsp70. The ADP-bound form of the complex remains associated with the protein substrate until GrpE displaces the ADP. This causes loss of Hsp40 followed by dissociation of Hsp70. The Hsp70 binds another ATP and the cycle can be repeated. GrpE (or its equivalent) is found only in bacteria, mitochondria, and chloroplasts; in other locations, the dissociation reaction is coupled to ATP hydrolysis in a more complex way.

Protein folding is accomplished by multiple cycles of association and dissociation. As the protein chain lengthens, Hsp70 (DnaK) may dissociate from one binding site and then reassociate with another, thus releasing parts of the substrate protein to fold correctly in an ordered manner. Finally, the intact protein is released from the ribosome, folded into its mature conformation.

Different members of the Hsp70 class function on various types of target proteins. Cytosolic proteins (the eponymous Hsp70 and a related protein called Hsc70) act on nascent proteins on ribosomes. Variants in the ER (called BiP or Grp78 in higher eukaryotes, called Kar2 in *S. cerevisiae)*, or in mitochondria or chloroplasts, function in a rather similar manner on proteins as they emerge into the interior of the organelle on passing through the membrane.

What feature does Hsp70 recognize in a target protein? It binds to a linear stretch of amino acids embedded in a hydrophobic context. This is precisely the sort of motif that is buried in the hydrophobic core of a properly folded mature protein. Its exposure therefore indicates that the protein is nascent or denatured. Motifs of this nature occur about every 40 amino acids. Binding to the motif prevents it from misaggregating with another one.

This mode of action explains how the Hsp70 protein BiP can fulfill two functions: to assist in oligomerization and/or folding of newly

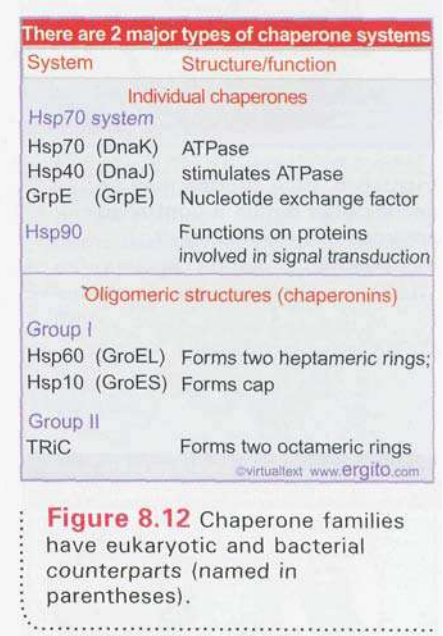


Figure 8.12 Chaperone families have eukaryotic and bacterial counterparts (named in parentheses).

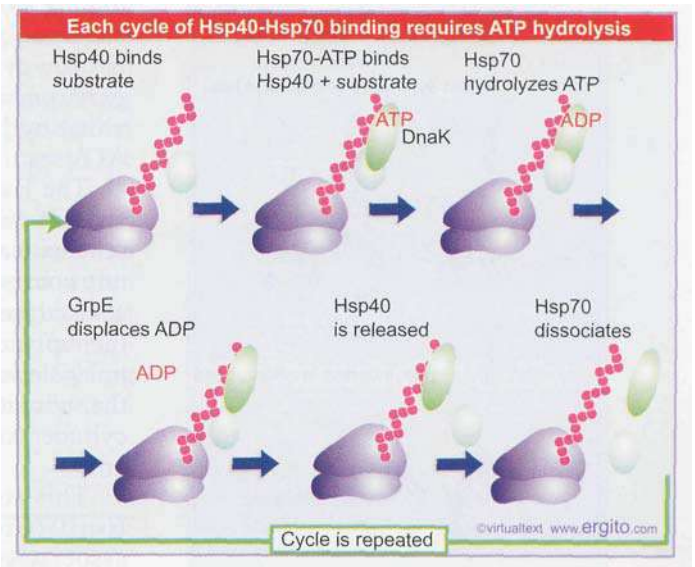


Figure 8.13 Hsp40 binds the substrate and then Hsp70. ATP hydrolysis drives conformational change. GrpE displaces the ADP; this causes the chaperones to be released. Multiple cycles of association and dissociation may occur during the folding of a substrate protein.

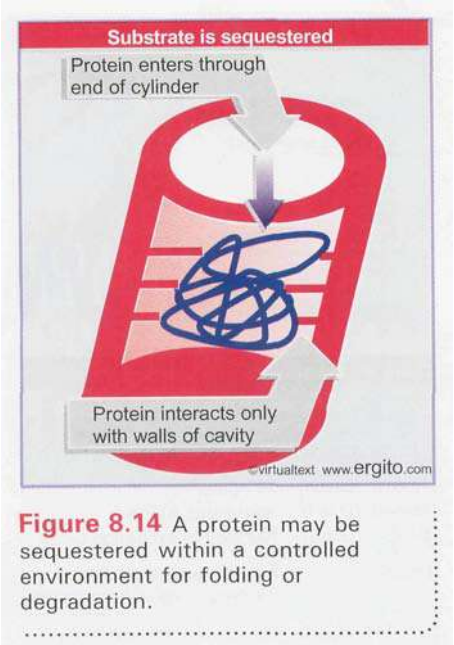


Figure 8.14 A protein may be sequestered within a controlled environment for folding or degradation.

translocated proteins in the ER; and to remove misfolded proteins. Suppose that BiP recognizes certain peptide sequences that are inaccessible within the conformation of a mature, properly folded protein. These sequences are exposed and attract BiP when the protein enters the ER lumen in an essentially one-dimensional form. And if a protein is misfolded or denatured, they may become exposed on its surface instead of being properly buried.

8.7 Hsp60/GroEL forms an oligomeric ring structure

Key Concepts

- Hsp60/GroEL forms an oligomeric structure consisting of 14 subunits arranged in two inverted heptameric rings.
- A GroES heptamer forms a dome that caps one end of the double ring.
- A substrate protein undergoes a cycle of folding in the cavity of one of the Hsp60/GroEL rings. It is released and rebound for further cycles until it reaches mature conformation.
- Hydrolysis of ATP provides energy for the folding cycles.

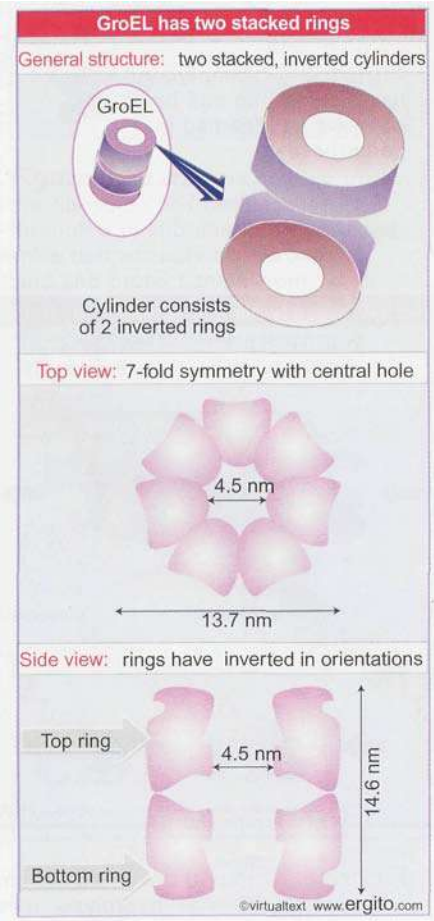


Figure 8.15 GroEL forms an oligomer of two rings, each comprising a hollow cylinder made of 7 subunits.

Large (oligomeric) structures with hollow cavities are often used for handling the folding or degradation of proteins. The typical structure is a ring of many subunits, forming a doughnut or cylinder. Figure 8.14 shows that the target protein is in effect placed in a controlled environment—the cavity—where it is closely associated with the surrounding protein. This creates a high local concentration of binding sites and supports cooperative interactions. In the case of folding, the closed environment prevents the target protein from forming wrongful interactions with other proteins, which may be important in driving folding along the proper pathway. In the case of degradation, isolation presumably makes for a more controlled process than would be possible in open cytosol (see *8.32 The proteasome is a large machine that degrades ubiquitinated proteins).* The energy for these processes is provided by hydrolysis of ATP—typically the subunits of the ring are j ATPases.

The Hsp60 class of chaperones forms a large apparatus that consists of two types of subunit. Figure 8.15 illustrates the structure schematically. Hsp60 itself (known as GroEL in *E. coli*) forms a structure consisting of 14 subunits that are arranged in two heptameric rings stacked on top of each other in inverted orientation. This means that the top and bottom surfaces of the double ring are the same. The central hole is blocked at the equator of each ring by the COOH ends of the subunits, which protrude into the interior. So the ends of the double cylinder form symmetrical cavities extending about half way into each unit.

This structure associates with a heptamer formed of subunits of Hsp10 (GroES in *E. coli*). A single GroES heptamer forms a dome that associates with one surface of the double ring, as shown in Figure 8.16. The dome sits over the central cavity, thus capping one opening of the cylinder. The dome is hollow and in effect extends the cavity into the closed surface. We can distinguish the two rings of GroEL as the proximal ring (bound to GroES) or the distal ring (not bound to GroES). The entire GroEL/GroES structure has a mass ~10^6 daltons, comparable to a small ribosomal subunit. GroEL is sometimes called a chaperonin, and GroES is called a co-chaperonin, because GroEL

By Book_Crazy [IND]

plays the essential role in guiding folding, but GroES is required for its activity.

GroEL binds to many unfolded proteins, probably by recognizing a condensed "molten globule" state. Interaction with the substrate is based on hydrophobic interactions between surfaces of the substrate and residues of GroEL that are exposed in its central cavity. Substrates may be provided by proteins that have become denatured; or they may be transferred to GroEL by other chaperones—for example, Hsp70 may assist a nascent protein in folding, but then passes it on to GroEL for the process to be completed when it is released from the ribosome.

The key reactions in substrate binding and folding are illustrated in **Figure 8.17.** The reaction starts when substrate and ATP are bound to the same ring of GroEL. This defines the proximal ring. Then GroES caps this ring. Binding of GroES induces a conformational change in the proximal GroEL ring, increasing the volume of the central cavity. This also changes the environment for the substrate. The hydrophobic residues in GroEL that had previously bound substrate are involved in binding to GroES. The result is that the substrate now finds itself in a hydrophilic environment that forces a change in its conformation.

ATP plays an important role in GroEL function. Each subunit of GroEL has a molecule of ATP. The presence of ATP on the subunits of the proximal ring is required for folding to occur. Hydrolysis is required for the transition to the next stage. Hydrolysis of the ATP in the proximal ring changes the properties of the distal ring in such a way as to allow substrate and ATP to bind to it. This in turn triggers the dissociation of the substrate and GroES from the proximal ring. Now the situation at the start of the cycle has been restored. The ring that was the distal ring in the previous cycle is bound to substrate and ATP, and becomes the proximal ring for the next cycle. So the rings of GroEL alternate as proximal and distal.

An important question in the action of this (and other macromolecular) chaperones is whether their action is processive. Does a substrate enter the central cavity, undergo multiple cycles of folding within it, and become released in mature form? Or does it undergo a single folding cycle, after which it is released? Typically it will still have improperly folded regions, and therefore will be bound again for another folding cycle. This process will continue until the protein has reached a mature conformation that does not offer a substrate to the chaperone.

These models have been tested by using a mutant GroEL that can bind unfolded proteins but cannot release them. When this "trap GroEL" is added to wild-type GroEL that is actively engaged with a substrate, it blocks the appearance of mature protein. This suggests that the substrate has been released before folding was completed. The simplest explanation is that substrate protein is released after each folding cycle. One cycle of folding, ATP hydrolysis, and release takes about 15 sec *in vitro*.

Figure 8.16 Two rings of GroEL associate back to back to form a hollow cylinder. GroES forms a dome that covers the central cavity on one side. Protein substrates bind to the cavity in the proximal ring.

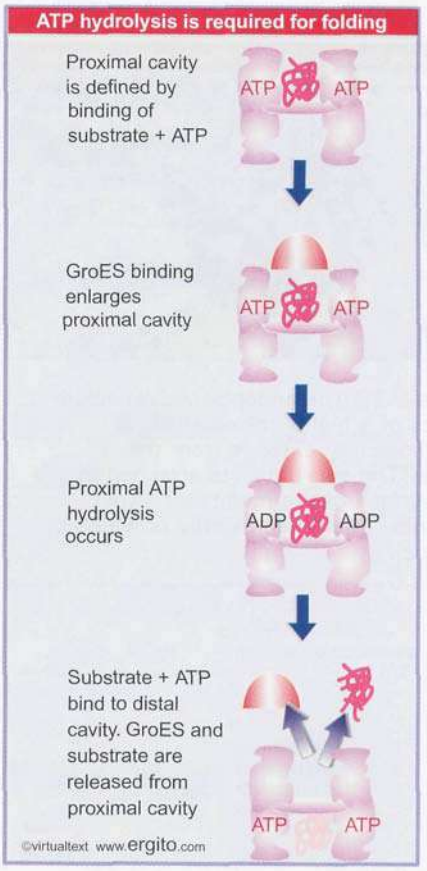


Figure 8.17 Protein folding occurs in the proximal GroEL ring and requires ATP. Release of substrate and GroES requires ATP hydrolysis in the distal ring.

8.8 Signal sequences initiate translocation

Key Concepts

- Proteins associate with the ER system only co-translationally.
- The signal sequence of the substrate protein is responsible for membrane association.

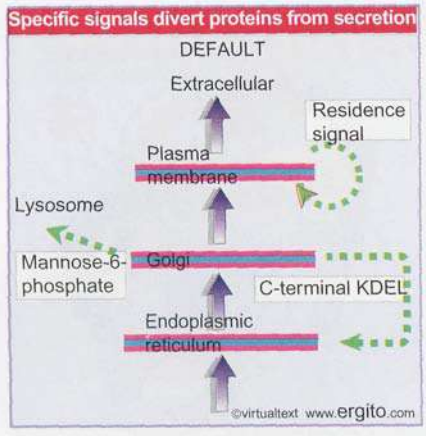


Figure 8.18 Proteins that enter the ER-Golgi pathway may flow through to the plasma membrane or may be diverted to other destinations by specific signals.

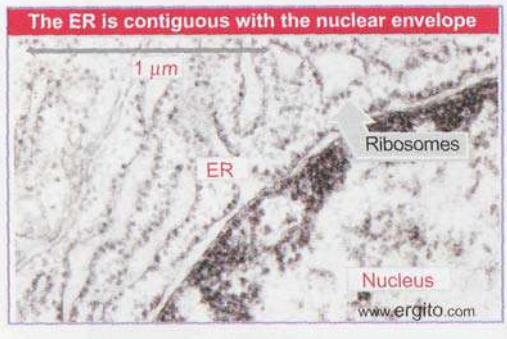


Figure 8.19 The endoplasmic reticulum consists of a highly folded sheet of membranes that extends from the nucleus. The small objects attached to the outer surface of the membranes are ribosomes. Photograph kindly provided by Lelio Orci.

Proteins that associate with membranes via N-terminal leaders use a hierarchy of signals to find their final destination. In the case of the reticuloendothelial system, the ultimate location of a protein depends on how it is directed as it transits the endoplasmic reticulum and Golgi apparatus. The leader sequence itself introduces the protein to the membrane; the intrinsic consequence of the interaction is for the protein to pass through the membrane into the compartment on the other side. For a protein to reside within the membrane, a further signal is required to stop passage through the membrane. Other types of signals are required for a protein to be sorted to a particular destination, that is, to remain within the membrane or lumen of some particular compartment. The general process of finding its ultimate destination by transport through successive membrane systems is called protein sorting or targeting, and is discussed in *27 Protein trafficking.*

The overall nature of the pathway is summarized in **Figure 8.18.** The "default pathway" takes a protein through the ER, into the Golgi, and on to the plasma membrane. Proteins that reside in the ER possess a C-terminal tetrapeptide (KDEL, which actually provides a signal for them to return to the ER from the Golgi). The signal that diverts a protein to the lysosome is a covalent modification: the addition of a particular sugar residue. Other signals are required for a protein to become a permanent constituent of the Golgi or the plasma membrane. We discuss direction to these locations in *27 Protein trafficking.*

There is a common starting point for proteins that associate with, or pass through, the reticuloendothelial system of membranes. *These proteins can associate with the membrane only while they are being synthesized.* The ribosomes synthesizing these proteins become associated with the endoplasmic reticulum, enabling the nascent protein to be co-translationally transferred to the membrane. Regions in which ribosomes are associated with the ER are sometimes called the "rough ER," in contrast with the "smooth ER" regions that lack associated polysomes and which have a tubular rather than sheet-like appearance. **Figure 8.19** shows ribosomes in the act of transferring nascent proteins to ER membranes.

The proteins synthesized at the rough endoplasmic reticulum pass from the ribosome directly to the membrane. Then they are transferred to the Golgi apparatus, and finally are directed to their ultimate destination, such as the lysosome or secretory vesicle or plasma membrane. The process occurs within a membranous environment as the proteins are carried between organelles in small membrane-coated vesicles (see *27 Protein trafficking.)*

Co-translational insertion is directed by a signal sequence. Usually this is a cleavable leader sequence of 15-30 N-terminal amino acids. At or close to the N-terminus are several polar residues, and within the leader is a hydrophobic core consisting exclusively or very largely of hydrophobic amino acids. There is no other conservation of sequence. **Figure 8.20** gives an example.

The signal sequence is both necessary and sufficient to sponsor transfer of any attached polypeptide into the target membrane. A signal sequence added to the N-terminus of a globin protein, for example,

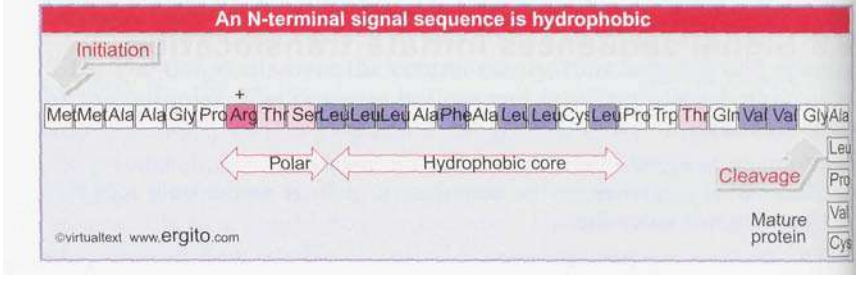


Figure 8.20 The signal sequence of bovine growth hormone consists of the N-terminal 29 amino acids and has a central highly hydrophobic region, preceded or flanked by regions containing polar amino acids.

By Book_Crazy [IND]

causes it to be secreted through cellular membranes instead of remaining in the cytosol.

The signal sequence provides the connection that enables the ribosomes to attach to the membrane. There is no intrinsic difference between free ribosomes (synthesizing proteins in the cytosol) and ribosomes that are attached to the ER. A ribosome starts synthesis of a protein without knowing whether the protein will be synthesized in the cytosol or transferred to a membrane. It is the synthesis of a signal sequence that causes the ribosome to associate with a membrane.

8.9 The signal sequence interacts with the SRP

Key Concepts

- The signal sequence binds to the SRP (signal recognition particle).
- Signal-SRP binding causes protein synthesis to pause.
- Protein synthesis resumes when the SRP binds to the SRP receptor in the membrane.
- The signal sequence is cleaved from the translocating protein by the signal peptidase located on the "inside" face of the membrane.

Protein translocationcan be divided into two general stages: first, ribosomes carrying nascent polypeptides associate with the membranes; second, the nascent chain is transferred to the channel and translocates through it.

The attachment of ribosomes to membranes requires the signal recognition particle (SRP). The SRP has two important abilities:

- It can bind to the signal sequence of a nascent secretory protein.
- It can bind to a protein (the SRP receptor) located in the membrane.

The SRP and SRP receptor function catalytically to transfer a ribosome carrying a nascent protein to the membrane. The first step is the recognition of the signal sequence by the SRP. Then the SRP binds to the SRP receptor and the ribosome binds to the membrane. The stages of translation of membrane proteins are summarized in **Figure 8.21.**

The role of the SRP receptor in protein translocation is transient. When the SRP binds to the signal sequence, it arrests translation. This usually happens when ~70 amino acids have been incorporated into the polypeptide chain (at this point the 25 residue leader has become exposed, with the next ~40 amino acids still buried in the ribosome).

Then when the SRP binds to the SRP receptor, the SRP releases the signal sequence. The ribosome becomes bound by another component of the membrane. At this point, translation can resume. When the ribosome has been passed on to the membrane, the role of SRP and SRP receptor has been played. They now recycle, and are free to sponsor the association of another nascent polypeptide with the membrane.

This process may be needed to control the conformation of the protein. If the nascent protein were released into the cytoplasm, it could take up a conformation in which it might be unable to traverse the membrane. The ability of the SRP to inhibit translation while the ribosome is being handed over to the membrane is therefore important in preventing the protein from being released into the aqueous environment.

The signal peptide is cleaved from a translocating protein by a complex of 5 proteins called the signal peptidase. The complex is several times more abundant than the SRP and SRP receptor. Its amount is roughly equivalent to the amount of bound ribosomes, suggesting that it functions in a structural capacity. It is located on the lumenal face of the

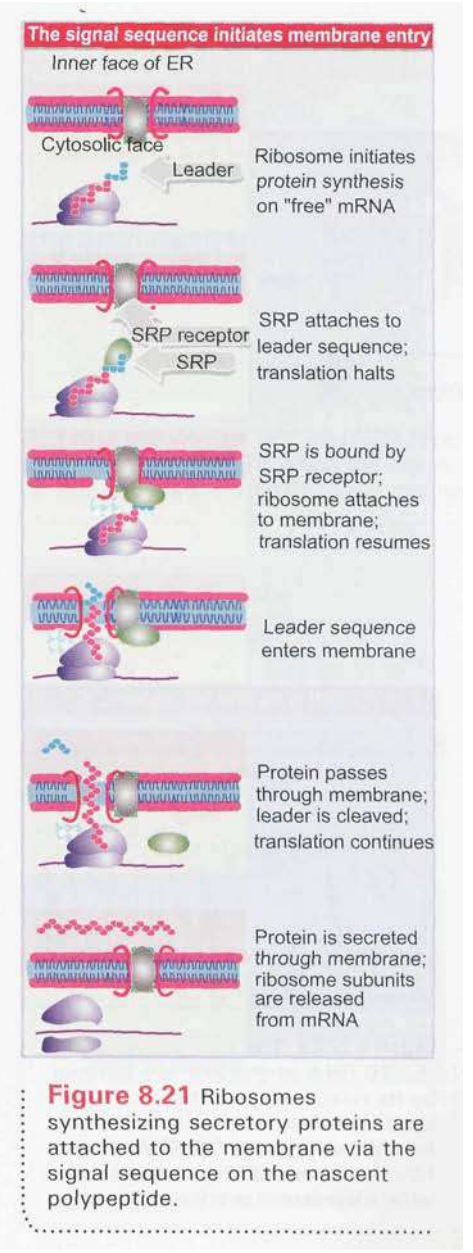


Figure 8.21 Ribosomes synthesizing secretory proteins are attached to the membrane via the signal sequence on the nascent polypeptide.

ER membrane, which implies that the entire signal sequence must cross the membrane before the cleavage event occurs. Homologous signal peptidases can be recognized in eubacteria, archaea, and eukaryotes.

8.10 The SRP interacts with the SRP receptor

Key Concepts

- The SRP is a complex of 7S RNA with 6 proteins.
- The bacterial equivalent to the SRP is a complex of 4.5S RNA with two proteins.
- The SRP receptor is a dimer.
- GTP hydrolysis releases the SRP from the SRP receptor after their interaction.

The interaction between the SRP and the SRP receptor is the key event in eukaryotic translation in transferring a ribosome carrying a nascent protein to the membrane. An analogous interacting system exists in bacteria, although its role is more restricted.

The SRP is an 11S ribonucleoprotein complex, containing 6 proteins (total mass 240 kD) and a small (305 base, 100 kD) 7S RNA. **Figure 8.22** shows that the 7S RNA provides the structural backbone of the particle; the individual proteins do not assemble in its absence.

The 7S RNA of the SRP particle is divided into two parts. The 100 bases at the 5' end and 45 bases at the 3' end are closely related to the sequence of Alu RNA, a common mammalian small RNA. They therefore define the Alu domain. The remaining part of the RNA comprises the S domain.

Different parts of the SRP structure depicted in Figure 8.22 have separate functions in protein targeting. SRP54 can be crosslinked to the signal sequence of a nascent protein; it is directly responsible for recognition of the substrate protein. The SRP68-SRP72 dimer binds to the central region of the RNA; it is needed for recognizing the SRP receptor. The SRP9-SRP14 dimer binds at the other end of the molecule; it is responsible for elongation arrest.

The SRP receptor is a dimer containing subunits SRα (72 kD) and SRβ (30 kD). The β subunit is an integral membrane protein. The amino-terminal end of the large a subunit is anchored by the β subunit. The bulk of the a protein protrudes into the cytosol. A large part of the sequence of the cytoplasmic region of the protein resembles a nucleic acid-binding protein, with many positive residues. This suggests the possibility that the SRP receptor recognizes the 7S RNA in the SRP.

There is a counterpart to SRP in bacteria, although it contains fewer components. *E. coli* contains a 4.5S RNA that associates with ribosomes and is homologous to the 7S RNA of the SRP. It associates with two proteins: Ffh is homologous to SRP54. FtsY is homologous to the a subunit of the SRP receptor. In fact, FtsY replaces the functions of both the a and β SRP subunits; its N-terminal domain substitutes for SRPβ in membrane targeting, and the C-terminal domain interacts with the target protein. The role of this complex is more limited than that of SRP-SRP receptor. It is probably required to keep some (but not all) secreted proteins in a conformation that enables them to interact with the secretory apparatus. This could be the original connection between protein synthesis and secretion; in eukaryotes the SRP has acquired the additional roles of causing translational arrest and targeting to the membrane.

Why should the SRP have an RNA component? The answer must lie in the evolution of the SRP: it must have originated very early in evolution, in an RNA-dominated world, presumably in conjunction with a

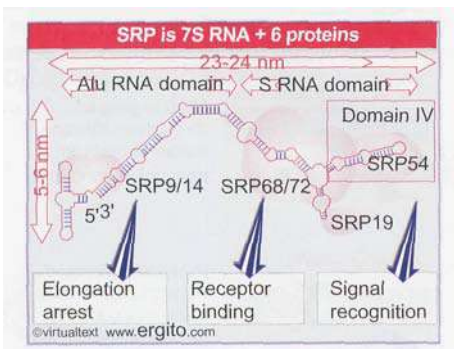


Figure 8.22 The two domains of the 7S RNA of the SRP are defined by its relationship to the Alu sequence. Five of the six proteins bind directly to the 7S RNA. Each function of the SRP is associated with a particular protein(s).

ribosome whose functions were mostly carried out by RNA. The crystal structure of the complex between the protein-binding domain of 4.5S RNA and the RNA-binding domain of Ffh suggests that RNA continues to play a role in the function of SRR

The 4.5S RNA has a region (domain IV) that is very similar to domain IV in 7S RNA (see Figure 8.22). Ffh consists of three domains (N, G, and M). The M domain (named for a high content of methionines) performs the key binding functions. It has a hydrophobic pocket that binds the signal sequence of a target protein. The hydrophobic side chains of the methionine residues create the pocket by projecting into a cleft in the protein structure. Next to the pocket is a helix-turn-helix motif that is typical of DNA-binding proteins (see *12.12 Repressor uses a helix-turn-helix motif to bind DNA).*

The crystal structure shows that the helix-loop-helix of the M domain binds to a duplex region of the 4.5S RNA in domain IV. The negatively charged backbone of the RNA is adjacent to the hydrophobic pocket. This raises the possibility that a signal sequence actually binds to both the protein and RNA components of the SRR The positively charged sequences that start the signal sequence (see Figure 8.20) could interact with the RNA, while the hydrophobic region of the signal sequence could sit in the pocket.

GTP hydrolysis plays an important role in inserting the signal sequence into the membrane. Both the SRP and the SRP receptor have GTPase capability. The signal-binding subunit of the SRP, SRP54, is a GTPase. And both subunits of the SRP receptor are GTPases. All of the GTPase activities are necessary for a nascent protein to be transferred to the membrane. **Figure 8.23** shows that the SRP starts out with GDP when it binds to the signal sequence. The ribosome then stimulates replacement of the GDP with GTP. The signal sequence inhibits hydrolysis of the GTP. This ensures that the complex has GTP bound when it encounters the SRP receptor.

For the nascent protein to be transferred from the SRP to the membrane, the SRP must be released from the SRP receptor. **Figure 8.24** shows that this requires hydrolysis of the GTPs of both the SRP and the SRP receptor. The reaction has been characterized in the bacterial system, where it has the unusual feature that Ffh activates hydrolysis by FtsY, and FtsY reciprocally activates hydrolysis by Ffh.

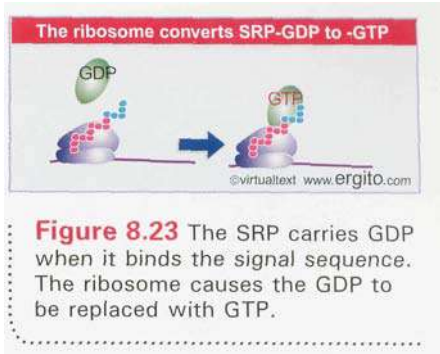


Figure 8.23 The SRP carries GDP when it binds the signal sequence. The ribosome causes the GDP to be replaced with GTP.

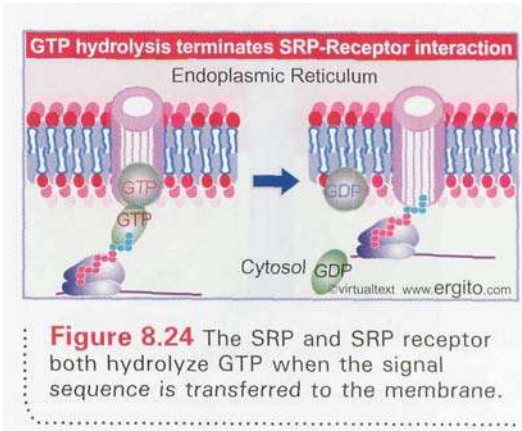


Figure 8.24 The SRP and SRP receptor both hydrolyze GTP when the signal sequence is transferred to the membrane.

8.11 The translocon forms a pore

Key Concepts

- The Sec61 trimeric complex provides the channel for proteins to pass through a membrane.
- A translocating protein passes directly from the ribosome to the translocon without exposure to the cytosol.

There is a basic problem in passing a (largely) hydrophilic protein through a hydrophobic membrane. The energetics of the interaction between the charged protein and the hydrophobic lipids are highly unfavorable. However, a protein in the process of translocation across the ER membrane can be extracted by denaturants that are effective in an aqueous environment. The same denaturants do not extract proteins that are resident components of the membrane. This suggests the model for translocation illustrated in **Figure** 8.25, in which proteins that are part of the ER membrane form an aqueous channel through the bilayer. A translocating protein moves through this channel, interacting with the resident proteins rather than with the lipid bilayer. The channel is sealed

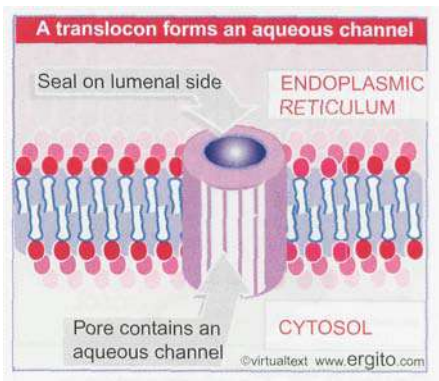


Figure 8.25 The translocon is a trimer of Sec61 that forms a channel through the membrane. It is sealed on the lumenal (ER) side.

on the lumenal side to stop free transfer of ions between the ER and the cytosol.

The channel through the membrane is called the translocon. Its components have been identified in two ways. Resident ER membrane proteins that are crosslinked to translocating proteins are potential subunits of the channel. And *sec* mutants in yeast (named because they fail to secrete proteins) include a class that cause precursors of secreted or membrane proteins to accumulate in the cytosol. These approaches together identify the **Sec61 complex,** which consists of three transmembrane proteins: Sec61α,β,γ. Sec61 is the major component of the translocon. In detergent (which provides a hydrophobic milieu that mimics the effect of a surrounding membrane), Sec61 forms cylindrical oligomers with a diameter of ~85 A and a central pore of ~20 A. Each oligomer consists of 3-4 heterotrimers.

Is the channel a preexisting structure (as implied in the figure) or might it be assembled in response to the association of a hydrophobic signal sequence with the lipid bilayer? Channels can be detected by their ability to allow the passage of ions (measured as a localized change in electrical conductance). Ion-conducting channels can be detected in the ER membrane, and their state depends on protein translocation. This demonstrates that the channel is a permanent feature of the membrane.

A channel opens when a nascent polypeptide is transferred from a ribosome to the ER membrane. The translocating protein fills the channel completely, so ions cannot pass through during translocation. But if the protein is released by treatment with puromycin, then the channel becomes freely permeable. If the ribosomes are removed from the membrane, the channel closes, suggesting that the open state requires the presence of the ribosome. This suggests that the channel is controlled in response to the presence of a translocating protein.

Measurements of the abilities of fluorescence quenching agents of different sizes to enter the channel suggest that it is large, with an internal diameter of 40-60 A. This is much larger than the diameter of an extended α-helical stretch of protein. It is also larger than the pore seen in direct views of the channel; this discrepancy remains to be explained.

The aqueous environment of an amino acid in a protein can be measured by incorporating variant amino acids that have photoreactive residues. The fluorescence of these residues indicates whether they are in an aqueous or hydrophobic environment. Experiments with such probes show that when the signal sequence is first synthesized in the ribosome, it is in an aqueous state, but is not accessible to ions in the cytosol. It remains in the aqueous state throughout its interaction with a membrane. This suggests that the translocating protein travels directly from an enclosed tunnel in the ribosome into an aqueous channel in the membrane.

In fact, access to the pore is controlled (or "gated") on *both* sides of the membrane. Before attachment of the ribosome, the pore is closed on the lumenal side. **Figure 8.26** shows that when the ribosome attaches, it seals the pore on the cytosolic side. When the nascent protein reaches a length of ~70 amino acids, that is, probably when it extends fully across the channel, the pore opens on the lumenal side. So at all times, the pore is closed on one side or the other, maintaining the ionic integrities of the separate compartments.

The translocon is versatile, and can be used by translocating proteins in several ways:

- It is the means by which nascent proteins are transferred from cytosolic ribosomes to the lumen of endoplasmic reticulum (see next section).
- It is also the route by which integral membrane proteins of the ER system are transferred to the membrane; this requires the channel to

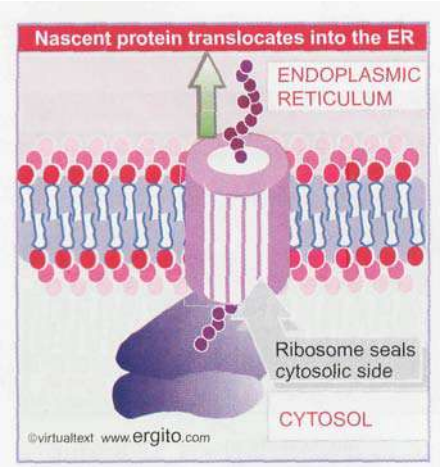


Figure 8.26 A nascent protein is transferred directly from the ribosome to the translocon. The ribosomal seals the channel on the cytosolic side.

open or disaggregate in some unknown way so that the protein can move laterally into the lipid bilayer (see *8.16 How do proteins insert into membranes?).*

Proteins can also be transferred from the ER back to the cytosol; this is known as reverse translocation (see *8.13 Reverse translocation sends proteins to the cytosol for degradation).*

8.12 Translocation requires insertion into the translocon and (sometimes) a ratchet in the ER

Key Concepts

- The ribosome, SRP, and SRP receptor are sufficient to insert a nascent protein into a translocon.
- Proteins that are inserted post-translationally require additional components in the cytosol and Bip in the ER.
- Bip is a ratchet that prevents a protein from slipping backward.

The translocon and the SRP receptor are the basic components required for co-translational translocation. When the Sec61 complex is incorporated into artificial membranes together with the SRP receptor, it can support translocation of some nascent proteins. Other nascent proteins require the presence of an additional component, TRAM, which is a major protein that becomes crosslinked to a translocating nascent chain. TRAM stimulates the translocation of all proteins.

The components of the translocon and their functions are summarized in Figure 8.27. The simplicity of this system makes several important points. We visualize Sec61 as forming the channel and also as interacting with the ribosome. The initial targeting is made when the SRP recognizes the signal sequence as the newly synthesized protein begins to emerge from the ribosome. The SRP binds to the SRP receptor, and the signal sequence is transferred to the translocon. When the signal sequence enters the translocon, the ribosome attaches to Sec61, forming a seal so that the pore is not exposed to the cytosol. Cleavage of the signal peptide does not occur in this system, and therefore cannot be necessary for translocation *per se*. In this system, components on the lumenal side of the membrane are not needed for translocation.

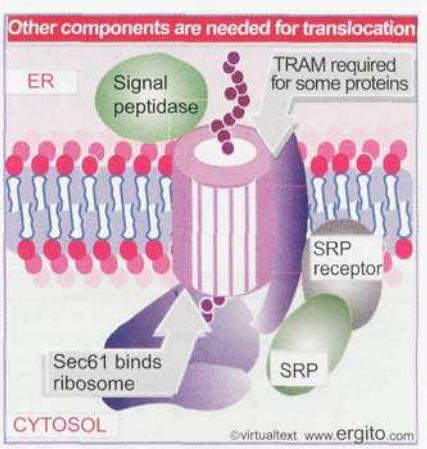


Figure 8.27 Translocation requires the translocon, SRP, SRP receptor, Sec61, TRAM, and signal peptidase.

Of course, the efficiency of the *in vitro* system is relatively low. Additional components could be required *in vivo* to achieve efficient transfer or to prevent other cellular proteins from interfering with the process.

A more complex apparatus is required in certain cases in which a protein is inserted into a membrane post-translationally. The same Sec61 complex forms the channel, but four other Sec proteins are also required, and in addition the chaperone BiP (a member of the Hsp70 class) and a supply of ATP are required on the lumenal side of the membrane. Figure 8.28 shows that BiP behaves as ratchet. In the absence of BiP, Brownian motion allows the protein to slip back into the cytosol. But BiP grabs the protein as it exits the pore into the endoplasmic reticulum. This stops the protein from moving backward. BiP does not pull the protein through; it just stops it from sliding back. (The reason why BiP is required for post-translational translocation but not for co-translational translocation may be that a newly synthesized protein is continuously extruded from the ribosome and therefore cannot slip backward.)

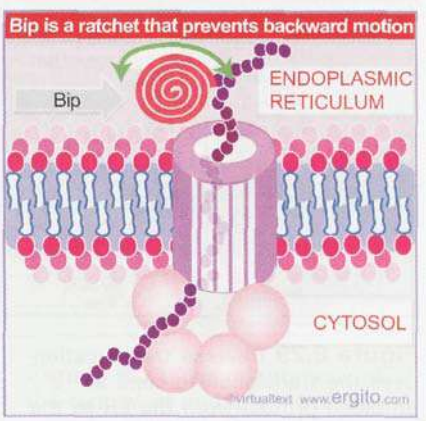


Figure 8.28 BiP acts as a ratchet to prevent backward diffusion of a translocating protein.

8.13 Reverse translocation sends proteins to the cytosol for degradation

Key Concepts

- Sec61 translocons can be used for reverse translocation of proteins from the ER into the cytosol.

Several important activities occur within the endoplasmic reticulum. Proteins move through the ER en route to a variety of destinations (see *27 Protein trafficking*). They are glycosylated and folded into their final conformations. The ER provides a "quality control" system in which misfolded proteins are identified and degraded. However, the degradation itself does not occur in the ER, but may require the protein to be exported back to the cytosol.

The first indication that ER proteins are degraded in the cytosol and not in the ER itself was provided by evidence for the involvement of the proteasome, a large protein aggregate with several proteolytic activities (see *8.32 The proteasome is a large machine that degrades ubiquitinated proteins*). Inhibitors of the proteasome prevent the degradation of aberrant ER proteins. Proteins are marked for cleavage by the proteasome when they are modified by the addition of ubiquitin, a small polypeptide chain (see *8.31 Ubiquitination targets proteins for degradation*). The important point to note now is that ubiquitination and proteasomal degradation both occur in the cytosol (with a minor proportion in the nucleus).

Transport from the ER back into the cytosol occurs by a reversal of the usual process of import. This is called reverse translocation. The Sec61 translocon is used. The conditions are different; for example, the translocon is not associated with a ribosome. Some mutations in Sec61 prevent reverse translocation, but do not prevent forward translocation. This could be either because there is some difference in the process or (more likely) because these regions interact with other components that are necessary for reverse translocation.

Figure 8.29 points out that we do not know how the channel is opened to allow insertion of the protein on the ER side. Special components are presumably involved. One model is that misfolded or misassembled proteins are recognized by chaperones, which transfer them to the translocon. In one particular case, human cytomegalovirus (CMV) codes for cytosolic proteins that destroy newly synthesized MHC class I (cellular major histocompatibility complex) proteins. This requires a viral protein product (US2), which is a membrane protein that functions in the ER. It interacts with the MHC proteins and probably conveys them into the translocon for reverse translocation.

The system involved in the degradation of aberrant ER proteins can be identified by mutations (in yeast) that lead to accumulation of aberrant proteins. Usually a protein that misfolds (produced by a mutated gene) is degraded instead of being transported through the ER. Yeast mutants that cannot degrade the substrate fall into two classes: some identify components of the proteolytic apparatus, such as the enzymes involved in ubiquitination; other identify components of the transport apparatus, including Sec61, BiP, and Sec63. There is also a protein in the ER membrane that functions on the cytosolic side to localize the ubiquitination enzymes at the translocon. In fact, retrograde transport into the cytosol cannot occur in the absence of this protein, which suggests that there is a mechanical link between retrograde transport and degradation.

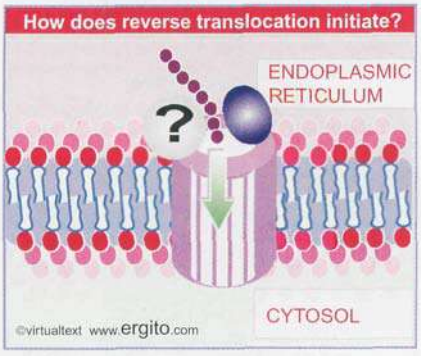


Figure 8.29 Reverse translocation uses the translocon to send an unfolded protein from the ER to the cytosol, where it is degraded. The mechanism of putting the translocon into reverse is not known.

By Book_Crazy [IND]

8.14 Proteins reside in membranes by means of hydrophobic regions

Key Concepts

- Group I proteins have the N-terminus on the far side of the membrane; group II proteins have the opposite orientation.
- Some proteins have multiple membrane-spanning domains.

All biological membranes contain proteins, which are held in the lipid bilayer by noncovalent interactions. The operational definition of an integral membrane protein is that it requires disruption of the lipid bilayer in order to be released from the membrane. A common feature in such proteins is the presence of at least one transmembrane domain, consisting of an α-helical stretch of 21-26 hydrophobic amino acids. A sequence that fits the criteria for membrane insertion can be identified by a hydropathy plot, which measures the cumulative hydrophobicity of a stretch of amino acids. A protein that has domains exposed on both sides of the membrane is called a transmembrane protein. The association of a protein with a membrane takes several forms. The topography of a membrane protein depends on the number and arrangement of transmembrane regions.

When a protein has a single transmembrane region, its position determines how much of the protein is exposed on either side of the membrane. A protein may have extensive domains exposed on both sides of the membrane or may have a site of insertion close to one end, so that little or no material is exposed on one side. The length of the N-terminal or C-terminal tail that protrudes from the membrane near the site of insertion varies from insignificant to quite bulky.

Figure 8.30 shows that proteins with a single transmembrane domain fall into two classes. Group I proteins in which the N-terminus faces the extracellular space are more common than group II proteins in which the orientation has been reversed so that the N-terminus faces the cytoplasm. Orientation is determined during the insertion of the protein into the endoplasmic reticulum.

Figure **8.31** shows orientations for proteins that have multiple membrane-spanning domains. An odd number means that both termini of the protein are on opposite sides of the membrane, whereas an even number implies that the termini are on the same face. The extent of the domains exposed on one or both sides is determined by the locations of the transmembrane domains. Domains at either terminus may be exposed, and internal sequences between the domains "loop out" into the extracellular space or cytoplasm. One common type of structure is the 7-membrane passage or "serpentine" receptor; another is the 12-membrane passage component of an ion channel.

Does a transmembrane domain itself play any role in protein function besides allowing the protein to insert into the lipid bilayer? In the simple group I or II proteins, it has little or no additional function; often it can be replaced by any other transmembrane domain. However, transmembrane domains play an important role in the function of proteins that make multiple passes through the membrane or that have subunits that oligomerize within the membrane. The transmembrane domains in such cases often contain polar residues, which are not found in the single membrane-spanning domains of group I and group II proteins. Polar regions in the membrane-spanning domains do not interact with the lipid bilayer, but instead interact with one another. This enables them to form a polar pore or channel within the lipid bilayer. Interaction between such transmembrane domains can create a hydrophilic passage through the hydrophobic interior of the membrane. This

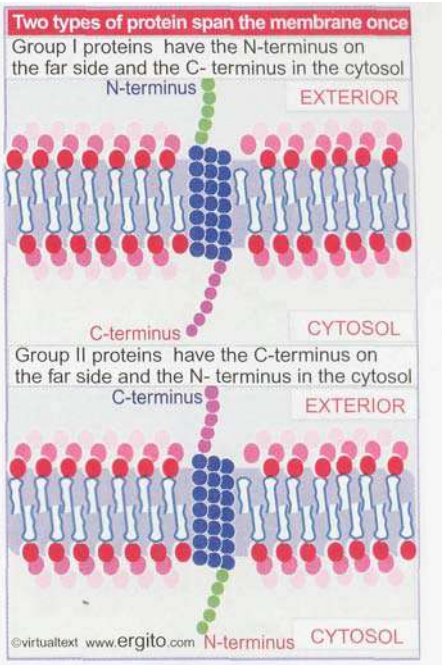


Figure 8.30 Group I and group II transmembrane proteins have opposite orientations with regard to the membrane.

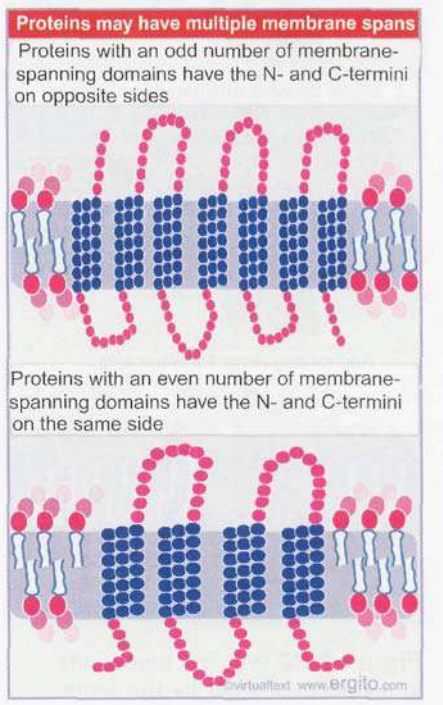


Figure 8.31 The orientations of the termini of multiple membrane-spanning proteins depends on whether there is an odd or even number of transmembrane segments.

can allow highly charged ions or molecules to pass through the membrane, and is important for the function of ion channels and transport of ligands. Another case in which conformation of the transmembrane domains is important is provided by certain receptors that bind lipophilic ligands. In such cases, the transmembrane domains (rather than the extracellular domains) bind the ligand within the plane of the membrane.

8.15 Anchor sequences determine protein orientation

Key Concepts

- An anchor sequence halts the passage of a protein through the translocon. Typically this is located at the C-terminal end and results in a group I orientation in which the N-terminus has passed through the membrane.
- A combined signal-anchor sequence can be used to insert a protein into the membrane and anchor the site of insertion. Typically this is internal and results in a group II orientation in which the N-terminus is cytosolic.

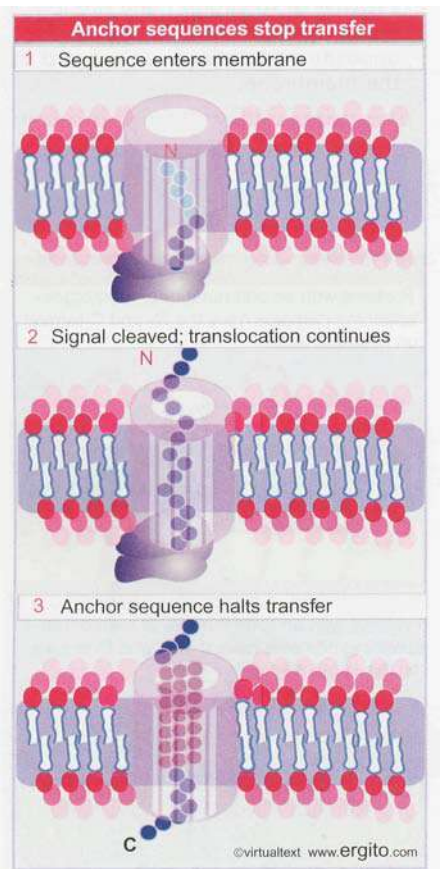


Figure 8.32 Proteins that reside in membranes enter by the same route as secreted proteins, but transfer is halted when an anchor sequence passes into the membrane. If the anchor is at the C-terminus, the bulk of the protein passes through the membrane and is exposed on the far surface.

Proteins that are secreted from the cell pass through a membrane while remaining in the aqueous channel of the translocon. By contrast, proteins that reside in membranes start the process in the same way, but then transfer from the aqueous channel into the hydrophobic environment. The challenge in accounting for insertion of proteins into membranes is to explain what distinguishes transmembrane proteins from secreted proteins, and causes this transfer. The pathway by which proteins of either type I or type II are inserted into the membrane follows the same initial route as that of secretory proteins, relying on a signal sequence that functions co-translationally. But proteins that are to remain within the membrane possess a second, stop-transfer signal. This takes the form of a cluster of hydrophobic amino acids adjacent to some ionic residues. The cluster serves as an anchor that latches on to the membrane and stops the protein from passing right through.

A surprising property of anchor sequences is that they can function as signal sequences when engineered into a different location. When placed into a protein lacking other signals, such a sequence may sponsor membrane translocation. One possible explanation for these results is that the signal sequence and anchor sequence interact with some common component of the apparatus for translocation. Binding of the signal sequence initiates translocation, but the appearance of the anchor sequence displaces the signal sequence and halts transfer.

The insertion of type I proteins is illustrated in **Figure 8.32.** The signal sequence is N-terminal. The location of the anchor signal determines when transfer of the protein is halted. When the anchor sequence takes root in the membrane, domains on the N-terminal side will be located in the lumen, while domains on the C-terminal side are located facing the cytosol.

A common location for a stop-transfer sequence of this type is at the C-terminus. As shown in the figure, transfer is halted only as the last sequences of the protein enter the membrane. This type of arrangement is responsible for the location in the membrane of many proteins, including cell surface proteins. Most of the protein sequence is exposed on the lumenal side of the membrane, with a small or negligible tail facing the cytosol.

Type II proteins do not have a cleavable leader sequence at the N-terminus. Instead the signal sequence is combined with an anchor sequence. We imagine that the general pathway for the integration of type II proteins into the membrane involves the steps illustrated in **Figure 8.33.**

Membrane insertion starts by the insertion of the signal sequence in the form of a hairpin loop. The signal sequence enters the membrane, but the joint signal-anchor sequence does not pass through. Instead it stays in the membrane (perhaps interacting directly with the lipid bilayer), while the rest of the growing polypeptide continues to loop into the endoplasmic reticulum.

The signal-anchor sequence is usually internal, and its location determines which parts of the protein remain in the cytosol and which are extracellular. Essentially all the N-terminal sequences that precede the signal-anchor are exposed to the cytosol. Usually the cytosolic tail is short, ~6-30 amino acids. In effect the N-terminus remains constrained while the rest of the protein passes through the membrane. This reverses the orientation of the protein with regard to the membrane.

The combined signal-anchor sequences of type II proteins resemble cleavable signal sequences. **Figure 8.34** gives an example. Like cleavable leader sequences, the amino acid composition is more important than the actual sequence. The regions at the extremities of the signal-anchor carry positive charges; the central region is uncharged and resembles a hydrophobic core of a cleavable leader. Mutations to introduce charged amino acids in the core region prevent membrane insertion; mutations on either side prevent the anchor from working, so the protein is secreted or located in an incorrect compartment.

The distribution of charges around the anchor sequence has an important effect on the orientation of the protein. More positive charges are usually found on the cytoplasmic side (N-terminal side in type II proteins). If the positive charges are removed by mutation, the orientation of the protein can be reversed. The effect of charges on orientation is summarized by the "positive inside" rule, which says that the side of the anchor with the most positive charges will be located in the cytoplasm. The positive charges in effect provide a hook that latches on to the cytoplasmic side of the membrane, controlling the direction in which the hydrophobic region is inserted. and thus determining the orientation of the protein.

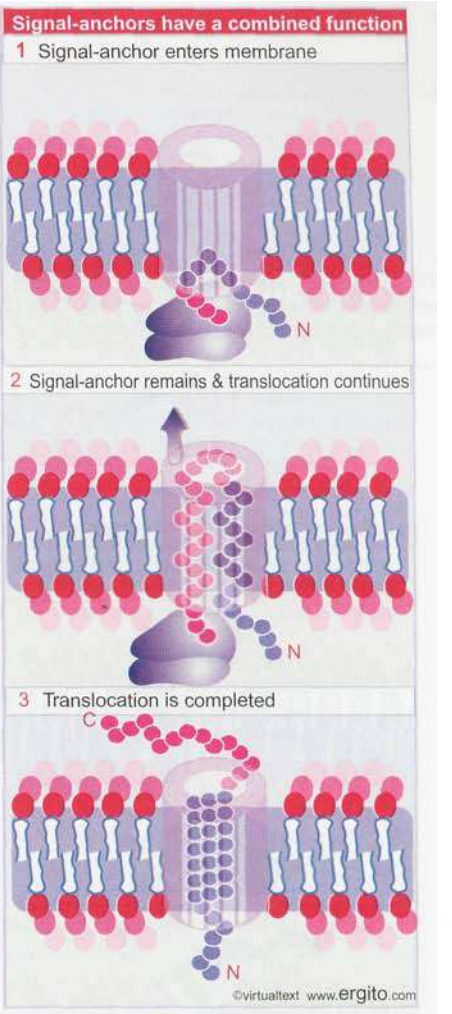


Figure 8.33 A combined signal-anchor sequence causes a protein to reverse its orientation, so that the N-terminus remains on the inner face and the C-terminus is exposed on the outer face of the membrane.

8.16 How do proteins insert into membranes?

Key Concepts

- Transfer of transmembrane domains from the translocon into the lipid bilayer is triggered by the interaction of the transmembrane region with the translocon.

We have a reasonable understanding of the processes by which secreted proteins pass through membranes and of how this relates to the insertion of the single-membrane spanning group I and group II proteins. We cannot yet explain the details of insertion of proteins with multiple membrane-spanning domains.

We understand how a secreted protein passes through a membrane without any conflict, but it is difficult to apply the same model to a protein that resides in the membrane. **Figure 8.35** illustrates the difference between the organization of a translocating protein, which is protected from the lipid bilayer by the aqueous channel, and a transmembrane protein, which has a hydrophobic segment directly in contact with the membrane. We do not know how a protein is transferred from its passage through the proteinaceous channel into the lipid bilayer itself. One possibility is that there is some mechanism for transferring hydrophobic transmembrane domains directly from the channel into the membrane, as suggested in **Figure 8.36.** This idea is supported by observations of an *in vitro* system which measured transfer into a lipid environment for proteins with different transmembrane domains. When the domain passed a threshold of hydrophobicity, the protein could

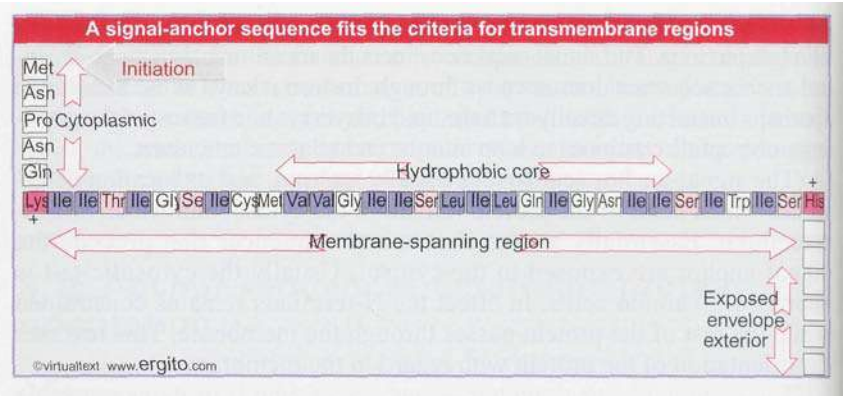


Figure 8.34 The signal-anchor of influenza neuraminidase is located close to the N-terminus and has a hydrophobic core.

pass from a channel consisting of Sec61 and TRAM into the lipid bilayer. The simplest explanation is that the structure of the channel allows the translocating protein to contact the lipid bilayer, so that a sufficiently hydrophobic segment can simply partition directly into the lipid. An alternative is that hydrophobic domains cause the channel to disaggregate, exposing the hydrophobic amino acids to the lipid bilayer.

It has always been a common assumption that, whatever the exact mechanism for transferring the transmembrane segment into the membrane, it is triggered by the presence of the transmembrane sequence in the pore. However, changes in the pore occur earlier in response to the synthesis of the transmembrane sequence in the ribosome. When a secreted protein passes through the pore, the channel remains sealed on the cytosolic side but opens on the lumenal side after synthesis of the first 70 residues. But as soon as a transmembrane sequence has been fully synthesized, that is, while it is still entirely within the ribosome, the pore closes on the lumenal side. How this change relates to the transfer of the transmembrane sequence into the membrane is not clear.

The process of insertion into a membrane has been characterized for both type I proteins and type II proteins, in which there is a single transmembrane domain. How is a protein with multiple membrane-spanning regions inserted into a membrane? Much less is known about this process, but we assume that it relies on sequences that provide signal and/or anchor capabilities. One model is to suppose that there is an alternating series of signal and anchor sequences. Translocation is initiated at the first signal sequence and continues until stopped by the first anchor. Then it is reinitiated by a subsequent signal sequence, until stopped by the next anchor. It is possible that there are multiple pathways for integration into the membrane, because in some cases a transmembrane domain seems to move into the lipid bilayer as soon as it enters the translocon, but in other cases there can be a delay until other transmembrane regions have been synthesized.

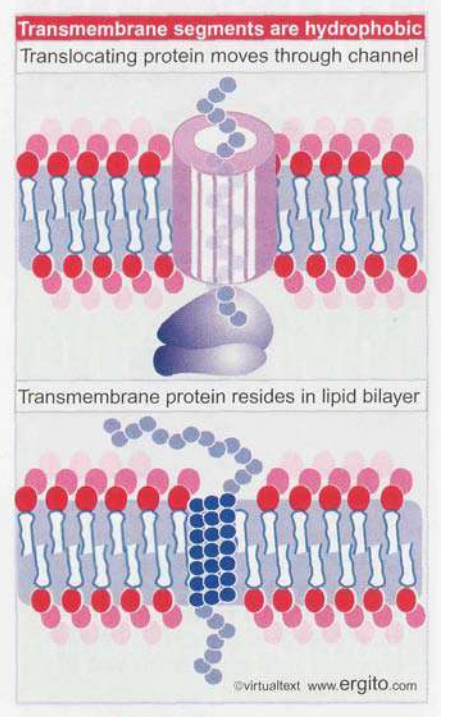


Figure 8.35 How does a transmembrane protein make the transition from moving through a proteinaceous channel to interacting directly with the lipid bilayer?

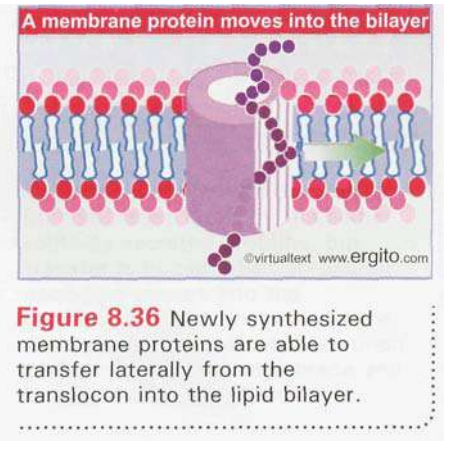


Figure 8.36 Newly synthesized membrane proteins are able to transfer laterally from the translocon into the lipid bilayer.

8.17 Post-translational membrane insertion depends on leader sequences

Key Concepts

- N-terminal leader sequences provide the information that allows proteins to associate with mitochondrial or chloroplast membranes.

Mitochondria and chloroplasts synthesize only some of their proteins. Mitochondria synthesize only ~10 organelle proteins; chloroplasts synthesize ~50 proteins. The majority of organelle proteins are synthesized in the cytosol by the same pool of free ribosomes that synthesize cytosolic proteins. They must then be imported into the organelle.

Many proteins that enter mitochondria or chloroplasts by a post-translational process have leader sequences that are responsible for primary recognition of the outer membrane of the organelle. As shown in the simplified diagram of **Figure 8.37,** the leader sequence initiates the interaction between the precursor and the organelle membrane. The protein passes through the membrane, and the leader is cleaved by a protease on the organelle side.

The leaders of proteins imported into mitochondria and chloroplasts usually have both hydrophobic and basic amino acids. They consist of stretches of uncharged amino acids interrupted by basic amino acids, and they lack acidic amino acids. There is little other homology. An example is given in **Figure 8.38.** Recognition of the leader does not depend on its exact sequence, but rather on its ability to form an amphipathic helix, in which one face has hydrophobic amino acids, and the other face presents the basic amino acids.

The leader sequence contains all the information needed to localize an organelle protein. The ability of a leader sequence can be tested by constructing an artificial protein in which a leader from an organelle protein is joined to a cytosolic protein. The experiment is performed by constructing a hybrid gene, which is then translated into the hybrid protein.

Several leader sequences have been shown by such experiments to function independently to target any attached sequence to the mitochondrion or chloroplast. For example, if the leader sequence given in Figure 8.38 is attached to the cytosolic protein DHFR (dihydrofolate reductase), the DHFR becomes localized in the mitochondrion.

The leader sequence and the transported protein represent domains that fold independently. Irrespective of the sequence to which it is attached, the leader must be able to fold into an appropriate structure to be recognized by receptors on the organelle envelope. The attached polypeptide sequence plays no part in recognition of the envelope.

What restrictions are there on transporting a hydrophilic protein through the hydrophobic membrane? An insight into this question is given by the observation that methotrexate, a ligand for the enzyme DHFR, blocks transport into mitochondria of DHFR fused to a mitochondrial leader. The tight binding of methotrexate prevents the enzyme from unfolding when it is translocated through the membrane. So although the sequence of the transported protein is irrelevant for targeting purposes, in order to follow its leader through the membrane, it requires the flexibility to assume an unfolded conformation.

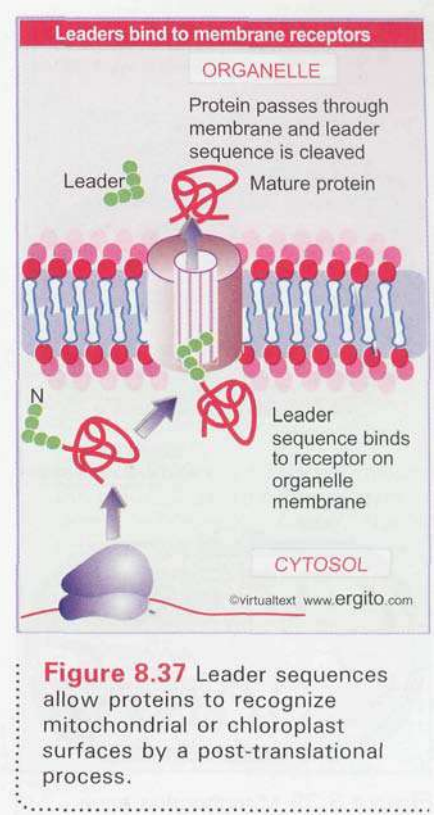


Figure 8.37 Leader sequences allow proteins to recognize mitochondrial or chloroplast surfaces by a post-translational process.

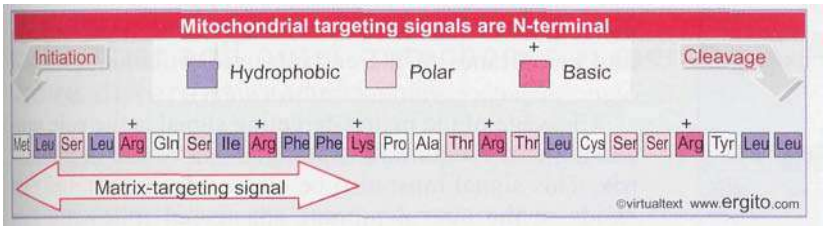


Figure 8.38 The leader sequence of yeast cytochrome *c* oxidase subunit IV consists of 25 neutral and basic amino acids. The first 12 amino acids are sufficient to transport any attached polypeptide into the mitochondrial matrix.

8.18 A hierarchy of sequences determines location within organelles

Key Concepts

- The N-terminal part of a leader sequence targets a protein to the mitochondrial matrix or chloroplast lumen.
- An adjacent sequence can control further targeting, to a membrane or the intermembrane spaces.
- The sequences are cleaved successively from the protein.

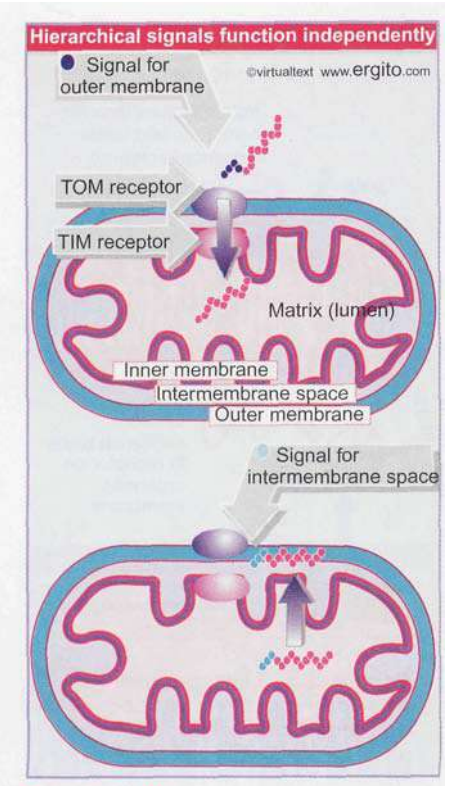


Figure 8.39 Mitochondria have receptors for protein transport in the outer and inner membranes. Recognition at the outer membrane may lead to transport through both receptors into the matrix, where the leader is cleaved. If it has a membrane-targeting signal, it may be re-exported.

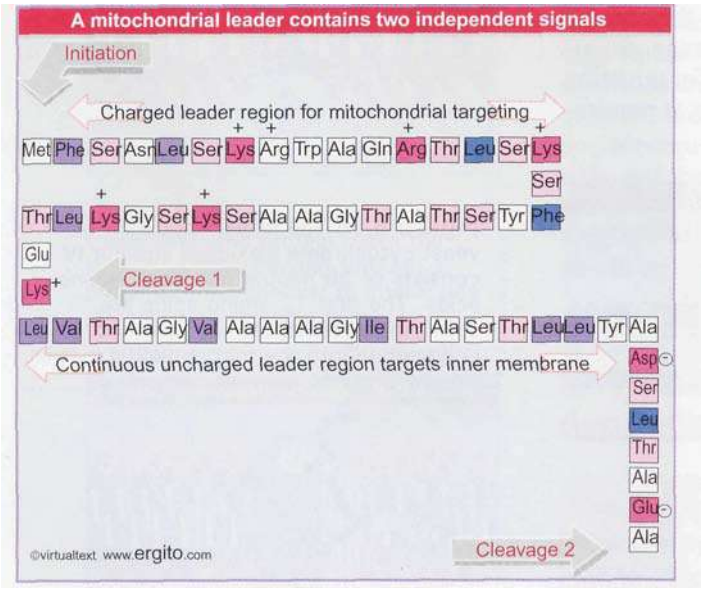


Figure 8.40 The leader of yeast cytochrome *c*1 contains an N-terminal region that targets the protein to the mitochondrion, followed by a region that targets the (cleaved) protein to the inner membrane. The leader is removed by two cleavage events.

The mitochondrion is surrounded by an envelope consisting of two membranes. Proteins imported into mitochondria may be located in the outer membrane, the intermembrane space, the inner membrane, or the matrix. A protein that is a component of one of the membranes may be oriented so that it faces one side or the other.

What is responsible for directing a mitochondrial protein to the appropriate compartment? The "default" pathway for a protein imported into a mitochondrion is to move through both membranes into the matrix. This property is conferred by the N-terminal part of the leader sequence. A protein that is localized within the intermembrane space or in the inner membrane itself requires an additional signal, which specifies its destination within the organelle. A multipart leader contains signals that function in a hierarchical manner, as summarized in **Figure 8.39.** The first part of the leader targets the protein to the organelle, and the second part is required if its destination is elsewhere than the matrix. The two parts of the leader are removed by successive cleavages.

Cytochrome *c1* is an example. It is bound to the inner membrane and faces the intermembrane space. Its leader sequence consists of 61 amino acids, and can be divided into regions with different functions. The sequence of the first 32 amino acids alone, or even the N-terminal half of this region, can transport DHFR all the way into the matrix. So the first part of the leader sequence (32 N-terminal amino acids) comprises a matrix-targeting signal. But the intact leader transports an attached sequence—such as murine DHFR—into the intermembrane space.

What prevents the protein from proceeding past the intermembrane space when it has an intact leader? The region following the matrix-targeting signal (comprising 19 amino acids of the leader) provides another signal that localizes the protein at the inner membrane or within the intermembrane space. For working purposes, we call this the membrane-targeting signal.

The two parts of a leader that contains both types of signal have different compositions. As indicated in **Figure 8.40,** the 35 N-terminal amino acids resemble other organelle leader sequences in the high content of uncharged amino acids, punctuated by basic amino acids. The next 19 amino acids, however, comprise an uninterrupted stretch of uncharged amino acids, long enough to span a lipid bilayer. This membrane-targeting signal resembles the sequences that are involved in protein translocation into membranes of the endoplasmic reticulum (see *8.8 Signal sequences initiate translocation).*

Cleavage of the matrix-targeting signal is the sole processing event required for proteins that reside in the matrix. This signal must also be cleaved from proteins that reside in the intermembrane space; but following this cleavage, the membrane-targeting signal (which is now the N-terminal sequence of the protein) directs the protein to its destination in the outer membrane, intermembrane space, or inner membrane. Then it in turn is cleaved.

The N-terminal matrix-targeting signal functions in the same manner for all mitochondrial proteins. Its recognition by a receptor on the outer membrane leads to transport through the two membranes. And the same protease is involved in cleaving the matrix-targeting signal, irrespective of the final destination of the protein. This protease is a water soluble, Mg2+-dependent enzyme that is located in the matrix. So the N-terminal sequence must reach the matrix, even if the protein ultimately will reside in the intermembrane space.

Residence in the matrix occurs in the absence of any other signal. If there is a membrane-targeting signal, however, it is activated by cleavage of the matrix-targeting signal. Then the remaining part of the leader (which is now N-terminal) causes the protein to take up its final destination.

The nature of the membrane-targeting signal is controversial. One model holds that the entire protein enters the matrix, after which the membrane-targeting signal causes it to be re-exported into or through the inner membrane. An alternative model proposes that the membrane-targeting sequence simply prevents the rest of the protein from following the leader through the inner membrane into the matrix. Whichever model applies, another protease (located within the intermembrane space) completes the removal of leader sequences.

Passage through chloroplast membranes is achieved in a similar manner. **Figure 8.41** illustrates the variety of locations for chloroplast proteins. They pass the outer and inner membranes of the envelope into the stroma, a process involving the same types of passage as into the mitochondrial matrix. But some proteins are transported yet further, across the stacks of the thylakoid membrane into the lumen. Proteins destined for the thylakoid membrane or lumen must cross the stroma en route.

Chloroplast targeting signals resemble mitochondrial targeting signals. The leader consists of ~50 amino acids, and the N-terminal half is needed to recognize the chloroplast envelope. A cleavage between positions 20-25 occurs during or following passage across the envelope, and proteins destined for the thylakoid membrane or lumen have a new N-terminal leader that guides recognition of the thylakoid membrane. There are several (at least four) different systems in the chloroplast that catalyze import of proteins into the thylakoid membrane.

The general principle governing protein transport into mitochondria and chloroplasts therefore is that the N-terminal part of the leader targets a protein to the organelle matrix, and an additional sequence (within the leader) is needed to localize the protein at the outer membrane, intermembrane space, or inner membrane. The additional sequence may function when it becomes N-terminal, after the first cleavage event.

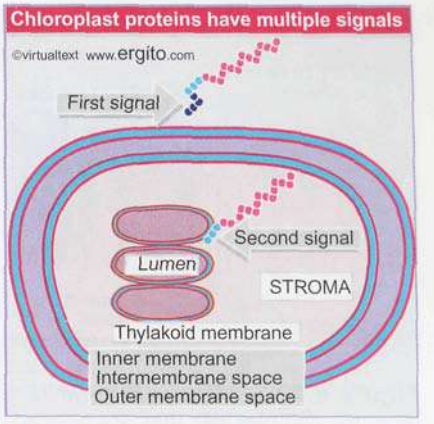


Figure 8.41 A protein approaches the chloroplast from the cytosol with a ~50 residue leader. The N-terminal half of the leader sponsors passage into the envelope or through it into the stroma. Cleavage occurs during envelope passage.

8.19 Inner and outer mitochondrial membranes have different translocons

Key Concepts

- Transport through the outer and inner mitochondrial membranes uses different receptor complexes.
- The TOM (outer membrane) complex is a large complex in which substrate proteins are directed to the Tom40 channel by one of two subcomplexes.
- Different TIM (inner membrane) complexes are used depending on whether the substrate protein is targeted to the inner membrane or to the lumen.
- Proteins pass directly from the TOM to the TIM complex.

There are different receptors for transport through each membrane in the chloroplast and mitochondrion. In the chloroplast they are called TOC and TIC, and in the mitochondrion they are called TOM and TIM, referring to the outer and inner membranes, respectively.

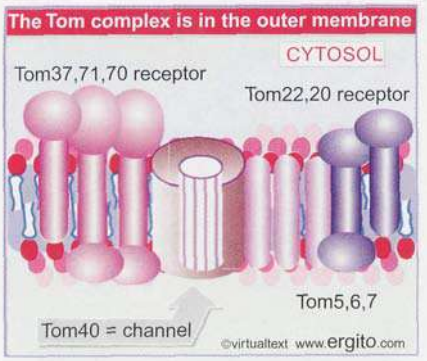


Figure 8.42 TOM proteins form receptor complex(es) that are needed for translocation across the mitochondrial outer membrane.

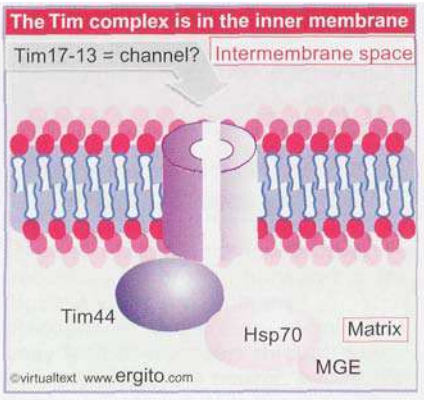


Figure 8.43 Tim proteins form the complex for translocation across the mitochondrial inner membrane.

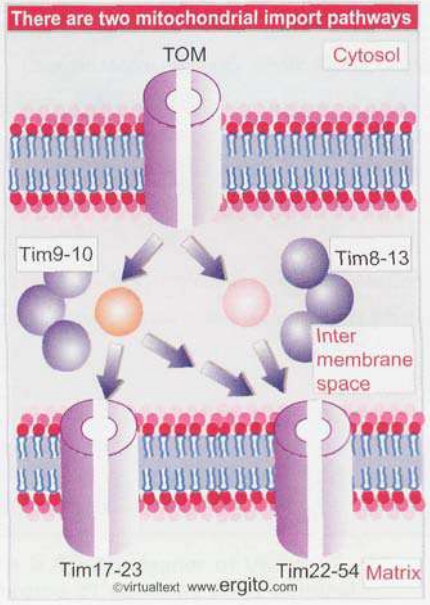


Figure 8.44 Tim9-10 takes proteins from TOM to either TIM complex, and Tim8-13 takes proteins to Tim22-54.

The TOM complex consists of ~9 proteins, many of which are integral membrane proteins. A general model for the complex is shown in **Figure 8.42.** The TOM aggregate has a size of >5 00 kD, with a diameter of ~138 A, and forms an ion-conducting channel. A complex contains 2-3 individual rings of diameter 75 A, each with a pore of diameter 20 A.

Tom40 is deeply imbedded in the membrane and provides the channel for translocation. It contacts preproteins as they pass through the outer membrane. It binds to three smaller proteins, Tom5,6,7, which may be components of the channel or assembly factors. There are two subcomplexes that provide surface receptors. Tom20,22 form a subcomplex with exposed domains in the cytosol. Most proteins that are imported into mitochondria are recognized by the Tom20,22 subcomplex, which is the primary receptor and recognizes the N-terminal sequence of the translocating protein. Tom37,70,71 provides a receptor for a smaller number of proteins that have internal targeting sequences.

When a protein is translocated through the TOM complex, it passes from a state in which it is exposed to the cytosol into a state in which it is exposed to the intermembrane space. However, it is not usually released, but instead is transferred directly to the TIM complex. It is possible to trap intermediates in which the leader is cleaved by the matrix protease, while a major part of the precursor remains exposed on the cytosolic surface of the envelope. This suggests that a protein spans the two membranes during passage. The TOM and TIM complexes do not appear to interact directly (or at least do not form a detectable stable complex), and they may therefore be linked simply by a protein in transit. When a translocating protein reaches the intermembrane space, the exposed residues may immediately bind to a TIM complex, while the rest of the protein continues to translocate through the TOM complex.

There are two TIM complexes in the inner membrane.

The Tim 17-23 complex translocates proteins to the lumen. Substrates are recognized by their possession of a positively charged N-terminal signal. Tim 17-23 are transmembrane proteins that comprise the channel. **Figure 8.43** shows that they are associated with Tim44 on the matrix side of the membrane. Tim 44 in turn binds the chaperone Hsp70. This is also associated with another chaperone, Mge, the counterpart to bacterial GrpE. This association ensures that when the imported protein reaches the matrix, it is bound by the Hsp70 chaperone. The high affinity of Hsp70 for the unfolded conformation of the protein as it emerges from the inner membrane helps to "pull" the protein through the channel.

A major chaperone activity in the mitochondrial matrix is provided by Hsp60 (which forms the same sort of structure as its counterpart GroEL). Association with Hsp60 is necessary for joining of the subunits of imported proteins that form oligomeric complexes. An imported protein may be "passed on" from Hsp70 to Hsp60 in the process of acquiring its proper conformation.

The Tim22-54 complex translocates proteins that reside in the inner membrane.

How does a translocating protein finds its way from the TOM complex to the appropriate TIM complex? Two protein complexes in the intermembrane space escort a translocating protein from TIM to TOM. The Tim9-10 and Tim8-13 complexes act as escorts for different sets of substrate proteins. Tim9-10 may direct its substrates to either Tim22-54 or Tim23-17, while Tim8-13 directs substrates only to Tim22-54. Some substrates do not use either Tim9-10 or Tim8-13, so other pathways must also exist. The pathways are summarized in **Figure 8.44.**

What is the role of the escorting complexes? They may be needed to help the protein exit from the TOM complex as well as for recognizing

the TIM complex. **Figure 8.45** shows that a translocating protein may pass directly from the TOM channel to the Tim9,10 complex, and then into the Tim22-54 channel.

A mitochondrial protein folds under different conditions before and after its passage through the membrane. Ionic conditions and the chaperones that are present are different in the cytosol and in the mitochondrial matrix. It is possible that a mitochondrial protein can attain its mature conformation *only* in the mitochondrion.

8.20 Peroxisomes employ another type of translocation system

Key Concepts

- Proteins are imported into peroxisomes in their fully folded state.
- They have either a PTS1 sequence at the C-terminus or a PTS2 sequence at the N-terminus.
- The receptor Pex5p binds the PTS1 sequence, and the receptor Pex7p binds the PTS2 sequence.
- The receptors are cytosolic proteins that shuttle into the peroxisome carrying a substrate protein and then return to the cytosol.

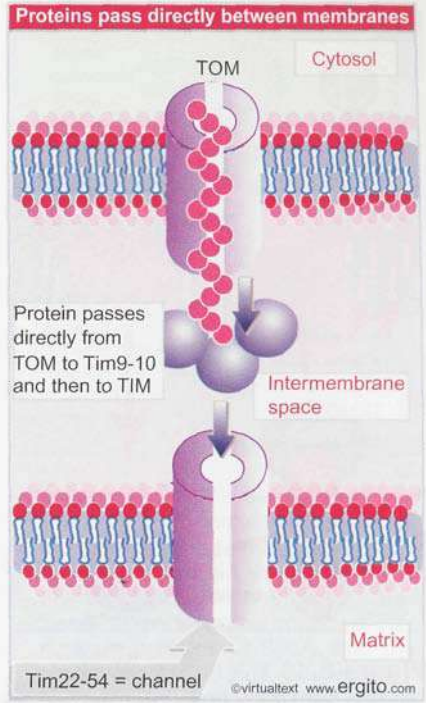


Figure 8.45 A translocating protein may be transferred directly from TOM to Tim22-54.

Peroxisomes are small bodies (05-1.5 μm diameter) enclosed by a single membrane. They contain enzymes concerned with oxygen utilization, which convert oxygen to hydrogen peroxide by removing hydrogen atoms from substrates. Catalase then uses the hydrogen peroxide to oxidize a variety of other substrates. Their activities are crucial for the cell. Since the fatal disease of Zellweger syndrome was found to be caused by lack of peroxisomes, > 15 human diseases have been linked to disorders in peroxisome function.

All of the components of the peroxisome are imported from the cytosol. Proteins that are required for peroxisome formation are called peroxins. Twenty three genes coding for peroxins have been identified, and human peroxisomal diseases have been mapped to 12 complementation groups, most identified with specific genes. Peroxisomes appear to be absent from cells that have null mutations in some of these genes. In some of these cases, introduction of a wild-type gene leads to the reappearance of peroxisomes. It has generally been assumed that, like other membrane-bounded organelles, peroxisomes can arise only by duplication of pre-existing peroxisomes. But these results raised the question of whether it might be possible to assemble them *de novo* from their components. In at least some cases, however, the absence of peroxins leaves the cells with peroxisomal ghosts—empty membrane bodies. Even when they cannot be easily seen, it is hard to exclude the possibility that there is some remnant that serves to regenerate the peroxisomes.

Transport of proteins to peroxisomes occurs post-translationally. Proteins that are imported into the matrix have either of two short sequences, called PTS1 and PTS2. The PTS1 signal is a tri- or tetrapeptide at the C-terminus. It was originally characterized as the sequence SKL (Ser-Lys-Leu) , but now a large variety of sequences have been shown to act as a PTS1 signal. The addition of a suitable sequence to the C-terminus of cytosolic proteins is sufficient to ensure their import into the organelle. The PTS2 signal is a sequence of 9 amino acids, again with much diversity, and this can be located near the N-terminus or internally. It is possible there may be a third type of sequence called PTS3.

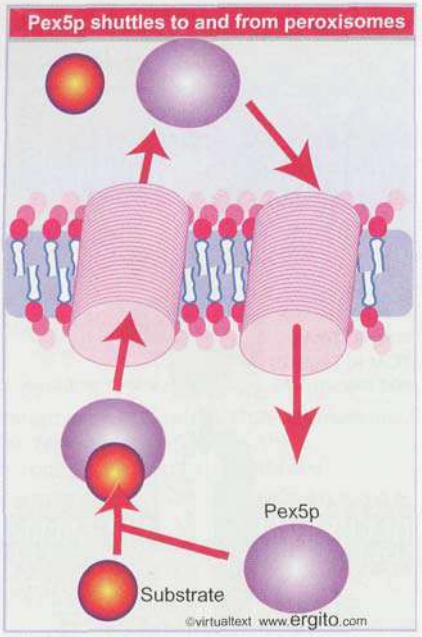


Figure 8.46 The Pex5p receptor binds a substrate protein in the cytosol, carries it across the membrane into the peroxisome, and then returns to the cytosol.

Several peroxisomal proteins are necessary for the import of proteins from the cytosol. The peroxisomal receptors that bind the two types of signals are called Pex5p and Pex7p, respectively. The other proteins are part of membrane-associated complexes concerned with the translocation reaction.

Transport into the peroxisome has unusual features that mark important differences from the system used for transport into other organelles.

Proteins can be imported into the peroxisome in their mature, fully-folded state. This contrasts with the requirement to unfold a protein for passage into the ER or mitochondrion, where it passes through a channel in the membrane into the organelle in something akin to an unfolded thread of amino acids. It is not clear how the structure of a preexisting channel could expand to permit this. One possibility is to resurrect an old idea and to suppose that the channel assembles around the substrate protein when it associates with the membrane.

The Pex5p and Pex7p receptors are not integral membrane proteins, but are largely cytosolic, with only a small proportion associated with peroxisomes. They behave in the same way, cycling between the peroxisome and the cytosol. **Figure 8.46** shows that the receptor binds a substrate protein in the cytosol, takes it to the peroxisome, moves with it through the membrane into the interior, and then returns to the cytosol to undertake another cycle. This shuttling behavior resembles the carrier system for import into the nucleus (see *8.28 Transport receptors carry cargo proteins through the pore).*

The import pathways converge at the perixosomal membrane, where Pex5p and Pex7p both interact with the same membrane protein complex, consisting of Pexl4p and Pexl3p. The receptors dock with this complex, and then several other peroxins are involved with the process of transport into the lumen. The details of the transport process are not yet clear.

Proteins that are incorporated into the peroxisomal membrane have a sequence called the mPTS, but little is known about the process of integration. Pex3p may be a key protein, because in its absence other proteins are not found in peroxisomal membranes. Pex3p has its own mPTS, which raises the question of how it enters the membrane. Perhaps it interacts with Pex3p that is already in the membrane. This bears on the question of whether peroxisomes can ever assemble *de novo*.

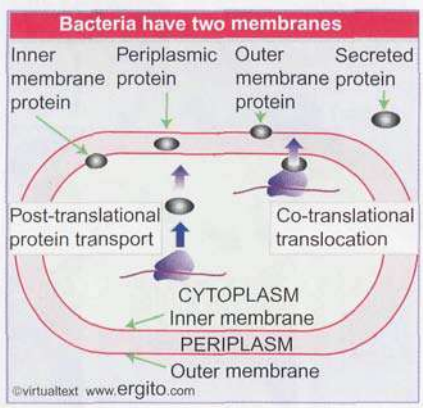


Figure 8.47 Bacterial proteins may be exported either post-translationally or co-translationally, and may be located within either membrane or the periplasmic space, or may be secreted.

8.21 Bacteria use both co-translational and post-translational translocation

Key Concepts

- Bacterial proteins that are exported to or through membranes use both post-translational and co-translational mechanisms.

The bacterial envelope consists of two membrane layers. The space between them is called the periplasm. Proteins are exported from the cytoplasm to reside in the envelope or to be secreted from the cell. The mechanisms of secretion from bacteria are similar to those characterized for eukaryotic cells, and we can recognize some related components. **Figure 8.47** shows that proteins that are exported from the cytoplasm have one of four fates:

- to be inserted into the inner membrane.
- to be translocated through the inner membrane to rest in the periplasm.
- to be inserted into the outer membrane.
- to be translocated through the outer membrane into the medium.

Different protein complexes in the inner membrane are responsible for transport of proteins depending on whether their fate is to pass through or stay within the inner membrane. This resembles the situation in mitochondria, where different complexes in each of the inner and outer membranes handle different subsets of protein substrates depending on their destinations (see *8.17 Post-translational membrane insertion depends on leader sequences)* A difference from import into organelles is that transfer in *E. coli* may be either co- or post-translational. Some proteins are secreted both co-translationally and post-translationally, and the relative kinetics of translation versus secretion through the membrane could determine the balance.

Exported bacterial proteins have N-terminal leader sequences, with a hydrophilic N-terminus and an adjacent hydrophobic core. The leader is cleaved by a signal peptidase that recognizes precursor forms of several exported proteins. The signal peptidase is an integral membrane protein, located in the inner membrane. Mutations in N-terminal leaders prevent secretion; they are suppressed by mutations in other genes, which are thus defined as components of the protein export apparatus. Several genes given the general description *sec* are implicated in coding for components of the secretory apparatus by the occurrence of mutations that block secretion of many or all exported proteins.

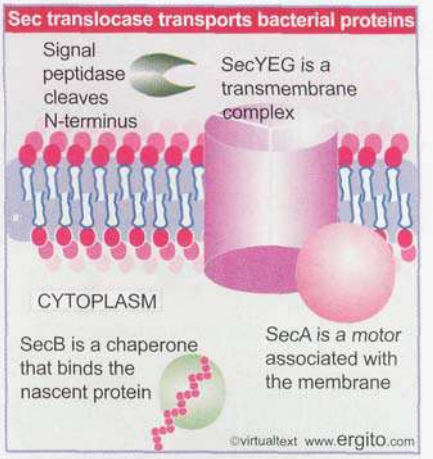


Figure 8.48 The Sec system has the SecYEG translocon embedded in the membrane, the SecA associated protein that pushes proteins through the channel, the SecB chaperone that transfers nascent proteins to SecA, and the signal peptidase that cleaves the N-terminal signal from the translocated protein.

8.22 The Sec system transports proteins into and through the inner membrane

Key Concepts

- The bacterial SecYEG translocon in the inner membrane is related to the eukaryotic Sec61 translocon.
- Various chaperones are involved in directing secreted proteins to the translocon.

There are several systems for transport through the inner membrane. The best characterized is the Sec system, whose components are shown in **Figure 8.48.** The translocon that is embedded in the membrane consists of three subunits that are related to the components of mammalian/yeast Sec61. Each of the subunits is an integral transmembrane protein. (SecY has 10 transmembrane segments and SecE has 3 transmembrane segments.) The functional translocon is a trimer with one copy of each subunit. The major pathway for directing proteins to the translocon consists of SecB and SecA. SecB is a chaperone that binds to the nascent protein to control its folding. It transfers the protein to SecA, which in turn transfers it to the translocon.

Figure 8.49 shows that there are two predominant ways of directing proteins to the Sec channel:

- the SecB chaperone;
- and the 4.5S RNA-based SRP.

Several chaperones can increase the efficiency of bacterial protein export by preventing premature folding; they include "trigger factor" (characterized as a chaperone that assists export), GroEL (see earlier), and SecB (identified as the product of one of the *sec* mutants). Although SecB is the least abundant of these proteins, it has the major role in promoting export. It has two functions. First, it behaves as a chaperone and binds to a nascent protein to retard folding. It cannot reverse the change in structure of a folded protein, so it does not function as an unfolding factor. Its role is therefore to inhibit improper folding of the

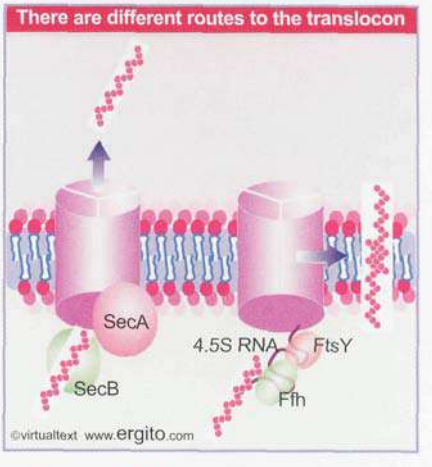


Figure 8.49 SecB/SecA transfer proteins to the translocon in order for them to pass through the membrane. 4.5S RNA transfers proteins that enter the membrane.

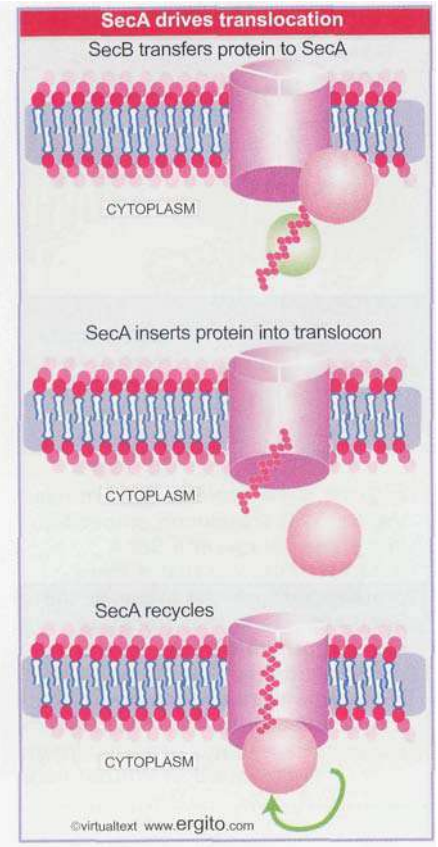


Figure 8.50 SecB transfers a nascent protein to SecA, which inserts the protein into the channel. Translocation requires hydrolysis of ATP and a protonmotive force. SecA undergoes cycles of association and dissociation with the channel and provides the motive force to push the protein through.

newly synthesized protein. Second, it has an affinity for the protein SecA. This allows it to target a precursor protein to the membrane. The SecB-SecYEG pathway is used for translocation of proteins that are secreted into the periplasm and is summarized in **Figure 8.50.**

SecA is a large peripheral membrane protein that has alternative ways to associate with the membrane. As a peripheral membrane protein, it associates with the membrane by virtue of its affinity for acidic lipids and for the SecY component of the translocon., which are part of a multisubunit complex that provides the translocase function. However, in the presence of other proteins (SecD and SecF), SecA can be found as a membrane-spanning protein. It probably provides the motor that pushes the substrate protein through the SecYEG translocon.

SecA recognizes both SecB and the precursor protein that it chaperones; probably features of the mature protein sequence as well as its leader are required for recognition. SecA has an ATPase activity that depends upon binding to lipids, SecY, and a precursor protein. The ATPase functions in a cyclical manner during translocation. After SecA binds a precursor protein, it binds ATP, and ~20 amino acids are translocated through the membrane. Hydrolysis of ATP is required to release the precursor from SecA. Then the cycle may be repeated. Precursor protein is bound again to provide the spur to bind more ATP, translocate another segment of protein, and release the precursor. SecA may alternate between the peripheral and integral membrane forms during translocation; with each cycle, a 30 kD domain of SecA may insert into the membrane and then retract.

Another process can also undertake translocation. When a precursor is released by SecA, it can be driven through the membrane by a protonmotive force (that is, an electrical potential across the membrane). This process cannot initiate transfer through the membrane, but can continue the process initiated by a cycle of SecA ATPase action. So after or between cycles of the SecA-ATP driven reaction, the protonmotive force can drive translocation of the precursor.

The *E. coli* ribonucleoprotein complex of 4.5S RNA with Ffh and FtsY proteins is a counterpart to the eukaryotic SRP (see *8.10 The SRP interacts with the SRP receptor).* It probably plays the role of keeping the nascent protein in an appropriate conformation until it interacts with other components of the secretory apparatus. It is needed for the secretion of some, but not all, proteins. As we see in Figure 8.49, its substrates are integral membrane proteins. The basis for differential selection of substrates is that the *E. coli* SRP recognizes an anchor sequence in the protein (anchor sequences by definition are present only in integral membrane proteins). Chloroplasts have counterparts to the Ffh and FtsY proteins, but do not require an RNA component.

8.23 Sec-independent translation systems in *E. coli*

Key Concepts

- *E. coli* and organelles have related systems for protein translocation.
- One system allows certain proteins to insert into membranes without a translocation apparatus.
- YidC is homologous to a mitochondrial system for transferring proteins into the inner membrane.
- The tat system transfers proteins with a twin arginine motif into the periplasmic space.

The most striking alternative system for protein translocation in *E. coli* is revealed by the coat protein of phage M13. **Figure 8.51**

By Book_Crazy [IND]

shows that this does not appear to require any translocation apparatus! It can insert post-translationally into protein-free liposomes. Targeting the protein to the membrane requires specific sequences (comprising basic residues) in the N- and C-terminal regions of the protein. They may interact with negatively charged heads of phospholipids. Then the protein enters the membrane by using hydrophobic groups in its N-terminal leader sequence and an internal anchor sequence. Hydrophobicity is the main driving force for translocation, but it can be assisted by a protonmotive force that is generated between the positively charged periplasmic side of the membrane and an acidic region in the protein. This drives the protein through the membrane, and leader peptidase can then cleave the N-terminal sequence. The generality of this mechanism in bacteria is unclear; it may apply only to the special case of bacteriophage coat proteins. Some chloroplast proteins may insert into the thylakoid membrane by a similar pathway.

Mutations in the gene *yidC* block insertion of proteins into the inner membrane. YidC is homologous to the protein Oxa1p that is required when proteins are inserted into the inner mitochondrial membrane from the matrix. It can function either independently of SecYEG or in conjunction with it. The insertion of some of the YidC-dependent proteins requires SecYEG, suggesting that YidC acts in conjunction with the translocon to divert the substrate into membrane insertion as opposed to secretion. Other proteins whose insertion depends on YidC do not require SecYEG: it seems likely that some other (unidentified) functions are required instead of the translocon.

The *Tat* system is named for its ability to transport proteins bearing a twin arginine targeting motif. It is responsible for translocation of proteins that have tightly bound cofactors. This may mean that they have limitations on their ability to unfold for passage through the membrane. This would be contrary to the principle of most translocation systems, where the protein passes through the membrane in an unfolded state, and then must be folded into its mature conformation after passage. This system is related to a system in the chloroplast thylakoid lumen called Hcf106. Both of these systems transport proteins into the periplasm.

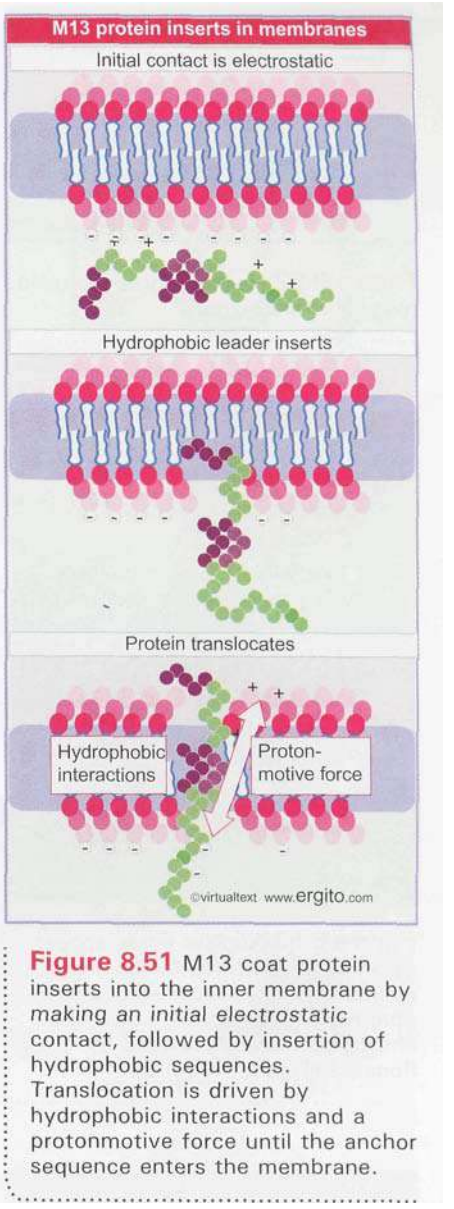


Figure 8.51 M13 coat protein inserts into the inner membrane by *making an initial electrostatic* contact, followed by insertion of hydrophobic sequences. Translocation is driven by hydrophobic interactions and a protonmotive force until the anchor sequence enters the membrane.

8.24 Pores are used for nuclear import and export

Key Concepts

- The same nuclear pores are used for importing proteins into the nucleus and for exporting proteins and RNA from the nucleus.

The nucleus is segregated from the cytoplasm by a layer of two membranes that constitute the nuclear envelope. The inner membrane contacts the nuclear lamina, providing in effect a surface layer for the nucleus. The outer membrane is continuous with the endoplasmic reticulum in the cytosol. The space between the two membranes is continuous with the lumen of the endoplasmic reticulum. The two membranes come into contact at openings called nuclear pore complexes. At the center of each complex is a pore that provides a water-soluble channel between nucleus and cytoplasm. This means that the nucleus and cytosol have the same ionic milieu. There are ~3000 pore complexes on the nuclear envelope of an animal cell.

Transport between nucleus and cytoplasm proceeds in both directions. Since all proteins are synthesized in the cytosol, any proteins required in the nucleus must be transported there. Since all RNA is synthesized in the

Nuclear pores work in both directions

Direction	Substrate	Passages /pore/min
Import	Histones	100
	Nonhistone proteins	100
	Ribosomal proteins	150
Export	Ribosomal subunits	~5
	mRNA	<1

Figure 8.52 Nuclear pores are used for import and export.

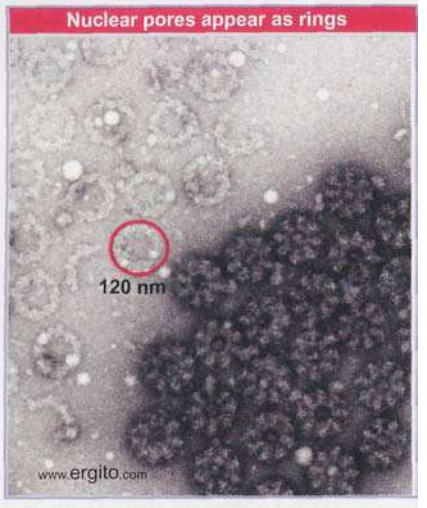


Figure 8.53 Nuclear pores appear as annular structures by electron microscopy. The circle around one pore has a diameter of 120 nm. Photograph kindly provided by Ronald Milligan.

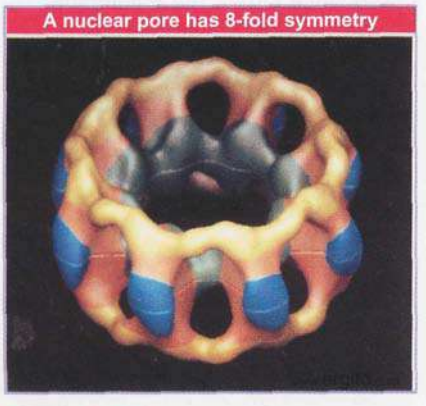


Figure 8.54 A model for the nuclear pore shows 8-fold symmetry. Two rings form the upper and lower surfaces (shown in yellow); they are connected by the spokes (shown in green on the inside and blue on the outside). Photograph kindly provided by Ronald Milligan.

nucleus, the entire cytoplasmic complement of RNA (mRNA, rRNA, tRNA, and other small RNAs) must be derived by export from the nucleus. The nuclear pores are used for both import and export of material. **Figure 8.52** summarizes the frequency with which the pores are used for some of the more prominent substrates.

We can form an impression of the magnitude of import by considering the histones, the major protein components of chromatin. In a dividing cell, enough histones must be imported into the nucleus during the period of DNA synthesis to associate with a diploid complement of chromosomes. Since histones form about half the protein mass of chromatin, we may conclude that overall about 200 chromosomal protein molecules must be imported through each pore per minute.

Uncertainties about the processing and stability of mRNA make it more difficult to calculate the number of mRNA molecules exported, but to account for the ~250,000 molecules of mRNA per cell probably requires ~1 event per pore per minute. The major RNA synthetic activity of the nucleus is of course the production of rRNA, which is exported in the form of assembled ribosomal subunits. Just to double the number of ribosomes during one cell cycle would require the export of ~5 ribosomal subunits (60S and 40S) through each pore per minute.

For ribosomal proteins to assemble with the rRNA, they must first be imported into the nucleus. So ribosomal proteins must shuttle into the nucleus as free proteins and out again as assembled ribosomal subunits. Given ~80 proteins per ribosome, their import must be comparable in magnitude to that of the chromosomal proteins.

8.25 Nuclear pores are large symmetrical structures

Key Concepts

- The nuclear pore is an annular structure with 8-fold symmetry.

How does a nuclear pore accommodate the transit of material of varied sizes and characteristics in either direction? Nuclear pore complexes have a uniform appearance when examined by microscopy. The pores can be released from the nuclear envelope by detergent, and **Figure 8.53** shows that they appear as annular structures, consisting of rosettes made of 8 spokes. **Figure 8.54** shows a model for the pore based on three-dimensional reconstruction of electron microscopic images. It consists of an upper ring and a lower ring, connected by a lattice of 8 structures.

The basis for the 8-fold symmetry is explained in terms of individual components in the schematic view from above shown in **Figure 8.55.** This includes the central structure of Figure 8.54, and extends it with an internal transporter and surrounding radial arms. The outside of the pore complex as such consists of a ring of diameter ~120 nm. The ring itself consists of 8 subunits. The 8 radial arms outside the ring may be responsible for anchoring the pore complex in the nuclear envelope; they penetrate the membrane. The 8 interior spokes project from the ring, closing the opening to a diameter of ~48 nm. Within this region is the transporter, which contains a pore that approximates a cylinder <10 nm in diameter.

The pore provides a passage across the outer and inner membranes of the nuclear envelope. As illustrated in **Figure 8.56,** the side view has two-fold symmetry about a horizontal axis in the plane of the nuclear envelope. There are matching annuli at the outer and inner membranes, comprising the surfaces that project into the cytosol and into the nucleus, and each is connected to the spokes, which form a central ring. (Only 2 of

the 8 spokes are seen in this side view.) The spokes are symmetrical about the horizontal axis. The central pore projects for the distance across the envelope. Sometimes material can be seen within the pore, but it has been difficult to equate such views with the transport of any particular material.

The size of the nuclear pore complex corresponds to a total mass ~50 X 10^6 daltons (compare this with the 80S ribosome at 4 X 10^6 daltons). We can identify the smallest repeating component by using the 8-fold symmetry as seen in cross-section (see Figure 8.55) and the 2-fold symmetry seen from the side (see Figure 8.56). This divides the scaffold into 16 identical units. Each of these units consists of ~30 different proteins, most often each present in 1-2 copies per unit. The central pore constitutes only a small part of the overall complex.

Individual protein components of the nuclear pore can be localized by immunoelectron microscopy. Most proteins are found on both sides of the pore; very few are found on only one side. This supports the view of the complex as a symmetrical structure built from identical assemblies. Some of the pore complex proteins are transmembrane proteins; they probably help to anchor the complex in the envelope.

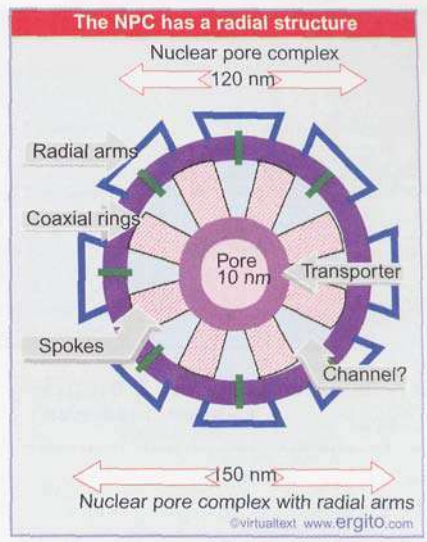


Figure 8.55 The outsides of the nuclear coaxial (cytoplasmic and nucleoplasmic) rings are connected to radial arms. The interior is connected to spokes that project towards the transporter that contains the central pore.

8.26 The nuclear pore is a size-dependent sieve for smaller material

Key Concepts

- The central channel is large enough to allow proteins of <50 kD to pass.
- Larger proteins must require the channel to open wider for passage.

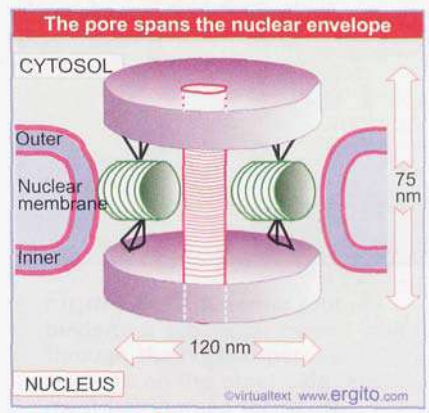


Figure 8.56 The nuclear pore complex spans the nuclear envelope by means of a triple ring structure.

The ability of compounds to diffuse freely through the pores is limited by their size. **Figure 8.57** summarizes the results of two sets of experiments in which material was injected into the cytoplasm, and its entry into the nucleus was followed over 24 hours. Using dextrans (large saccharides) of different sizes shows that the smallest size equilibrates very rapidly, with just over half of the material localized in the nucleus within minutes. As the size of the dextran increases, entry into the nucleus becomes limited. By a diameter of 7 nm, virtually no dextran can pass through the nuclear pore. Analogous results are obtained with proteins (see panel on right of figure), where there is a progressive reduction in the proportion in the nucleus as the size increases.

It is convenient to consider the material in three size classes:

Molecules of <5000 daltons that are injected into the cytoplasm appear virtually instantaneously in the nucleus: we may conclude that the nuclear envelope is freely permeable to ions, nucleotides, and other small molecules.

Proteins of 5-50 kD diffuse at a rate that is inversely related to their size, presumably determined by random contacts with, and passage through, the pore. It takes a few hours for the levels of an injected protein to equilibrate between cytoplasm and nucleus. We may conclude that small proteins can enter the nucleus by passive diffusion (but they may also be actively transported). The nuclear envelope in effect provides a mesh or molecular sieve that permits passage of material <50 kD. A globular protein of 50 kD in mass would have a diameter of ~5 nm if it were spherical.

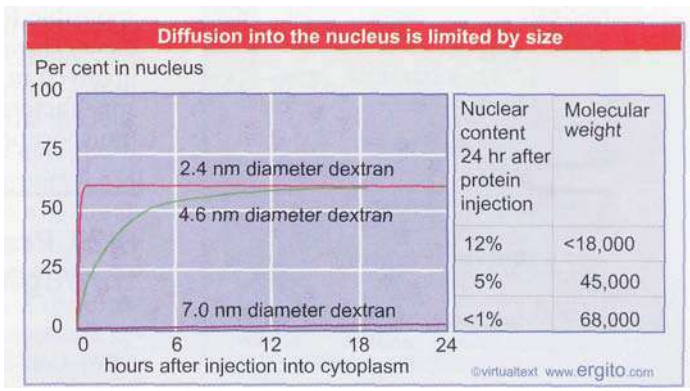


Figure 8.57 Small molecules and macromolecules can enter the nucleus freely, but diffusion is limited by size.

- Proteins >50 kD in size do not enter the nucleus by passive diffusion; a mechanism of active transport must be required for their passage.

Proteins are transported through the central pore. Electron-dense objects can sometimes be seen in the pore, and are usually assumed to represent material being transported (although this has not been proven). It is likely that the size limits on diffusion into the nucleus are determined by the diameter of the central channel. However, surrounding this channel are eight smaller openings, and it remains possible that smaller material can pass through them.

The ability of small proteins to diffuse through the pore means that, in the absence of any intervention, they will equilibrate between cytosol and nucleus. However, the distribution will be influenced by other interactions; for example, a small protein that is a component of chromatin could be bound to chromatin after it has diffused into the nucleus, and therefore will be largely localized in the nucleus. Larger proteins must use an active transport mechanism that overcomes the apparent size restriction of the pores. Also, active transport must be used for any protein that requires transport against a concentration gradient (for example, a protein that is localized freely in the nucleoplasm).

Transport through the pore has been characterized by using colloidal gold particles coated with a nuclear protein. When these particles are injected into the cytoplasm, they cluster at the nuclear pores, and then accumulate in the nucleus. This suggests that the pore structure can widen to accommodate objects of the size of the coated gold particles (~20 nm). Similar experiments have shown that gold particles coated with polynucleotides can be exported from the nucleus via pores. Following a simultaneous injection of RNA-gold particles of one size together with protein-gold particles of another size, pores can be seen to have both sizes of particles, which suggests that the same pores can be used for export and import.

The rigidity of the gold particle excludes the possibility that transport through the pore requires the protein to change into a conformation with a diameter physically smaller than the pore. We conclude that the nuclear pore has a "gating" mechanism that allows the interior to expand as material passes through. Pores engaged in transporting material appear to be opened to a diameter of ~20 nm, possibly by a mechanism akin to the iris of a camera lens. It is possible that two irises, one connected with the cytoplasmic ring and one connected with the nucleoplasmic ring, open in turn as material proceeds through the pore. Very large substrates, such as exported ribonucleoprotein particles, may have to change their conformation to conform with the limit of 20 nm.

We believe that all pores are identical. The nuclear pore complex provides a structural framework that supports the proteins actually responsible for binding and transporting material into (or out of) the nucleus. However, it does not include all of the active components that are involved in binding and translocation from one side to the other. Accessory factors that associate with the pore are responsible for the actual transport process.

8.27 Proteins require signals to be transported through the pore

Key Concepts

- The NILS and NES consist of short sequences that are necessary and sufficient for proteins to be transported through the pores into or out of the nucleus.

To be transported through the nuclear pore, a protein must have a special signal in its sequence. The most common motif responsible for import into the nucleus is called the *nuclear localization signal* (NLS). Its presence in a cytosolic protein is necessary and sufficient to sponsor import into the nucleus. Mutation of the signal can prevent the protein from entering the nucleus.

The summary of nuclear localization signals in **Figure 8.58** shows that there is no apparent conservation of sequence of NLS signals; perhaps the shape of the region or its basicity are the important features. Many NLS sequences take the form of a short, rather basic stretch of amino acids. Often there is a proline residue to break α-helix formation upstream of the basic residues. Hydrophobic residues are rare. Some NLS signals are bipartite and require two separate short clusters. Competition studies suggest that NLS sequences are interchangeable, suggesting that they are all recognized by the same import system.

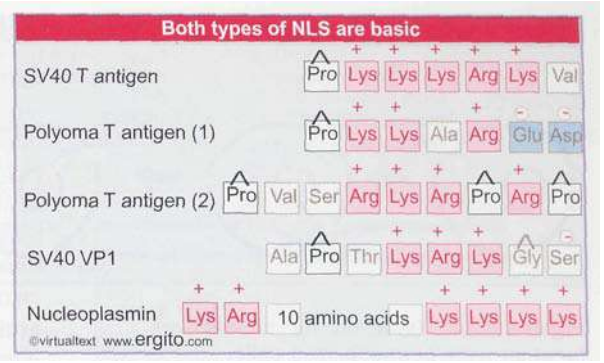


Figure 8.58 Nuclear localization signals have basic residues.

Many exported proteins have a common type of signal that is necessary and sufficient for the protein to move from the nucleus to the cytosol. It is called an NES (nuclear export signal), and typically consists of an ~10 amino acid sequence. The only common feature in the NES sequences in different proteins is a pattern of conserved leucines.

A protein may have both an NLS and an NES, the former used for its import into the nucleus, and the latter for its export. They may function constitutively, or their use can be regulated, for example, by association with other proteins that obscure or expose the relevant sequences.

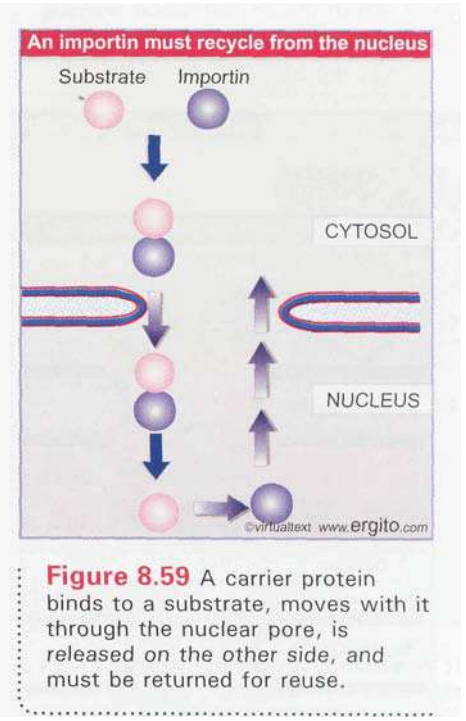


Figure 8.59 A carrier protein binds to a substrate, moves with it through the nuclear pore, is released on the other side, and must be returned for reuse.

8.28 Transport receptors carry cargo proteins through the pore

Key Concepts

- Transport receptors have the dual properties of recognizing **NLS** or NES sequences and binding to the nuclear pore.
- Exportins transport substrates from nucleus to cytoplasm; importins transport substrates from cytoplasm to nucleus.
- Exportins and importins interact with nucleoporins in the pore.

The basic principle of import and export is illustrated in **Figure 8.59.** A carrier protein (or transport receptor) takes the substrate through the pore. The transport receptor must then be returned across the membrane to function in another cycle. The transport receptors are classified according to the direction in which they transport the cargo. Importins bind the cargo in the cytoplasm and release it in the nucleus. Exportins bind the cargo in the nucleus and release it in the cytoplasm.

There are multiple pathways for import and export. Transport for all of the substrates for any particular pathway can be inhibited by saturating that pathway with one of its substrates. **Figure 8.60** summarizes the independent pathways. At least two different types of pathways exist for import of proteins; and each class of RNA is exported by a different system. Each transport receptor recognizes a particular type of sequence in its substrate, thus defining the specificity of the system.

The first handle on the process of import was provided by systems for transport that depend upon the presence of an NLS in the substrate protein. An *in vitro* assay for nuclear pore import has been developed by using permeabilized cells. When cells are treated with digitonin, the

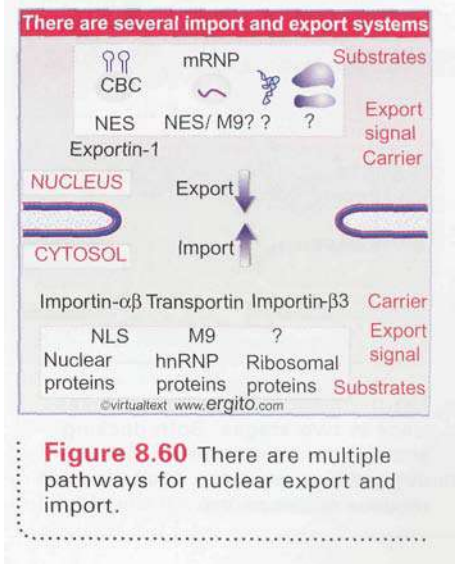


Figure 8.60 There are multiple pathways for nuclear export and import.

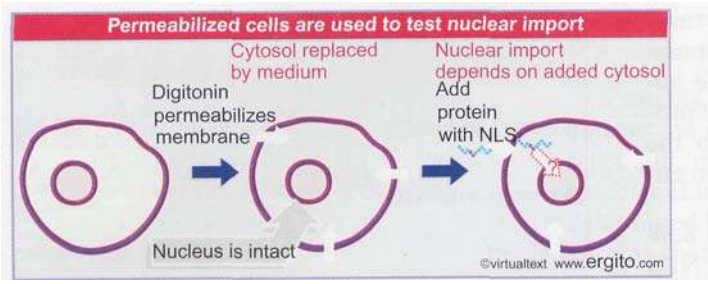


Figure 8.61 The assay for nuclear pore function uses permeabilized cells.

plasma membrane becomes permeable, but the nuclear envelope remains intact. Labeled proteins can be imported into the nucleus in a process that is dependent upon the provision of cytosolic components. **Figure 8.61** shows how this system has been used to characterize the transport process.

Transport can be divided into two stages: docking, which consists of binding to the pore; and translocation, which consists of movement through it. In the absence of ATP, proteins containing a nuclear import signal can bind at the pore, but they remain at the cytoplasmic face. A cytosolic fraction is needed for binding. When ATP is provided, proteins can be translocated through the pore. A different cytosolic fraction is needed to support translocation. The need for cytosolic fractions at both stages reinforces the view of the pore as a structure that provides the framework for transport, but that does not provide all of the necessary facilities for handling the substrates.

Figure 8.62 summarizes the functions of the components involved in nuclear import. The transport receptor is the key intermediate in the docking reaction. It can bind to both the nuclear pore and the cargo protein. Some transport receptors are single proteins, such as transportin or importin-β3, which undertake both binding reactions. Others are dimers in which one subunit binds to the pore, and the other is an adaptor that binds to the cargo protein. In the best characterized case, importin-αβ has a β subunit that binds to the nuclear pore. The a subunit binds proteins that have an NLS sequence. The single protein receptors transportin and importin-β3 are related to the β subunit of importin-αβ.

Translocation through the pore is inhibited by wheat germ agglutinin, a lectin (glycoprotein). The component proteins of the pore that bind to lectins were originally called nucleoporins. Note that nucleoporin has since come to be used to mean any component of the nuclear pore complex. The lectin-binding components are localized at or near the region of the central pore, and appear to be located on both sides of the nuclear envelope. When they are removed, pore complexes remain normal in appearance, but can no longer function to transport large material. Material smaller than the pore size continues to be able to move through by diffusion. When the lectin-binding proteins are added back, they restore full activity to the deficient pores. This suggests that they are needed for active transport of material larger than the resting diameter. Some of these proteins have some simple peptide repeating motifs (GKFG, FG, FXFG), and it is probably these motifs that bind the importin-β carrier proteins. They are called FG-nucleoporins.

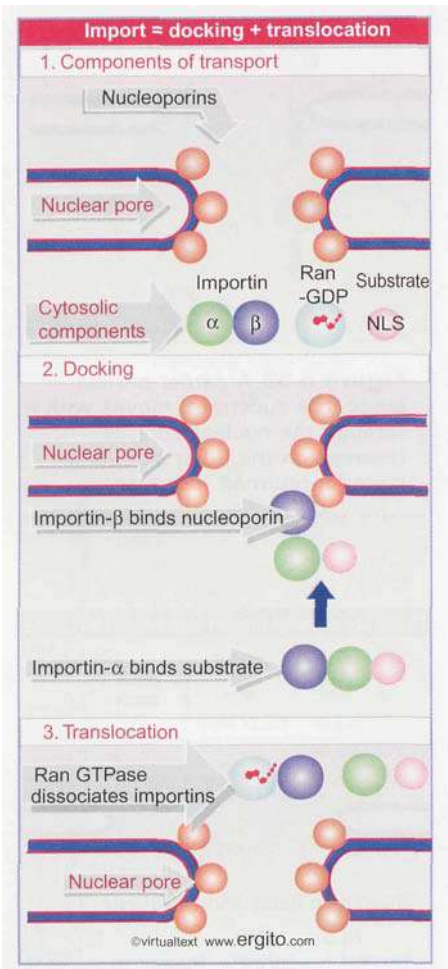


Figure 8.62 Nuclear import takes place in two stages. Both docking and translocation depend on cytosolic components. Translocation requires nucleoporins.

8.29 Ran controls the direction of transport

Key Concepts

- The nucleus contains Ran-GTP, which stabilizes export complexes, while the cytosol contains Ran-GDP, which stabilizes import complexes.
- Movement through the nuclear pore does not involve a motor.

How does material move through the pore? One of the striking features of the composition of the nuclear pore complex is the lack of any protein with motor activity. This suggests that movement must depend on the affinity of the carrier proteins for the pore itself. But now we have to explain how transport can work for both import and export. The answer lies in the properties of the monomeric G-protein, Ran.

By Book_Crazy [IND]

The cytosolic fraction that supports translocation has two active components. One is Ran; the other may be involved in targeting Ran to the nuclear pore. Ran is a typical monomeric G-protein that can exist in either the GTP-bound or GDP-bound state. Its GTPase activity generates Ran-GDP. Then an exchange factor is needed to replace GDP with GTP to regenerate Ran-GTP.

The directionality of nuclear import is controlled by the state of Ran. **Figure 8.63** shows that conditions in the nucleus and cytosol differ so that typically there is Ran-GTP in the nucleus, but there is Ran-GDP in the cytosol. The reason for this difference is an asymmetric distribution of two proteins that act on Ran. The nucleus contains Ran-GEF, which stimulates replacement of GDP by GTP, thus converting Ran-GDP to Ran-GTP. (In fact, this protein, also known as Rcc1, is localized to chromatin.) The cytoplasm contains Ran-GAP, which causes the GTP to be hydrolyzed to GDP. The Ran-GAP is localized on the surface of the cytoplasmic side of the nuclear pore complex, together with a Ran-binding protein (RanBP1) that stimulates its activity.

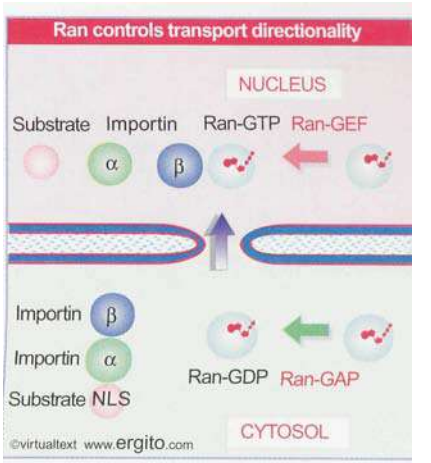


Figure 8.63 The state of the guanine nucleotide bound to Ran controls directionality of nuclear import and export.

Export complexes are stable in the presence of Ran-GTP, whereas import complexes are stable in the presence of Ran-GDP. So export complexes are driven to form in the nucleus and dissociate in the cytosol, whereas the reverse is true of import complexes. The reaction has been best characterized for the complex of importin-αβ with an NLS-containing protein. The triple complex is stable in the presence of Ran-GDP, and thus can form in the cytosol. However, Ran-GTP causes importin-α to dissociate from importin-β. This leads to release of the substrate protein in the nucleus. The effect of Ran-GTP in causing the importin dimer to release its substrate is also important at mitosis, when importins release proteins that trigger the attachment of microtubules to the spindle (see *29.23 A monomeric G protein controls spindle assembly).*

The crystal structures of complexes containing importin-β show how it binds to importin-a and to Ran. Importin-β consists of a series of repeating units coiled into a superhelix. The individual repeating unit (called HEAT) itself consists of two α-helices (HEAT-A and HEAT-B). Importin-a has a similar structure, but in addition has a domain (IBB for importin-binding domain) that binds importin-β. **Figure 8.64** shows that importin-β winds around the IBB of importin-a, making a tightly integrated dimer.

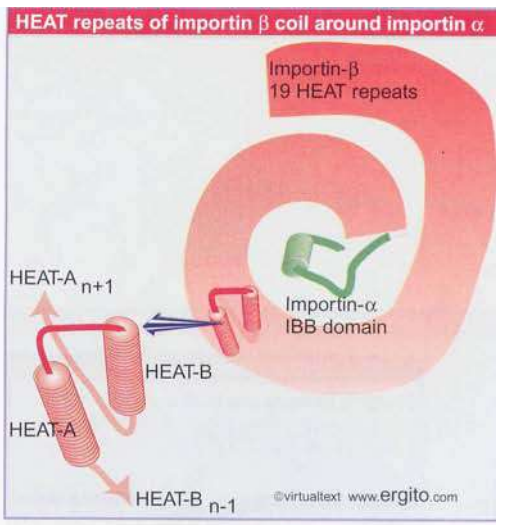


Figure 8.64 Importin-β consists of 19 HEAT repeats organized in a superhelix. Each HEAT unit consists of two α-helices (A and B) lying at an angle to one another. Importin-β is folded tightly around the IBB domain of importin-α.

Figure **8.65** shows that Ran-GTP binds tightly to two of the HEAT repeats in importin-β. Its binding site partially overlaps the binding site of importin-a. This explains why binding of Ran-GTP displaces importin-a (and the NLS of the cargo protein) from importin-β. There is a large structural change in Ran when GTP is hydrolyzed to GDP (involving the unfolding of a helical region), which explains why the displacement reaction is specific for Ran-GTP.

As pointed out in Figure 8.59, in order to function more than once, an importin must return to the cytosol after taking its substrate into the nucleus. In effect, when an importin is released in the nucleus, it must become a substrate for an exportin! Such a reaction has been characterized for importin-a, which is bound in the presence of Ran-GTP by a protein called CAS. CAS behaves in a similar manner to importin-β, except that it moves in the opposite direction. Like exportins, it dissociates from its substrate (importin-a) when Ran-GTP is hydrolyzed in the cytosol.

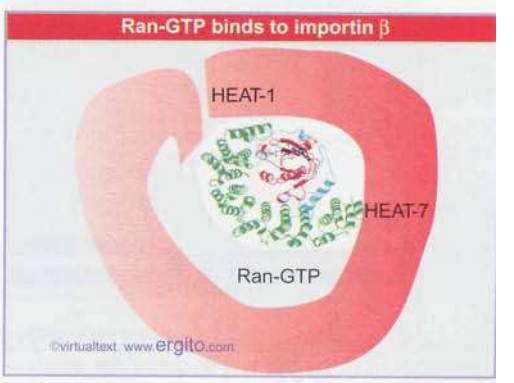


Figure 8.65 Importin-β binds Ran-GTP through close contacts to the N-terminal HEAT repeats and to repeats 7-8.

The NES binds to exportin-1, which is related in sequence to importin-β. Exportin-1 is needed for the export of U snRNAs, some proteins, and possibly (some) mRNAs. It binds the NES motif; it binds nucleoporins; and it binds Ran-GTP. The complex forms in the presence of Ran-GTP; hydrolysis of the GTP to generate Ran-GDP is accompanied by dissociation of the complex. So Ran controls directionality of export in the reverse sense from its control of import: because Ran-GTP is high in the nucleus, the complex forms there; because it becomes Ran-GDP in the cytosol, the complex dissociates there.

It was thought for a long time that energy for transport would be provided by the hydrolysis of GTP by Ran. However, this is not exactly true. Transport through the pore does not itself require energy. But the hydrolysis of GTP is required for a second cycle of transport, because it is necessary to generate Ran-GDP.

How does the importin-substrate complex cross the nuclear pore? It has a considerable distance to travel, ~200 nm. The pore itself has a symmetrical structure, and very few of its component proteins are concentrated on one side or the other. There are no obvious markers that might be used to indicate directionality. However, it is controversial whether the directionality of transport is intrinsic to the pore. At one time it was thought that it could be reversed by reversing the relative concentrations of Ran-GDP and Ran-GTP on either side, but more recent experiments suggest that it cannot. This leaves us with two types of model: there is some unknown intrinsic directionality resulting from an uneven distribution of proteins within the pore; or transport is stochastic, with receptor-substrate complexes bumbling through the pore until they emerge on the other side, where conditions ensure dissociation.

Importins and exportins move through the pore by a process of facilitated diffusion, but the details are still unclear. The crucial property that enables the importins and exportins to translocate through the pore is their ability to interact with the FG-nucleoporins by binding to the hydrophobic (Phe-Gly) repeats. The various transport receptors compete with one another to bind to the FG-nucleoporins, which argues that there is a common mechanism of translocation. FG-nucleoporins are distributed throughout the central pore, with a concentration at the midpoint. One model suggests that the hydrophobic repeats interact in this region to form a large mesh. The ability of transport receptors to interact with the repeats allows them to "dissolve" in the mesh and thus to pass through it. The different rates at which different substrates are translocated by the pore is determined by how easily they can be incorporated into the mesh. Some crucial questions remain unanswered, especially what determines directionality within the pore itself, since the process seems to be too rapid to be the result of a random walk that is ended by dissociation of the complex on the appropriate side of the pore.

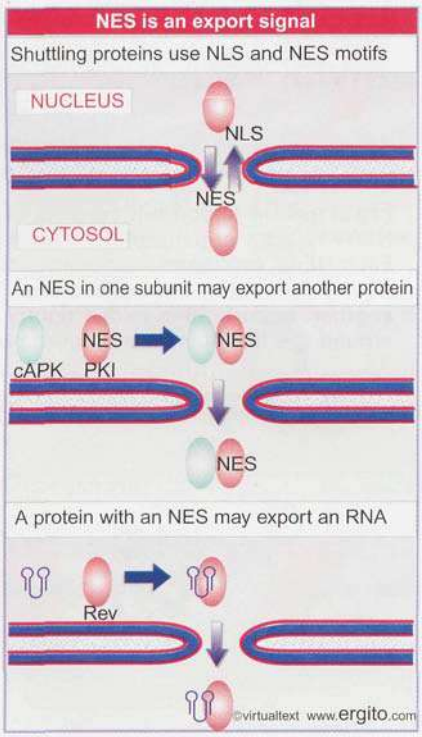


Figure 8.66 The common feature in proteins that are exported from the nucleus to the cytosol is the presence of an NES.

8.30 RNA is exported by several systems

Key Concepts

- There are (at least) three export systems for RNA.
- Each consists of an exportin that binds particular types of RNA.

Export systems have similar components to import systems. **Figure 8.66** illustrates some examples of export systems. The major substrates for export from the nucleus are ribonucleoproteins—ribosomes, mRNPs, snRNPs, and tRNA-protein complexes. In the first three cases, one of the protein components of the complex may be responsible for export (for example, for snRNAs it is the cap-binding complex (CBC). tRNA is bound directly by a specific export protein.

Export of mRNA has similar requirements to import. Mutations in a yeast FG-nucleoporin block export of RNA from the nucleus without affecting import of proteins. This suggests the possibility that the apparatus is similar for both import and export, but could have components that confer directionality or specificity for particular substrates. There is evidence for diversity in the export apparatus; using an assay for export of microinjected RNAs from the nucleus of the Xenopus oocyte,

By Book_Crazy [IND]

tRNAs, other small RNAs, and mRNAs each saturate transport only of their own class. This suggests that there are at least three groups of exported RNAs.

Some proteins "shuttle" between the cytoplasm and nucleus; they remain only briefly in either compartment before cycling back to the other. This behavior is characteristic of certain proteins that are bound to poly(A)$^+$ RNA in both the nucleus and the cytoplasm. The motif responsible for transport in one such protein (M9) has a single amino acid stretch that functions as both an import and export signal, and is therefore responsible for movement in both directions.

One particular issue with the export of mRNA is how to distinguish the final, processed mRNA from precursors that are not fully processed (for example, which retain some introns). Part of the answer may lie in the relative timing of events. Processing is connected with transcription in such a way that it is likely to be completed by the time the RNA is released from DNA. However, there are also some specific links that may connect export to the preceding events. One of the mRNA-binding proteins that is exported from the nucleus bound to mRNA attaches to the RNA via an interaction with the transcription apparatus when transcription is initiated. This suggests that the proteins involved in exporting mRNA may become complexed with it at a very early stage of its production. And then mRNAs that are spliced may require splicing to occur in order for other components of the export apparatus to bind to the mRNA (see *24.10 Splicing is connected to export of mRNA).* The proteins bound to the mRNA then interact with a protein called Mex67 (in yeast) or Tap (in animal cells), which is unrelated to the exportin or importin families, but can interact directly with nucleoporins.

8.31 Ubiquitination targets proteins for degradation

Key Concepts

Ubiquitin is added to proteins that are targeted for degradation by an apparatus consisting of three components.

A major pathway for protein degradation involves two stages: first the protein is targeted; and then it is proteolysed by a large complex that we describe in the next section. A small polypeptide called ubiquitin is connected by a covalent link to the substrate protein that is to be degraded. **Figure 8.67** shows that there are three components of the ubiquitination system.

- The ubiquitin-activating enzyme, E1, utilizes the cleavage of ATP to link itself via a high energy thiolester bond from a Cys residue to the C-terminal Gly residue of ubiquitin.
- The ubiquitin is then transferred to the ubiquitin-conjugating enzyme, E2.
- The ubiquitin ligase, E3, transfers the ubiquitin from E2 to form an isopeptide bond to the e NH_2 group of a Lys in the substrate protein.

Ubiquitin is released from a degraded substrate by an isopeptidase.

Responsibility for choosing substrate proteins to be ubiquitinated lies with both E2 and E3. In many cases, E3 selects the substrate protein by binding to it before transfer of ubiquitin is initiated. A cell may contain several E3 proteins that use different criteria for selecting substrates. There are also multiple varieties of E2, and they also may play a role in targeting substrate proteins, sometimes independently of E3.

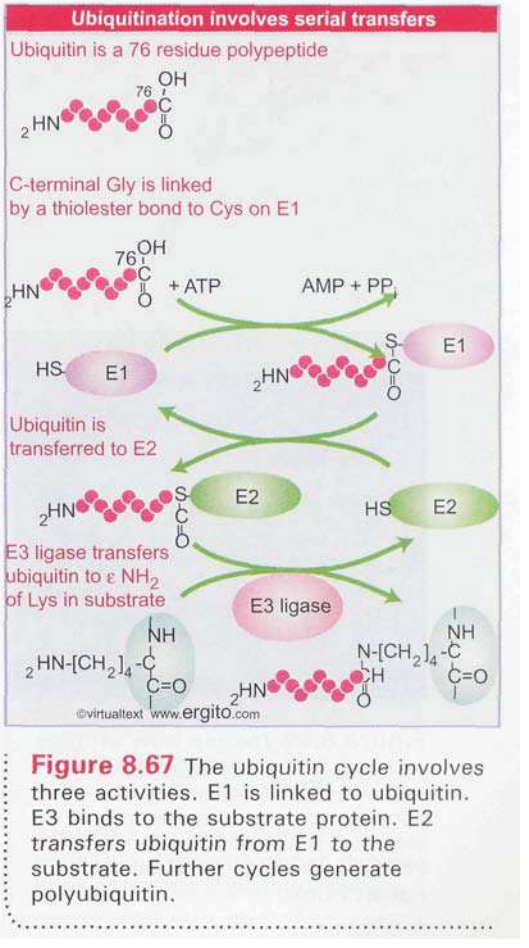


Figure 8.67 The ubiquitin cycle involves three activities. E1 is linked to ubiquitin. E3 binds to the substrate protein. E2 transfers ubiquitin from E1 to the substrate. Further cycles generate polyubiquitin.

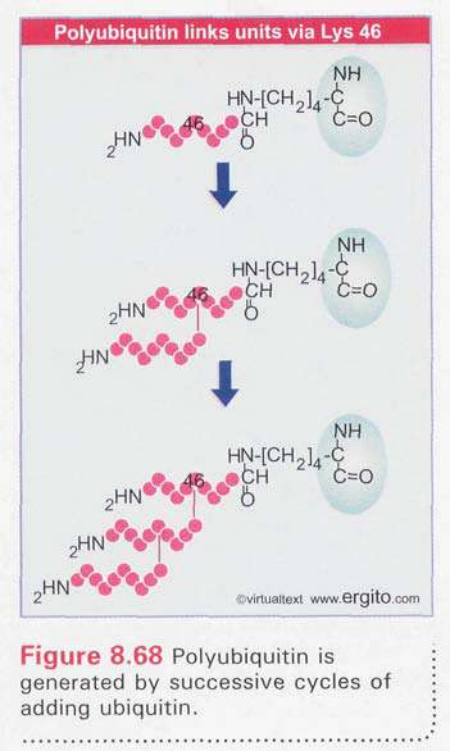


Figure 8.68 Polyubiquitin is generated by successive cycles of adding ubiquitin.

The addition of a single ubiquitin residue to a substrate protein is not sufficient to cause its degradation. **Figure 8.68** shows that further ubiquitin residues are added to form a polyubiquitin chain, in which each additional ubiquitin is added to the Lys at position 46 of the preceding ubiquitin. The formation of polyubiquitin is a signal for the proteasome to degrade the protein.

Targeting for degradation by the proteasome in the cytosol is the major function of ubiquitination. However, it also targets plasma membrane proteins for degradation in lysosomes, and possibly may have other regulatory effects.

In addition to ubiquitin, there are ubiquitin-like proteins that modify target proteins in a similar way. The best characterized of these is SUMO (also known as Sentrin). A difference between the ubiquitin system and the SUMO system is that only a single SUMO residue is added to a target protein, compared with polyubiquitination. The consequences of sumoylation are not entirely clear; it may be concerned with protein localization or with protection against ubiquitination. SUMO has fewer targets than ubiquitin, but they often include important cellular proteins.

Ubiquitination (or sumoylation) can be reversed by proteases that cleave the conjugate from the target protein, so there is the potential for significant complexity in regulation.

8.32 The proteasome is a large machine that degrades ubiquitinated proteins

Key Concepts

- Protein degradation in the cytosol is catalyzed by the proteasome.
- The 20S proteasome consists of four stacked rings each containing 7 subunits.
- The 26S proteasome is formed in eukaryotes when caps associate with one or both ends of the 20S proteasome.
- Caps pass ubiquitinated proteins to the core for proteolysis.
- The proteasome contains several different protease activities.

Figure 8.69 The top view of the archaeal 20S proteasome shows a hollow cylinder consisting of heptameric rings of α subunits (red) and β subunits (blue). Photograph kindly provided by Robert Huber.

What determines the stability of proteins? A cell contains many proteases, with varying specificities. We may divide them into three general groups:

- Some proteases are involved in specific processing events to generate mature proteins that are smaller than the precursors. Such proteases are involved in a variety of activities, including cleaving the signal sequence from a secreted protein, and cleaving cytosolic enzymes into their mature forms. The caspase group of proteases are involved a pathway that leads to cell death (see *29.27 A common pathway for apoptosis functions via caspases).*
- Lysosomes are membrane-bounded organelles that degrade proteins imported into the cell; we discuss this process in the context of protein transport through membranes in *27.15 Receptors recycle via endocytosis*.
- The proteasome is a large complex that degrades cytosolic proteins. It is involved in both general degradation (the complete conversion of a protein into small fragments) and in certain specific processing events. The major substrates for complete degradation are proteins that have been misfolded—this is basically a quality control system—and certain proteins whose degradation is a regulatory event, for example, to allow progress through the cell cycle.

By Book_Crazy [IND]

The proteasome was originally discovered as a large complex that degrades proteins conjugated to ubiquitin. It exists in two forms. A 20S complex of ~700 kD has protease activity. Additional proteins convert the complex to a 26S form of ~2000 kD; they are regulatory subunits that confer specificity—for example, for binding to ubiquitin conjugates. ATP cleavage is required for the conversion from 20S to 26S, and is also required later in the reaction for cleaving peptide bonds, releasing the products, etc.

The 20S complex takes the form of a hollow cylinder, and the additional components of the 26S complex are attached to the ends of the cylinder, making a dumbbell. Basically, the active sites are contained in the interior of a barrel, and access is obtained through relatively narrow channels, typically allowing only access only to unfolded proteins. This protects normal, mature proteins from adventitious degradation.

This general type of structure is common to ATP-dependent proteases. For example, the ClpAP protease in *E. coli*, which is not related by sequence to the proteasome, has a structure in which the ClpP protease forms two rings of 7 subunits each, with the proteolytic activities contained in a central cavity. ClpA is hexameric and stacks on to the ClpP complex (which implies an interesting symmetry mismatch in the ClpAP complex). ClpA provides the ATPase activity. It unfolds the substrates and translocates them in denatured form through a narrow passage into the ClpP cavity, where they are degraded. Degradation is processive; once a substrate has been admitted to the central cavity, the reaction proceeds to its end.

The simplest proteasome is found in the archaea. **Figure 8.69** shows the top view of the crystal structure of the 20S assembly. It consists of two types of subunits, organized in the form α_7–β_7–β_7–α_7, where each septamer forms a ring. **Figure 8.70** shows the side view of the backbone. The a subunits form the two outer rings (on top and on the bottom), and the β subunits form the two inner rings. The β subunits have the protease activities, and the active sites are located at the N-terminal ends that project into the interior. The opening of ~20 A restricts the entrance for substrates. A yet simpler structure is found in *E. coli,* where a protein related to the β subunit, HslV, forms a structure of two six-member rings with a proteolytic core.

The eukaryotic 20S proteasome is more complex, consisting of 7 different α subunits and 7 different β subunits. **Figure 8.71** shows that it has the same general structure of α–β–β–α rings. The rings in each half of the structure are organized in the opposite rotational sense. A significant structural difference with the archaeal proteasome is that the central hole is occluded, so that there is no obvious entrance from the ends of the cylinder. This probably means that the structure is rearranged at some point to allow entrance from the ends.

The eukaryotic 26S proteasome is formed when the 19S caps associate with the 20S core, binding to one or both ends, to form an elongated structure of ~45 nm in length. The 19S caps are found only in eukaryotic (not archaeal or bacterial) proteasomes. The caps recognize ubiquitinated proteins, and pass them to the 20S core for proteolysis. The 19S caps contain ~18 subunits, several of which are ATPases; presumably the hydrolysis of ATP provides energy for handling the substrate proteins.

The hydrolytic mechanism of the proteasome is different from that of other proteases. The active site of a catalytic β-subunit is an N-terminal threonine; the hydroxyl group of the threonine attacks the peptide bond of the substrate. The proteasome contains several protease activities, with different specificities, for example, for cleaving after basic, acidic, or hydrophobic amino acids, allowing it to attack a variety of types of targets. Proteolytic activities with different substrate specificities may be provided by different β subunits. More than one β subunit

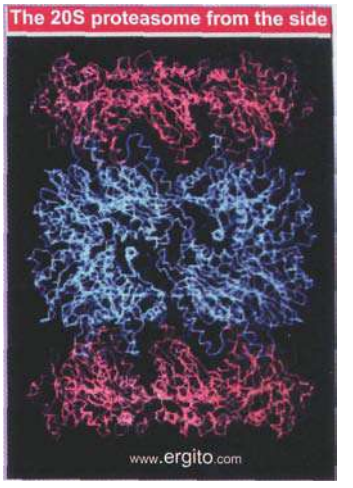


Figure 8.70 The side view of the archaeal 20S proteasome shows the rings of α subunits (red) and β subunits (blue). Photograph kindly provided by Robert Huber.

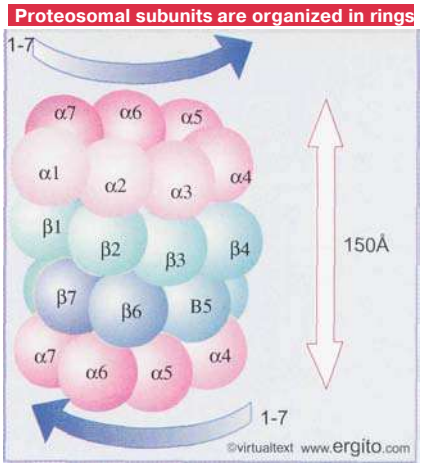


Figure 8.71 The eukaryotic 20S proteasome consists of two dimeric rings organized in counter-rotation.

may be needed for a particular enzymatic activity. The peptide products typically are octa- and nona-peptides. Proteasomes function processively, that is, a substrate is degraded to completion within the cavity, without any intermediates being released. Basically the central chamber traps proteins until they have been degraded to fragments below a certain size.

Inhibitors of the proteasome block the degradation of most cellular proteins, showing that it is responsible for bulk degradation. In fact, a significant proportion of newly synthesized proteins are immediately degraded by the proteasome (which casts a light on the efficiency of the production of proteins). It is also responsible for cleaving antigens in cells of the immune system to generate the small peptides that are presented on the surface of the cell to provoke the immune response (see *26.18 T cell receptors are related to immunoglobulins).* The peptide fragments are then transported by TAP (the transporter associated with antigen processing) from the cytosol into the ER, where they are bound by MHC molecules. Other reactions in which target proteins are completely degraded include the removal of cell cycle regulators; in particular, cyclins are degraded during mitosis (see *29.18 Protein degradation is important in mitosis),* and replication control proteins are degraded during the phase of DNA synthesis. In addition to these reactions, the proteasome may undertake specific processing events, for example, cleaving a precursor to a transcription factor to generate the active protein. The means by which these activities are regulated remain to be discovered.

8.33 Summary

A protein that is inserted into, or passes through, a membrane has a signal sequence that is recognized by a receptor that is part of the membrane or that can associate with it. The protein passes through an aqueous channel that is created by transmembrane protein(s) that reside in the membrane. In almost all cases, the protein passes through the channel in an unfolded form, and association with chaperones when it emerges is necessary in order to acquire the correct conformation. The major exception is the peroxisome, where an imported protein in its mature conformation binds to a cytosolic protein that carries it through the channel in the membrane.

Synthesis of proteins in the cytosol starts on "free" ribosomes. Proteins that are secreted from the cell or that are inserted into membranes of the reticuloendothelial system start with an N-terminal signal sequence that causes the ribosome to become attached to the membrane of the endoplasmic reticulum. The protein is translocated through the membrane by co-translational transfer. The process starts when the signal sequence is recognized by the SRP (a ribonucleoprotein particle), which interrupts translation. The SRP binds to the SRP receptor in the ER membrane, and transfers the signal sequence to the Sec61/TRAM receptor in the membrane. Synthesis resumes, and the protein is translocated through the membrane while it is being synthesized, although there is no energetic connection between the processes. The channel through the membrane provides a hydrophilic environment, and is largely made of the protein Sec61.

A secreted protein passes completely through the membrane into the ER lumen. Proteins that are integrated into membranes can be divided into two general types based on their orientation. For type I integral membrane proteins, the N-terminal signal sequence is cleaved, and transfer through the membrane is halted later by an anchor sequence. The protein becomes oriented in the membrane with its N-terminus on the far side and its C-terminus in the cytosol. Type

II proteins do not have a cleavable N-terminal signal, but instead have a combined signal-anchor sequence, which enters the membrane and becomes embedded in it, causing the C-terminus to be located on the far side, while the N-terminus remains in the cytosol. The orientation of the signal-anchor is determined by the "positive inside" rule that the side of the anchor with more positive charges will be located in the cytoplasm. Proteins that have single transmembrane spanning regions move laterally from the channel into the lipid bilayer. Proteins may have multiple membrane-spanning regions, with loops between them protruding on either side of the membrane. The mechanism of insertion of multiple segments is unknown.

In the absence of any particular signal, a protein is released into the cytosol when its synthesis is completed. Proteins are imported post-translationally into mitochondria or chloroplasts. They possess N-terminal leader sequences that target them to the outer membrane of the organelle envelope; then they are transported through the outer and inner membranes into the matrix. Translocation requires ATP and a potential across the inner membrane. The N-terminal leader is cleaved by a protease within the organelle. Proteins that reside within the membranes or intermembrane space possess a signal (which becomes N-terminal when the first part of the leader is removed) that either causes export from the matrix to the appropriate location or which halts transfer before all of the protein has entered the matrix. Control of folding, by Hsp70 and Hsp60 in the mitochondrial matrix, is an important feature of the process.

Mitochondria and chloroplasts have separate receptor complexes that create channels through each of the outer and inner membranes. All imported proteins pass directly from the TOM complex in the outer membrane to a TIM complex in the inner membrane. Proteins that reside in the inter-membrane space or in the outer membrane are re-exported from the TIM complex after entering the matrix. The TOM complex uses different receptors for imported proteins depending on whether they have N-terminal or internal signal sequences, and directs both types into the Tom40 channel. There are two TIM receptors in the inner membrane, one used for proteins whose ultimate destination is the inner matrix, the other used for proteins that are re-exported to the inter-membrane space or outer membrane.

Bacteria have components for membrane translocation that are related to those of the co-translational eukaryotic system, but translocation often occurs by a post-translational mechanism. SecY/E provide the translocase, and SecA associates with the channel and is involved in inserting and propelling the substrate protein. SecB is a chaperone that brings the protein to the channel. Some integral membrane proteins are inserted into the channel by an interaction with an apparatus resembling the SRP, consisting of 4.5S RNA and the Fth and FtsY proteins. The protein YidC is homologous to a mitochondrial protein and is required for insertion of some membrane proteins.

Nuclear pore complexes are massive structures embedded in the nuclear membrane, and are responsible for all transport of protein into the nucleus and RNA out of the nucleus. They have 8-fold symmetry seen in cross section and 2-fold symmetry viewed from the side. Each nuclear pore complex contains a central pore, which forms a channel of diameter <10 nm. The central channel can be opened to a diameter of ~20 nm to allow passage of larger material, some of which may need to undergo conformational changes to fit. The proteins of the complex are called nucleoporins; a subset called FG-nucleoporins have hydrophobic repeated sequences of Phe-Gly, are found in the central pore, and may be important in the translocation process.

Proteins that are actively transported into the nucleus require specific NLS sequences, which are short, but do not seem to share common features except for their basicity. Proteins that are exported

from the nucleus have specific NES sequences, which share a pattern of leucine residues. Transport is a two stage process, involving docking followed by translocation. The docking reaction is undertaken by a transport receptor. Importins carry proteins into the nucleus, and exportins carry proteins out of the nucleus. The best characterized transport receptor is importin-αβ, which has subunits that bind to the substrate protein and to a nucleoporin protein in the pore, respectively. Other transport receptors consist of single proteins that have both functions. The direction of translocation is controlled by Ran. The presence of Ran-GDP in the cytosol destabilizes export complexes. The presence of Ran-GTP in the nucleus destabilizes import complexes. This ensures release of substrate on the appropriate side of the nuclear envelope. ATP is required for translocation only in order to support the regeneration of Ran-GTP from Ran-GDP; energy is not required for the translocation process itself. The mechanism of translocation is not understood in detail, but is likely to involve interactions of the transport receptors with the FG-nucleoporins.

The major system responsible for bulk degradation of proteins, but also for certain specific processing events, is the proteasome, a large complex that contains several protease activities. It acts upon substrate proteins that have been conjugated to ubiquitin through an isopeptide bond, and upon which a polyubiquitin chain has formed.

References

8.4 Chaperones may be required for protein folding

exp Horwich, A. (2002). The Discovery of Protein Folding by Chaperonins (www.ergito.com/lookup.jsp?expt = horwich)

rev Ellis, R. J. and van der Vies, S. M. (1991). Molecular chaperones. Ann. Rev. Biochem. 60, 321-347.
Fersht, A. R. and Daggett, V. (2002). Protein folding and unfolding at atomic resolution. Cell 108, 573-582.
Hartl, F. U. and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. Science 295, 1852-1858.

ref Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science 181, 223-230.
van den Berg, B., Ellis, R. J., and Dobson, C. M. (1999). Effects of macromolecular crowding on protein folding and aggregation. EMBO J. 18, 6927-6933.

8.5 Chaperones are needed by newly synthesized and by denatured proteins

rev Frydman, J. (2001). Folding of newly translated proteins in vivo: the role of molecular chaperones. Ann. Rev. Biochem. 70, 603-647.
Moarefi, I. and Hartl, F. U. (2001). Hsp90: a specialized but essential protein-folding tool. J. Cell Biol. 154, 267-273.

ref Queitsch, C, Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. Nature 417, 618-624.
Rutherford, S. L. and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. Nature 396, 336-342.

8.6 The Hsp70 family is ubiquitous.

rev Bukau, B. and Horwich, A. L. (1998). The Hsp70 and Hsp60 chaperone machines. Cell 92, 351-366.
Frydman, J. (2001). Folding of newly translated proteins in vivo: the role of molecular chaperones. Ann. Rev. Biochem. 70, 603-647.
Georgopoulos, C. and Welch, W. J. (1993). Role of the major heat shock proteins as molecular chaperones. Ann. Rev. Cell Biol. 9, 601-634.
Hartl, F.-U. (1966). Molecular chaperones in cellular protein folding. Nature 381, 571-580.

ref Blond-Elguindi, S., Cwirla, S. E., Dower, W. J., Lipshutz, R. J., Sprang, S. R., Sambrook, J. F., and Gething, M. J. (1993). Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. Cell 75, 717-728.
Flaherty, K. M., DeLuca-Flaherty, C, and McKay, D. B. (1990). Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. Nature 346, 623-628.
Flynn, G. C, Pohl, J., Flocco, M. T., and Rothman, J. E. (1991). Peptide-binding specificity of the molecular chaperone BiP. Nature 353, 726-730.
Zhu, X., Zhao, X., Burkholder, W. F., Gragerov, A., Ogata, C. M., Gottesman, M. E., Hendrickson, and, Gottesman, M. E. (1996). Structural analysis of substrate binding by the molecular chaperone DnaK. Science 272, 1606-1614.

8.7 Hsp60/GroEL forms an oligomeric ring structure

rev Horwich A. L., Weber-Ban E. U., Finley D. (1999). Chaperone rings in protein folding and degradation. Proc. Nat. Acad. Sci. USA 96, 1 1033-1 1040.

ref Braig, K. et al. (1994). The crystal structure of the bacterial chaperonin GroEL at 28 A. Nature 371, 578-586.
Chen, S. et al. (1994). Location of a folding protein and shape changes in GroEL-GroES complexes imaged by cryo-electron microscopy. Nature 371, 261-264.
Hunt, J. F. et al. (1996). The crystal structure of the GroES co-chaperonin at 28 A resolution. Nature 379, 37-45.
Rye H. S., Roseman A. M., Chen S., Furtak K., Fenton W. A., Saibil H. R., and Horwich A. L. (1999). GroEL-GroES cycling: ATP and nonnative polypeptide direct alternation of folding-active rings. Cell 97, 325-338.

Mayhew, M. et al. (1996). Protein folding in the central cavity of the GroEL-GroES chaperonin complex. Nature 379, 420-426.
Rye, H. S, et al. (1997). Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL. Nature 388, 792-798.
Weissman, J. S. et al. (1994). GroEL-mediated protein folding proceeds by multiple rounds of binding and release of nonnative forms. Cell 78, 693-702.
Weissman, J. A. et al. (1995). Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES. Cell 83, 577-587.
Xu, Z., Horwich, A. L, and Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature 388, 741-750.

8.8 Signal sequences initiate translocation

rev Lee, C. and Beckwith, J. (1986). Cotranslational and posttranslational protein translocation in prokaryotic systems. Ann. Rev. Cell Biol. 2, 315-336.

ref Blobel, G. and Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell Biol. 67, 835-851.

Lingappa, V. R., Chaidez, J., Yost, C. S., and Hedgpeth, J. (1984). Determinants for protein localization: beta-lactamase signal sequence directs globin across microsomal membranes. Proc. Nat. Acad. Sci. USA 81, 456-460.

Palade, G. (1975). Intracellular aspects of the process of protein synthesis. Science 189, 347-358.

von Heijne, G. (1985). Signal sequences. The limits of variation. J. Mol. Biol. 184, 99-105.

8.9 The signal sequence interacts with the SRP

rev Walter, P. and Johnson, A. E. (1994). Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Ann. Rev. Cell Biol. 10, 87-119.

ref Tjalsma, H., Bolhuis, A., van Roosmalen, M. L., Wiegert, T,, Schumann, W., Broekhuizen, C. P., Quax, W. J., Venema, G., Bron, S., and van Dijl, M. (1998). Functional analysis of the secretory precursor processing machinery of Bacillus subtilis: identification of a eubacterial homolog of archaeal and eukaryotic signal peptidases. Genes Dev. 12, 2318-2331.

Walter, P. and Blobel, G. (1981). Translocation of proteins across the ER III SRP causes signal sequence and site specific arrest of chain elongation that is released by microsomal membranes. J. Cell Biol. 91, 557-561.

8.10 The SRP interacts with the SRP receptor

rev Keenan, R. J., Freymann, D. M., Stroud, R. M., and Walter, P. (2001). The signal recognition particle. Ann. Rev. Biochem. 70, 755-775.

ref Batey, R. T., Rambo, R. P., Lucast, L., Rha, B., and Doudna, J. A. (2000). Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science 287, 1232-1239.
Keenan, R. J., Freymann, D. M., Walter, P., and Stroud, R. M. (1998). Crystal structure of the signal sequence-binding subunit of the signal recognition particle. Cell 94, 181-191.

Powers, T. and Walter, P. (1995). Reciprocal stimulation of GTP hydrolysis by two directly interacting GTPases. Science 269, 1422-1424.

Siegel, V. and Walter, P. (1988). Each of the activities of SRP is contained within a distinct domain: analysis of biochemical mutants of SRP. Cell 52, 39-49.

Tajima, S., Lauffer, L., Rath, V. L., and Walter, P. (1986). The signal recognition particle receptor is a complex that contains two distinct polypeptide chains. J. Cell Biol. 103, 1167-1178.
Walter, P. and Blobel, G. (1981). Translocation of proteins across the ER III SRP causes signal sequence and site specific arrest of chain elongation that is released by microsomal membranes. J. Cell Biol. 91, 557-561.

Walter, P. and Blobel, G. (1982). Signal recognition particle contains a 7S RNA essential for protein translocation across the ER. Nature 299, 691-698.
Zopf, D., Bernstein, H. D., Johnson, A. E., and Walter, P. (1990). The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be crosslinked to a signal sequence. EMBO J. 9, 4511-4517.

8.11 The translocon forms a pore

ref Crowley, K. S. (1994). Secretory proteins move through the ER membrane via an aqueous, gated pore. Cell 78, 461-471.

Deshaies, R. J. and Schekman, R. (1987). A yeast mutant defective at an early stage in import of secretory protein precursors into the endoplasmic reticulum. J. Cell Biol. 105, 633-645.
Esnault, Y., Blondel, M. O., Deshaies, R. J., Scheckman, R., and Kepes, F. (1993). The yeast SSS1 gene is essential for secretory protein translocation and encodes a conserved protein of the endoplasmic reticulum. EMBO J. 12, 4083-4093.
Hanein, D. et al. (1996). Oligomeric rings of the Sec61p complex induced by ligands required for protein translocation. Cell 87, 721-732.
Liao, S., Lin, J., Do, H., and Johnson, A. E. (1997). Both lumenal and cytosolic gating of the aqueous ER translocon pore are regulated from inside the ribosome during membrane protein integration. Cell 90, 31-41.

Mothes, W., Prehn, S., and Rapoport, T. A. (1994). Systematic probing of the environment of a translocating secretory protein during translocation through the ER membrane. EMBO J. 13, 3973-3982.
Simon, S. M. and Blobel, G. (1991). A protein-conducting channel in the endoplasmic reticulum. Cell 65, 371-380.

8.12 Translocation requires insertion into the translocon and (sometimes) a ratchet in the ER

rev Rapoport, T. A., Jungnickel, B., and Kutay, U. (1996). Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. Ann. Rev. Biochem. 65, 271-303.
Walter, P. and Lingappa, V. (1986). Mechanism of protein translocation across the endoplasmic reticulum membrane. Ann. Rev. Cell Biol. 2, 499-516.

ref Gorlich, D. and Rapoport, T. A. (1993). Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. Cell 75, 615-630.
Matlack, K. E., Misselwitz, B., Plath, K., and Rapoport, T. A. (1999). BiP acts as a molecular ratchet during posttranslational transport of prepro-alpha factor across the ER membrane. Cell 97, 553-564.

8.13 Reverse translocation sends proteins to the cytosol for degradation

rev Tsai, B., Ye, Y., and Rapoport, T. A. (2002). Retro-translocation of proteins from the endoplasmic reticulum into the cytosol. Nat. Rev. Mol. Cell Biol. 3, 246-255.

ref Johnson, A. E. and Haigh, N. G. (2000). The ER translocon and retrotranslocation: is the shift into reverse manual or automatic? Cell 102, 709-712.
Wiertz, E. J. H. J. et al. (1996). Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. Nature 384, 432-438.
Wilkinson, B. M., Tyson, J. R., Reid, P. J., and Stirling, C. J. (2000). Distinct domains within yeast Sec61p involved in post-translational translocation and protein dislocation. J. Biol. Chem. 275, 521-529.
Zhou, M. and Schekman, R. (1999). The engagement of Sec61p in the ER dislocation process. Mol. Cell 4, 925-934.

8.16 How do proteins insert into membranes?

rev Hegde, R. S. and Lingappa, V. R. (1997). Membrane protein biogenesis: regulated complexity at the endoplasmic reticulum. Cell 91, 575-582.
Wickner, W. T. and Lodish, H. (1985). Multiple mechanisms of protein insertion into and across membranes. Science 230, 400-407.

ref Borel, A. C, and Simon, S. M. (1996). Biogenesis of polytopic membrane proteins: membrane segments assemble within translocation channels prior to membrane integration. Cell 85, 379-389.
Do, H., Do, H., Lin, J., Andrews, D. W., and Johnson, A. E. (1996). The cotranslational integration of membrane proteins into the phospholipid bilayer is a multistep process. Cell 85, 369-378.
Heinrich, S. U., Mothes, W., Brunner, J., and Rapoport, T. A. (2000). The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. Cell 102, 233-244.
Kim, P. K., Janiak-Spens, F., Trimble, W. S., Leber, B., and Andrews, D. W. (1997). Evidence for multiple mechanisms for membrane binding and integration via carboxyl-terminal insertion sequences. Biochemistry 36, 8873-8882.
Liao, S., Lin, J., Do, H., and Johnson, A. E. (1997). Both lumenal and cytosolic gating of the aqueous ER translocon pore are regulated from inside the ribosome during membrane protein integration. Cell 90, 31-41.
Mothes, W., Heinrich, S. U., Graf, R., Nilsson, I., von Heijne, G., Brunner, J., and Rapoport, T. A. (1997). Molecular mechanism of membrane protein integration into the endoplasmic reticulum. Cell 89, 523-533.

8.17 Post-translational membrane insertion depends on leader sequences

rev Baker, K. P. and Schatz, G. (1991). Mitochondrial proteins essential for viability mediate protein import into yeast mitochondria. Nature 349, 205-208.
Schatz, G. and Dobberstein, B. (1996). Common principles of protein translocation across membranes. Science 271, 1519-1526.

ref Eilers, M. and Schatz, G. (1986). Binding of a specific ligand inhibits import of a purified precursor protein into mitochondria. Nature 322, 228-232.

8.18 A hierarchy of sequences determines location within organelles

rev Cline, K. and Henry, R. (1996). Import and routing of nucleus-encoded chloroplast proteins. Ann. Rev. Cell Dev. Biol. 12, 1-26.

ref Hartl, F.-U. et al. (1988). Successive translocation into and out of the mitochondrial matrix: targeting of proteins to the intermembrane space by a bipartite signal peptide. Cell 51, 1027-1037.
van Loon, A. P. G. M. et al. (1986). The presequences of two imported mitochondrial proteins contain information for intracellular and intramitochondrial sorting. Cell 44, 801-812.

8.19 Inner and outer mitochondrial membranes have different translocons

rev Dalbey, R. E. and Kuhn, A. (2000). Evolutionarily related insertion pathways of bacterial, mitochondrial, and thylakoid membrane proteins. Ann. Rev. Cell Dev. Biol. 16, 51-87.
Neupert, W. (1997). Protein import into mitochondria. Ann. Rev. Biochem. 66, 863-917.
Neupert, W. and Brunner, M. (2002). The protein import motor of mitochondria. Nat. Rev. Mol. Cell Biol. 3, 555-565.

ref Leuenberger, D., Bally, N. A., Schatz, G., and Koehler, C. M. (1999). Different import pathways through the mitochondrial intermembrane space for inner membrane proteins. EMBO J. 18, 4816-4822.
Ostermann, J., Horwich, A. L., Neupert, W., and Hartl, F. U. (1989). Protein folding in mitochondria requires complex formation with hsp60 and ATP hydrolysis. Nature 341, 125-130.

8.20 Peroxisomes employ another type of translocation system

rev Purdue, P. E. and Lazarow, P. B. (2001). Peroxisome biogenesis. Ann. Rev. Cell Dev. Biol. 17, 701-752.

ref Dodt, G. and Gould, S. J. (1996). Multiple PEX genes are required for proper subcellular distribution and stability of Pex5p, the PTS1 receptor: evidence that PTS1 protein import is mediated by a cycling receptor. J. Cell Biol. 135, 1763-1774.
Elgersma, Y., Vos, A., van den Berg, M., van Roermund, C. W., van der Sluijs, P., Distel, B., and Tabak, H. F. (1996). Analysis of the carboxyl-terminal peroxisomal targeting signal 1 in a homologous context in *S. cerevisiae*. J. Biol. Chem. 271, 26375-26382.
Elgersma, Y., Elgersma-Hooisma, M., Wenzel, T., McCaffery, J. M., Farquhar, M. G., and Subramani, S. (1998). A mobile PTS2 receptor for peroxisomal protein import in Pichia pastoris. J. Cell Biol. 140, 807-820.
Goldfischer, S., Moore, C. L., Johnson, A. B., Spiro, A. J., Valsamis, M. P., Wisniewski, H. K., Ritch, R. H., Norton, W. T., Rapin, I., and Gartner, L. M. (1973). Peroxisomal and mitochondrial defects in the cerebro-hepato-renal syndrome. Science 182, 62-64.
Gould, S. J., Keller, G. A., Hosken, N., Wilkinson, J., and Subramani, S. (1989). A conserved tripeptide sorts proteins to peroxisomes. J. Cell Biol. 108, 1657-1664.
Matsuzono, Y., Kinoshita, N., Tamura, S., Shimozawa, N., Hamasaki, M., Ghaedi, K., Wanders, R. J., Suzuki, Y., Kondo, N., and Fujiki, Y. (1999). Human PEX19: cDNA cloning by functional complementation, mutation analysis in a patient with Zellweger syndrome, and potential role in peroxisomal membrane assembly. Proc. Nat. Acad. Sci. USA 96, 2116-2121.
South, S. T. and Gould, S. J. (1999). Peroxisome synthesis in the absence of preexisting peroxisomes. J. Cell Biol. 144, 255-266.
Subramani, S., Koller, A., and Snyder, W. B. (2000). Import of peroxisomal matrix and membrane proteins. Ann. Rev. Biochem. 69, 399-418.
Walton, P. A., Hill, P. E., and Hill, S. (1995). Import of stably folded proteins into peroxisomes. Mol. Biol. Cell 6, 675-683.

8.22 The *Sec* system transports proteins into and through the inner membrane

rev Lee, C. and Beckwith, J. (1986). Cotranslational and posttranslational protein translocation in prokaryotic systems. Ann. Rev. Cell Biol. 2, 315-336.
Oliver, D. (1985). Protein secretion in *E. coli*. Ann. Rev. Immunol. 39, 615-648.

ref Beck, K., Wu, L. F., Brunner, J., and Muller, M. (2000). Discrimination between SRP- and SecA/SecB-dependent substrates involves selective recognition of nascent chains by SRP and trigger factor. EMBO J. 19, 134-143.
Brundage, L. et al. (1990). The purified *E. coli* integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. Cell 62, 649-657.
Collier, D. N. et al. (1988). The antifolding activity of SecB promotes the export of the *E. coli* maltose-binding protein. Cell 53, 273-283.
Crooke, E. et al. (1988). ProOmpA is stabilized for membrane translocation by either purified *E. coli* trigger factor or canine signal recognition particle. Cell 54, 1003-1011.
Valent, Q. A., Scotti, P. A., High, S., von Heijne, G., Lentzen, G., Wintermeyer, W., Oudega, B., and Luirink, J. (1998). The *E. coli* SRP and SecB targeting pathways converge at the translocon. EMBO J. 17, 2504-2512.
Yahr, T. L. and Wickner, W. T. (2000). Evaluating the oligomeric state of SecYEG in preprotein translocase. EMBO J. 19, 4393-4401.

8.23 Sec-independent translation systems in *E. coli*

rev Dalbey, R. E. and Kuhn, A. (2000). Evolutionarily related insertion pathways of bacterial, mitochondrial, and thylakoid membrane proteins. Ann. Rev. Cell Dev. Biol. 16, 51-87.
Dalbey, R. E. and Robinson, C. (1999). Protein translocation into and across the bacterial plasma membrane and the plant thylakoid membrane. Trends Biochem. Sci. 24, 17-22.

ref Beck, K., Wu, L. F., Brunner, J., and Muller, M. (2000). Discrimination between SRP- and SecA/SecB-dependent substrates involves selective recognition of nascent chains by SRP and trigger factor. EMBO J. 19, 134-143.
Samuelson, J. C, Chen, M., Jiang, F., Moller, I., Wiedmann, M., Kuhn, A., Phillips, G. J., and Dalbey, R. E. (2000). YidC mediates membrane protein insertion in bacteria. Nature 406, 637-641.
Scotti, P. A., Urbanus, M. L., Brunner, J., de Gier, J. W., von Heijne, G., van der Does, C, Driessen, A. J., Oudega, B., and Luirink, J. (2000). YidC, the *E. coli* homologue of mitochondrial Oxa1p, is a component of the Sec translocase. EMBO J. 19, 542-549.
Soekarjo, M., Eisenhawer, M., Kuhn, A., and Vogel, H. (1996). Thermodynamics of the membrane insertion process of the M13 procoat protein, a lipid bilayer traversing protein containing a leader sequence. Biochemistry 35, 1232-1241.

8.25 Nuclear pores are large symmetrical structures

rev Davis, L. I. (1995). The nuclear pore complex. Ann. Rev. Biochem. 64, 865-896.
Forbes, D. J. (1992). Structure and function of the nuclear pore complex. Ann. Rev. Cell Biol. 8, 495-527.

ref Hinshaw, J. E., Carragher, B. 0., and Milligan, R. A. (1992). Architecture and design of the nuclear pore complex. Cell 69, 1133-1141.
Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. 148, 635-651.
Yang, Q., Rout, M. P., and Akey, C. W. (1998). Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. Mol. Cell 1, 223-234.

8.26 The nuclear pore is a size-dependent sieve for smaller material

rev Dingwall, C. and Laskey, R. A. (1986). Protein import into the cell nucleus. Ann. Rev. Cell Biol. 2, 367-390.

ref Akey, C. W. and Goldfarb, D. S. (1989). Protein import through the nuclear pore complex is a multistep process. J. Cell Biol. 109, 971-982.
Dworetzky, S. I. and Feldherr, C. M. (1988). Translocation of RNA-coated gold particles through the nuclear pores of oocytes. J. Cell Biol. 106, 575-584.
Feldherr, C. M., Kallenbach, E., and Schultz, N. (1984). Movement of a karyophilic protein through the nuclear pores of oocytes. J. Cell Biol. 99, 2216-2222.
Paine, P. L. (1975). Nucleocytoplasmic movement of fluorescent tracers microinjected into living salivary gland cells. J. Cell Biol. 66, 652-657.

8.27 Proteins require signals to be transported through the pore

ref Bogerd, H. P., Fridell, R. A., Benson, R. E., Hua, J., and Cullen, B. R. (1996). Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel *in vitro* randomization-selection assay. Mol. Cell Biol. 16, 4207-4214.
Dingwall, C, Sharnick, S. V., and Laskey, R. A. (1982). A polypeptide domain that specifies migration of nucleoplasmin into the nucleus. Cell 30, 449-458.
Fischer, U., Huber, J., Boelens, W. C, Mattaj, I. W., and Luhrmann, R. (1995). The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. Cell 82, 475-483.
Kalderon, D., Richardson, W. D., Markham, A. F., and Smith, A. E. (1984). Sequence requirements for nuclear location of simian virus 40 large-T antigen. Nature 311, 33-38.
Kalderon, D., Roberts, B. L., Richardson, W. D., and Smith, A. E. (1984). A short amino acid sequence able to specify nuclear location. Cell 39, 499-509.
Michael, W. M., Choi, M. and Dreyfuss, G. (1995). A nuclear export signal in hnRNP A1: a signal-mediated, temperature-dependent nuclear protein export pathway. Cell 83, 415-422.
Robbins, J. et al. (1991). Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequences. Cell 64, 615-623.

8.28 Transport receptors carry cargo proteins through the pore

rev Gorlich, D. and Kutay, U. (1999). Transport between the cell nucleus and cytoplasm. Ann. Rev. Cell Dev. Biol. 15, 607-660.

ref Adam, S. A., Marr, R. S., and Gerace, L. (1990). Nuclear protein import in permeabilized mammalian cells requires soluble cytoplasmic factors. J. Cell Biol. 111, 807-816.
Moore, M. S. and Blobel, G. (1992). The two steps of nuclear import, targeting to the nuclear envelope and translocation through the nuclear pore, require different cytosolic factors. Cell 69, 939-950.
Moroianu, J., Blobel, G., and Radu, A. (1995). Previously identified protein of uncertain function is importin α docks import substrate at nuclear pore complexes. Proc. Nat. Acad. Sci. USA 92, 2008-2011.

Newmeyer, D. D. and Forbes, D. J. (1988). An NEM-sensitive cytosolic factor necessary for nuclear protein import: requirement in a signal-mediated binding to the nuclear pore. Cell 52, 641-653.
Radu, A., Moore, M. S., and Blobel, G. (1995). The peptide repeat domain of nucleoporin Nup98 functions as a docking site in transport across the nuclear pore complex. Cell 81, 215-222.
Richardson, W. D. et al. (1988). Nuclear protein migration involves two steps: rapid binding at the nuclear envelope followed by slower translocation through nuclear pores. Cell 52, 655-664.

8.29 Ran controls the direction of transport

rev Gorlich, D. and Kutay, U. (1999). Transport between the cell nucleus and cytoplasm. Ann. Rev. Cell Dev. Biol. 15, 607-660.
Komeili, A. and O'Shea, E. K. (2001). New perspectives on nuclear transport. Ann. Rev. Genet. 35, 341-364.

ref Chook, Y. M. and Blobel, G. (1999). Structure of the nuclear transport complex karyopherin-beta2-Ran x GppNHp. Nature 399, 230-237.
Cingolani, G., Petosa, C, Weis, K., Muller, C. W., Cingolani, G., Petosa, C, Weis, K., and Muller, C. W. (1999). Structure of importin-beta bound to the IBB domain of importin-alpha. Nature 399, 221-229.
Gorlich, D., Pante, N., Kutay, U., Aebi, U., and Bischoff, F. R. (1996). Identification of different roles for RanGDP and RanGTP in nuclear protein import. EMBO J. 15, 5584-5594.
Keminer, O., Siebrasse, J. P., Zerf, K., and Peters, R. (1999). Optical recording of signal-mediated protein transport through single nuclear pore complexes. Proc. Nat. Acad. Sci. USA 96, 11842-11847.
Kutay, U. et al. (1997). Export of importin α from the nucleus is mediated by a specific nuclear transport factor. Cell 90, 1061-1071.
Nachury, M. V. and Weis, K. (1999). The direction of transport through the nuclear pore can be inverted. Proc. Nat. Acad. Sci. USA 96, 9622-9627.
Rexach, M. and Blobel, G. (1995). Protein import into nuclei: association and dissociation reactions involving transport substrate, transport factors, and nucleoporins. Cell 83, 683-692.
Ribbeck, K. and Gorlich, D. (2001). Kinetic analysis of translocation through nuclear pore complexes. EMBO J. 20, 1320-1330.
Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. 148, 635-651.

8.30 RNA is exported by several systems

rev Gorlich, D. and Mattaj, I. W. (1996). Nucleo-cytoplasmic transport. Science 271, 1513-1518.
Reed, R. and Hurt, E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell 108, 523-531.

ref Fornerod, M. et al. (1997). CRM1 is an export receptor for leucine-rich nuclear export signals. Cell 90, 1051-1060.
Lei, E. P., Krebber, H., and Silver, P. A. (2001). Messenger RNAs are recruited for nuclear export during transcription. Genes Dev. 15, 1771-1782.
Stade, K., Ford, C. S., Guthrie, C, and Weis, K. (1997). Exportin 1 (Crm1p) is an essential nuclear export factor. Cell 90, 1041-1050.

8.31 Ubiquitination targets proteins for degradation

exp Hershko, A. and Ciechenover, A. (2002). Ubiquitin Conjugation as a Proteolytic Signal: The First Experiments (www.ergito.com/lookup.jsp?expt = Ciechanover)

rev Ciechanover, A. (1994). The ubiquitin-proteasome proteolytic pathway. Cell 79, 13-21.
Hershko, A., and Ciechanover, A. (1998). The ubiquitin system. Ann. Rev. Biochem. 67, 425-479.
Jentsch, S. (1992). The ubiquitin-conjugation system. Ann. Rev. Genet. 26, 179-207.
Muller, S., Hoege, C, Pyrowolakis, G., and Jentsch, S. (2001). SUMO, ubiquitin's mysterious cousin. Nat. Rev. Mol. Cell Biol. 2, 202-210.
Pickart, C. M. (2001). Mechanisms underlying ubiquitination. Ann. Rev. Biochem. 70, 503-533.
Weissman, A. M. (2001). Themes and variations on ubiquitylation. Nat. Rev. Mol. Cell Biol. 2, 169-178.

ref Chau, V. et al. (1989). A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science 243, 1576-1583.
Ciechanover, A. et al. (1980). ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein degradation. Proc. Nat. Acad. Sci. USA 77, 1365-1368.

8.32 The proteasome is a large machine that degrades ubiquitinated proteins

rev Baumeister, W. et al. (1998). The proteasome: paradigm of a self-compartmentalizing protease. Cell 92, 367-380.
Coux, O., Tanaka, K., and Goldberg, A. L (1996). Structure and functions of the 20S and 26S proteasomes. Ann. Rev. Biochem. 65, 801-847.
Voges, D., Zwickl, P., and Baumeister, W. (1999). The 26S proteasome: a molecular machine designed for controlled proteolysis. Ann. Rev. Biochem. 68, 1015-1068.

ref Eytan, E. et al. (1989). ATP-dependent incorporation of 20S protease into the 26S complex that degrades proteins conjugated to ubiquitin. Proc. Nat. Acad. Sci. USA 86, 7751-7755.
Groll, M. et al. (1997). Structure of 20S proteasome from yeast at 24 Å resolution. Nature 386, 463-471.
Ishikawa, T., Beuron, F., Kessel, M., Wickner, S., Maurizi, M. R., and Steven, A. C. (2001). Translocation pathway of protein substrates in ClpAP protease. Proc. Nat. Acad. Sci. USA 98, 4328-4333.
Lowe, J. et al. (1995). Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 34 A resolution. Science 268, 533-539.
Reid, B. G., Fenton, W. A., Horwich, A. L., and Weber-Ban, E. U. (2001). ClpA mediates directional translocation of substrate proteins into the ClpP protease. Proc. Nat. Acad. Sci. USA 98, 3768-3772.
Reits, E. A., Vos, J. C, Gromme, M., and Neefjes, J. (2000). The major substrates for TAP *in vitro are* derived from newly synthesized proteins. Nature 404, 774-778.
Schubert, U., Anton, L. C, Gibbs, J., Norbury, C. C, Yewdell, J. W., and Bennink, J. R. (2000). Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. Nature 404, 770-774.
Wang, J., Hartling, J. A., and Flanagan, J. M. (1997). The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis. Cell 91, 447-456.

Chapter 9

Transcription

9.1 Introduction

Transcription involves synthesis of an RNA chain representing one strand of a DNA duplex. By "representing" we mean that the RNA is *identical in sequence* with one strand of the DNA, which is called the coding strand. It is *complementary* to the other strand, which provides the template strand for its synthesis. **Figure** 9.1 recapitulates the relationship between double-stranded DNA and its single-stranded RNA transcript.

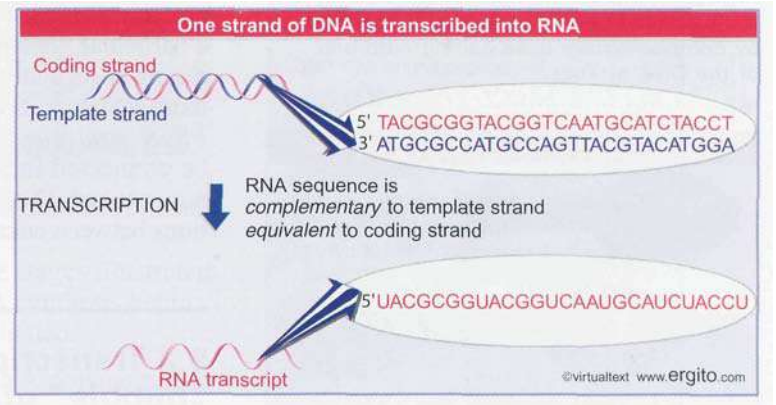


Figure 9.1 The function of RNA polymerase is to copy one strand of duplex DNA into RNA.

RNA synthesis is catalyzed by the enzyme RNA polymerase. Transcription starts when RNA polymerase binds to a special region, the promoter, at the start of the gene. The promoter surrounds the first base pair that is transcribed into RNA, the startpoint. From this point, RNA polymerase moves along the template, synthesizing RNA, until it reaches a terminator sequence. This action defines a transcription unit that extends from the promoter to the terminator. The critical feature of the transcription unit, depicted in **Figure 9.2,** is that it constitutes a stretch of DNA *expressed via the production of a single RNA molecule*. A transcription unit may include more than one gene.

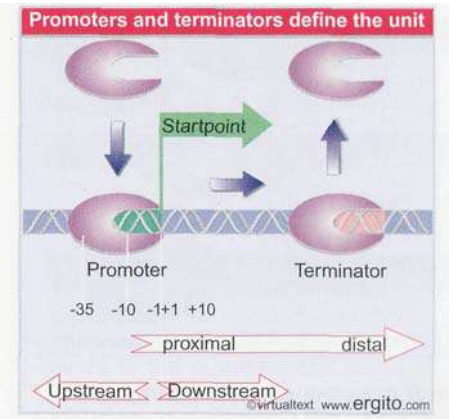


Figure 9.2 A transcription unit is transcribed into a single RNA.

Sequences prior to the startpoint are described as upstream of it; those after the startpoint (within the transcribed sequence) are downstream of it. Sequences are conventionally written so that transcription proceeds from left (upstream) to right (downstream). This corresponds to writing the mRNA in the usual 5′→3′ direction.

Often the DNA sequence is written to show only the coding strand, which has the same sequence as the RNA. Base positions are numbered in both directions away from the startpoint, which is assigned the value +1; numbers are increased going downstream. The base before the startpoint is numbered - 1 , and the negative numbers increase going upstream. (There is no base assigned the number 0.)

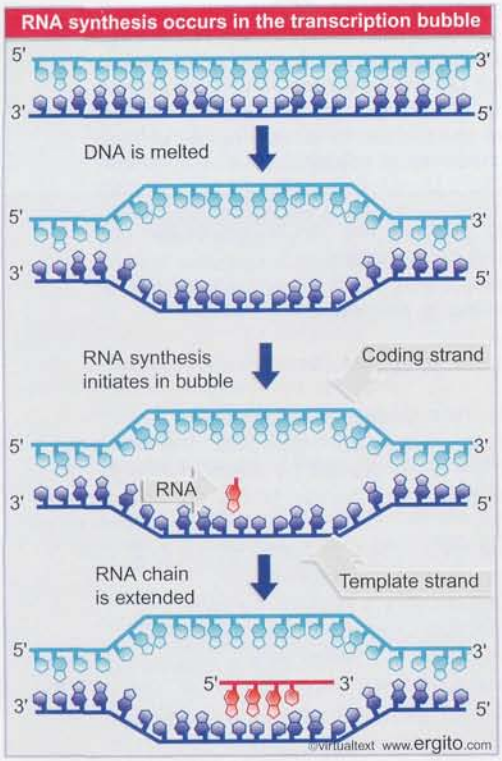


Figure 9.3 DNA strands separate to form a transcription bubble. RNA is synthesized by complementary base pairing with one of the DNA strands.

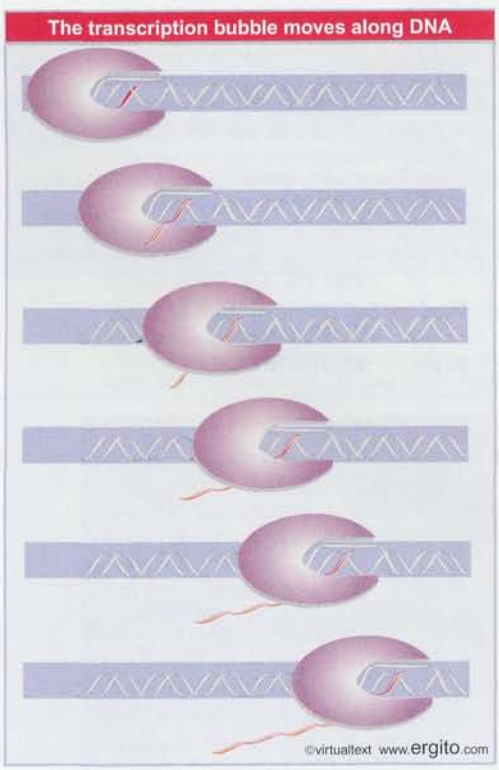


Figure 9.4 As the transcription bubble progresses, the DNA duplex reforms behind it, displacing the RNA in the form of a single polynucleotide chain.

The immediate product of transcription is called the primary transcript. It would consist of an RNA extending from the promoter to the terminator, possessing the original 5' and 3' ends. However, the primary transcript is almost always unstable. In prokaryotes, it is rapidly degraded (mRNA) or cleaved to give mature products (rRNA and tRNA). In eukaryotes, it is modified at the ends (mRNA) and/or cleaved to give mature products (all RNA).

Transcription is the first stage in gene expression, and the principal step at which it is controlled. Regulatory proteins determine whether a particular gene is available to be transcribed by RNA polymerase. The initial (and often the only) step in regulation is the decision on whether or not to transcribe a gene. Most regulatory events occur at the initiation of transcription, although subsequent stages in transcription (or other stages of gene expression) are sometimes regulated.

Within this context, there are two basic questions in gene expression:

- How does RNA polymerase find promoters on DNA? This is a particular example of a more general question: how do proteins distinguish their specific binding sites in DNA from other sequences?
- How do regulatory proteins interact with RNA polymerase (and with one another) to activate or to repress specific steps in the initiation, elongation, or termination of transcription?

In this chapter, we analyze the interactions of bacterial RNA polymerase with DNA, from its initial contact with a gene, through the act of transcription, culminating in its release when the transcript has been completed. *10 The operon* discusses the various means by which regulatory proteins can assist or prevent bacterial RNA polymerase from recognizing a particular gene for transcription. *11 Regulatory circuits* discusses other means of regulation, including the use of small RNAs, and considers how these interactions can be connected into larger regulatory networks. In *12 Phage strategies* we consider how individual regulatory interactions can be connected into more complex networks. In 2/ *Promoters and Enhancers* and *22 Activating transcription,* we consider the analogous reactions between eukaryotic RNA polymerases and their templates.

9.2 Transcription occurs by base pairing in a "bubble" of unpaired DNA

Key Concepts

- RNA polymerase separates the two strands of DNA in a transient "bubble" and uses one strand as a template to direct synthesis of a complementary sequence of RNA.
- The length of the bubble is ~12-14 bp, and the length of RNA-DNA hybrid within it is ~8-9 bp.

Transcription takes place by the usual process of complementary base pairing. **Figure 9.3** illustrates the general principle of transcription. RNA synthesis takes place within a "transcription bubble," in which DNA is transiently separated into its single strands, and the template strand is used to direct synthesis of the RNA strand.

The RNA chain is synthesized from the 5' end toward the 3' end. The 3'–OH group of the last nucleotide added to the chain reacts with an incoming nucleoside 5' triphosphate. The incoming nucleotide loses its terminal two phosphate groups (γ and β); its a group is used in the phosphodiester bond linking it to the chain. The overall reaction rate is ~40 nucleotides/second at 37°C (for the bacterial RNA polymerase);

this is about the same as the rate of translation (15 amino acids/sec), but much slower than the rate of DNA replication (800 bp/sec).

RNA polymerase creates the transcription bubble when it binds to a promoter. **Figure 9.4** shows that as RNA polymerase moves along the DNA, the bubble moves with it, and the RNA chain grows longer. The process of base pairing and base addition within the bubble is catalyzed and scrutinized by the enzyme.

The structure of the bubble within RNA polymerase is shown in the expanded view of **Figure 9.5.** As RNA polymerase moves along the DNA template, it unwinds the duplex at the front of the bubble (the unwinding point), and rewinds the DNA at the back (the rewinding point). The length of the transcription bubble is ~12-14 bp, but the length of the RNA-DNA hybrid region within it is shorter.

There is a major change in the topology of DNA extending over ~1 turn, but it is not clear how much of this region is actually base paired with RNA at any given moment. Certainly the RNA-DNA hybrid is short and transient. As the enzyme moves on, the DNA duplex reforms, and the RNA is displaced as a free polynucleotide chain. About the last 25 ribonucleotides added to a growing chain are complexed with DNA and/or enzyme at any moment.

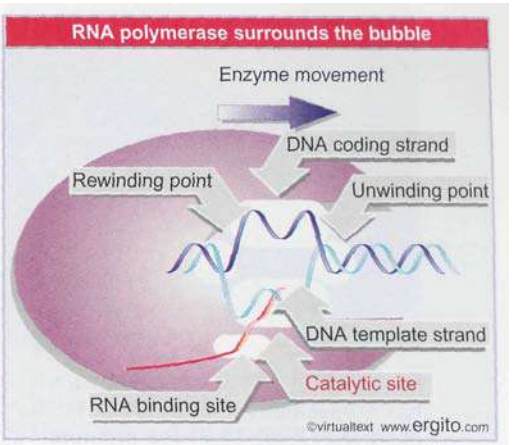


Figure 9.5 During transcription, the bubble is maintained within bacterial RNA polymerase, which unwinds and rewinds DNA, maintains the conditions of the partner and template DNA strands, and synthesizes RNA.

9.3 The transcription reaction has three stages

Key Concepts

- RNA polymerase initiates transcription after binding to a promoter site on DNA.
- During elongation the transcription bubble moves along DNA and the RNA chain is extended in the 5'-3' direction.
- Transcription stops, the DNA duplex reforms and RNA polymerase dissociates at a terminator site.

The transcription reaction can be divided into the stages illustrated in **Figure 9.6,** in which a bubble is created. RNA synthesis begins, the bubble moves along the DNA, and finally is terminated:

- *Template recognition* begins with the binding of RNA polymerase to the double-stranded DNA at a promoter to form a "closed complex". Then the strands of DNA are separated to form the "open complex" that makes the template strand available for base pairing with ribonucleotides. The transcription bubble is created by a local unwinding that begins at the site bound by RNA polymerase.
- Initiation describes the synthesis of the first nucleotide bonds in RNA. The enzyme remains at the promoter while it synthesizes the first ~9 nucleotide bonds. The initiation phase is protracted by the occurrence of abortive events, in which the enzyme makes short transcripts, releases them, and then starts synthesis of RNA again. The initiation phase ends when the enzyme succeeds in extending the chain and clears the promoter. *The sequence of DNA needed for RNA polymerase to bind to the template and accomplish the initiation reaction defines the promoter.* Abortive initiation probably involves synthesizing an RNA chain that fills the active site. If the RNA is released, the initiation is aborted and must start again. Initiation is accomplished if and when the enzyme manages to move along the template to move the next region of the DNA into the active site.
- During elongation the enzyme moves along the DNA and extends the growing RNA chain. As the enzyme moves, it unwinds the DNA helix to expose a new segment of the template in single-stranded

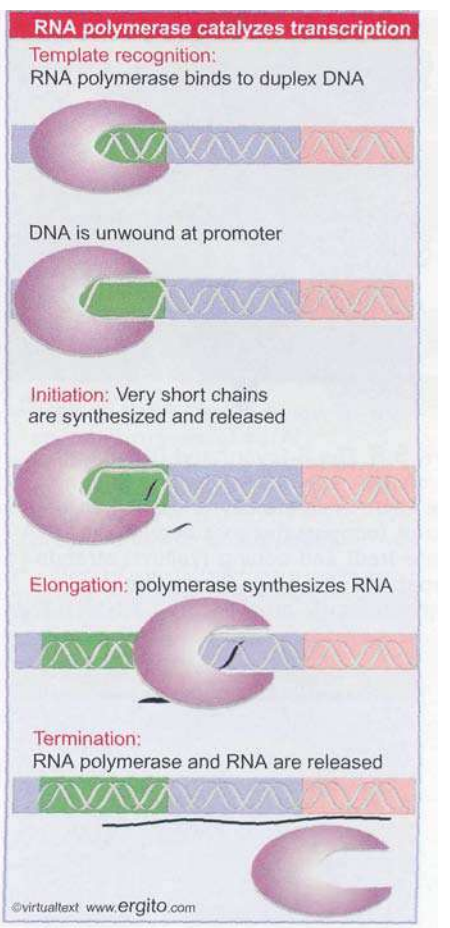


Figure 9.6 Transcription has four stages, which involve different types of interaction between RNA polymerase and DNA. The enzyme binds to the promoter and melts DNA, remains stationary during initiation, moves along the template during elongation, and dissociates at termination.

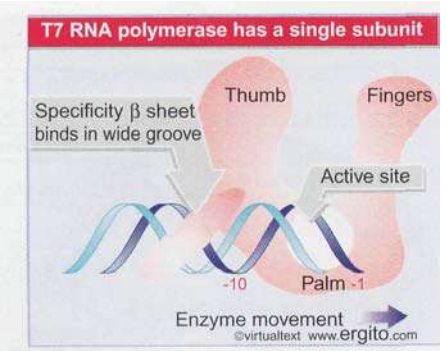


Figure 9.7 T7 RNA polymerase has a specificity loop that binds positions -7 to -11 of the promoter while positions -1 to -4 enter the active site.

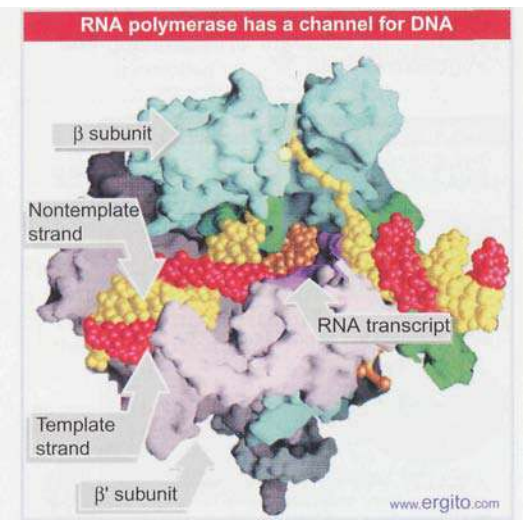


Figure 9.8 The β (cyan) and β′ subunit (pink) of RNA polymerase have a channel for the DNA template. Synthesis of an RNA transcript (copper) has just begun; the DNA template (red) and coding (yellow) strands are separated in a transcription bubble. Photograph kindly provided by Seth Darst.

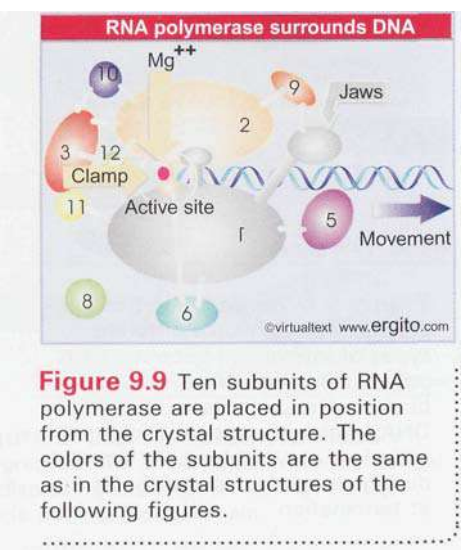


Figure 9.9 Ten subunits of RNA polymerase are placed in position from the crystal structure. The colors of the subunits are the same as in the crystal structures of the following figures.

condition. Nucleotides are covalently added to the 3' end of the growing RNA chain, forming an RNA-DNA hybrid in the unwound region. Behind the unwound region, the DNA template strand pairs with its original partner to reform the double helix. The RNA emerges as a free single strand. *Elongation involves the movement of the transcription bubble by a disruption of DNA structure, in which the template strand of the transiently unwound region is paired with the nascent RNA at the growing point.*

- Termination involves recognition of the point at which no further bases should be added to the chain. To terminate transcription, the formation of phosphodiester bonds must cease, and the transcription complex must come apart. When the last base is added to the RNA chain, the transcription bubble collapses as the RNA-DNA hybrid is disrupted, the DNA reforms in duplex state, and the enzyme and RNA are both released. *The sequence of DNA required for these reactions defines the terminator.*

The traditional view of elongation has been that it is a monotonic process, in which the enzyme moves forward 1 bp along DNA for every nucleotide added to the RNA chain. Changes in this pattern occur in certain circumstances, in particular when RNA polymerase pauses. One type of pattern is for the "front end" of the enzyme to remain stationary while the "back end" continues to move, thus compressing the footprint on DNA. After movement of several base pairs, the "front end" is released, restoring a footprint of full length. This gave rise to the "inchworm" model of transcription, in which the enzyme proceeds discontinuously, alternatively compressing and releasing the footprint on DNA. However, it may be the case that these events describe an aberrant situation rather than normal transcription.

9.4 Phage T7 RNA polymerase is a useful model system

Key Concepts

- T3 and T7 phage RNA polymerases are single polypeptides with minimal activities in recognizing a small number of phage promoters.
- Crystal structures of T7 RNA polymerase with DNA identify the DNA-binding region and the active site.

The existence of very small RNA polymerases, comprising single polypeptide chains coded by certain phages, gives some idea of the "minimum" apparatus necessary for transcription. These RNA polymerases recognize just a few promoters on the phage DNA; and they have no ability to change the set of promoters to which they respond. They provide simple model systems for characterizing the binding of RNA polymerase to DNA and the initiation reaction.

The RNA polymerases coded by the related phages T3 and T7 are single polypeptide chains of < 100 kD each. They synthesize RNA at rates of ~200 nucleotides/second at 37°C, more rapidly than bacterial RNA polymerase.

The T7 RNA polymerase is homologous to DNA polymerases, and has a similar structure, in which DNA lies in a "palm" surrounded by "fingers" and a "thumb" (see Figure 14.7). We now have a direct view of the active site from a crystal structure of a phage T7 RNA polymerase engaged in transcription.

The T7 RNA polymerase recognizes its target sequence in DNA by binding to bases in the major groove at a position upstream from the startpoint, as shown in **Figure 9.7.** The enzyme uses a *specificity loop* that is formed by a β ribbon. This feature is unique to the RNA polymerase (it is not found in DNA polymerases). The common point with all RNA polymerases is that the enzyme recognizes specific bases in DNA that are upstream of the sequence that is transcribed.

When transcription initiates, the conformation of the enzyme remains essentially the same while several nucleotides are added, and the transcribed template strand is "scrunched" in the active site. The active site can hold a transcript of 6-9 nucleotides. The transition from initiation to elongation is defined as the point when the enzyme begins to move along DNA. This occurs when the nascent transcript extends beyond the active site and interacts with the specificity loop. The RNA emerges to the surface of the enzyme when 12-14 nucleotides have been synthesized. These features are similar to those displayed by bacterial RNA polymerase.

9.5 A model for enzyme movement is suggested by the crystal structure

We now have much information about the structure and function of RNA polymerase as the result of the crystal structures of the bacterial and yeast enzymes. Bacterial RNA polymerase has overall dimensions of ~90 X 95 X 160 A. Eukaryotic RNA polymerase is larger but less elongated. Structural analysis shows that they share a common type of structure, in which there is a "channel" or groove on the surface ~25 A wide that could be the path for DNA. This is illustrated in **Figure 9.8** for the example of bacterial RNA polymerase. The length of the groove could hold 16 bp in the bacterial enzyme, and ~25 bp in the eukaryotic enzyme, but this represents only part of the total length of DNA bound during transcription.

The yeast enzyme is a large structure with 12 subunits (see *21.2 Eukaryotic RNA polymerases consist of many subunits).* Ten subunits of the yeast RNA polymerase II have been located on the crystal structure, as shown in **Figure 9.9.** The catalytic site is formed by a cleft between the two large subunits (#1 and #2), which grasp DNA downstream in "jaws" as it enters the RNA polymerase. Subunits 4 and 7 are missing from this structure; they form a subcomplex that dissociates from the complete enzyme. The structure is generally similar to that of bacterial RNA polymerase. This can be seen more clearly in the crystal structure of **Figure 9.10.** RNA polymerase surrounds the DNA, as seen in the view of **Figure 9.11.** A catalytic Mg^{2+} ion is found at the active site. The DNA is clamped in position at the active site by subunits **1,** 2, and 6. **Figure 9.12** shows that DNA is forced to take a turn at the entrance to the site, because of an adjacent wall of protein. The length of the RNA hybrid is limited by another protein obstruction, called the rudder. Nucleotides probably enter the active site from below, via pores through the structure.

The expanded view of the active site in **Figure 9.13** shows that the transcription bubble includes 9 bp of DNA-RNA hybrid. Where the DNA takes its turn, the bases downstream are flipped out of the DNA helix. As the enzyme moves along DNA, the base in the template strand at the start of the turn will be flipped to face the nucleotide entry site. The RNA-DNA hybrid is 9 bp long, and the 5' end of the RNA is forced to leave the DNA when it hits the protein rudder (see Figure 9.12).

Once DNA has been melted, the individual strands have a flexible structure in the transcription bubble. This enables DNA to take its turn in the active site. But before transcription starts, the DNA double helix is a relatively rigid straight structure. How does this structure enter the

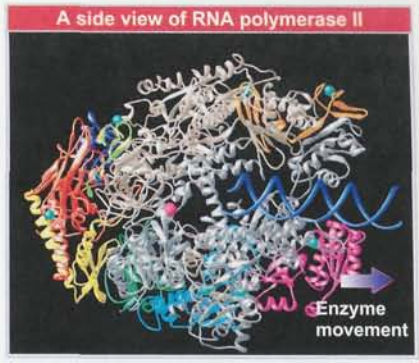


Figure 9.10 The side view of the crystal structure of RNA polymerase II from yeast shows that DNA is held downstream by a pair of jaws. Photograph kindly provided by Roger Kornberg.

Figure 9.11 The end view of the crystal structure of RNA polymerase II from yeast shows that DNA is surrounded by ~270° of protein. Photograph kindly provided by Roger Kornberg.

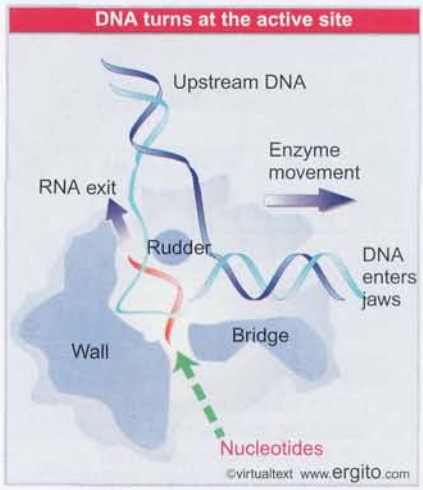


Figure 9.12 DNA is forced to make a turn at the active site by a wall of protein. Nucleotides may enter the active site through a pore in the protein.

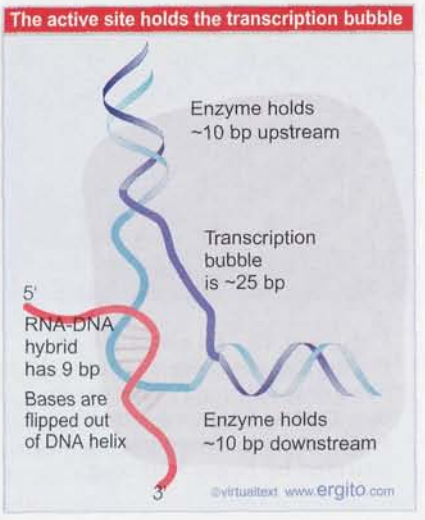


Figure 9.13 An expanded view of the active site shows the sharp turn in the path of DNA.

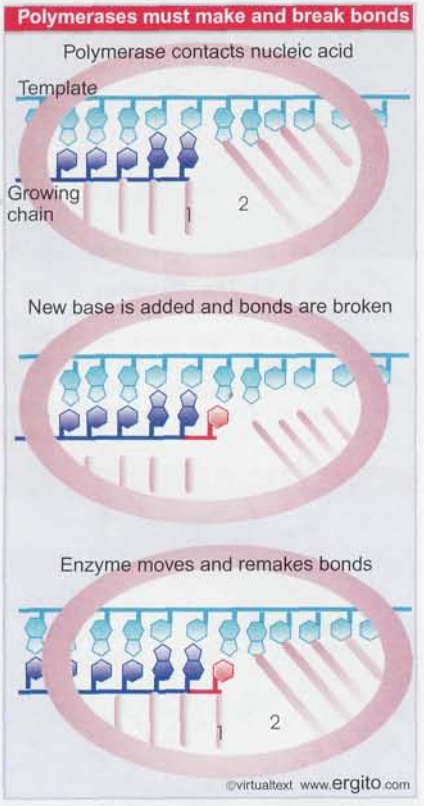


Figure 9.14 Movement of a nucleic acid polymerase requires breaking and remaking bonds to the nucleotides that occupy fixed positions relative to the enzyme structure. The nucleotides in these positions change each time the enzyme moves a base along the template.

polymerase without being blocked by the wall? The answer is that a large conformational shift must occur in the enzyme. Adjacent to the wall is a clamp. In the free form of RNA polymerase, this clamp swings away from the wall to allow DNA to follow a straight path through the enzyme. After DNA has been melted to create the transcription bubble, the clamp must swing back into position against the wall.

One of the dilemmas of any nucleic acid polymerase is that the enzyme must make tight contacts with the nucleic acid substrate and product, but must break these contacts and remake them with each cycle of nucleotide addition. Consider the situation illustrated in **Figure 9.14.** A polymerase makes a series of specific contacts with the bases at particular positions. For example, contact "1" is made with the base at the end of the growing chain, and contact "2" is made with the base in the template strand that is complementary to the next base to be added. But the bases that occupy these locations in the nucleic acid chains change every time a nucleotide is added!

The top and bottom panels of the figure show the same situation: a base is about to be added to the growing chain. The difference is that the growing chain has been extended by one base in the bottom panel. The geometry of both complexes is exactly the same, but contacts "1" and "2" in the bottom panel are made to bases in the nucleic acid chains that are located one position farther along the chain. The middle panel shows that this must mean that, after the base is added, and before the enzyme moves relative to the nucleic acid, the contacts made to specific positions must be broken so that they can be remade to bases that occupy those positions after the movement.

The RNA polymerase structure suggests an insight into how the enzyme retains contact with its substrate while breaking and remaking bonds. A structure in the protein called the bridge is adjacent to the active site (see Figure 9.12). This feature is found in both the bacterial and yeast enzymes, but it has different shapes in the different crystal structures. In one it is bent, and in the other it is straight. **Figure 9.15** suggests that the change in conformation of the bridge structure is closely related to translocation of the enzyme along the nucleic acid.

At the start of the cycle of translocation, the bridge has a straight conformation adjacent to the nucleotide entry site. This allows the next nucleotide to bind at the nucleotide entry site. The bridge is in contact with the newly added nucleotide. Then the protein moves one base pair along the substrate. The bridge changes its conformation, bending to keep contact with the newly added nucleotide. In this conformation, the bridge obscures the nucleotide entry site. To end the cycle, the bridge returns to its straight conformation, allowing access again to the nucleotide entry site. The bridge acts as a ratchet that releases the DNA and RNA strands for translocation while holding on to the end of the growing chain.

9.6 Bacterial RNA polymerase consists of multiple subunits

Key Concepts

- Bacterial RNA core polymerases are ~500 kD multisubunit complexes with the general structure $\alpha_2\beta\beta'$.
- DNA is bound in a channel and is contacted by both the β and subunits.

The best characterized RNA polymerases are those of eubacteria, for which *E. coli* is a typical case. *A single type of RNA poly-*

merase *appears to be responsible for almost all synthesis of mRNA, and all rRNA and tRNA, in a eubacterium.* About 7000 RNA polymerase molecules are present in an *E. coli* cell. Many of them are engaged in transcription; probably 2000-5000 enzymes are synthesizing RNA at any one time, the number depending on the growth conditions.

The complete enzyme or holoenzyme in *E. coli* has a molecular weight of ~465 kD. Its subunit composition is summarized in **Figure 9.16.**

The β and β′ subunits together make up the catalytic center. Their sequences are related to those of the largest subunits of eukaryotic RNA polymerases (see *21.2 Eukaryotic RNA polymerases consist of many subunits),* suggesting that there are common features to the actions of all RNA polymerases. The β subunit can be crosslinked to the template DNA, the product RNA, and the substrate ribonucleotides; mutations in *rpoB* affect all stages of transcription. Mutations in *rpoC* show that β′ also is involved at all stages.

The a subunit is required for assembly of the core enzyme. When phage T4 infects *E. coli,* the a subunit is modified by ADP-ribosylation of an arginine. The modification is associated with a reduced affinity for the promoters formerly recognized by the holoenzyme, suggesting that the a subunit plays a role in promoter recognition. The a subunit also plays a role in the interaction of RNA polymerase with some regulatory factors.

The a subunit is concerned specifically with promoter recognition, and we have more information about its functions than on any other subunit (see next section).

The crystal structure of the bacterial enzyme (Figure 9.8) shows that the channel for DNA lies at the interface of the β and β′ subunits. (The α subunits are not visible in this view.) The DNA is unwound at the active site, where an RNA chain is being synthesized. Crosslinking experiments identify the points at which the RNA polymerase subunits contact DNA. These are summarized in **Figure 9.17.** The β and β′ subunits contact DNA at many points downstream of the active site. They make several contacts with the coding strand in the region of the transcription bubble, thus stabilizing the separated single strands. The RNA is contacted largely in the region of the transcription bubble.

The drug rifampicin (a member of the rifamycin antibiotic family) blocks transcription by bacterial RNA polymerase. It is a major drug used against tuberculosis. The crystal structure of RNA polymerase bound to rifampicin explains its action: it binds in a pocket of the *ft* subunit, >12 A away from the active site, but in a position where it blocks the path of the elongating RNA. By preventing the RNA chain from extending beyond 2-3 nucleotides, this blocks transcription.

Originally defined simply by its ability to incorporate nucleotides into RNA under the direction of a DNA template, the enzyme RNA polymerase now is seen as part of a more complex apparatus involved in transcription. *The ability to catalyze RNA synthesis defines the minimum component that can be described as RNA polymerase.* It supervises the base pairing of the substrate ribonucleotides with DNA and catalyzes the formation of phosphodiester bonds between them.

All of the subunits of the basic polymerase that participate in elongation are necessary for initiation and termination. But transcription units differ in their dependence on additional polypeptides at the initiation and termination stages. Some of these additional polypeptides are needed at all genes, but others may be needed specifically for initiation or termination at particular genes. The analogy with the division of labors between the ribosome and the protein synthesis factors is obvious.

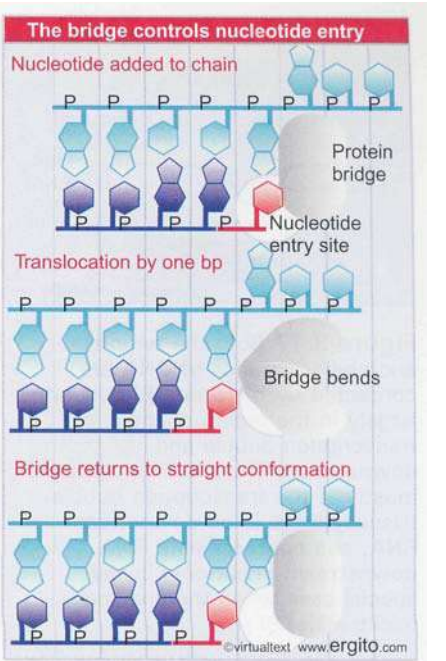


Figure 9.15 The RNA polymerase elongation cycle starts with a straight bridge adjacent to the nucleotide entry site. After nucleotide addition, the enzyme moves one base pair and bridge bends as it retains contact with the newly added nucleotide. When the bridge is released, the cycle can start again.

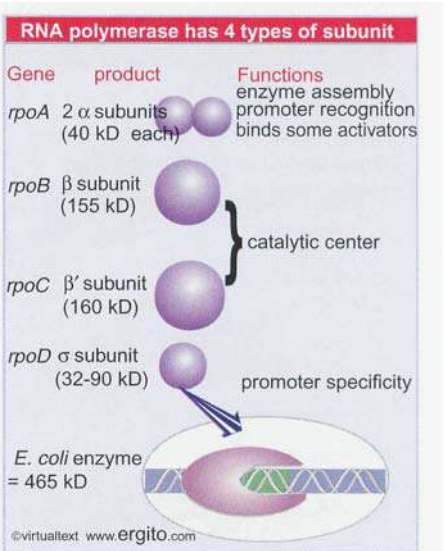


Figure 9.16 Eubacterial RNA polymerases have four types of subunit; α, β, and β have rather constant sizes in different bacterial species, but σ varies more widely.

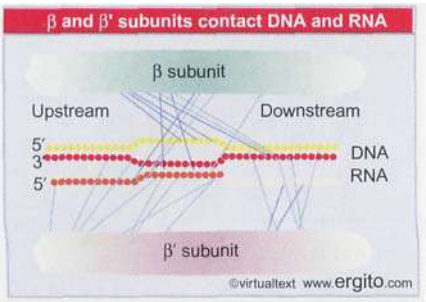


Figure 9.17 Both the template and coding strands of DNA are contacted by the β and β′ subunits largely in the region of the transcription bubble and downstream. The RNA is contacted mostly in the transcription bubble. (Usually there is no downstream RNA, and contacts with RNA downstream occur only in the special case when the enzyme backtracks.)

E. coli RNA polymerase can transcribe any one of many (>1000) transcription units. The enzyme therefore requires the ability to interact with a variety of host and phage functions that modify its intrinsic transcriptional activities. The complexity of the enzyme therefore at least in part reflects its need to interact with regulatory factors, rather than any demand inherent in its catalytic activity.

9.7 RNA polymerase consists of the core enzyme and sigma factor

Key Concepts

- Bacterial RNA polymerase can be divided into the $\alpha_2\beta\beta'$ core enzyme that catalyzes transcription and the sigma subunit that is required only for initiation.
- Sigma factor changes the DNA-binding properties of RNA polymerase so that its affinity for general DNA is reduced and its affinity for promoters is increased.
- Binding constants of RNA polymerase for different promoters vary over 6 orders of magnitude, corresponding to the frequency with which transcription is initiated at each promoter.

The holoenzyme ($\alpha_2\beta\beta'\sigma$) can be separated into two components, the core enzyme ($\alpha_2\beta\beta'$) and the sigma factor (the a polypeptide). *Only the holoenzyme can initiate transcription. Sigma factor ensures that bacterial RNA polymerase binds in a stable manner to DNA only at promoters.* The sigma "factor" is usually released when the RNA chain reaches 8-9 bases, leaving the core enzyme to undertake elongation. *Core enzyme has the ability to synthesize RNA on a DNA template, but cannot initiate transcription at the proper sites.*

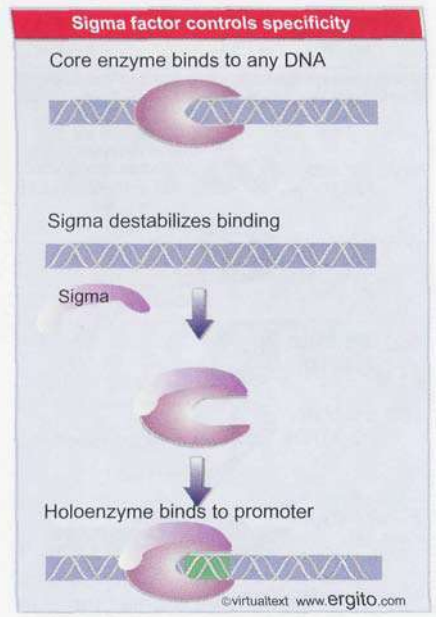


Figure 9.18 Core enzyme binds indiscriminately to any DNA. Sigma factor reduces the affinity for sequence-independent binding, and confers specificity for promoters.

The core enzyme has a general affinity for DNA, in which electrostatic attraction between the basic protein and the acidic nucleic acid plays a major role. Any (random) sequence of DNA that is bound by core polymerase in this general binding reaction is described as a loose binding site. No change occurs in the DNA, which remains duplex. The complex at such a site is stable, with a half-life for dissociation of the enzyme from DNA ~60 minutes. *The core enzyme does not distinguish between promoters and other sequences of DNA.*

Figure 9.18 shows that sigma factor introduces a major change in the affinity of RNA polymerase for DNA. *The holoenzyme has a drastically reduced ability to recognize loose binding sites*—that is, to bind to any general sequence of DNA. The association constant for the reaction is reduced by a factor of ~10^4, and the half-life of the complex is <1 second. So sigma factor destabilizes the general binding ability very considerably.

But sigma factor also *confers the ability to recognize specific binding sites.* The holoenzyme binds to promoters very tightly, with an association constant increased from that of core enzyme by (on average) 1000 times and with a half-life of several hours.

The specificity of holoenzyme for promoters compared to other sequences is ~10^7, but this is only an average, because there is wide variation in the rate at which the holoenzyme binds to different promoter sequences. This is an important parameter in determining the efficiency of an individual promoter in initiating transcription. The binding constants range from ~10^{12} to ~10^6. Other factors also affect the frequency of initiation, which varies from ~1/sec (rRNA genes) to ~1/30 min (the *lacI* promoter).

By Book_Crazy [IND]

9.8 The association with sigma factor changes at initiation

Key Concepts

- When RNA polymerase binds to a promoter, it separates the DNA strands to form a transcription bubble and incorporates up to 9 nucleotides into RNA.
- There may be a cycle of abortive initiations before the enzyme moves to the next phase.
- Sigma factor may be released from RNA polymerase when the nascent RNA chain reaches 8-9 bases in length.

We can now describe the stages of transcription in terms of the interactions between different forms of RNA polymerase and the DNA template. The initiation reaction can be described by the parameters that are summarized in **Figure 9.19:**

- The holoenzyme·promoter reaction starts by forming a closed binary complex. "Closed" means that the DNA remains duplex. Because the formation of the closed binary complex is reversible, it is usually described by an equilibrium constant (K_B). There is a wide range in values of the equilibrium constant for forming the closed complex.
- The closed complex is converted into an open complex by "melting" of a short region of DNA within the sequence bound by the enzyme. The series of events leading to formation of an open complex is called tight binding. For strong promoters, conversion into an open binary complex is irreversible, so this reaction is described by a rate constant (k_2). This reaction is fast. Sigma factor is involved in the melting reaction (see *9.16 Substitution of sigma factors may control initiation).*
- The next step is to incorporate the first two nucleotides; then a phosphodiester bond forms between them. This generates a ternary complex that contains RNA as well as DNA and enzyme. Formation of the ternary complex is described by the rate constant k_i; this is even faster than the rate constant k_2. Further nucleotides can be added without any enzyme movement to generate an RNA chain of up to 9 bases. After each base is added, there is a certain probability that the enzyme will release the chain. This comprises an abortive initiation, after which the enzyme begins again with the first base. A cycle of abortive initiations usually occurs to generate a series of very short oligonucleotides.
- When initiation succeeds, sigma is no longer necessary, and the enzyme makes the transition to the elongation ternary complex of core polymerase·DNA·nascent RNA. The critical parameter here is *how long it takes for the polymerase to leave the promoter so another polymerase can initiate.* This parameter is the promoter clearance time; its minimum value of 1-2 sec establishes the maximum frequency of initiation as <1 event per second. The enzyme then moves along the template, and the RNA chain extends beyond 10 bases.

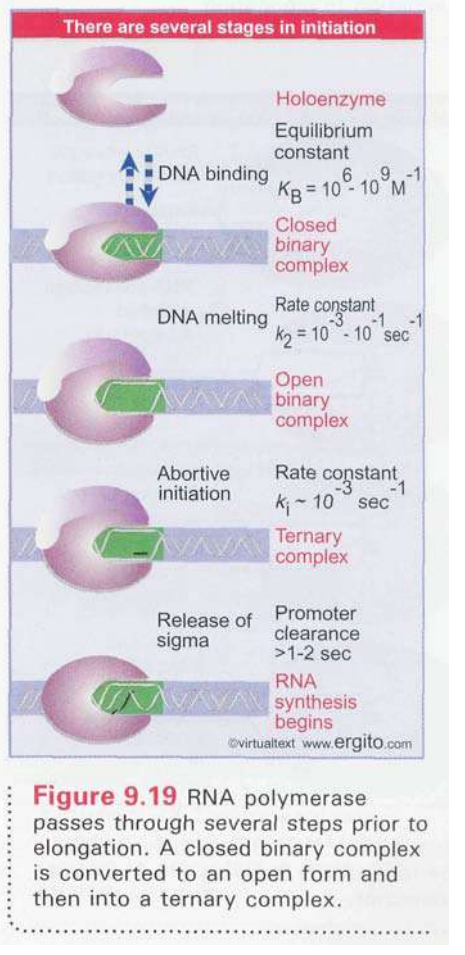


Figure 9.19 RNA polymerase passes through several steps prior to elongation. A closed binary complex is converted to an open form and then into a ternary complex.

When RNA polymerase binds to DNA, the elongated dimension of the protein extends along the DNA, but some interesting changes in shape occur during transcription. Transitions in shape and size identify three forms of the complex, as illustrated in **Figure 9.20:**

- When RNA polymerase holoenzyme initially binds to DNA, it covers some 75-80 bp, extending from -55 to +20. (The long dimension of RNA polymerase (160 A) could cover ~50 bp of DNA in extended

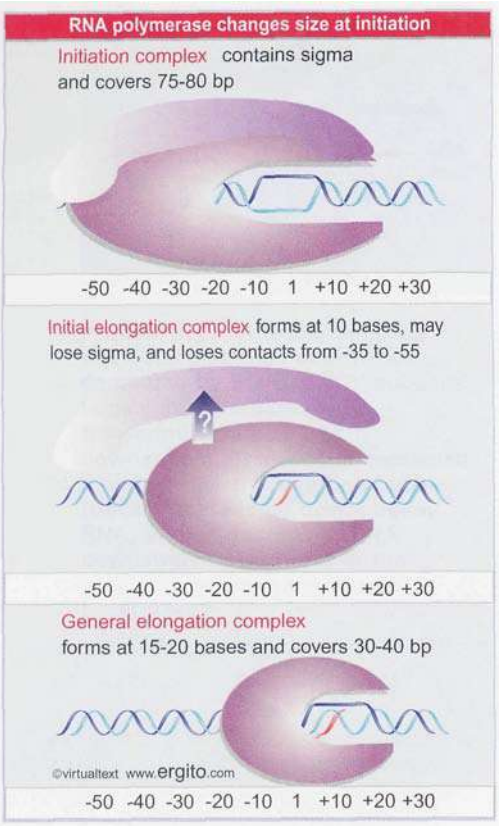


Figure 9.20 The length of DNA bound by RNA polymerase changes as it moves from initiation to elongation.

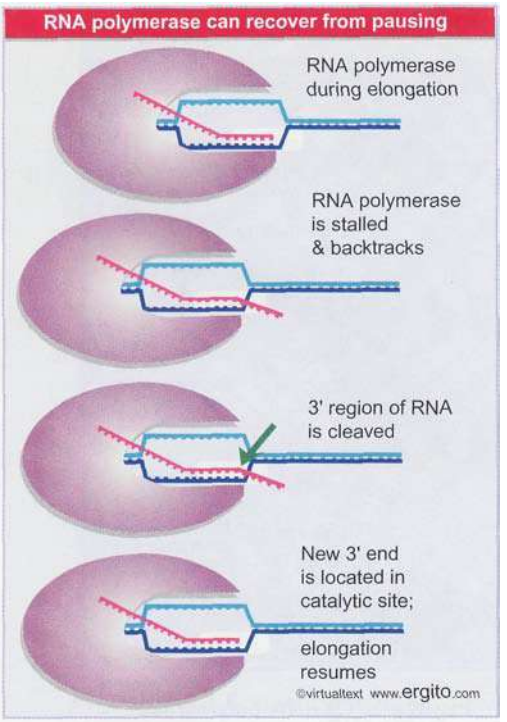


Figure 9.21 A stalled RNA polymerase can be released by cleaving the 3′ end of the transcript.

form, which implies that binding of a longer stretch of DNA must involve some bending of the nucleic acid.)

- The shape of the RNA polymerase changes at the transition from initiation to elongation. This is associated with the loss of contacts in the -55 to -35 region, leaving only ~60 bp of DNA covered by the enzyme. This corresponds with the concept that the more upstream part of the promoter is involved in initial recognition by RNA polymerase, but is not required for the later stages of initiation (*9.13 Promoter efficiencies can be increased or decreased by mutation*).
- When the RNA chain extends to 15-20 bases, the enzyme makes a further transition, to form the complex that undertakes elongation; now it covers 30-40 bp (depending on the stage in the elongation cycle).

It has been a tenet of transcription since soon after the discovery of sigma factor that it is released after initiation. However, this may not be strictly true. Direct measurements of elongating RNA polymerase complexes show that ~70% of them retain sigma factor. Since a third of elongating polymerases lack sigma, the original conclusion is certainly correct that it is not necessary for elongation. In those cases where it remains associated with core enzyme, the nature of the association has almost certainly changed (see *9.11 Sigma factor controls binding to DNA*).

9.9 A stalled RNA polymerase can restart

Key Concepts

- An arrested RNA polymerase can restart transcription by cleaving the RNA transcript to generate a new 3' end.

RNA polymerase must be able to handle situations when transcription is blocked. This can happen, for example, when DNA is damaged. A model system for such situations is provided by arresting elongation *in vitro* by omitting one of the necessary precursor nucleotides. When the missing nucleotide is restored, the enzyme can overcome the block by cleaving the 3′ end of the RNA, to create a new 3′ terminus for chain elongation. The cleavage involves accessory factors in addition to the enzyme itself. In the case of *E. coli* RNA polymerase, the proteins GreA and GreB release the RNA polymerase from elongation arrest. In eukaryotic cells, RNA polymerase II requires an accessory factor (TFIIS), which enables the polymerase to cleave a few ribonucleotides from the 3' terminus of the RNA product.

The catalytic site of RNA polymerase undertakes the actual cleavage in each case. There have been differences of opinion concerning the change in the enzyme that occurs at this time. One view is that there is an internal reorganization of structure, in which the catalytic center moves relative to the rest of the enzyme. The alternative model shown in **Figure 9.21** suggests that the enzyme as a whole "backtracks" on the DNA. The 3' terminus of the RNA is exposed in single-stranded form, and the RNA-DNA hybrid region reverses its position. Cleavage restores a normal elongation complex. This model is supported by more recent measurements showing a constant distance between the catalytic center and the "front end".

The reason for this reaction may be that stalling causes the template to be mispositioned, so that the 3′ terminus is no longer located in the active site. Cleavage and backtracking is necessary to place the terminus in the right location for addition of further bases.

We see therefore that RNA polymerase has the facility to unwind and rewind DNA, to hold the separated strands of DNA and the RNA product, to catalyze the addition of ribonucleotides to the growing RNA chain, and to adjust to difficulties in progressing by cleaving the RNA product and restarting RNA synthesis (with the assistance of some accessory factors).

9.10 How does RNA polymerase find promoter sequences?

Key Concepts

- The rate at which RNA polymerase binds to promoters is too fast to be accounted for by random diffusion.
- RNA polymerase probably binds to random sites on DNA and exchanges them with other sequences very rapidly until a promoter is found.

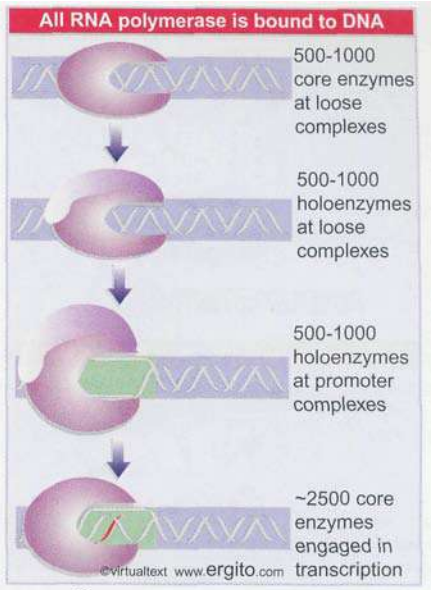


Figure 9.22 Core enzyme and holoenzyme are distributed on DNA, and very little RNA polymerase is free.

How is RNA polymerase distributed in the cell? A (somewhat speculative) picture of the enzyme's situation is depicted in Figure **9.22:**

- Excess core enzyme exists largely as closed loose complexes, because the enzyme enters into them rapidly and leaves them slowly. There is very little, if any, free core enzyme.
- There is enough sigma factor for about one third of the polymerases to exist as holoenzymes, and they are distributed between loose complexes at nonspecific sites and binary complexes (mostly closed) at promoters.
- About half of the RNA polymerases consist of core enzymes engaged in transcription.
- How much holoenzyme is free? We do not know, but we suspect that the amount is very small.

RNA polymerase must find promoters within the context of the genome. Suppose that a promoter is a stretch of ~60 bp; how is it distinguished from the 4×10^6 bp that comprise the *E. coli* genome? The next three figures illustrate the principle of some possible models.

Figure 9.23 shows the simplest model for promoter binding, in which RNA polymerase moves by random diffusion. Holoenzyme very rapidly associates with, and dissociates from, loose binding sites. So it could continue to make and break a series of closed complexes until (by chance) it encounters a promoter. Then its recognition of the specific sequence would allow tight binding to occur by formation of an open complex.

For RNA polymerase to move from one binding site on DNA to another, it must dissociate from the first site, find the second site, and then associate with it. Movement from one site to another is limited by the speed of diffusion through the medium. Diffusion sets an upper limit for the rate constant for associating with a 60 bp target of $<10^8\ M^{-1}\ sec^{-1}$. But the actual forward rate constant for some promoters *in vitro* appears to be $\sim 10^8\ M^{-1}\ sec^{-1}$, at or above the diffusion limit. If this value applies *in vivo*, the time required for random cycles of successive association and dissociation at loose binding sites is too great to account for the way RNA polymerase finds its promoter.

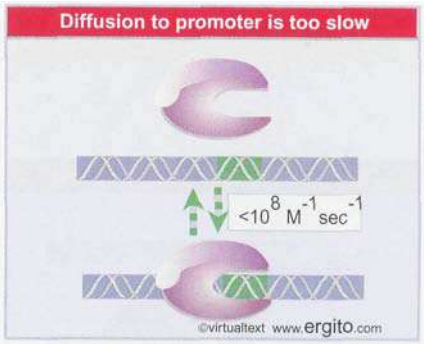


Figure 9.23 The forward rate constant for RNA polymerase binding to promoters is faster than random diffusion.

RNA polymerase must therefore use some other means to seek its binding sites. **Figure 9.24** shows that the process could be speeded up if the initial target for RNA polymerase is the whole genome, not just a

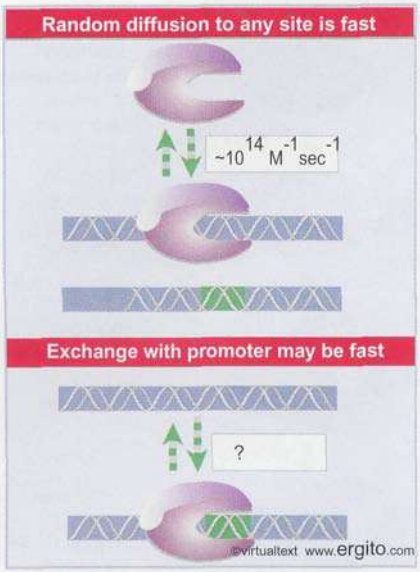


Figure 9.24 RNA polymerase binds very rapidly to random DNA sequences and could find a promoter by direct displacement of the bound DNA sequence.

specific promoter sequence. By increasing the target size, the rate constant for diffusion to DNA is correspondingly increased, and is no longer limiting.

If this idea is correct, a free RNA polymerase binds DNA and then remains in contact with it. How does the enzyme move from a random (loose) binding site on DNA to a promoter? The most likely model is to suppose that the bound sequence is directly displaced by another sequence. Having taken hold of DNA, the enzyme exchanges this sequence with another sequence very rapidly, and continues to exchange sequences until a promoter is found. Then the enzyme forms a stable complex, after which initiation occurs. The search process becomes much faster because association and dissociation are virtually simultaneous, and time is not spent commuting between sites. Direct displacement can give a "directed walk," in which the enzyme moves preferentially from a weak site to a stronger site.

Another idea supposes that the enzyme slides along the DNA by a one-dimensional random walk, as shown in **Figure 9.25,** being halted only when it encounters a promoter. However, there is no evidence that RNA polymerase (or other DNA-binding proteins) can function in this manner.

Figure 9.25 RNA polymerase does not slide along DNA.

9.11 Sigma factor controls binding to DNA

Key Concepts

- A change in association between sigma and holoenzyme changes binding affinity for DNA so that core enzyme can move along DNA.

RNA polymerase encounters a dilemma in reconciling its needs for initiation with those for elongation. Initiation requires tight binding *only* to particular sequences (promoters), while elongation requires close association with *all* sequences that the enzyme encounters during transcription. **Figure 9.26** illustrates how the dilemma is solved by the reversible association between sigma factor and core enzyme. As mentioned previously (see *9.8 The association with sigma factor changes at initiation),* sigma factor is either released following initiation or changes its association with core enzyme so that it no longer participates in DNA binding. Because there are fewer molecules of sigma than of core enzyme, the utilization of core enzyme requires that sigma recycles. This occurs immediately after initiation (as shown in the figure) in about one third of cases; presumably sigma and core dissociate at some later point in the other cases.

Irrespective of the exact timing of its release from core enzyme, sigma factor is involved only in initiation. It becomes unnecessary when abortive initiation is concluded and RNA synthesis has been successfully initiated. We do not know whether the state of polymerase changes as a consequence of overcoming abortive initiation, or whether instead it is the change in state that ends abortive initiation and allows elongation to commence.

When sigma factor is released from core enzyme, it becomes immediately available for use by another core enzyme. Whether sigma is released or remains more loosely associated with core enzyme, the core enzyme in the ternary complex is bound very tightly to DNA. It is essentially "locked in" until elongation has been completed. When transcription terminates, the core enzyme is released. It is then "stored" by binding to a loose site on DNA. If it has lost its sigma factor, it must

find another sigma factor in order to undertake a further cycle of transcription.

Core enzyme has a high intrinsic affinity for DNA, which is increased by the presence of nascent RNA. But its affinity for loose binding sites is too high to allow the enzyme to distinguish promoters efficiently from other sequences. By reducing the stability of the loose complexes, sigma allows the process to occur much more rapidly; and by stabilizing the association at tight binding sites, the factor drives the reaction irreversibly into the formation of open complexes. When the enzyme releases sigma (or changes its association with it), it reverts to a general affinity for all DNA, irrespective of sequence, that suits it to continue transcription.

What is responsible for the ability of holoenzyme to bind specifically to promoters? Sigma factor has domains that recognize the promoter DNA. As an independent polypeptide, sigma does not bind to DNA, but when holoenzyme forms a tight binding complex, σ contacts the DNA in the region upstream of the startpoint. This difference is due to a change in the conformation of sigma factor when it binds to core enzyme. The N-terminal region of free sigma factor suppresses the activity of the DNA-binding region; when sigma binds to core, this inhibition is released, and it becomes able to bind specifically to promoter sequences (see also Figure 9.36 later). The inability of free sigma factor to recognize promoter sequences may be important: if σ could freely bind to promoters, it might block holoenzyme from initiating transcription.

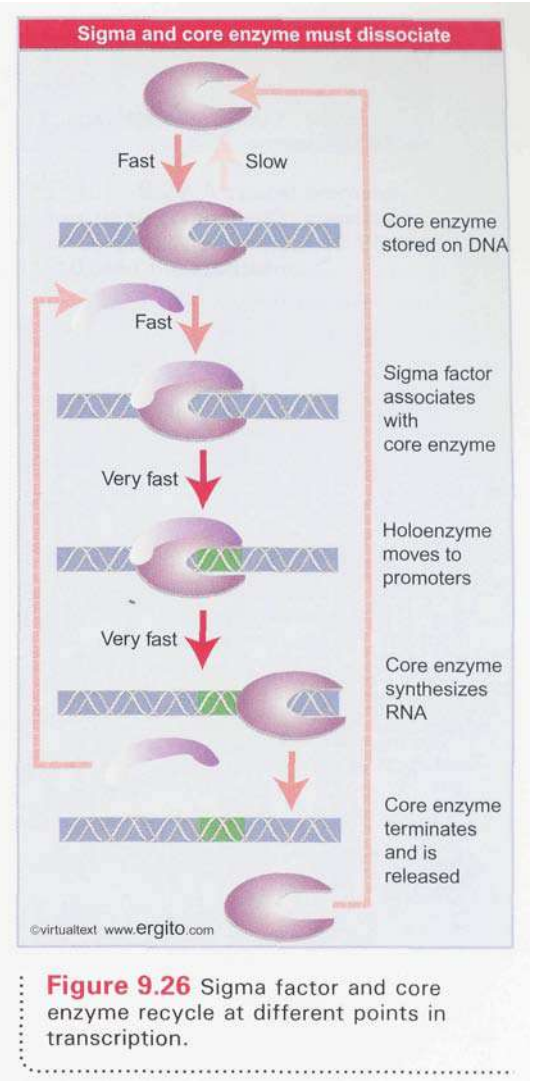


Figure 9.26 Sigma factor and core enzyme recycle at different points in transcription.

9.12 Promoter recognition depends on consensus sequences

Key Concepts

- A promoter is defined by the presence of short consensus sequences at specific locations.
- The promoter consensus sequences consist of a purine at the startpoint, the hexamer TATAAT centered at –10, and another hexamer centered at -35.
- Individual promoters usually differ from the consensus at one or more positions.

As a sequence of DNA whose function is to be *recognized by proteins*, a promoter differs from sequences whose role is to be transcribed or translated. The information for promoter function is provided directly by the DNA sequence: its structure is the signal. This is a classic example of a *cis*-acting site, as defined previously in Figure 1.41 and Figure 1.42. By contrast, expressed regions gain their meaning only after the information is transferred into the form of some other nucleic acid or protein.

A key question in examining the interaction between an RNA polymerase and its promoter is how the protein recognizes a specific promoter sequence. Does the enzyme have an active site that distinguishes the chemical structure of a particular sequence of bases in the DNA double helix? How specific are its requirements?

One way to design a promoter would be for a particular sequence of DNA to be recognized by RNA polymerase. Every promoter would consist of, or at least include, this sequence. In the bacterial genome, the minimum length that could provide an adequate signal is 12 bp. (Any shorter sequence is likely to occur—just by chance—a sufficient number of additional times to provide false signals. The minimum length required for

unique recognition increases with the size of genome.) The 12 bp sequence need not be contiguous. If a specific number of base pairs separates two constant shorter sequences, their combined length could be less than 12 bp, since the *distance* of separation itself provides a part of the signal (even if the intermediate *sequence* is itself irrelevant).

Attempts to identify the features in DNA that are necessary for RNA polymerase binding started by comparing the sequences of different promoters. Any essential nucleotide sequence should be present in all the promoters. Such a sequence is said to be conserved. However, a conserved sequence need not necessarily be conserved at every single position; some variation is permitted. How do we analyze a sequence of DNA to determine whether it is sufficiently conserved to constitute a recognizable signal?

Putative DNA recognition sites can be defined in terms of an idealized sequence that represents the base most often present at each position. A consensus sequence is defined by aligning all known examples so as to maximize their homology. For a sequence to be accepted as a consensus, each particular base must be reasonably predominant at its position, and most of the actual examples must be related to the consensus by rather few substitutions, say, no more than 1-2.

The striking feature in the sequence of promoters in *E. coli* is the *lack of any extensive conservation of sequence* over the 60 bp associated with RNA polymerase. The sequence of much of the binding site is irrelevant. But some short stretches within the promoter are conserved, and they are critical for its function. *Conservation of only very short consensus sequences is a typical feature of regulatory sites (such as promoters) in both prokaryotic and eukaryotic genomes.*

There are four (perhaps five) conserved features in a bacterial promoter: the startpoint; the –10 sequence; the -35 sequence; and the separation between the —10 and —35 sequences:

- The startpoint is usually (>90% of the time) a purine. It is common for the startpoint to be the central base in the sequence CAT, but the conservation of this triplet is not great enough to regard it as an obligatory signal.
- Just upstream of the startpoint, a 6 bp region is recognizable in almost all promoters. The center of the hexamer generally is close to 10 bp upstream of the startpoint; the distance varies in known promoters from position -18 to -9. Named for its location, the hexamer is often called the -10 sequence. Its consensus is *TATAAT*, and can be summarized in the form

$$T_{80}\,A_{95}\,T_{45}\,A_{60}\,A_{50}\,T_{96}$$

where the subscript denotes the percent occurrence of the most frequently found base, varying from 45-96%. (A position at which there is no discernible preference for any base would be indicated by N.) If the frequency of occurrence indicates likely importance in binding RNA polymerase, we would expect the initial highly conserved TA and the final almost completely conserved T in the —10 sequence to be the most important bases.

Another conserved hexamer is centered ~35 bp upstream of the startpoint. This is called the -35 sequence. The consensus is *TTGACA*; in more detailed form, the conservation is

$$T_{82}\,T_{84}\,G_{78}\,A_{65}\,C_{54}\,A_{45}$$

The distance separating the —35 and —10 sites is between 16-18 bp in 90% of promoters; in the exceptions, it is as little as 15 or as great as 20 bp. *Although the actual sequence in the intervening region is*

unimportant, the distance is critical in holding the two sites at the appropriate separation for the geometry of RNA polymerase.

- Some promoters have an A-T-rich sequence located farther upstream. This is called the UP element. It interacts with the a subunit of the RNA polymerase. It is typically found in promoters that are highly expressed, such as the promoters for rRNA genes.

The optimal promoter is a sequence consisting of the -35 hexamer, separated by 17 bp from the -10 hexamer, lying 7 bp upstream of the startpoint. The structure of a promoter, showing the permitted range of variation from this optimum, is illustrated in **Figure 9.27.**

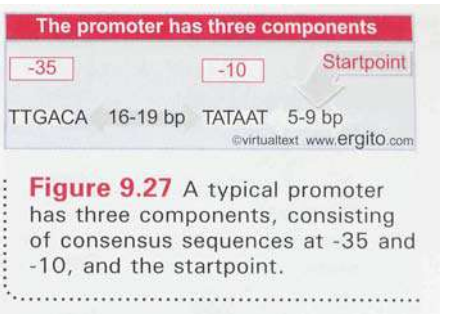


Figure 9.27 A typical promoter has three components, consisting of consensus sequences at -35 and -10, and the startpoint.

9.13 Promoter efficiencies can be increased or decreased by mutation

Key Concepts

- Down mutations to decrease promoter efficiency usually decrease conformance to the consensus sequences, whereas up mutations have the opposite effect.
- Mutations in the -35 sequence usually affect initial binding of RNA polymerase.
- Mutations in the –10 sequence usually affect the melting reaction that converts a closed complex to an open one.

Mutations are a major source of information about promoter function. Mutations in promoters affect the level of expression of the gene(s) they control, without altering the gene products themselves. Most are identified as bacterial mutants that have lost, or have very much reduced, transcription of the adjacent genes. They are known as down mutations. Less often, mutants are found in which there is increased transcription from the promoter. They have up mutations.

It is important to remember that "up" and "down" mutations are defined relative to the *usual* efficiency with which a particular promoter functions. This varies widely. So a change that is recognized as a down mutation in one promoter might never have been isolated in another (which in its wild-type state could be even less efficient than the mutant form of the first promoter). Information gained from studies *in vivo* simply identifies the overall direction of the change caused by mutation.

Is the most effective promoter one that has the actual consensus sequences? This expectation is borne out by the simple rule that up mutations usually increase homology with one of the consensus sequences or bring the distance between them closer to 17 bp. Down mutations usually decrease the resemblance of either site with the consensus or make the distance between them more distant from 17 bp. Down mutations tend to be concentrated in the most highly conserved positions, which confirms their particular importance as the main determinant of promoter efficiency. However, there are occasional exceptions to these rules.

To determine the absolute effects of promoter mutations, we must measure the affinity of RNA polymerase for wild-type and mutant promoters *in vitro*. There is ~ 100-fold variation in the rate at which RNA polymerase binds to different promoters *in vitro*, which correlates well with the frequencies of transcription when their genes are expressed *in vivo*. Taking this analysis further, we can investigate the stage at which a mutation influences the capacity of the promoter. Does it change the affinity of the promoter for binding RNA polymerase? Does it leave the

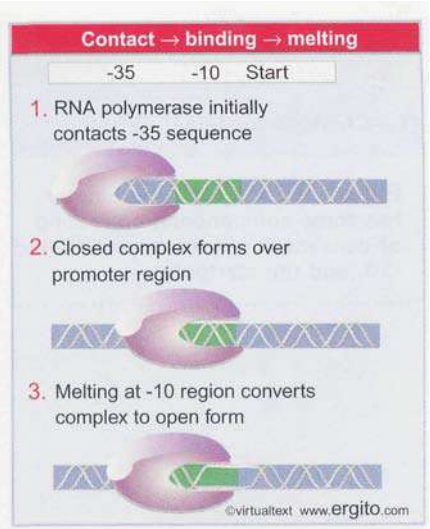


Figure 9.28 The –35 sequence is used for initial recognition, and the –10 sequence is used for the melting reaction that converts a closed complex to an open complex.

enzyme able to bind but unable to initiate? Is the influence of an ancillary factor altered?

By measuring the kinetic constants for formation of a closed complex and its conversion to an open complex, as defined in Figure 9.19, we can dissect the two stages of the initiation reaction:

- Down mutations in the —35 sequence reduce the rate of closed complex formation (they reduce K_B), but do not inhibit the conversion to an open complex.
- Down mutations in the —10 sequence do not affect the initial formation of a closed complex, but they slow its conversion to the open form (they reduce k_2).

These results suggest the model shown in **Figure 9.28.** The function of the -35 sequence is to provide the signal for recognition by RNA polymerase, while the –10 sequence allows the complex to convert from closed to open form. We might view the -35 sequence as comprising a "recognition domain," while the -10 sequence comprises an "unwinding domain" of the promoter.

The consensus sequence of the -10 site consists exclusively of AT base pairs, which assists the initial melting of DNA into single strands. The lower energy needed to disrupt A·T pairs compared with G·C pairs means that a stretch of A·T pairs demands the minimum amount of energy for strand separation.

The sequence immediately around the startpoint influences the initiation event. And the initial transcribed region (from +1 to +30) influences the rate at which RNA polymerase clears the promoter, and therefore has an effect upon promoter strength. So the overall strength of a promoter cannot be predicted entirely from its —35 and –10 consensus sequences.

A "typical" promoter relies upon its -35 and -10 sequences to be recognized by RNA polymerase, but one or the other of these sequences can be absent from some (exceptional) promoters. In at least some of these cases, the promoter cannot be recognized by RNA polymerase alone, and the reaction requires ancillary proteins, which overcome the deficiency in intrinsic interaction between RNA polymerase and the promoter.

9.14 RNA polymerase binds to one face of DNA

Key Concepts

- The consensus sequences at -35 and -10 provide most of the contact points for RNA polymerase in the promoter.
- The points of contact lie on one face of the DNA.

The ability of RNA polymerase (or indeed any protein) to recognize DNA can be characterized by footprinting. A sequence of DNA bound to the protein is *partially* digested with an endonuclease to attack individual phosphodiester bonds within the nucleic acid. Under appropriate conditions, any particular phosphodiester bond is broken in some, but not in all, DNA molecules. The positions that are cleaved are recognized by using DNA labeled on one strand at one end only. The principle is the same as that involved in DNA sequencing; partial cleavage of an end-labeled molecule at a susceptible site creates a fragment of unique length.

As **Figure 9.29** shows, following the nuclease treatment, the broken DNA fragments are recovered and electrophoresed on a gel that sepa-

rates them according to length. Each fragment that retains a labeled end produces a radioactive band. The position of the band corresponds to the number of bases in the fragment. The shortest fragments move the fastest, so distance from the labeled end is counted up from the bottom of the gel.

In a free DNA, *every* susceptible bond position is broken in one or another molecule. But when the DNA is complexed with a protein, the region covered by the DNA-binding protein is protected in every molecule. So two reactions are run in parallel: a control of DNA alone; and an experimental mixture containing molecules of DNA bound to the protein. *When a bound protein blocks access of the nuclease to DNA, the bonds in the bound sequence fail to be broken in the experimental mixture.*

In the control, every bond is broken, generating a series of bands, one representing each base. There are 31 bands in the figure. In the protected fragment, bonds cannot be broken in the region bound by the protein, so bands representing fragments of the corresponding sizes are not generated. The absence of bands 9-18 in the figure identifies a protein-binding site covering the region located 9-18 bases from the labeled end of the DNA. By comparing the control and experimental lanes with a sequencing reaction that is run in parallel it becomes possible to "read off" the corresponding sequence directly, thus identifying the nucleotide sequence of the binding site.

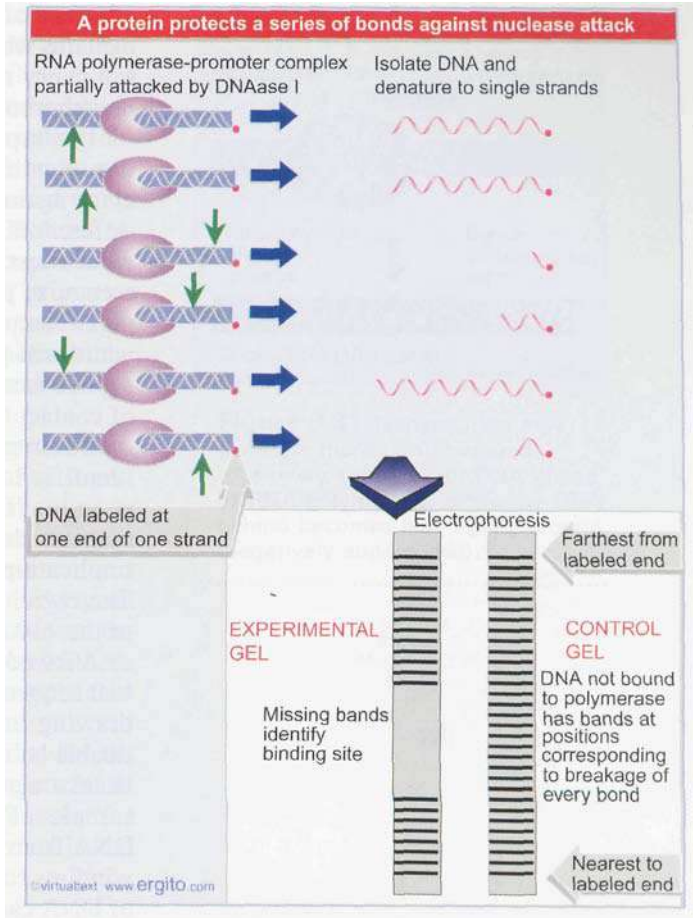


Figure 9.29 Footprinting identifies DNA-binding sites for proteins by their protection against nicking.

As described previously (see Figure 9.20), RNA polymerase initially binds the region from -50 to +20. The points at which RNA polymerase actually contacts the promoter can be identified by modifying the footprinting technique to treat RNA polymerase·promoter complexes with reagents that modify particular bases. We can perform the experiment in two ways:

- The DNA can be modified before it is bound to RNA polymerase. If the modification prevents RNA polymerase from binding, we have identified a base position where contact is essential.
- The RNA polymerase·DNA complex can be modified. Then we compare the pattern of protected bands with that of free DNA and the unmodified complex. Some bands disappear, identifying sites at which the enzyme has protected the promoter against modification. Other bands increase in intensity, identifying sites at which the DNA must be held in a conformation in which it is more exposed.

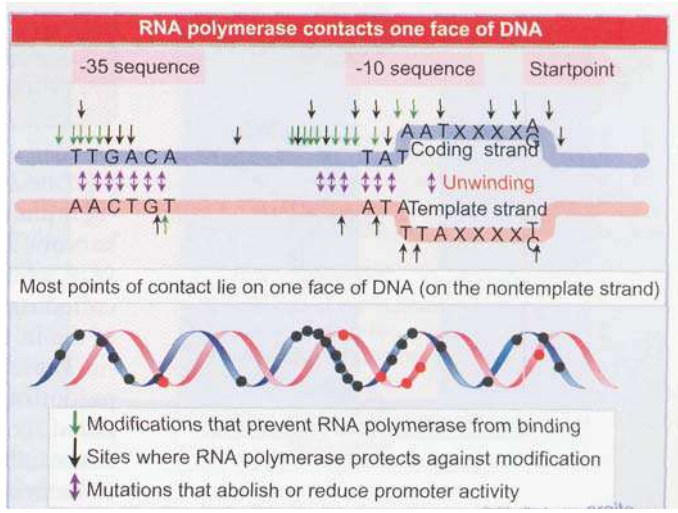


Figure 9.30 One face of the promoter contains the contact points for RNA.

These changes in sensitivity reveal the geometry of the complex, as summarized in **Figure 9.30** for a typical promoter. The regions at –35 and –10 contain most of the contact points for the enzyme. Within these regions, the same sets of positions tend both to prevent binding if previously modified, and to show increased or decreased susceptibility to modification after binding. Although the points of contact do not coincide completely with sites of mutation, they occur in the same limited region.

It is noteworthy that the same *positions* in different promoters provide the contact points, even though a different base is present. This indicates that there is a common mechanism for RNA polymerase binding, although the reaction does not depend on the

presence of particular bases at some of the points of contact. This model explains why some of the points of contact are not sites of mutation. Also, not every mutation lies in a point of contact; they may influence the neighborhood without actually being touched by the enzyme.

It is especially significant that the experiments with prior modification identify *only* sites in the same region that is protected by the enzyme against subsequent modification. These two experiments measure different things. Prior modification identifies all those sites that the enzyme must recognize in order to bind to DNA. Protection experiments recognize all those sites that actually make contact in the binary complex. The protected sites include all the recognition sites and also some additional positions, which suggests that the enzyme first recognizes a set of bases necessary for it to "touch down," and then extends its points of contact to additional bases.

The region of DNA that is unwound in the binary complex can be identified directly by chemical changes in its availability. When the strands of DNA are separated, the unpaired bases become susceptible to reagents that cannot reach them in the double helix. Such experiments implicate positions between -9 and +3 in the initial melting reaction. The region unwound during initiation therefore includes the right end of the –10 sequence and extends just past the startpoint.

Viewed in three dimensions, the points of contact upstream of the -10 sequence all lie on one face of DNA. This can be seen in the lower drawing in Figure 9.30, in which the contact points are marked on a double helix viewed from one side. Most lie on the coding strand. These bases are probably recognized in the initial formation of a closed binary complex. This would make it possible for RNA polymerase to approach DNA from one side and recognize that face of the DNA. As DNA unwinding commences, further sites that originally lay on the other face of DNA can be recognized and bound.

9.15 Supercoiling is an important feature of transcription

Key Concepts

- Negative supercoiling increases the efficiency of some promoters by assisting the melting reaction.
- Transcription generates positive supercoils ahead of the enzyme and negative supercoils behind it, and these must be removed by gyrase and topoisomerase.

The importance of strand separation in the initiation reaction is emphasized by the effects of supercoiling. Both prokaryotic and eukaryotic RNA polymerases can initiate transcription more efficiently *in vitro* when the template is supercoiled, presumably because the supercoiled structure requires less free energy for the initial melting of DNA in the initiation complex.

The efficiency of some promoters is influenced by the degree of supercoiling. The most common relationship is for transcription to be aided by negative supercoiling. We understand in principle how this assists the initiation reaction. But why should some promoters be influenced by the extent of supercoiling while others are not? One possibility is that the dependence of a promoter on supercoiling is determined by its sequence. This would predict that some promoters have sequences that are easier to melt (and are therefore less dependent on supercoiling), while others have more difficult sequences (and have a

greater need to be supercoiled). An alternative is that the location of the promoter might be important if different regions of the bacterial chromosome have different degrees of supercoiling.

Supercoiling also has a continuing involvement with transcription. As RNA polymerase transcribes DNA, unwinding and rewinding occurs, as illustrated in Figure 9.4. This requires that either the entire transcription complex rotates about the DNA or the DNA itself must rotate about its helical axis. The consequences of the rotation of DNA are illustrated in **Figure 9.31** in the *twin domain* model for transcription. As RNA polymerase pushes forward along the double helix, it generates positive supercoils (more tightly wound DNA) ahead and leaves negative supercoils (partially unwound DNA) behind. For each helical turn traversed by RNA polymerase, +1 turn is generated ahead and -1 turn behind.

Transcription therefore has a significant effect on the (local) structure of DNA. As a result, the enzymes gyrase (introduces negative supercoils) and topoisomerase I (removes negative supercoils) are required to rectify the situation in front of and behind the polymerase, respectively. Blocking the activities of gyrase and topoisomerase causes major changes in the supercoiling of DNA. For example, in yeast lacking an enzyme that relaxes negative supercoils, the density of negative supercoiling doubles in a transcribed region. A possible implication of these results is that transcription is responsible for generating a significant proportion of the supercoiling that occurs in the cell.

A similar situation occurs in replication, when DNA must be unwound at a moving replication fork, so that the individual single strands can be used as templates to synthesize daughter strands. Solutions for the topological constraints associated with such reactions are indicated later in Figure 15.20.

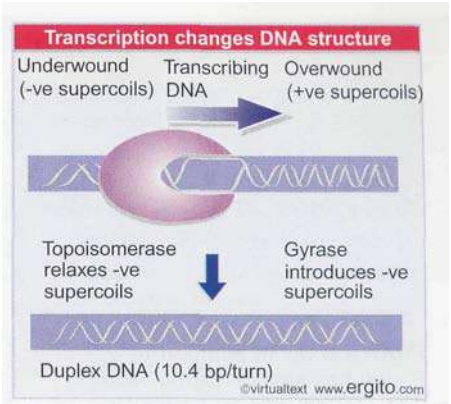


Figure 9.31 Transcription may generate more tightly wound (positively supercoiled) DNA ahead of RNA polymerase, while the DNA behind becomes less tightly wound (negatively supercoiled).

9.16 Substitution of sigma factors may control initiation

Key Concepts

- *E. coli* has several sigma factors, each of which causes RNA polymerase to initiate at a set of promoters defined by specific -35 and -10 sequences.
- cr^{70} is used for general transcription, and the other sigma factors are activated by special conditions.

The division of labors between a core enzyme that undertakes chain elongation and a sigma factor involved in site selection immediately raises the question of whether there is more than one type of sigma, each specific for a different class of promoters. **Figure 9.32** shows the principle of a system in which a substitution of the sigma factor changes the choice of promoter.

E. coli uses alternative sigma factors to respond to general environmental changes. They are listed in **Figure** 9.33. (They are named either by molecular weight of the product or for the gene.) The general factor, responsible for transcription of most genes under normal conditions, is a^{70}. The alternative sigma factors σ^{S}, cr^{32}, CT^{E}, and cr^{54} are activated in response to environmental changes; a^{28} is used for expression of flagellar genes during normal growth, but its level of expression responds to changes in the environment. All the sigma factors except σ^{54} belong to the same protein family and function in the same general manner.

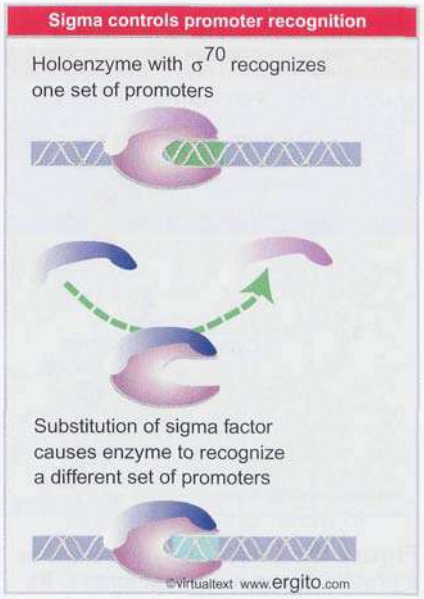


Figure 9.32 The sigma factor associated with core enzyme determines the set of promoters where transcription is initiated.

E. coli has several sigma factors

Gene	Factor	Use
rpoD	σ^{70}	general
rpoS	σ^{S}	stress
rpoH	σ^{32}	heat shock
rpoE	σ^{E}	heat shock
rpoN	σ^{54}	nitrogen
fliA	σ^{28} (σ^{F})	flagellar

©virtualtext www.ergito.com

Figure 9.33 In addition to σ^{70}, *E. coli* has several sigma factors that are induced by particular environmental conditions. (A number in the name of a factor indicates its mass.)

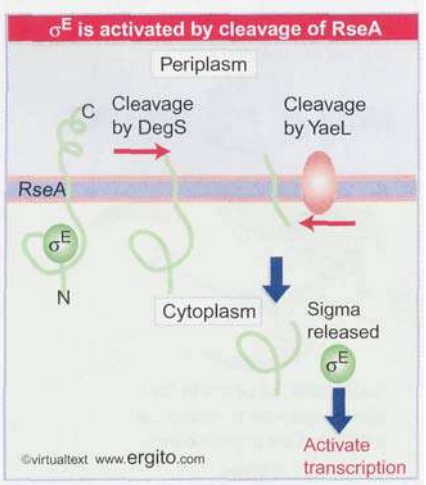


Figure 9.34 RseA is synthesized as a protein in the inner membrane. Its cytoplasmic domain binds the σ^{E} factor. RseA is cleaved sequentially in the periplasmic space and then in the cytoplasm. The cytoplasmic cleavage releases σ^{E}.

Temperature fluctuation is a common type of environmental challenge. Many organisms, both prokaryotic and eukaryotic, respond in a similar way. Upon an increase in temperature, synthesis of the proteins currently being made is turned off or down, and a new set of proteins is synthesized. The new proteins are the products of the heat shock genes. They play a role in protecting the cell against environmental stress, and are synthesized in response to other conditions as well as heat shock. Several of the heat shock proteins are chaperones. In *E. coli*, the expression of 17 heat shock proteins is triggered by changes at transcription. The gene *rpoH* is a regulator needed to switch on the heat shock response. Its product is cr , which functions as an alternative sigma factor that causes transcription of the heat shock genes.

The heat shock response is accomplished by increasing the amount of CT^{32} when the temperature increases, and decreasing its activity when the temperature change is reversed. The basic signal that induces production of a^{32} is the accumulation of unfolded (partially denatured) proteins that results from increase in temperature. The a^{32} protein is unstable, which is important in allowing its quantity to be increased or decreased rapidly. cr^{70} and a^{32} can compete for the available core enzyme, so that the set of genes transcribed during heat shock depends on the balance between them.

Changing sigma factors is a serious matter that has widespread implications for gene expression in the bacterium. It is not surprising, therefore, that the production of new sigma factors can be the target of many regulatory circuits. The factor σ^{S} is induced when bacteria make the transition from growth phase to stationary phase, and also in other stress conditions. It is controlled at two levels. Translation of the *rpoS* mRNA is increased by low temperature or high osmolarity. Proteolysis of the protein product is inhibited by carbon starvation (the typical signal of stationary phase) and by high temperature.

Another group of heat-regulated genes is controlled by the factor CT^{E}. It responds to more extreme temperature shifts than a^{32}, and is induced by accumulation of unfolded proteins in the periplasmic space or outer membrane. It is controlled by the intricate circuit summarized in Figure 9.34. σ^{E} binds to a protein (RseA) that is located in the inner membrane. As a result, it cannot activate transcription. The accumulation of unfolded proteins activates a protease (DegS) in the periplasmic space, which cleaves off the C-terminal end of the RseA protein. This cleavage activates another protein in the inner membrane (YaeL) which cleaves the N-terminal region of RseA. When this happens, the σ^{E} factor is released, and can then activate transcription. The net result is that the accumulation of unfolded proteins at the periphery of the bacterium is responsible for activating the set of genes controlled by the sigma factor.

This circuit has two interesting parallels with other regulatory circuits. The response to unfolded proteins in eukaryotic cells also uses a pathway in which an unfolded protein (within the endoplasmic reticulum) activates a membrane protein. In this case, the membrane protein is an endonuclease that cleaves an RNA, leading ultimately to a change in splicing that causes the production of a transcription factor (see *24.17 The unfolded protein response is related to tRNA splicing*). And a more direct parallel is with the first case to be discovered in which cleavage of a membrane protein activates a transcription factor. In this case, the transcription factor itself is synthesized as a membrane protein, and the level of sterols in the membrane controls the activation of proteases that release the transcription factor from the cytosolic domain of the protein.

Another sigma factor is used under conditions of nitrogen starvation. *E. coli* cells contain a small amount of σ^{54}, which is activated when ammonia is absent from the medium. In these conditions, genes are turned on to allow utilization of alternative nitrogen sources. Coun-

terparts to this sigma factor have been found in a wide range of bacteria, so it represents a response mechanism that has been conserved in evolution.

Another case of evolutionary conservation of sigma factors is presented by the factor σ^F, which is present in small amounts and causes RNA polymerase to transcribe genes involved in chemotaxis and flagellar structure. Its counterpart in *B. subtilis* is σ^D, which controls flagellar and motility genes; factors with the same promoter specificity are present in many species of bacteria.

Each sigma factor causes RNA polymerase to initiate at a particular set of promoters. By analyzing the sequences of these promoters, we can show that each set is identified by unique sequence elements. Indeed, the sequence of each type of promoter ensures that it is recognized only by RNA polymerase directed by the appropriate sigma factor. We can deduce the general rules for promoter recognition from the identification of the genes responding to the sigma factors found in *E. coli* and those involved in sporulation in *B. subtilis* (see *9.19 Sporulation is controlled by sigma factors).*

A significant feature of the promoters for each enzyme is that *they have the same size and location relative to the startpoint, and they show conserved sequences only around the usual centers of –35 and –10.* (σ^{54} is an exception for which the consensus sequences are closer together, and are positioned at -24 and –12; see next section.) As summarized in **Figure 9.35,** the consensus sequences for each set of promoters are different from one another at either or both of the -35 and -10 positions. This means that an enzyme containing a particular sigma factor can recognize only its own set of promoters, so that transcription of the different groups is mutually exclusive. Substitution of one sigma factor by another therefore turns off transcription of the old set of genes as well as turning on transcription of a new set of genes. (Some genes are expressed by RNA polymerases with different sigma factors because they have more than one promoter, each with a different set of consensus sequences.)

Sigma factors recognize promoters by their consensus sequences

Gene	Factor	-35 Sequence	Separation	-10 Sequence
rpoD	σ^{70}	TTGACA	16-18 bp	TATAAT
rpoH	σ^{32}	CCCTTGAA	13-15 bp	CCCGATNT
rpoN	σ^{54}	CTGGNA	6 bp	TTGCA
fliA	σ^{28} (σ^F)	CTAAA	15 bp	GCCGATAA
sigH	σ^H	AGGANPuPu	11-12 bp	GCTGAATCA

Figure 9.35 *E. coli* sigma factors recognize promoters with different consensus sequences.

9.17 Sigma factors directly contact DNA

Key Concepts

a^{70} changes its structure to release its DNA-binding regions when it associates *with core enzyme.*
σ^{70} binds both the -35 and -10 sequences.

The definition of a series of different consensus sequences recognized at -35 and -10 by holoenzymes containing different sigma factors (see Figure 9.35) carries the immediate implication that the sigma subunit must itself contact DNA in these regions. This suggests the general principle that there is a common type of relationship between sigma and core enzyme, in which the sigma factor is positioned in such a way as to make critical contacts with the promoter sequences in the vicinity of –35 and –10.

Direct evidence that sigma contacts the promoter directly at both the -35 and –10 consensus sequences is provided by mutations in sigma that suppress mutations in the consensus sequences. When a mutation at a particular position in the promoter prevents recognition by RNA polymerase, and a compensating mutation in sigma factor allows the polymerase to use the mutant promoter, the most likely explanation is that the relevant base pair in DNA is contacted by the amino acid that has been substituted.

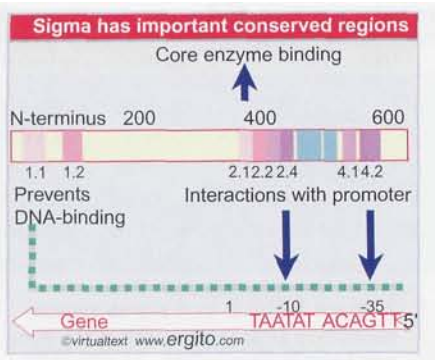


Figure 9.36 A map of the *E. coli* σ^{70} factor identifies conserved regions. Regions 2.1 and 2.2 contact core polymerase, 2.3 is required for melting, and 2.4 and 4.2 contact the –10 and –35 promoter elements.

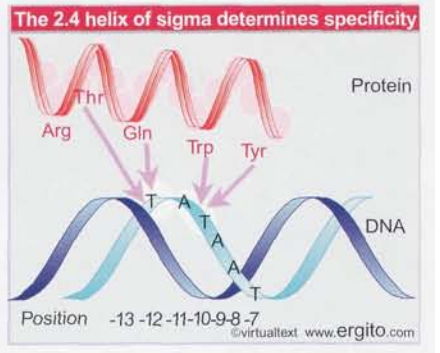


Figure 9.37 Amino acids in the 2.4 α-helix of σ^{70} contact specific bases in the coding strand of the –10 promoter sequence.

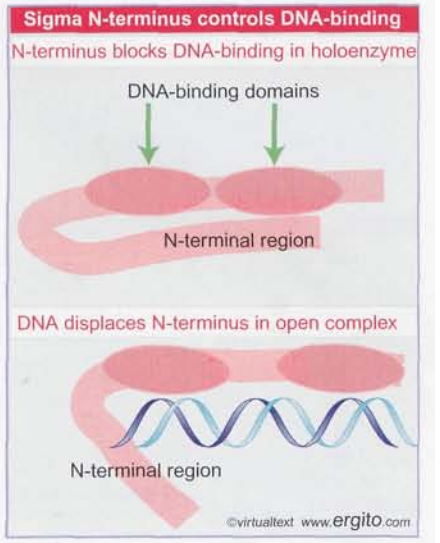


Figure 9.38 The N-terminus of sigma blocks the DNA-binding regions from binding to DNA. When an open complex forms, the N-terminus swings 20 Å away, and the two DNA-binding regions separate by 15 Å.

Comparisons of the sequences of several bacterial sigma factors identify regions that have been conserved. Their locations in *E. coli* σ^{70} are summarized in **Figure 9.36.** The crystal structure of a sigma factor fragment from the bacterium *Thermus aquaticus* shows that these regions fold into three independent domains in the protein: domain σ_2 contains 1.2-2.4, σ_3 contains 3.0-1.3, and σ_4 contains 4.1-4.2.

Figure 9.36 shows that two short parts of regions 2 and 4 (named 2.4 and 4.2) are involved in contacting bases in the –10 and —35 elements, respectively. Both of these regions form short stretches of α-helix in the protein. Experiments with heteroduplexes show that a^{70} makes contacts with bases principally on the coding strand, and it continues to hold these contacts after the DNA has been unwound in this region. This suggests that sigma factor could be important in the melting reaction.

The use of α-helical motifs in proteins to recognize duplex DNA sequences is common (see *12.12 Repressor uses a helix-turn-helix motif to bind DNA).* Amino acids separated by 3-4 positions lie on the same face of an α-helix and are therefore in a position to contact adjacent base pairs. **Figure 9.37** shows that amino acids lying along one face of the 2.4 region α-helix contact the bases at positions -12 to -10 of the –10 promoter sequence.

Region 2.3 resembles proteins that bind single-stranded nucleic acids, and is involved in the melting reaction. Regions 2.1 and 2.2 (which is the most highly conserved part of sigma) are involved in the interaction with core enzyme. It is assumed that all sigma factors bind the same regions of the core polymerase (ensuring that the reactions are competitive).

The N-terminal region of cr^{70} has important regulatory functions. If it is removed, the shortened protein becomes able to bind specifically to promoter sequences. This suggests that the N-terminal region behaves as an autoinhibition domain. It occludes the DNA-binding domains when a^{70} is free. Association with core enzyme changes the conformation of sigma so that the inhibition is released, and the DNA-binding domains can contact DNA.

Figure 9.38 schematizes the conformational change in sigma at open complex formation. When sigma binds to the core polymerase, the N-terminal domain swings ~20 A away from the DNA-binding domains, and the DNA-binding domains separate from one another by ~15 A, presumably to acquire a more elongated conformation appropriate for contacting DNA. Mutations in either the -10 or -35 sequences prevent an (N-terminal-deleted) a^{70} from binding to DNA, which suggests that cr^{70} contacts both sequences simultaneously. This implies that the sigma factor must have a rather elongated structure, extending over the ~68 A of two turns of DNA.

In the free holoenzyme, the N-terminal domain is located in the active site of the core enzyme components, essentially mimicking the location that DNA will occupy when a transcription complex is formed. When the holoenzyme forms an open complex on DNA, the N-terminal sigma domain is displaced from the active site. Its relationship with the rest of the protein is therefore very flexible, and changes when sigma binds to core enzyme, and again when the holoenzyme binds to DNA.

Comparisons of the crystal structures of the core enzyme and holoenzyme show that sigma factor lies largely on the surface of the core enzyme. **Figure 9.39** shows that it has an elongated structure that extends past the DNA-binding site. This places it in a position to contact DNA during the initial binding. The DNA helix has to move some 16 A from the initial position in order to enter the active site. **Figure 9.40** illustrates this movement, looking in cross-section down the helical axis of the DNA.

An interesting difference in behavior is found with the a^{54} factor. This causes RNA polymerase to recognize promoters that have a distinct

consensus sequence, with a conserved element at -12 and another close by at -24 (given in the "-35" column of Figure 9.33). So the geometry of the polymerase-promoter complex is different under the direction of this sigma factor. Another difference in the mechanism of regulation is that high level transcription directed by σ^{54} requires other activators to bind to sites that are quite distant from the promoter. This contrasts with the other types of bacterial promoter, where the regulator sites are always in close proximity to the promoter. The behavior of CT^{54} itself is different from other sigma factors, most notably in its ability to bind to DNA independently of core polymerase. In these regards, a^{54} is more like the eukaryotic regulators we discuss in *21 Promoters and Enhancers* than the typical prokaryotic regulators discussed in *10 The operon*.

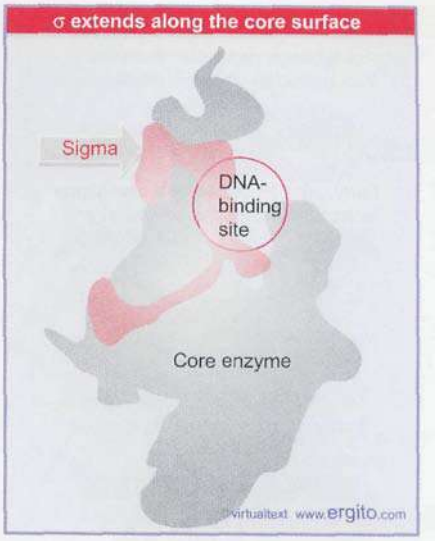


Figure 9.39 The sigma factor has an elongated structure that extends along the surface of the core subunits when the holoenzyme is formed.

9.18 Sigma factors may be organized into cascades

Key Concepts

- A cascade of sigma factors is created when one sigma factor is required to transcribe the gene coding for the next sigma factor.
- The early genes of phage SPO1 are transcribed by host RNA polymerase.
- One of the early genes codes for a sigma factor that causes RNA polymerase to transcribe the middle genes.
- Two of the middle genes code for subunits of a sigma factor that causes RNA polymerase to transcribe the late genes.

Sigma factors are used extensively to control initiation of transcription in the bacterium *B. subtilis,* where ~10 different a factors are known. Some are present in vegetative cells; other are produced only in the special circumstances of phage infection or the change from vegetative growth to sporulation.

The major RNA polymerase found in *B. subtilis* cells engaged in normal vegetative growth has the same structure as that of *E. coli*, $\alpha_2\beta\beta'\sigma$. Its sigma factor (described as σ^{43} or σ^{A}) recognizes promoters with the same consensus sequences used by the *E. coli* enzyme under direction from a^{70}. Several variants of the RNA polymerase that contain other sigma factors are found in much smaller amounts. The variant enzymes recognize different promoters on the basis of consensus sequences at –35 and –10.

Transitions from expression of one set of genes to expression of another set are a common feature of bacteriophage infection. In all but the very simplest cases, the development of the phage involves shifts in the pattern of transcription during the infective cycle. These shifts may be accomplished by the synthesis of a phage-encoded RNA polymerase or by the efforts of phage-encoded ancillary factors that control the bacterial RNA polymerase. A well characterized example of control via the production of new sigma factors occurs during infection of *B. subtilis* by phage SPO1.

The infective cycle of SPO1 passes through three stages of gene expression. Immediately on infection, the early genes of the phage are transcribed. After 4-5 minutes, the early genes cease transcription and the middle genes are transcribed. Then at 8-12 minutes, middle gene transcription is replaced by transcription of late genes.

The early genes are transcribed by the holoenzyme of the host bacterium. They are essentially indistinguishable from host genes whose promoters have the intrinsic ability to be recognized by the RNA polymerase $\alpha_2\beta\beta'\sigma^{43}$.

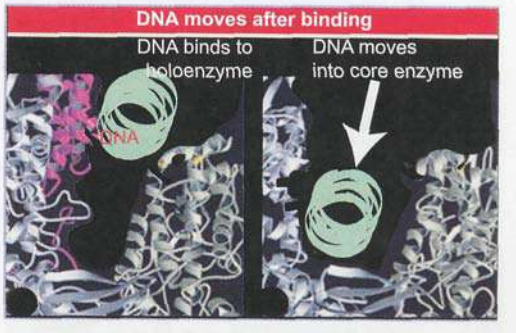


Figure 9.40 DNA initially contacts the sigma sigma factor (pink) and the core enzyme (gray). It moves deeper into the core enzyme to make contacts at the –10 sequence. When sigma is released, the width of the passage containing DNA increases.

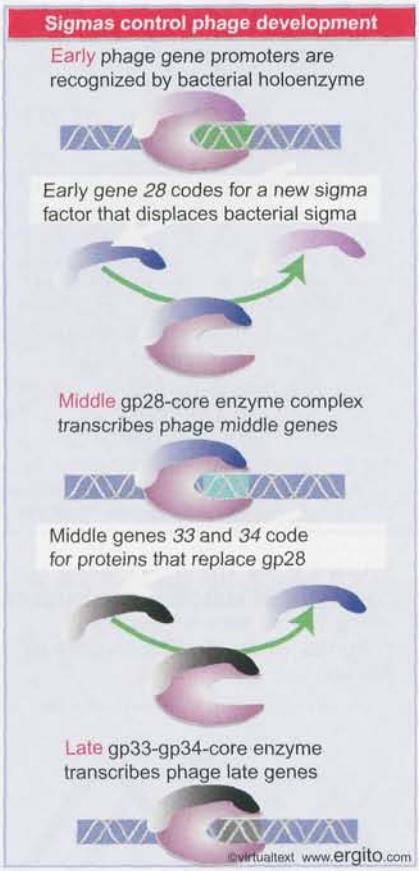


Figure 9.41 Transcription of phage SPO1 genes is controlled by two successive substitutions of the sigma factor that change the initiation specificity.

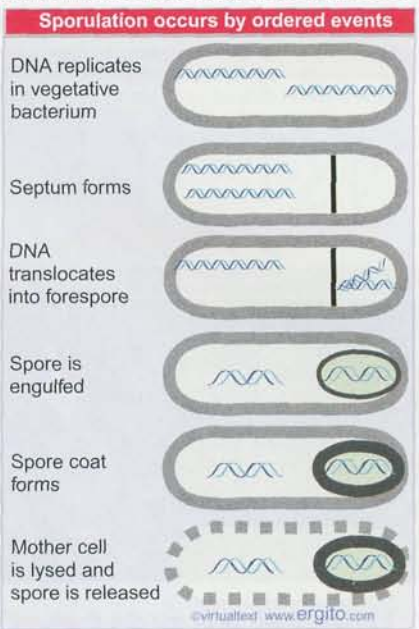


Figure 9.42 Sporulation involves the differentiation of a vegetative bacterium into a mother cell that is lysed and a spore that is released.

Expression of phage genes is required for the transitions to middle and late gene transcription. Three regulatory genes, named *28*, *33*, and *34*, control the course of transcription. Their functions are summarized in **Figure 9.41.** The pattern of regulation creates a cascade, in which the host enzyme transcribes an early gene whose product is needed to transcribe the middle genes; and then two of the middle genes code for products that are needed to transcribe the late genes.

Mutants in the early gene *28* cannot transcribe the middle genes. The product of gene *28* (called gp28) is a protein of 26 kD that replaces the host sigma factor on the core enzyme. *This substitution is the sole event required to make the transition from early to middle gene expression.* It creates a holoenzyme that can no longer transcribe the host genes, but instead specifically transcribes the middle genes. We do not know how gp28 displaces σ^{43}, or what happens to the host sigma polypeptide.

Two of the middle genes are involved in the next transition. Mutations in either gene *33* or *34* prevent transcription of the late genes. The products of these genes form a dimer that replaces gp28 on the core polymerase. Again, we do not know how gp33 and gp34 exclude gp28 (or any residual host CT^{43}), *but once they have bound to the core enzyme, it is able to initiate transcription only at the promoters for late genes.*

The successive replacements of sigma factor have dual consequences. Each time the subunit is changed, the RNA polymerase becomes able to recognize a new class of genes, *and* it no longer recognizes the previous class. These switches therefore constitute global changes in the activity of RNA polymerase. Probably all or virtually all of the core enzyme becomes associated with the sigma factor of the moment; and the change is irreversible.

9.19 Sporulation is controlled by sigma factors

Key Concepts

- Sporulation divides a bacterium into a mother cell that is lysed and a spore that is released.
- Each compartment advances to the next stage of development by synthesizing a new sigma factor that displaces the previous sigma factor.
- Communication between the two compartments coordinates the timing of sigma factor substitutions.

Perhaps the most extensive example of switches in sigma factors is provided by sporulation, an alternative lifestyle available to some bacteria. At the end of the vegetative phase in a bacterial culture, logarithmic growth ceases because nutrients in the medium become depleted. This triggers sporulation, as illustrated in **Figure 9.42.** DNA is replicated, a genome is segregated at one end of the cell, and eventually it is surrounded by the tough spore coat. When the septum forms, it generates two independent compartments, the mother cell and the forespore. At the start of the process, one chromosome is attached to each pole of the cell. The growing septum traps part of one chromosome in the forespore, and then a translocase (SpoIIIE) pumps the rest of the chromosome into the forespore.

Sporulation takes ~8 hours. It can be viewed as a primitive sort of differentiation, in which a parent cell (the vegetative bacterium) gives rise to two different daughter cells with distinct fates: the mother cell is eventually lysed, while the spore that is released has an entirely different structure from the original bacterium.

By Book_Crazy [IND]

Sporulation involves a drastic change in the biosynthetic activities of the bacterium, in which many genes are involved. The basic level of control lies at transcription. Some of the genes that functioned in the vegetative phase are turned off during sporulation, but most continue to be expressed. In addition, the genes specific for sporulation are expressed only during this period. At the end of sporulation, ~40% of the bacterial mRNA is sporulation-specific.

New forms of the RNA polymerase become active in sporulating cells; they contain the same core enzyme as vegetative cells, but have different proteins in place of the vegetative CT^{43}. The changes in transcriptional specificity are summarized in **Figure 9.43.** The principle is that in each compartment the existing sigma factor is successively displaced by a new factor that causes transcription of a different set of genes. Communication between the compartments occurs in order to coordinate the timing of the changes in the forespore and mother cell.

The sporulation cascade is initiated when environmental conditions trigger a phosphorelay, in which a phosphate group is passed along a series of proteins until it reaches SpoOA. (Several gene products are involved in this process, whose complexity may reflect the need to avoid mistakes in triggering sporulation unnecessarily.) SpoOA is a transcriptional regulator whose activity is affected by phosphorylation. In the phosphorylated form, it activates transcription of two operons, each of which is transcribed by a different form of the host RNA polymerase. Under the direction of phosphorylated SpoOA, host enzyme utilizing the general CT^{43} transcribes the gene coding for the factor σ^F; and host enzyme under the direction of a minor factor, CT^H, transcribes the gene coding for the factor pro-σ^E. Both of these new sigma factors are produced before septum formation, but become active later.

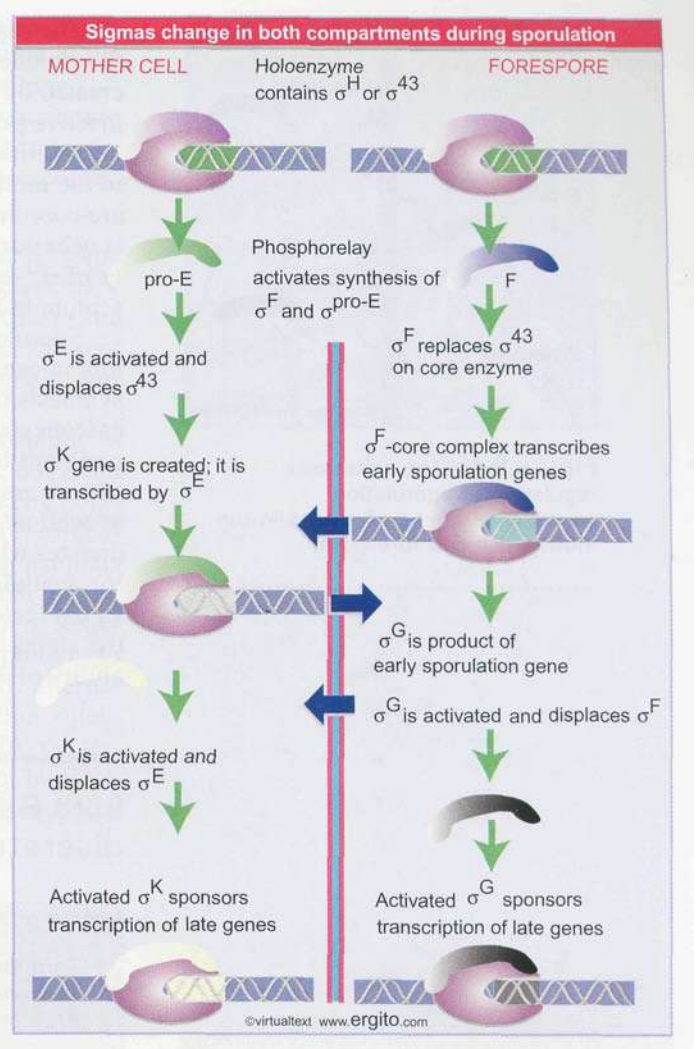


Figure 9.43 Sporulation involves successive changes in the sigma factors that control the initiation specificity of RNA polymerase. The cascades in the forespore (left) and the mother cell (right) are related by signals passed across the septum (indicated by horizontal arrows).

σ^F is the first factor to become active in the forespore compartment. It is inhibited by an anti-sigma factor that binds to it; in the forespore, an anti-anti-sigma factor removes the inhibitor. This reaction is controlled by a series of phosphorylation/dephosphorylation events. The initial determinant is a phosphatase (SpoIIE) that is an integral membrane protein, and which accumulates at the pole, with the result that its phosphatase domain becomes more concentrated in the forespore. It dephosphorylates, and thereby activates, SpoIIAA, which in turn displaces the anti-sigma factor SpoIIAB from the complex of SpoIIAB-σ^F. Release of σ^F activates it.

Activation of σ^F is the start of sporulation. Under the direction of σ^F, RNA polymerase transcribes the first set of sporulation genes instead of the vegetative genes it was previously transcribing. The replacement reaction probably affects only part of the RNA polymerase population, since σ^F is produced only in small amounts. Some vegetative enzyme remains present during sporulation. The displaced CT^{43} is not destroyed, but can be recovered from extracts of sporulating cells.

Two regulatory events follow from the activity of σ^F, as detailed in Figure **9.44.** In the forespore itself, another factor, σ^G, is the product of one of the early sporulation genes. σ^G is the factor that causes RNA polymerase to transcribe the late sporulation genes in the forespore. Another early sporulation gene product is responsible for communicating with the mother cell compartment. σ^F activates SpoIIR, which is secreted from the forespore. It then activates the membrane-bound protein SpoIIGA to cleave the inactive precursor pro-σ^E into the active factor σ^E in the mother cell. (Any σ^E that is produced in the forespore is degraded by forespore-specific functions.)

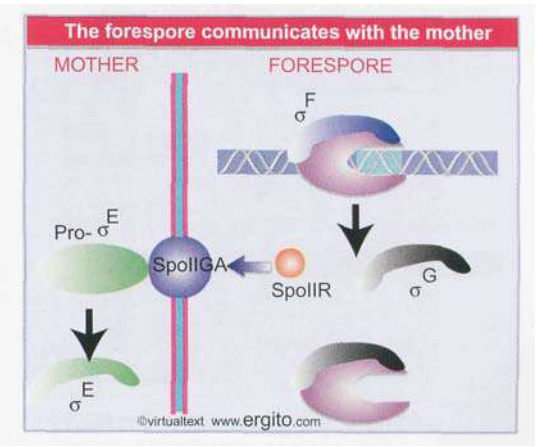


Figure 9.44 σ^F triggers synthesis of the next sigma factor in the forespore (σ^G) and turns on SpoIIR which causes SpoIIGA to cleave pro-σ^E.

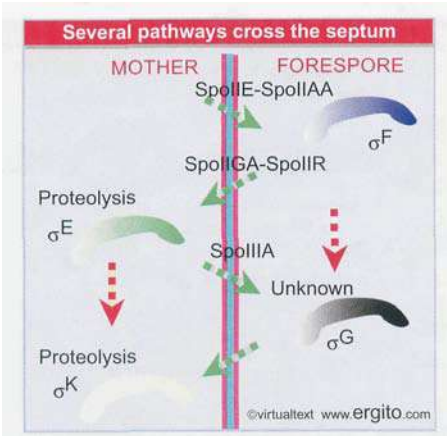


Figure 9.45 The crisscross regulation of sporulation coordinates timing of events in the mother cell and forespore.

The cascade continues when σ^E in turn is replaced by σ^K. (Actually the production of σ^K is quite complex, because first its gene must be created by a recombination event!) This factor also is synthesized as an inactive precursor (pro-σ^K) that is activated by a protease. Once σ^K has been activated, it displaces CT^E and causes transcription of the late genes in the mother cell. The timing of these events in the two compartments are coordinated by further signals. The activity of σ^E in the mother cell is necessary for activation of σ^G in the forespore; and in turn the activity of σ^G is required to generate a signal that is transmitted across the septum to activate σ^K.

Sporulation is thus controlled by two cascades, in which sigma factors in each compartment are successively activated, each directing the synthesis of a particular set of genes. **Figure 9.45** outlines how the two cascades are connected by the transmission of signals from one compartment to the other. As new sigma factors become active, old sigma factors are displaced, so that transitions in sigma factors turn genes off as well as on. The incorporation of each factor into RNA polymerase dictates when its set of target genes is expressed; and the amount of factor available influences the level of gene expression. More than one sigma factor may be active at any time, and the specificities of some of the sigma factors overlap. We do not know what is responsible for the ability of each sigma factor to replace its predecessor.

9.20 Bacterial RNA polymerase terminates at discrete sites

Key Concepts

- Termination may require both recognition of the terminator sequence in DNA and the formation of a hairpin structure in the RNA product.

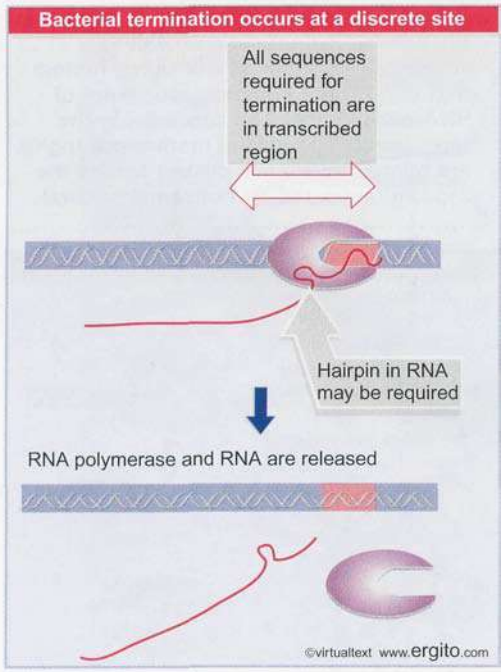


Figure 9.46 The DNA sequences required for termination are located prior to the terminator sequence. Formation of a hairpin in the RNA may be necessary.

Once RNA polymerase has started transcription, the enzyme moves along the template, synthesizing RNA, until it meets a terminator (?) sequence. At this point, the enzyme stops adding nucleotides to the growing RNA chain, releases the completed product, and dissociates from the DNA template. Termination requires that all hydrogen bonds holding the RNA-DNA hybrid together must be broken, after which the DNA duplex reforms.

It is difficult to define the termination point of an RNA molecule that has been synthesized in the living cell. It is always possible that the 3′ end of the molecule has been generated by *cleavage* of the primary transcript, and therefore does not represent the actual site at which RNA polymerase terminated.

The best identification of termination sites is provided by systems in which RNA polymerase terminates *in vitro*. Because the ability of the enzyme to terminate is strongly influenced by parameters such as the ionic strength, its termination at a particular point *in vitro* does not prove that this same point is a natural terminator. But we can identify authentic 3′ ends when the same end is generated *in vitro* and *in vivo*.

Figure 9.46 summarizes the two types of feature found in bacterial terminators.

- Terminators in bacteria and their phages have been identified as sequences that are needed for the termination reaction *(in vitro* or *in vivo)*. The sequences at prokaryotic terminators show no similarities beyond the point at which the last base is added to the RNA. The

responsibility for termination lies with the *sequences already transcribed* by RNA polymerase. So termination relies on scrutiny of the template or product that the polymerase is currently transcribing.

- Many terminators require a hairpin to form in the secondary structure of the RNA being transcribed. *This indicates that termination depends on the RNA product and is not determined simply by scrutiny of the DNA sequence during transcription.*

Terminators vary widely in their efficiencies of termination. At some terminators, the termination event can be *prevented* by specific ancillary factors that interact with RNA polymerase. Antitermination causes the enzyme to continue transcription past the terminator sequence, an event called readthrough (the same term used to describe a ribosome's suppression of termination codons).

In approaching the termination event, we must regard it not simply as a mechanism for generating the 3' end of the RNA molecule, but as an opportunity to control gene expression. So the stages when RNA polymerase associates with DNA (initiation) or dissociates from it (termination) both are subject to specific control. There are interesting parallels between the systems employed in initiation and termination. Both require breaking of hydrogen bonds (initial melting of DNA at initiation, RNA-DNA dissociation at termination); and both require additional proteins to interact with the core enzyme. In fact, they are accomplished by alternative forms of the polymerase. However, whereas initiation relies solely upon the interaction between RNA polymerase and duplex DNA, the termination event involves recognition of signals in the transcript by RNA polymerase or by ancillary factors as well as the recognition of sequences in DNA.

9.21 There are two types of terminators in *E. coli*

Key Concepts

Intrinsic terminators consist of a G·C-rich hairpin in the RNA product followed by a U-rich region in which termination occurs.

Terminators are distinguished in *E. coli* according to whether RNA polymerase requires any additional factors to terminate *in vitro:*

- A core enzyme can terminate *in vitroat* certain sites in the absence of any other factor. These sites are called intrinsic terminators.
- Rho-dependent terminators are defined by the need for addition of rho factor (p) *in vitro;* and mutations show that the factor is involved in termination *in vivo.*

Intrinsic terminators have the two structural features evident in Figure 9.47: a hairpin in the secondary structure; and a region that is rich in U residues at the very end of the unit. Both features are needed for termination. The hairpin usually contains a G-C-rich region near the base of the stem. The typical distance between the hairpin and the U-rich region is 7-9 bases. There are ~l 100 sequences in the *E. coli* genome that fit these criteria, suggesting that about half of the genes have intrinsic terminators.

Point mutations that prevent termination occur within the stem region of the hairpin. What is the effect of a hairpin on transcription? Probably all hairpins that form in the RNA product cause the polymerase to slow (and perhaps to pause) in RNA synthesis.

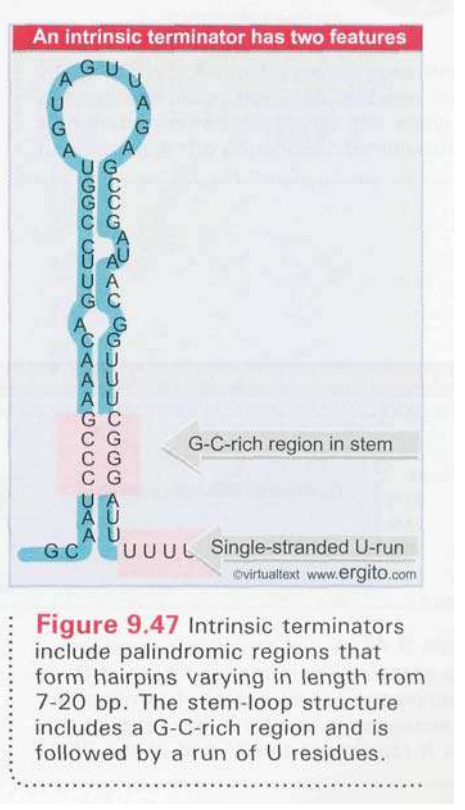


Figure 9.47 Intrinsic terminators include palindromic regions that form hairpins varying in length from 7-20 bp. The stem-loop structure includes a G-C-rich region and is followed by a run of U residues.

Pausing creates an opportunity for termination to occur. Pausing occurs at sites that resemble terminators but have an increased separation (typically 10-11 bases) between the hairpin and the U-run. But if the pause site does not correspond to a terminator, usually the enzyme moves on again to continue transcription. The length of the pause varies, but at a typical terminator lasts ~60 seconds.

A downstream U-rich region destabilizes the RNA-DNA hybrid when RNA polymerase pauses at the hairpin. The rU·dA RNA-DNA hybrid has an unusually weak base-paired structure; it requires the least energy of any RNA-DNA hybrid to break the association between the two strands. When the polymerase pauses, the RNA-DNA hybrid unravels from the weakly bonded rU·dA terminal region. Often the actual termination event takes place at any one of several positions toward or at the end of the U-rich region, as though the enzyme "stutters" during termination. The U-rich region in RNA corresponds to an A-T-rich region in DNA, so we see that A-T-rich regions are important in intrinsic termination as well as initiation.

Both the sequence of the hairpin and the length of the U-run influence the efficiency of termination. However, termination efficiency *in vitro* varies from 2-90%, and does hot correlate in any simple way with the constitution of the hairpin or the number of U residues in the U-rich region. The hairpin and U-region are therefore necessary, but not sufficient, and additional parameters influence the interaction with RNA polymerase. In particular, the sequences both upstream and downstream of the intrinsic terminator influence its efficiency.

Less is known about the signals and ancillary factors involved in termination for eukaryotic polymerases. Each class of polymerase uses a different mechanism (see *24 RNA splicing and processing).*

9.22 How does rho factor work?

Key Concepts

- The rho factor is a terminator protein that binds to nascent RNA and; translocates to a sequence that is rich in C and poor in G residues preceding the actual termination site in the RNA.

Rho factor is an essential protein in *E. coli.* It functions solely at the stage of termination. It is a ~275 kD hexamer of identical subunits. The subunit has an RNA-binding domain and an ATP hydrolysis domain. Rho is a member of the family of hexameric ATP-dependent helicases that function by passing nucleic acid through the hole in the middle of the hexamer formed from the RNA-binding domains of the subunits. Rho functions as an ancillary factor for RNA polymerase; typically its maximum activity *in vitro* is displayed when it is present at ~ 10% of the concentration of the RNA polymerase.

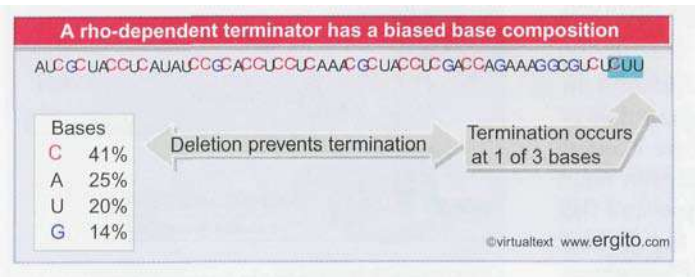


Figure 9.48 A rho-dependent terminator has a sequence rich in C and poor in G preceding the actual site(s) of termination. The sequence is shown in the form of the RNA. It represents the 3' end of the RNA.

Rho-dependent terminators account for about half of *E. coli* terminators. They were discovered in phage genomes, where they have been most fully characterized. The sequences required for rho-dependent termination are 50-90 bases long and lie upstream of the termination site. Their common feature is that the RNA is rich in C residues and poor in G residues. An example is given in **Figure 9.48;** C is by far the most common base (41%) and G is the least common base (14%). As a general rule the efficiency of a rho-dependent terminator increases with the length of the C-rich/G-poor region.

By Book_Crazy [IND]

Does the rho factor act via recognizing DNA, RNA, or RNA polymerase? The "hot pursuit" model for rho action is shown in Figure 9.49. An individual rho factor acts processively on a single RNA substrate. Rho's key function is its helicase activity, for which energy is provided by an RNA-dependent ATP hydrolysis. The initial binding site for rho is an extended (~70 nucleotide) single-stranded region in the RNA upstream of the terminator. Rho binds to RNA and then uses its ATPase activity to provide the energy to translocate along the RNA until it reaches the RNA-DNA helical region, where it unwinds the duplex structure.

How does rho catch up with RNA polymerase? Probably it simply moves along the transcript faster than RNA polymerase moves along the DNA. The enzyme pauses when it reaches a terminator, and termination occurs if rho catches it there. Pausing is therefore important in rho-dependent termination, just as in intrinsic termination, because it gives time for the other necessary events to occur.

These abilities suggest that rho can directly gain access to the stretch of RNA-DNA hybrid in the transcription bubble and cause it to unwind. We do not know whether this action is sufficient to release the transcript or whether rho also interacts with RNA polymerase to help release RNA.

The idea that rho moves along RNA leads to an important prediction about the relationship between transcription and translation. Rho must first have access to a binding sequence on RNA, and then must be able to move along the RNA. Either or both of these conditions may be prevented if ribosomes are translating an RNA. So the ability of the rho factor to reach RNA polymerase at a terminator depends on what is happening in translation.

This model explains a puzzling phenomenon. In some cases, a nonsense mutation in one gene of a transcription unit prevents the expression of subsequent genes in the unit. This effect is called polarity. A common cause is the absence of the mRNA corresponding to the subsequent (distal) parts of the unit.

Suppose that there are rho-dependent terminators *within* the transcription unit, that is, before the terminator that *usually* is used. The consequences are illustrated in Figure 9.50. Normally these earlier terminators are not used, because the ribosomes prevent rho from reaching RNA polymerase. But a nonsense mutation releases the ribosomes, so that rho is free to attach to and/or move along the mRNA, enabling it to act on RNA polymerase at the terminator. As a result, the enzyme is released, and the distal regions of the transcription unit are never expressed. (Why should there be internal terminators? Perhaps they are simply sequences that by coincidence mimic the usual rho-dependent terminator.) Some stable RNAs that have extensive secondary structure are preserved from polar effects, presumably because the structure impedes rho attachment or movement.

rho mutations show wide variations in their influence on termination. The basic nature of the effect is a failure to terminate. But the magnitude of the failure, as seen in the percent of readthrough *in vivo*, depends on the particular target locus. Similarly, the need for rho factor *in vitro* is variable. Some (rho-dependent) terminators require relatively high concentrations of rho, while others function just as well at lower levels. This suggests that different terminators require different levels of rho factor for termination, and therefore respond differently to the residual levels of rho factor in the mutants (*rho* mutants are usually leaky).

Some *rho* mutations can be suppressed by mutations in other genes. This approach provides an excellent way to identify proteins that interact with rho. The β subunit of RNA polymerase is implicated by two types of mutation. First, mutations in the *rpoB* gene can reduce

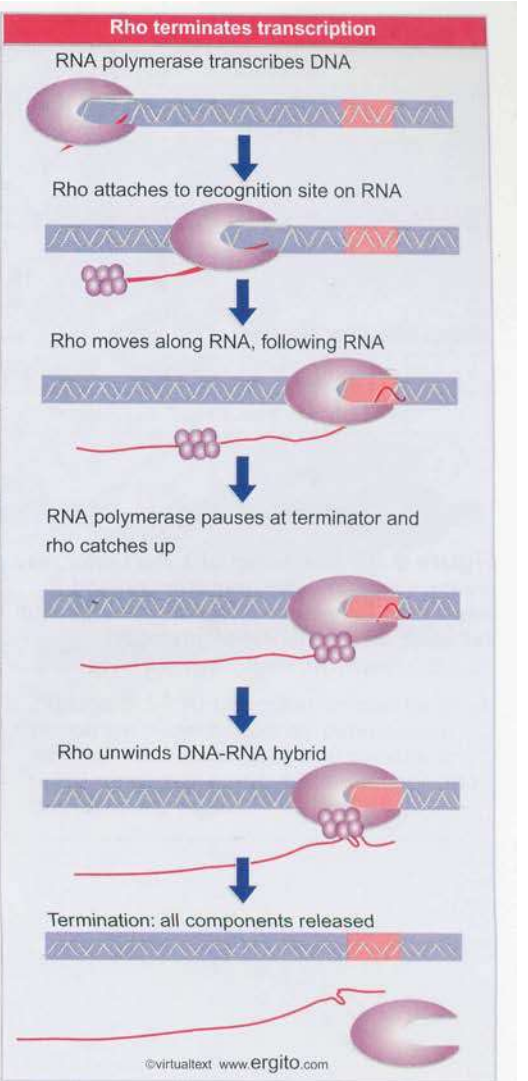


Figure 9.49 A rho factor pursues RNA polymerase along the RNA and can cause termination when it catches the enzyme pausing at a rho-dependent terminator.

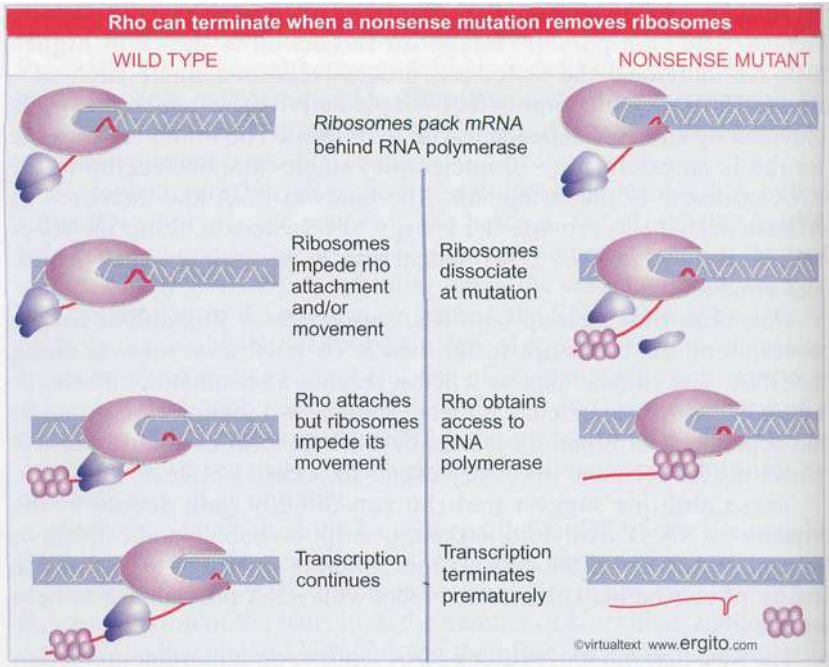


Figure 9.50 The action of a rho factor may create a link between transcription and translation when a rho-dependent terminator lies soon after a nonsense mutation.

termination at a rho-dependent site. Second, mutations in *rpoB* can restore the ability to terminate transcription at rho-dependent sites in *rho* mutant bacteria. However, we do not know what function the interaction plays.

9.23 Antitermination is a regulatory event

Key Concepts

- Termination is prevented when antitermination proteins act on RNA polymerase to cause it to readthrough a specific terminator or terminators.
- Phage lambda has two antitermination proteins, pN and pQ, that act on different transcription units.

Antitermination is used as a control mechanism in both phage regulatory circuits and bacterial operons. **Figure 9.51** shows that antitermination controls the ability of the enzyme to read past a terminator into genes lying beyond. In the example shown in the figure, the default pathway is for RNA polymerase to terminate at the end of region 1. But antitermination allows it to continue transcription through region 2. Because the promoter does not change, both situations produce an RNA with the same 5' sequences; the difference is that after antitermination the RNA is extended to include new sequences at the 3' end.

Antitermination was discovered in bacteriophage infections. A common feature in the control of phage infection is that very few of the phage genes (the "early" genes) can be transcribed by the bacterial host RNA polymerase. Among these genes, however, are regulator(s) whose product(s) allow the next set of phage genes to be expressed (see *12.4 Two types of regulatory event control the lytic cascade).* One of these types of regulator is an antitermination protein. **Figure 9.52** shows that it enables RNA polymerase to read through a terminator, extending the RNA transcript. In the absence of the antitermination protein, RNA polymerase terminates at the terminator (top panel). When the antiter-

By Book_Crazy [IND]

initiation protein is present, it continues past the terminator (middle panel).

The best characterized example of antitermination is provided by phage lambda, with which the phenomenon was discovered. It is used at two stages of phage expression. The antitermination protein produced at each stage is specific for the particular transcription units that are expressed at that stage, as summarized in the bottom panel of Figure 9.52.

The host RNA polymerase initially transcribes two genes, which are called the immediate early genes. The transition to the next stage of expression is controlled by preventing termination at the ends of the immediate early genes, with the result that the delayed early genes are expressed. (We discuss the overall regulation of lambda development in *12 Phage strategies.*)

The regulator gene that controls the switch from immediate early to delayed early expression is identified by mutations in lambda gene *N* that can transcribe *only* the immediate early genes; they proceed no further into the infective cycle. There are two transcription units of immediate early genes (transcribed from the promoters P_L and P_R). Transcription by *E. coli* RNA polymerase itself stops at the terminators at the ends of these transcription units (t_{L1} and t_{R1}, respectively.) Both terminators depend on rho; in fact, these were the terminators with which rho was originally identified. The situation is changed by expression of the *N* gene. The product pN is an antitermination protein that acts on both of the immediate early transcription units, and allows RNA polymerase to read through the terminators into the delayed early genes beyond them.

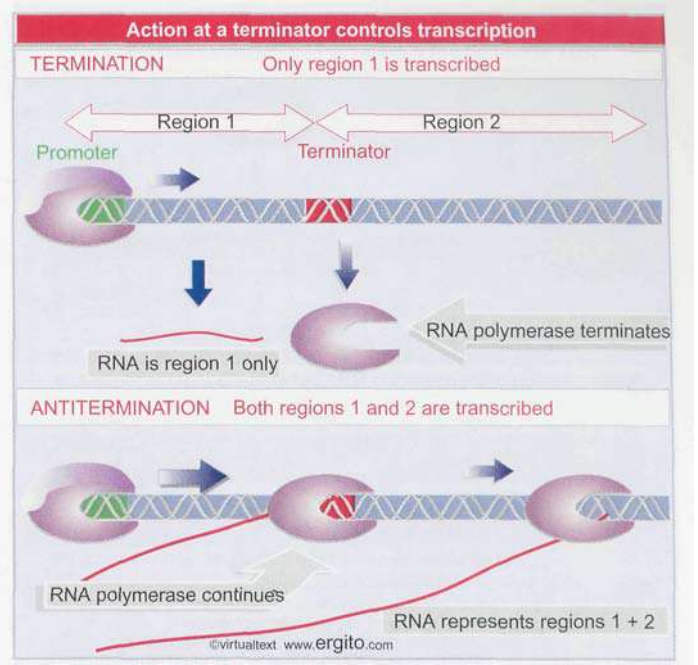


Figure 9.51 Antitermination can be used to control transcription by determining whether RNA polymerase terminates or reads through a particular terminator into the following region.

Like other phages, still another control is needed to express the late genes that code for the components of the phage particle. This switch is regulated by gene *Q*, itself one of the delayed early genes. Its product, pQ, is another antitermination protein, one that specifically allows RNA polymerase initiating at another site, the late promoter $P_{R'}$, to read through a terminator that lies between it and the late genes.

The different specificities of pN and pQ establish an important general principle: *RNA polymerase interacts with transcription units in such a way that an ancillary factor can sponsor antitermination specifically for some transcripts.* Termination can be controlled with the same sort of precision as initiation.

9.24 Antitermination requires sites that are independent of the terminators

Key Concepts

- The site where an antiterminator protein acts is upstream of the terminator site in the transcription unit.
- The location of the antiterminator site varies in different cases, and can be in the promoter or within the transcription unit.

What sites are involved in controlling the specificity of antitermination? The antitermination activity of pN is highly specific, but *the antitermination event is not determined by the terminators* t_{L1} *and* t_{R1}; *the recognition site needed for antitermination lies upstream in the transcription unit, that is, at a different place from the terminator*

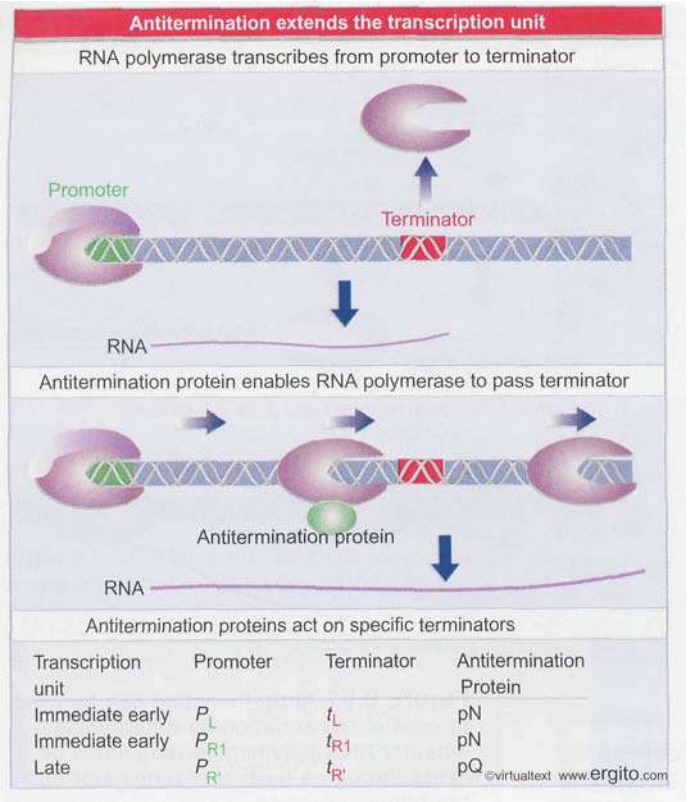


Figure 9.52 An antitermination protein can act on RNA polymerase to enable it to readthrough a specific terminator.

site at which the action eventually is accomplished. This conclusion establishes a general principle. When we know the site on DNA at which some protein exercises its effect, we cannot assume that this coincides with the DNA sequence that it initially recognizes. They may be separate. **Figure 9.53** shows the locations of the sites required for antitermination in phage lambda.

The recognition sites required for pN action are called *nut* (for *N utilization*). The sites responsible for determining leftward and rightward antitermination are described as *nutL* and *nutR*, respectively. Mapping of *nut* mutations locates *nutL* between the startpoint of P_L and the beginning of the *N* coding region. By contrast, *nutR* lies between the end of the *cro* gene and t_{L1}. This means that the two *nut* sites lie in different positions relative to the organization of their transcription units. Whereas *nutL* is near the promoter, *nutR* is near to the terminator. (*qut* is different yet again and lies within the promoter.)

How does antitermination occur? When pN recognizes the *nut* site, it must act on RNA polymerase to ensure that the enzyme can no longer respond to the terminator. The variable locations of the *nut* sites indicate that this event is linked neither to initiation nor to termination, but can occur to RNA polymerase as it elongates the RNA chain past the *nut* site. As illustrated in **Figure 9.54,** the polymerase then becomes a juggernaut that continues past the terminator, heedless of its signal. (This reaction involves antitermination at rho-dependent terminators, but pN also suppresses termination at intrinsic terminators.)

Is the ability of pN to recognize a short sequence within the transcription unit an example of a more widely used mechanism for antitermination? Other phages, related to lambda, have different *N* genes and different antitermination specificities. The region of the phage genome in which the *nut* sites lie has a different sequence in each of these phages, and each phage must therefore have characteristic *nut* sites recognized specifically by its own pN. Each of these pN products must have the same general ability to interact with the transcription apparatus in an antitermination capacity, but has a different specificity for the sequence of DNA that activates the mechanism.

9.25 Termination and anti-termination factors interact with RNA polymerase

Key Concepts

- Several bacterial proteins are required for lambda pN to interact with RNA polymerase.
- These proteins are also involved in antitermination in the *rrn* operons of the host bacterium.
- The lambda antiterminator pQ has a different mode of interaction that involves binding to DNA at the promoter.

Termination and antitermination are closely connected, and involve bacterial and phage proteins that interact with RNA polymerase. Several proteins concerned with termination have been

identified by isolating mutants of *E. coli* in which pN is ineffective. Several of these mutations lie in the *rpoB* gene. This argues that pN (like the rho factor) interacts with the β subunit of the core enzyme. Other *E. coli* mutations that prevent pN function identify the *nus* loci: *nusA, nusB, nusE,* and *nusG.* (The term *nus* is an acronym for *N* utilization substance.)

A lambda *nut* site consists of two sequence elements, called *boxA* and *boxB.* Sequence elements related to *boxA* are also found in bacterial operons. The *boxA* element is required for binding bacterial proteins that are necessary for antitermination in both phage and bacterial operons. The *boxB* element is specific to the phage genome, and mutations in *boxB* abolish the ability of pN to cause antitermination.

The *nus* loci code for proteins that form part of the transcription apparatus, but that are not isolated with the RNA polymerase enzyme. The *nusA, nusB,* and *nusG* functions are concerned solely with the termination of transcription. The *nusE* loci codes for ribosomal protein S10; the relationship between its location in the 30S subunit and its function in termination is not clear. The Nus proteins bind to RNA polymerase at the *nut* site, as summarized in **Figure 9.55.**

NusA is a general transcription factor that increases the efficiency of termination, probably by enhancing RNA polymerase's tendency to pause at terminators (and indeed at other regions of secondary structure; see below). NusB and S10 form a dimer that binds specifically to RNA containing a *boxA* sequence. NusG may be concerned with the general assembly of all the Nus factors into a complex with RNA polymerase.

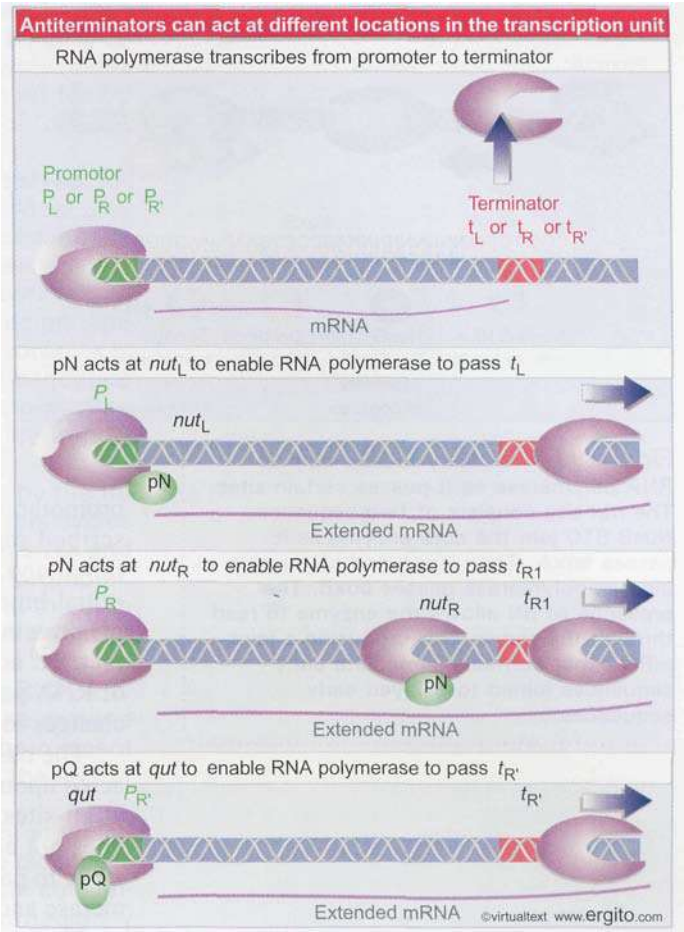


Figure 9.53 Host RNA polymerase transcribes lambda genes and terminates at f sites. pN allows it to read through terminators in the L and R1 units; pQ allows it to read through the R' terminator. The sites at which pN acts *{nut)* and at which pQ acts (*qut*) are located at different relative positions in the transcription units.

Intrinsic and rho-dependent terminators have different requirements for the Nus factors. NusA is required for termination at intrinsic terminators, and the reaction can be prevented by pN. At rho-dependent terminators, all 4 Nus proteins are required, and again pN alone can inhibit the reaction. The common feature of pN at both types of terminator is to prevent the role of NusA in termination. Binding of pN to NusA inhibits the ability of NusA to bind RNA, which is necessary for termination.

Antitermination occurs in the *rrn* (rRNA) operons of *E. coli,* and involves the same *nus* functions. The leader regions of the *rrn* operons contain *boxA* sequences; NusB-S10 dimers recognize these sequences and bind to RNA polymerase as it elongates past *boxA.* This changes the properties of RNA polymerase in such a way that it can now read through rho-dependent terminators that are present within the transcription unit.

The *boxA* sequence of lambda RNA does not bind NusB-S10, and is probably enabled to do so by the presence of NusA and pN; the *boxB* sequence could be required to stabilize the reaction. So variations in *boxA* sequences may determine which particular set of factors is required for antitermination. The consequences are the same: when RNA polymerase passes the *nut* site, it is modified by addition of appropriate factors, and fails to terminate when it subsequently encounters the terminator sites.

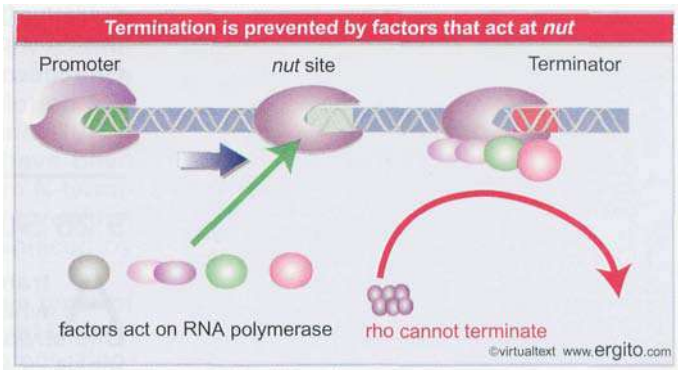


Figure 9.54 Ancillary factors bind to RNA polymerase as it passes the *nut* site. They prevent rho from causing termination when the polymerase reaches the terminator.

Antitermination in lambda requires pN to bind to RNA polymerase in a manner that depends on the sequence of the transcription unit. Does pN recognize the *boxB* site in DNA or in the RNA transcript? It does

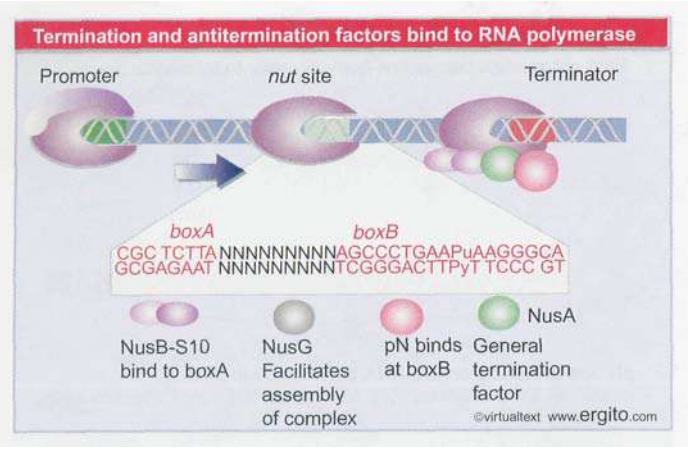


Figure 9.55 Ancillary factors bind to RNA polymerase as it passes certain sites. The *nut* site consists of two sequences. NusB-S10 join the core enzyme as it passes boxA. Then NusA and pN protein bind as polymerase passes boxB. The presence of pN allows the enzyme to read through the terminator, producing a joint mRNA that contains immediate early sequences joined to delayed early sequences.

not bind independently to either type of sequence, but does bind to a transcription complex when the core enzyme passes the *boxB* site. The pN protein has separate domains that recognize the *boxB* RNA sequence and the NusA protein. After joining the transcription complex, pN remains associated with the core enzyme, in effect becoming an additional subunit whose presence changes recognition of terminators. It is possible that pN in fact continues to bind to both the *boxB* RNA sequence and to RNA polymerase, maintaining a loop in the RNA; thus the role of *boxB* RNA would partly be to tether pN in the vicinity, effectively increasing its local concentration.

The pQ protein, which prevents termination later in phage infection by acting at *qut*, has a different mode of action. The *qut* sequence lies at the start of the late transcription unit. The upstream part of *qut* lies within the promoter, while the downstream part lies at the beginning of the transcribed region. This implies that pQ action involves recognition of DNA, and implies that its mechanism of action, at least concerning the initial binding to the complex, must be different from that of pN. The pQ protein interacts with the holoenzyme during the initiation phase. In fact, σ^{70} is required for the interaction with pQ. This reinforces the view of RNA polymerase as an interactive structure in which conformational changes induced at one phase may affect its activity at a later phase.

The basic action of pQ is to interfere with pausing; and once pQ has acted upon RNA polymerase, the enzyme shows much reduced pausing at all sites, including rho-dependent and intrinsic terminators. So pQ does not act directly on termination *per se,* but instead allows the enzyme to pass the terminator more quickly, thus depriving the core polymerase and/or accessory factor of the opportunity to cause termination.

The general principle is that RNA polymerase may exist in forms that are competent to undertake particular stages of transcription, and its activities at these stages can be changed only by modifying the appropriate form. So substitutions of sigma factors may change one initiation-competent form into another; and additions of Nus factors may change the properties of termination-competent forms.

Termination seems to be closely connected with the mode of elongation. In its basic transcription mode, core polymerase is subject to many pauses during elongation; and pausing at a terminator site is the prerequisite for termination to occur. Under the influence of factors such as NusA, pausing becomes extended, increasing the efficiency of termination; while under the influence of pN or pQ, pausing is abbreviated decreasing the efficiency of termination. Because recognition sites for these factors are found only in certain transcription units, pausing and consequently termination are altered only in those units.

9.26 Summary

A transcription unit comprises the DNA between a promoter, where transcription initiates, and a terminator, where it ends. One strand of the DNA in this region serves as a template for synthesis of a complementary strand of RNA. The RNA-DNA hybrid region is short and transient, as the transcription "bubble" moves along DNA. The RNA polymerase holoenzyme that synthesizes bacterial RNA can be separated into two components. A core enzyme is a multimer of structure $\alpha_2\beta\beta'$ that is responsible for elongating the RNA chain. A sigma factor (σ) is a single subunit that is required at the stage of initiation for recognizing the promoter.

By Book_Crazy [IND]

Core enzyme has a general affinity for DNA. The addition of sigma factor reduces the affinity of the enzyme for nonspecific binding to DNA, but increases its affinity for promoters. The rate at which RNA polymerase finds its promoters is too great to be accounted for by diffusion and random contacts with DNA; direct exchange of DNA sequences held by the enzyme may be involved.

Bacterial promoters are identified by two short conserved sequences centered at -35 and -10 relative to the startpoint. Most promoters have sequences that are well related to the consensus sequences at these sites. The distance separating the consensus sequences is 16-18 bp. RNA polymerase initially "touches down" at the -35 sequence and then extends its contacts over the -10 region. The enzyme covers ~77 bp of DNA. The initial "closed" binary complex is converted to an "open" binary complex by melting of a sequence of ~12 bp that extends from the -10 region to the startpoint. The AT-rich base pair composition of the -10 sequence may be important for the melting reaction.

The binary complex is converted to a ternary complex by the incorporation of ribonucleotide precursors. There are multiple cycles of abortive initiation, during which RNA polymerase synthesizes and releases very short RNA chains without moving from the promoter. At the end of this stage, there is a change in structure, and the core enzyme contracts to cover ~50 bp. The sigma factor is either released (30% of cases) or changes its form of association with the core enzyme. Then the core enzyme moves along DNA, synthesizing RNA. A locally unwound region of DNA moves with the enzyme. The enzyme contracts further in size to cover only 30-40 bp when the nascent chain has reached 15-20 nucleotides; then it continues to the end of the transcription unit.

The "strength" of a promoter describes the frequency at which RNA polymerase initiates transcription; it is related to the closeness with which its -35 and -10 sequences conform to the ideal consensus sequences, but is influenced also by the sequences immediately downstream of the startpoint. Negative supercoiling increases the strength of certain promoters. Transcription generates positive supercoils ahead of RNA polymerase and leaves negative supercoils behind the enzyme.

The core enzyme can be directed to recognize promoters with different consensus sequences by alternative sigma factors. In *E. coli,* these sigma factors are activated by adverse conditions, such as heat shock or nitrogen starvation. *B. subtilis* contains a single major sigma factor with the same specificity as the *E. coli* sigma factor, and also contains a variety of minor sigma factors. Another series of factors is activated when sporulation is initiated; sporulation is regulated by two cascades in which sigma factor replacements occur in the forespore and mother cell. A cascade for regulating transcription by substitution of sigma factors is also used by phage SPO1.

The geometry of RNA polymerase-promoter recognition is similar for holoenzymes containing all sigma factors (except σ^{54}). Each sigma factor causes RNA polymerase to initiate transcription at a promoter that conforms to a particular consensus at -35 and -10. Direct contacts between sigma and DNA at these sites have been demonstrated for *E. coli* a^{70}. The σ^{70} factor of *E. coli* has an N-terminal autoinhibitory domain that prevents the DNA-binding regions from recognizing DNA. The autoinhibitory region is displaced by DNA when the holoenzyme forms an open complex.

Bacterial RNA polymerase terminates transcription at two types of sites. Intrinsic terminators contain a GC-rich hairpin followed by a U-rich region. They are recognized *in vitro* by core enzyme alone. Rho-dependent terminators require a rho factor both *in vitro* and *in vivo;* they have a stretch of 50-90 nucleotides preceding the site of termination that is rich in C and poor in G residues. The rho factor is an essential protein that acts as an ancillary termination factor. Rho binds to a single-stranded stretch of RNA and translocates along the RNA until it reaches the RNA-DNA hybrid region in the transcription bubble of

RNA polymerase. Rho has a hexameric ATP-dependent helicase activity that separates the RNA from the DNA. In both types of termination, pausing by RNA polymerase is important in order to allow time for the actual termination event to occur.

The Nus factors are required for termination. NusA is required for intrinsic terminators, and in addition NusB-S10 is required for rho-dependent terminators. The NusB-S10 dimer recognizes the *boxA* sequence of a *nut* site in the elongating RNA; NusA joins subsequently.

Antitermination is used by some phages to regulate progression from one stage of gene expression to the next. The lambda gene *N* codes for an antitermination protein (pN) that is necessary to allow RNA polymerase to read through the terminators located at the ends of the immediate early genes. Another antitermination protein, pQ, is required later in phage infection. pN and pQ act on RNA polymerase as it passes specific sites (*nut* and *qut,* respectively). These sites are located at different relative positions in their respective transcription units. The pN protein recognizes RNA polymerase carrying NusA when the enzyme passes the sequence *boxB.* The pN protein then binds to the complex and prevents termination by antagonizing the action of NusA when the polymerase reaches the rho-dependent terminator.

References

9.2 Transcription occurs by base pairing in a "bubble" of unpaired DNA

Losick, R. and Chamberlin, M. (1976). RNA Polymerase. Cold Spring Harbor Symp. Quant. Biol.

ref Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., and Darst, S. A. (2000). A structural model of transcription elongation. Science 289, 619-625.

9.3 The transcription reaction has three stages

ref Rice, G. A., Kane, C. M., and Chamberlin, M. (1991). Footprinting analysis of mammalian RNA polymerase II along its transcript: an alternative view of transcription elongation. Proc. Nat. Acad. Sci. USA 88, 4245-281.

Wang, D. et al. (1995). Discontinuous movements of DNA and RNA in RNA polymerase accompany formation of a paused transcription complex. Cell 81, 341-350.

9.4 Phage T7 RNA polymerase is a useful model system

ref Cheetham, G. M. T. and Steitz, T. A. (1999). Structure of a transcribing T7 RNA polymerase initiation complex. Science 286, 2305-2309.

Cheetham, G. M., Jeruzalmi, D., and Steitz, T. A. (1999). Structural basis for initiation of transcription from an RNA polymerase-promoter complex. Nature 399, 80-83.

Temiakov, D., Mentesana, D., Temiakov, D., Ma, K., Mustaev, A., Borukhov, S., and McAllister, W. T. (2000). The specificity loop of T7 RNA polymerase interacts first with the promoter and then with the elongating transcript, suggesting a mechanism for promoter clearance. Proc. Nat. Acad. Sci. USA 97, 14109-14114.

9.5 A model for enzyme movement is suggested by the crystal structure

ref Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R., and Kornberg, R. D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science 288, 640-649.

Cramer, P., Bushnell, P., and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 A resolution. Science 292, 1863-1876.

Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science 292, 1876-1882.

9.6 Bacterial RNA polymerase consists of multiple subunits

rev Helmann, J. D. and Chamberlin, M. (1988). Structure and function of bacterial sigma factors. Ann. Rev. Biochem. 57, 839-872.

ref Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A., and Darst, S. A. (2001). Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. Cell 104, 901-912.

Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., and Darst, S. A. (2000). A structural model of transcription elongation. Science 289, 619-625.

Zhang, G., Campbell, E. A., Zhang, E. A., Minakhin, L., Richter, C, Severinov, K., and Darst, S. A. (1999). Crystal structure of *Thermusaquaticus* core RNA polymerase at 3.3 A resolution. Cell 98, 81 1-824.

9.7 RNA polymerase consists of the core enzyme and sigma factor

ref Travers, A. A. and Burgess, R. R. (1969). Cyclic reuse of the RNA polymerase sigma factor. Nature 222, 537-540.

9.8 The association with sigma factor changes at initiation

ref Bar-Nahum, G. and Nudler, E. (2001). Isolation and characterization of sigma(70)-retaining transcription elongation complexes from *E. coli.* Cell 106, 443-451.

Krummel, B. and Chamberlin, M. J. (1989). RNA chain initiation by *E. coli* RNA polymerase. Structural transitions of the enzyme in early ternary complexes. Biochemistry 28, 7829-7842.

Mukhopadhyay, J., Kapanidis, A. N., Mekler, V., Kortkhonjia, E., Ebright, Y. W., and Ebright, R. H. (2001). Translocation of sigma(70) with RNA Polymerase during Transcription. Fluorescence Resonance Energy Transfer Assay for Movement Relative to DNA. Cell 106, 453-463.

9.9 A stalled RNA polymerase can restart

Nudler, E. et al. (1997). The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. Cell 89, 33-41.

Sigma factor controls binding to DNA

Bar-Nahum, G. and Nudler, E. (2001). Isolation and characterization of sigma(70)-retaining transcription *elongation* complexes from *E. coli*. Cell 106, 443-451.

Kortkhonjia, E., Ebright, Y. W., and Ebright, R. H. (2001). Translocation of sigma(70) with RNA Polymerase during Transcription. Fluorescence Resonance Energy Transfer Assay for Movement Relative to DNA. Cell 106, 453-463.

.12 Promoter recognition depends on consensus sequences

rev McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. Ann. Rev. Biochem. 54, 171-204.

ref Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C, Ishihama, A., Severinov, K., and Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science 262, 1407-1413.

.13 Promoter efficiencies can be increased or decreased by mutation

rev McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. Ann. Rev. Biochem. 54, 171-204.

).14 RNA polymerase binds to one face of DNA

ref Siebenlist, U., Simpson, R. B., and Gilbert, W. (1980). *E. coli* RNA polymerase interacts homologously with two different promoters. Cell 20, 269-281.

9.15 Supercoiling is an important feature of transcription

ref Wu, H.-Y. et al. (1988). Transcription generates positively and negatively supercoiled domains in the template. Cell 53, 433-440.

9.16 Substitution of sigma factors may control initiation

rev Hengge-Aronis, R. (2002). Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. Microbiol. Mol. Biol. Rev. 66, 373-393.

ref Alba, B. M., Onufryk, C, Lu, C. Z., and Gross, C. A. (2002). DegS and YaeL participate sequentially in the cleavage of RseA to activate the sigma(E)-dependent extracytoplasmic stress response. Genes Dev. 16, 2156-2168.

Grossman, A. D., Erickson, J. W., and Gross, C. A. (1984). The htpR gene product of *E. coli* is a sigma factor for heat-shock promoters. Cell 38, 383-390.

Kanehara, K., Ito, K., and Akiyama, Y. (2002). YaeL (EcfE) activates the sigma(E) pathway of stress response through a site-2 cleavage of anti-sigma(E), RseA. Genes Dev. 16, 2147-2155.

Sakai, J., Duncan, E. A., Rawson, R. B., Hua, X., Brown, M. S., and Goldstein, J. L. (1996). Sterol-regulated release of SREBP-2 from cell membranes requires two sequential cleavages, one within a transmembrane segment. Cell 85, 1037-1046.

9.17 Sigma factors directly contact DNA

ref Campbell, E. A., Muzzin, O., Chlenov, M., Sun, J. L., Olson, C. A., Weinman, O., Trester-Zedlitz, M. L., and Darst, S, A. (2002). Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol. Cell 9, 527-539.

Dombrowski, A. J. et al. (1992). Polypeptides containing highly conserved regions of transcription initiation factor a^{70} exhibit specificity of binding to promoter DNA. Cell 70, 501-512.

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A. N., Niu, W., Ebright, Y. W., Levy, R., and Ebright, R. H. (2002). Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. Cell 108, 599-614.

Vassylyev, S3, G. Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S., Yokoyama ST. (2002). Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. Nature 417, 712-719.

9.19 Sporulation is controlled by sigma factors

rev Errington, J. (1993). *B. subtilis* sporulation: regulation of gene expression and control of morphogenesis. Microbiol. Rev. 57, 1-33.

Haldenwang, W. G. (1995). The sigma factors of *B. subtilis.* Microbiol. Rev. 59, 1-30.

Losick, R. et al. (1986). Genetics of endospore formation in *B. subtilis.* Ann. Rev. Genet. 20, 625-669.

Losick, R. and Stragier, P. (1992). Crisscross regulation of cell-type specific gene expression during development in *B. subtilis.* Nature 355, 601-604.

Stragier, P. and Losick, R. (1996). Molecular genetics of sporulation in *B. subtilis.* Ann. Rev. Genet. 30, 297-341.

ref Haldenwang, W. G. and Losick, R. (1980). A novel RNA polymerase sigma factor from *B. subtilis.* Proc. Nat. Acad. Sci. USA 77, 7000-7004.

Haldenwang, W. G., Lang, N., and Losick, R. (1981). A sporulation-induced sigma-like regulatory protein from *B. subtilis.* Cell 23, 615-624.

9.20 Bacterial RNA polymerase terminates at discrete sites

rev Adhya, S. and Gottesman, M. (1978). Control of transcription termination. Ann. Rev. Biochem. 47, 967-996.

Friedman, D. I., Imperiale, M. J., and Adhya, S. L. (1987). RNA 3' end formation in the control of gene expression. Ann. Rev. Genet. 21, 453-488.

Platt, T. (1986). Transcription termination and the regulation of gene expression. Ann. Rev. Biochem. 55, 339-372.

9.21 There are two types of terminators in *E. coli*

rev von Hippel, P. H. (1998). An integrated model of the transcription complex in elongation, termination, and editing. Science 281, 660-665.

ref Lee, D. N., Phung, L., Stewart, J., and Landick, R. (1990). Transcription pausing by *E. coli* RNA polymerase is modulated by downstream DNA sequences. J. Biol. Chem. 265, 15145-15153.

Lesnik, E. A., Sampath, R., Levene, H. B., Henderson, T. J., McNeil, J.A., and Ecker, D. J. (2001). Prediction of rho-independent transcriptional terminators in *E. coli*. Nuc. Acids Res. 29, 3583-3594.

Reynolds, R., Bermadez-Cruz, R. M., and Chamberlin, M. J. (1992). Parameters affecting transcription termination by *E. coli* RNA polymerase. I. Analysis of 13 rho-independent terminators. J. Mol. Biol. 224, 31-51.

9.22 How does rho factor work?

rev Das, A. (1993). Control of transcription termination by RNA-binding proteins. Ann. Rev. Biochem. 62, 893-930.

Richardson, J. P. (1996). Structural organization of transcript/on termination factor Rho. J. Biol. Chem. 271, 1251-1254.

von Hippel, P. H. (1998). An integrated model of the transcription complex in elongation, termination, and

ref Brennan, C. A., Dombroski, A. J., and Platt, T. (1987). Transcription termination factor rho is an RNA-DNA helicase. Cell 48, 945-952.

Geiselmann, J., Wang, Y., Seifried, S. E., and von Hippel, P. H. (1993). A physical model for the translocation and helicase activities of *E. coli* transcription termination protein Rho. Proc. Nat. Acad. Sci. USA 90, 7754-7758.

Roberts, J. W. (1969). Termination factor for RNA synthesis. Nature 224, 1168-1174.

9.25 Termination and anti-termination factors interact with RNA polymerase

rev Greenblatt, J., Nodwell, J. R., and Mason, S. W. (1993). Transcriptional antitermination. Nature 364 401-406.

ref Legault, P., Li, J., Mogridge, J., Kay, L. E., and Greenblatt, J. (1998). NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. Cell 93, 289-299.

Mah, T. F., Kuznedelov, K., Mushegian, A., Severinov, K., and Greenblatt, J. (2000). The alpha subunit of *E. coli* RNA polymerase activates RNA binding by NusA. Genes Dev. 14, 2664-2675.
Mogridge, J., Mah, J., and Greenblatt, J. (1995). A protein-RNA interaction network facilitates the template-independent cooperative assembly on RNA polymerase of a stable antitermination complex containing the lambda N protein. Genes Dev. 9, 2831-2845.

Olson, E. R., Flamm, E. L, and Friedman, D. I. (1982). Analysis of nutR: a region of phage lambda required for antitermination of transcription. Cell 31, 61-70.

Chapter 10

The operon

10.1 Introduction

Gene expression can be controlled at any of several stages, which we divide broadly into transcription, processing, and translation:

Transcription often is controlled at the stage of initiation. Transcription is not usually controlled at elongation, but may be controlled at termination to determine whether RNA polymerase is allowed to proceed past a terminator to the gene(s) beyond.

In eukaryotic cells, processing of the RNA product may be regulated at the stages of modification, splicing, transport, or stability. In bacteria, an mRNA is in principle available for translation as soon as (or even while) it is being synthesized, and these stages of control are not available.

Translation may be regulated, usually at the stages of initiation and termination (like transcription). Regulation of initiation is formally analogous to the regulation of transcription: the circuitry can be drawn in similar terms for regulating initiation of transcription on DNA or initiation of translation on RNA.

The basic concept for how transcription is controlled in bacteria was provided by the classic formulation of the model for control of gene expression by Jacob and Monod in 1961. They distinguished between two types of sequences in DNA: sequences that code for trans-acting products; and cis-acting sequences that function exclusively within the DNA. Gene activity is regulated by the specific interactions of the *trans*-acting products (usually proteins) with the *cis*-acting sequences (usually sites in DNA). In more formal terms:

A gene is a sequence of DNA that codes for a diffusible product. This product may be protein (as in the case of the majority of genes) or may be RNA (as in the case of genes that code for tRNA and rRNA). *The crucial feature is that the product diffuses away from its site of synthesis to act elsewhere.* Any gene product that is free to diffuse to find its target is described as *trans*-acting.

The description *cis*-acting applies to any sequence of DNA that is not converted into any other form, but that functions exclusively as a DNA sequence *in situ*, affecting only the DNA to which it is

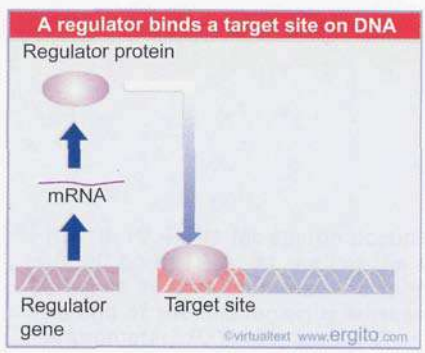


Figure 10.1 A regulator gene codes for a protein that acts at a target site on DNA.

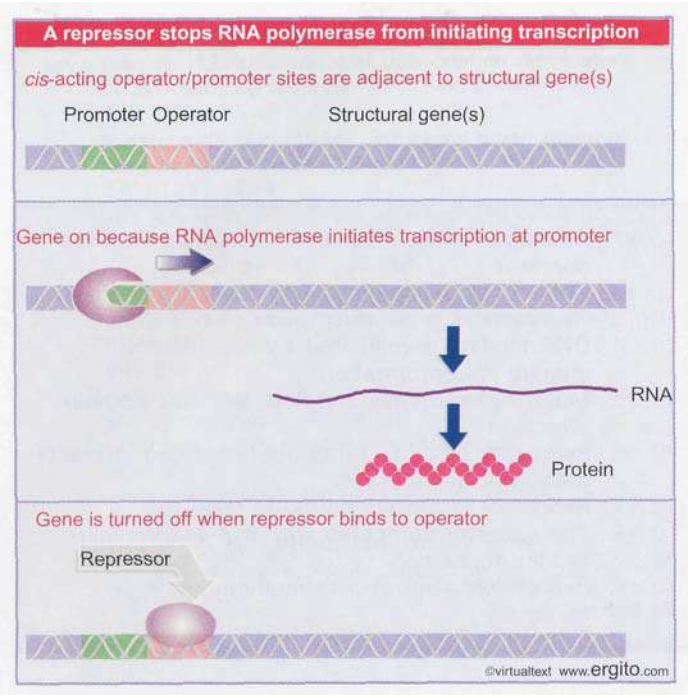


Figure 10.2 In negative control, a trans-acting repressor binds to the c/s-acting operator to turn off transcription.

physically linked. (In some cases, a *cis*-acting sequence functions in an RNA rather than in a DNA molecule.)

To help distinguish between the components of regulatory circuits and the genes that they regulate, we sometimes use the terms structural gene and regulator gene. A structural gene is simply any gene that codes for a protein (or RNA) product. Structural genes represent an enormous variety of protein structures and functions, including structural proteins, enzymes with catalytic activities, and regulatory proteins. A regulator gene simply describes a gene that codes for a protein (or an RNA) involved in regulating the expression of other genes.

The simplest form of the regulatory model is illustrated in Figure **10.1:** *a regulator gene codes for a protein that controls transcription by binding to particular site(s) on DNA.* This interaction can regulate a target gene in either a positive manner (the interaction turns the gene on) or in a negative manner (the interaction turns the gene off). The sites on DNA are usually (but not exclusively) located just upstream of the target gene.

The sequences that mark the beginning and end of the transcription unit, the promoter and terminator, are examples of *cis*-acting sites. *A promoter serves to initiate transcription only of the gene or genes physically connected to it on the same stretch of DNA.* In the same way, a terminator can terminate transcription only by an RNA polymerase that has traversed the preceding gene(s). In their simplest forms, promoters and terminators are c/s-acting elements that are recognized by the same *trans*-acting species, that is, by RNA polymerase (although other factors also participate at each site).

Additional *cis*-acting regulatory sites are often juxtaposed to, or interspersed with, the promoter. A bacterial promoter may have one or more such sites located close by, that is, in the immediate vicinity of the startpoint. A eukaryotic promoter is likely to have a greater number of sites, spread out over a longer distance.

10.2 Regulation can be negative or positive

Key Concepts

- In negative regulation a repressor protein binds to an operator to prevent a gene from being expressed.
- In positive regulation a transcription factor is required to bind at the promoter in order to enable RNA polymerase to initiate transcription.

A classic mode of control in bacteria is *negative:* a repressor protein prevents a gene from being expressed. Figure 10.2 shows that the "default state" for such a gene is to be expressed via the recognition of its promoter by RNA polymerase. Close to the promoter is another *cis*-acting site called the operator, which is the target for the repressor protein. When the repressor binds to the operator, RNA polymerase is prevented from initiating transcription, and *gene expression is therefore turned off.*

An alternative mode of control is *positive.* This is used in bacteria (probably) with about equal frequency to negative control, and it is the most common mode of control in eukaryotes. A transcription factor is required to assist RNA polymerase in initiating at the promoter. Figure 10.3 shows that the typical default state of a eukaryotic gene is inactive:

polymerase cannot by itself initiate transcription at the promoter. Several *trans-acting* factors have target sites in the vicinity of the promoter, and *binding of some or all of these factors enables RNA polymerase to initiate transcription.*

The unifying theme is that regulatory proteins are *trans-acting* factors that recognize *cis*-acting elements (usually) upstream of the gene. The consequences of this recognition are to activate or to repress the gene, depending on the individual type of regulatory protein. A typical feature is that the protein functions by recognizing a very short sequence in DNA, usually <10 bp in length, although the protein actually binds over a somewhat greater distance of DNA. *The bacterial promoter is* an example: although RNA polymerase covers >70 bp of DNA *at initiation, the crucial sequences that it recognizes* are the hexamers centered at -35 and -10.

A significant difference in gene organization between prokaryotes and eukaryotes is that structural genes in bacteria are organized in clusters, while those in eukaryotes occur individually. Clustering of structural genes allows them to be coordinately controlled by means of interactions at a single promoter: as a result of these interactions, the entire set of genes is either transcribed or not transcribed. In this chapter, we discuss this mode of control and its use by bacteria. The means employed to coordinate control of dispersed eukaryotic genes are discussed in *22 Activating transcription.*

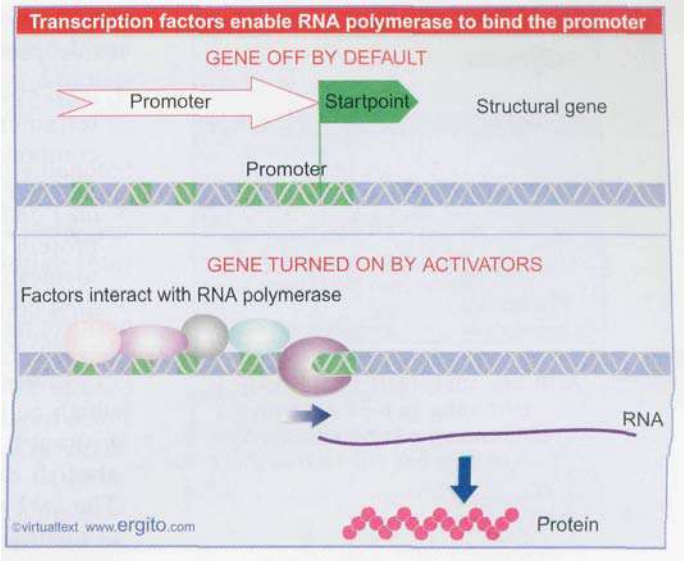


Figure 10.3 In positive control, Trans-acting factors must bind to *cis*-acting sites in order for RNA polymerase to initiate transcription at the promoter.

10.3 Structural gene clusters are coordinately

Key Concepts

- Genes coding for proteins that function in the same pathway may be located adjacent to one another and controlled as a single unit that is transcribed into a polycistronic mRNA.

Bacterial structural genes are often organized into clusters that include genes coding for proteins whose functions are related. It is common for the genes coding for the enzymes of a metabolic pathway to be organized into such a cluster. In addition to the enzymes actually involved in the pathway, other related activities may be included in the unit of coordinate control; for example, the protein responsible for transporting the small molecule substrate into the cell.

The cluster of the three *lac* structural genes, *lacZYA*, is typical. **Figure** 104 summarizes the organization of the structural genes, their associated *cis*-acting regulatory elements, and the *trans*-acting regulatory gene. *The key feature is that the cluster is transcribed into a single polycistronic mRNA from a promoter where initiation of transcription is regulated.*

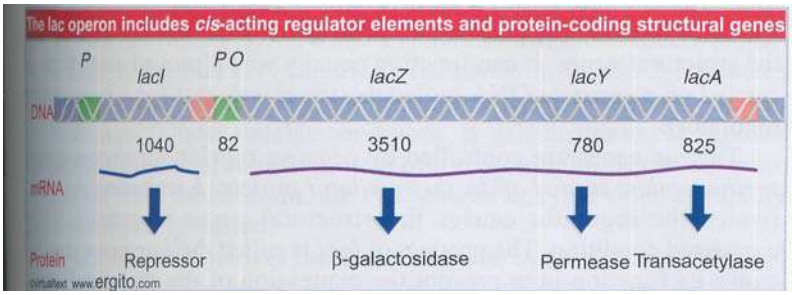


Figure 10.4 The *lac* operon occupies ~6000 bp of DNA. At the left the *lacI* gene has its own promoter and terminator. The end of the *lacI* region is adjacent to the promoter, P. The operator, *O*, occupies the first 26 bp of the transcription unit. The long *lacZ* gene starts at base 39, and is followed by the *lacY* and *lacA* genes and a terminator.

The protein products enable cells to take up and metabolize β-galactosides, such as lactose. The roles of the three structural genes are:

- *lacZ* codes for the enzyme β-galactosidase, whose active form is a tetramer of ~500 kD. The enzyme breaks a β-galactoside into its component sugars. For example, lactose is cleaved into glucose and galactose (which are then further metabolized).
- *lacY* codes for the β-galactoside permease, a 30 kD membrane-bound protein constituent of the transport system. This transports β-galactosides into the cell.
- *lacA* codes for β-galactoside transacetylase, an enzyme that transfers an acetyl group from acetyl-CoA to β-galactosides.

Mutations in either *lacZ* or *lacY* can create the *lac* genotype, in which cells cannot utilize lactose. (The genotypic description *"lac"*\ without a qualifier indicates loss-of-function.) The *lacZ* mutations abolish enzyme activity, directly preventing metabolism of lactose.l The *lacY* mutants cannot take up lactose from the medium. (No defect is identifiable in *lacA* cells, which is puzzling. It is possible that the acetylation reaction gives an advantage when the bacteria grow in the presence of certain analogs of β-galactosides that cannot be metabolized, because the modification results in detoxification and excretion.)

The entire system, including structural genes and the elements that control their expression, forms a common unit of regulation; this is called an operon. The activity of the operon is controlled by regulator gene(s), whose protein products interact with the *cis*-acting control elements.

10.4 The *lac* genes are controlled by a repressor

Key Concepts

- Transcription of the *lacZYA* gene cluster is controlled by a repressor protein that binds to an operator that overlaps the promoter at the start of the cluster.
- The repressor protein is a tetramer of identical subunits coded by the gene *lacI*.

We can distinguish between structural genes and regulator genes by the effects of mutations. A mutation in a structural gene deprives the cell of the particular protein for which the gene codes. But a mutation in a regulator gene influences the expression of all the structural genes that it controls. The consequences of a regulatory mutation reveal the type of regulation.

Transcription of the *lacZYA* genes is controlled by a regulator protein synthesized by the *lacI* gene. It happens that *lacI* is located adjacent to the structural genes, but it comprises an independent transcription unit with its own promoter and terminator. Since *lacI* specifies a diffusible product, in principle it need not be located near the structural genes; it can function equally well if moved elsewhere,™ carried on a separate DNA molecule (the classic test for a *trans*-acting regulator).

The *lac* genes are controlled by negative regulation: *they are transcribed unless turned off by the regulator protein*. A mutation that inactivates the regulator causes the structural genes to remain in the expressed condition. The product of *lacI* is called the *Lac repressor*, because its function is to prevent the expression of the structural genes

The repressor is a tetramer of identical subunits of 38 kD each. There are ~10 tetramers in a wild-type cell. The regulator gene is not controlled by the availability of lactose. It is transcribed into a monocistronic mRNA at a rate that appears to be governed simply by the affinity of its promoter for RNA polymerase.

The repressor functions by binding to an operator (formally denoted O_{lac}) at the start of the *lacZYA* cluster. The operator lies between the promoter (P_{lac}) and the structural genes *(lacZYA). When the repressor binds at the operator, it prevents RNA polymerase from initiating transcription at the promoter.* **Figure 10.5** expands our view of the region at the start of the *lac* structural genes. The operator extends from position -5 just upstream of the mRNA startpoint to position +21 within the transcription unit. So it overlaps the right end of the promoter. We discuss the relationship between repressor and RNA polymerase in more detail in *10.10 Repressor protein binds to the operator* and *10.16 Repressor binds to three operators and interacts with RNA polymerase.*

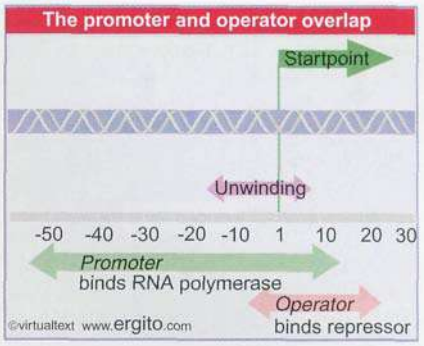


Figure 10.5 Repressor and RNA polymerase bind at sites that overlap around the transcription startpoint of the *lac* operon.

10.5 The *lac* operon can be induced

Key Concepts

- Small molecules that induce an operon are identical with or related to the substrate for its enzymes.
- β-galactosides are the substrates for the enzymes coded by *lacZYA.*
- In the absence of β-galactosides the *lac* operon is expressed only at a very low (basal) level.
- Addition of β-galactosides induces transcription of all three genes of the operon.
- Because the *lac* mRNA is extremely unstable, induction can be rapidly reversed.
- The same types of systems that allow substrates to induce operons coding for metabolic enzymes can be used to allow end-products to repress the operons that code for biosynthetic enzymes.

Bacteria need to respond swiftly to changes in their environment. Fluctuations in the supply of nutrients can occur at any time; survival depends on the ability to switch from metabolizing one substrate to another. Yet economy also is important, since a bacterium that indulges in energetically expensive ways to meet the demands of the environment is likely to be at a disadvantage. So a bacterium avoids synthesizing the enzymes of a pathway in the absence of the substrate; but is ready to produce the enzymes if the substrate should appear.

The synthesis of enzymes in response to the appearance of a specific substrate is called induction. This type of regulation is widespread in bacteria, and occurs also in unicellular eukaryotes (such as yeasts). The lactose system of *E. coli* provides the paradigm for this sort of control mechanism.

When cells of *E. coli* are grown in the absence of a β-galactoside, there is no need for β-galactosidase, and they contain very few molecules of the enzyme —say, 5. When a suitable substrate is added, the enzyme activity appears very rapidly in the bacteria. Within 2-3 minutes some enzyme is present, and soon there are ~5000 molecules of enzyme per bacterium. (Under suitable conditions, β-galactosidase can account for 5-10% of the total soluble protein of the bacterium.) If the substrate is removed from the medium, the synthesis of enzyme stops as rapidly as it had originally started.

Figure 10.6 summarizes the essential features of induction. Control of transcription of the *lac* genes responds very rapidly to the inducer, as

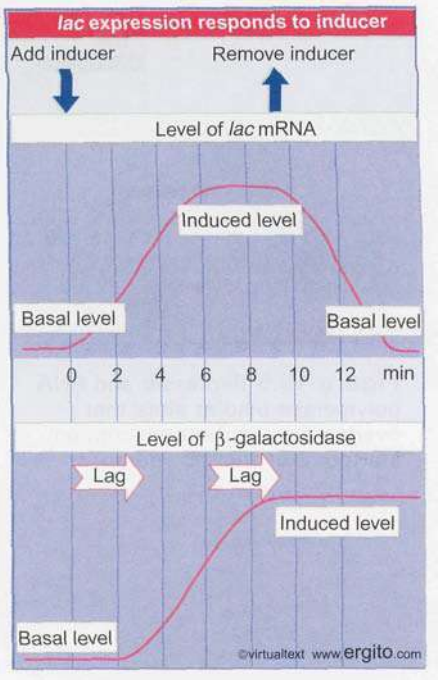


Figure 10.6 Addition of inducer results in rapid induction of lac mRNA, and is followed after a short lag by synthesis of the enzymes; removal of inducer is followed by rapid cessation of synthesis.

shown in the upper part of the figure. In the absence of inducer, the operon is transcribed at a very low basal level. Transcription is stimulated as soon as inducer is added; the amount of *lac* mRNA increases rapidly to an induced level that reflects a balance between synthesis and degradation of the mRNA.

The *lac* mRNA is extremely unstable, and decays with a half-life of only ~3 minutes. This feature allows induction to be reversed rapidly. Transcription ceases as soon as the inducer is removed; and in a very short time all the *lac* mRNA has been destroyed, and the cellular content has returned to the basal level.

The production of protein is followed in the lower part of the figure. Translation of the *lac* mRNA produces β-galactosidase (and the products of the other *lac* genes). There is a short lag between the appearance of *lac* mRNA and appearance of the first completed enzyme molecules (it is ~2 min after the rise of mRNA from basal level before protein begins to increase). There is a similar lag between reaching maximal induced levels of mRNA and protein. When inducer is removed, synthesis of enzyme ceases almost immediately (as the mRNA is degraded), but the β-galactosidase in the cell is more stable than the mRNA, so the enzyme activity remains at the induced level for longer.

This type of rapid response to changes in nutrient supply not only provides the ability to metabolize new substrates, but also is used to shut off endogenous synthesis of compounds that suddenly appear in the medium. For example, *E. coli* synthesizes the amino acid tryptophan through the action of the enzyme tryptophan synthetase. But if tryptophan is provided in the medium on which the bacteria are growing, the production of the enzyme is immediately halted. This effect is called repression. It allows the bacterium to avoid devoting its resources to unnecessary synthetic activities.

Induction and repression represent the same phenomenon. In one case the bacterium adjusts its ability to use a given substrate (such as lactose) for growth; in the other it adjusts its ability to synthesize a particular metabolic intermediate (such as an essential amino acid). The trigger for either type of adjustment is the small molecule that is the substrate for the enzyme, or the product of the enzyme activity, respectively. Small molecules that cause the production of enzymes able t metabolize them are called inducers. Those that prevent the productio of enzymes able to synthesize them are called corepressors.

10.6 Repressor is controlled by a small molecule inducer

Key Concepts

- An inducer functions by converting the repressor protein into an inactive form.
- Repressor has two binding sites, one for the operator and another for the inducer.
- Repressor is inactivated by an allosteric interaction in which binding of inducer at its site changes the properties of the DNA-binding site.

The ability to act as inducer or corepressor is highly specific. Only the substrate/product or a closely related molecule can serve j *the activity of the small molecule does not depend on its interactio with the target enzyme.* Some inducers resemble the natural induce:

of the *lac* operon, but cannot be metabolized by the enzyme. The example *par excellence* is isopropylthiogalactoside (IPTG), one of several thiogalactosides with this property. Although it is not recognized by β-galactosidase, IPTG is a very efficient inducer of the *lac* genes.

Molecules that induce enzyme synthesis but are not metabolized are called gratuitous inducers. They are extremely useful because they remain in the cell in their original form. (A real inducer would be metabolized, interfering with study of the system.) The existence of gratuitous inducers reveals an important point. *The system must possess some component, distinct from the target enzyme, that recognizes the appropriate substrate; and its ability to recognize related potential substrates is different from that of the enzyme.*

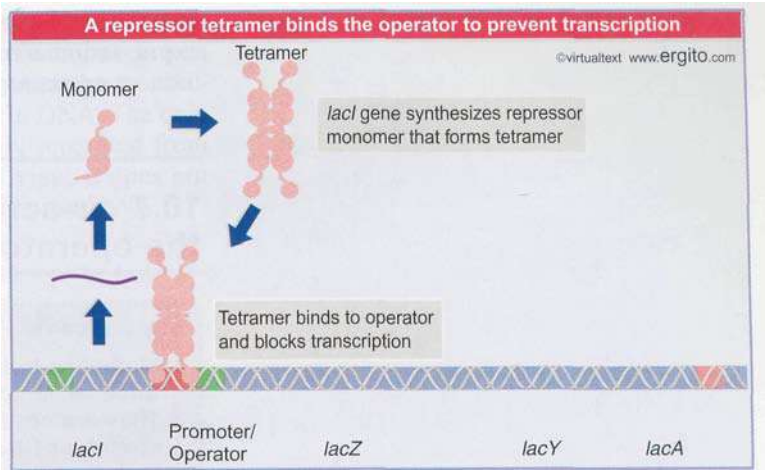


Figure 10.7 Repressor maintains the *lac* operon in the inactive condition by binding to the operator. The shape of the repressor is represented as a series of connected domains as revealed by its crystal structure.

The component that responds to the inducer is the repressor protein coded by *lacI*. The *lacZYA* structural genes are transcribed into a single mRNA from a promoter just upstream of *lacZ*. The state of the repressor determines whether this promoter is turned off or on:

- Figure 10.7 shows that in the absence of an inducer, the genes are not transcribed, because repressor protein is in an active form that is bound to the operator.
- Figure 10.8 shows that when an inducer is added, the repressor is converted into an inactive form that leaves the operator. Then transcription starts at the promoter and proceeds through the genes to a terminator located beyond the 3' end of *lacA*.

The crucial features of the control circuit reside in the dual properties of the repressor: it can prevent transcription; and it can recognize the small-molecule inducer. The repressor has two binding sites, one for the operator and one for the inducer. When the inducer binds at its site, it changes the conformation of the protein in such a way as to influence the activity of the operator-binding site. The ability of one site in the protein to control the activity of another is called allosteric control.

Induction accomplishes a coordinate regulation: *all the genes are expressed (or not expressed) in unison.* The mRNA is translated sequentially from its 5' end, which explains why induction always causes the appearance of β-galactosidase, β-galactoside permease, and β-galactoside transacetylase, in that order. Translation of a common mRNA explains why the relative amounts of the three enzymes always remain the same under varying conditions of induction.

Induction throws a switch that causes the genes to be transcribed. Inducers vary in their effectiveness, and other factors influence the absolute level of transcription or translation, but the relationship between the three genes is predetermined by their organization.

We notice a potential paradox in the constitution of the operon. The lactose operon contains the structural gene (*lacZ*) coding for the β-galactosidase activity needed to metabolize the sugar; it also includes the gene (*lacY*) that codes *for the protein needed to transport the substrate into* the cell. But if the operon is in a repressed state, how does the inducer enter the cell to start the process of induction?

Two features ensure that there is always a minimal amount of the protein present in the cell, enough to start the process off. There is a

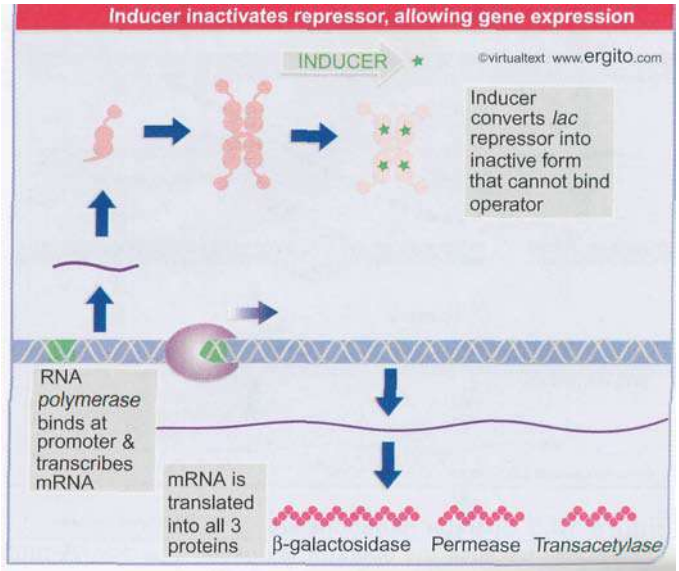


Figure 10.8 Addition of inducer converts repressor to an inactive form that cannot bind the operator. This allows RNA polymerase to initiate transcription.

basal level of expression of the operon: even when it is not induced, it is expressed at a residual level (0.1% of the induced level). And some inducer enters anyway via another uptake system.

10.7 *cis*-acting constitutive mutations identify the operator

Key Concepts

- Mutations in the operator cause constitutive expression of all three *lac* structural genes.
- They are c/s-acting and affect only those genes on the contiguous stretch of DNA.

Mutations in the regulatory circuit may either abolish expression of the operon or cause unregulated expression. Mutants that cannot be expressed at all are called uninducible. The continued expression of a gene that does not respond to regulation is called constitutive gene expression, and mutants with this property are called constitutive mutants.

Components of the regulatory circuit of the operon can be identified by mutations that *affect the expression of all the structural genes and map outside them.* They fall into two classes. The promoter and the operator are identified as targets for the regulatory proteins (RNA polymerase and repressor, respectively) by c/s-acting mutations. And the locus *lacI* is identified as the gene that codes for the repressor protein by mutations that eliminate the *trans-acting* product.

The operator was originally identified by constitutive mutations, denoted O^c, whose distinctive properties provided the first evidence for *an element that functions without being represented in a diffusible product.*

The structural genes contiguous with an O^c mutation are expressed constitutively because the mutation changes the operator so that the repressor no longer binds to it. So the repressor cannot prevent RNA polymerase from initiating transcription. The operon is transcribed constitutively, as illustrated in Figure 10.9.

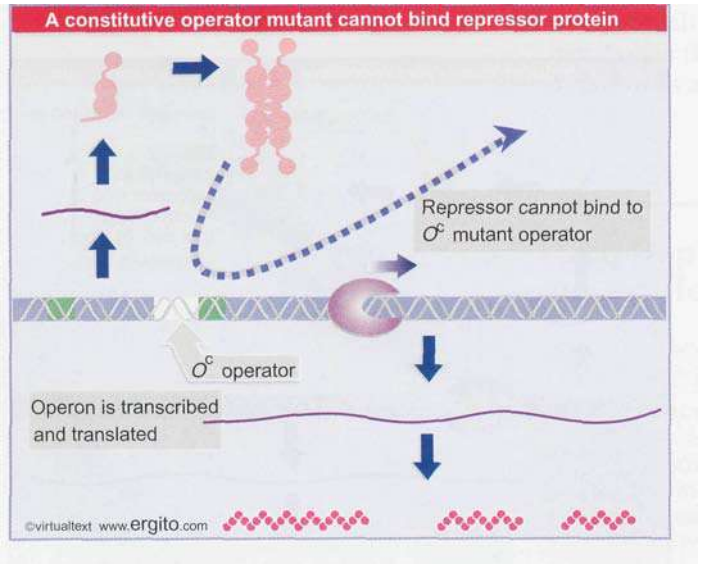


Figure 10.9 Operator mutations are constitutive because the operator is unable to bind repressor protein; this allows RNA polymerase to have unrestrained access to the promoter. The *CP* mutations are c/s-acting, because they affect only the contiguous set of structural genes.

The operator can control only the lac *genes that are adjacent to it.* If a second *lac* operon is introduced into the bacterium on an independent molecule of DNA, it has its own operator. Neither operator is influenced by the other. So if one operon has a wild-type operator, it will be repressed under the usual conditions, while a second operon with an O^c mutation will be expressed in its characteristic fashion.

These properties define the operator as a typical cis-acting site, whose function depends upon recognition of its DNA sequence by some *trans-acting* factor. The operator controls the adjacent genes irrespective of the presence in the cell of other alleles of the site. A mutation in such a site, for example, the O^c mutation, is formally described as cis-dominant.

A mutation in a *cis-acting* site cannot be assigned to a complementation group. (The ability to complement defines genes that are expressed as diffusible products.) When two c/s-acting sites lie close together—for example, a promoter and an operator—we cannot classify the mutations by a complementation test. We are restricted to distinguishing them by their effects on the phenotype.

cis-dominance is a characteristic of any site that is *physically contiguous with the sequences it controls.* If a control site functions as part of a polycistronic mRNA, mutations in it will display *exactly the same pattern* of *cis*-dominance as they would if functioning in DNA. The critical feature is that the control site cannot be physically separated from the genes that it regulates. From the genetic point of view, it does not matter whether the site and genes are together on DNA or on RNA.

10.8 trans-acting mutations identify the regulator gene

Key Concepts

- Mutations in the *lacI* gene are frans-acting and affect expression of all *lacZYA* clusters in the bacterium.
- Mutations that eliminate *lacI* function cause constitutive expression and are recessive.
- Mutations in the DNA-binding site of the repressor are constitutive because the repressor cannot bind the operator.
- Mutations in the inducer-binding site of the repressor prevent it from being inactivated and cause uninducibility.
- Mutations in the promoter are uninducible and *cis*-acting.

Constitutive transcription is also caused by mutations of the *lacI* type, which are caused by loss of function (including deletions of the gene). When the repressor is inactive or absent, transcription can initiate at the promoter. **Figure 10.10** shows that the $lacI^-$ mutants express the structural genes all the time (constitutively), *irrespective of whether the inducer is present or absent,* because the repressor is inactive.

The two types of constitutive mutations can be distinguished genetically. O^c mutants are *cis*-dominant, whereas $lacI^-$ mutants are recessive. This means that the introduction of a normal, $lacI^+$ gene restores control, irrespective of the presence of the defective $lacI^-$ gene.

Mutants of the operon that are uninducible fall into the same two types of genetic classes as the constitutive mutants:

- Promoter mutations are *cis*-acting. If they prevent RNA polymerase from binding at P_{lac}, they render the operon nonfunctional because it cannot be transcribed.
- Mutations that abolish the ability of repressor to bind the inducer are described as $lacI^s$. They are *trans-acting.* The repressor is "locked in" to the active form that recognizes the operator and prevents transcription. The addition of inducer has no effect because its binding site is absent, and therefore it is impossible to convert the repressor to the inactive form. The mutant repressor binds to all *lac* operators in the cell to prevent their transcription, and cannot be pried off, irrespective of the properties of any wild-type repressor protein that is present, so it is genetically dominant.

The two types of mutations in *lacI* can be used to identify the individual active sites in the repressor protein. The *DNA-binding site* recognizes the sequence of the operator. It is identified by constitutive point mutations that prevent repressor from binding to DNA to block RNA polymerase. The *inducer-binding site* is identified by point mutations that cause uninducibility, because inducer cannot bind to trigger the allosteric change in the DNA-binding site.

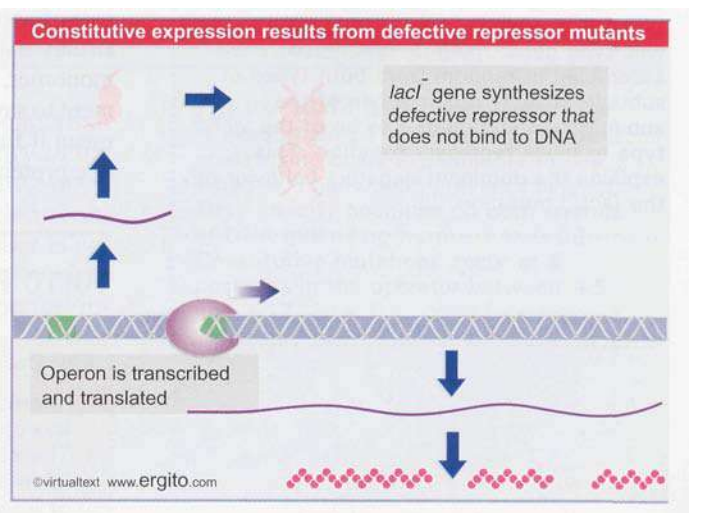


Figure 10.10 Mutations that inactivate the *lacI* gene cause the operon to be constitutively expressed, because the mutant repressor protein cannot bind to the operator.

10.9 Multimeric proteins have special genetic properties

Key Concepts

- Active repressor is a tetramer of identical subunits.
- When mutant and wild-type subunits are present, a single *lacI*$^{-d}$ mutant subunit can inactivate a tetramer whose other subunits are wild-type.
- *lacI*$^{-d}$ mutations occur in the DNA-binding site. Their effect is explained by the fact that repressor activity requires all DNA-binding sites in the tetramer to be active.

An important feature of the repressor is that it is multimeric. Repressor subunits associate at random in the cell to form the active protein tetramer. When two different alleles of the *lacI* gene are present, the subunits made by each can associate to form a heterotetramer, whose properties differ from those of either homotetramer. This type of interaction between subunits is a characteristic feature of multimeric proteins and is described as interallelic complementation.

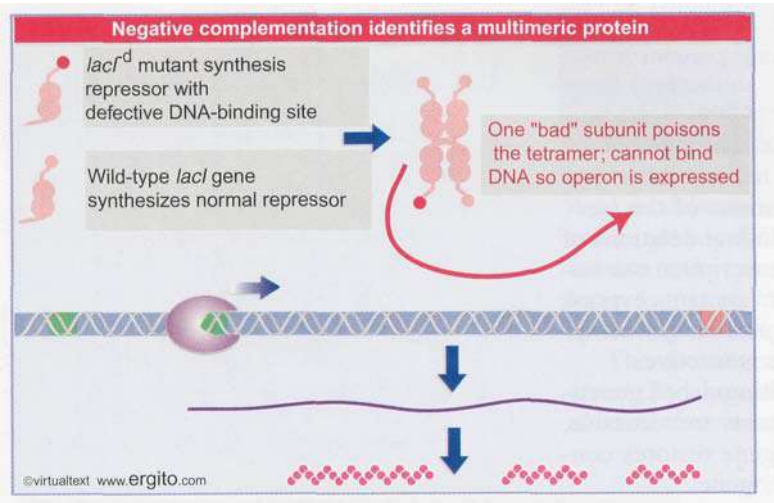


Figure 10.11 A *lacI*$^{-d}$ mutant gene makes a monomer that has a damaged DNA binding site (shown by the red circle). When it is present in the same cell as a wild-type gene, multimeric repressors are assembled at random from both types of subunits. It only requires one of the subunits of the multimer to be of the *lacI*$^{-d}$ type to block repressor function. This explains the dominant negative behavior of the *lacI*$^{-d}$ mutation.

Negative complementation occurs between some repressor mutants, as seen in the combination of *lacI*$^{-d}$ with *lacI*$^{+}$ genes. The *lacI*$^{-d}$ mutation alone results in the production of a repressor that cannot bind the operator, and is therefore constitutive like the *lacI*$^{-}$ alleles. Because the *lacI*$^{-}$ type of mutation inactivates the repressor, it is usually recessive to the wild type. However, the *-d* notation indicates that this variant of the negative type is dominant when paired with a wild-type allele. Such mutations are called dominant negative.

Figure **10.11** explains this phenomenon. The reason for the dominance is that the *lacI*$^{-d}$ allele produces a "bad" subunit, which is not only itself unable to bind to operator DNA, but is also able as part of a tetramer to prevent any "good" subunits from binding. This demonstrates that the repressor tetramer as a whole, rather than the individual monomer, is needed to achieve repression. In fact, we may reverse the argument to say that, whenever a protein has a dominant negative form, this must mean it functions as part of a multimer. The production of dominant negative proteins has become an important technique in eukaryotic genetics.

10.10 Repressor protein binds to the operator

Key Concepts

- Repressor protein binds to the double-stranded DNA sequence of the operator.
- The operator is a palindromic sequence of 26 bp.
- Each inverted repeat of the operator binds to the DNA-binding site of one repressor subunit.

The repressor was isolated originally by purifying the component able to bind the gratuitous inducer IPTG. (Because the amount of repressor in the cell is so small, in order to obtain enough material it

was necessary to use a promoter up mutation to increase *lacI* transcription, and to place this *lacI* locus on a DNA molecule present in many copies per cell. This results in an overall overproduction of 100-1000-fold.)

The repressor binds to double-stranded DNA containing the sequence of the wild-type *lac* operator. The repressor does not bind DNA from an O^c mutant. The addition of IPTG releases the repressor from operator DNA *in vitro.* The *in vitro* reaction between repressor protein and operator DNA therefore displays the characteristics of control inferred *in vivo;* so it can be used to establish the basis for repression.

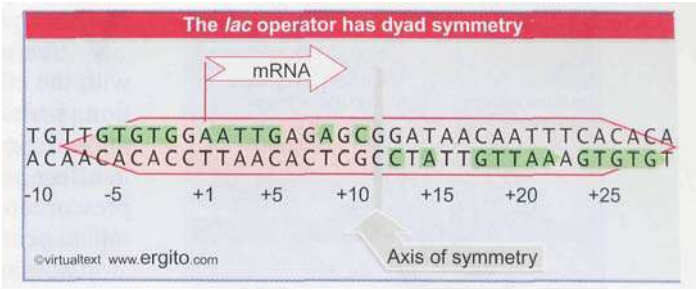


Figure 10.12 The *lac* operator has a symmetrical sequence. The sequence is numbered relative to the startpoint for transcription at +1. The pink arrows to left and right identify the two dyad repeats. The green blocks indicate the positions of identity.

How does the repressor recognize the specific sequence of operator DNA? The operator has a feature common to many recognition sites for bacterial regulator proteins: it is a palindrome. The inverted repeats are highlighted in Figure 10.12. Each repeat can be regarded as a half-site of the operator.

We can use the same approaches to define the points that the repressor contacts in the operator that we used for analyzing the polymerase-promoter interaction (see *9.14 RNA polymerase binds to one face of DNA).* Deletions of material on either side define the end points of the region; constitutive point mutations identify individual base pairs that must be crucial. Experiments in which DNA bound to repressor is compared with unbound DNA for its susceptibility to methylation or UV cross-linking identify bases that are either protected or more susceptible when associated with the protein.

Figure 10.13 shows that the region of DNA protected from nucleases by bound repressor lies within the region of symmetry, comprising the 26 bp region from -5 to +21. The area identified by constitutive mutations is even smaller. Within a central region extending over the 13 bp from +5 to +17, there are eight sites at which single base-pair substitutions cause constitutivity. This emphasizes the same point made by the promoter mutations summarized earlier in Figure 9.30. *A small number of essential specific contacts within a larger region can be responsible for sequence-specific association of DNA with protein.*

The symmetry of the DNA sequence reflects the symmetry in the protein. Each of the identical subunits in a repressor tetramer has a DNA-binding site. Two of these sites contact the operator in such a way that each inverted repeat of the operator makes the same pattern of contacts with a repressor monomer. This is shown by symmetry in the contacts that repressor makes with the operator (the pattern between +1 and +6 is identical with that between +21 and +16) and by matching constitutive mutations in each inverted repeat. (However, the operator is not perfectly symmetrical; the left side binds more strongly than the right side to the repressor. A stronger operator would be created by a perfect inverted duplication of the left side.)

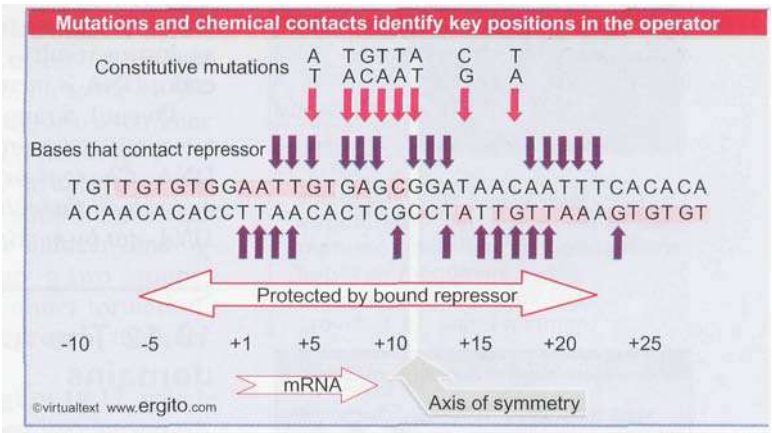


Figure 10.13 Bases that contact the repressor can be identified by chemical crosslinking or by experiments to see whether modification prevents binding. They identify positions on both strands of DNA extending from +1 to +23. Constitutive mutations occur at 8 positions in the operator between +5 and +17.

.11 Binding of inducer releases repressor m the operator

i Key Concepts

- Inducer binding causes a change in repressor conformation that reduces its affinity for DNA and releases it from the operator.

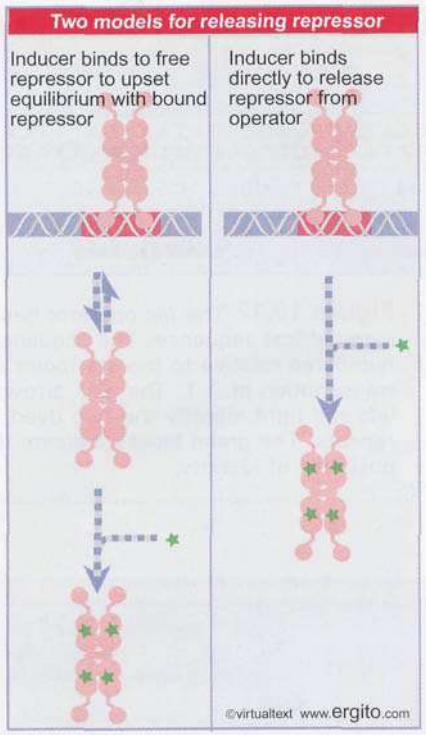


Figure 10.14 Does the inducer bind to free repressor to upset an equilibrium (left) or directly to repressor bound at the operator (right)?

Various inducers cause characteristic reductions in the affinity of the repressor for the operator *in vitro*. These changes correlate with the effectiveness of the inducers *in vivo*. This suggests that induction results from a reduction in the attraction between operator and repressor. So when inducer enters the cell, it binds to free repressors and in effect prevents them from finding their operators. But consider a repressor tetramer that is already bound tightly to the operator. How does inducer cause this repressor to be released?

Two models for repressor action are illustrated in **Figure 10.14:**

- The equilibrium model (left) calls for repressor bound to DNA to be in rapid equilibrium with free repressor. Inducer would bind to the free form of repressor, and thus unbalance the equilibrium by preventing reassociation with DNA.
- But the rate of dissociation of the repressor from the operator is much too slow to be compatible with this model (the half-life *in vitro* in the absence of inducer is >15 min). This means that instead the *inducer must bind directly to repressor protein complexed with the operator.* As indicated in the model on the right, inducer binding must produce a change in the repressor that makes it release the operator. Indeed, addition of IPTG causes an immediate destabilization of the repressor-operator complex *in vitro*.

Binding of the repressor-IPTG complex to the operator can be studied by using greater concentrations of the protein in the methylation protection/enhancement assay. The large amount compensates for the low affinity of the repressor-IPTG complex for the operator. The complex makes exactly the same pattern of contacts with DNA as the free repressor. An analogous result is obtained with mutant repressors whose affinity for operator DNA is increased; they too make the same pattern of contacts.

Overall, a range of repressor variants whose affinities for the operator span seven orders of magnitude all make the same contacts with DNA. *Changes in the affinity of the repressor for DNA must therefore occur by influencing the general conformation of the protein in binding DNA, not by making or breaking one or a few individual bonds.*

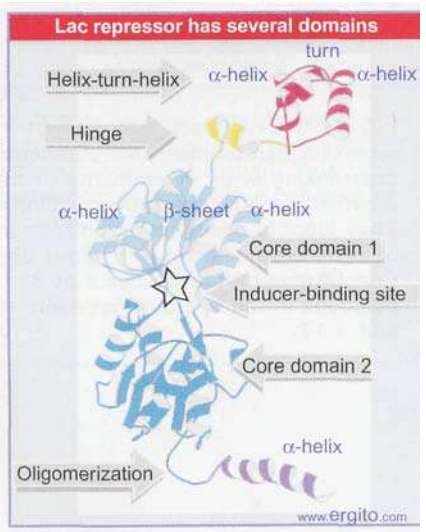


Figure 10.15 The structure of a monomer of Lac repressor identifies several independent domains. Photograph kindly provided by Mitchell Lewis.

10.12 The repressor monomer has several domains

Key Concepts

- A single repressor subunit can be divided into the N-terminal DNA-binding domain, a hinge, and the core of the protein.
- The DNA-binding domain contains two short α-helical regions that bind the major groove of DNA.
- The inducer-binding site and the regions responsible for multimerization are located in the core.

The repressor has several domains. The DNA-binding domain occupies residues 1-59. It is known as the headpiece. It can be cleaved from the remainder of the monomer, which is known as the *core*, by trypsin. The crystal structure illustrated in **Figure 10.15** offers a more detailed account of these regions.

The N-terminus of the monomer consists of two α-helices separated by a turn. This is a common DNA-binding motif, known as the HTH (helix-turn-helix); the two α-helices fit into the major groove of DNA, where they make contacts with specific bases (see *12.12 Repressor uses a helix-turn-helix motif to bind DNA*). This region is connected by a *hinge* to the main body of the protein. In the DNA-binding form of

repressor, the hinge forms a small α-helix (as shown in the figure); but when the repressor is not bound to DNA, this region is disordered. The HTH and hinge together correspond to the headpiece.

The bulk of the core consists of two regions with similar structures (core domains 1 and 2). Each has a six-stranded parallel β-sheet sandwiched between two α-helices on either side. The inducer binds in a cleft between the two regions.

At the C-terminus, there is an α-helix that contains two leucine heptad repeats. This is the oligomerization domain. The oligomerization helices of four monomers associate to maintain the tetrameric structure.

10.13 Repressor is a tetramer made of two dimers

Key Concepts

Monomers form a dimer by making contacts between core domain 2 and between the oligomerization helices.
Dimers form a tetramer by interactions between the oligomerization helices.

Figure **10.16** shows the structure of the tetrameric core (using a different modeling system from Figure 10.15). It consists in effect of two dimers. The body of the dimer contains a loose interface between the N-terminal regions of the core monomers, a cleft at which inducer binds, and a hydrophobic core (top). The C-terminal regions of each monomer protrude as parallel helices. (The headpiece would join on to the N-terminal regions at the top.) Together the dimers interact to form a tetramer (center) that is held together by a C-terminal bundle of 4 helices.

Sites of mutations are shown by beads on the structure at the bottom. *lacI*^{S} mutations map in two groups: yellow shows those that affect the dimer interface, and gray shows those in the inducer-binding cleft. *lacI* mutations that affect oligomerization map in two groups. White shows mutations in core domain 2 that prevent dimer formation. Purple *shows* those in the oligomerization helix that prevent tetramer formation from dimers.

From these data we can derive the schematic of **Figure 10.17,** which shows how the monomers are organized into the tetramer. Two monomers form a dimer by means of contacts at core domain 2 and in the oligomerization helix. The dimer has two DNA-binding domains at one end of the structure, and the oligomerization helices at the other end. Two dimers then form a tetramer by interactions at the oligomerization interface.

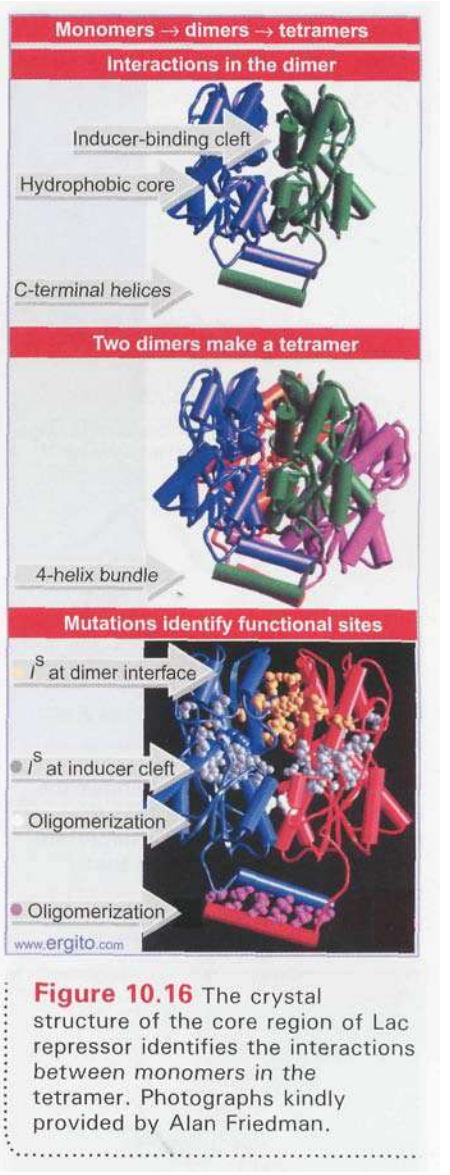


Figure 10.16 The crystal structure of the core region of Lac repressor identifies the interactions between monomers in the tetramer. Photographs kindly provided by Alan Friedman.

10.14 DNA-binding is regulated by an allosteric change in conformation

Key Concepts

- The DNA-binding domain of a monomer inserts into the major groove of DNA.
- Active repressor has a conformation in which the two DNA-binding domains of a dimer can insert into successive turns of the double helix.
- Inducer binding changes the conformation so that the two DNA-binding sites are not in the right geometry to make simultaneous contacts.

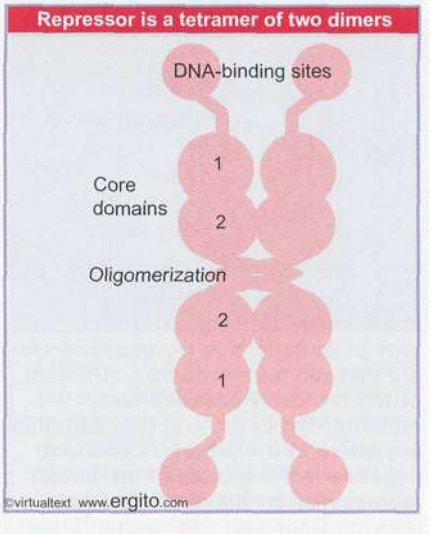


Figure 10.17 The repressor tetramer consists of two dimers.

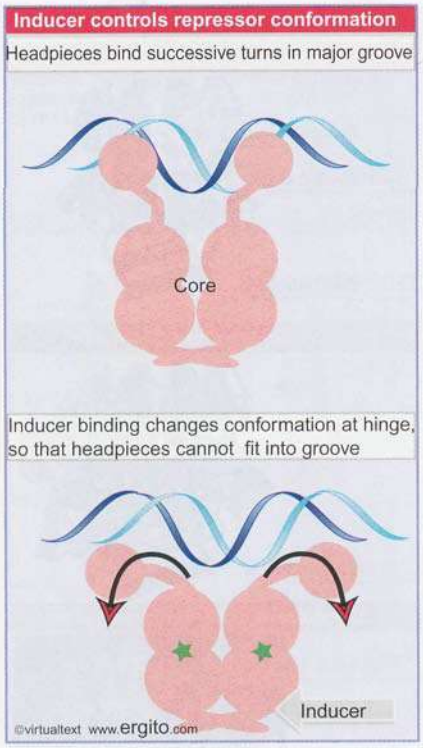


Figure 10.18 Inducer changes the structure of the core so that the headpieces of a repressor dimer are no longer in an orientation that permits binding to DNA.

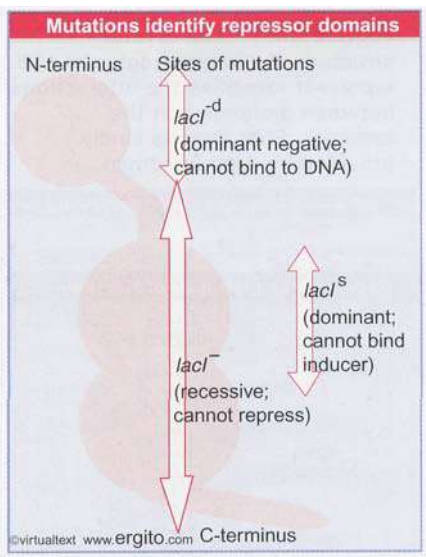


Figure 10.19 The locations of three type of mutations in lactose repressor are mapped on the domain structure of the protein. Recessive *lacI*$^-$ mutants that cannot repress can map anywhere in the protein. Dominant negative *lacI*$^{-d}$ mutants that cannot repress map to the DNA-binding domain. Dominant *lacI*^{s} mutants that cannot induce because they do not bind inducer map to core domain 1.

Early work suggested a model in which the headpiece is relatively independent of the core. It can bind to operator DNA by making the same pattern of contacts with a half-site as intact repressor. However, its affinity for DNA is many orders of magnitude less than that of intact repressor. The reason for the difference is that the dimeric form of intact repressor allows two headpieces to contact the operator simultaneously, each binding to one half-site. **Figure 10.18** shows that the two DNA-binding domains in a dimeric unit contact DNA by inserting into successive turns of the major groove. This enormously increases affinity for the operator.

Binding of inducer causes an immediate conformational change in the repressor protein. Binding of two molecules of inducer to the repressor tetramer is adequate to release repression. Binding of inducer changes the orientation of the headpieces relative to the core, with the result that the two headpieces in a dimer can no longer bind DNA simultaneously. This eliminates the advantage of the multimeric repressor, and reduces the affinity for the operator.

10.15 Mutant phenotypes correlate with the domain structure

Key Concepts

- Different types of mutations occur in different domains of the repressor subunit.

Mutations in the Lac repressor identified the existence of different domains even before the structure was known. We can now explain the nature of the mutations more fully by reference to the structure, as summarized in **Figure 10.19.**

Recessive mutations of the *lacI*$^-$ type can occur anywhere in the bulk of the protein. Basically any mutation that inactivates the protein will have this phenotype. The more detailed mapping of mutations on to the crystal structure in Figure 10.16 identifies specific impairments for some of these mutations; for example, those that affect oligomerization.

The special class of dominant-negative *lacI*$^{-d}$ mutations lie in the DNA-binding site of the repressor subunit (see *10.9 Multimeric proteins have special genetic properties).* This explains their ability to prevent mixed tetramers from binding to the operator; a reduction in the number of binding sites reduces the specific affinity for the operator. The role of the N-terminal region in specifically binding DNA is shown also by its location as the site of occurrence of "tight binding" mutations. These increase the affinity of the repressor for the operator, sometimes so much that it cannot be released by inducer. They are rare.

Uninducible *lacI*^{s} mutations map in a region of the core domain extending from the inducer-binding site to the hinge. One group lies in amino acids that contact the inducer, and these mutations function by preventing binding of inducer. The remaining mutations lie at sites that must be involved in transmitting the allosteric change in conformation to the hinge when inducer binds.

10.16 Repressor binds to three operators and interacts with RNA polymerase

Key Concepts

- Each dimer in a repressor tetramer can bind an operator, so that the tetramer can bind two operators simultaneously.
- Full repression requires the repressor to bind to an additional operator downstream or upstream as well as to the operator at *lacZ*.
- Binding of repressor at the operator stimulates binding of RNA polymerase at the promoter.

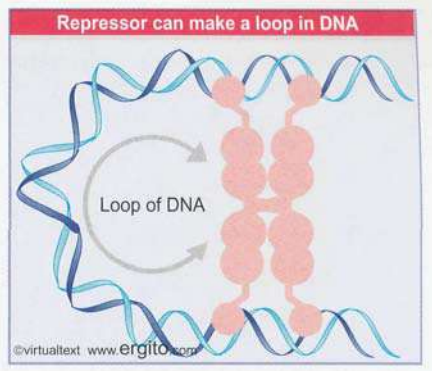


Figure 10.20 If both dimers in a repressor tetramer bind to DNA, the DNA between the two binding sites is held in a loop.

The allosteric transition that results from binding of inducer occurs in the repressor dimer. So why is a tetramer required to establish full repression?

Each dimer can bind an operator sequence. This enables the intact repressor to bind to two operator sites simultaneously. In fact, there are two further operator sites in the initial region of the *lac* operon. The original operator, *Ol*, is located just at the start of the *lacZ* gene. It has the strongest affinity for repressor. Weaker operator sequences (sometimes called pseudo-operators) are located on either side; *O2* is 410 bp downstream of the startpoint, and *03* is 83 bp upstream of it.

Figure 10.20 shows what happens when a DNA-binding protein can bind simultaneously to two separated sites on DNA. The DNA between the two sites forms a loop from a base where the protein has bound the two sites. The length of the loop depends on the distance between the two binding sites. When Lac repressor binds simultaneously to *01* and to one of the other operators, it causes the DNA between them to form a rather short loop, significantly constraining the DNA structure. A scale model for binding of tetrameric repressor to two operators is shown in **Figure 10.21.**

Binding at the additional operators affects the level of repression. Elimination of either the downstream operator (*O2*) or the upstream operator (03) reduces the efficiency of repression by 2-4X. However, if *both 02* and *03* are eliminated, repression is reduced 100×. *This suggests that the ability of the repressor to bind to one of the two other operators as well as to Ol is important for establishing repression.* We do not know how or why this simultaneous binding increases repression.

We know most about the direct effects of binding of repressor to the operator *(Ol)*. It was originally thought that repressor binding would occlude RNA polymerase from binding to the promoter. However, we now know that the two proteins may be bound to DNA simultaneously, and *the binding of repressor actually enhances the binding of RNA polymerase!* But the bound enzyme is prevented from initiating transcription.

The equilibrium constant for RNA polymerase binding alone to the *lac* promoter is $1.9 \times 10^7\ M^{-1}$. The presence of repressor increases this constant by two orders of magnitude to $2.5 \times 10^9\ M^{-1}$. In terms of the range of values for the equilibrium constant K_B given in Figure 9.20, repressor protein effectively converts the formation of closed complex by RNA polymerase at the *lac* promoter from a weak to a strong interaction.

What does this mean for induction of the operon? The higher value for K_B means that, when occupied by repressor, the promoter is 100 times more likely to be bound by an RNA polymerase. And by allowing RNA polymerase to be bound at the same time as repressor, it becomes possible for transcription to begin immediately upon induction, instead of waiting for an RNA polymerase to be captured.

The repressor in effect causes RNA polymerase to be stored at the promoter. The complex of RNA polymerase·repressor·DNA is blocked at the closed stage. When inducer is added, the repressor is released,

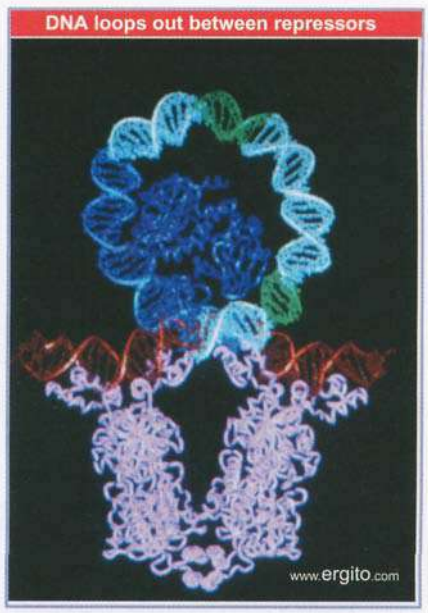


Figure 10.21 When a repressor tetramer binds to two operators, the stretch of DNA between them is forced into a tight loop. (The blue structure in the center of the looped DNA represents CAP, another regulator protein that binds in this region). Photograph kindly provided by Mitchell Lewis.

and the closed complex is converted to an open complex that initiates transcription. The overall effect of repressor has been to speed up the induction process.

Does this model apply to other systems? The interaction between RNA polymerase, repressor, and the promoter/operator region is distinct in each system, because the operator does not always overlap with the same region of the promoter (see Figure 10.26). For example, in phage lambda, the operator lies in the upstream region of the promoter, and binding of repressor occludes the binding of RNA polymerase (see *12 Phage strategies).* So a bound repressor does not interact with RNA polymerase in the same way in all systems.

10.17 Repressor is always bound to DNA

Key Concepts

- Proteins that have a high affinity for a specific DNA sequence also have a low affinity for other DNA sequences.
- Every base pair in the bacterial genome is the start of a low-affinity binding-site for repressor.
- The large number of low-affinity sites ensures that all repressor protein is bound to DNA.
- Repressor binds to the operator by moving from a low-affinity site rather than by equilibrating from solution.

Probably all proteins that have a high affinity for a specific sequence also possess a low affinity for any (random) DNA sequence. A large number of low-affinity sites will compete just as we for a repressor tetramer as a small number of high-affinity sites. Ther is only one high-affinity site in the *E. coli* genome: the operator. The re mainder of the DNA provides low-affinity binding sites. Every bas pair in the genome starts a new low-affinity site. (Just moving one bas pair along the genome, out of phase with the operator itself, creates low-affinity site!) So there are 4.2 X 10^6 low-affinity sites.

The large number of low-affinity sites means that, even in the ab sence of a specific binding site, all or virtually all repressor is bound t DNA; none is free in solution. **Figure 10.22** shows how the equation o describing the equilibrium between free repressor and DNA-bound repressor can be rearranged to give the proportion of free repressor.

Applying the parameters for the *lac* system, we find that

- The nonspecific equilibrium binding constant is K_A = *2* X 10^6 M^{-1}.
- The concentration of nonspecific binding sites is 4 X 10^6 in a bacterial volume of 10^{-15} liter, which corresponds to [DNA] = 7 X 10^{-3} M (a very high concentration).

Substituting these values gives: Free / Bound repressor = 10^{-4}.

So all but 0.01% of repressor is bound to (random) DNA. Since there are ~10 molecules of repressor per cell, this is tantamount to saying that there is no free repressor protein. This has an important implication for the interaction of repressor with the operator: it means that we are concerned with the *partitioning* of the repressor on DNA, in which the single high-affinity site of the operator *competes* with the large number of low-affinity sites.

In this competition, the absolute values of the association constants for operator and random DNA are not important; what is important is the ratio of K_{sp} (the constant for binding a specific site) to K_{nsp} (the constant for binding any random DNA sequence), that is, the specificity.

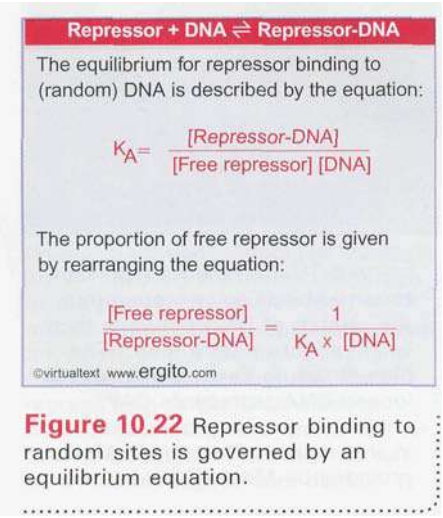


Figure 10.22 Repressor binding to random sites is governed by an equilibrium equation.

10.18 The operator competes with low-affinity sites to bind repressor

Key Concepts

- In the absence of inducer, the operator has an affinity for repressor that is $10^7\times$ that of a low affinity site.
- The level of 10 repressor tetramers per cell ensures that the operator is bound by repressor 96% of the time.
- Induction reduces the affinity for the operator to $10^4\times$ that of low-affinity sites, so that only 3% of operators are bound.
- Induction causes repressor to move from the operator to a low-affinity site by direct displacement.
- These parameters could be changed by a reduction in the effective concentration of DNA *in vivo.*

We can define the parameters that influence the ability of a regulator protein to saturate its target site by comparing the equilibrium equations for specific and nonspecific binding. As might be expected intuitively, the important parameters are as follows:

- The size of the genome dilutes the ability of a protein to bind specific target sites.
- The specificity of the protein counters the effect of the mass of DNA.
- The amount of protein that is required increases with the total amount of DNA in the genome and decreases with the specificity.
- The amount of protein also must be in reasonable excess of the total number of specific target sites, so we expect regulators with many targets to be found in greater quantities than regulators with fewer targets.

Figure 10.23 compares the equilibrium constants for *lac* repressor/operator binding with repressor/general DNA binding. From these constants, we can deduce how repressor is partitioned between the operator and the rest of DNA, and what happens to the repressor when inducer causes it to dissociate from the operator.

Repressor binds $\sim 10^7$ times better to operator DNA than to any random DNA sequence of the same length. So the operator comprises a single high-affinity site that will compete for the repressor 10^7 better than any low-affinity (random) site. How does this ensure that the repressor can maintain effective control of the operon?

Using the specificity, we can calculate the distribution between random sites and the operator, and can express this in terms of occupancy of the operator. If there are 10 molecules of *lac* repressor per cell with a specificity for the operator of 10^7, the operator will be bound by repressor 96% of the time. The role of specificity explains two features of the *lac* repressor-operator interaction:

- When inducer binds to the repressor, the affinity for the operator is reduced by $\sim 10^3$-fold. The affinity for general DNA sequences remains unaltered. So the specificity is now only 10^4, which is insufficient to capture the repressor against competition from the excess of 4.2×10^6 low-affinity sites. Only 3% of operators would be bound under these conditions.
- Mutations that reduce the affinity of the operator for the repressor by as little as 20-30X have sufficient effect to be constitutive. Within the genome, the mutant operators can be overwhelmed by the preponderance of random sites. The occupancy of the operator is reduced to ~50% if the repressor's specificity is reduced just 10X.

The consequence of these affinities is that in an uninduced cell, one tetramer of repressor usually is bound to the operator. All or almost all of

Repressor specifically binds operator DNA

DNA	Repressor	Repressor + inducer
Operator	2×10^{13}	2×10^{10}
Other DNA	2×10^6	2×10^6
Specificity	10^7	10^4

Figure 10.23 Lac repressor binds strongly and specifically to its operator, but is released by inducer. All equilibrium constants are in M^{-1}.

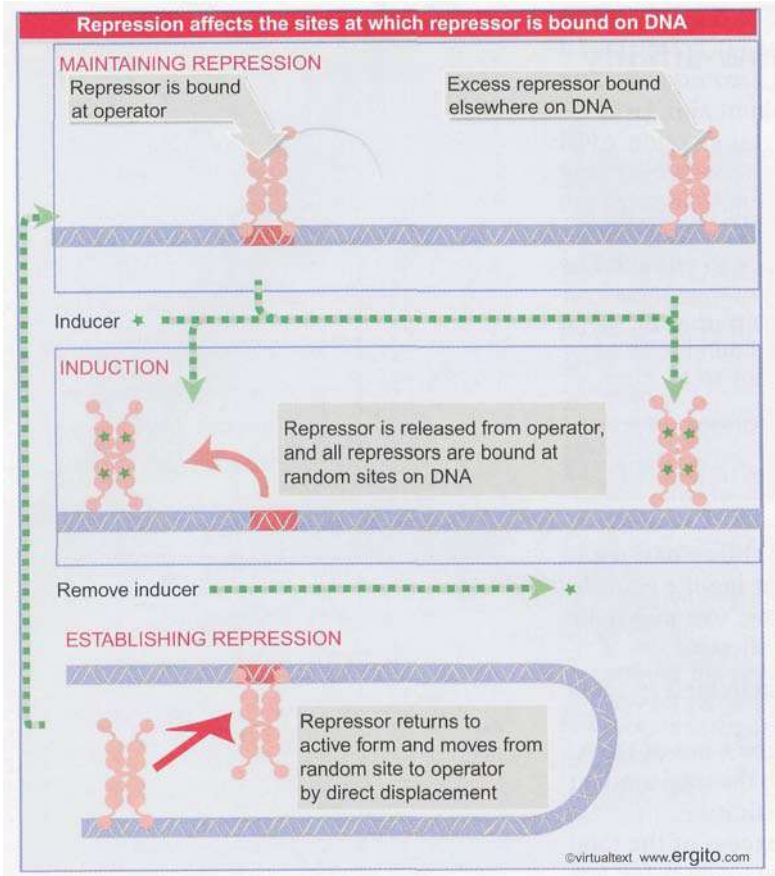


Figure 10.24 Virtually all the repressor in the cell is bound to DNA.

the remaining tetramers are bound at random to other regions of DNA, as illustrated in Figure 10.24. There are likely to be very few or no repressor tetramers free within the cell.

The addition of inducer abolishes the ability of repressor to bind specifically at the operator. Those repressors bound at the operator are released, and bind to random (low-affinity) sites. So in an induced cell, the repressor tetramers are "stored" on random DNA sites. In a noninduced cell, a tetramer is bound at the operator, while the remaining repressor molecules are bound to nonspecific sites. *The effect of induction is therefore to change the distribution of repressor on DNA, rather than to generate free repressor.*

When inducer is removed, repressor recovers its ability to bind specifically to the operator, and does so very rapidly. This must involve its movement from a nonspecific "storage" site on DNA. What mechanism is used for this rapid movement? The ability to bind to the operator very rapidly is not consistent with the time that would be required for multiple cycles of dissociation and reassociation with nonspecific sites on DNA. The discrepancy excludes random-hit mechanisms for finding the operator, suggesting that the repressor can move directly from a random site on DNA to the operator. This is the same issue that we encountered previously with the ability of RNA polymerase to find its promoters (see Figure 9.23 and Figure 9.24). The same solution is likely: movement could be accomplished by direct displacement from site to site (as indicated in Figure 10.24). A displacement reaction might be aided by the presence of more binding sites per tetramer (four) than are actually needed to contact DNA at any one time (two).

The parameters involved in finding a high-affinity operator in the face of competition from many low-affinity sites pose a dilemma for repressor. Under conditions of repression, there must be high specificity for the operator. But under conditions of induction, this specificity must be relieved. Suppose, for example, that there were 1000 molecules of repressor per cell. Then only 0.04% of operators would be free under conditions of repression. But upon induction only 40% of operators would become free. We therefore see an inverse correlation between the ability to achieve complete repression and the ability to relieve repression effectively. We assume that the number of repressors synthesized *in vivo* has been subject to selective forces that balance these demands.

The difference in expression of the lactose operon between its induced and repressed states *in vivo* is actually $10^3\times$. In other words, even when inducer is absent, there is a basal level of expression of ~0.1% of the induced level. This would be reduced if there were more repressor protein present, increased if there were less. So it could be impossible to establish tight repression if there were fewer repressors than the 10 found per cell; and it might become difficult to induce the operon if there were too many.

It is possible to introduce the *lac* operator-repressor system into the mouse. When the *lac* operator is connected to a tyrosinase reporter gene, the enzyme is induced by the addition of IPTG. This means that the repressor is finding its target in a genome 10^3 times larger than that of *E. coli*. Induction occurs at approximately the same concentration of IPTG as in bacteria. However, we do not know the concentration of Lac repressor and how effectively the target is induced.

In order to extrapolate *in vivo* from the affinity of a DNA-protein interaction *in vitro*, we need to know the effective *concentration* of DNA *in vivo*. The "effective concentration" differs from the mass/ volume because of several factors. The effective concentration is increased, for example, by molecular crowding, which occurs when polyvalent cations neutralize ~90% of the charges on DNA, and the nucleic acid collapses into condensed structures. The major force that decreases the effective concentration is the inaccessibility of DNA that results from occlusion or sequestration by DNA-binding proteins.

One way to determine the effective concentration is to compare the rate of a reaction *in vitro* and *in vivo* that depends on DNA concentration. This has been done using intermolecular recombination between two DNA molecules. To provide a control, the same reaction is followed as an intramolecular recombination, that is, the two recombining sites are presented on the same DNA molecule. We assume that concentration is the same *in vivo* and *in vitro* for the *intramolecular* reaction, and therefore any difference in the ratio of intermolecular/ intramolecular recombination rates can be attributed to a change in the effective concentration *in vivo*. The results of such a comparison suggest that the effective concentration of DNA is reduced >10-fold *in vivo*.

This could affect the rates of reactions that depend on DNA concentration, including DNA recombination, and protein-DNA binding. It emphasizes the problem encountered by all DNA-binding proteins in finding their targets with sufficient speed, and reinforces the conclusion that diffusion is not adequate (see Figure 9.23).

10.19 Repression can occur at multiple loci

[Key Concepts

A repressor will act on all loci that have a copy of its target operator sequence.

The *lac* repressor acts only on the operator of the *lacZYA* cluster. However, some repressors control dispersed structural genes by binding at more than one operator. An example is the *trp* repressor, which controls three unlinked sets of genes:

- operator at the cluster of structural genes *trpEDBCA* controls coordinate synthesis of the enzymes that synthesize tryptophan from chorismic acid.
- An operator at another locus controls the *aroH* gene, which codes for one of the three enzymes that catalyze the initial reaction in the common pathway of aromatic amino acid biosynthesis.

The *trpR* regulator gene is repressed by its own product, the *trp* repressor. So the repressor protein acts to reduce its own synthesis.

Operators for TrpR have related sequences

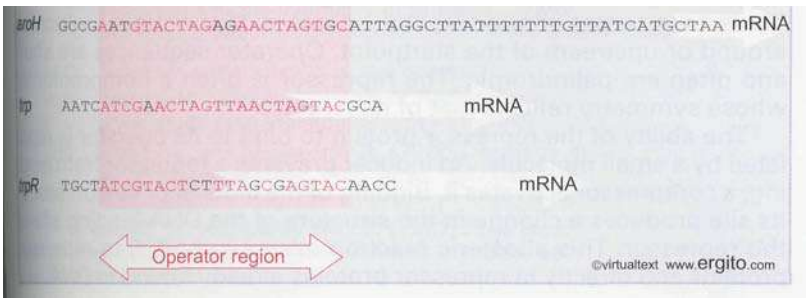


Figure 10.25 The *trp* repressor recognizes operators at three loci. Conserved bases are shown in red. **The** location of the startpoint and mRNA varies, as indicated by the white arrows.

By Book_Crazy [IND]

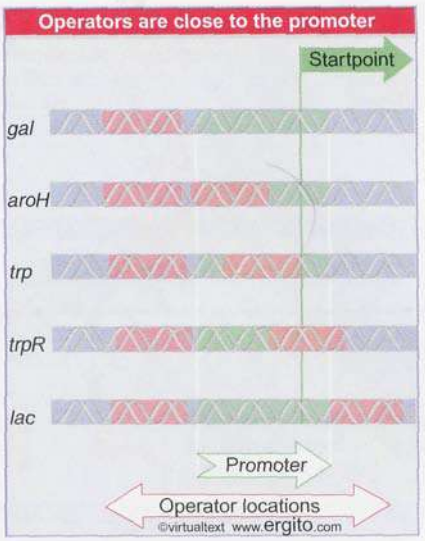


Figure 10.26 Operators may lie at various positions relative to the promoter.

This circuit is an example of autogenous control. Such circuits are quite common in regulatory genes, and may be either negative or positive (see *11.11 r-protein synthesis is controlled by autogenous regulation* and *12.9 Repressor maintains an autogenous circuit).*

A related 21 bp operator sequence is present at each of the three loci at which the *trp* repressor acts. The conservation of sequence is indicated in **Figure 10.25**. Each operator contains appreciable (but not identical) dyad symmetry. The features conserved at all three operators include the important points of contact for *trp* repressor. This explains how one repressor protein acts on several loci: *each locus has a copy of a specific DNA-binding sequence recognized by the repressor* (just as each promoter shares consensus sequences with other promoters).

Figure 10.26 summarizes the variety of relationships between operators and promoters. A notable feature of the dispersed operators recognized by TrpR is their presence at different locations within the promoter in each locus. In *trpR* the operator lies between positions -12 and +9, while in the *trp* operon it occupies positions -23 to -3, but in the *aroH* locus it lies farther upstream, between -49 and -29. In other cases, the operator lies downstream from the promoter (as in *lac*), or apparently just upstream of the promoter (as in *gal*, where the nature of the repressive effect is not quite clear). The ability of the repressors to act at operators whose positions are different in each target promoter suggests that there could be differences in the exact mode of repression, the common feature being that RNA polymerase is prevented from initiating transcription at the promoter.

10.20 Summary

Transcription is regulated by the interaction between *trans*-acting factors and c/s-acting sites. A *trans*-acting factor is the product of a regulator gene. It is usually protein but can be RNA. Because it diffuses in the cell, it can act on any appropriate target gene. A c/s-acting site in DNA (or RNA) is a sequence that functions by being recognized *in situ.* It has no coding function and can regulate only those sequences that are physically contiguous with it. Bacterial genes coding for proteins whose functions are related, such as successive enzymes in a pathway, may be organized in a cluster that is transcribed into a polycistronic mRNA from a single promoter. Control of this promoter regulates expression of the entire pathway. The unit of regulation, containing structural genes and c/s-acting elements, is called the operon.

Initiation of transcription is regulated by interactions that occur in the vicinity of the promoter. The ability of RNA polymerase to initiate at the promoter is prevented or activated by other proteins. Genes that are active unless they are turned off are said to be under negative control. Genes that are active only when specifically turned on are said to be under positive control. The type of control can be determined by the dominance relationships between wild type and mutants that are constitutive/derepressed (permanently on) or uninducible/super-repressed (permanently off).

A repressor protein prevents RNA polymerase either from binding to the promoter or from activating transcription. The repressor binds to a target sequence, the operator, that usually is located around or upstream of the startpoint. Operator sequences are short and often are palindromic. The repressor is often a homomultimer whose symmetry reflects that of its target.

The ability of the repressor protein to bind to its operator is regulated by a small molecule. An inducer prevents a repressor from binding; a corepressor activates it. Binding of the inducer or corepressor to its site produces a change in the structure of the DNA-binding site of the repressor. This allosteric reaction occurs in both free repressor proteins and directly in repressor proteins already bound to DNA.

The lactose pathway operates by induction, when an inducer β-galactoside prevents the repressor from binding its operator; transcription and translation of the *lacZ* gene then produce β-galactosidase, the enzyme that metabolizes β-galactosides. The tryptophan pathway operates by repression; the corepressor (tryptophan) activates the repressor protein, so that it binds to the operator and prevents expression of the genes that code for the enzymes that biosynthesize tryptophan. A repressor can control multiple targets that have copies of an operator consensus sequence.

A protein with a high affinity for a particular target sequence in DNA has a lower affinity for all DNA. The ratio defines the specificity of the protein. Because there are many more nonspecific sites (any DNA sequence) than specific target sites in a genome, a DNA-binding protein such as a repressor or RNA polymerase is "stored" on DNA; probably none or very little is free. The specificity for the target sequence must be great enough to counterbalance the excess of nonspecific sites over specific sites. The balance for bacterial proteins is adjusted so that the amount of protein and its specificity allow specific recognition of the target in "on" conditions, but allow almost complete release of the target in "off" conditions.

References

10.1 Introduction

ref Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318-389.

10.2 Regulation can be negative or positive

rev Miller, J. and Reznikoff, W. (1978). The Operon. Cold Spring Harbor Symp. Quant. Biol.

10.4 The *lac* genes are controlled by a repressor

rev Barkleya, M. D. and Bourgeois, S. (1978). Repressor recognition of operator and effectors. In The Operon, Eds. Miller, J. and Reznikoff, W. Cold Spring Harbor Laboratory, New York 177-220.

Beckwith, J. (1978). *lac:* the genetic system. In The Operon, Eds. Miller, J. and Reznikoff, W. Cold Spring Harbor Laboratory, New York 11-30.

Beyreuther, K. (1978). Chemical structure and functional organization of lac repressor from *E. coli.* In The Operon, Eds. Miller, J. and Reznikoff, W. Cold Spring Harbor Laboratory, New York 123-154.

Miller, J. H. (1978). The *lacI* gene: its role in lac operon control and its use as a genetic system. In The Operon, Eds. Miller, J. and Reznikoff, W. Cold Spring Harbor Laboratory, New York 31-88.

Weber, K. and Geisler, N. (1978). Lac repressor fragments produced *in vivo* and *in vitro:* an approach to the understanding of the interaction of repressor and DNA. In The Operon, Eds. Miller, J. and Reznikoff, W. Cold Spring Harbor Laboratory, New York 155-176.

ref Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318-389.

10.10 Repressor protein binds to the operator

exp Ptashne, M. (2002). Isolation of Repressor (www.ergito.com/lookup.jsp?expt = ptashne)

ref Gilbert, W. and Muller-Hill, B. (1966). Isolation of the lac repressor. Proc. Nat. Acad. Sci. USA 56, 1891-1898.

Gilbert, W. and Muller-Hill, B. (1967). The lac operator is DNA. Proc. Nat. Acad. Sci. USA 58, 2415-2421.

10.12 The repressor monomer has several domains

ref Friedman, A. M., Fischmann, T. 0., and Steitz, T. A. (1995). Crystal structure of lac repressor core tetramer and its implications for DNA looping. Science 268, 1721-1727.

Lewis, M. et al. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 271, 1247-1254.

10.15 Mutant phenotypes correlate with the domain structure

rev Pace, H. C, Kercher, M. A., Lu, P., Markiewicz, P., Miller, J. H., Chang, G., and Lewis, M. (1997). Lac repressor genetic map in real space. Trends Biochem. Sci. 22, 334-339.

ref Markiewicz, P., Kleina, L. G., Cruz, C, Ehret, S., and Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *E. coli lac* repressors reveals essential and non-essential residues, as well as spacers which do not require a specific sequence. J. Mol. Biol. 240, 421-433.

Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. (1996). Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J. Mol. Biol. 261, 509-523.

10.16 Repressor binds to three operators and interacts with RNA polymerase

ref Oehler, S. (1990). The three operators of the lac operon cooperate in repression. EMB0 J. 9, 973-979.

10.18 The operator competes with low-affinity sites to bind repressor

ref Cronin, C. A., Gluba, W., and Scrable, H. (2001). The lac operator-repressor system is functional in the mouse. Genes Dev. 15, 1506-1517.

Hildebrandt, E. R. et al. (1995). Comparison of recombination *in vitro* and in *E. coli* cells: measure of the effective concentration of DNA in vivo. Cell 81, 331-340.

Lin, S.-y. and Riggs, A. D. (1975). The general affinity of lac repressor for *E. coli* DNA: implications for gene regulation in prokaryotes and eukaryotes. Cell 4, 107-111.

Chapter 11

Regulatory circuits

11.1 Introduction

The basic concept of genetic regulation in bacteria is that the expression of a gene may be controlled by a regulator that interacts with a specific sequence or structure in DNA or mRNA at some stage prior to the synthesis of protein. This is an extremely flexible idea with many ramifications. The stage of expression that is controlled can involve transcription, when the target for regulation is DNA; or it can be at translation, when the target for regulation is RNA. When transcription is involved, the level of control can be at initiation or at termination. The regulator can be a protein or an RNA. "Controlled" can mean that the regulator turns off (represses) the target or that it turns on (activates) the target. Expression of many genes can be coordinately controlled by a single regulator gene on the principle that each target contains a copy of the sequence or structure that the regulator recognizes. Regulators may themselves be regulated, most typically in response to small molecules whose supply responds to environmental conditions. Regulators may be controlled by other regulators to make complex circuits.

Let's compare the ways that different types of regulators work.

Protein regulators work on the principle of allostery. The protein has two binding sites, one for a nucleic acid target, the other for a small molecule. Binding of the small molecule to its site changes the conformation in such a way as to alter the affinity of the other site for the nucleic acid. The way in which this happens is known in detail for the Lac repressor (see *10.14 DNA-binding is regulated by an allosteric change in conformation).* Protein regulators are often multimeric, with a symmetrical organization that allows two subunits to contact a palindromic target on DNA. This can generate cooperative binding effects that create a more sensitive response to regulation.

RNA regulators use changes in secondary structure as the guiding principle. An RNA regulator recognizes its target by the familiar principle of complementary base pairing. Figure 11.1 shows that the regulator is usually a small RNA molecule with extensive secondary structure, but with a single-stranded region(s) that is complementary to a single-stranded region in its target. The formation of a double helical region between regulator and target can have two types of consequence:

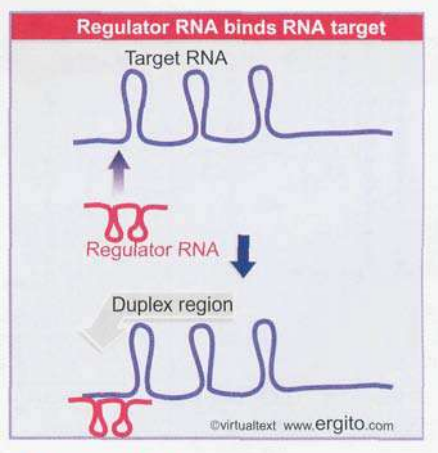


Figure 11.1 A regulator RNA is a small RNA with a single-stranded region that can pair with a single-stranded region in a target RNA.

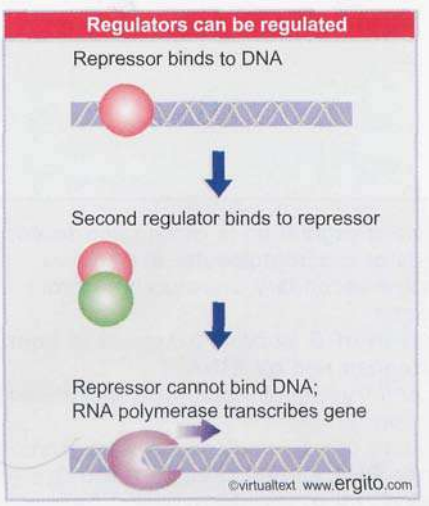


Figure 11.2 A protein that antagonizes a repressor can turn a gene on.

- Formation of the double helical structure may itself be sufficient. In some cases, protein(s) can bind only to the single-stranded form of the target sequence, and are therefore prevented by acting by duplex formation. In other cases, the duplex region becomes a target for binding, for example, by nucleases that degrade the RNA and therefore prevent its expression.
- Duplex formation may be important because it sequesters a region of the target RNA that would otherwise participate in some alternative secondary structure.

Going beyond the interactions in which a protein or RNA regulates expression of a single gene, we find that bacteria have responses in which the expression of many genes is coordinated. The simplest form of regulating multiple genes occurs when they all have a copy of the same *cis*-acting regulatory elements. Genes that have a copy of the same operator sequence are coordinately repressed by a single repressor protein. Genes that have the same type of promoter are coordinately activated by the production of a sigma factor that causes RNA polymerase to use that promoter.

Regulatory networks afe created when one regulator is required for the production of another. This happens in situations in which an ordered temporal expression of genes is required, for example, during phage infection (see *12.3 Lytic development is controlled by a cascade)* or during the development of a new cell type, such as a spore (see *9.18 Sigma factors may be organized into cascades).* These use one of the simplest relationships between regulators, in which one regulator is necessary for expression of the next regulator in a series, creating a cascade.

Another type of relationship occurs when one regulator directly regulates the activity of another regulator. For example, a protein that represses or activates expression of a gene may itself by inhibited by an "anti-regulator" that responds to some other signal. Figure 11.2 shows an example. A series of such relationships can be extended indefinitely. Some circuits are controlled by a series of regulators each of which antagonizes another. Such circuits allow the cell to control the target set of genes in response to multiple stimuli, since each stimulus can feed into the regulatory circuit at a different point.

A special type of circuit is created when a protein regulates expression of the gene that codes for it. This is called autogenous control. It allows a protein to regulate its own level of expression without reference to any other circuit. It is often used for regulating levels of proteins that are assembled into macromolecular complexes.

11.2 Distinguishing positive and negative control

Key Concepts

- Induction can be achieved by inactivating a repressor or activating an activator.
- Repression can be achieved by activating a repressor or inactivating an activator.

Positive and negative control systems are defined by the response of the operon when no regulator protein is present. The characteristics of the two types of control system are mirror images.

Genes under negative control are expressed unless they an switched off by a repressor protein (see Figure 10.2). Any action that interferes with gene expression can provide a negative control.

Typically a repressor protein either binds to DNA to prevent RNA polymerase from initiating transcription, or binds to mRNA to prevent a ribosome from initiating translation.

Negative control provides a fail-safe mechanism: if the regulator protein is inactivated, the system functions and so the cell is not deprived of these enzymes. It is easy to see how this might evolve. Originally a system functions constitutively, but then cells able to interfere specifically with its expression acquire a selective advantage by virtue of their increased efficiency.

For genes under positive control, expression is possible only when an active regulator protein is present. The mechanism for controlling an individual operon is an exact counterpart of negative control, but instead of *interfering* with initiation, the regulator protein is *essential* for it. It interacts with DNA and with RNA polymerase to *assist the initiation event* (see Figure 10.3). The use of sigma factors to regulate transcription formally is an example of positive control. A positive regulator protein that responds to a small molecule is usually called an activator.

It is less obvious how positive control evolved, since the cell must have had the ability to express the regulated genes even before any control existed. Presumably some component of the control system must have changed its role. Perhaps originally it was used as a regular part of the apparatus for gene expression; then later it became restricted to act only in a particular system or systems.

Operons are defined as inducible or repressible by the nature of their response to the small molecule that regulates their expression. Inducible operons function only in the *presence* of the small-molecule inducer. Repressible operons function only in the *absence* of the small-molecule corepressor (so called to distinguish it from the repressor protein).

The terminology used for repressible systems describes the active state of the operon as derepressed; this has the same meaning as *induced.* The condition in which a (mutant) operon cannot be derepressed is sometimes called super-repressed; this is the exact counterpart of *uninducible.*

Either positive or negative control could be used to achieve either induction or repression by utilizing appropriate interactions between the regulator protein and the small-molecule inducer or corepressor. Figure 11.3 summarizes four simple types of control circuit. Induction is achieved when an inducer inactivates a repressor protein or activates an activator protein. Repression is accomplished when a corepressor activates a repressor protein or inactivates an activator protein.

The *trp* operon is a repressible system. Tryptophan is the end product of the reactions catalyzed by a series of biosynthetic enzymes. Both the activity and the synthesis of the tryptophan enzymes are controlled by the level of tryptophan in the cell.

Tryptophan functions as a corepressor that activates a repressor protein. This is the classic mechanism for repression, as seen in Figure 11.3 (lower left). In conditions when the supply of tryptophan is plentiful, the operon is repressed because the repressor protein·corepressor complex is bound at the operator. When tryptophan is in short supply, the corepressor is inactive, therefore has reduced specificity for the operator, and is stored elsewhere on DNA.

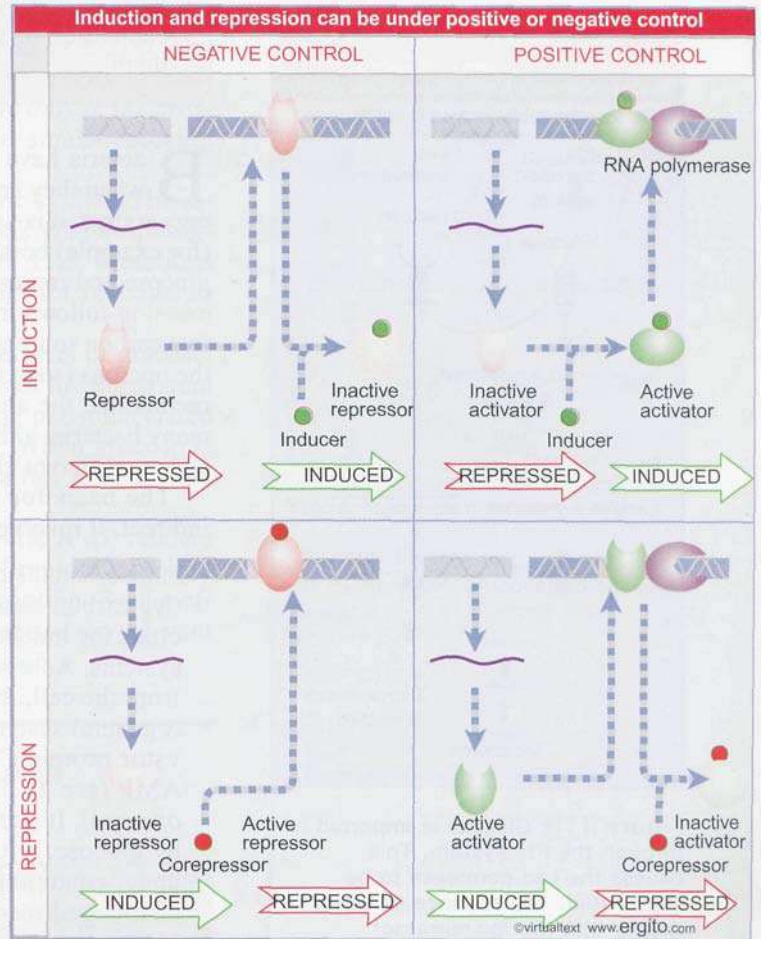


Figure 11.3 Control circuits are versatile and can be designed to allow positive or negative control of induction or repression.

Deprivation of repressor causes ~70-fold increase in the frequency of initiation events at the *trp* promoter. Even under repressing conditions, the structural genes continue to be expressed at a low basal level (sometimes also called the repressed level). The efficiency of repression at the operator is much lower than in the *lac* operon (where the basal level is only ~ 1/1000 of the induced level).

We have treated both induction and repression as phenomena that rely upon allosteric changes induced in regulator proteins by small molecules. Other types of interactions also can be used to control the activities of regulator proteins. One example is OxyR, a transcriptional activator of genes induced by hydrogen peroxide. The OxyR protein is directly activated by oxidation, so it provides a sensitive measure of oxidative stress. Another common type of signal is phosphorylation of a regulator protein.

11.3 Glucose repression controls use of carbon sources

Key Concepts

- *E. coli* uses glucose in preference to other carbon sources when it has a choice.
- Glucose prevents uptake of alternative carbon sources from the medium.
- Exclusion of the alternative carbon sources from the cell prevents expression of the operons coding for the enzymes that metabolize them.

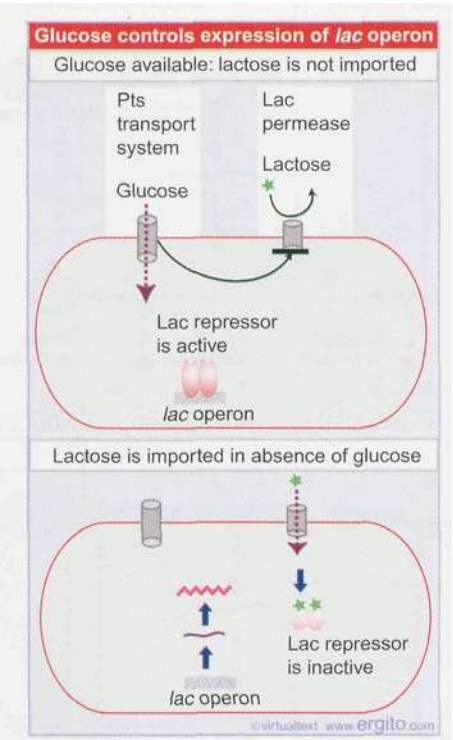


Figure 11.4 Glucose is imported through the Pts system. This causes the Lac permease to be inactivated. The absence of lactose means that the *lac* repressor switches off the operon. When glucose is absent, lactose can be imported, and inactivates the *lac* repressor, so the operon is switched on.

Bacteria have distinct preferences among potential carbon sources when they are offered a choice. When glucose is available as an energy source, it is used in preference to other sugars. So when *E. coli* finds (for example) both glucose and lactose in the medium, it metabolizes the glucose and represses the use of lactose. The phenomenon of glucose repression follows from the ability of glucose to prevent the use of alternative carbon sources. It describes the general repression of transcription of the operons (such as *lac*, *gal*, and *ara)* coding for the enzymes required to metabolize the alternative carbon sources. The same effect is found in many bacteria, although different molecular mechanisms may be responsible in different classes of bacteria.

The basis for glucose repression in *E. coli* appears to be largely indirect. It involves two mechanisms:

- Inducer exclusion describes the ability of glucose to prevent alternative carbon sources from being taken up from the medium. These include the inducers of the operons coding for the alternative metabolic systems. As a result of the exclusion of the small molecule inducers from the cell, the operons cannot be transcribed.
- A general system for activating many operons is provided by the activator protein CRP, which is activated by the small molecule cyclic AMP (see *11.5 CRP functions in different ways in different target operons).* It was thought for many years that this system is inactivated by glucose, with the result that its target operons are not expressed under conditions of glucose repression. However, this view has been challenged recently, and the exact way in which it is controlled by glucose is not clear.

Figure 11.4 outlines the interactions involved in inducer exclusion. The key molecular component in inducer exclusion is the PTS (phosphoenolpyruvate:glycose phosphotransferase system), a complex of proteins in the bacterial membrane, which simultaneously phosphorylates

and transports sugars into the cell. One of the proteins of this complex (IIA^{Glc}, which is coded by the *err* gene) becomes dephosphorylated as a result of glucose transport. It then binds to the Lac permease and prevents it from importing lactose into the cell.

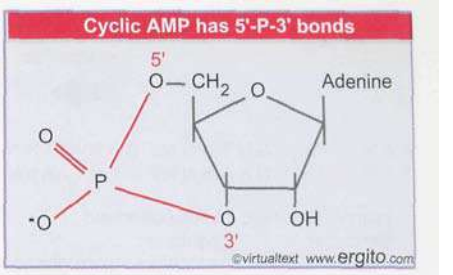


Figure 11.5 Cyclic AMP has a single phosphate group connected to both the 3′ and 5′ positions of the sugar ring.

11.4 Cyclic AMP is an inducer that activates CRP to act at many operons

Key Concepts

- CRP is an activator protein that binds to a target sequence at a promoter.
- A dimer of CRP is activated by a single molecule of cyclic AMP.

So far we have dealt with the promoter as a DNA sequence that is competent to bind RNA polymerase, which then initiates transcription. But there are some promoters at which RNA polymerase cannot initiate transcription without assistance from an ancillary protein. Such proteins are positive regulators, because their presence is necessary to switch on the transcription unit. Typically the activator overcomes a deficiency in the promoter, for example, a poor consensus sequence at −35 or −10.

One of the most widely acting activators is a protein called CRP activator that controls the activity of a large set of operons in *E. coli*. The protein is a positive control factor whose presence is necessary to initiate transcription at dependent promoters. CRP is active *only in the presence of cyclic AMP*, which behaves as the classic small-molecule inducer (see Figure 11.3, upper right).

Cyclic AMP is synthesized by the enzyme adenylate cyclase. The reaction uses ATP as substrate and introduces a 3′–5′ link via phosphodiester bonds, generating the structure drawn in **Figure 11.5.** Mutations in the gene coding for adenylate cyclase (*cya*⁻) do not respond to changes in glucose levels.

The level of cyclic AMP is inversely related to the level of glucose. The basis for this effect lies with the same component of the Pts system that is responsible for controlling lactose uptake. The phosphorylated form of protein IIA^{Glc} stimulates adenylate cyclase. When glucose is imported, the dephosphorylation of IIA^{Glc} leads to a fall in adenylate cyclase activity.

Figure 11.6 shows that reducing the level of cyclic AMP renders the (wild-type) protein unable to bind to the control region, which in turn prevents RNA polymerase from initiating transcription. So the effect of glucose in reducing cyclic AMP levels is to deprive the relevant operons of a control factor necessary for their expression.

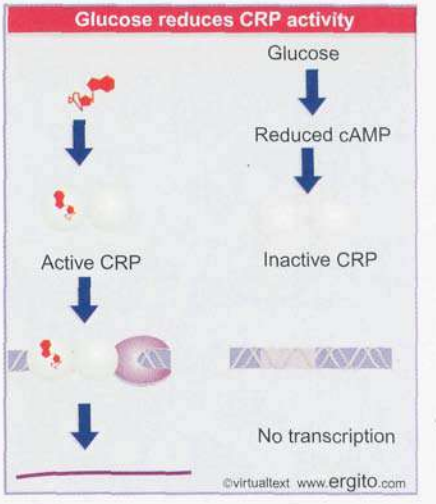


Figure 11.6 By reducing the level of cyclic AMP, glucose inhibits the transcription of operons that require CRP activity.

11.5 CRP functions in different ways in different target operons

Key Concepts

- CRP-binding sites lie at highly variable locations relative to the promoter.
- CRP interacts with RNA polymerase, but the details of the interaction depend on the relative locations of the CRP-binding site and the promoter.

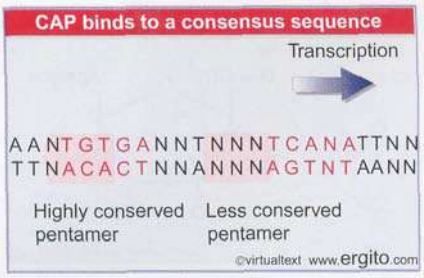


Figure 11.7 The consensus sequence for CRP contains the well conserved pentamer TGTGA and (sometimes) an inversion of this sequence (TCANA).

The CRP factor binds to DNA, and complexes of cyclic AMP·CRP·DNA can be isolated at each promoter at which it functions. The factor is a dimer of two identical subunits of 22.5 kD, which can be activated by a single molecule of cyclic AMP. A CRP monomer contains a DNA-binding region and a transcription-activating region.

A CRP dimer binds to a site of ~22 bp at a responsive promoter. The binding sites include variations of the consensus sequence given in **Figure 11.7.** Mutations preventing CRP action usually are located within the well conserved pentamer,

TGTGA
ACACT

which appears to be the essential element in recognition. CRP binds most strongly to sites that contain two (inverted) versions of the pentamer, because this enables both subunits of the dimer to bind to the DNA. Many binding sites lack the second pentamer, however, and in these the second subunit must bind a different sequence (if it binds to DNA). The hierarchy of binding affinities for CRP helps to explain why different genes are activated by different levels of cyclic AMP *in vivo*.

The action of CRP has the curious feature that its binding sites lie at different locations relative to the startpoint in the various operons that it regulates. The TGTGA pentamer may lie in either orientation. The three examples summarized in **Figure 11.8** encompass the range of locations:

- The CRP-binding site is adjacent to the promoter, as in the *lac* operon, in which the region of DNA protected by CRP is centered on –61. It is possible that two dimers of CRP are bound. The binding pattern is consistent with the presence of CRP largely on one face of DNA, the same face that is bound by RNA polymerase. This location would place the two proteins just about within reach of each other.
- Sometimes the CRP-binding site lies within the promoter, as in the *gal* locus, where the CRP-binding site is centered on –41. It is likely that only a single CRP dimer is bound, probably in quite intimate contact with RNA polymerase, since the CRP-binding site extends well into the region generally protected by the RNA polymerase.
- In other operons, the CRP-binding site lies well upstream of the promoter. In the *ara* region, the binding site for a single CRP is the farthest from the startpoint, centered at -92.

Dependence on CRP is related to the intrinsic efficiency of the promoter. No CRP-dependent promoter has a good -35 sequence and some also lack good -10 sequences. In fact, we might argue that effective control by CRP would be difficult if the promoter had effective -35 and –10 regions that interacted independently with RNA polymerase.

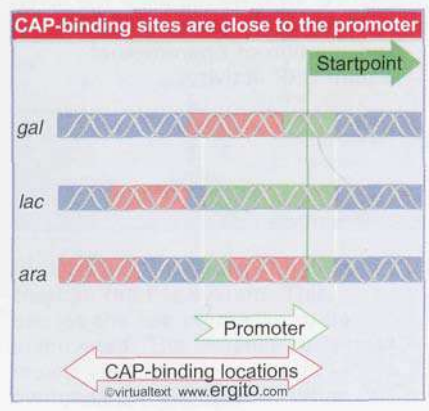


Figure 11.8 The CRP protein can bind at different sites relative to RNA polymerase.

There are in principle two ways in which CRP might activate transcription: it could interact directly with RNA polymerase; or it could act upon DNA to change its structure in some way that assists RNA polymerase to bind. In fact, CRP has effects upon both RNA polymerase and DNA.

Binding sites for CRP at most promoters resemble either *lac* (centered at –61) or *gal* (centered at –41 bp). The basic difference between them is that in the first type (called class I) the CRP-binding site is entirely upstream of the promoter, whereas in the second type (called class II) the CRP-binding site overlaps the binding site for RNA polymerase. (The interactions at the *ara* promoter may be different.)

In both types of promoter, the CRP binding site is centered an integral number of turns of the double helix from the startpoint. This suggests that CRP is bound to the same face of DNA as RNA polymerase. However, the nature of the interaction between CRP and RNA polymerase is different at the two types of promoter.

When the a subunit of RNA polymerase has a deletion in the C-terminal end, transcription appears normal except for the loss of ability to be ac-

tivated by CRP. CRP has an "activating region", which consists of a small exposed loop of ~ 10 amino acids, that is required for activating both types of its promoters. The activating region is a small patch of amino acids that interacts directly with the a subunit of RNA polymerase to stimulate the enzyme. At class I promoters, this interaction is sufficient. At class II promoters, a second interaction is also required, involving another region of CRP and the N-terminal region of the RNA polymerase a subunit.

Experiments using CRP dimers in which only one of the subunits has a functional transcription-activating region shows that, when CRP is bound at the *lac* promoter, only the activating region of the subunit nearer the startpoint is required, presumably because it touches RNA polymerase. This offers an explanation for the lack of dependence on the orientation of the binding site: the dimeric structure of CRP ensures that one of the subunits is available to contact RNA polymerase, no matter which subunit binds to DNA and in which orientation.

The effect upon RNA polymerase binding depends on the relative locations of the two proteins. At class I promoters, where CRP binds adjacent to the promoter, it increases the rate of initial binding to form a closed complex. At class II promoters, where CRP binds within the promoter, it increases the rate of transition from the closed to open complex.

11.6 CRP bends DNA

Key Concepts

- CRP introduces a 90° bend into DNA at its binding site.

The structure of the CRP-DNA complex is interesting: *the DNA has a bend.* Proteins may distort the double helical structure of DNA when they bind, and several regulator proteins induce a bend in the axis.

Figure 11.9 illustrates a technique that can be used to measure the extent and location of a bend. A target sequence containing the site is cut with different restriction enzymes to generate a set of fragments all of the same length, but each containing the protein-binding site at a different location.

The fragments move at different speeds in an electrophoretic gel, depending on the position of the bend. (If there is no bend, all fragments move at the same rate.) The greatest impediment to motion, causing the lowest mobility, happens when the bend is in the center of the

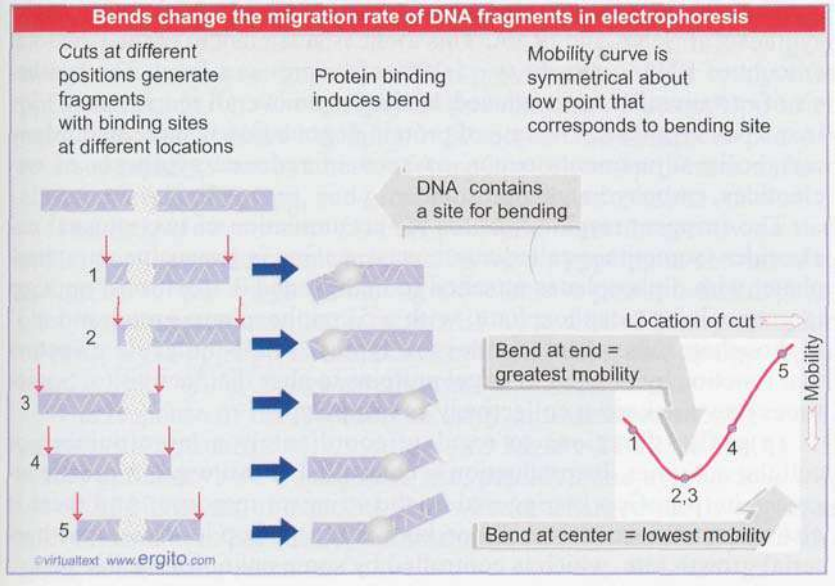


Figure 11.9 Gel electrophoresis can be used to analyze bending.

By Book_Crazy [IND]

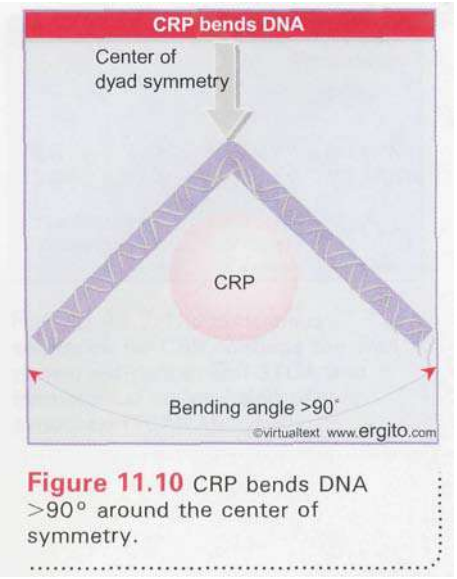


Figure 11.10 CRP bends DNA >90° around the center of symmetry.

DNA fragment. The least impediment to motion, allowing the greatest mobility, happens when the bend is at one end.

The results are analyzed by plotting mobility against the site of restriction cutting. The low point on the curve identifies the situation in which the restriction enzyme has cut the sequence immediately adjacent to the site of bending.

For the interaction of CRP with the *lac* promoter, this point lies at the center of dyad symmetry. The bend is quite severe, >90°, as illustrated in the model of **Figure 11.10.** There is therefore a dramatic change in the organization of the DNA double helix when CRP protein binds. The mechanism of bending is to introduce a sharp kink within the TGTGA consensus sequence. When there are inverted repeats of the consensus, the two kinks in each copy present in a palindrome cause the overall 90° bend. It is possible that the bend has some direct effect upon transcription, but it could be the case that it is needed simply to allow CRP to contact RNA polymerase at the promoter.

Whatever the exact means by which CRP activates transcription at various promoters, it accomplishes the same general purpose: to turn off alternative metabolic pathways when they become unnecessary because the cell has an adequate supply Of glucose. Again, this makes the point that coordinate control, of either negative or positive type, can extend over dispersed loci by repetition of binding sites for the regulator protein.

11.7 The stringent response produces (p)ppGpp

Key Concepts

- Poor growth conditions cause bacteria to produce the small molecule regulators ppGpp and pppGpp.

When bacteria find themselves in such poor growth conditions that they lack a sufficient supply of amino acids to sustain protein synthesis, they shut down a wide range of activities. This is called the stringent response. We can view it as a mechanism for surviving hard times: the bacterium husbands its resources by engaging in only the minimum of activities until nutrient conditions improve, when it reverses the response and again engages its full range of metabolic activities.

The stringent response causes a massive (10-20×) reduction in the synthesis of rRNA and tRNA. This alone is sufficient to reduce the total amount of RNA synthesis to ~5-10% of its previous level. The synthesis of certain mRNAs is reduced, leading to an overall reduction of ~3× in mRNA synthesis. The rate of protein degradation is increased. Many metabolic adjustments occur, as seen in reduced synthesis of nucleotides, carbohydrates, lipids, etc.

The stringent response causes the accumulation of two unusual nucleotides (sometimes called alarmones). ppGpp is guanosine tetraphosphate, with diphosphates attached to both 5' and 3' positions. pppGpp is guanosine pentaphosphate, with a 5' triphosphate group and a 3' diphosphate. These nucleotides are typical small-molecule effectors that function by binding to target proteins to alter their activities. Sometimes they are known collectively as (p)ppGpp.

(p)ppGpp functions to regulate coordinately a large number of cellular activities. Its production is controlled in two ways. A drastic increase in (p)ppGpp is triggered by the stringent response. And there is also a general inverse correlation between (p)ppGpp levels and the bacterial growth rate, which is controlled by some unknown means.

By Book_Crazy [IND]

11.8 (p)ppGpp is produced by the ribosome

Key Concepts

- The stringent factor RelA is a (p)ppGpp synthetase that is associated with ~5% of ribosomes.
- RelA is activated when the A site is occupied by an uncharged tRNA.
- One (p)ppGpp is produced every time an uncharged tRNA enters the A site.

Deprivation of any one amino acid, or mutation to inactivate any aminoacyl-tRNA synthetase, is sufficient to initiate the stringent response. The trigger that sets the entire series of events in train is *the presence of uncharged tRNA in the A site of the ribosome.* Under normal conditions, of course, only aminoacyl-tRNA is placed in the A site by EF-Tu (see *6.10 Elongation factor Tu loads aminoacyl-tRNA into the A site).* But when there is no aminoacyl-tRNA available to respond to a particular codon, the uncharged tRNA becomes able to gain entry. Of course, this blocks any further progress by the ribosome; and it triggers an idling reaction.

The components involved in producing (p)pGpp via the idling reaction have been identified through the existence of relaxed (re/) mutants. *rel* mutations abolish the stringent response, so that starvation for amino acids does not cause any reduction in stable RNA synthesis or alter any of the other reactions that are usually seen.

The most common site of relaxed mutation lies in the gene *relA*, which codes for a protein called the stringent factor. This factor is associated with the ribosomes, although the amount is rather low—say, < 1 molecule for every 200 ribosomes. So perhaps only a minority of the ribosomes are able to produce the stringent response.

Ribosomes obtained from stringent bacteria can synthesize ppGpp and pppGpp *in vitro*, provided that the A site is occupied by an uncharged tRNA *specifically responding to the codon.* Ribosomes extracted from relaxed mutants cannot perform this reaction; but they are able to do so if the stringent factor is added.

Figure 11.11 shows the pathways for synthesis of (p)ppGpp. The stringent factor (RelA) is an enzyme that catalyzes the synthetic reaction in which ATP is used to donate a pyrophosphate group to the 3′ position of either GTP or GDP. The formal name for this activity is (p)ppGpp synthetase.

How is ppGpp removed when conditions return to normal? A gene called *spoT* codes for an enzyme that provides the major catalyst for ppGpp degradation. The activity of this enzyme causes ppGpp to be rapidly degraded, with a half-life of ~20 sec; so the stringent response is reversed rapidly when synthesis of (p)ppGpp ceases. *spoT* mutants have elevated levels of ppGpp, and grow more slowly as a result.

The RelA enzyme uses GTP as substrate more frequently, so that pppGpp is the predominant product. However, pppGpp is converted to ppGpp by several enzymes; among those able to perform this dephosphorylation are the translation factors EF-Tu and EF-G. The production of ppGpp via pppGpp is the most common route, *and ppGpp is the usual effector of the stringent response.*

The response of the ribosome to entry of uncharged tRNA is compared with normal protein synthesis in **Figure 11.12.** When EF-Tu places aminoacyl-tRNA in the A site, peptide bond synthesis is followed by ribosomal movement. But when uncharged tRNA is paired with the codon in the A site, the ribosome remains stationary and engages in the idling reaction.

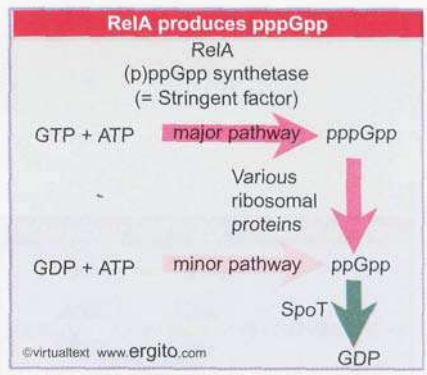


Figure 11.11 Stringent factor catalyzes the synthesis of pppGpp and ppGpp; ribosomal proteins can dephosphorylate pppGpp to ppGpp.

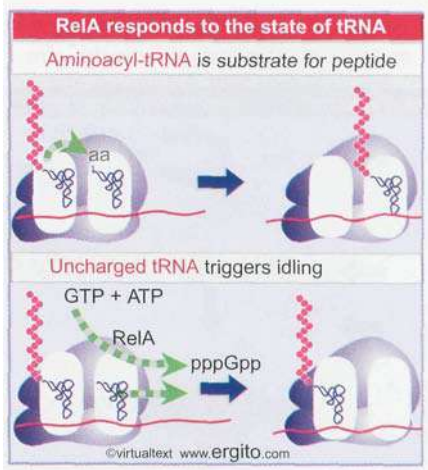


Figure 11.12 In normal protein synthesis, the presence of aminoacyl-tRNA in the A site is a signal for peptidyl transferase to transfer the polypeptide chain, followed by movement catalyzed by EF-G; but under stringent conditions, the presence of uncharged tRNA causes RelA protein to synthesize (p)ppGpp and to expel the tRNA.

How does the state of the ribosome control the activity of RelA enzyme? An indication of the nature of the interaction is revealed by relaxed mutations in another locus, originally called *relC*, which turns out to be the same as *rplK*, which codes for the 50S subunit protein L11. This protein is located in the vicinity of the A and P sites, in a position to respond to the presence of a properly paired but uncharged tRNA in the A site. A conformational change in this protein or some other component could activate the RelA enzyme, so that the idling reaction occurs instead of polypeptide transfer from the peptidyl-tRNA.

One round of (p)ppGpp synthesis is associated with release of the uncharged tRNA from the A site, so that synthesis of (p)ppGpp is a continuing response to the level of uncharged tRNA. So under limiting conditions, a ribosome stalls when no aminoacyl-tRNA is available to respond to the codon in the A site. Entry of uncharged tRNA triggers the synthesis of a (p)ppGpp molecule, and the resulting expulsion of the uncharged tRNA allows the situation to be reassessed. Depending upon the availability of aminoacyl-tRNA, the ribosome resumes polypeptide synthesis or undertakes another idling reaction.

11.9 ppGpp has many effects

Key Concepts

- ppGpp inhibits transcription of rRNA.

What does ppGpp do? It is an effector for controlling several reactions, including the inhibition of transcription. Many effects have been reported, among which two stand out:

- *Initiation of transcription is specifically inhibited at the promoters of operons coding for rRNA.* Mutations of stringently regulated promoters can abolish stringent control, which suggests that the effect requires an interaction with specific promoter sequences.
- *The elongation phase of transcription of many or most templates is reduced by ppGpp.* The cause is increased pausing by RNA polymerase. This effect is responsible for the general reduction in transcription efficiency when ppGpp is added *in vitro*.

The use of ppGpp is just one aspect of a more general regulatory network that relates production of ribosomes to the growth rate. The level of protein synthesis increases in proportion with the growth rate. This is accomplished by increasing the production of ribosomes as cells grow more rapidly. The cell therefore needs some general indicator of growth rate that can be used to control the synthesis of ribosomes. The indicator appears to be NTP levels, and the target for their action is the control of transcription of rRNA.

Figure 11.13 summarizes the systems that are used to control rRNA transcription in response to growth rate. Under conditions of starvation, ppGpp is produced, and (among its various actions) inhibits initiation at the promoters of the *rrn* loci that code for rRNA. As growth rate increases, the levels of ATP and GTP increase. These increase the rate of initiation at the *rrn* promoters.

The *rrn* promoters in *E. coli* form atypical open complexes with RNA polymerase. The open complexes are unusually unstable. The result is that the main factor governing the rate of initiation becomes the decay rate of the open complex. Increased concentration of the initiating nucleotide (which is ATP at six of the *rrn* promoters and GTP at the seventh) drives the initiation reaction forward by stabilizing the open complex.

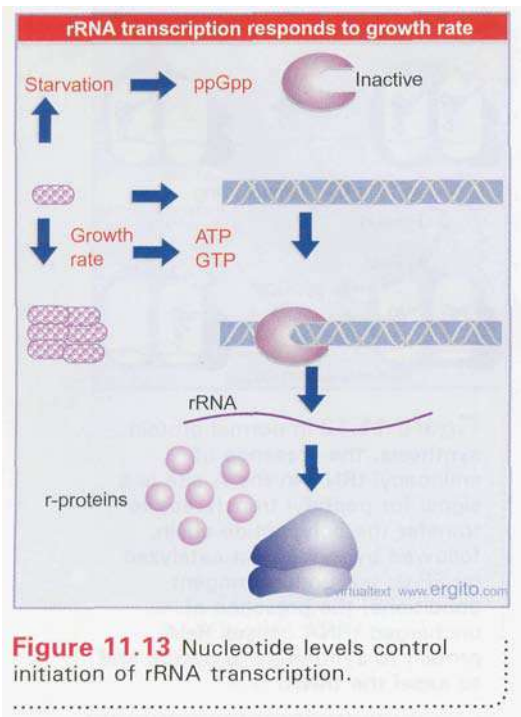


Figure 11.13 Nucleotide levels control initiation of rRNA transcription.

By Book_Crazy [IND]

The level of rRNA controls the production of ribosomes (by a feedback loop in which the absence of rRNA inhibits synthesis of ribosomal proteins, see Figure 11.18). This means that the production of ribosomes, and thus the level of protein synthesis, in turn respond to the levels of ATP and GTP, which reflect the nutritional condition of the cell. So concentrations of particular nucleotides control ribosome synthesis in response to normal changes in growth rate and the more extreme conditions of starvation.

11.10 Translation can be regulated

Key Concepts

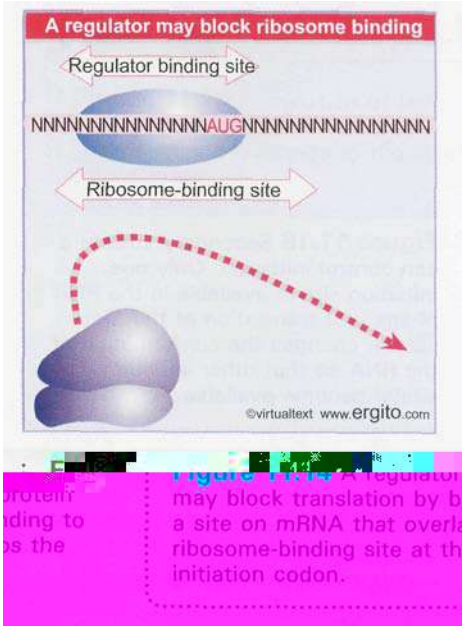
- A repressor protein can regulate translation by preventing a ribosome from binding to an initiation codon.
- Accessibility of initiation codons in a polycistronic mRNA can be controlled by changes in the structure of the mRNA that occur as the result of translation.

Translational control is a notable feature of operons coding for components of the protein synthetic apparatus. The operon provides an arrangement for *coordinate* regulation of a group of structural genes. But, superimposed on it, further controls, such as those at the level of translation, may create *differences* in the extent to which individual genes are expressed.

A similar type of mechanism is used to achieve translational control in several systems. *Repressor function is provided by a protein that binds to a target region on mRNA to prevent ribosomes from recognizing the initiation region.* Formally this is equivalent to a repressor protein binding to DNA to prevent RNA polymerase from utilizing a promoter. **Figure 11.14** illustrates the most common form of this interaction, in which the regulator protein binds directly to a sequence that includes the AUG initiation codon, thereby preventing the ribosome from binding.

Some examples of translational repressors and their targets are summarized in **Figure 11.15**. A classic example is the coat protein of the RNA phage R17; it binds to a hairpin that encompasses the ribosome binding site in the phage mRNA. Similarly the T4 RegA protein binds to a consensus sequence that includes the AUG initiation codon in several T4 early mRNAs; and T4 DNA polymerase binds to a sequence in its own mRNA that includes the Shine-Dalgarno element needed for ribosome binding.

Another form of translational control occurs when translation of one cistron requires changes in secondary structure that depend on translation of a preceding cistron. This happens during translation of the RNA phages, whose cistrons always are expressed in a set order. **Figure 11.16** shows that the phage RNA takes up a secondary structure in which only one initiation sequence is accessible; the second cannot be recognized by ribosomes because it is base paired with other regions of the RNA. However, translation of the first cistron disrupts the secondary structure, allowing ribosomes to bind to the initiation site of the next cistron. In this mRNA, secondary structure controls translatability.



may block translation by b
a site on mRNA that overl
ribosome-binding site at th
initiation codon.

Translational repressors bind to mRNA

Repressor	Target Gene	Site of Action
R17 coat protein	R17 replicase	hairpin that includes ribosome binding site
T4 RegA	early T4 mRNAs	various sequences including initiation codon
T4 DNA polymerase	T4 DNA polymerase	Shine-Dalgarno sequence
T4 p32	gene 32	single-stranded 5' leader

©virtualtext www.ergito.com

Figure 11.15 Proteins that bind to sequences within the initiation regions of mRNAs may function as translational repressors.

By Book_Crazy [IND]

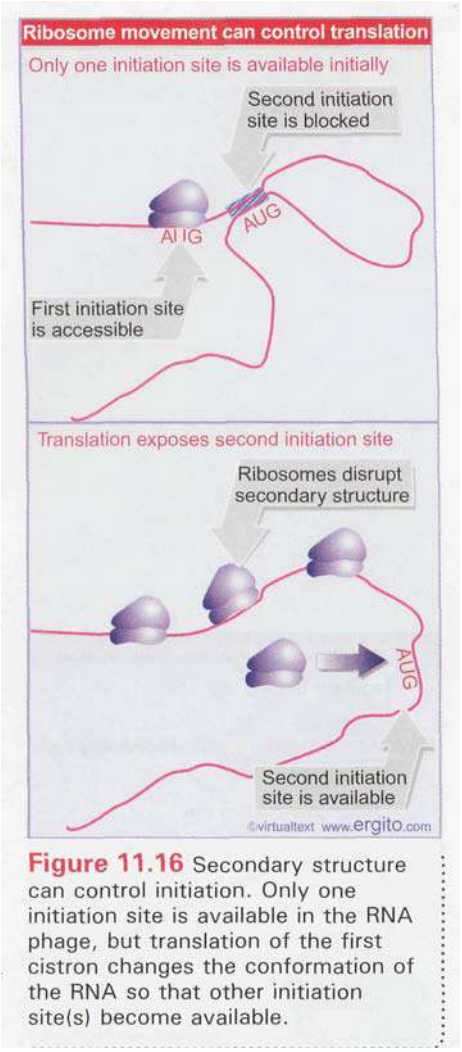


Figure 11.16 Secondary structure can control initiation. Only one initiation site is available in the RNA phage, but translation of the first cistron changes the conformation of the RNA so that other initiation site(s) become available.

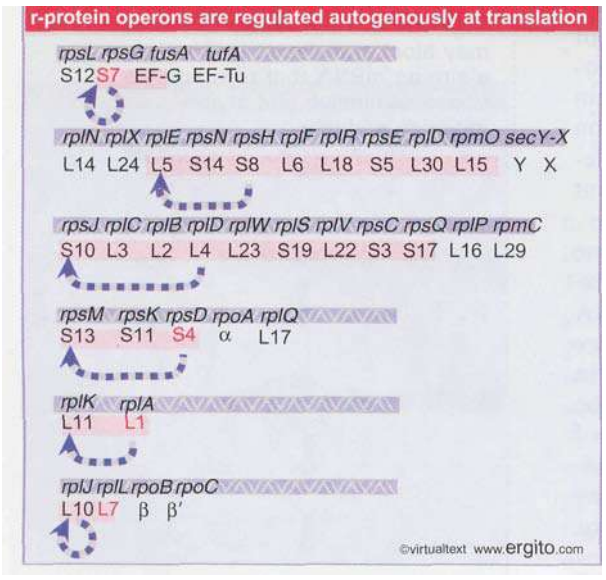


Figure 11.17 Genes for ribosomal proteins, protein synthesis factors, and RNA polymerase subunits are interspersed in a small number of operons that are autonomously regulated. The regulator is named in red; the proteins that are regulated are shaded in pink.

11.11 r-protein synthesis is controlled by autogenous regulation

Key Concepts

- Translation of an r-protein operon can be controlled by a product of the operon that binds to a site on the polycistronic mRNA.

About 70 or so proteins constitute the apparatus for bacterial gene expression. The ribosomal proteins are the major component, together with the ancillary proteins involved in protein synthesis. The subunits of RNA polymerase and its accessory factors make up the remainder. The genes coding for ribosomal proteins, protein-synthesis factors, and RNA polymerase subunits all are intermingled and organized into a small number of operons. Most of these proteins are represented only by single genes in *E. coli*.

Coordinate controls ensure fhat these proteins are synthesized in amounts appropriate for the growth conditions: when bacteria grow more rapidly, they devote a greater proportion of their efforts to the production of the apparatus for gene expression. An array of mechanisms is used to control the expression of the genes coding for this apparatus and to ensure that the proteins are synthesized at comparable levels that are related to the levels of the rRNAs.

The organization of six operons is summarized in Figure 11.17. About half of the genes for ribosomal proteins (r-proteins) map in four operons that lie close together (named *str, spc, S10,* and a simply for the first one of the functions to have been identified in each case). The *rif* and *L11* operons lie together at another location.

Each operon codes for a variety of functions. The *str* operon has genes for small subunit ribosomal proteins as well as for EF-Tu and EF-G. The *spc* and *S10* operons have genes interspersed for both small and large ribosomal subunit proteins. The a operon has genes for proteins of both ribosomal subunits as well as for the a subunit of RNA polymerase. The *rif* locus has genes for large subunit ribosomal proteins and for the β and *ft* subunits of RNA polymerase.

All except one of the ribosomal proteins are needed in equimolar amounts, which must be coordinated with the level of rRNA. The dispersion of genes whose products must be equimolar, and their intermingling with genes whose products are needed in different amounts, pose some interesting problems for coordinate regulation.

A feature common to all of the operons described in Figure 11.17 is regulation of some of the genes by one of the products. In each case, the gene coding for the regulatory product is itself one of the targets for regulation. Autogenous regulation occurs whenever a protein (or RNA) regulates its own production. In the case of the r-protein operons, the regulatory protein inhibits expression of a contiguous set of genes within the operon, so this is an example of negative autogenous regulation.

In each case, *accumulation of the protein inhibits further synthesis of itself and of some other gene products.* The effect often is exercised at the level of translation of the polycistronic mRNA. Each of the regulators is a ribosomal protein that binds directly to rRNA. *Its effect on translation is a result of its ability also to bind to its own mRNA.* The sites on mRNA at which these proteins bind either overlap the sequence where translation is initiated or lie nearby and

By Book_Crazy [IND]

probably influence the accessibility of the initiation site by inducing conformational changes. For example, in the S10 operon, protein L4 acts at the very start of the mRNA to inhibit translation of S10 and the subsequent genes. The inhibition may result from a simple block to ribosome access, as illustrated previously in Figure 11.15, or it may prevent a subsequent stage of translation. In *two cases (including S4 in the a* operon), the regulatory protein stabilizes a particular secondary structure in the mRNA that prevents the initiation reaction from continuing after the 30S subunit has bound.

The use of r-proteins that bind rRNA to establish autogenous regulation immediately suggests that this provides a mechanism to link r-protein synthesis to rRNA synthesis. A generalized model is depicted in **Figure 11.18.** Suppose that the binding sites for the autogenous regulator r-proteins on rRNA are much stronger than those on the mRNAs. Then so long as any free rRNA is available, the newly synthesized r-proteins will associate with it to start ribosome assembly. There will be no free r-protein available to bind to the mRNA, so its translation will continue. But as soon as the synthesis of rRNA slows or stops, free r-proteins begin to accumulate. Then they are available to bind their mRNAs, repressing further translation. This circuit ensures that each r-protein operon responds in the same way to the level of rRNA: as soon as there is an excess of r-protein relative to rRNA, synthesis of the protein is repressed.

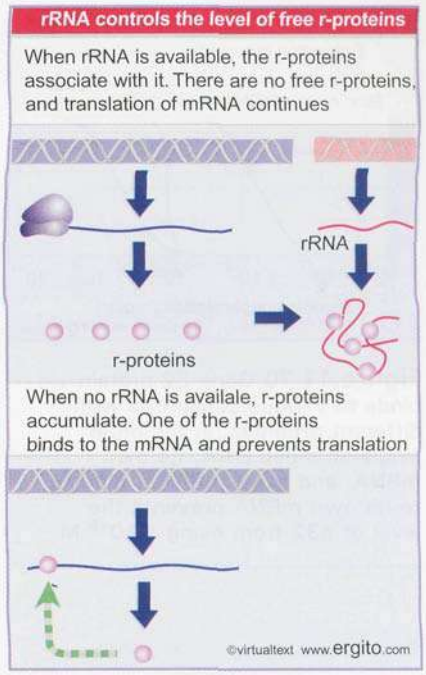


Figure 11.18 Translation of the r-protein operons is autogenously controlled and responds to the level of rRNA.

11.12 Phage T4 p32 is controlled by an autogenous circuit

Key Concepts

* p32 binds to its own mRNA to prevent initiation of translation.

Autogenous regulation has been placed on a quantitative basis for gene *32* of phage T4. The protein (p32) plays a central role in genetic recombination, DNA repair, and replication, in which its function is exercised by virtue of its ability to bind to single-stranded DNA. Nonsense mutations cause the inactive protein to be overproduced. *So when the function of the protein is prevented, more of it is made.* This effect occurs at the level of translation; the gene *32* mRNA is stable, and remains so irrespective of the behavior of the protein product.

Figure 11.19 presents a model for the gene *32* control circuit. When single-stranded DNA is present in the phage-infected cell, it sequesters p32. However, in the absence of single-stranded DNA, or at least in conditions in which there is a surplus of p32, the protein prevents translation of its own mRNA. The effect is mediated directly by p32 binding to mRNA to prevent initiation of translation. Probably this occurs at an A·T-rich region that surrounds the ribosome binding site.

Two features of the binding of p32 to the site on mRNA are required to make the control loop work effectively:

- The affinity of p32 for the site on gene *32* mRNA must be significantly lower than its affinity for single-stranded DNA. The equilibrium constant for binding RNA is in fact almost two orders of magnitude below that for single-stranded DNA.
- But the affinity of p32 for the mRNA must be significantly greater than the affinity for other RNA sequences. It is influenced by base composition and by secondary structure; an important aspect of the

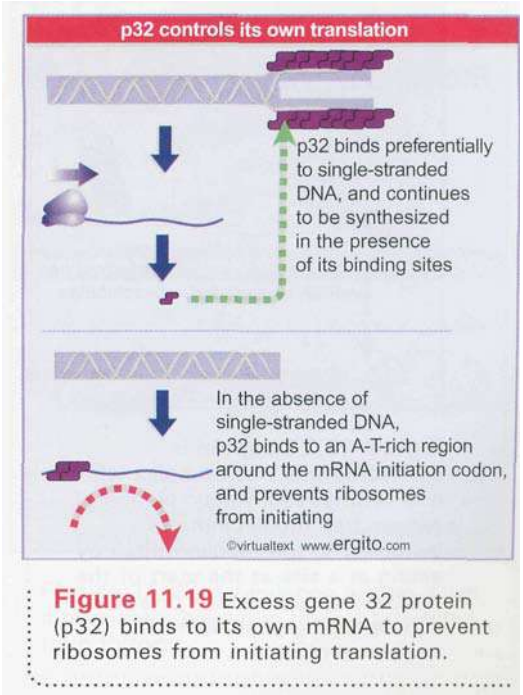


Figure 11.19 Excess gene 32 protein (p32) binds to its own mRNA to prevent ribosomes from initiating translation.

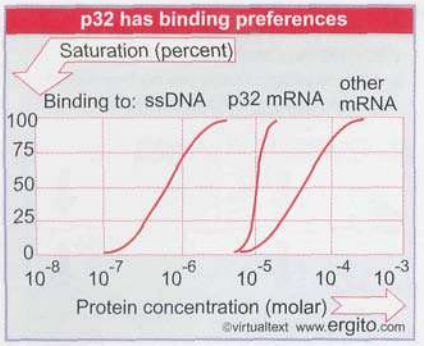


Figure 11.20 Gene 32 protein binds to various substrates with different affinities, in the order single-stranded DNA, its own mRNA, and other mRNAs. Binding to its own mRNA prevents the level of p32 from rising $>10^{-6}$ M.

binding to gene *32* mRNA is that the regulatory region has an extended sequence lacking secondary structure.

Using the known equilibrium constants, we can plot the binding of p32 to its target sites as a function of protein concentration. **Figure 11.20** shows that at concentrations below 10^{-6} M, p32 binds to single-stranded DNA. At concentrations $>10^{-6}$ M, it binds to gene *32* mRNA. At yet greater concentrations, it binds to other mRNA sequences, with a range of affinities.

These results imply that the level of p32 should be autoregulated to be $<10^{-6}$ M, which corresponds to ~2000 molecules per bacterium. This fits well with the measured level of 1000-2000 molecules/cell.

A feature of autogenous control is that each regulatory interaction is unique: a protein acts only on the mRNA responsible for its own synthesis. Phage T4 provides an example of a more general translational regulator, coded by the gene *regA*, which represses the expression of several genes that are transcribed during early infection. RegA protein prevents the translation of mRNAs for these genes by competing with 30S subunits for the initiation sites on the mRNA. Its action is a direct counterpart to the function of a repressor protein that binds multiple operators.

11.13 Autogenous regulation is often used to control synthesis of macromolecular assemblies

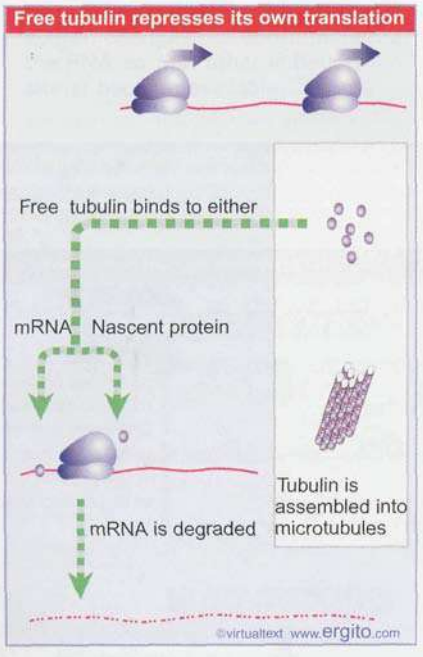


Figure 11.21 Tubulin is assembled into microtubules when it is synthesized. Accumulation of excess free tubulin induces instability in the tubulin mRNA by acting at a site at the start of the reading frame in mRNA or at the corresponding position in the nascent protein.

Key Concepts

- The precursor to microtubules, free tubulin protein, inhibits translation of tubulin mRNA.

Autogenous regulation is a common type of control among proteins that are incorporated into macromolecular assemblies. The assembled particle itself may be unsuitable as a regulator, because it is too large, too numerous, or too restricted in its location. But the need for synthesis of its components may be reflected in the pool of free precursor subunits. If the assembly pathway is blocked for any reason, free subunits accumulate and shut off the unnecessary synthesis of further components.

Eukaryotic cells have a common system in which autogenous regulation of this type occurs. Tubulin is the monomer from which microtubules, a major filamentous system of all eukaryotic cells, are synthesized. The production of tubulin mRNA is controlled by the free tubulin pool. When this pool reaches a certain concentration, the production of further tubulin mRNA is prevented. Again, the principle is the same: tubulin sequestered into its macromolecular assembly plays no part in regulation, but the level of the free precursor pool determines whether further monomers are added to it.

The target site for regulation is a short sequence at the start of the coding region. We do not know yet what role this sequence plays, but two models are illustrated in **Figure 11.21**. Tubulin may bind directly to the mRNA; or it may bind to the nascent polypeptide representing this region. Whichever model applies, excess tubulin causes tubulin mRNA that is located on polysomes to be degraded, so the consequence of the reaction is to make the tubulin mRNA unstable.

Autogenous control is an *intrinsically* self-limiting system, by contrast with the *extrinsic* control that we discussed previously. A repressor protein's ability to bind an operator may be controlled by the level of an extraneous small molecule, which activates or inhibits its activity. But in the case of autogenous regulation, the critical parameter is the concentration of the protein itself.

11.14 Alternative secondary structures control attenuation

Key Concepts

- Termination of transcription can be attenuated by controlling formation of the necessary hairpin structure in RNA.
- The most direct mechanisms for attenuation involve proteins that either stabilize or destabilize the hairpin.

RNA structure provides an opportunity for regulation in both prokaryotes and eukaryotes. Its most common role occurs when an RNA molecule can take up alternative secondary structures by utilizing different schemes for intramolecular base pairing. The properties of the alternative conformations may be different. This type of mechanism can be used to regulate the termination of transcription, when the alternative structures differ in whether they permit termination. Another means of controlling conformation (and thereby function) is provided by the cleavage of an RNA: by removing one segment of an RNA, the conformation of the rest may be altered. It is possible also for a (small) RNA molecule to control the activity of a target RNA by base pairing with it: the role of the small RNA is directly analogous to that of a regulator protein (see *11.19 Small RNA molecules can regulate translation).* The ability of an RNA to shift between different conformations with regulatory consequences is the nucleic acid's alternative to the allosteric changes of conformation that regulate protein function. Both these mechanisms allow an interaction at one site in the molecule to affect the structure of another site.

Several operons are regulated by attenuation, a mechanism that controls the ability of RNA polymerase to read through an attenuator, which is an intrinsic terminator located at the beginning of a transcription unit. *The principle of attenuation is that some external event controls the formation of the hairpin needed for intrinsic termination.* If the hairpin is allowed to form, termination prevents RNA polymerase from transcribing the structural genes. If the hairpin is prevented from forming, RNA polymerase elongates through the terminator, and the genes are expressed. Different types of mechanisms are used in different systems for controlling the structure of the RNA.

Attenuation may be regulated by proteins that bind to RNA, either to stabilize or to destabilize formation of the hairpin required for termination. **Figure 11.22** shows an example in which a protein prevents formation of the terminator hairpin. The activity of such a protein may be intrinsic or may respond to a small molecule in the same manner as a repressor protein responds to corepressor.

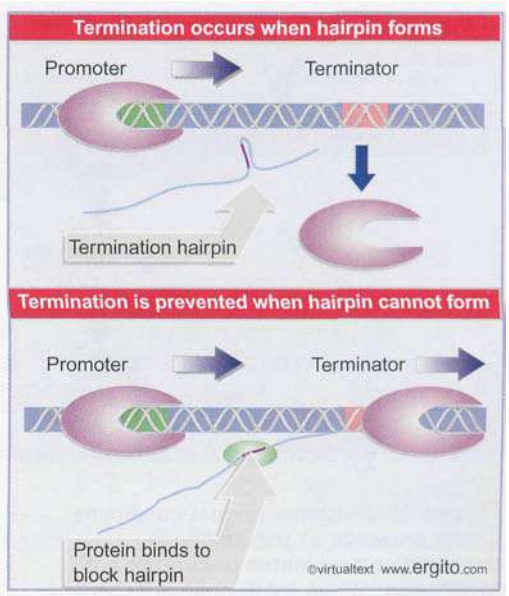


Figure 11.22 Attenuation occurs when a terminator hairpin in RNA is prevented from forming.

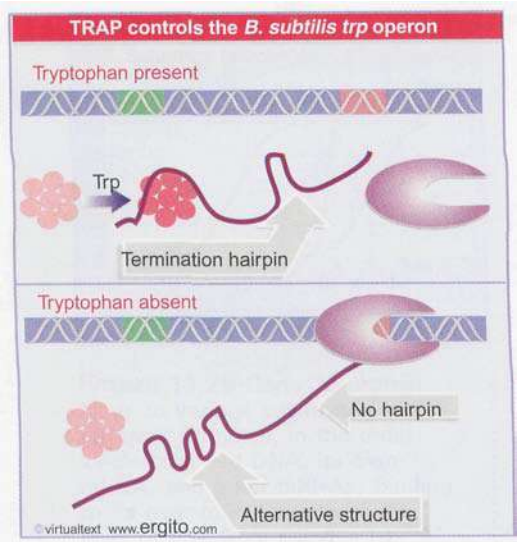


Figure 11.23 TRAP is activated by tryptophan and binds to trp mRNA. This allows the termination hairpin to form, with the result that RNA polymerase terminates, and the genes are not expressed. In the absence of tryptophan, TRAP does not bind, and the mRNA adopts a structure that prevents the terminator hairpin from forming.

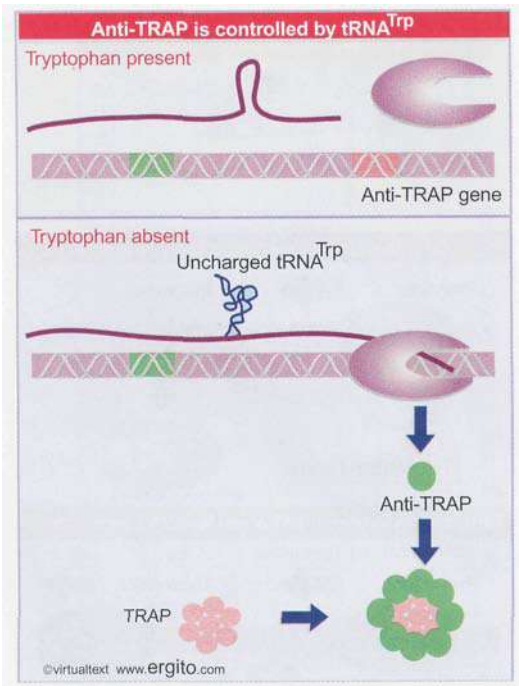


Figure 11.24 Under normal conditions (in the presence of tryptophan) transcription terminates before the anti-TRAP gene. When tryptophan is absent, uncharged $tRNA^{Trp}$ base pairs with the anti-TRAP mRNA, preventing formation of the terminator hairpin, thus causing expression of anti-TRAP.

11.15 Termination of *B. subtilis trp* genes is controlled by tryptophan and by $tRNA^{Trp}$

Key Concepts

- A terminator protein called TRAP is activated by tryptophan to prevent transcription of *trp* genes.
- Activity of TRAP is (indirectly) inhibited by uncharged $tRNA^{Trp}$.

The circuitry that controls transcription via termination can use both direct and indirect means to respond to the level of small molecule products or substrates.

In *B. subtilis*, a protein called TRAP (formerly called MtrB) is activated by tryptophan to bind to a sequence in the leader of the nascent transcript. TRAP forms a multimer of 11 subunits. Each subunit binds a single tryptophan amino acid and a trinucleotide (GAG or UAG) of RNA. The RNA is wound in a circle around the protein. **Figure 11.23** shows that the result is to ensure the availability of the regions that are required to form the terminator hairpin. The termination of transcription then prevents production of the tryptophan biosynthetic enzymes. In effect, TRAP is a terminator protein that responds to the level of tryptophan. In the absence of TRAP, an alternative secondary structure precludes the formation of the terminator hairpin.

However, the TRAP protein in turn is also controlled by $tRNA^{Trp}$. **Figure 11.24** shows that uncharged $tRNA^{Trp}$ binds to the mRNA for a protein called anti-TRAP. This is necessary to suppress formation of a termination hairpin in the mRNA. The result is the synthesis of anti-TRAP, which binds to TRAP, and prevents it from repressing the tryptophan operon. By this complex series of events, the absence of tryptophan generates the uncharged tRNA, which causes synthesis of anti-TRAP, which prevents function of TRAP, which causes expression of tryptophan genes.

Expression of the *B. subtilis trp* genes is therefore controlled by both tryptophan and $tRNA^{Trp}$. When tryptophan is present, there is no need for it to be synthesized. This is accomplished when tryptophan activates TRAP and therefore inhibits expression of the enzymes that synthesize tryptophan. The presence of uncharged $tRNA^{Trp}$ indicates that there is a shortage of tryptophan. The uncharged tRNA activates the anti-TRAP, and thereby activates transcription of the *trp* genes.

11.16 The *E. coli tryptophan* operon is controlled by attenuation

Key Concepts

- An attenuator (intrinsic terminator) is located between the promoter and the first gene of the *trp* cluster.
- The absence of tryptophan suppresses termination and results in a 10× increase in transcription.

A complex regulatory system is used in *E. coli* (where attenuation was originally discovered). The changes in secondary structure

By Book_Crazy [IND]

that control attenuation are determined by the position of the ribosome on mRNA. **Figure 11.25** shows that termination requires that *the ribosome can translate a leader segment that precedes the trp genes in the mRNA.* When the ribosome translates the leader region, a termination hairpin forms at terminator 1. But when the ribosome is prevented from translating the leader, the termination hairpin does not form, and RNA polymerase transcribes the coding region. *This mechanism of antitermination therefore depends upon the ability of external circumstances to influence ribosome movement in the leader region.*

The *trp* operon consists of five structural genes arranged in a contiguous series, coding for the three enzymes that convert chorismic acid to tryptophan. **Figure 11.26** shows that transcription starts at a promoter at the left end of the cluster. *trp* operon expression is controlled by two separate mechanisms. Repression of expression is exercised by a repressor protein (coded by the unlinked gene *trpR)* that binds to an operator that is adjacent to the promoter. Attenuation controls the progress of RNA polymerase into the operon by regulating whether termination occurs at a site preceding the first structural gene.

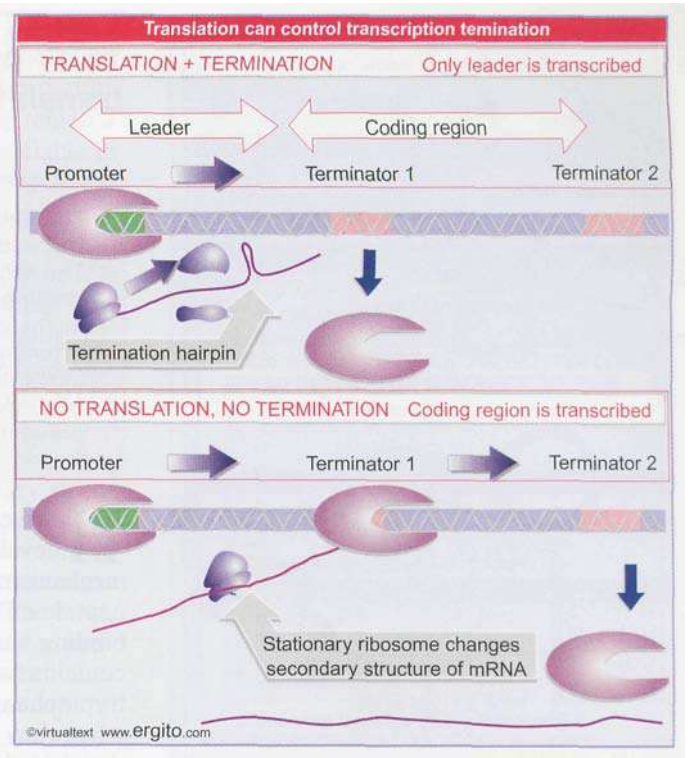


Figure 11.25 Termination can be controlled via changes in RNA secondary structure that are determined by ribosome movement.

Attenuation was first revealed by the observation that deleting a sequence between the operator and the *trpE* coding region can increase the expression of the structural genes. This effect is independent of repression: both the basal and derepressed levels of transcription are increased. So this site influences events that occur *after* RNA polymerase has set out from the promoter (irrespective of the conditions prevailing at initiation).

An attenuator (intrinsic terminator) is located between the promoter and the *trpE* gene. It provides a barrier to transcription into the structural genes. RNA polymerase terminates there, either *in vivo* or *in vitro,* to produce a 140-base transcript.

Termination at the attenuator responds to the level of tryptophan, as illustrated in **Figure 11.27.** In the presence of adequate amounts of tryptophan, termination is efficient. But in the absence of tryptophan, RNA polymerase can continue into the structural genes.

Repression and attenuation respond in the same way to the level of tryptophan. When tryptophan is present, the operon is repressed; and most of the RNA polymerases that escape from the promoter then terminate at the attenuator. When tryptophan is removed, RNA polymerase has free access to the promoter, and also is no longer compelled to terminate prematurely.

Attenuation has ~10× effect on transcription. When tryptophan is present, termination is effective, and the attenuator allows only ~10% of the RNA polymerases to proceed. In the absence of tryptophan, attenuation allows virtually all of the polymerases to proceed. Together with the ~70X increase in initiation of transcription that results from the release of repression, this allows an ~700-fold range of regulation of the operon.

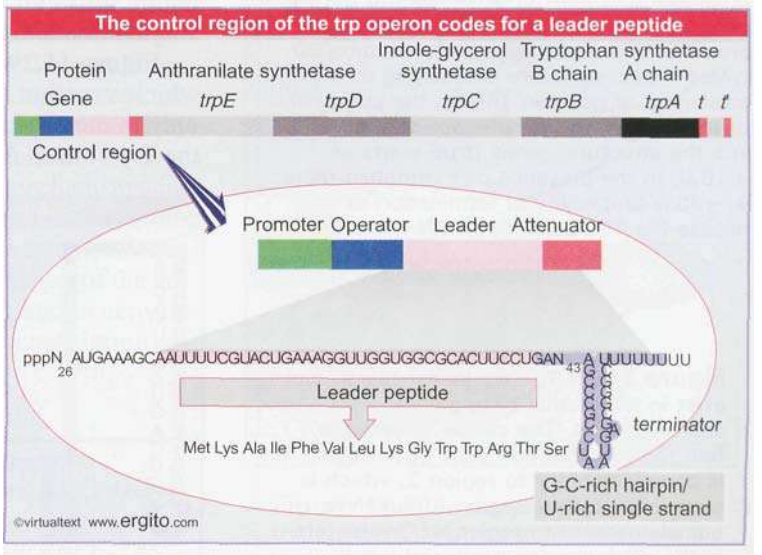


Figure 11.26 The *trp* operon consists of five contiguous structural genes preceded by a control region that includes a promoter, operator, leader peptide coding region, and attenuator.

By Book_Crazy [IND]

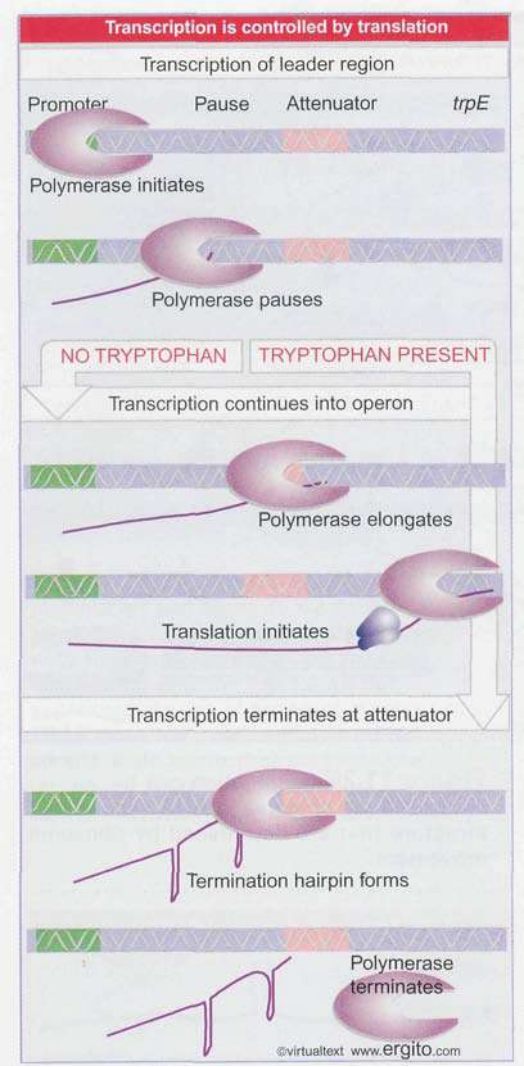


Figure 11.27 An attenuator controls the progression of RNA polymerase into the *trp* genes. RNA polymerase initiates at the promoter and then proceeds to position 90, where it pauses before proceeding to the attenuator at position 140. In the absence of tryptophan, the polymerase continues into the structural genes (*trpE* starts at +163). In the presence of tryptophan there is ~90% probability of termination to release the 140-base leader RNA.

11.17 Attenuation can be controlled by translation

Key Concepts

- The leader region of the *trp* operon has a 14-codon open reading frame that includes two codons for tryptophan.
- The structure of RNA at the attenuator depends on whether this reading frame is translated.
  - In the presence of tryptophan, the leader is translated, and the attenuator is able to form the hairpin that causes termination.
  - In the absence of tryptophan, the ribosome stalls at the tryptophan codons and an alternative secondary structure prevents formation of the hairpin, so that transcription continues.

How can termination of transcription at the attenuator respond to the level of tryptophan? The sequence of the leader region suggests a mechanism. It has a short coding sequence that could represent a leader peptide of 14 amino acids. Figure 11.26 shows that it contains a ribosome binding site whose AUG codon is followed by a short coding region that contains two successive codons for tryptophan. When the cell runs out of tryptophan, ribosomes initiate translation of the leader peptide, but stop when they reach the Trp codons. The sequence of the mRNA suggests that this ribosome stalling influences termination at the attenuator.

The leader sequence can be written in alternative base-paired structures. The ability of the ribosome to proceed through the leader region controls transitions between these structures. The structure determines whether the mRNA can provide the features needed for termination.

Figure 11.28 draws these structures. In the first, region *1* pairs with region 2; and region *3* pairs with region *4*. The pairing of regions *3* and *4* generates the hairpin that precedes the U_8 sequence: this is the essential signal for intrinsic termination. Probably the RNA would take up this structure in lieu of any outside intervention.

A different structure is formed if region *1* is prevented from pairing with region *2*. In this case, region *2* is free to pair with region *3*. Then region *4* has no available pairing partner; so it is compelled to remain single-stranded. So the terminator hairpin cannot be formed.

Figure 11.29 shows that the position of the ribosome can determine which structure is formed in such a way that termination is attenuated only in the absence of tryptophan. The crucial feature is the position of the Trp codons in the leader peptide coding sequence.

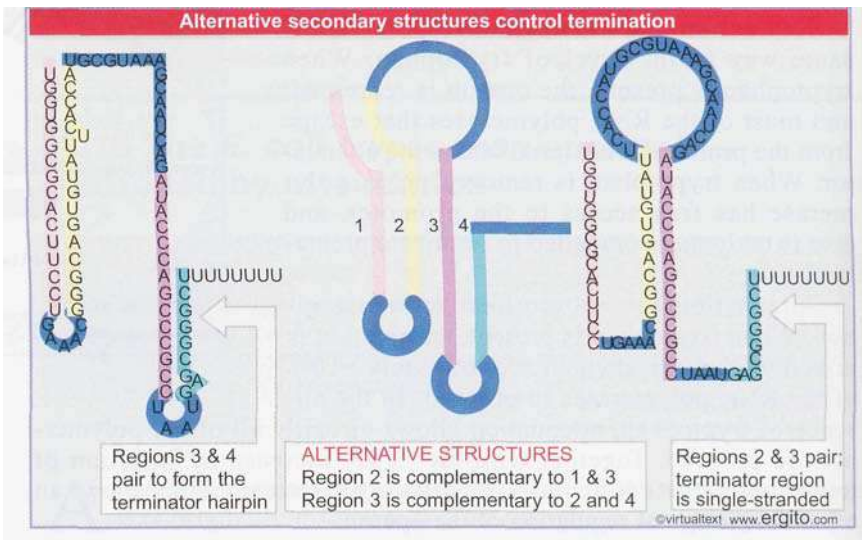


Figure 11.28 The trp leader region can exist in alternative base-paired conformations. The center shows the four regions that can base pair. Region 1 is complementary to region 2, which is complementary to region 3, which is complementary to region 4. On the left is the conformation produced when region 1 pairs with region 2, and region 3 pairs with region 4. On the right is the conformation when region 2 pairs with region 3, leaving regions 1 and 4 unpaired.

By Book_Crazy [IND]

When tryptophan is present, ribosomes are able to synthesize the leader peptide. They continue along the leader section of the mRNA to the UGA codon, which lies between regions *1* and *2*. As shown in the lower part of the figure, by progressing to this point, the ribosomes extend over region *2* and prevent it from base pairing. The result is that region 3 is available to base pair with region *4*, generating the terminator hairpin. Under these conditions, therefore, RNA polymerase terminates at the attenuator.

When there is no tryptophan, ribosomes stall at the Trp codons, which are part of region *1*, as shown in the upper part of the figure. So region *1* is sequestered within the ribosome and cannot base pair with region *2*. This means that regions *2* and *3* become base paired before region *4* has been transcribed. This compels region *4* to remain in a single-stranded form. In the absence of the terminator hairpin, RNA polymerase continues transcription past the attenuator.

Control by attenuation requires a precise timing of events. For ribosome movement to determine formation of alternative secondary structures that control termination, *translation of the leader must occur at the same time when RNA polymerase approaches the terminator site. A* critical event in controlling the timing is the presence of a site that causes the RNA polymerase to pause at base 90 along the leader. The RNA polymerase remains paused until a ribosome translates the leader peptide. Then the polymerase is released and moves off toward the attenuation site. By the time it arrives there, secondary structure of the attenuation region has been determined.

Figure 11.30 summarizes the role of Trp-tRNA in controlling expression of the operon. *By providing a mechanism to sense the inadequacy of the supply of Trp-tRNA, attenuation responds directly to the need of the cell for tryptophan in protein synthesis.*

How widespread is the use of attenuation as a control mechanism for bacterial operons? It is used in at least six operons that code for enzymes concerned with the biosynthesis of amino acids. So a feedback from the level of the amino acid available for protein synthesis (as represented by the availability of aminoacyl-tRNA) to the production of the enzymes may be common.

The use of the ribosome to control RNA secondary structure in response to the availability of an aminoacyl-tRNA establishes an inverse relationship between the presence of aminoacyl-tRNA and the transcription of the operon, equivalent to a situation in which aminoacyl-tRNA functions as a corepressor of transcription. Since the regulatory mechanism is mediated by changes in the formation of duplex regions, attenuation provides a striking example of the importance of secondary structure in the termination event, and of its use in regulation.

E. coli and *B. subtilis* therefore use the same types of mechanisms, involving control of mRNA structure in response to the presence or absence of a tRNA, but they have combined the individual interactions in different ways. The end result is the same: to inhibit production of the enzymes when there is an excess supply of the amino acid, and to activate production when a shortage is indicated by the accumulation of uncharged $tRNA^{Trp}$.

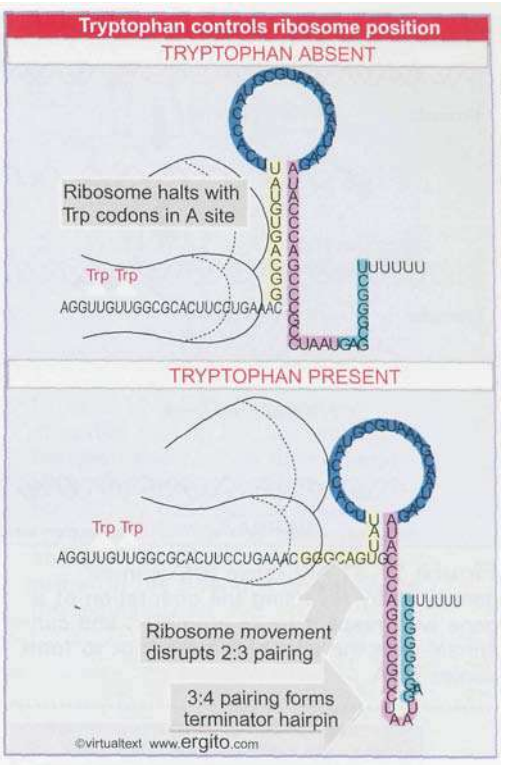


Figure 11.29 The alternatives for RNA polymerase at the attenuator depend on the location of the ribosome, which determines whether regions 3 and 4 can pair to form the terminator hairpin.

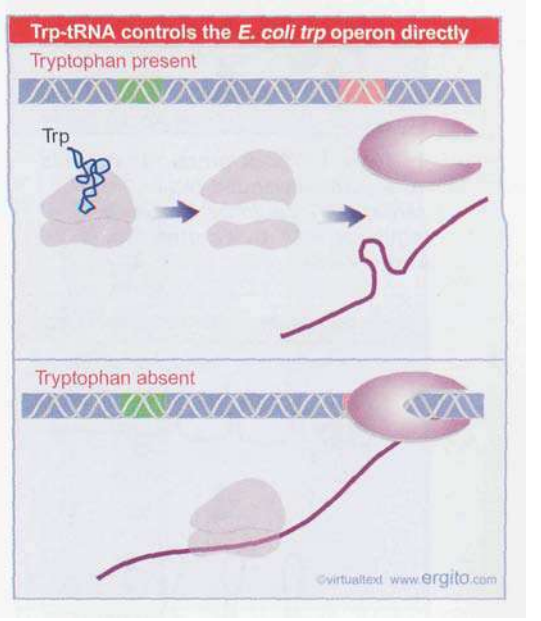


Figure 11.30 In the presence of tryptophan tRNA, ribosomes translate the leader peptide and are released. This allows hairpin formation, so that RNA polymerase terminates. In the absence of tryptophan tRNA, the ribosome is blocked, the termination hairpin cannot form, and RNA polymerase continues.

11.18 Antisense RNA can be used to inactivate gene expression

Key Concepts

- Antisense genes block expression of their targets when introduced into eukaryotic cells.

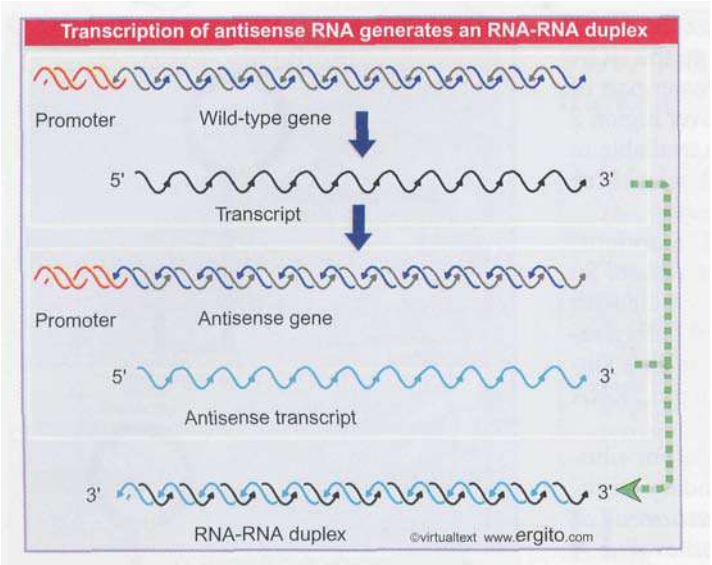


Figure 11.31 Antisense RNA can be generated by reversing the orientation of a gene with respect to its promoter, and can anneal with the wild-type transcript to form duplex RNA.

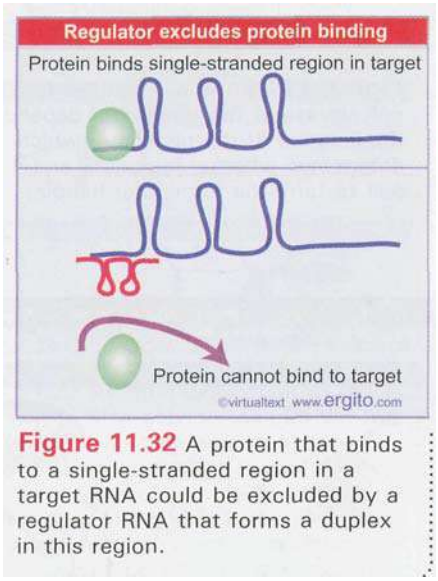


Figure 11.32 A protein that binds to a single-stranded region in a target RNA could be excluded by a regulator RNA that forms a duplex in this region.

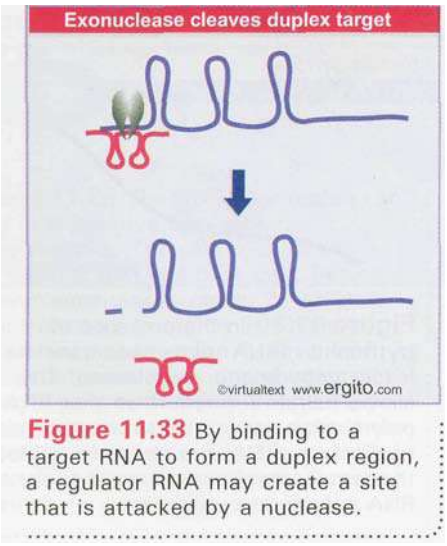


Figure 11.33 By binding to a target RNA to form a duplex region, a regulator RNA may create a site that is attacked by a nuclease.

Base pairing offers a powerful means for one RNA to control the activity of another. There are many cases in both prokaryotes and eukaryotes where a (usually rather short) single-stranded RNA base pairs with a complementary region of an mRNA, and as a result prevents expression of the mRNA. One of the early illustrations of this effect was provided by an artificial situation, in which antisense genes were introduced into eukaryotic cells.

Antisense genes are constructed by reversing the orientation of a gene with regard to its promoter, so that the "antisense" strand is transcribed, as illustrated in Figure 11.31. Synthesis of antisense RNA can inactivate a target RNA in either prokaryotic or eukaryotic cells. An antisense RNA is in effect a synthetic RNA regulator. An antisense thymidine kinase gene inhibits synthesis of thymidine kinase from the endogenous gene. Quantitation of the effect is not entirely reliable, but it seems that an excess (perhaps a considerable excess) of the antisense RNA may be necessary.

At what level does the antisense RNA inhibit expression? It could in principle prevent transcription of the authentic gene, processing of its RNA product, or translation of the messenger. Results with different systems show that the inhibition depends on formation of RNA·RNA duplex molecules, but this can occur either in the nucleus or in the cytoplasm. In the case of an antisense gene stably carried by a cultured cell, sense-antisense RNA duplexes form in the nucleus, preventing normal processing and/or transport of the sense RNA. In another case, injection of antisense RNA into the cytoplasm inhibits translation by forming duplex RNA in the 5' region of the mRNA.

This technique offers a powerful approach for turning off genes at will; for example, the function of a regulatory gene can be investigated by introducing an antisense version. An extension of this technique is to place the antisense gene under control of a promoter itself subject to regulation. Then the target gene can be turned off and on by regulating the production of antisense RNA. This technique allows investigation of the importance of the timing of expression of the target gene.

11.19 Small RNA molecules can regulate translation

Key Concepts

- A regulator RNA functions by forming a duplex region with a target RNA.
- The duplex may block initiation of translation, cause termination of transcription, or create a target for an endonuclease.

Repressors and activators are *trans-acting* proteins. Yet the formal circuitry of a regulatory network could equally well be constructed by using an RNA as regulator. In fact, the original model for the operon left open the question of whether the regulator might be RNA or protein. Indeed, the construction of synthetic antisense RNAs turns out to mimic a class of RNA regulators that is becoming of increasing importance.

Like a protein regulator, a small regulator RNA is an independently synthesized molecule that diffuses to a target site consisting of a

By Book_Crazy [IND]

specific nucleotide sequence. The target for a regulator RNA is a single-stranded nucleic acid sequence. The regulator RNA functions complementarity with its target, at which it can form a double-stranded region.

We can imagine two general mechanisms for the action of a regulator RNA:

- Formation of a duplex region with the target nucleic acid directly prevents its ability to function, by forming or sequestering a specific site. Figure 11.32 illustrates the situation in which a protein that binds to single-stranded RNA is prevented from acting by formation of a duplex. Figure 11.33 shows the opposite type of relationship in which the formation of a double-stranded region creates a target site for an endonuclease that destroys the RNA target.
- Formation of a duplex region in one part of the target molecule changes the conformation of another region, thus indirectly affecting its function. Figure 11.34 shows an example. The mechanism is essentially similar to the use of secondary structure in attenuation (see *11.14 Alternative secondary structures control attenuation*), except that the interacting regions are on different RNA molecules instead of being part of the same RNA molecule.

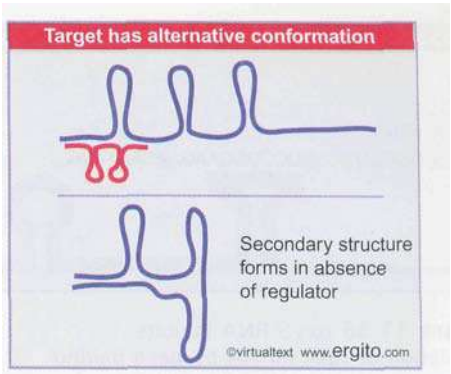


Figure 11.34 The secondary structure formed by base pairing between two regions of the target RNA may be prevented from forming by base pairing with a regulator RNA. In this example, the ability of the 3′ end of the RNA to pair with the 5′ end is prevented by the regulator.

The feature common to both types of RNA-mediated regulation is that changes in secondary structure of the target control its activity.

A difference between RNA regulators and the proteins that repress operons is that the RNA does not have allosteric properties; it cannot respond to other small molecules by changing its ability to recognize its target. It can be turned on by controlling transcription of its gene or it could be turned off by an enzyme that degrades the RNA regulator product.

11.20 Bacteria contain regulator RNAs

Key Concepts

- Bacterial regulator RNAs are called sRNAs.
- Several of the sRNAs are bound by the protein Hfq, which increases their effectiveness.
- The OxyS sRNA activates or represses expression of > 10 loci at the post-transcriptiona level.

In bacteria, regulator RNAs are short molecules, collectively known as sRNAs; *E. coli* contains at least 17 different sRNAs. Some of the sRNAs are general regulators that affect many target genes. Oxidative stress provides an interesting example of a general control system in which RNA is the regulator. When exposed to reactive oxygen species, bacteria respond by inducing antioxidant defense genes. Hydrogen peroxide activates the transcription activator OxyR, which controls the expression of several inducible genes. One of these genes is *oxyS*, which codes for a small RNA.

Figure 11.35 shows two salient features of the control of *oxyS* expression. In a wild-type bacterium under normal conditions, it is not expressed. The pair of gels on the left side of the figure shows that it is expressed at high levels in a mutant bacterium with a constitutively active *oxyR* gene. This identifies *oxyS* as a target for activation by *oxyR*. The pair of gels on the right side of the figure show that OxyS RNA is transcribed within 1 minute of exposure to hydrogen peroxide.

The OxyS RNA is a short sequence (109 nucleotides) that does not code for protein. It is a *trans*-acting regulator that affects gene

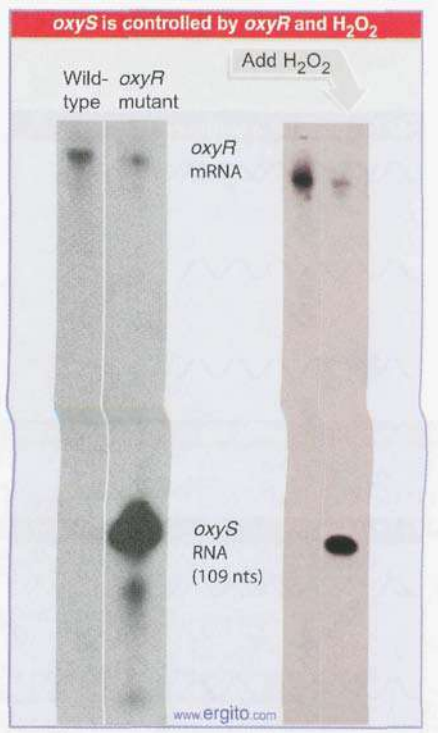


Figure 11.35 The gels on the left show that *oxyS* RNA is induced in an *oxyR* constitutive mutant. The gels on the right show that oxyS RNA is induced within 1 minute of adding hydrogen peroxide to a wild-type culture. Photograph kindly provided by Gisela Storz.

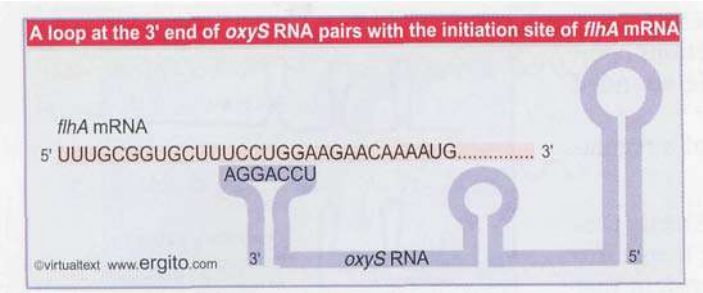


Figure 11.36 *oxyS* RNA inhibits translation of *flhA* mRNA by base pairing with a sequence just upstream of the AUG initiation codon.

expression at post-transcriptional levels. It has >10 target loci; at some of them, it activates expression, at others it represses expression. **Figure 11.36** shows the mechanism of repression of one target, the FlhA mRNA. Three stem-loop structures protrude in the secondary structure of OxyR mRNA, and the loop close to the 3' terminus is complementary to a sequence just preceding the initiation codon of FlhA mRNA. Base pairing between OxyS RNA and FlhA RNA prevents the ribosome from binding to the initiation codon, and therefore represses translation. There is also a second pairing interaction that involves a sequence within the coding region of FlhA.

Another target for *oxyS* is *rpoS,* the gene coding for an alternative sigma factor (which activates a general stress response). By inhibiting production of the sigma factor, *oxyS* ensures that the specific response to oxidative stress does not trigger the response that is appropriate for other stress conditions. The *rpoS* gene is also regulated by two other sRNAs (DsrA and RprA), which activate it. These three sRNAs appear to be global regulators that coordinate responses to various environmental conditions.

The actions of all three sRNAs are assisted by an RNA-binding protein called Hfq. The Hfq protein was originally identified as a bacterial host factor needed for replication of the RNA bacteriophage Qβ. It is related to the Sm proteins of eukaryotes that bind to many of the snRNAs (small nuclear RNAs) that have regulatory roles in gene expression (see *24.5 snRNAs are required for splicing).* Mutations in its gene have many effects, identifying it as a pleiotropic protein. Hfq binds to many of the sRNAs of *E. coli.* It increases the effectiveness of OxyS RNA by enhancing its ability to bind to its target mRNAs. The effect of Hfq is probably mediated by causing a small change in the secondary structure of OxyS RNA that improves the exposure of the single-stranded sequences that pair with the target mRNAs.

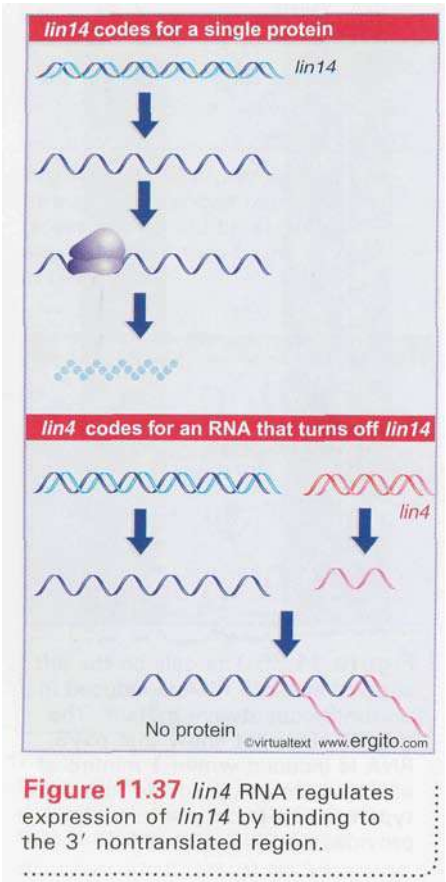


Figure 11.37 *lin4* RNA regulates expression of *lin14* by binding to the 3′ nontranslated region.

11.21 MicroRNAs are regulators in many eukaryotes

Key Concepts

- Animal and plant genomes code for many short (~22 base) RNA molecules, called microRNAs.
- MicroRNAs regulate gene expression by base pairing with complementary sequences in target mRNAs.

Very small RNAs are gene regulators in many eukaryotes. The first example was discovered in the nematode *C. elegans* as the result of the interaction between the regulator gene *Hn4* and its target gene, *Hnl4.* **Figure 11.37** illustrates the behavior of this regulatory system. The *Hnl4* target gene regulates larval development. Expression of *lin14* is controlled by *lin4,* which codes for a small transcript of 22 nucleotides. The *lin4* transcripts are complementary to a 10-base sequence that is repeated 7 times in the 3' nontranslated region of *Unl4.* Expression of *Hn4* represses expression of *lin 14* post-transcriptionally, most likely because the base pairing reaction between the two RNAs leads to degradation of the mRNA. This system is especially interesting in implicating the 3' end as a site for regulation.

By Book_Crazy [IND]

The *lin4* RNA is an example of a microRNA. There are ~55 genes in the *C. elegans* genome coding for microRNAs of 21-24 nucleotide length. They have varying patterns of expression during development and are likely to be regulators of gene expression. Many of the microRNAs of *C. elegans* are contained in a large (15S) ribonucleoprotein particle.

Many of the *C. elegans* microRNAs have homologues in mammals, so the mechanism may be widespread. They are also found in plants. Of 16 microRNAs in *Arabidopsis*, 8 are completely conserved in rice, suggesting widespread conservation of this regulatory mechanism.

The mechanism of production of the microRNAs is also widely conserved. In the example of *Un4*, the gene is transcribed into a transcript that forms a double-stranded region that becomes a target for a nuclease called Dicer. This has an N-terminal helicase activity, enabling it to unwind the double-stranded region, and two nuclease domains that are related to the bacterial ribonuclease III. Related enzymes are found in flies, worms, and plants. Cleavage of the initial transcript generates the active microRNA. Interfering with the enzyme activity blocks the production of microRNAs and causes developmental defects.

11.22 RNA interference is related to gene silencing

Key Concepts

RNA interference triggers degradation of mRNAs complementary to either strand of a short dsRNA.
dsRNA may cause silencing of host genes.

The regulation of mRNAs by microRNAs is mimicked by the phenomenon of RNA interference (RNAi). This was discovered when it was observed that antisense and sense RNAs can be equally effective in inhibiting gene expression. The reason is that preparations of either type of (supposedly) single-stranded RNA are actually contaminated by small amounts of double-stranded RNA.

Work with an *in vitro* system shows that the dsRNA is degraded by ATP-dependent cleavage to give oligonucleotides of 21-23 bases. The short RNA is sometimes called siRNA (short interfering RNA). **Figure 11.38** shows that the mechanism of cleavage involves making breaks relative to each 3' end of a long dsRNA to generate siRNA fragments with short (2 base) protruding 3' ends. The same enzyme (Dicer) that generates microRNAs is responsible for the cleavage.

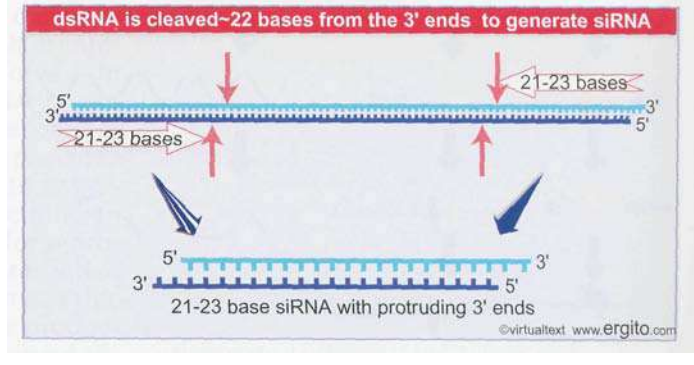


Figure 11.38 siRNA that mediates RNA interference is generated by cleaving dsRNA into smaller fragments. The cleavage reaction occurs 21-23 nucleotides from a 3' end. The siRNA product has protruding bases on its 3' ends.

RNAi occurs post-transcriptionally when an siRNA induces degradation of a complementary mRNA. **Figure 11.39** suggests that the siRNA may provide a template that directs a nuclease to degrade mRNAs that are complementary to one or both strands, perhaps by a process in which the mRNA pairs with the fragments. It is likely that a helicase is required to assist the pairing reaction. The siRNA directs cleavage of the mRNA in the middle of the paired segment. These reactions occur within a ribonucleoprotein complex called RISC (RNA-induced silencing complex).

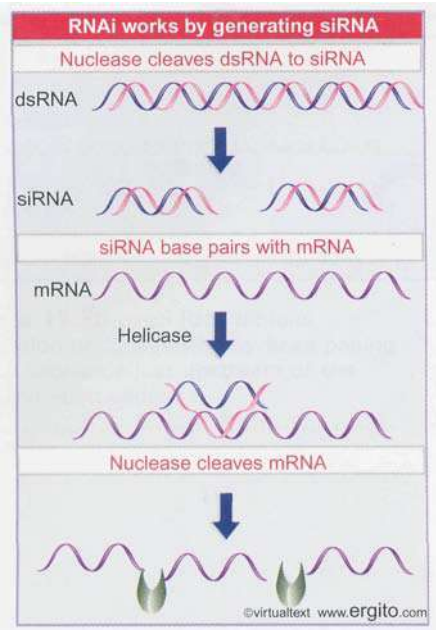


Figure 11.39 RNAi occurs when a dsRNA is cleaved into fragments that direct cleavage of the corresponding mRNA.

RNAi has become a powerful technique for ablating the expression of a specific target gene in invertebrate cells, especially in *C. elegans* and *D. melanogaster*. However, the technique has been limited in mammalian cells, which have a more generalized response to dsRNA of shutting down protein synthesis and degrading mRNA. **Figure 11.40** shows that this happens because of two reactions. The dsRNA activates the enzyme PKR, which inactivates the translation initiation factor eIF2a by phosphorylating it. And it activates 2'5' oligoadenylate synthetase, whose product activates RNAase L, which degrades all mRNAs. However, it turns out that these reactions require dsRNA that is longer than 26 nucleotides. If shorter dsRNA (21-23 nucleotides) is introduced into mammalian cells, it triggers the specific degradation of complementary RNAs just as with the RNAi technique in worms and flies. With this advance, it seems likely that RNAi will become the universal mechanism of choice for turning off the expression of a specific gene.

As an example of the progress being made with the technique, it has been possible to use RNAi for a systematic analysis of gene expression in *C. elegans*. Loss of function phenotypes can be generated by feeding worms with bacteria expressing a dsRNA that is homologous to a target gene. By making a library of bacteria in which each bacterium expresses a dsRNA corresponding to a different gene, worms have been screened for the effects of knocking out most (86%) of the genes.

RNA interference is related to natural processes in which gene expression is silenced. Plants and fungi show RNA silencing (sometimes called post-transcriptional gene silencing) in which dsRNA inhibits expression of a gene. The most common source of the RNA is a replicating virus. This mechanism may have evolved as a defense against viral infection. When a virus infects a plant cell, the formation of dsRNA triggers the suppression of expression from the plant genome. RNA silencing has the further remarkable feature that it is not limited to the cell in which the viral infection occurs: it can spread throughout the plant systemically. Presumably the propagation of the signal involves passage of RNA or fragments of RNA. It may require some of the same features that are involved in movement of the virus itself. It is possible that RNA silencing involves an amplification of the signal by an RNA-dependent RNA synthesis process in which a novel polymerase uses the siRNA as a primer to synthesize more RNA on a template of complementary RNA.

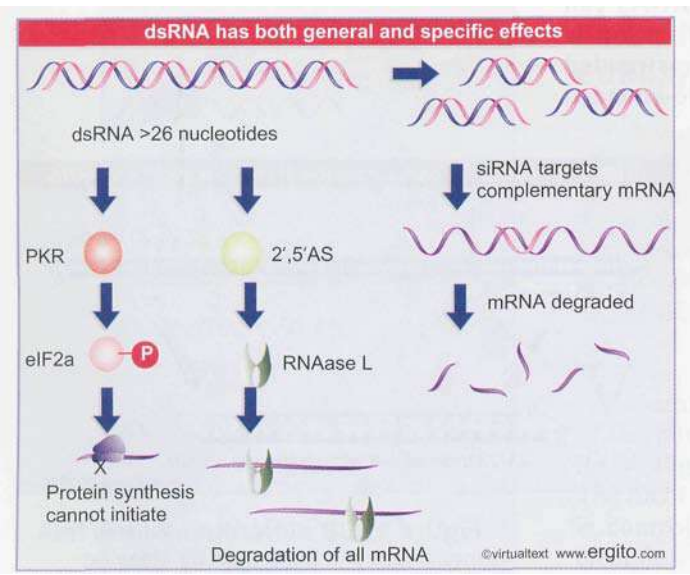


Figure 11.40 dsRNA inhibits protein synthesis and triggers degradation of all mRNA in mammalian cells as well as having sequence-specific effects.

A related process is the phenomenon of cosuppression in which introduction of a transgene causes the corresponding endogenous gene to be silenced. This has been largely characterized in plants. The implication is that the transgene must make both antisense and sense RNA copies, and this inhibits expression of the endogenous gene.

Silencing takes place by RNA-RNA interactions. It is also possible that dsRNA may inhibit gene expression by interacting with the DNA. If a DNA copy of a viroid RNA sequence is inserted into a plant genome, it becomes methylated when the viroid RNA replicates. This suggests that the RNA sequence could be inducing methylation of the DNA sequence. Similar targeting of methylation of DNA corresponding to sequences represented in dsRNA has been detected in plant cells. Methylation of DNA is associated with repression of transcription, so this could be another means of silencing genes represented in dsRNA (see *21.18 Gene expression is associated with demethylation*). Nothing is known about the mechanism.

11.23 Summary

Gene expression can be regulated positively by factors that activate a gene or negatively by factors that repress a gene. The first and most common level of control is at the initiation of transcription, but termination of transcription may be also be controlled. Translation may be controlled by regulators that interact with mRNA. The regulatory products may be proteins, which often are controlled by allosteric interactions in response to the environment, or RNAs, which function by base pairing with the target RNA to change its secondary structure. Regulatory networks can be created by linking regulators so that the production or activity of one regulator is controlled by another.

Bacteria respond to the supply of glucose by repressing the production of the enzyme systems that catabolize alternative carbon sources. Inducer exclusion is a major component of the response, and works by inhibiting the uptake of the other sugars into the bacterium, with the result that the operons coding for their enzyme systems fail to be turned on. Increase in glucose levels also may lead to a reduction in the level of the small nucleotide cyclic AMP, although this is now controversial.

Some promoters cannot be recognized by RNA polymerase (or are recognized only poorly) unless a specific activator protein is present. Activator proteins also may be regulated by small molecules. The CRP activator becomes able to bind to target sequences in the presence of cyclic AMP. All promoters that respond to CRP have at least one copy of the target sequence. Binding of CRP to its target involves bending DNA. Direct contact between one subunit of CRP and RNA polymerase is required to activate transcription.

A common means for controlling translation is for a regulator protein to bind to a site on the mRNA that overlaps the ribosome binding site at the initiation codon. This prevents ribosomes from initiating translation. RegA of T4 is a general regulator that functions on several target mRNAs at the level of translation. Most proteins that repress translation possess this capacity in addition to other functional roles; in particular, translation is controlled in some cases of autogenous regulation, when a gene product regulates expression of the operon containing its own gene.

The level of protein synthesis itself provides an important coordinating signal. Deficiency in aminoacyl-tRNA causes an idling reaction on the ribosome, which leads to the synthesis of the unusual nucleotide ppGpp. This is an effector that inhibits initiation of transcription at certain promoters; it also has a general effect in inhibiting elongation on all templates.

Attenuation is a mechanism that relies on regulation of termination to control transcription through bacterial operons. It is commonly used in operons that code for enzymes involved in biosynthesis of an amino acid. The polycistronic mRNA of the operon starts with a sequence that can form alternative secondary structures. One of the structures has a hairpin loop that provides an intrinsic terminator upstream of the structural genes; the alternative structure lacks the hairpin. Various types of interaction can be used to determine whether the hairpin forms. One is for a protein to bind to the mRNA to prevent formation of the alternative structure. In the *trp* operon of *B. subtilis,* the TRAP protein has this function; it is controlled by the anti-TRAP protein, whose production in turn is controlled by the level of uncharged aminoacyl-tRNA^{Trp}. In the *trp* operon of *E. coli,* the choice of which structure forms is controlled by the progress of translation through a short leader sequence that includes codons for the amino acid(s) that are the product of the system. In the presence of aminoacyl-tRNA bearing such amino acid(s), ribosomes translate the leader peptide, allowing a secondary structure to form that supports termination. In the absence of this aminoacyl-tRNA, the ribosome stalls,

resulting in a new secondary structure in which the hairpin needed for termination cannot form. The supply of aminoacyl-tRNA therefore (inversely) controls amino acid biosynthesis.

Small regulator RNAs are found in both bacteria and eukaryotes. *E. coli* has ~17 sRNA species. The oxyS sRNA controls about 10 target loci at the post-transcriptional level; some of them are repressed, and others are activated. Repression is caused when the sRNA binds to a target mRNA to form a duplex region that includes the ribosome-binding site. MicroRNAs are ~22 bases long and are produced in many eukaryotes by cleavage of a longer transcript. They function by base pairing with target mRNAs to form duplex regions that are susceptible to cleavage by endonucleases. The degradation of the mRNA prevents its expression. The technique of RNA interference is becoming the method of choice for inactivating eukaryotic genes. It uses the introduction of short dsRNA sequences with one strand complementary to the target RNA, and it works by inducing degradation of the targets. This may be related to a natural defense system in plants called RNA silencing.

References

11.3 Glucose repression controls use of carbon sources

rev Meadow, N. D., Fox, D. K., and Roseman, S. (1990). The bacterial phosphoenolpyruvate: glycose phosphotransferase system. Ann. Rev. Biochem. 59, 497-542.

Stalke, J. and Hillen, W. (2000). Regulation of carbon catabolism in *Bacillus* species. Ann. Rev. Microbiol. 54, 849-880.

11.5 CRP functions in different ways in different target operons

rev Botsford, J. L. and Harman, J. G. (1992). Cyclic AMP in prokaryotes. Microbiol. Rev. 56, 100-122.

Kolb, A. (1993). Transcriptional regulation by cAMP and its receptor protein. Ann. Rev. Biochem. 62, 749-795.

ref Niu, W., Kim, Y., Tau, G., Heyduk, T., and Ebright, R. H. (1996). Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. Cell 87, 1123-1134.

Zhou, Y., Busby, S., and Ebright, R. H. (1993). Identification of the functional subunit of a dimeric transcription activator protein by use of oriented heterodimers. Cell 73, 375-379.

Zhou, Y., Merkel, T. J., and Ebright, R. H. (1994). Characterization of the activating region of *E. coli* catabolite gene activator protein (CAP). II. Role at Class I and class II CAP-dependent promoters. J. Mol. Biol. 243, 603-610.

11.6 CRP bends DNA

ref Gaston, K. A. et al. (1990). Stringent spacing requirements for transcription activation by CRP. Cell 62, 733-743.

11.7 The stringent response produces (p)ppGpp

rev Cashel, M. and Rudd, K. E. (1987). The stringent response In *E. coli* and *S. typhimurium*In E. coli and S. typhimurium: Cellular and Molecular Biology, Ed. F. C. Neidhardt, American Society for Microbiology, Washington DC 1410-1429.

ref Cashel, M. and Gallant, J. (1969). Two compounds implicated in the function of the RC gene of *E. coli.* Nature 221, 838-841.

11.8 (p)ppGpp is produced by the ribosome

ref Haseltine, W. A. and Block, R. (1973). Synthesis of guanosine tetra and pentaphosphate requires the presence of a codon specific uncharged tRNA in the acceptor site of ribosomes. Proc. Nat. Acad. Sci. USA 70, 1564-1568.

11.9 ppGpp has many effects

rev Condon, C, Squires, C, and Squires, C. L. (1995). Control of rRNA transcription in *E. coli.* Microbiol. Rev. 59, 623-645.

11.11 r-protein synthesis is controlled by autogenous regulation

rev Nomura, M. et al. (1984). Regulation of the synthesis of ribosomes and ribosomal components. Ann. Rev. Biochem. 53, 75-117.

ref Baughman, G. and Nomura, M. (1983). Localization of the target site for translational regulation of the L11 operon and direct evidence for translational coupling in *E. coli.* Cell 34, 979-988.

11.13 Autogenous regulation is often used to control synthesis of macromolecular assemblies

rev Gold, L. (1988). Posttranscriptional regulatory mechanisms in *E. coli.* Ann. Rev. Biochem. 57, 199-223.

11.15 Termination of *B. subtilis trp* genes is controlled by tryptophan and by $tRNA^{Trp}$

rev Gollnick, P. (1994). Regulation of the *B. subtilis trp* operon by an RNA-binding protein. Mol. Microbiol. 11, 991-997.

ref Antson, A. A. et al. (1999). Structure of the *trp* RNA-binding attenuation protein, TRAP, bound to RNA. Nature 401, 235-242.

Babitzke, P. and Yanoksy, C. (1993). Reconstitution of *B. subtilis trp* attenuation *in vitro* with TRAP, the *trp* RNA-binding attenuation protein. Proc. Nat. Acad. Sci. USA 90, 133-137.

Otridge, J. and Gollnick, P. (1993). MtrB from *B. subtilis* binds specifically to *trp* leader RNA in a tryptophan-dependent manner. Proc. Nat. Acad. Sci. USA 90, 128-132.

Valbuzzi, A. and Yanofsky, C. (2001). Inhibition of the B. subtilis regulatory protein TRAP by the TRAP-inhibitory protein, AT. Science 293, 2057-2059.

11.16 The *E. coli* tryptophan operon is controlled by attenuation

rev Yanofsky, C. (1981). Attenuation in the control of expression of bacterial operons. Nature 289, 751-758.

11.17 Attenuation can be controlled by translation

rev Bauer, C. E. et al. (1983). Attenuation in bacterial operons. In Gene Function in Prokaryotes, Eds. J. Beckwith, J. E. Davies, and J. A. Gallant. Cold Spring Harb 65-89.

Landick, R. and Yanofsky, C. (1987). Transcription attenuation In *E. coli* and *S. typhimurium*In E. coli and S. typhimurium: Cellular and Molecular Biology, Ed. F. C. Neidhardt, American Society for Microbiology, Washington DC 1276-1301.

Yanofsky, C. and Crawford, I. P. (1987). The tryptophan operon. In E. coli and S. typhimurium: Cellular and Molecular Biology, Ed. F. C. Neidhardt, American Society for Microbiology, Washington DC 1453-1472.

ref Lee, F. and Yanofsky, C. (1977). Transcription termination at the *trp* operon attenuators of *E. coli* and S. *typhimurium:* RNA secondary structure and regulation of termination. Proc. Nat. Acad. Sci. USA 74, 4365-4368.

Zurawski, G. et al. (1978). Translational control of transcription termination at the attenuator of the *E. coli tryptophan* operon. Proc. Nat. Acad. Sci. USA 75, 5988-5991.

11.18 Antisense RNA can be used to inactivate gene expression

ref Izant, J. G. and Weintraub, H. (1984). Inhibition of thymidine kinase gene expression by antisense RNA: a molecular approach to genetic analysis. Cell 36, 1007-1015.

11.20 Bacteria contain regulator RNAs

rev Gottesman, S. (2002). Stealth regulation: biological circuits with small RNA switches. Genes Dev. 16, 2829-2842.

ref Altuvia, S., Zhang, A., Argaman, L., Tiwari, A., and Storz, G. (1998). The *E. coli* OxyS regulatory RNA represses *fhlA* translation by blocking ribosome binding. EMBO J. 17, 6069-6075.

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., and Storz, G. (1997). A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. Cell 90, 43-53.

Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G., and Valentin-Hansen, P. (2002). Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Mol. Cell 9, 23-30.

Wassarman, K. M., Repoila, F., Rosenow, C, Storz, G., and Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. 15, 1637-1651.

Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C, and Storz, G. (2002). The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. Mol. Cell 9, 1 1-22.

11.21 MicroRNAs are regulators in many eukaryotes

ref Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363-366.

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., and Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans.* Genes Dev. 15, 2654-2659.

Lau, N. C, Lim, l. e. E. P., Weinstein, E. G., and Bartel, d. a. V. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *C. elegans.* Science 294, 858-862.

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in *C. elegans.* Science 294, 862-864.

Lee, R. C, Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14.* Cell 75, 843-854.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. 16, 720-728.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. Genes Dev. 16, 1616-1626.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans.* Cell 75, 855-862:

1 1 .22 RNA interference is related to gene silencing

rev Matzke, M., Matzke, A. J., and Kooter, J. M. (2001). RNA: guiding gene silencing. Science 293, 1080-1083.

Schwartz, D. S. and Zamore, P. D. (2002). Why do miRNAs live in the miRNP? Genes Dev. 16, 1025-1031.

Sharp, P. A. (2001). RNA interference—2001. Genes Dev. 15, 485-490.

Ahlquist, P. (2002). RNA-Dependent RNA Polymerases, Viruses, and RNA Silencing. Science 296, 1270-1273.

Tijsterman, M., Ketting, R. F., and Plasterk, R. H. (2002). The genetics of RNA silencing. Ann. Rev. Genet. 36, 489-519.

ref Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl. T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411, 494-498.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *C. elegans.* Nature **391,** 806-811.

Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 286, 950-952.

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P., and Ahringer, J. (2003). Systematic functional analysis of the *C. elegans* genome using RNAi. Nature 421, 231-237.

Mette, M. F., Aufsatz, W., van der Winden, J., Matzke, M. A., and Matzke, A. J. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. EMBO J. 19, 5194-5201.

Montgomery, M. K., Xu, S., and Fire, A. (1998). RNA as a target of double-stranded RNA-mediated genetic interference in *C. elegans.* Proc. Nat. Acad. Sci. USA 95, 15502-15507.

Ngo, H., Tschudi, C, Gull, K., and Ullu, E. (1998). Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei.* Proc. Nat. Acad. Sci. USA 95, 14687-14692.

Voinnet, 0., Pinto, Y. M., and Baulcombe, D. C. (1999). Suppression of gene silencing: a general strategy used by diverse DNA and RNA viruses of plants. Proc. Nat. Acad. Sci. USA 96, 14147-14152.

Wassenegger, M., Heimes, S., Riedel, L, and Sanger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell 76, 567-576.

Waterhouse, P. M., Graham, M. W., and Wang, M. B. (1998). Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. Proc. Nat. Acad. Sci. USA 95, 13959-13964.

Zamore, P. D., Tuschl, T., Sharp, P. A., and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell 101, 25-33.

Chapter 12

Phage strategies

12.1 Introduction

Some phages have only a single strategy for survival. On infecting a susceptible host, they subvert its functions to the purpose of producing a large number of progeny phage particles. As the result of this lytic infection, the host bacterium dies. In the typical lytic cycle, the phage DNA (or RNA) enters the host bacterium, its genes are transcribed in a set order, the phage genetic material is replicated, and the protein components of the phage particle are produced. Finally, the host bacterium is broken open (*lysed*) to release the assembled progeny particles by the process of lysis.

Other phages have a dual existence. They are able to perpetuate themselves via the same sort of lytic cycle in what amounts to an open strategy for producing as many copies of the phage as rapidly as possible. But they also have an alternative form of existence, in which the phage genome is present in the bacterium in a latent form known as prophage. This form of propagation is called lysogeny.

In a lysogenic bacterium, the prophage is inserted into the bacterial genome, and is inherited in the same way as bacterial genes. The process by which it is converted from an independent phage genome into a prophage that is a linear part of the bacterial genome is described as integration. By virtue of its possession of a prophage, a lysogenic bacterium has immunity against infection by further phage particles of the same type. Immunity is established by a single integrated prophage, so usually a bacterial genome contains only one copy of a prophage of any particular type.

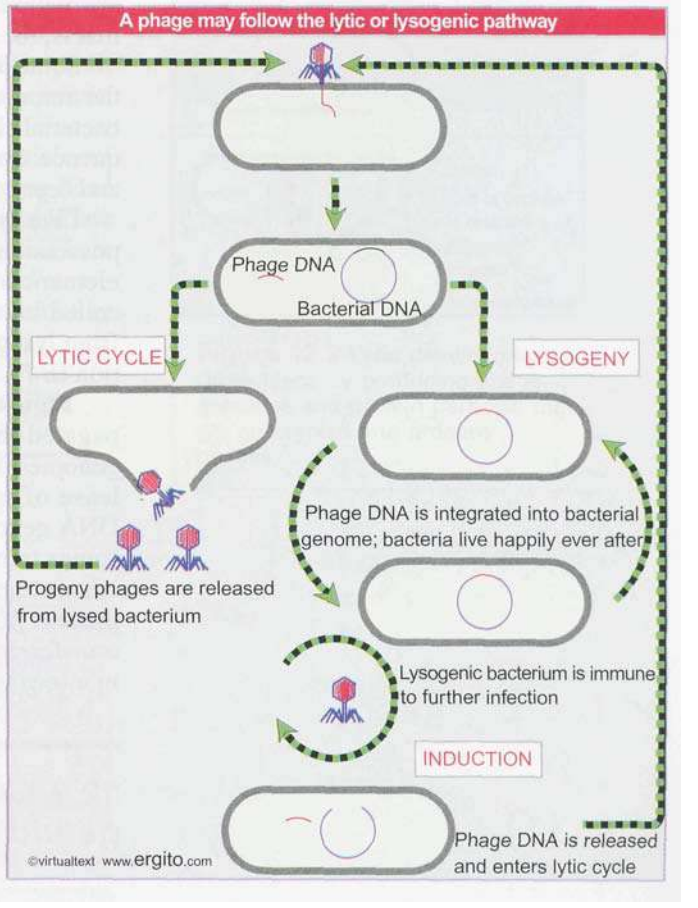


Figure 12.1 Lytic development involves the reproduction of phage particles with destruction of the host bacterium, but lysogenic existence allows the phage genome to be carried as part of the bacterial genetic information.

Transitions occur between the lysogenic and lytic modes of existence. **Figure 12.1** shows that when a phage produced by a lytic cycle enters a new bacterial host cell, it either repeats the lytic cycle or enters the lysogenic state. The outcome depends on the conditions of infection and the genotypes of phage and bacterium.

Figure 12.2 Several types of independent genetic units exist in bacteria.

Phages and plasmids live in bacteria

Type of Unit	Genome Structure	Mode of Propagation	Consequences
Lytic phage	ds- or ss-DNA or RNA linear or circular	Infects susceptible host	Usually kills host
Lysogenic phage	ds-DNA	Linear sequence in host chromosome	Immunity to infection
Plasmid	ds-DNA circle	Replicates at defined copy number May be transmissible	Immunity to plasmids in same group
Episome	ds-DNA circle	Free circle or linear integrated	May transfer host DNA

A prophage is freed from the restrictions of lysogeny by the process called induction. First the phage DNA is released from the bacterial chromosome by excision; then the free DNA proceeds through the lytic pathway.

The alternative forms in which these phages are propagated are determined by the regulation of transcription. Lysogeny is maintained by the interaction of a phage repressor with an operator. The lytic cycle requires a cascade of transcriptional controls. And the transition between the two life-styles is accomplished by the establishment of repression (lytic cycle to lysogeny) or by the relief of repression (induction of lysogen to lytic phage).

Another type of existence within bacteria is represented by plasmids. These are autonomous units that exist in the cell as extrachromosomal genomes. Plasmids are self-replicating circular molecules of DNA that are maintained in the cell in a stable and characteristic number of copies; that is, the number remains constant from generation to generation.

Some plasmids also have alternative life-styles. They can exist either in the autonomous extrachromosomal state; or they can be inserted into the bacterial chromosome, and then are carried as part of it like any other sequence. Such units are properly called episomes (but the terms "plasmid" and "episome" are sometimes used loosely as though interchangeable).

Like lysogenic phages, plasmids and episomes maintain a selfish possession of their bacterium and often make it impossible for another element of the same type to become established. This effect also is called immunity, although the basis for plasmid immunity is different from lysogenic immunity. (We discuss the control of plasmid perpetuation in *13 The replicon.)*

Figure 12.2 summarizes the types of genetic units that can be propagated in bacteria as independent genomes. Lytic phages may have genomes of any type of nucleic acid; they transfer between cells by release of infective particles. Lysogenic phages have double-stranded DNA genomes, as do plasmids and episomes. Some plasmids and episomes transfer between cells by a conjugative process (involving direct contact between donor and recipient cells). A feature of the transfer process in both cases is that on occasion some bacterial host genes are transferred with the phage or plasmid DNA, so these events play a role in allowing exchange of genetic information between bacteria.

12.2 Lytic development is divided into two periods

Key Concepts

- A phage infective cycle is divided into the early period (before replication) and the late period (after the onset of replication).
- A phage infection generates a pool of progeny phage genomes that replicate and recombine.

By Book_Crazy [IND]

Phage genomes of necessity are small. As with all viruses, they are restricted by the need to package the nucleic acid within the protein coat. This limitation dictates many of the viral strategies for reproduction. Typically a virus takes over the apparatus of the host cell, which then replicates and expresses phage genes instead of the bacterial genes.

Usually the phage includes genes whose function is to ensure preferential replication of phage DNA. These genes are concerned with the initiation of replication and may even include a new DNA polymerase. Changes are introduced in the capacity of the host cell to engage in transcription. They involve replacing the RNA polymerase or modifying its capacity for initiation or termination. The result is always the same: phage mRNAs are preferentially transcribed. So far as protein synthesis is concerned, usually the phage is content to use the host apparatus, redirecting its activities principally by replacing bacterial mRNA with phage mRNA.

Lytic development is accomplished by a pathway in which the phage genes are expressed in a particular order. This ensures that the right amount of each component is present at the appropriate time. The cycle can be divided into the two general parts illustrated in **Figure 12.3:**

- Early infection describes the period from entry of the DNA to the start of its replication.
- Late infection defines the period from the start of replication to the final step of lysing the bacterial cell to release progeny phage particles.

The early phase is devoted to the production of enzymes involved in the reproduction of DNA. These include the enzymes concerned with DNA synthesis, recombination, and sometimes modification. Their activities cause a *pool* of phage genomes to accumulate. In this pool, genomes are continually replicating and recombining, so that *the events of a single lytic cycle concern a population of phage genomes.*

During the late phase, the protein components of the phage particle are synthesized. Often many different proteins are needed to make up head and tail structures, so the largest part of the phage genome consists of late functions. In addition to the structural proteins, "assembly proteins" are needed to help construct the particle, although they are not themselves incorporated into it. By the time the structural components are assembling into heads and tails, replication of DNA has reached its maximum rate. The genomes then are inserted into the empty protein heads, tails are added, and the host cell is lysed to allow release of new viral particles.

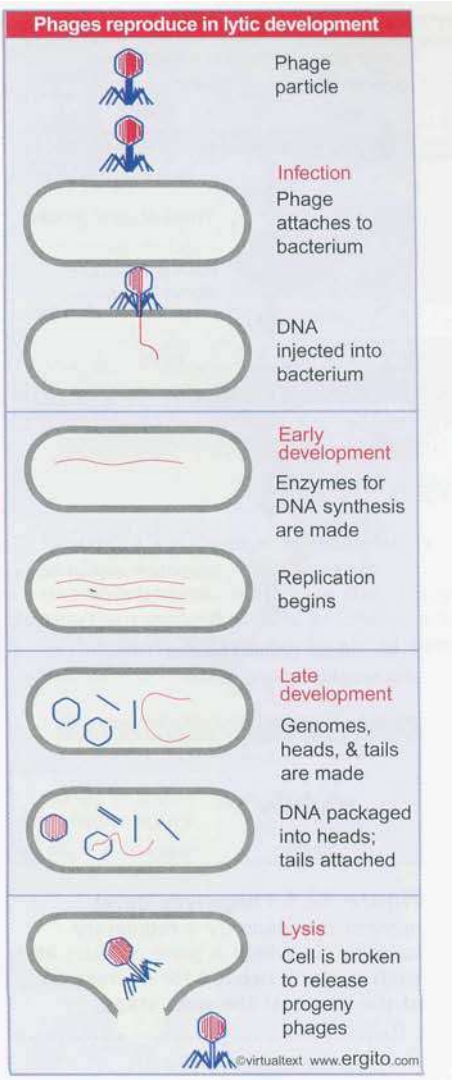


Figure 12.3 Lytic development takes place by producing phage genomes and protein particles that are assembled into progeny phages.

12.3 Lytic development is controlled by a cascade

Key Concepts

- The early genes transcribed by host RNA polymerase following infection include or comprise regulators required for expression of the middle set of phage genes.
- The middle group of genes include regulators to transcribe the late genes.
- This results in the ordered expression of groups of genes during phage infection.

The organization of the phage genetic map often reflects the sequence of lytic development. The concept of the operon is taken to somewhat of an extreme, in which the genes coding for proteins with related functions are clustered to allow their control with the maximum economy. This allows the pathway of lytic development to be controlled with a small number of regulatory switches.

By Book_Crazy [IND]

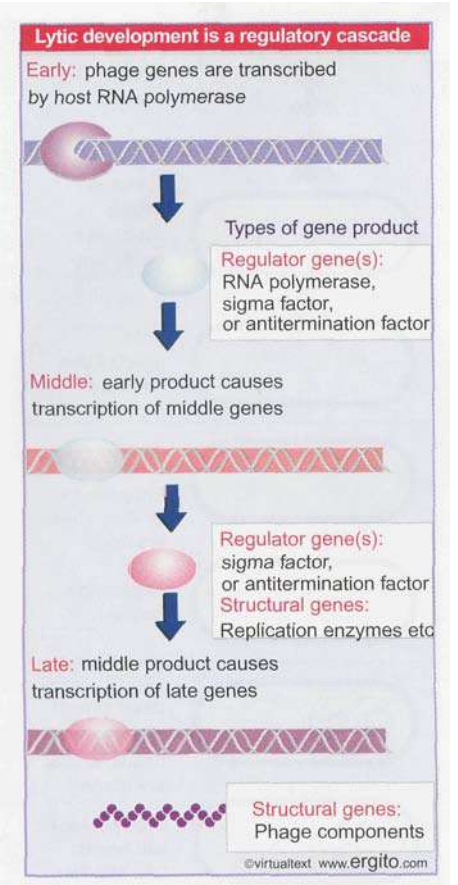


Figure 12.4 Phage lytic development proceeds by a regulatory cascade, in which a gene product at each stage is needed for expression of the genes at the next stage.

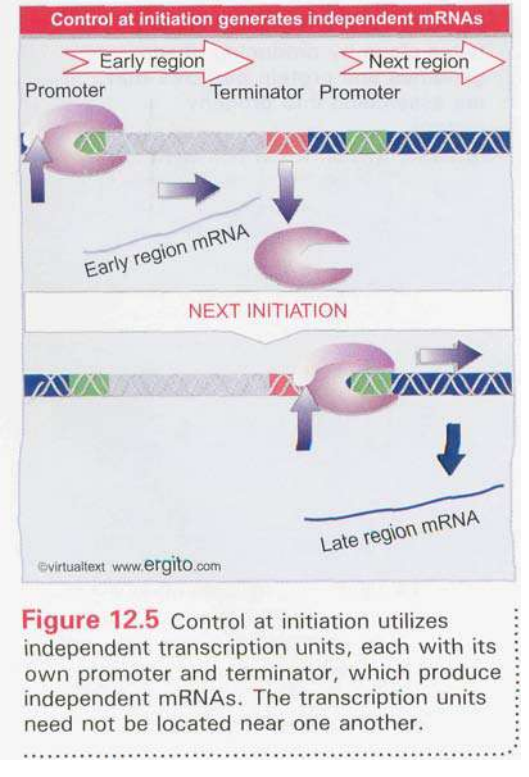


Figure 12.5 Control at initiation utilizes independent transcription units, each with its own promoter and terminator, which produce independent mRNAs. The transcription units need not be located near one another.

The lytic cycle is under positive control, so that each group of phage genes can be expressed only when an appropriate signal is given. **Figure 12.4** shows that the regulatory genes function in a cascade, in which a gene expressed at one stage is necessary for synthesis of the genes that are expressed at the next stage.

The first stage of gene expression necessarily relies on the transcription apparatus of the host cell. Usually only a few genes are expressed at this stage. Their promoters are indistinguishable from those of host genes. The name of this class of genes depends on the phage. In most cases, they are known as the early genes. In phage lambda, they are given the evocative description of immediate early. Irrespective of the name, they constitute only a preliminary, representing just the initial part of the early period. Sometimes they are exclusively occupied with the transition to the next period. At all events, *one of these genes always codes for a protein that is necessary for transcription of the next class of genes.*

This second class of genes is known variously as the delayed early or middle gene group. Its expression typically starts as soon as the regulator protein coded by the early gene(s) is available. Depending on the nature of the control circuit, the initial set of early genes may or may not continue to be expressed at this stage. If control is at initiation, the two events are independent (see Figure 12.5), and early genes can be switched off when middle genes are transcribed. If control is at termination, the early genes must continue to be expressed (see Figure 12.6). Often the expression of host genes is reduced. Together the two sets of early genes account for all necessary phage functions except those needed to assemble the particle coat itself and to lyse the cell.

When the replication of phage DNA begins, it is time for the late genes to be expressed. Their transcription at this stage usually is arranged by embedding a further regulator gene within the previous (delayed early or middle) set of genes. This regulator may be another antitermination factor (as in lambda) or it may be another sigma factor (as in SPO1).

A lytic infection often falls into three stages, as shown in Figure 12.4. The first stage consists of early genes transcribed by host RNA polymerase (sometimes the regulators are the only products at this stage). The second stage consists of genes transcribed under direction of the regulator produced in the first stage (most of these genes code for enzymes needed for replication of phage DNA). The final stage consists of genes for phage components, transcribed under direction of a regulator synthesized in the second stage.

The use of these successive controls, in which each set of genes contains a regulator that is necessary for expression of the next set, creates a cascade in which groups of genes are turned on (and sometimes off) at particular times. The means used to construct each phage cascade are different, but the results are similar, as the following sections show.

12.4 Two types of regulatory event control the lytic cascade

Key Concepts

- Regulator proteins used in phage cascades may sponsor initiation at new (phage) promoters or cause the host polymerase to read through transcription terminators.

At every stage of phage expression, one or more of the active genes is a regulator that is needed for the subsequent stage. The regulator may take the form of a new RNA polymerase, a sigma factor that redirects the specificity of the host RNA polymerase (see *9.18 Sigma factors*

may be organized into cascades), or an antitermination factor that allows it to read a new group of genes (see *9.23 Antitermination is a regulatory event).* The next two figures compare the use of switching at initiation or termination to control gene expression.

One mechanism for recognizing new phage promoters is to replace the sigma factor of the host enzyme with another factor that redirects its specificity in initiation (see Figure 9.32). An alternative mechanism is to synthesize a new phage RNA polymerase. In either case, the critical feature that distinguishes the new set of genes is their possession of *different promoters from those originally recognized by host RNA polymerase.* **Figure 12.5** shows that the two sets of transcripts are independent; as a consequence, early gene expression can cease after the new sigma factor or polymerase has been produced.

Antitermination provides an alternative mechanism for phages to control the switch from early genes to the next stage of expression. The use of antitermination depends on a particular arrangement of genes. **Figure 12.6** shows that the early genes lie adjacent to the genes that are to be expressed next, but are separated from them by terminator sites. *If termination is prevented at these sites, the polymerase reads through into the genes on the other side.* So in antitermination, the *same promoters* continue to be recognized by RNA polymerase. So the new genes are expressed only by extending the RNA chain to form molecules that contain the early gene sequences at the 5′ end and the new gene sequences at the 3′ end. Since the two types of sequence remain linked, early gene expression inevitably continues.

The regulator gene that controls the switch from immediate early to delayed early expression in phage lambda is identified by mutations in gene *N* that can transcribe *only* the immediate early genes; they proceed no further into the infective cycle (see Figure 9.53). The same effect is seen when gene *28* of phage SPO1 is mutated to prevent the production of σ^{gp28} (see Figure 9.41). From the genetic point of view, the mechanisms of new initiation and antitermination are similar. *Both are positive controls in which an early gene product must be made by the phage in order to express the next set of genes.* By employing either sigma factors or antitermination proteins with different specificities, a cascade for gene expression can be constructed.

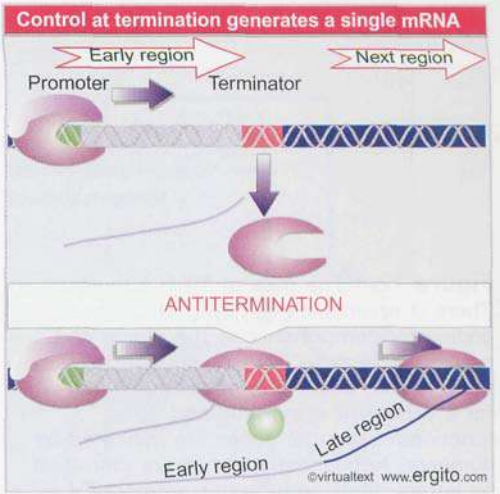


Figure 12.6 Control at termination requires adjacent units, so that transcription can read from the first gene into the next gene. This produces a single mRNA that contains both sets of genes.

12.5 The T7 and T4 genomes show functional clustering

Key Concepts

- Genes concerned with related functions are often clustered.
- Phages T7 and T4 are examples of regulatory cascades in which phage infection is divided into three periods.

The genome of phage T7 has three classes of genes, each constituting a group of adjacent loci. As **Figure 12.7** shows, the class I genes are the immediate early type, expressed by host RNA polymerase as soon as the phage DNA enters the cell. Among the products of these genes are a phage RNA polymerase and enzymes that interfere with host gene expression. The phage RNA polymerase is responsible for expressing the class II genes (concerned principally with DNA synthesis functions) and the class III genes (concerned with assembling the mature phage particle).

T4 has one of the larger phage genomes (165 kb), organized with extensive functional grouping of genes. **Figure 12.8** presents the genetic map. *Essential genes* are numbered: a mutation in any one of these loci prevents successful completion of the lytic cycle. *Nonessential genes* are indicated by three-letter abbreviations. (They are defined as nonessential

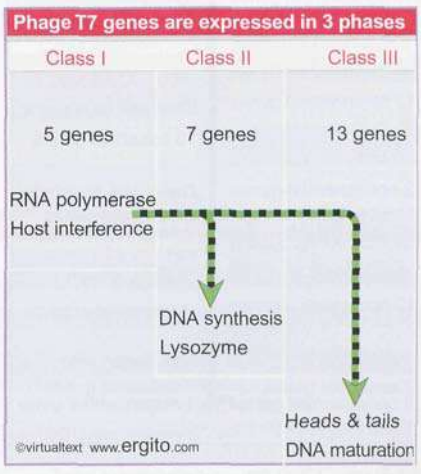


Figure 12.7 Phage T7 contains three classes of genes that are expressed sequentially. The genome is ~38 kb.

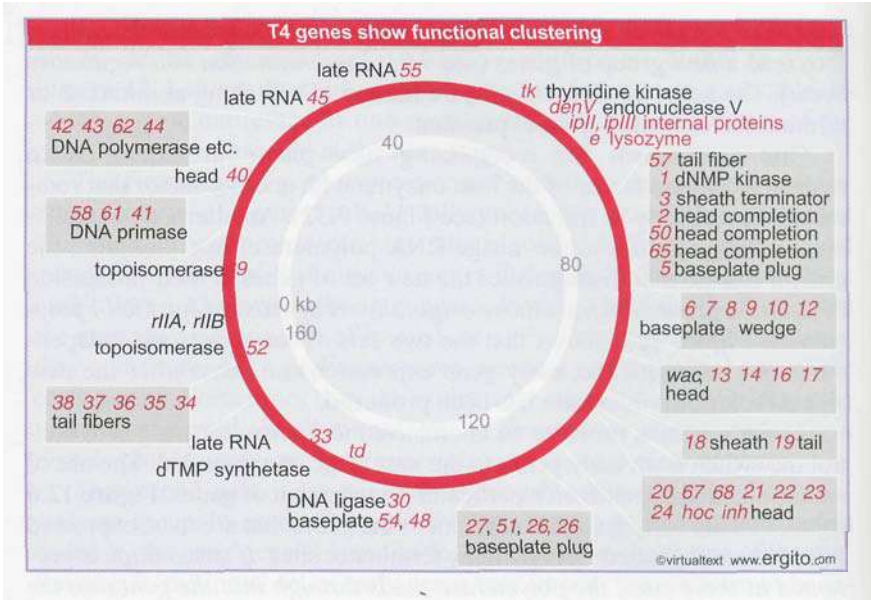


Figure 12.8 The map of T4 is circular. There is extensive clustering of genes coding for components of the phage and processes such as DNA replication, but there is also dispersion of genes coding for a variety of enzymatic and other functions. Essential genes are indicated by numbers. Nonessential genes are identified by letters. Only some representative T4 genes are shown on the map.

under the usual conditions of infection. We do not really understand the inclusion of many nonessential genes, but presumably they confer a selective advantage in some of T4's habitats. In smaller phage genomes, most or all of the genes are essential.)

There are three phases of gene expression. A summary of the functions of the genes expressed at each stage is given in **Figure 12.9.** The early genes are transcribed by host RNA polymerase. The middle genes are also transcribed by host RNA polymerase, but two phage-encoded products, MotA and AsiA, are also required. The middle promoters lack a consensus —30 sequence, and instead have a binding sequence for MotA. The phage protein is an activator that compensates for the deficiency in the promoter by assisting host RNA polymerase to bind. (This is similar to a mechanism employed by phage lambda, which is illustrated later in Figure 12.28.) The early and middle genes account for virtually all of the phage functions concerned with the synthesis of DNA, modifying cell structure, and transcribing and translating phage genes.

The two essential genes in the "transcription" category fulfill a regulatory function: their products are necessary for late gene expression. Phage T4 infection depends on a mechanical link between replication and late gene expression. Only actively replicating DNA can be used as template for late gene transcription. The connection is generated by introducing a new sigma factor and also by making other modifications in the host RNA polymerase so that it is active only with a template of replicating DNA. This link establishes a correlation between the synthesis of phage protein components and the number of genomes available for packaging.

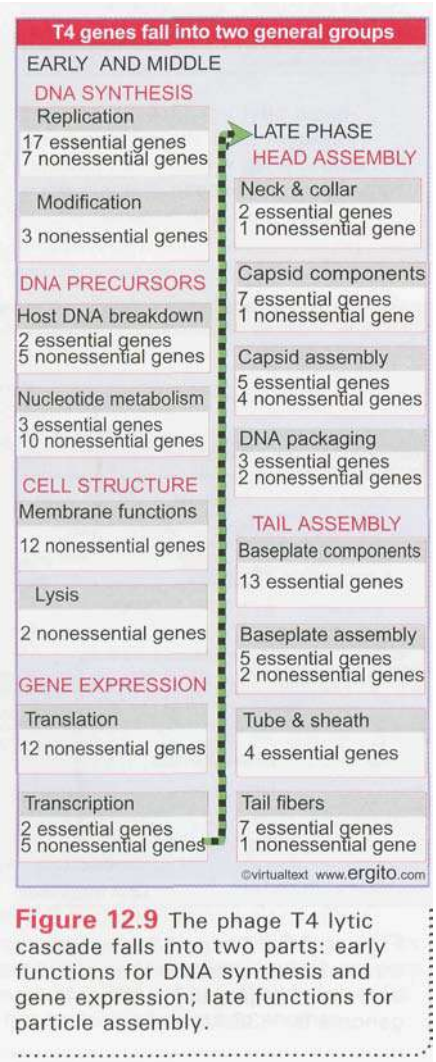


Figure 12.9 The phage T4 lytic cascade falls into two parts: early functions for DNA synthesis and gene expression; late functions for particle assembly.

12.6 Lambda immediate early and delayed early genes are needed for both lysogeny and the lytic cycle

Key Concepts

- Lambda has two immediate early genes, *N* and *cro,* which are transcribed by host RNA polymerase.
- *N* is required to express the delayed early genes.
- Three of the delayed early genes are regulators.
- Lysogeny requires the delayed early genes *cII-cIII.*
- The lytic cycle requires the immediate early gene *cro* and the delayed early gene *Q.*

By Book_Crazy [IND]

One of the most intricate cascade circuits is provided by phage lambda. Actually, the cascade for lytic development itself is straightforward, with two regulators controlling the successive stages of development. But the circuit for the lytic cycle is interlocked with the circuit for establishing lysogeny, as summarized in **Figure 12.10.**

When lambda DNA enters a new host cell, the lytic and lysogenic pathways start off the same way. Both require expression of the immediate early and delayed early genes. But then they diverge: lytic development follows if the late genes are expressed; lysogeny ensues if synthesis of the repressor is established.

Lambda has only two immediate early genes, transcribed independently by host RNA polymerase:

- *N* codes for an antitermination factor whose action at the *nut* sites allows transcription to proceed into the delayed early genes (see *9.24 Antitermination requires sites that are independent of the terminators).*
- *cro* has dual functions: it prevents synthesis of the repressor (a necessary action if the lytic cycle is to proceed); and it turns off expression of the immediate early genes (which are not needed later in the lytic cycle).

The delayed early genes include two replication genes (needed for lytic infection), seven recombination genes (some involved in recombination during lytic infection, two necessary to integrate lambda DNA into the bacterial chromosome for lysogeny), and three regulators. The regulators have opposing functions:

- The *cII-cIII* pair of regulators is needed to establish the synthesis of repressor.
- The *Q* regulator is an antitermination factor that allows host RNA polymerase to transcribe the late genes.

So the delayed early genes serve two masters: some are needed for the phage to enter lysogeny, the others are concerned with controlling the order of the lytic cycle.

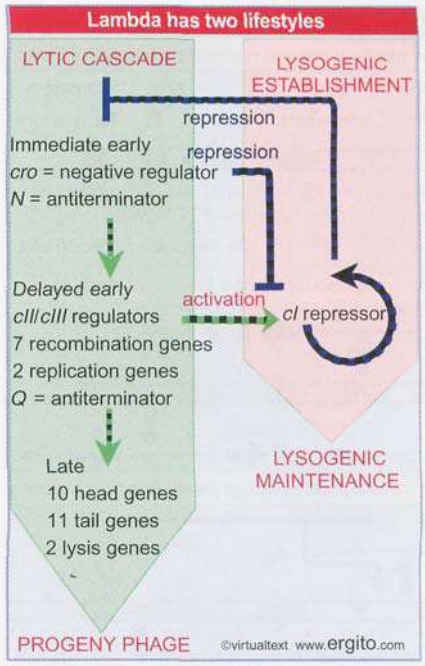


Figure 12.10 The lambda lytic cascade is interlocked with the circuitry for lysogeny.

12.7 The lytic cycle depends on antitermination

Key Concepts

- pN is an antitermination factor that allows RNA polymerase to continue transcription past the ends of the two immediate early genes.
- pQ is the product of a delayed early gene and is an antiterminator that allows RNA polymerase to transcribe the late genes.
- Because lambda DNA circularizes after infection, the late genes form a single transcription unit.

To disentangle the two pathways, let's first consider just the lytic cycle. **Figure 12.11** gives the map of lambda phage DNA. A group of genes concerned with regulation is surrounded by genes needed for recombination and replication. The genes coding for structural components of the phage are clustered. All of the genes necessary for the lytic cycle are expressed in polycistronic transcripts from three promoters.

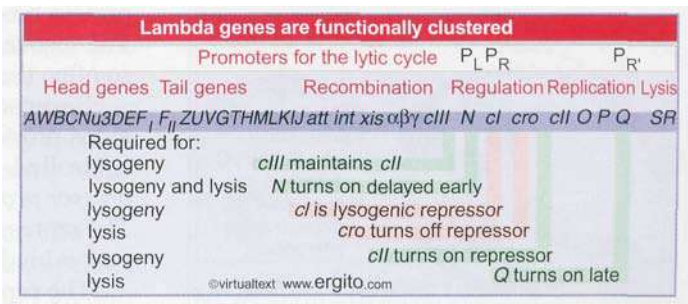


Figure 12.11 The lambda map shows clustering of related functions. The genome is 48,514 bp.

Figure 12.12 shows that the two immediate early genes, *N* and *cm*, are transcribed by host RNA polymerase. *N* is transcribed toward the left, and *cm* toward the right. Each transcript is terminated at the end of the gene. pN is the regulator that allows transcription to continue into the delayed early genes. It is an antitermination factor

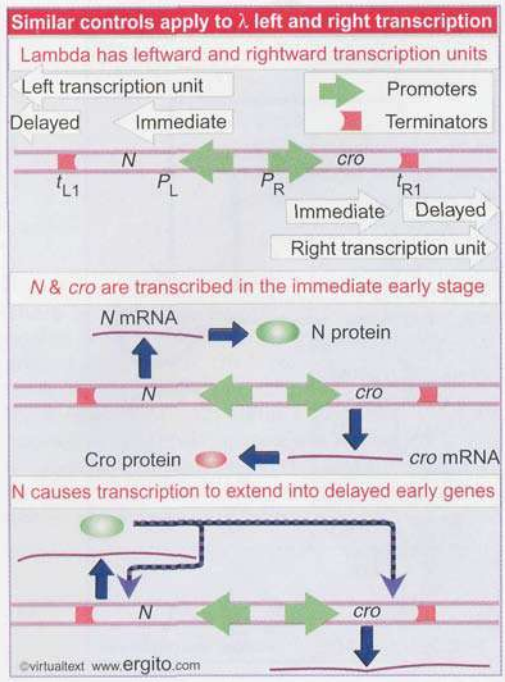


Figure 12.12 Phage lambda has two early transcription units; in the "leftward" unit, the "upper" strand is transcribed toward the left; in the "rightward" unit, the "lower" strand is transcribed toward the right.

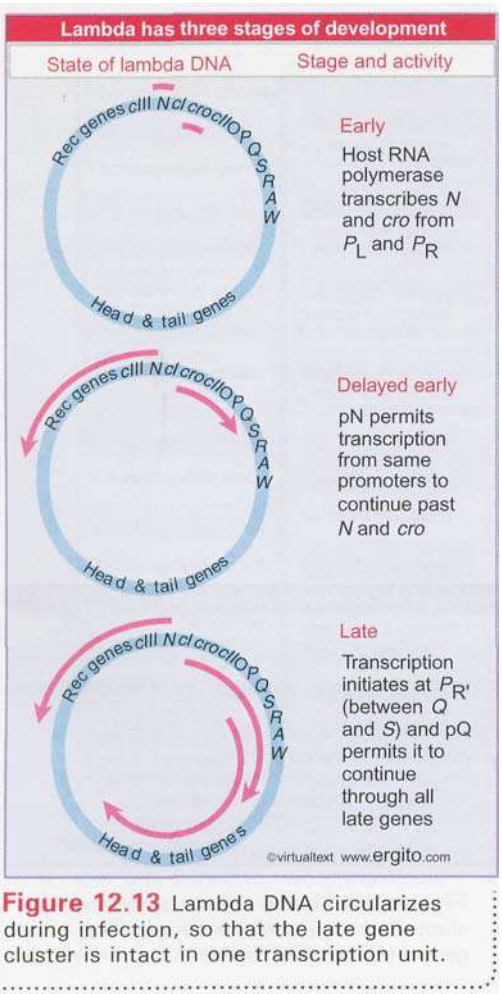


Figure 12.13 Lambda DNA circularizes during infection, so that the late gene cluster is intact in one transcription unit.

that suppresses use of the terminators t_L and t_R (see *9.25 Termination and anti-termination factors interact with RNA polymerase)*. In the presence of pN, transcription continues to the left of *N* into the recombination genes, and to the right of *cro* into the replication genes.

The map in Figure 12.11 gives the organization of the lambda DNA as it exists in the phage particle. But shortly after infection, the ends of the DNA join to form a circle. **Figure 12.13** shows the true state of lambda DNA during infection. The late genes are welded into a single group, containing the lysis genes *S-R* from the right end of the linear DNA, and the head and tail genes *A-J from* the left end.

The late genes are expressed as a single transcription unit, starting from a promoter $P_{R'}$ that lies between *Q* and *S*. The late promoter is used constitutively. However, in the absence of the product of gene *Q* (which is the last gene in the rightward delayed early unit), late transcription terminates at a site t_{R3}. The transcript resulting from this termination event is 194 bases long; it is known as 6S RNA. When pQ becomes available, it suppresses termination at t_{R3} and the 6S RNA is extended with the result that the late genes are expressed.

Late gene transcription does not seem to terminate at any specific point, but continues through all the late genes into the region beyond. A similar event happens with the leftward delayed early transcription, which continues past the recombination functions. Transcription in each direction is probably terminated before the polymerases could crash into each other.

12.8 Lysogeny is maintained by repressor protein

Key Concepts

- Mutants in the *cI* gene cannot maintain lysogeny.
- *cI* codes for a repressor protein that acts at the O_L and O_R operators to block transcription of the immediate early genes.
- Because the immediate early genes trigger a regulatory cascade, their repression prevents the lytic cycle from proceeding.

Looking at the lambda lytic cascade, we see that the entire program is set in train by initiating transcription at the two promoters P_L and P_R for the immediate early genes *N* and *cro*. Because lambda uses antitermination to proceed to the next stage of (delayed early) expression, the same two promoters continue to be used throughout the early period.

The expanded map of the regulatory region drawn in **Figure 12.14** shows that the promoters P_L and P_R lie on either side of the *cI* gene. Associated with each promoter is an operator (O_L, O_R) at which repressor protein binds to prevent RNA polymerase from initiating transcription. The sequence of each operator overlaps with the promoter that it controls; so often these are described as the P_L/O_L and P_R/O_R control regions.

Because of the sequential nature of the lytic cascade, the control regions provide a pressure point at which entry to the entire cycle can be controlled. *By denying RNA polymerase access to these promoters, a repressor protein prevents the phage genome from entering the lytic cycle.* The repressor functions in the same way as repressors of bacterial operons: it binds to specific operators.

The repressor protein is coded by the *cI* gene. Mutants in this gene cannot maintain lysogeny, but always enter the lytic cycle. Since the original isolation of the repressor protein, its characterization has shown how it both maintains the lysogenic state and provides immunity for a lysogen against superinfection by new phage lambda genomes.

By Book_Crazy [IND]

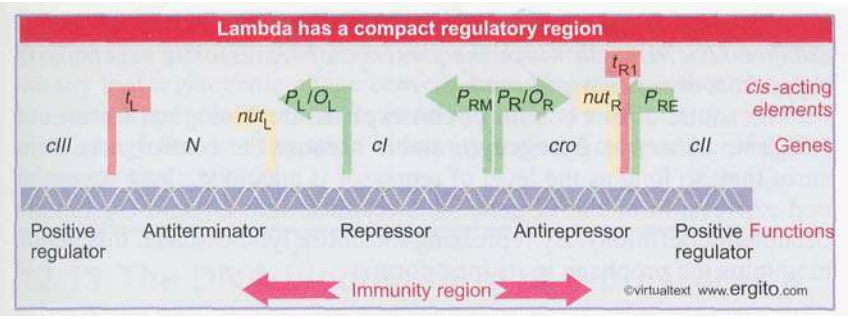


Figure 12.14 The lambda regulatory region contains a cluster of frans-acting functions and *cis*-acting elements.

When a bacterial culture is infected with a phage, the cells are lysed to generate regions that can be seen on a culture plate as small areas of clearing called plaques. With wild-type phages, the plaques are turbid or cloudy, because they contain some cells that have established lysogeny instead of being lysed. The effect of a *cI* mutation is to prevent lysogeny, so that the plaques contain only lysed cells. As a result, such an infection generates only clear plaques, and three genes (*cI, cII, cIII*) were named for their involvement in this phenotype. **Figure 12.15** compares wild-type and mutant plaques.

The *cI* gene is transcribed from a promoter P_{RM} that lies at its right end. (The subscript "RM" stands for repressor maintenance.) Transcription is terminated at the left end of the gene. The mRNA starts with the AUG initiation codon; because of the absence of the usual ribosome binding site, the mRNA is translated somewhat inefficiently, producing only a low level of repressor protein.

12.9 Repressor maintains an autogenous circuit

Key Concepts

- Repressor binding at O_L blocks transcription of gene *N* from P_L.
- Repressor binding at O_R blocks transcription of *cro* but also is required for transcription of *cI.*
- Repressor binding to the operators therefore simultaneously blocks entry to the lytic cycle and promotes its own synthesis.

The repressor binds independently to the two operators. It has a single function at O_L, but has dual functions at O_R. These are illustrated in the upper part of **Figure 12.16.**

At O_L the repressor has the same sort of effect that we have already discussed for several other systems: it prevents RNA polymerase from initiating transcription at P_L. This stops the expression of gene *N.* Since P_L is used for all leftward early gene transcription, this action prevents expression of the entire leftward early transcription unit. *So the lytic cycle is blocked before it can proceed beyond the early stages.*

At O_R, repressor binding prevents the use of P_R. So *cro* and the other rightward early genes cannot be expressed. (We see later why it is important to prevent the expression of *cro* when lysogeny is being maintained.)

But the presence of repressor at O_R also has another effect. The promoter for repressor synthesis, P_{RM}, is adjacent to the rightward operator O_R. It turns out that *RNA polymerase can initiate efficiently at* P_{RM} *only when repressor is bound at* O_R. The repressor behaves as a positive regulator protein that is necessary for transcription of the *cI* gene (see *12.15 Repressor at* O_R2 interacts with RNA polymerase at P_{RM}). *Since*

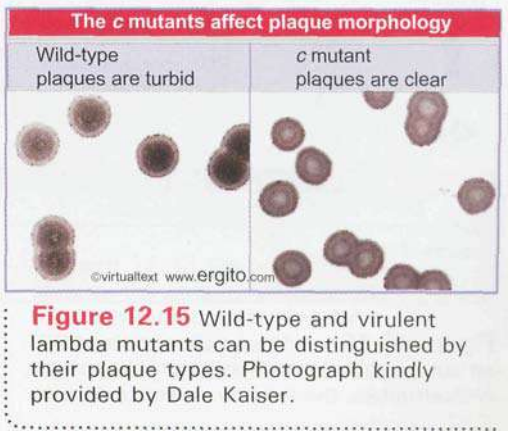


Figure 12.15 Wild-type and virulent lambda mutants can be distinguished by their plaque types. Photograph kindly provided by Dale Kaiser.

the repressor is the product of cI. *this interaction creates a positive autogenous circuit, in which the presence of repressor is necessary to support its own continued synthesis.*

The nature of this control circuit explains the biological features of lysogenic existence. Lysogeny is stable because the control circuit ensures that, so long as the level of repressor is adequate, there is continued expression of the *cI* gene. The result is that O_L and O_R remain occupied indefinitely. By repressing the entire lytic cascade, this action maintains the prophage in its inert form.

12.10 The repressor and its operators define the immunity region

Key Concepts

- Several lambdoid phages have different immunity regions.
- A lysogenic phage confers immunity to further infection by any other phage with the same immunity region.

The presence of repressor explains the phenomenon of immunity. If a second lambda phage DNA enters a lysogenic cell, repressor protein synthesized from the resident prophage genome will immediately bind to O_L and O_R in the new genome. This prevents the second phage from entering the lytic cycle.

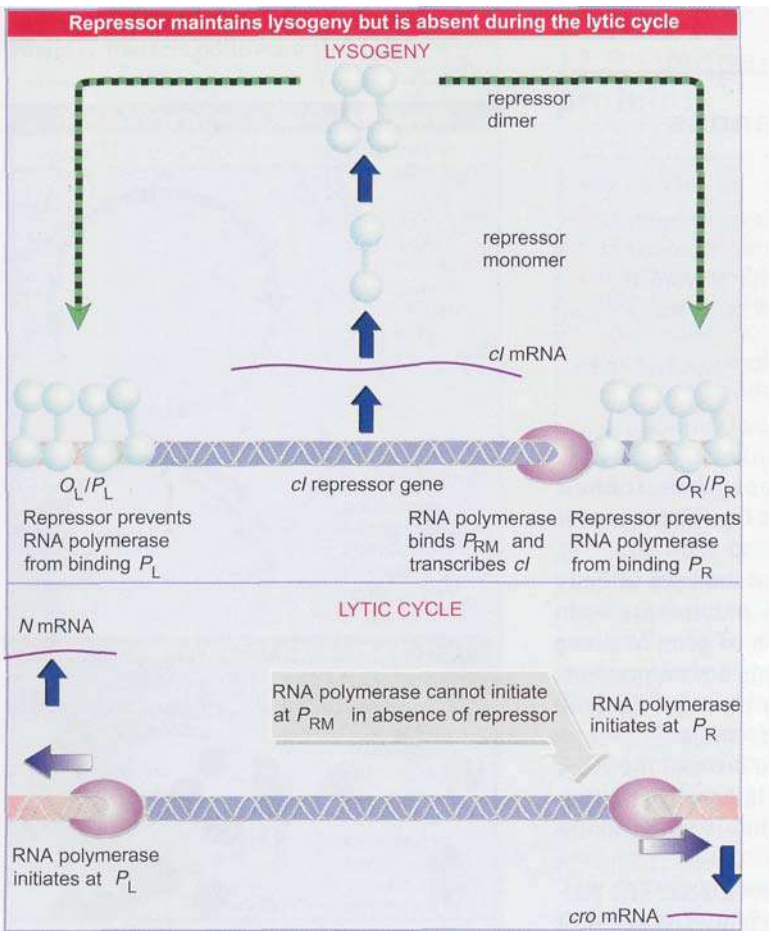


Figure 12.16 Lysogeny is maintained by an autogenous circuit (upper). If this circuit is interrupted, the lytic cycle starts (lower).

The operators were originally identified as the targets for repressor action by virulent mutations (*λvir*). These mutations prevent the repressor from binding at O_L or O_R, with the result that the phage inevitably proceeds into the lytic pathway when it infects a new host bacterium. And *λvir* mutants can grow on lysogens because the virulent mutations in O_L and O_R allow the incoming phage to ignore the resident repressor and thus to enter the lytic cycle. Virulent mutations in phages are the equivalent of operator-constitutive mutations in bacterial operons.

Prophage is induced to enter the lytic cycle when the lysogenic circuit is broken. This happens when the repressor is inactivated (see next section). The absence of repressor allows RNA polymerase to bind at P_L and P_R, starting the lytic cycle as shown in the lower part of Figure 12.16.

The autogenous nature of the repressor-maintenance circuit creates a sensitive response. Because the presence of repressor is necessary for its own synthesis, expression of the *cI* gene stops as soon as the existing repressor is destroyed. So no repressor is synthesized to replace the molecules that have been damaged. This enables the lytic cycle to start without interference from the circuit that maintains lysogeny.

The region including the left and right operators, the *cI* gene, and the *cro* gene determines the immunity of the phage. Any phage that possesses this region has the same type of immunity, because *it specifies both the repressor protein and the sites on which the repressor acts.* Accordingly, this is called the

immunity region (as marked in Figure 12.14). Each of the four lambdoid phages *ϕ80, 21, 434, and* A *has a unique immunity region. When* we say that a lysogenic phage confers immunity to any other phage of the same type, we mean more precisely that the immunity is to any other phage that has the same immunity region (irrespective of differences in other regions).

12.11 The DNA-binding form of repressor is a dimer

Key Concepts

- A repressor monomer has two distinct domains.
- The N-terminal domain contains the DNA-binding site.
- The C-terminal domain dimerizes.
- Binding to the operator requires the dimeric form so that two DNA-binding domains can contact the operator simultaneously.
- Cleavage of the repressor between the two domains reduces the affinity for the operator and induces a lytic cycle.

The repressor subunit is a polypeptide of 27 kD with the two distinct domains summarized in **Figure 12.17.**

- The N-terminal domain, residues 1-92, provides the operator-binding site.
- The C-terminal domain, residues 132-236, is responsible for dimerization.

The two domains are joined by a connector of 40 residues. When repressor is digested by a protease, each domain is released as a separate fragment.

Each domain can exercise its function independently of the other. The C-terminal fragment can form oligomers. The N-terminal fragment can bind the operators, although with a lower affinity than the intact repressor. So the information for specifically contacting DNA is contained within the N-terminal domain, but the efficiency of the process is enhanced by the attachment of the C-terminal domain.

The dimeric structure of the repressor is crucial in maintaining lysogeny. The induction of a lysogenic prophage to enter the lytic cycle is caused by cleavage of the repressor subunit in the connector region, between residues 111 and 113. (This is a counterpart to the allosteric change in conformation that results when a small-molecule inducer inactivates the repressor of a bacterial operon, a capacity that the lysogenic repressor does not have.) Induction occurs under certain adverse conditions, such as exposure of lysogenic bacteria to UV irradiation, which leads to proteolytic inactivation of the repressor.

In the intact state, dimerization of the C-terminal domains ensures that when the repressor binds to DNA its two N-terminal domains each contact DNA simultaneously. But cleavage releases the C-terminal domains from the N-terminal domains. As illustrated in **Figure 12.18** this means that the N-terminal domains can no longer dimerize; this upsets the equilibrium between monomers and dimers, so that repressor dissociates from DNA, allowing lytic infection to start. (Another relevant parameter is the loss of cooperative effects between adjacent dimers.)

The balance between lysogeny and the lytic cycle depends on the concentration of repressor. Intact repressor is present in a lysogenic cell at a concentration sufficient to ensure that the operators are occupied. But if the repressor is cleaved, this concentration is inadequate, because of the lower affinity of the separate N-terminal domain for the operator.

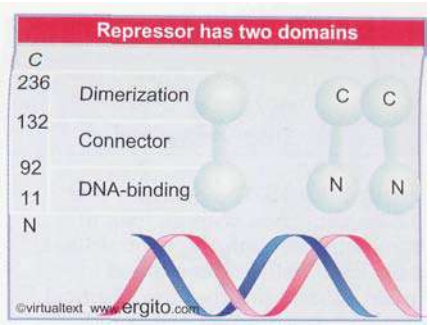


Figure 12.17 The N-terminal and C-terminal regions of repressor form separate domains. The C-terminal domains associate to form dimers; the N-terminal domains bind DNA.

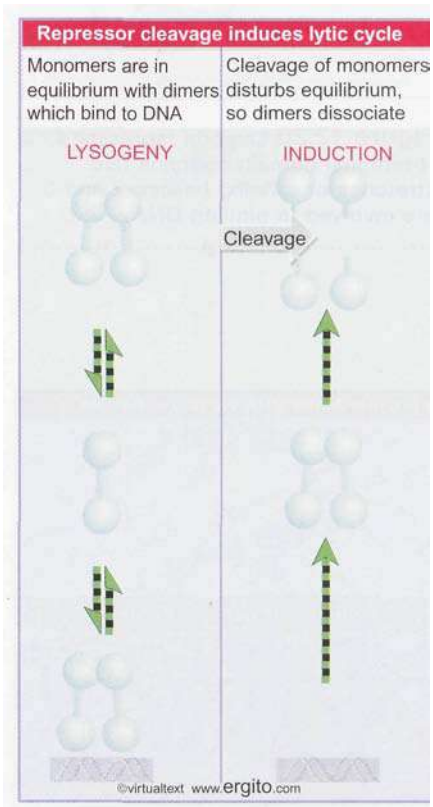


Figure 12.18 *Repressor dimers* bind to the operator. The affinity of the N-terminal domains for DNA is controlled by the dimerization of the C-terminal domains.

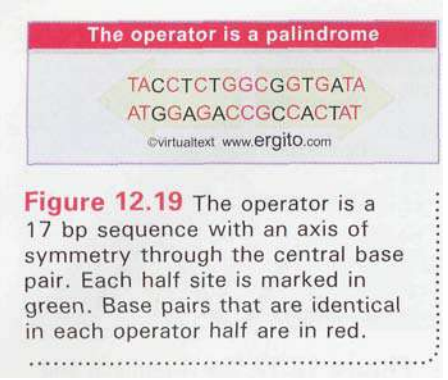


Figure 12.19 The operator is a 17 bp sequence with an axis of symmetry through the central base pair. Each half site is marked in green. Base pairs that are identical in each operator half are in red.

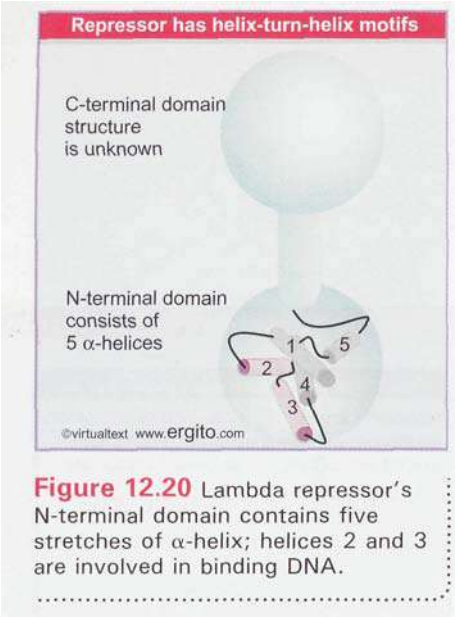


Figure 12.20 Lambda repressor's N-terminal domain contains five stretches of α-helix; helices 2 and 3 are involved in binding DNA.

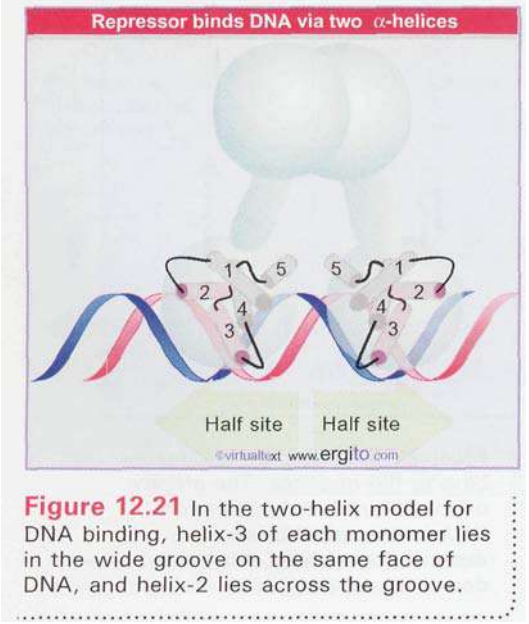


Figure 12.21 In the two-helix model for DNA binding, helix-3 of each monomer lies in the wide groove on the same face of DNA, and helix-2 lies across the groove.

Too high a concentration of repressor would make it impossible to induce the lytic cycle in this way; too low a level, of course, would make it impossible to maintain lysogeny.

12.12 Repressor uses a helix-turn-helix motif to bind DNA

Key Concepts

- Each DNA-binding region in the repressor contacts a half-site in the DNA.
- The DNA-binding site of repressor includes two short α-helical regions which fit into the successive turns of the major groove of DNA.
- A DNA-binding site is a (partially) palindromic sequence of 17 bp.

A repressor dimer is the unit that binds to DNA. It recognizes a sequence of 17 bp displaying partial symmetry about an axis through the central base pair. **Figure 12.19** shows an example of a binding site. The sequence on each side of the central base pair is sometimes called a "half-site". Each individual N-terminal region contacts a half-site. Several DNA-binding proteins that regulate bacterial transcription share a similar mode of holding DNA, in which the active domain contains two short regions of α-helix that contact DNA. (Some transcription factors in eukaryotic cells use a similar motif; see *22.14 Homeodomains bind related targets in DNA.*)

The N-terminal domain of lambda repressor contains several stretches of α-helix, arranged as illustrated diagrammatically in **Figure 12.20.** Two of the helical regions are responsible for binding DNA. The helix-turn-helix model for contact is illustrated in **Figure 12.21.** Looking at a single monomer, α-helix-3 consists of 9 amino acids, lying at an angle to the preceding region of 7 amino acids that forms α-helix-2. In the dimer, the two apposed helix-3 regions lie 34 A apart, enabling them to fit into successive major grooves of DNA. The helix-2 regions lie at an angle that would place them across the groove. The symmetrical binding of dimer to the site means that each N-terminal domain of the dimer contacts a similar set of bases in its half-site.

12.13 The recognition helix determines specificity for DNA

Key Concepts

- The amino acid sequence of the recognition helix makes contacts with particular bases in the operator sequence that it recognizes.

Related forms of the α-helical motifs employed in the helix-loop-helix of the lambda repressor are found in several DNA-binding proteins, including CRP, the *lac* repressor, and several other phage repressors. By comparing the abilities of these proteins to bind DNA, we can define the roles of each helix:

* Contacts between helix-3 and DNA rely on hydrogen bonds between the amino acid side chains and the exposed positions of the

base pairs. This helix is responsible for recognizing the specific target DNA sequence, and is therefore also known as the recognition helix.

- Contacts from helix-2 to the DNA take the form of hydrogen bonds connecting with the phosphate backbone. These interactions are necessary for binding, but do not control the specificity of target recognition. In addition to these contacts, a large part of the overall energy of interaction with DNA is provided by ionic interactions with the phosphate backbone.

What happens if we manipulate the coding sequence to construct a new protein by substituting the recognition helix in one repressor with the corresponding sequence from a closely related repressor? The specificity of the hybrid protein is that of its new recognition helix. *The amino acid sequence of this short region determines the sequence specificities of the individual proteins, and is able to act in conjunction with the rest of the polypeptide chain.*

Figure 12.22 shows the details of the binding to DNA of two proteins that bind similar DNA sequences. Both lambda repressor and Cro protein have a similar organization of the helix-turn-helix motif, although their individual specificities for DNA are not identical:

- Each protein uses similar interactions between hydrophobic amino acids to maintain the relationship between helix-2 and helix-3: repressor has an Ala-Val connection, while Cro has an Ala-Ile association.
- Amino acids in helix-3 of the repressor make contacts with specific bases in the operator. Three amino acids in repressor recognize three bases in DNA; the amino acids at these positions and also at additional positions in Cro recognize five (or possibly six) bases in DNA.

Two of the amino acids involved in specific recognition are identical in repressor and Cro (Gln and Ser at the N-terminal end of the helix), while the other contacts are different (Ala in repressor versus Lys and the additional Asn in Cro). Also, a Thr in helix-2 of Cro directly contacts DNA.

The interactions shown in the figure represent binding to the DNA sequence that each protein recognizes most tightly. The sequences shown at the bottom of the figure with the contact points in color differ at 3 of the 9 base pairs. The use of overlapping, but not identical contacts between amino acids and bases shows how related recognition helices confer recognition of related DNA sequences. This enables repressor and Cro to recognize the same set of sequences, but with different relative affinities for particular members of the group.

The bases contacted by helix-3 of repressor or Cro lie on one face of DNA, as can be seen from the positions indicated on the helical diagram in Figure 12.22. However, repressor makes an additional contact with the other face of DNA. Removing the last six N-terminal amino acids (which protrude from helix-1) eliminates some of the contacts. This observation provides the basis for the idea that the bulk of the N-terminal domain contacts one face of DNA, while the last six N-terminal amino acids form an "arm" extending around the back. **Figure 12.23** shows the view from the back. Lysine residues in the arm make contacts with G residues in the major groove, and also with the phosphate backbone. The interaction between the arm and DNA contributes heavily to DNA binding; the affinity of the armless repressor for DNA is reduced by ~1000-fold.

Bases that are not contacted directly by repressor protein may have an important effect on binding. The related phage *434* repressor binds DNA via a helix-turn-helix motif, and the crystal structure shows that helix-3 is positioned at each half-site so that it contacts the 5 outermost

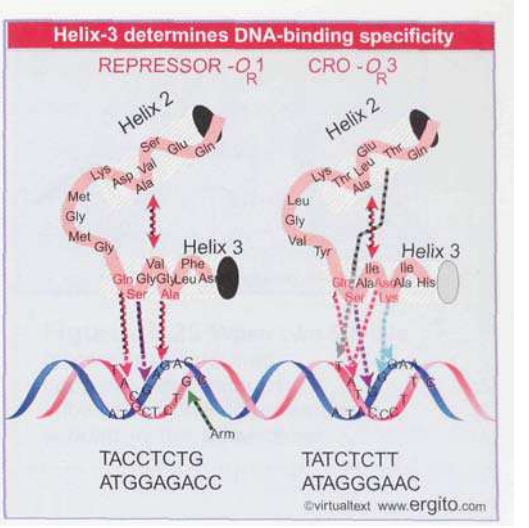


Figure 12.22 Two proteins that use the two-helix arrangement to contact DNA recognize lambda operators with affinities determined by the amino acid sequence of helix-3.

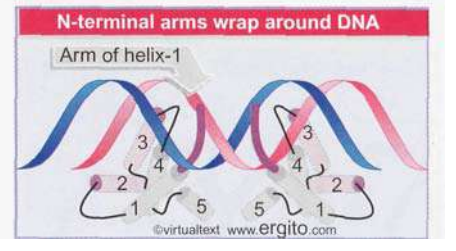


Figure 12.23 A view from the back shows that the bulk of the repressor contacts one face of DNA, but its N-terminal arms reach around to the other face.

base pairs but not the inner 2. However, operators with A·T base pairs at the inner positions bind *434* repressor more strongly than operators with G·C base pairs. The reason is that *434* repressor binding slightly twists DNA at the center of the operator, widening the angle between the two half-sites of DNA by ~3°. This is probably needed to allow each monomer of the repressor dimer to make optimal contacts with DNA. A-T base pairs allow this twist more readily than G·C pairs, thus affecting the affinity of the operator for repressor.

12.14 Repressor dimers bind cooperatively to the operator

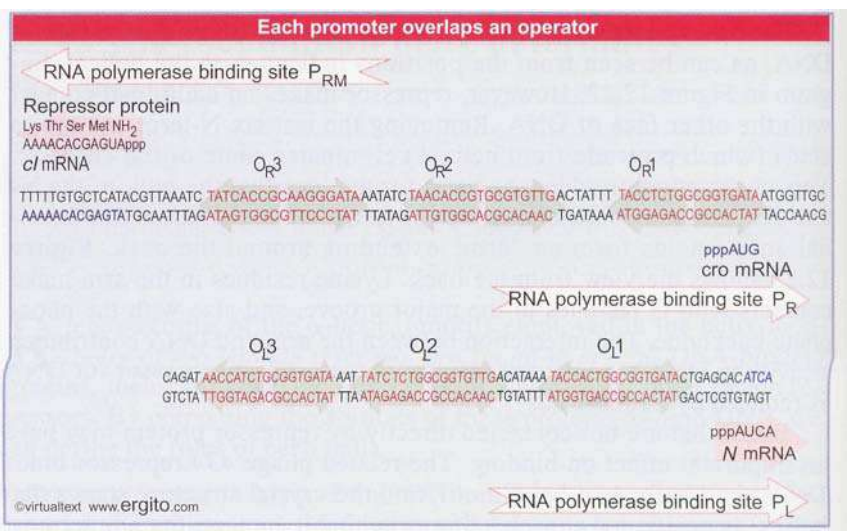
Key Concepts

- Repressor binding to one operator increases the affinity for binding a second repressor dimer to the adjacent operator.
- The affinity is 10× greater for O_L1 and O_R1 than other operators, so they are bound first.
- Cooperativity allows repressor to bind the 01/02 sites at lower concentrations.

Each operator contains three repressor-binding sites. As can be seen from **Figure 12.24,** no two of the six individual repressor-binding sites are identical, but they all conform with a consensus sequence. The binding sites within each operator are separated by spacers of 3-7 bp that are rich in A-T base pairs. The sites at each operator are numbered so that O_R consists of the series O_R1-O_R2-O_R3, while O_L consists of the series O_L1-O_L2-O_L3. In each case, site 1 lies closest to the startpoint for transcription in the promoter, and sites 2 and 3 lie farther upstream.

Faced with the triplication of binding sites at each operator, how does repressor decide where to start binding? At each operator, site 1 has a greater affinity (roughly tenfold) than the other sites for the repressor. So the repressor always binds first to O_L1 and O_R1.

Lambda repressor binds to subsequent sites within each operator in a cooperative manner. The presence of a dimer at site 1 greatly increases the affinity with which a second dimer can bind to site 2. When both sites 1 and 2 are occupied, this interaction does *not* extend farther, to site 3. At the concentrations of repressor usually found in a lysogen, both sites 1 and 2 are filled at each operator, but site 3 is not occupied.



By Book_Crazy [IND]

Figure 12.24 *Each operator contains* three repressor-binding sites, and overlaps with the promoter at which RNA polymerase binds. The orientation of O_L has been reversed from usual to facilitate comparison with O_R.

If site 1 is inactive (because of mutation), then repressor binds cooperatively to sites 2 and 3. That is, binding at site 2 assists another dimer to bind at site 3. This interaction occurs directly between repressor dimers and not via conformational change in DNA. The C-terminal domain is responsible for the cooperative interaction between dimers as well as for the dimer formation between subunits. **Figure 12.25** shows that it involves both subunits of each dimer, that is, each subunit contacts its counterpart in the other dimer, forming a tetrameric structure.

A result of cooperative binding is to increase the effective affinity of repressor for the operator at physiological concentrations. This enables a lower concentration of repressor to achieve occupancy of the operator. This is an important consideration in a system in which release of repression has irreversible consequences. In an operon coding for metabolic enzymes, after all, failure of repression will merely allow unnecessary synthesis of enzymes. But failure to repress lambda prophage will lead to induction of phage and lysis of the cell.

From the sequences shown in Figure 12.24, we see that O_L1 and O_R1 lie more or less in the center of the RNA polymerase binding sites of P_L and P_R, respectively. Occupancy of O_L1-O_L2 and O_R1-O_R2 thus physically blocks access of RNA polymerase to the corresponding promoters.

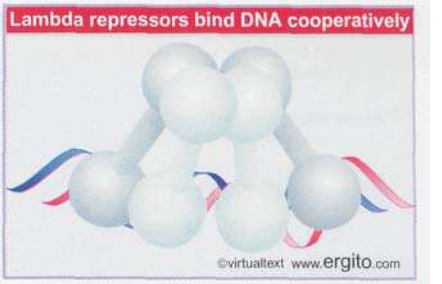


Figure 12.25 When two lambda repressor dimers bind cooperatively, each of the subunits of one dimer contacts a subunit in the other dimer.

12.15 Repressor at O_R2 interacts with RNA polymerase at P_{RM}

Key Concepts

- The DNA-binding region of repressor at O_R2 contacts RNA polymerase and stabilizes its binding to P_{RM}.
- This is the basis for the autogenous control of repressor maintenance.

A different relationship is shown between O_R and the promoter P_{RM} for transcription of *cI*. The RNA polymerase binding site is adjacent to O_R2. This explains how repressor autogenously regulates its own synthesis. When two dimers are bound at O_R1-O_R2, the dimer at O_R2 interacts with RNA polymerase (see Figure 12.16 in *12.9 Repressor maintains an autogenous circuit*). This effect resides in the amino terminal domain of repressor.

Mutations that abolish positive control map in the *cl* gene. One interesting class of mutants remain able to bind the operator to repress transcription, but cannot stimulate RNA polymerase to transcribe from P_{RM}. They map within a small group of amino acids, located on the outside of helix-2 or in the turn between helix-2 and helix-3. The mutations reduce the negative charge of the region; conversely, mutations that increase the negative charge enhance the activation of RNA polymerase. This suggests that the group of amino acids constitutes an "acidic patch" that functions by an electrostatic interaction with a basic region on RNA polymerase.

The location of these "positive control mutations" in the repressor is indicated on **Figure 12.26.** They lie at a site on repressor that is close to a phosphate group on DNA that is also close to RNA polymerase. So the group of amino acids on repressor that is involved in positive control is in a position to contact the polymerase. The interaction between repressor and polymerase is needed for the polymerase to make the transition from a closed complex to an open complex (see also Figure 12.29). The

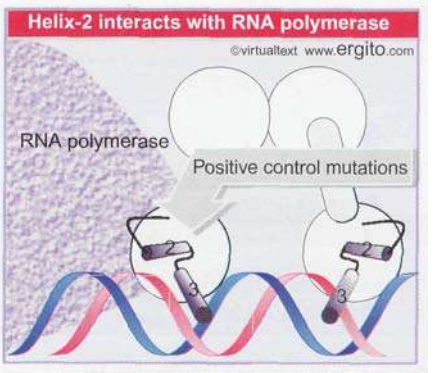


Figure 12.26 Positive control mutations identify a small region at helix-2 that interacts directly with RNA polymerase.

important principle is that *protein-protein interactions can release energy that is used to help to initiate transcription.*

What happens if a repressor dimer binds to O_R3? This site overlaps with the RNA polymerase binding site at P_{RM}. So if the repressor concentration becomes great enough to cause occupancy of O_R3, the transcription of c/is prevented. This leads in due course to a reduction in repressor concentration; O_R3 then becomes empty, and the autogenous loop can start up again because O_R2 remains occupied.

This mechanism could prevent the concentration of repressor from becoming too great, although it would require repressor concentration in lysogens to reach unusually high levels. In the formal sense, the repressor is an autogenous regulator of its own expression that functions positively at low concentrations and negatively at high concentrations.

Virulent mutations occur in sites 1 and 2 of both O_L and O_R. The mutations vary in their degree of virulence, according to the extent to which they reduce the affinity of the binding site for repressor, and also depending on the relationship of the affected site to the promoter. Consistent with the conclusion that O_R3 and O_L3 usually are not occupied, virulent mutations are not found in these sites.

12.16 The *cII* and *cIII* genes are needed to establish lysogeny

Key Concepts

- The delayed early gene products cII and cIII are necessary for RNA polymerase to initiate transcription at the promoter P_{RE}.
- cII acts directly at the promoter and cIII protects cII from degradation.
- Transcription from P_{RE} leads to synthesis of repressor and also blocks the transcription of *cro*.

The control circuit for maintaining lysogeny presents a paradox. *The presence of repressor protein is necessary for its own synthesis.* This explains how the lysogenic condition is perpetuated. But how is the synthesis of repressor established in the first place?

When a lambda DNA enters a new host cell, RNA polymerase cannot transcribe *cI*, because there is no repressor present to aid its binding at P_{RM}. But this same absence of repressor means that P_R and P_L are available. So the first event when lambda DNA infects a bacterium is for genes *N* and *cm* to be transcribed. Then pN allows transcription to be extended farther. This allows *cIII* (and other genes) to be transcribed on the left, while *cII* (and other genes) are transcribed on the right (see Figure 12.14).

The *cII* and *cIII* genes share with *cl* the property that mutations in them cause clear plaques. But there is a difference. The *cl* mutants can neither establish nor maintain lysogeny. The *cII* or *clll* mutants have some difficulty in establishing lysogeny, but once established, they are able to maintain it by the *cl* autogenous circuit.

This implicates the *cII* and *clll* genes as positive regulators whose products are needed for an alternative system for repressor synthesis. The system is needed only to *initiate* the expression of *cI* in order to circumvent the inability of the autogenous circuit to engage in *de novo* synthesis. They are not needed for continued expression.

The cII protein acts directly on gene expression. Between the *cro* and *cII* genes is another promoter, called P_{RE}. (The subscript "RE" stands for repressor establishment.) This promoter can be recognized by RNA polymerase only in the presence of cII, whose action is illustrated in **Figure 12.27.**

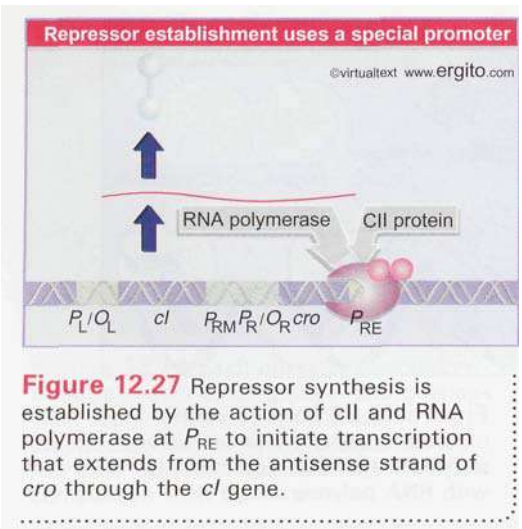


Figure 12.27 Repressor synthesis is established by the action of cII and RNA polymerase at P_{RE} to initiate transcription that extends from the antisense strand of *cro* through the *cl* gene.

By Book_Crazy [IND]

The cII protein is extremely unstable *in vivo*, because it is degraded as the result of the activity of a host protein called HflA. The role of cIII is to protect cII against this degradation.

Transcription from P_{RE} promotes lysogeny in two ways. Its direct effect is that *cI* is translated into repressor protein. An indirect effect is that transcription proceeds through the *cro* gene in the "wrong" direction. So the 5' part of the RNA corresponds to an antisense transcript of *cro;* in fact, it hybridizes to authentic *cro* mRNA, inhibiting its translation. This is important because *cro* expression is needed to enter the lytic cycle (see *12.19 The cro repressor is needed for lytic infection).*

The *cI* coding region on the P_{RE} transcript is very efficiently translated, in contrast with the weak translation of the P_{RM} transcript. In fact, repressor is synthesized ~7-8 times more effectively via expression from P_{RE} than from P_{RM}. This reflects the fact that the P_{RE} transcript has an efficient ribosome-binding site, whereas the P_{RM} transcript has no ribosome-binding site and actually starts with the AUG initiation codon.

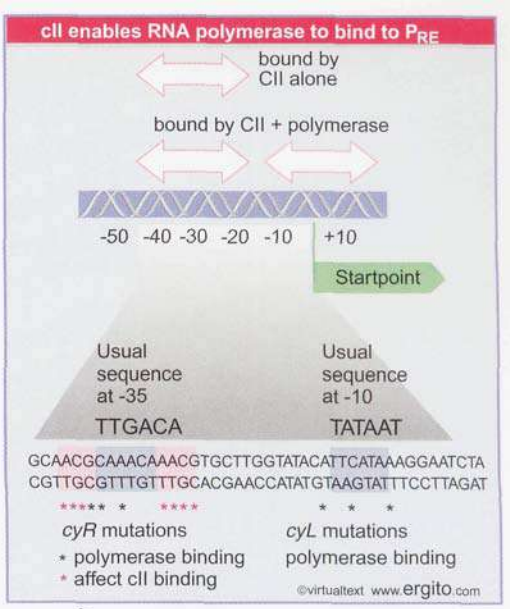


Figure 12.28 RNA polymerase binds to P_{RE} only in the presence of cII, which contacts the region around –35.

12.17 A poor promoter requires cII protein

Key Concepts

- P_{RE} has atypical sequences at -10 and -35.
- RNA polymerase binds the promoter only in the presence of cII.
- cII binds to sequences close to the -35 region.

The P_{RE} promoter has a poor fit with the consensus at –10 and lacks a consensus sequence at -35. This deficiency explains its dependence on *cII*. The promoter cannot be transcribed by RNA polymerase alone *in vitro,* but can be transcribed when cII is added. The regulator binds to a region extending from about -25 to -45. When RNA polymerase is added, an additional region is protected, extending from -12 to +13. As summarized in **Figure 12.28,** the two proteins bind to overlapping sites.

The importance of the -35 and –10 regions for promoter function, in spite of their lack of resemblance with the consensus, is indicated by the existence of *cy* mutations. These have effects similar to those of *cII* and *cIII* mutations in preventing the establishment of lysogeny; but they are ris-acting instead of *trans-acting*. They fall into two groups, *cyL* and *cyR*, localized at the consensus operator positions of –10 and -35.

The *cyL* mutations are located around –10, and probably prevent RNA polymerase from recognizing the promoter.

The *cyR* mutations are located around -35, and fall into two types, affecting either RNA polymerase or cII binding. Mutations in the center of the region do not affect cI 1 binding; presumably they prevent RNA polymerase binding. On either side of this region, mutations in short tetrameric repeats, TTGC, prevent cII from binding. Each base in the tetramer is 10 bp (one helical turn) separated from its homologue in the other tetramer, so that when cII recognizes the two tetramers, it lies on one face of the double helix.

Positive control of a promoter implies that an accessory protein has increased the efficiency with which RNA polymerase initiates transcription. **Figure 12.29** reports that either or both stages of the interaction between promoter and polymerase can be the target for regulation. Initial binding to form a closed complex or its conversion into an open complex can be enhanced.

Positive regulation influences initiation

Promoter	Regulator	Polymerase Binding (equilibrium constant, K_B)	Closed-Open Conversion (rate constant, k_2)
P_{RM}	repressor	no effect	11X
P_{RE}	CII	100X	100X

©virtualtext www.ergito.com

Figure 12.29 Positive regulation can influence RNA polymerase at either stage of initiating transcription.

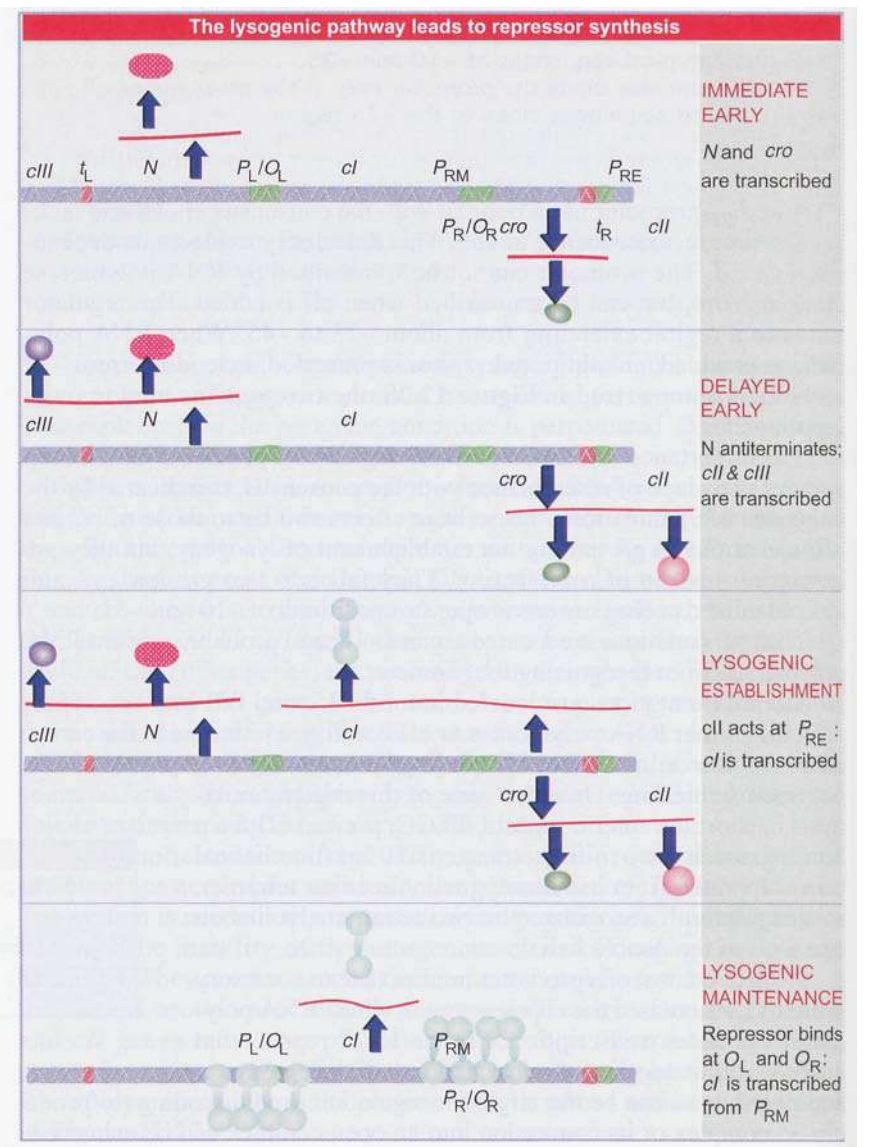
12.18 Lysogeny requires several events

Key Concepts

- cII/cIII cause repressor synthesis to be established and also trigger inhibition of late gene transcription.
- Establishment of repressor turns off immediate and delayed early gene expression.
- Repressor turns on the maintenance circuit for its own synthesis.
- Lambda DNA is integrated into the bacterial genome at the final stage in establishing lysogeny.

Now we can see how lysogeny is established during an infection. **Figure 12.30** recapitulates the early stages and shows what happens as the result of expression of *cIII* and *cII*. The presence of cII allows P_{RE} to be used for transcription extending through *cI*. Repressor protein is synthesized in high amounts from this transcript. Immediately it binds to O_L and O_R.

By directly inhibiting any further transcription from P_L and P_R, repressor binding turns off the expression of all phage genes. This halts



By Book_Crazy [IND]

Figure 12.30 A cascade is needed to establish lysogeny, but then this circuit is switched off and replaced by the autogenous repressor-maintenance circuit.

the synthesis of cII and cIII, which are unstable; they decay rapidly, with the result that P_{RE} can no longer be used. So the synthesis of repressor via the establishment circuit is brought to a halt.

But repressor now is present at O_R. It switches on the maintenance circuit for expression from P_{RM}. Repressor continues to be synthesized, although at the lower level typical of P_{RM} function. So the establishment circuit starts off repressor synthesis at a high level; then repressor turns off all other functions, while at the same time turning on the maintenance circuit, which functions at the low level adequate to sustain lysogeny.

We shall not now deal in detail with the other functions needed to establish lysogeny, but we can just briefly remark that the infecting lambda DNA must be inserted into the bacterial genome (see *15.16 Specialized recombination involves specific sites).* The insertion requires the product of gene *int,* which is expressed from its own promoter P_I, at which cII also is necessary. The sequence of P_I shows homology with P_{RE} in the cII binding site (although not in the -10 region). The functions necessary for establishing the lysogenic control circuit are therefore under the same control as the function needed to integrate the phage DNA into the bacterial genome. So the establishment of lysogeny is under a control that ensures all the necessary events occur with the same timing.

Emphasizing the tricky quality of lambda's intricate cascade, we now know that cII promotes lysogeny in another, indirect manner. It sponsors transcription from a promoter called $P_{anti\text{-}Q}$, which is located within the *Q* gene. This transcript is an antisense version of the *Q* region, and it hybridizes with *Q* mRNA to prevent translation of Q protein, whose synthesis is essential for lytic development. So the same mechanisms that directly promote lysogeny by causing transcription of the *cI* repressor gene also indirectly help lysogeny by inhibiting the expression of *cro* (see above) and *Q,* the regulator genes needed for the antagonistic lytic pathway.

12.19 The cro repressor is needed for lytic infection

Key Concepts

- Cro binds to the same operators as repressor but with different affinities.
- When Cro binds to O_R3, it prevents RNA polymerase from binding to P_{RM}, and blocks maintenance of repressor.
- When Cro binds to other operators at O_R or O_L, it prevents RNA polymerase from expressing immediate early genes, which (indirectly) blocks repressor establishment.

Lambda has the alternatives of entering lysogeny or starting a lytic infection. Lysogeny is initiated by establishing an autogenous maintenance circuit that inhibits the entire lytic cascade through applying pressure at two points. The program for establishing lysogeny proceeds through some of the same events that are required for the lytic cascade (expression of delayed early genes via expression of *N* is needed). We now face a problem. How does the phage enter the lytic cycle?

The key influence on the lytic cycle is the role of gene *cro,* which codes for another repressor. *Cro is responsible for preventing the synthesis of the repressor protein;* this action shuts off the possibility of establishing lysogeny. *cro* mutants usually establish lysogeny rather than entering the lytic pathway, because they lack the ability to switch events away from the expression of repressor.

Cro forms a small dimer (the subunit is 9 kD) that acts within the immunity region. It has two effects:

- It prevents the synthesis of repressor via the maintenance circuit; that is, it prevents transcription via P_{RM}.
- It also inhibits the expression of early genes from both P_L and P_R.

This means that, when a phage enters the lytic pathway, Cro has responsibility both for preventing the synthesis of repressor and (subsequently) for turning down the expression of the early genes.

Cro achieves its function by binding to the same operators as (*cI*) repressor protein. Cro includes a region with the same general structure as the repressor; a helix-2 is offset at an angle from recognition helix-3. (The remainder of the structure is different, demonstrating that the helix-turn-helix motif can operate within various contexts.) Like repressor, Cro binds symmetrically at the operators.

The sequences of Cro and repressor in the helix-turn-helix region are related, explaining their ability to contact the same DNA sequences (see Figure 12.22). Cro makes similar contacts to those made by repressor, but binds to only one face of DNA; it lacks the N-terminal arms by which repressor reaches around to the other side.

How can two proteins have the same sites of action, yet have such opposite effects? The answer lies in the different affinities that each protein has for the individual binding sites within the operators. Let us

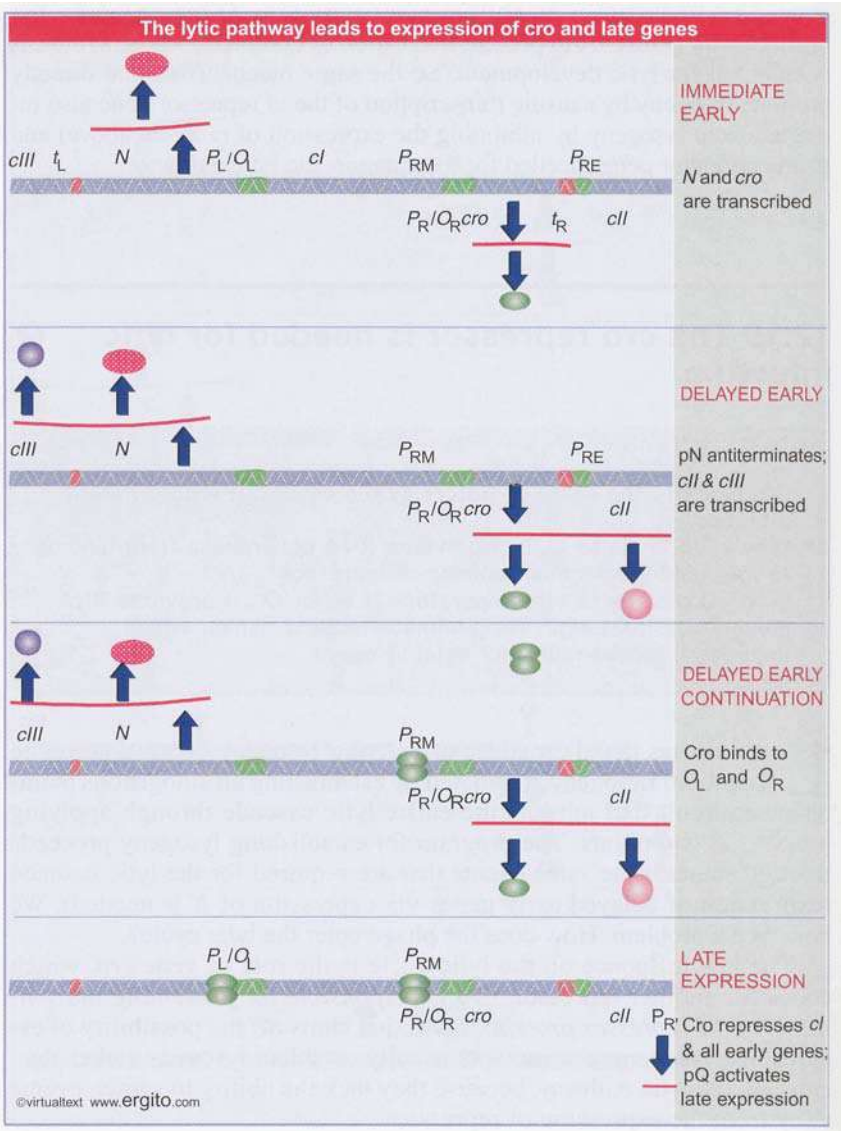


Figure 12.31 The lytic cascade requires Cro protein, which directly prevents repressor maintenance via P_{RM}, as well as turning off delayed early gene expression, indirectly preventing repressor establishment.

By Book_Crazy [IND]

just consider O_R, where more is known, and where Cro exerts both its effects. The series of events is illustrated in **Figure 12.31.** (Note that the first two stages are identical to those of the lysogenic circuit shown in Figure 12.30.)

The affinity of Cro for O_R3 is greater than its affinity for O_R2 or O_R1. So it binds first to O_R3. This inhibits RNA polymerase from binding to P_{RM}. So Cro's first action is to prevent the maintenance circuit for lysogeny from coming into play.

Then Cro binds to O_R2 or O_R1. Its affinity for these sites is similar, and there is no cooperative effect. Its presence at either site is sufficient to prevent RNA polymerase from using P_R. This in turn stops the production of the early functions (including Cro itself). Because cII is unstable, any use of P_{RE} is brought to a halt. So the two actions of Cro together block *all* production of repressor.

So far as the lytic cycle is concerned, Cro turns down (although it does not completely eliminate) the expression of the early genes. Its incomplete effect is explained by its affinity for O_R1 and O_R2, which is about eight times lower than that of repressor. This effect of Cro does not occur until the early genes have become more or less superfluous, because pQ is present; by this time, the phage has started late gene expression, and is concentrating on the production of progeny phage particles.

12.20 What determines the balance between lysogeny and the lytic cycle?

Key Concepts

- The delayed early stage when both Cro and repressor are being expressed is common to lysogeny and the lytic cycle.
- The critical event is whether cII causes sufficient synthesis of repressor to overcome the action of Cro.

The programs for the lysogenic and lytic pathways are so intimately related that it is impossible to predict the fate of an individual phage genome when it enters a new host bacterium. Will the antagonism between repressor and Cro be resolved by establishing the autogenous maintenance circuit shown in Figure 12.30, or by turning off repressor synthesis and entering the late stage of development shown in Figure 12.31?

The same pathway is followed in both cases right up to the brink of decision. Both involve the expression of the immediate early genes and extension into the delayed early genes. The difference between them comes down to the question of whether repressor or Cro will obtain occupancy of the two operators.

The early phase during which the decision is taken is limited in duration in either case. No matter which pathway the phage follows, expression of all early genes will be prevented as P_L and P_R *are* repressed; and, as a consequence of the disappearance of cII and cIII, production of repressor via P_{RE} will cease.

The critical question comes down to whether the cessation of transcription from P_{RE} is followed by activation of P_{RM} and the establishment of lysogeny, or whether P_{RM} fails to become active and the pQ regulator commits the phage to lytic development. **Figure 12.32** shows the critical stage, at which both repressor and Cro are being synthesized.

The initial event in establishing lysogeny is the binding of repressor at O_L1 and O_R1. Binding at the first sites is rapidly succeeded by

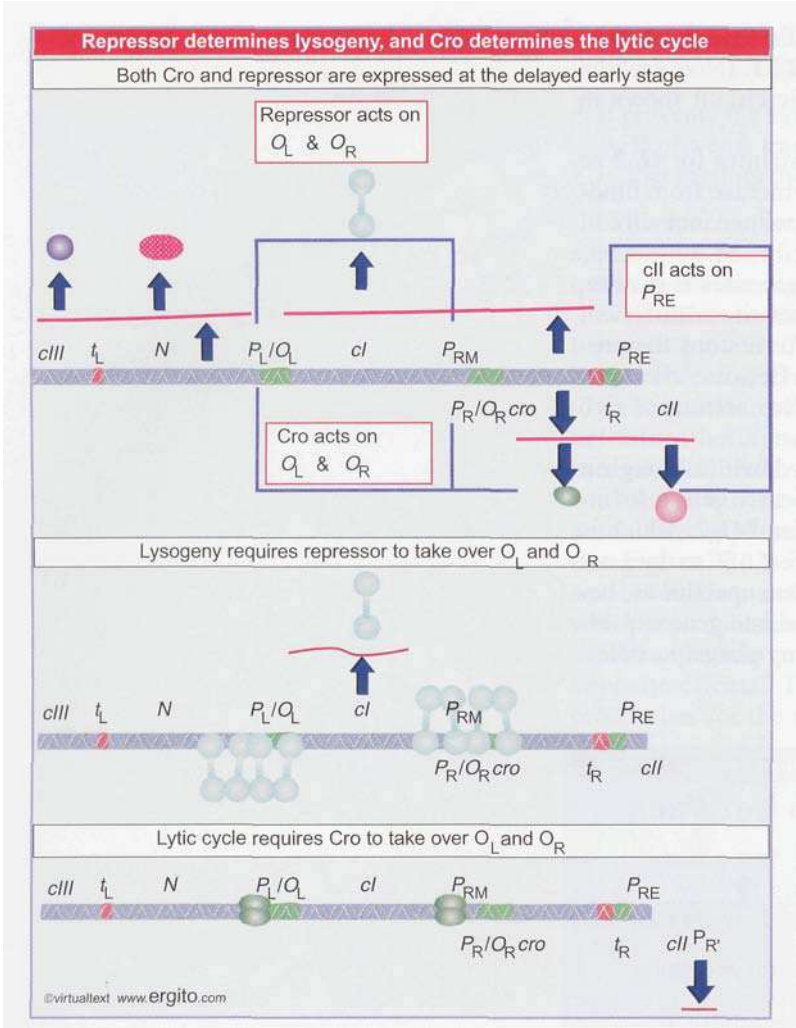


Figure 12.32 The critical stage in deciding between lysogeny and lysis is when delayed early genes are being expressed. If cII causes sufficient synthesis of repressor, lysogeny will result because repressor occupies the operators. Otherwise Cro occupies the operators, resulting in a lytic cycle.

cooperative binding of further repressor dimers at O_L2 and O_R2. This shuts off the synthesis of Cro and starts up the synthesis of repressor via P_{RM}.

The initial event in entering the lytic cycle is the binding of Cro at O_R3. This stops the lysogenic-maintenance circuit from starting up at P_{RM}. Then Cro must bind to O_R1 or O_R2, and to O_L1 or O_L2, to turn down early gene expression. By halting production of cII and cIII, this action leads to the cessation of repressor synthesis via P_{RE}. The shutoff of repressor establishment occurs when the unstable cII and cIII proteins decay.

The critical influence over the switch between lysogeny and lysis is cII. If cII is active, synthesis of repressor via the establishment promoter is effective; and, as a result, repressor gains occupancy of the operators. If cII is not active, repressor establishment fails, and Cro binds to the operators.

The level of cII protein under any particular set of circumstances determines the outcome of an infection. Mutations that increase the stability of cII increase the frequency of lysogenization. Such mutations occur in *cII* itself or in other genes. The cause of cII's instability is its susceptibility to degradation by host proteases. Its level in the cell is influenced by *cIII* as well as by host functions.

The effect of the lambda protein cIII is secondary: it helps to protect cII against degradation. Although the presence of cIII does not guarantee the survival of cII, in the absence of cIII, cII is virtually always inactivated.

Host gene products act on this pathway. Mutations in the host genes *hflA* and *hflB* increase lysogeny—*hfl* stands for *h*igh *f*requency *l*ysogenization. The mutations stabilize cII because they inactivate host protease(s) that degrade it.

The influence of the host cell on the level of cII provides a route for the bacterium to interfere with the decision-taking process. For example, host proteases that degrade cII are activated by growth on rich medium, so lambda tends to lyse cells that are growing well, but is more likely to enter lysogeny on cells that are starving (and which lack components necessary for efficient lytic growth).

12.21 Summary

Phages have a lytic life cycle, in which infection of a host bacterium is followed by production of a large number of phage particles, lysis of the cell, and release of the viruses. Some phages also can exist in lysogenic form, in which the phage genome is integrated into the bacterial chromosome and is inherited in this inert, latent form like any other bacterial gene.

Lytic infection falls typically into three phases. In the first phase a small number of phage genes are transcribed by the host RNA polymerase. One or more of these genes is a regulator that controls expression of the group of genes expressed in the second phase. The pattern is repeated in the second phase, when one or more genes is a regulator needed for expression of the genes of the third phase. Genes of the first two phases code for enzymes needed to reproduce phage DNA; genes of the final phase code for structural

components of the phage particle. It is common for the very early genes to be turned off during the later phases.

In phage lambda, the genes are organized into groups whose expression is controlled by individual regulatory events. The immediate early gene *N* codes for an antiterminator that allows transcription of the leftward and rightward groups of delayed early genes from the early promoters P_R and P_L. The delayed early gene *Q* has a similar antitermination function that allows transcription of all late genes from the promoter $P_{R'}$. The lytic cycle is repressed, and the lysogenic state maintained, by expression of the *cI* gene, whose product is a repressor protein that acts at the operators O_R and O_L to prevent use of the promoters P_R and P_L, respectively. A lysogenic phage genome expresses only the *cI* gene, from its promoter P_{RM}. Transcription from this promoter involves positive autogenous regulation, in which repressor bound at O_R activates RNA polymerase at P_{RM}.

Each operator consists of three binding sites for repressor. Each site is palindromic, consisting of symmetrical half-sites. Repressor functions as a dimer. Each half binding site is contacted by a repressor monomer. The N-terminal domain of repressor contains a helix-turn-helix motif that contacts DNA. Helix-3 is the recognition helix, responsible for making specific contacts with base pairs in the operator. Helix-2 is involved in positioning helix-3; it is also involved in contacting RNA polymerase at P_{RM}. The C-terminal domain is required for dimerization. Induction is caused by cleavage between the N- and C-terminal domains, which prevents the DNA-binding regions from functioning in dimeric form, thereby reducing their affinity for DNA and making it impossible to maintain lysogeny. Repressor-operator binding is cooperative, so that once one dimer has bound to the first site, a second dimer binds more readily to the adjacent site.

The helix-turn-helix motif is used by other DNA-binding proteins, including lambda Cro, which binds to the same operators, but has a different affinity for the individual operator sites, determined by the sequence of helix-3. Cro binds individually to operator sites, starting with O_R3, in a noncooperative manner. It is needed for progression through the lytic cycle. Its binding to O_R3 first prevents synthesis of repressor from P_{RM}; then its binding to O_R2 and O_R1 prevents continued expression of early genes, an effect also seen in its binding to O_L1 and O_L2.

Establishment of repressor synthesis requires use of the promoter P_{RE}, which is activated by the product of the *cII* gene. The product of *cIII* is required to stabilize the *cII* product against degradation. By turning off c//and *cIII* expression, Cro acts to prevent lysogeny. By turning off all transcription except that of its own gene, repressor acts to prevent the lytic cycle. The choice between lysis and lysogeny depends on whether repressor or Cro gains occupancy of the operators in a particular infection. The stability of cII protein in the infected cell is a primary determinant of the outcome.

References

12.4 Two types of regulatory event control the lytic cascade

rev Greenblatt, J., Nodwell, J. R., and Mason, S. W. (1993). Transcriptional antitermination. Nature 364, 401-406.

12.6 Lambda immediate early and delayed early genes are needed for both lysogeny and the lytic cycle

rev Ptashne, M. (1992). Genetic Switch: Phage Lambda and Higher Organisms (Cell Press and Blackwell Scientific, Cambridge).

12.8 Lysogeny is maintained by repressor protein

exp Ptashne, M. (2002). Isolation of Repressor (www.ergito.com/lookup.jsp?expt = ptashne)

ref Pirrotta, V., Chadwick, P., and Ptashne, M. (1970). Active form of two coliphage repressors. Nature 227, 41-44.

Ptashne, M. (1967). Isolation of the *X* phage repressor. Proc. Nat. Acad. Sci. USA 57, 306-313.

Ptashne, M. (1967). Specific binding of the *X* phage repressor to DNA. Nature 214, 232-234.

12.10 The repressor and its operators define the immunity region

rev Friedman, D. I. and Gottesman, M. (1982). Lytic mode of lambda development In Lambda. In Lambda II, Eds. R. W. Hendrix, J. W. Roberts, F. W. Stahl and R. A. Weisberg, Cold Spring Harbor 21-51.

12.11 The DNA-binding form of repressor is a dimer

ref Pabo, C. 0. and Lewis, M. (1982). The operator-binding domain of *X* repressor: structure and DNA recognition. Nature 298, 443-447.

12.12 Repressor uses a helix-turn-helix motif to bind DNA

ref Sauer, R. T. et al. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature 298, 447-451.

12.13 The recognition helix determines specificity for DNA

ref Brennan, R. G. et al. (1990). Protein-DNA conformational changes in the crystal structure of a λCro-operator complex, Proc. Nat. Acad. Sci. USA 87, 8165-8169.

Wharton, R. L., Brown, E. L., and Ptashne, M. (1984). Substituting an α-helix switches the sequence specific DNA interactions of a repressor. Cell 38, 361-369.

12.14 Repressor dimers bind cooperatively to the operator

ref Bell, C. E., Frescura, P., Hochschild, A., and Lewis, M. (2000). Crystal structure of the lambda repressor C-terminal domain provides a model for cooperative operator binding. Cell 101, 801-811.

Johnson, A. D., Meyer, B. J., and Ptashne, M. (1979). Interactions between DNA-bound repressors govern regulation by the phage λrepressor. Proc. Nat. Acad. Sci. USA 76, 5061-5065.

Chapter 13

The replicon

13.1 Introduction

Whether a cell has only one chromosome (as in prokaryotes) or has many chromosomes (as in eukaryotes), the entire genome must be replicated precisely once for every cell division. How is the act of replication linked to the cell cycle?

Two general principles are used to compare the state of replication with the condition of the cell cycle:

- *Initiation of DNA replication commits the cell (prokaryotic or eukaryotic) to a further division.* From this standpoint, the number of descendants that a cell generates is determined by a series of decisions on whether or not to initiate DNA replication. Replication is controlled at the stage of initiation. *Once replication has started, it continues until the entire genome has been duplicated.*
- If replication proceeds, the consequent division cannot be permitted to occur until the replication event has been completed. Indeed, the completion of replication may provide a trigger for cell division. Then the duplicate genomes are segregated one to each daughter cell. The unit of segregation is the chromosome.

In prokaryotes, the initiation of replication is a single event involving a unique site on the bacterial chromosome, and the process of division is accomplished by the development of a septum that grows from the cell wall and divides the cell into two. In eukaryotic cells, initiation of replication is identified by the start of S phase, a protracted period during which DNA synthesis occurs, and which involves many individual initiation events. The act of division is accomplished by the reorganization of the cell at mitosis. In this chapter, we are concerned with the regulation of DNA replication. How is a cycle of replication initiated? What controls its progress and how is its termination signaled? In *29 Cell cycle and growth regulation*, we discuss the regulatory processes in eukaryotic cells that control entry into S phase and into mitosis, and also the "checkpoints" that postpone these actions until the appropriate conditions have been fulfilled.

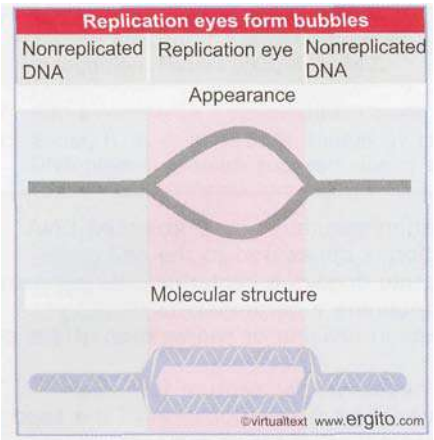


Figure 13.1 Replicated DNA is seen as a replication eye flanked by nonreplicated DNA.

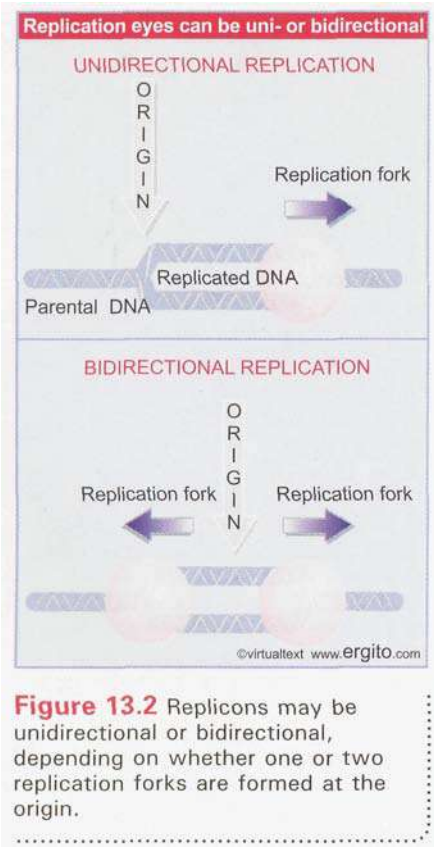


Figure 13.2 Replicons may be unidirectional or bidirectional, depending on whether one or two replication forks are formed at the origin.

The unit of DNA in which an individual act of replication occurs is called the replicon. Each replicon "fires" once and only once in each cell cycle. The replicon is defined by its possession of the control elements needed for replication. It has an origin at which replication is initiated. It may also have a terminus at which replication stops.

Any sequence attached to an origin—or, more precisely, not separated from an origin by a terminus—is replicated as part of that replicon. The origin is a *cis*-acting site, able to affect only that molecule of DNA on which it resides.

(The original formulation of the replicon (in prokaryotes) viewed it as a unit possessing both the origin *and* the gene coding for the regulator protein. Now, however, "replicon" is usually applied to eukaryotic chromosomes to describe a unit of replication that contains an origin; *trans-acting* regulator protein(s) may be coded elsewhere.)

A genome in a prokaryotic cell constitutes a single replicon; so the units of replication and segregation coincide. Initiation at a single origin sponsors replication of the entire genome, once for every cell division. Each haploid bacterium has a single chromosome, so this type of replication control is called single copy.

Bacteria may contain additional genetic information in the form of plasmids. *A plasmid is an autonomous circular DNA genome that constitutes a separate replicon* (see Figure 12.2). A plasmid replicon may show single copy control, which means that it replicates once every time the bacterial chromosome replicates. Or it may be under multicopy control, when it is present in a greater number of copies than the bacterial chromosome. Each phage or virus DNA also constitutes a replicon, able to initiate many times during an infectious cycle. Perhaps a better way to view the prokaryotic replicon, therefore, is to reverse the definition: *any DNA molecule that contains an origin can be replicated autonomously in the cell.*

A major difference in the organization of bacterial and eukaryotic genomes is seen in their replication. Each eukaryotic chromosome contains a large number of replicons. So the unit of segregation includes many units of replication. This adds another dimension to the problem of control. All the replicons on a chromosome must be fired during one cell cycle, although they are not active simultaneously, but are activated over a fairly protracted period. *Yet each of these replicons must be activated no more than once in each cell cycle.*

Some signal must distinguish replicated from nonreplicated replicons, so that replicons do not fire a second time. And because many replicons are activated independently, another signal must exist to indicate when the entire process of replicating all replicons has been completed.

We have begun to collect information about the construction of individual replicons, but we still have little information about the relationship between replicons. We do not know whether the pattern of replication is the same in every cell cycle. Are all origins always used or are some origins sometimes silent? Do origins always fire in the same order? If there are different classes of origins, what distinguishes them?

In contrast with nuclear chromosomes, which have a single-copy type of control, the DNA of mitochondria and chloroplasts may be regulated more like plasmids that exist in multiple copies per bacterium. There are multiple copies of each organelle DNA per cell, and the control of organelle DNA replication must be related to the cell cycle.

In all these systems, the key question is to define the sequences that function as origins and to determine how they are recognized by the appropriate proteins of the apparatus for replication. We start by considering the basic construction of replicons and the various forms that they take; following the consideration of the origin, we turn to the question of how replication of the genome is coordinated with bacterial division, and what is responsible for segregating the genomes to daughter bacteria.

13.2 Replicons can be linear or circular

Key Concepts

- A replicated region appears as an eye within nonreplicated DNA.
- A replication fork is initiated at the origin and then moves sequentially along DNA.
- Replication is unidirectional when a single replication fork is created at an origin.
- Replication is bidirectional when an origin creates two replication forks that move in opposite directions.

A molecule of DNA engaged in replication has two types of regions. **Figure 13.1** shows that when replicating DNA is viewed by electron microscopy, the replicated region appears as a replication eye within the nonreplicated DNA. The nonreplicated region consists of the parental duplex; this opens into the replicated region where the two daughter duplexes have formed.

The point at which replication is occurring is called the replication fork (sometimes also known as the growing point). *A replication fork moves sequentially along the DNA, from its starting point at the origin.* The origin may be used to start either unidirectional replication or bidirectional replication. The type of event is determined by whether one or two replication forks set out from the origin. In unidirectional replication, one replication fork leaves the origin and proceeds along the DNA. In bidirectional replication, two replication forks are formed; they proceed away from the origin in opposite directions.

The appearance of a replication eye does not distinguish between unidirectional and bidirectional replication. As depicted in **Figure 13.2,** the eye can represent either of two structures. If generated by unidirectional replication, the eye represents one fixed origin and one moving replication fork. If generated by bidirectional replication, the eye represents a pair of replication forks. In either case, the progress of replication expands the eye until ultimately it encompasses the whole replicon.

When a replicon is circular, the presence of an eye forms the 0-structure drawn in **Figure 13.3.** The successive stages of replication of the circular DNA of polyoma virus are visualized by electron microscopy in **Figure 13.4.**

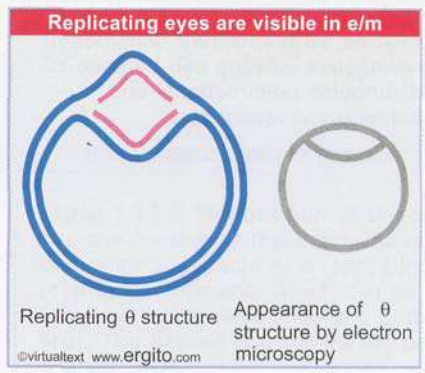


Figure 13.3 A replication eye forms a theta structure in circular DNA.

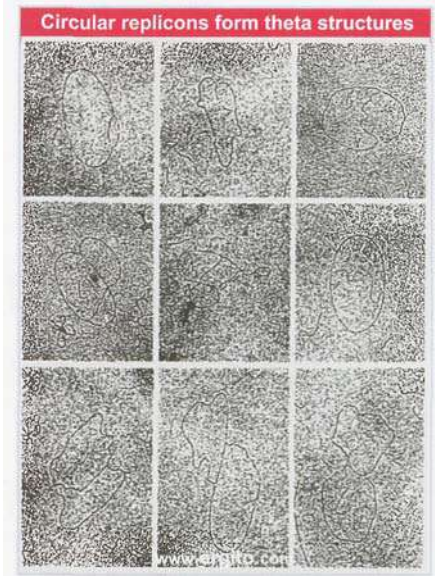


Figure 13.4 The replication eye becomes larger as the replication forks proceed along the replicon. Note that the "eye" becomes larger than the nonreplicated segment. The two sides of the eye can be defined because they are both the same length. Photograph kindly provided by Bernard Hirt.

13.3 Origins can be mapped by autoradiography and electrophoresis

Key Concepts

- Replication fork movement can be detected by autoradiography using radioactive pulses.
- Replication forks create Y-shaped structures that change the electrophoretic migration of DNA fragments.

Whether a replicating eye has one or two replication forks can be determined in two ways. The choice of method depends on whether the DNA is a defined molecule or an unidentified region of a cellular genome.

With a defined linear molecule, we can use electron microscopy to measure the distance of each end of the eye from the end of the DNA. Then the positions of the ends of the eyes can be compared in molecules

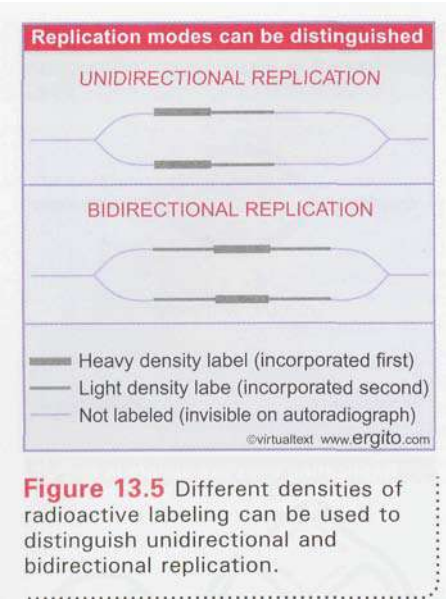


Figure 13.5 Different densities of radioactive labeling can be used to distinguish unidirectional and bidirectional replication.

that have eyes of different sizes. If replication is unidirectional, only one of the ends will move; the other is the fixed origin. If replication is bidirectional, both will move; the origin is the point midway between them.

With undefined regions of large genomes, two successive pulses of radioactivity can be used to label the movement of the replication forks. If one pulse has a more intense label than the other, they can be distinguished by the relative intensities of labeling. These can be visualized by autoradiography. **Figure 13.5** shows that unidirectional replication causes one type of label to be followed by the other at *one* end of the eye. Bidirectional replication produces a (symmetrical) pattern at *both* ends of the eye. This is the pattern usually observed in replicons of eukaryotic chromosomes.

A more recent method for mapping origins with greater resolution takes advantage of the effects that changes in shape have upon electrophoretic migration of DNA. **Figure 13.6** illustrates the two dimensional mapping technique, in which restriction fragments of replicating DNA are electrophoresed in a first dimension that separates by mass, and a second dimension where movement is determined more by shape. Different types of replicating molecules follow characteristic paths, measured by their deviation from the line that would be followed by a linear molecule of DNA that doubled in size.

A simple Y-structure, in which one fork moves along a linear fragment, follows a continuous path. An inflection point occurs when all three branches are the same length, and the structure therefore deviates most extensively from linear DNA. Analogous considerations determine the paths of double Y-structures or bubbles. An asymmetric bubble follows a discontinuous path, with a break at the point at which the bubble is converted to a Y-structure as one fork runs off the end.

Taken together, the various techniques for characterizing replicating DNA show that origins are most often used to initiate bidirectional replication. From this level of resolution, we must now proceed to the molecular level, to identify the *cis*-acting sequences that comprise the origin, and the *trans*-acting factors that recognize it.

13.4 The bacterial genome is a single circular replicon

Key Concepts

- Bacterial replicons are usually circles that are replicated bidirectionally from a single origin.
- The origin of *E. coli*, *oriC*, is 245 bp in length.
- The two replication forks usually meet halfway round the circle, but there are *ter* sites that cause termination if they go too far.

To be properly inherited, a bacterial replicon should support several functions:

- Initiating a replication cycle.
- Controlling the frequency of initiation events.
- Segregating replicated chromosomes to daughter cells.

The first two functions both are properties of the origin. Segregation could be an independent function, but in prokaryotic systems it is usually determined by sequences in the vicinity of the origin. Origins in eukaryotes do not function in segregation, but are concerned only with replication.

By Book_Crazy [IND]

As a general principle, the DNA constituting an origin can be isolated by its ability to support replication of any DNA sequence to which it is joined. When DNA from the origin is cloned into a molecule that lacks an origin, this will create a plasmid capable of autonomous replication *only if the DNA from the origin contains all the sequences needed to identify itself as an authentic origin for replication.*

Origins now have been identified in bacteria, yeast, chloroplasts, and mitochondria, although not in higher eukaryotes. A general feature is that the overall sequence composition is A·T-rich. We assume this is related to the need to melt the DNA duplex to initiate replication.

The genome of *E. coli* is replicated bidirectionally from a single origin, identified as the genetic locus *oriC*. The addition of *oriC* to any piece of DNA creates an artificial plasmid that can replicate in *E. coli*. By reducing the size of the cloned fragment of *oriC*, the region required to initiate replication has been equated with a fragment of 245 bp. (We discuss the properties of *oriC* and its interaction with the replication apparatus in more detail in *14.15 Creating the replication forks at an origin.)*

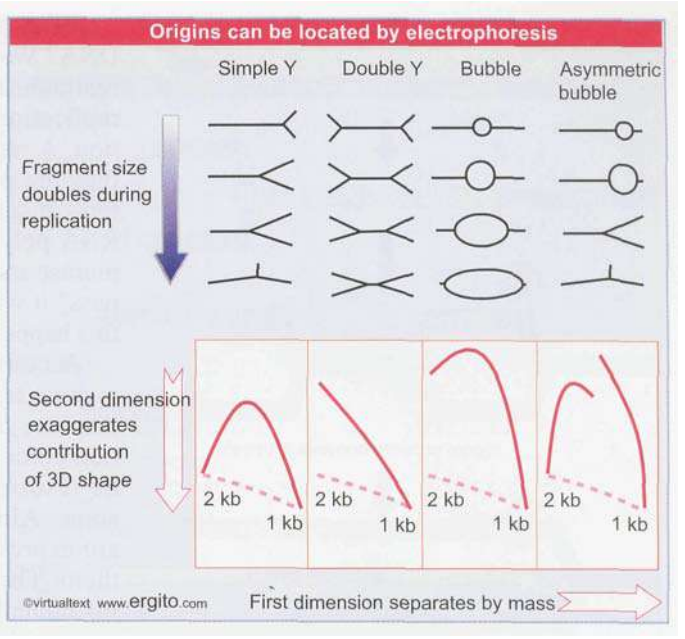


Figure 13.6 The position of the origin and the number of replicating forks determine the shape of a replicating restriction fragment, which can be followed by its electrophoretic path (solid line). The dashed line shows the path for a linear DNA.

Prokaryotic replicons are usually circular, so that the DNA forms a closed circle with no free ends. Circular structures include the bacterial chromosome itself, all plasmids, and many bacteriophages. They are also common in chloroplast and mitochondrial DNAs. Replication of a circular molecule avoids the problem of how to replicate the ends of a linear molecule, but poses the problem of how to terminate replication.

The bacterial chromosome is replicated bidirectionally as a single unit from *oriC*. Two replication forks initiate at *oriC* and move around the genome (at approximately the same speed) to a meeting point. Termination occurs in a discrete region. One interesting question is what ensures that the DNA is replicated right across the region where the forks meet. Following the termination of DNA replication itself, enzymes that manipulate higher-order structure of DNA are required for the two daughter chromosomes to be physically separated.

Sequences that cause termination are called *ter* sites. A *ter* site contains a short (~23 bp) sequence that causes termination *in vitro*. The termination sequences function in only one orientation. The *ter* site is recognized by a protein (called Tus in *E. coli* and *RTF* in *B. subtilis)* that recognizes the consensus sequence and prevents the replication fork from proceeding (see *14.17 The primosome is needed to restart replication).* However, deletion of the *ter* sites does not prevent normal replication cycles from occurring, although it does affect segregation of the daughter chromosomes (see *13.19 Chromosomal segregation may require site-specific recombination).*

Termination in *E. coli* and *B. subtilis* has the interesting features reported in **Figure 13.7.** We know that the replication forks usually meet and halt replication at a point midway round the chromosome from the origin. But two termination regions *(terE,D,A* and *terC,B* in *E. coli,* and *terI, terII* and also some other sites in *B. subtilis)* have been identified, located ~100 kb on either side of this meeting point. Each contains multiple terminators. Each terminus is specific for one direction of fork movement, and they are arranged in such a way that each fork would have to pass the other in order to reach the terminus to which it is susceptible. This arrangement creates a "replication fork trap". If for some reason one fork is delayed, so that the forks fail to meet at the usual central position, the more rapid fork will be trapped at the *ter* region to wait for the arrival of the slow fork.

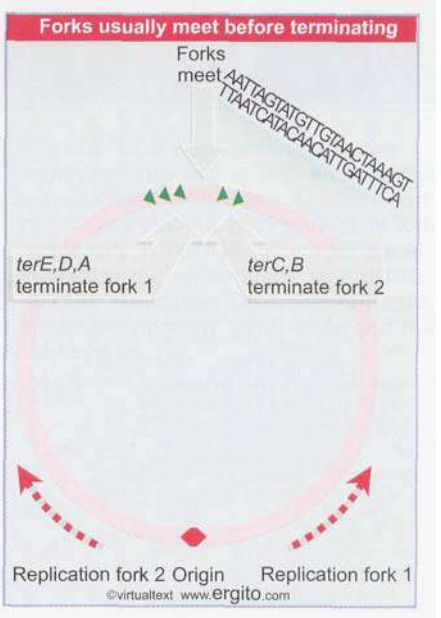


Figure 13.7 Replication termini in *E. coli* are located beyond the point at which the replication forks actually meet.

By Book_Crazy [IND]

What happens when a replication fork encounters a protein bound to DNA? We assume that repressors (for example) are displaced and then reattach. A particularly interesting question is what happens when a replication fork encounters an RNA polymerase engaged in transcription. A replication fork moves >10X faster than RNA polymerase. If they are proceeding in the same direction, either the replication fork must displace the polymerase or it must slow down as it waits for the RNA polymerase to reach its terminator. It appears that a DNA polymerase moving in the same direction as an RNA polymerase can "bypass" it without disrupting transcription, but we do not understand how this happens.

A conflict arises when the replication fork meets an RNA polymerase traveling in the opposite direction, that is, toward it. Can it displace the RNA polymerase? Or do both replication and transcription come to a halt? An indication that these encounters cannot easily be resolved is provided by the organization of the *E. coli* chromosome. Almost all active transcription units are oriented so that they are expressed in the same direction as the replication fork that passes them. The exceptions all comprise small transcription units that are infrequently expressed. The difficulty of generating inversions containing highly expressed genes argues that head-on encounters between a replication fork and a series of transcribing RNA polymerases may be lethal.

13.5 Each eukaryotic chromosome contains many replicons

Key Concepts

- Eukaryotic replicons are 40-100 kb in length.
- A chromosome is divided into many replicons.
- Individual replicons are activated at characteristic times during S phase.
- Regional activation patterns suggest that replicons near one another are activated at the same time.

In eukaryotic cells, the replication of DNA is confined to part of the cell cycle. S phase usually lasts a few hours in a higher eukaryotic cell. Replication of the large amount of DNA contained in a eukaryotic chromosome is accomplished by dividing it into many individual replicons. Only some of these replicons are engaged in replication at any point in S phase. Presumably each replicon is activated at a specific time during S phase, although the evidence on this issue is not decisive.

The start of S phase is signaled by the activation of the first replicons. Over the next few hours, initiation events occur at other replicons in an ordered manner. Much of our knowledge about the properties of the individual replicons is derived from autoradiographic studies, generally using the types of protocols illustrated in Figure 13.5 and Figure 13.6. Chromosomal replicons usually display bidirectional replication.

How large is the average replicon, and how many are there in the genome? A difficulty in characterizing the individual unit is that adjacent replicons may fuse to give large replicated eyes, as illustrated in **Figure 13.8.** The approach usually used to distinguish individual replicons from fused eyes is to rely on stretches of DNA in which several replicons can be seen to be active, presumably captured at a stage when

all have initiated around the same time, but before the forks of adjacent units have met.

In groups of active replicons, the average size of the unit is measured by the distance between the origins (that is, between the midpoints of adjacent replicons). The rate at which the replication fork moves can be estimated from the maximum distance that the autoradiographic tracks travel during a given time.

Individual replicons in eukaryotic genomes are relatively small, typically ~40 kb in yeast or fly, ~100 kb in animals cells. However, they can vary >10-fold in length within a genome. The rate of replication is ~2000 bp/min, which is much slower than the 50,000 bp/min of bacterial replication fork movement.

From the speed of replication, it is evident that a mammalian genome could be replicated in ~1 hour if all replicons functioned simultaneously. But S phase actually lasts for >6 hours in a typical somatic cell, which implies that no more than 15% of the replicons are likely to be active at any given moment. There are some exceptional cases, such as the early embryonic divisions of *Drosophila* embryos, where the duration of S phase is compressed by the simultaneous functioning of a large number of replicons.

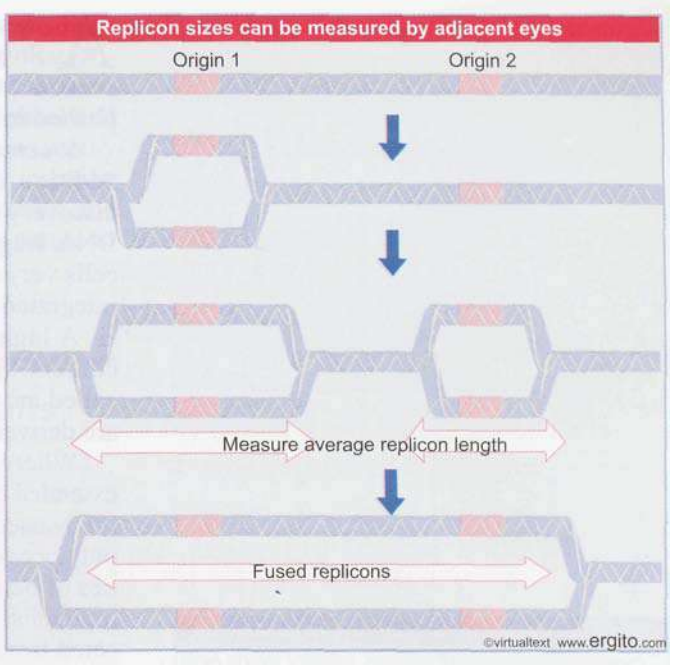


Figure 13.8 Measuring the size of the replicon requires a stretch of DNA in which adjacent replicons are active.

How are origins selected for initiation at different times during S phase? In *S. cerevisiae*, the default appears to be for origins to replicate early, but *cis*-acting sequences can cause origins linked to them to replicate at late times.

Available evidence suggests that chromosomal replicons do not have termini at which the replication forks cease movement and (presumably) dissociate from the DNA. It seems more likely that a replication fork continues from its origin until it meets a fork proceeding toward it from the adjacent replicon. We have already mentioned the potential topological problem of joining the newly synthesized DNA at the junction of the replication forks.

The propensity of replicons located in the same vicinity to be active at the same time could be explained by "regional" controls, in which groups of replicons are initiated more or less coordinately, as opposed to a mechanism in which individual replicons are activated one by one in dispersed areas of the genome. Two structural features suggest the possibility of large-scale organization. Quite large regions of the chromosome can be characterized as "early replicating" or "late replicating," implying that there is little interspersion of replicons that fire at early or late times. And visualization of replicating forks by labeling with DNA precursors identifies 100-300 "foci" instead of uniform staining; each focus shown in Figure 13.9 probably contains >300 replication forks. The foci could represent fixed structures through which replicating DNA must move.

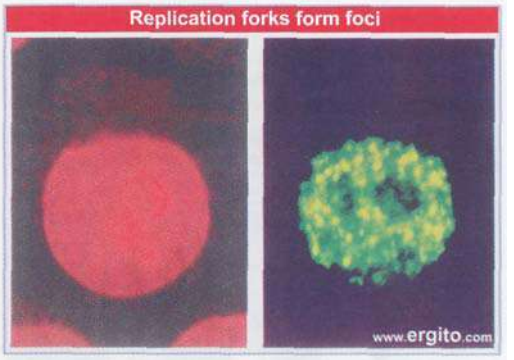


Figure 13.9 Replication forks are organized into foci in the nucleus. Cells were labeled with BrdU. The leftmost panel was stained with propidium iodide to identify bulk DNA. The right panel was stained using an antibody to BrdU to identify replicating DNA. Photographs kindly provided by A. D. Mills and Ron Laskey.

13.6 Replication origins can be isolated in yeast

Key Concepts

- Origins in *S. cerevisiae* are short A·T-rich sequences that have an essential 11 bp sequence.
- The ORC is a complex of 6 proteins that binds to an *ARS*.

By Book_Crazy [IND]

Any segment of DNA that has an origin should be able to replicate. So although plasmids are rare in eukaryotes, it may be possible to construct them by suitable manipulation *in vitro*. This has been accomplished in yeast, although not in higher eukaryotes.

S. cerevisiae mutants can be "transformed" to the wild phenotype by addition of DNA that carries a wild-type copy of the gene. The discovery of yeast origins resulted from the observation that some yeast DNA fragments (when circularized) are able to transform defective cells very efficiently. These fragments can survive in the cell in the unintegrated (autonomous) state, that is, as self-replicating plasmids.

A high-frequency transforming fragment possesses a sequence that confers the ability to replicate efficiently in yeast. This segment is called an ARS (for autonomously replicating sequence). *ARS* elements are derived from origins of replication.

Where *ARS* elements have been systematically mapped over extended chromosomal regions, it seems that only some of them are actually used to initiate replication. The others are silent, or possibly used only occasionally. If it is true that some origins have varying probabilities of being used, it follows that there can be no fixed termini between replicons. In this case, a given region of a chromosome could be replicated from different origins in different cell cycles.

An *ARS* element consists of an A·T-rich region that contains discrete sites in which mutations affect origin function. Base composition rather than sequence may be important in the rest of the region. **Figure 13.10** shows a systematic mutational analysis along the length of an origin. Origin function is abolished completely by mutations in a 14 bp "core" region, called the A domain, that contains an 11 bp consensus sequence consisting of A·T base pairs. This consensus sequence (sometimes called the *ACS* for ARS consensus sequence) is the only homology between known *ARS* elements.

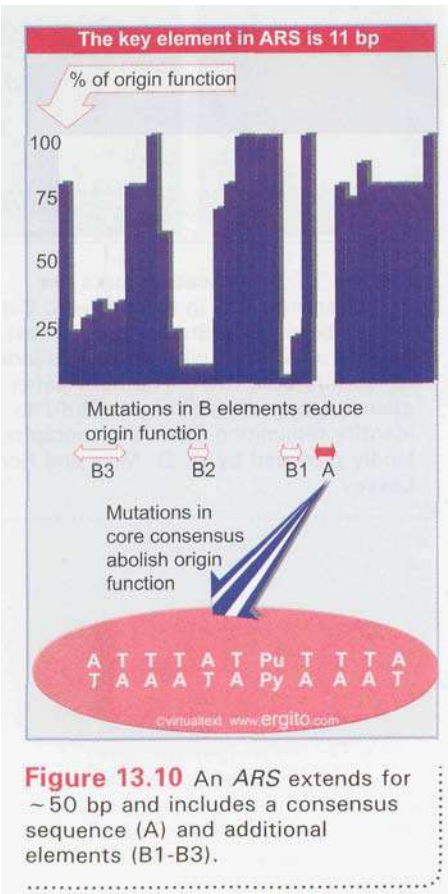


Figure 13.10 An *ARS* extends for ~50 bp and includes a consensus sequence (A) and additional elements (B1-B3).

Mutations in three adjacent elements, numbered B1-B3, reduce origin function. An origin can function effectively with any 2 of the B elements, so long as a functional A element is present. (Imperfect copies of the core consensus, typically conforming at 9/11 positions, are found close to, or overlapping with, each B element, but they do not appear to be necessary for origin function.)

The ORC (origin recognition complex) is a complex of 6 proteins with a mass of ~400 kD. ORC binds to the A and B1 elements on the A-T-rich strand, and is associated with *ARS* elements throughout the cell cycle. This means that initiation depends on changes in its condition rather than *de novo* association with an origin (see *14.21 Licensing factor consists of MCM proteins*).By counting the number of sites to which ORC binds, we can estimate that there are about 400 origins of replication in the yeast genome. This means that the average length of a replicon is ~35,000 bp. Counterparts to ORC are found in higher eukaryotic cells.

ORC was first found in *S. cerevisiae* (where it is called scORC), but similar complexes have now been characterized in *S. pombe* (spORC), *Drosophila* (DmORC) and *Xenopus* (XlORC). All of the ORC complexes bind to DNA. Although none of the binding sites have been characterized in the same detail as in *S. cerevisiae*, in several cases they are at locations associated with the initiation of replication. It seems clear that ORC is an initiation complex whose binding identifies an origin of replication. However, details of the interaction are clear only in *S. cerevisiae;* it is possible that additional components are required to recognize the origin in the other cases.

ARS elements satisfy the classic definition of an origin as a *cis-acting* sequence that causes DNA replication to initiate. Are similar elements to be found in higher eukaryotes? Difficulties in finding

sequences comparable to *ARS* elements that can support the existence of plasmids in higher eukaryotic cells suggest the possibility that origins may be more complex (or determined by features other than discrete *cis*-acting sequences). There are suggestions that some animal cell replicons may have complex patterns of initiation: in some cases, many small replication bubbles are found in one region, posing the question of whether there are alternative or multiple starts to replication, and whether there is a small discrete origin. It is fair to say that the nature of the higher eukaryotic origin remains to be established.

13.7 D loops maintain mitochondrial origins

Key Concepts

- Mitochondria use different origin sequences to initiate replication of each DNA strand.
- Replication of the H-strand is initiated in a D-loop.
- Replication of the L-strand is initiated when its origin is exposed by the movement of the first replication fork.

The origins of replicons in both prokaryotic and eukaryotic chromosomes are static structures: they comprise sequences of DNA that are recognized in duplex form and used to initiate replication at the appropriate time. Initiation requires separating the DNA strands and commencing bidirectional DNA synthesis. A different type of arrangement is found in mitochondria.

Replication starts at a specific origin in the circular duplex DNA. But initially only one of the two parental strands (the H strand in mammalian mitochondrial DNA) is used as a template for synthesis of a new strand. Synthesis proceeds for only a short distance, displacing the original partner (L) strand, which remains single-stranded, as illustrated in **Figure 13.11.** The condition of this region gives rise to its name as the *displacement* or D loop.

DNA polymerases cannot initiate synthesis, but require a priming 3' end (see *14.8 Priming is required to start DNA synthesis).* Replication at the H strand origin is initiated when RNA polymerase transcribes a primer. 3' ends are generated in the primer by an endonuclease that cleaves the DNA-RNA hybrid at several discrete sites. The endonuclease is specific for the triple structure of DNA-RNA hybrid plus the displaced DNA single strand. The 3' end is then extended into DNA by the DNA polymerase.

A single D loop is found as an opening of 500-600 bases in mammalian mitochondria. The short strand that maintains the D loop is unstable and turns over; it is frequently degraded and resynthesized to maintain the opening of the duplex at this site. Some mitochondrial DNAs possess several D loops, reflecting the use of multiple origins. The same mechanism is employed in chloroplast DNA, where (in higher plants) there are two D loops.

To replicate mammalian mitochondrial DNA, the short strand in the D loop is extended. The displaced region of the original L strand becomes longer, expanding the D loop. This expansion continues until it reaches a point about two-thirds of the way around the circle. Replication of this region exposes an origin in the displaced L strand. Synthesis of an H strand initiates at this site, which is used by a special primase that synthesizes a short RNA. The RNA is then extended by DNA polymerase, proceeding around the displaced single-stranded L template in the opposite direction from L-strand synthesis.

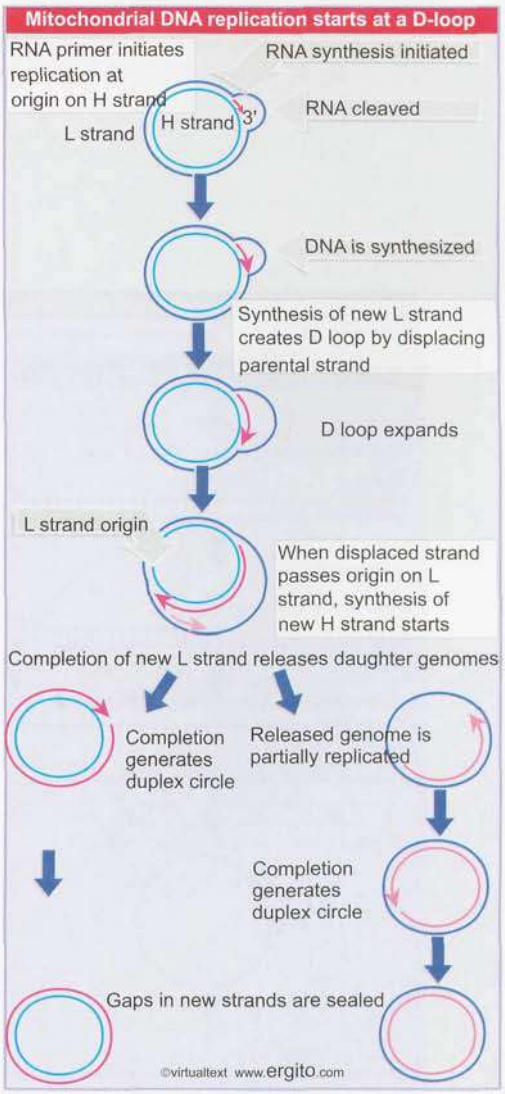


Figure 13.11 The D loop maintains an opening in mammalian mitochondrial DNA, which has separate origins for the replication of each strand.

By Book_Crazy [IND]

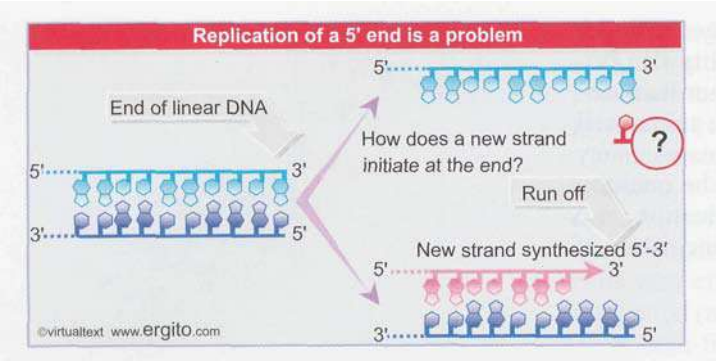


Figure 13.12 Replication could run off the 3' end of a newly synthesized linear strand, but could it initiate at a 5' end?

Because of the lag in its start, H-strand synthesis has proceeded only a third of the way around the circle when L-strand synthesis finishes. This releases one completed duplex circle and one gapped circle, which remains partially single-stranded until synthesis of the H strand is completed. Finally, the new strands are sealed to become covalently intact.

The existence of D loops exposes a general principle. *An origin can he a sequence of DNA that serves to initiate DNA synthesis using one strand as template.* The opening of the duplex does not necessarily lead to the initiation of replication on the other strand. In the case of mitochondrial DNA replication, the origins for replicating the complementary strands lie at different locations. Origins that sponsor replication of only one strand are also found in the rolling circle mode of replication (see *13.10 Rolling circles produce multimers of a replicon).*

13.8 The ends of linear DNA are a problem for replication

Key Concepts

- Special arrangements must be made to replicate the DNA strand with a 5′ end.

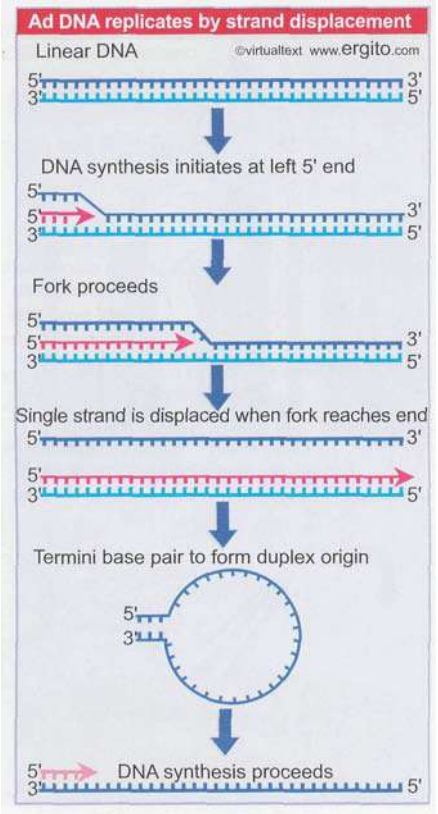


Figure 13.13 Adenovirus DNA *replication is initiated separately at* the two ends of the molecule and proceeds by strand displacement.

None of the replicons that we have considered so far have a linear end: either they are circular (as in the *E. coli* or mitochondrial genomes) or they are part of longer segregation units (as in eukaryotic chromosomes). But linear replicons occur, in some cases as single extrachromosomal units, and of course at the ends of eukaryotic chromosomes.

The ability of all known nucleic acid polymerases, DNA or RNA, to proceed only in the 5'- 3' direction poses a problem for synthesizing DNA at the end of a linear replicon. Consider the two parental strands depicted in **Figure 13.12.** The lower strand presents no problem: it can act as template to synthesize a daughter strand that runs right up to the end, where presumably the polymerase falls off. But to synthesize a complement at the end of the upper strand, synthesis must start right at the very last base (or else this strand would become shorter in successive cycles of replication).

We do not know whether initiation right at the end of a linear DNA is feasible. We usually think of a polymerase as binding at a site *surrounding* the position at which a base is to be incorporated. So a special mechanism must be employed for replication at the ends of linear replicons. Several types of solution may be imagined to accommodate the need to copy a terminus:

- The problem may be circumvented by converting a linear replicon into a circular or multimeric molecule. Phages such as T4 or lambda use such mechanisms (see *13.10 Rolling circles produce multimers of a replicon).*
- The DNA may form an unusual structure—for example, by creating a hairpin at the terminus, so that there is no free end. Formation of a crosslink is involved in replication of the linear mitochondrial DNA of *Paramecium.*

By Book_Crazy [IND]

Instead of being precisely determined, the end may be variable. Eukaryotic chromosomes may adopt this solution, in which the number of copies of a short repeating unit at the end of the DNA changes (see *19.18 Telomeres are synthesized by a ribonucleoprotein enzyme).* A mechanism to add or remove units makes it unnecessary to replicate right up to the very end.

A protein may intervene to make initiation possible at the actual terminus. Several linear viral nucleic acids have proteins that are *covalently linked to the 5′ terminal base.* The best characterized examples are adenovirus DNA, phage ϕ29 DNA, and poliovirus RNA.

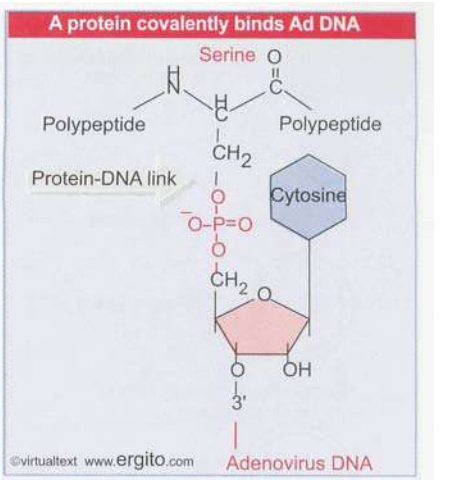


Figure 13.14 The 5′ terminal phosphate at each end of adenovirus DNA is covalently linked to serine in the 55 kD Ad-binding protein.

13.9 Terminal proteins enable initiation at the ends of viral DNAs

Key Concepts

- A terminal protein binds to the 5′ end of DNA and provides a cytidine nucleotide with a 3′-OH end that primes replication.

An example of initiation at a linear end is provided by adenovirus and ϕ29 DNAs, which actually replicate from both ends, using the mechanism of strand displacement illustrated in **Figure 13.13.** The same events can occur independently at either end. Synthesis of a new strand starts at one end, displacing the homologous strand that was previously paired in the duplex. When the replication fork reaches the other end of the molecule, the displaced strand is released as a free single strand. It is then replicated independently; this requires the formation of a duplex origin by base pairing between some short complementary sequences at the ends of the molecule.

In several viruses that use such mechanisms, a protein is found covalently attached to each 5′ end. In the case of adenovirus, a terminal protein is linked to the mature viral DNA via a phosphodiester bond to serine, as indicated in **Figure 13.14.**

How does the attachment of the protein overcome the initiation problem? The terminal protein has a dual role: it carries a cytidine nucleotide that provides the primer; and it is associated with DNA polymerase. In fact, linkage of terminal protein to a nucleotide is undertaken by DNA polymerase in the presence of adenovirus DNA. This suggests the model illustrated in **Figure 13.15.** The complex of polymerase and terminal protein, bearing the priming C nucleotide, binds to the end of the adenovirus DNA. The free 3'-OH end of the C nucleotide is used to prime the elongation reaction by the DNA polymerase. This generates a new strand whose 5' end is covalently linked to the initiating C nucleotide. (The reaction actually involves displacement of protein from DNA rather than binding *de novo.* The 5' end of adenovirus DNA is bound to the terminal protein that was used in the previous replication cycle. The old terminal protein is displaced by the new terminal protein for each new replication cycle.)

Terminal protein binds to the region located between 9 and 18 bp from the end of the DNA. The adjacent region, between positions 17 and 48, is essential for the binding of a host protein, nuclear factor I, which is also required for the initiation reaction. The initiation complex may therefore form between positions 9 and 48, a fixed distance from the actual end of the DNA.

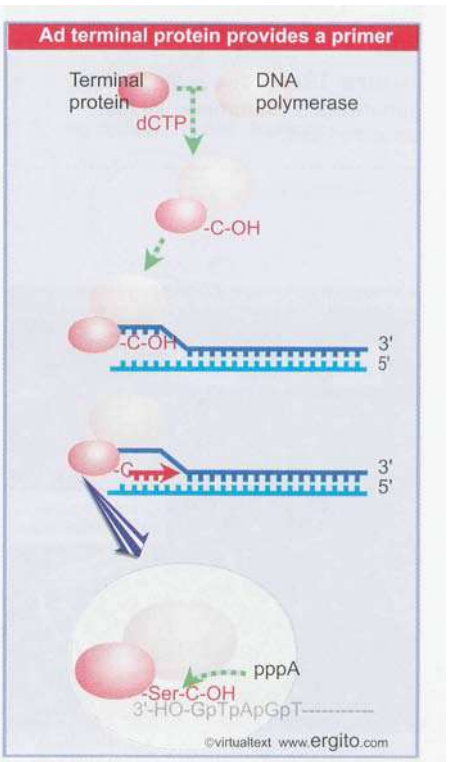


Figure 13.15 Adenovirus terminal protein binds to the 5′ end of DNA and provides a C-OH end to prime synthesis of a new DNA strand.

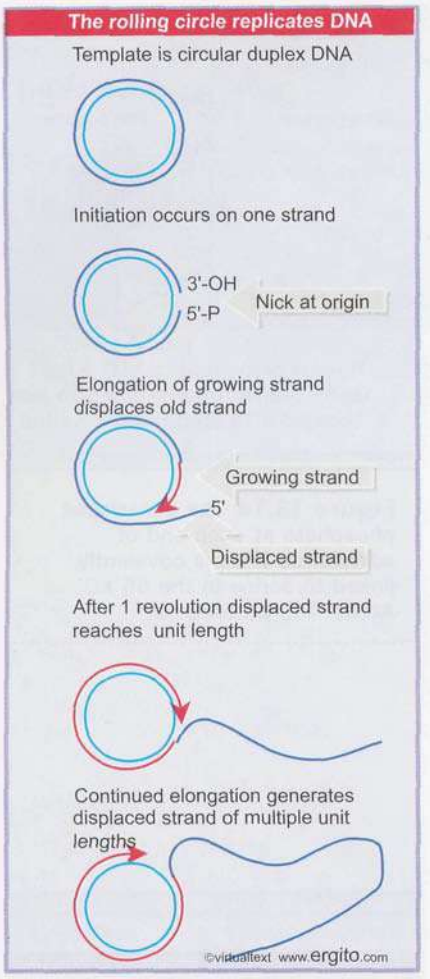


Figure 13.16 The rolling circle generates a multimeric single-stranded tail.

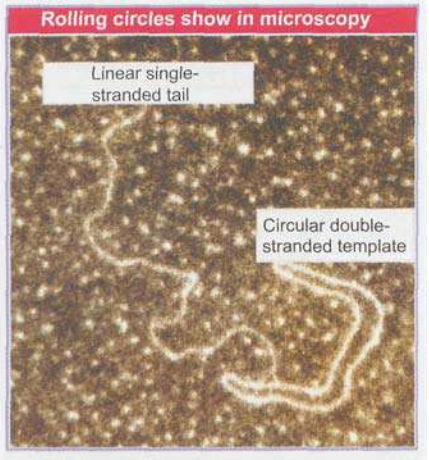


Figure 13.17 A rolling circle appears as a circular molecule with a linear tail by electron microscopy. Photograph kindly provided by David Dressler.

13.10 Rolling circles produce multimers of a replicon

Key Concepts

- A rolling circle generates single-stranded multimers of the original sequence.

The structures generated by replication depend on the relationship between the template and the replication fork. The critical features are whether the template is circular or linear, and whether the replication fork is engaged in synthesizing both strands of DNA or only one.

Replication of only one strand is used to generate copies of some circular molecules. A nick opens one strand, and then the free 3′-OH end generated by the nick is extended by the DNA polymerase. The newly synthesized strand displaces the original parental strand. The ensuing events are depicted in **Figure 13.16.**

This type of structure is called a rolling circle, because the growing point can be envisaged as rolling around the circular template strand. It could in principle continue to do so indefinitely. As it moves, the replication fork extends the outer strand and displaces the previous partner. An example is shown in the electron micrograph of **Figure 13.17.**

Because the newly synthesized material is covalently linked to the original material, the displaced strand has the original unit genome at its 5' end. The original unit is followed by any number of unit genomes, synthesized by continuing revolutions of the template. Each revolution displaces the material synthesized in the previous cycle.

The rolling circle is put to several uses *in vivo.* Some pathways that are used to replicate DNA are depicted in **Figure 13.18.**

Cleavage of a unit length tail generates a copy of the original circular replicon in linear form. The linear form may be maintained as a single strand or may be converted into a duplex by synthesis of the complementary strand (which is identical in sequence to the template strand of the original rolling circle).

The rolling circle provides a means for amplifying the original (unit) replicon. This mechanism is used to generate amplified rDNA in the *Xenopus* oocyte. The genes for rRNA are organized as a large number of contiguous repeats in the genome. A single repeating unit from the genome is converted into a rolling circle. The displaced tail, containing many units, is converted into duplex DNA; later it is cleaved from the circle so that the two ends can be joined together to generate a large circle of amplified rDNA. The amplified material therefore consists of a large number of identical repeating units.

13.11 Rolling circles are used to replicate phage genomes

Key Concepts

- The ϕX A protein is a c/s-acting relaxase that generates single-stranded circles from the tail produced by rolling circle replication.

By Book_Crazy [IND]

Replication by rolling circles is common among bacteriophages. Unit genomes can be cleaved from the displaced tail, generating monomers that can be packaged into phage particles or used for further replication cycles. A more detailed view of a phage replication cycle that is centered on the rolling circle is given in **Figure 13.19.**

Phage ϕX174 consists of a single-stranded circular DNA, known as the plus (+) strand. A complementary strand, called the minus (-) strand, is synthesized. This action generates the duplex circle shown at the top of the figure, which is then replicated by a rolling circle mechanism.

The duplex circle is converted to a covalently closed form, which becomes supercoiled. A protein coded by the phage genome, the A protein, nicks the (+) strand of the duplex DNA at a specific site that defines the origin for replication. After nicking the origin, the A protein remains connected to the 5' end that it generates, while the 3' end is extended by DNA polymerase.

The structure of the DNA plays an important role in this reaction, for the DNA can be nicked *only when it is negatively supercoiled* (wound about its axis in space in the opposite sense from the handedness of the double helix; see *15.12 Supercoiling affects the structure of DNA).* The A protein is able to bind to a single-stranded decamer fragment of DNA that surrounds the site of the nick. This suggests that the supercoiling is needed to assist the formation of a single-stranded region that provides the A protein with its binding site. (An enzymatic activity in which a protein cleaves duplex DNA and binds to a released 5' end is sometimes called a relaxase.) The nick generates a 3'-OH end and a 5′-phosphate end (covalently attached to the A protein), both of which have roles to play in ϕX174 replication.

Using the rolling circle, the 3'-OH end of the nick is extended into a new chain. The chain is elongated around the circular (-) strand template, until it reaches the starting point and displaces the origin. Now the A protein functions again. It remains connected with the rolling circle as well as to the 5' end of the displaced tail, and it is therefore in the vicinity as the growing point returns past the origin. So the same A protein is available again to recognize the origin and nick it, now attaching to the end generated by the new nick. The cycle can be repeated indefinitely.

Following this nicking event, the displaced single (+) strand is freed as a circle. The A protein is involved in the circularization. In fact, the joining of the 3' and 5' ends of the (+) strand product is accomplished by the A protein as part of the reaction by which it is released at the end of one cycle of replication, and starts another cycle.

The A protein has an unusual property that may be connected with these activities. It is *cis*-acting *in vivo.* (This behavior is not reproduced *in vitro,* as can be seen from its activity on any DNA template in a cell-free system.) *The implication is that in vivo the A protein synthesized by a particular genome can attach only to the DNA of that genome.* We do not know how this is accomplished. However, its activity *in vitro* shows how it remains associated with the same parental (-) strand template. The A protein has two active sites; this may allow it to cleave the "new" origin while still retaining the "old" origin; then it ligates the displaced strand into a circle.

The displaced (+) strand may follow either of two fates after circularization. During the replication phase of viral infection, it may be used as a template to synthesize the complementary (-) strand. The duplex circle may then be used as a rolling circle to generate more progeny. During phage morphogenesis, the displaced (+) strand is packaged into the phage virion.

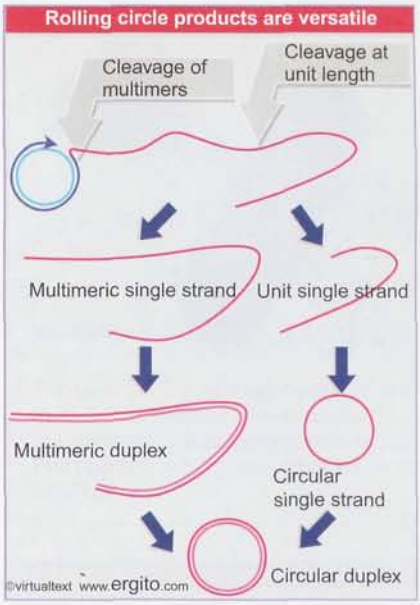


Figure 13.18 The fate of the displaced tail determines the types of products generated by rolling circles. Cleavage at unit length generates monomers. Cleavage of multimers generates a series of tandemly repeated copies of the original unit.

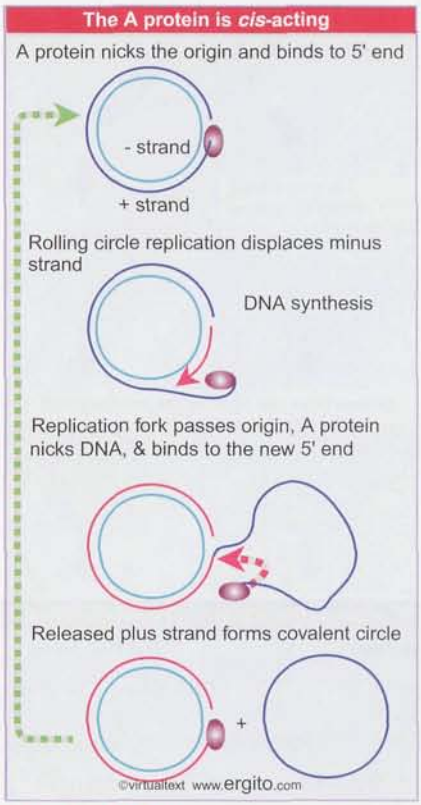


Figure 13.19 ϕX174 RF DNA is a template for synthesizing single-stranded viral circles. The A protein remains attached to the same genome through indefinite revolutions, each time nicking the origin on the viral (+) strand and transferring to the new 5′ end.

By Book_Crazy [IND]

13.12 The F plasmid is transferred by conjugation between bacteria

Key Concepts

- A free F factor is a replicon that is maintained at the level of one plasmid per bacterial chromosome.
- An F factor can integrate into the bacterial chromosome, in which case its own replication system is suppressed.
- The F factor codes for specific pili that form on the surface of the bacterium.
- An F-pilus enables an F-positive bacterium to contact an F-negative bacterium and to initiate conjugation.

Another example of a connection between replication and the propagation of a genetic unit is provided by bacterial conjugation, in which a plasmid genome or host chromosome is transferred from one bacterium to another.

Conjugation is mediated by the F plasmid, which is the classic example of an episome, an element that may exist as a free circular plasmid, or that may become integrated into the bacterial chromosome as a linear sequence (like a lysogenic bacteriophage). The F plasmid is a large circular DNA, ~100 kb in length.

The F factor can integrate at several sites in the *E. coli* chromosome, often by a recombination event involving certain sequences (called IS sequences; see *16.5 Transposons cause rearrangement of DNA)* that are present on both the host chromosome and F plasmid. In its free (plasmid) form, the F plasmid utilizes its own replication origin *(oriV)* and control system, and is maintained at a level of one copy per bacterial chromosome. When it is integrated into the bacterial chromosome, this system is suppressed, and F DNA is replicated as a part of the chromosome.

The presence of the F plasmid, whether free or integrated, has important consequences for the host bacterium. Bacteria that are F-positive are able to conjugate (or mate) with bacteria that are F-negative. Conjugation involves a contact between donor (F-positive) and recipient (F-negative) bacteria; contact is followed by transfer of the F factor. If the F factor exists as a free plasmid in the donor bacterium, it is transferred as a plasmid, and the infective process converts the F-negative recipient into an F-positive state. If the F factor is present in an integrated form in the donor, the transfer process may also cause some or all of the bacterial chromosome to be transferred. Many plasmids have conjugation systems that operate in a generally similar manner, but the F factor was the first to be discovered, and remains the paradigm for this type of genetic transfer.

A large (~33 kb) region of the F plasmid, called the transfer region, is required for conjugation. It contains ~40 genes that are required for the transmission of DNA; their organization is summarized in **Figure 13.20.** The genes are named as *tra* and *trb* loci. Most of them are expressed coordinately as part of a single 32 kb transcription unit (the *traY-I* unit). *traM* and *traJ* are expressed separately. *traJ* is a regulator that turns on both *traM and traY-I.* On the opposite strand, *finP* is a regulator that codes for a small antisense RNA that turns off *traJ.* Its activity requires expression of another gene, *finO.* Only four of the *tra* genes in the major transcription unit are concerned directly with the transfer of DNA; most are concerned with the properties of the bacterial cell surface and with maintaining contacts between mating bacteria.

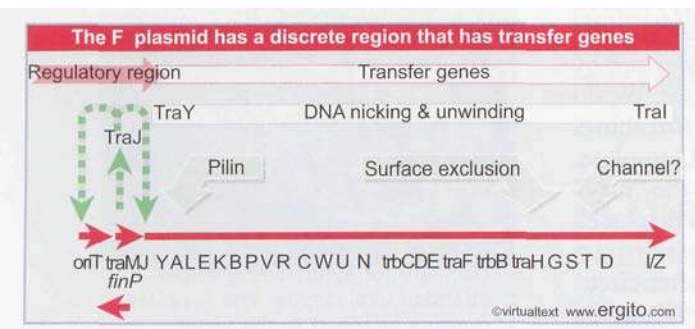


Figure 13.20 The *tra* region of the F plasmid contains the genes needed for bacterial conjugation.

By Book_Crazy [IND]

F-positive bacteria possess surface appendages called pili (singular pilus) that are coded by the F factor. The gene *traA* codes for the single subunit protein, pilin, that is polymerized into the pilus. At least 12 *tra* genes are required for the modification and assembly of pilin into the pilus. The F-pili are hair-like structures, 2-3 μm long, that protrude from the bacterial surface. A typical F-positive cell has 2-3 pili. The pilin subunits are polymerized into a hollow cylinder, ~8 nm in diameter, with a 2 nm axial hole.

Mating is initiated when the tip of the F-pilus contacts the surface of the recipient cell. **Figure 13.21** shows an example of *E. coli* cells beginning to mate. A donor cell does not contact other cells carrying the F factor, because the genes *traS* and *traT* code for "surface exclusion" proteins that make the cell a poor recipient in such contacts. This effectively restricts donor cells to mating with F-negative cells. (And the presence of F-pili has secondary consequences; they provide the sites to which RNA phages and some single-stranded DNA phages attach, so F-positive bacteria are susceptible to infection by these phages, whereas F-negative bacteria are resistant.)

The initial contact between donor and recipient cells is easily broken, but other *tra* genes act to stabilize the association, bringing the mating cells closer together. The F pili are essential for initiating pairing, but retract or disassemble as part of the process by which the mating cells are brought into close contact. There must be a channel through which DNA is transferred, but the pilus itself does not appear to provide it. TraD is an inner membrane protein in F^+ bacteria that is necessary for transport of DNA and it may provide or be part of the channel.

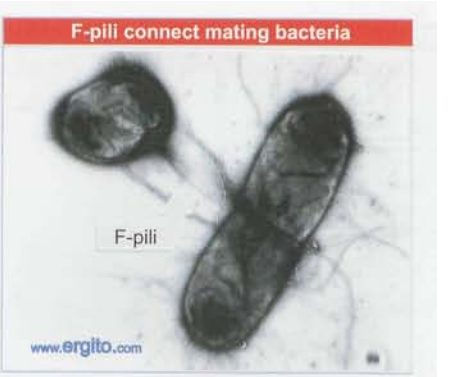


Figure 13.21 Mating bacteria are initially connected when donor F pili contact the recipient bacterium. Photograph kindly provided by Ron Skurray.

13.13 Conjugation transfers single-stranded DNA

Key Concepts

- Transfer of an F factor is initiated when rolling circle replication begins at *oriT*.
- The free 5' end initiates transfer into the recipient bacterium.
- The transferred DNA is converted into double-stranded form in the recipient bacterium.
- When an F factor is free, conjugation "infects" the recipient bacterium with a copy of the F factor.
- When an F factor is integrated, conjugation causes transfer of the bacterial chromosome until the process is interrupted by (random) breakage of the contact between donor and recipient bacteria.

Transfer of the F factor is initiated at a site called *ori T,* the origin of transfer, which is located at one end of the transfer region. The transfer process may be initiated when TraM recognizes that a mating pair has formed. Then TraY binds near *oriT* and causes TraI to bind. TraI is a relaxase, like ϕX174 A protein. TraI nicks *oriT* at *a* unique site (called *nic),* and then forms a covalent link to the 5′ end that has been generated. TraI also catalyzes the unwinding of ~200 bp of DNA (this is a helicase activity; see *14.7 The ϕX model system shows how single-stranded DNA is generated for replication).* **Figure 13.22** shows that the freed 5′ end leads the way into the recipient bacterium. A complement for the transferred single strand is synthesized in the recipient bacterium, which as a result is converted to the F-positive state.

A complementary strand must be synthesized in the donor bacterium to replace the strand that has been transferred. If this happens concomitantly with the transfer process, the state of the F plasmid will resemble the rolling circle of Figure 13.16 (and will not generate the

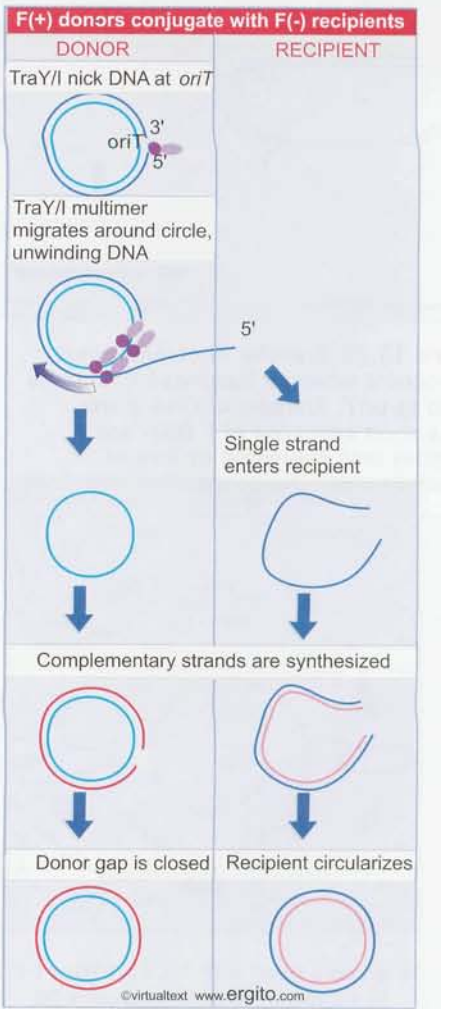


Figure 13.22 Transfer of DNA occurs when the F factor is nicked at *oriT* and a single strand is led by the 5′ end into the recipient. Only one unit length is transferred. Complementary strands are synthesized to the single strand remaining in the donor and to the strand transferred into the recipient.

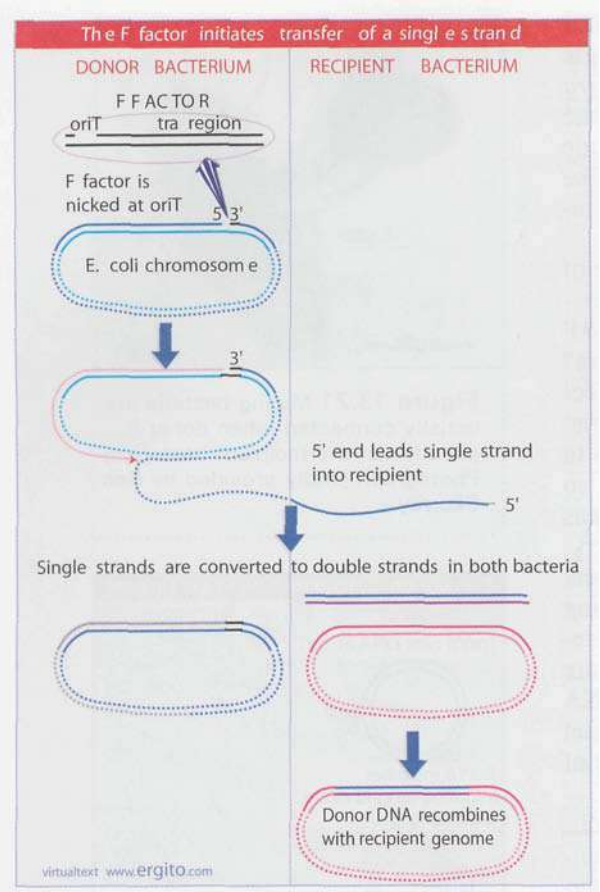


Figure 13.23 Transfer of chromosomal DNA occurs when an integrated F factor is nicked at *oriT*. Transfer of DNA starts with a short sequence of F DNA and continues until prevented by loss of contact between the bacteria.

extensive single-stranded regions shown in Figure 13.22). Conjugating DNA usually appears like a rolling circle, but replication as such is not necessary to provide the driving energy, and single-strand transfer is independent of DNA synthesis. Only a single unit length of the F factor is transferred to the recipient bacterium. This implies that some (unidentified) feature terminates the process after one revolution, after which the covalent integrity of the F plasmid is restored.

When an integrated F plasmid initiates conjugation, the orientation of transfer is directed away from the transfer region, into the bacterial chromosome. **Figure 13.23** shows that, following a short leading sequence of F DNA, bacterial DNA is transferred. The process continues until it is interrupted by the breaking of contacts between the mating bacteria. It takes ~100 minutes to transfer the entire bacterial chromosome, and under standard conditions, contact is often broken before the completion of transfer.

Donor DNA that enters a recipient bacterium is converted to double-stranded form, and may recombine with the recipient chromosome. (Note that two recombination events are required to insert the donor DNA.) So conjugation affords a means to exchange genetic material between bacteria (a contrast with their usual asexual growth). A strain of *E. coli* with an integrated F factor supports such recombination at relatively high frequencies (compared to strains that lack integrated F factors); such strains are described as Hfr (for high frequency recombination). Each position of integration for the F factor gives rise to a different Hfr strain, with a characteristic pattern of transferring bacterial markers to a recipient chromosome.

Contact between conjugating bacteria is usually broken before transfer of DNA is complete. As a result, the probability that a region of the bacterial chromosome will be transferred depends upon its distance from *oriT*. Bacterial genes located close to the site of F integration (in the direction of transfer) enter recipient bacteria first, and are therefore found at greater frequencies than those located farther away that enter later. This gives rise to a gradient of transfer frequencies around the chromosome, declining from the position of F integration. Marker positions on the donor chromosome can be assayed in terms of the time at which transfer occurs, and this gave rise to the standard description of the *E. coli* chromosome as a map divided into 100 minutes. The map refers to transfer times from a particular Hfr strain; the starting point for the gradient of transfer is different for each Hfr strain, being determined by the site where the F factor has integrated into the bacterial genome.

13.14 Replication is connected to the cell cycle

Key Concepts

- The doubling time of *E. coli* can vary over a 10× range, depending on growth conditions.
- It requires 40 minutes to replicate the bacterial chromosome (at normal temperature).
- Completion of a replication cycle triggers a bacterial division 20 minutes later.
- If the doubling time is <60 minutes, a replication cycle is initiated before the division resulting from the previous replication cycle.
- Fast rates of growth therefore produce multiforked chromosomes.
- A replication cycle is initiated at a constant ratio of mass/number of chromosome origins.
- There is one origin per unit cell of 1.7 μm in length.

By Book_Crazy [IND]

Bacteria have two links between replication and cell growth:

- The frequency of initiation of cycles of replication is adjusted to fit the rate at which the cell is growing.
- The completion of a replication cycle is connected with division of the cell.

The rate of bacterial growth is assessed by the doubling time, the period required for the number of cells to double. The shorter the doubling time, the faster the growth rate. *E. coli* cells can grow at rates ranging from doubling times as fast as 18 minutes to slower than 180 minutes. Because the bacterial chromosome is a single replicon, the frequency of replication cycles is controlled by the number of initiation events at the single origin. The replication cycle can be defined in terms of two constants:

- C is the fixed time of ~40 minutes required to replicate the entire bacterial chromosome. Its duration corresponds to a rate of replication fork movement of ~50,000 bp/minute. (The rate of DNA synthesis is more or less invariant at a constant temperature; it proceeds at the same speed unless and until the supply of precursors becomes limiting.)
- *D* is the fixed time of ~20 minutes that elapses between the completion of a round of replication and the cell division with which it is connected. This period may represent the time required to assemble the components needed for division.

(The constants C and *D* can be viewed as representing the maximum speed with which the bacterium is capable of completing these processes. They apply for all growth rates between doubling times of 18 and 60 minutes, but both constant phases become longer when the cell cycle occupies >60 minutes.)

A cycle of chromosome replication must be initiated a fixed time before a cell division, $C + D = 60$ minutes. For bacteria dividing more frequently than every 60 minutes, a cycle of replication must be initiated before the end of the preceding division cycle.

Consider the example of cells dividing every 35 minutes. The cycle of replication connected with a division must have been initiated 25 minutes before the preceding division. This situation is illustrated in Figure 13.24, which shows the chromosomal complement of a bacterial cell at 5-minute intervals throughout the cycle.

At division (35/0 minutes), the cell receives a partially replicated chromosome. The replication fork continues to advance. At 10 minutes, when this "old" replication fork has not yet reached the terminus, initiation occurs at both origins on the partially replicated chromosome. The start of these "new" replication forks creates a multiforked chromosome.

At 15 minutes—that is, at 20 minutes before the next division—the old replication fork reaches the terminus. Its arrival allows the two daughter chromosomes to separate; each of them has already been partially replicated by the new replication forks (which now are the only replication forks). These forks continue to advance.

At the point of division, the two partially replicated chromosomes segregate. This recreates the point at which we started. The single replication fork becomes "old," it terminates at 15 minutes, and 20 minutes later there is a division. We see that the initiation event occurs $1^{25}/_{35}$ cell cycles before the division event with which it is associated.

The general principle of the link between initiation and the cell cycle is that, as cells grow more rapidly (the cycle is shorter), the initiation event occurs an increasing number of cycles before the related

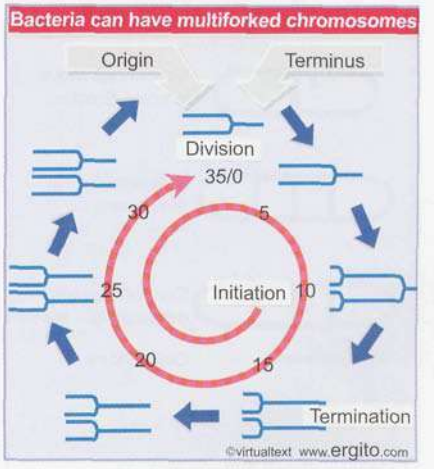


Figure 13.24 The fixed interval of 60 minutes between initiation of replication and cell division produces multiforked chromosomes in rapidly growing cells. Note that only the replication forks moving in one direction are shown; actually the chromosome is replicated symmetrically by two sets of forks moving in opposite directions on circular chromosomes.

division. There are correspondingly more chromosomes in the individual bacterium. This relationship can be viewed as the cell's response to its inability to reduce the periods of C and D to keep pace with the shorter cycle.

How does the cell know when to initiate the replication cycle? The initiation event occurs at a constant ratio of cell mass to the number of chromosome origins. Cells growing more rapidly are larger and possess a greater number of origins. The growth of the bacterium can be described in terms of the unit cell, an entity 1.7 μm long. A bacterium contains one origin per unit cell; a rapidly growing cell with two origins will be 1.7-3.4 μm long. In terms of Figure 13.24, it is at the point 10 minutes after division that the cell mass has increased sufficiently to support an initiation event at both available origins.

How is cell mass titrated? An initiator protein could be synthesized continuously throughout the cell cycle; accumulation of a critical amount would trigger initiation. This explains why protein synthesis is needed for the initiation event. An alternative possibility is that an inhibitor protein might be synthesized at a fixed point, and diluted below an effective level by the increase in cell volume.

13.15 The septum divides a bacterium into progeny each containing a chromosome

Key Concepts

- Septum formation is initiated at the annulus, which is a ring around the cell where the structure of the envelope is altered.
- New annuli are initiated at 50% of the distance from the septum to each end of the bacterium.
- When the bacterium divides, each daughter has an annulus at the midcenter position.
- Septation starts when the cell reaches a fixed length.
- The septum consists of the same peptidoglycans that comprise the bacterial envelope.

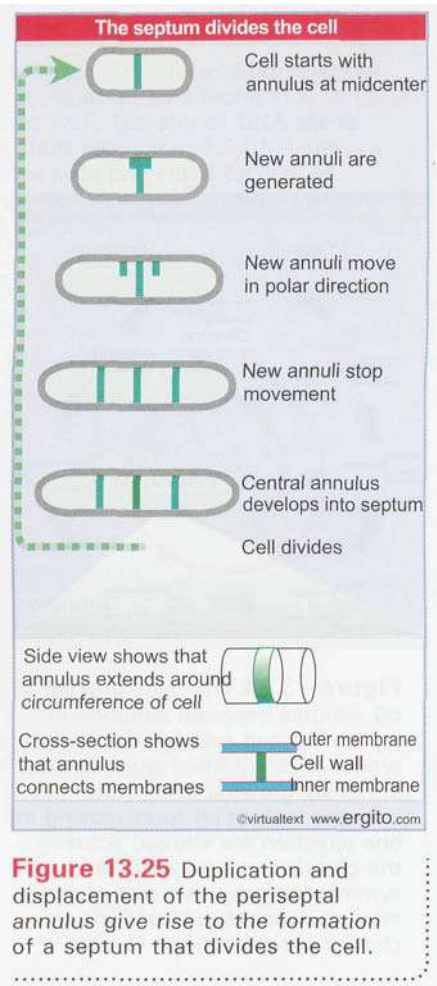


Figure 13.25 Duplication and displacement of the periseptal annulus give rise to the formation of a septum that divides the cell.

Chromosome segregation in bacteria is especially interesting because the DNA itself is involved in the mechanism for partition. (This contrasts with eukaryotic cells, in which segregation is achieved by the complex apparatus of mitosis.) The bacterial apparatus is quite accurate, however; anucleate cells form <0.03% of a bacterial population.

The division of a bacterium into two daughter cells is accomplished by the formation of a septum, a structure that forms in the center of the cell as an invagination from the surrounding envelope. The septum forms an impenetrable barrier between the two parts of the cell and provides the site at which the two daughter cells eventually separate entirely. Two related questions address the role of the septum in division: what determines the location at which it forms; and what ensures that the daughter chromosomes lie on opposite sides of it?

The formation of the septum is preceded by the organization of the periseptal annulus. This is observed as a zone in *E. coli* or *S. typhimurium* in which the structure of the envelope is altered so that the inner membrane is connected more closely to the cell wall and outer membrane layer. As its name suggests, the annulus extends around the cell. **Figure 13.25** summarizes its development.

The annulus starts at a central position in a new cell. As the cell grows, two events occur. A septum forms at the midcell position defined by the annulus. And new annuli form on either side of the initial

By Book_Crazy [IND]

annulus. These new annuli are displaced from the center and move along the cell to positions at 1/4 and 3/4 of the cell length. These will become the midcell positions after the next division. The displacement of the periseptal annulus to the correct position may be the crucial event that ensures the division of the cell into daughters of equal size. (The mechanism of movement is unknown.) Septation begins when the cell reaches a fixed length (2L), and the distance between the new annuli is always L. We do not know how the cell measures length, but the relevant parameter appears to be linear distance as such (not area or volume).

The septum consists of the same components as the cell envelope: there is a rigid layer of peptidoglycan in the periplasm, between the inner and outer membranes. The peptidoglycan is made by polymerization of tri- or pentapeptide-disaccharide units in a reaction involving connections between both types of subunit (transpeptidation and transglycosylation). The rod-like shape of the bacterium is maintained by a pair of activities, PBP2 and RodA. They are interacting proteins, coded by the same operon. RodA is a member of the SEDS family (SEDS stands for shape, elongation, division, and sporulation) that is present in all bacteria that have a peptidoglycan cell wall. Each SEDS protein functions together with a specific transpeptidase, which catalyzes the formation of the cross-links in the peptidoglycan. PBP2 (penicillin-binding protein 2) is the transpeptidase that interacts with RodA. Mutations in the gene for either protein cause the bacterium to lose its extended shape, becoming round. This demonstrates the important principle that shape and rigidity can be determined by the simple extension of a polymeric structure. Another enzyme is responsible for generating the peptidoglycan in the septum (see *13.17 FtsZ is necessary for septum formation).* The septum initially forms as a double layer of peptidoglycan, and the protein EnvA is required to split the covalent links between the layers, so that the daughter cells may separate.

The behavior of the periseptal annulus suggests that the mechanism for measuring position is associated with the cell envelope. It is plausible to suppose that the envelope could also be used to ensure segregation of the chromosomes. A direct link between DNA and the membrane could account for segregation. If daughter chromosomes are attached to the membrane, they could be physically separated when the septum forms. **Figure 13.26** shows that the formation of a septum could segregate the chromosomes into the different daughter cells if the origins are connected to sites that lie on either side of the periseptal annulus.

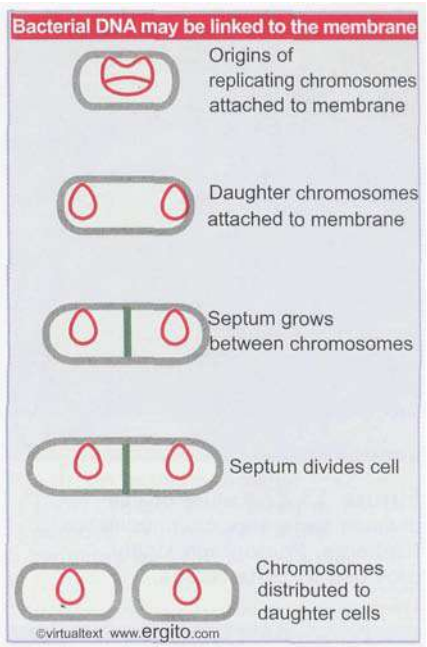


Figure 13.26 Attachment of bacterial DNA to the membrane could provide a mechanism for segregation.

13.16 Mutations in division or segregation affect cell shape

Key Concepts

- *fts* mutants form long filaments because the septum fails to form to divide the daughter bacteria.
- Minicells form in mutants that produce too many septa; they are small and lack DNA.
- Anucleate cells of normal size are generated by partition mutants in which the duplicate chromosomes fail to separate.

A difficulty in isolating mutants that *affect cell* division is that mutations in the critical functions may be lethal and/or pleiotropic. For example, if formation of the annulus occurs at a site that is essential for overall growth of the envelope, it would be difficult to distinguish

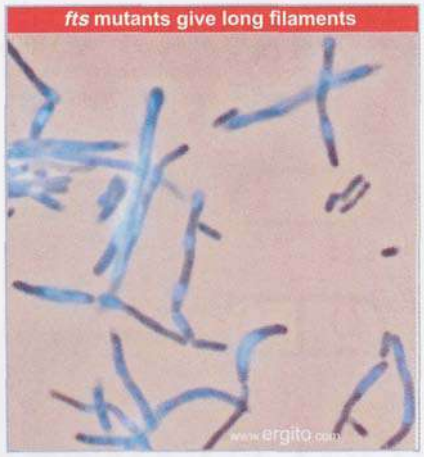


Figure 13.27 Failure of cell division generates multinucleated filaments. Photograph kindly provided by Sota Hiraga.

mutations that specifically interfere with annulus formation from those that inhibit envelope growth generally. Most mutations in the division apparatus have been identified as conditional mutants (whose division is affected under nonpermissive conditions; typically they are temperature sensitive). Mutations that affect cell division or chromosome segregation cause striking phenotypic changes. **Figure 13.27** and **Figure 13.28** illustrate the opposite consequences of failure in the division process and failure in segregation:

- *Long filaments* form when septum formation is inhibited, but chromosome replication is unaffected. The bacteria continue to grow, and even continue to segregate their daughter chromosomes, but septa do not form, so the cell consists of a very long filamentous structure, with the nucleoids (bacterial chromosomes) regularly distributed along the length of the cell. This phenotype is displayed by *fts* mutants (named for temperature-sensitive filamentation), which identify defect(s) that lie in the division process itself.
- Minicells form when septum formation occurs too frequently or in the wrong place, with the result that one of the new daughter cells lacks a chromosome. The minicell has a rather small size, and lacks DNA, but otherwise appears morphologically normal. Anucleate cells form when segregation is aberrant; like minicells, they lack a chromosome, but because septum formation is normal, their size is unaltered. This phenotype is caused by *par* (partition) mutants (named because they are defective in chromosome segregation).

13.17 FtsZ is necessary for septum formation

Key Concepts

- The product of *ftsZ* is required for septum formation at pre-existing sites.
- FtsZ is a GTPase that forms a ring on the inside of the bacterial envelope. It is connected to other cytoskeletal components.

Figure 13.28 *E. coli* generate anucleate cells when chromosome segregation fails. Cells with chromosomes stain blue; daughter cells lacking chromosomes have no blue stain. This field shows cells of the *mukB* mutant; both normal and abnormal divisions can be seen. Photograph kindly provided by Sota Hiraga.

The gene *ftsZ* plays a central role in division. Mutations in *ftsZ* block septum formation and generate filaments. Overexpression induces minicells, by causing an increased number of septation events per unit cell mass. *ftsZ* mutants act at stages varying from the displacement of the periseptal annuli to septal morphogenesis. FtsZ is therefore required for usage of pre-existing sites for septum formation, but does not itself affect the formation of the periseptal annuli or their localization.

FtsZ functions at an early stage of septum formation. Early in the division cycle, FtsZ is localized throughout the cytoplasm. As the cell elongates and begins to constrict in the middle, FtsZ becomes localized in a ring around the circumference. The structure is sometimes called the Z-ring. **Figure 13.29** shows that it lies in the position of the mid-center annulus of Figure 13.25. The formation of the Z-ring is the rate-limiting step in septum formation. In a typical division cycle, it forms in the center of cell 1 -5 min after division, remains for 15 min, and then quickly constricts to pinch the cell into two.

The structure of FtsZ resembles tubulin, suggesting that assembly of the ring could resemble the formation of microtubules in eukaryotic cells. FtsZ has GTPase activity, and GTP cleavage is used to support the oligomerization of FtsZ monomers into the ring structure. The Z-ring appears to be a dynamic structure, in which there is continuous exchange of subunits with a cytoplasmic pool.

Two other proteins needed for division, ZipA and FtsA, interact directly and independently with FtsZ. ZipA is an integral membrane protein, located in the inner bacterial membrane. It provides the means for linking FtsZ to the membrane. FtsA is a cytosolic protein, but is often found associated with the membrane. The Z-ring can form in the absence of either ZipA or FtsA, but cannot form if both are absent. This suggests that they have overlapping roles in stabilizing the Z-ring, and perhaps in linking it to the membrane.

The products of several other *fts* genes join the Z-ring in a defined order after FtsA has been incorporated. They are all transmembrane proteins. The final structure is sometimes called the septal ring. It consists of a multiprotein complex that is presumed to have the ability to constrict the membrane. One of the last components to be incorporated into the septal ring is FtsW, which is a protein belonging to the SEDS family *ftsW* is expressed as part of an operon with *ftsI*, which codes for a transpeptidase (also called PBP3 for penicillin-binding protein 3), a membrane-bound protein that has its catalytic site in the periplasm. FtsW is responsible for incorporating FtsI into the septal ring. This suggests a model for septum formation in which the transpeptidase activity then causes the peptidoglycan to grow inward, thus pushing the inner membrane and pulling the outer membrane.

FtsZ is the major cytoskeletal component of septation. It is common in bacteria, and is found also in chloroplasts. **Figure 13.30** shows the localization of the plant homologues to a ring at the mid-point of the chloroplast. Chloroplasts also have other genes related to the bacterial division genes. Consistent with the common evolutionary origins of bacteria and chloroplasts, the apparatus for division seems generally to have been conserved. Mitochondria, which also share an evolutionary origin with bacteria, usually do not have FtsZ. Instead, they use a variant of the protein dynamin, which is involved in pinching off vesicles from membranes of eukaryotic cytoplasm (see *27.5 Different types of coated vesicles exist in each pathway*). This functions from the outside of the organelle, squeezing the membrane to generate a constriction. The common feature, then, in the division of bacteria, chloroplasts, and mitochondria is the use of a cytoskeletal protein that forms a ring round the organelle, and either pulls or pushes the membrane to form a constriction.

Figure 13.29 Immunofluorescence with an antibody against FtsZ shows that it is localized at the mid-cell. Photograph kindly provided by William Margolin.

Figure 13.30 Immunofluorescence with antibodies against the *Arabidopsis* proteins FtsZ1 and FtsZ2 show that they are localized at the mid point of the chloroplast (top panel). The bright field image (lower panel) shows the outline of the chloroplast more clearly. Photograph kindly provided by Katherine Osteryoung.

13.18 *min* genes regulate the location of the septum

Key Concepts

- The location of the septum is controlled by *minC,D,E.*
- The number and location of septa is determined by the ratio of MinE/MinC,D.
- The septum forms where MinE is able to form a ring.
- At normal concentrations, MinC/D allows a mid-center ring, but prevents additional rings of MinE from forming at the poles.

Information about the localization of the septum is provided by minicell mutants. The original minicell mutation lies in the locus *minB*; deletion of *minB* generates minicells by allowing septation to occur at the poles as well as (or instead of) at midcell. This suggests that the cell possesses the ability to initiate septum formation either at midcell or at the poles; and the role of the wild-type *minB* locus is to suppress septation at the poles. In terms of the events depicted in Figure 13.25, this

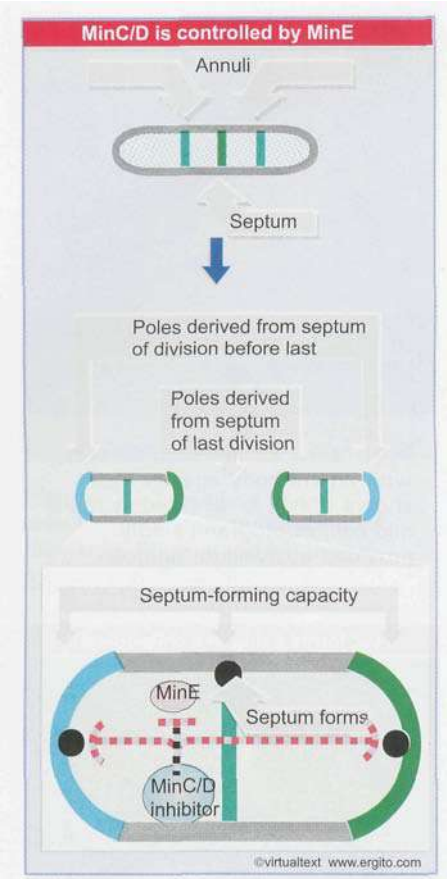


Figure 13.31 MinC/D is a division inhibitor, whose action is confined to the polar sites by MinE.

implies that a newborn cell has potential septation sites associated both with the annulus at mid-center and with the poles. One pole was formed from the septum of the previous division; the other pole represents the septum from the division before that. Perhaps the poles retain remnants of the annuli from which they were derived, and these remnants can nucleate septation.

The *minB* locus consists of three genes, *minC,D,E*. Their roles are summarized in **Figure 13.31.** The products of *minC* and *minD* form a division inhibitor. MinD is required to activate MinC, which prevents FtsZ from polymerizing into the Z-ring.

Expression of MinCD in the absence of MinE, or overexpression even in the presence of MinE, causes a generalized inhibition of division. The resulting cells grow as long filaments without septa. Expression of MinE at levels comparable to MinCD confines the inhibition to the polar regions, so restoring normal growth. MinE protects the midcell sites from inhibition. Overexpression of MinE induces minicells, because the presence of excess MinE counteracts the inhibition at the poles as well as at midcell, allowing septa to form at both locations.

The determinant of septation at the proper (midcell) site is therefore the ratio of MinCD to MinE. The wild-type level prevents polar septation, while permitting midcell septation. The effects of MinC/D and MinE are inversely related; absence of MinCD or too much MinE causes indiscriminate septation, forming minicells; too much MinCD or absence of MinE inhibits midcell as well as polar sites, resulting in filamentation.

MinE forms a ring at the septal position. Its accumulation suppresses the action of MinCD in the vicinity, thus allowing formation of the septal ring (which includes FtsZ and ZipA). Curiously, MinD is required for formation of the MinE ring.

13.19 Chromosomal segregation may require site-specific recombination

Key Concepts

- The Xer site-specific recombination system acts on a target sequence near the chromosome terminus to recreate monomers if a generalized recombination event has converted the bacterial chromosome to a dimer.

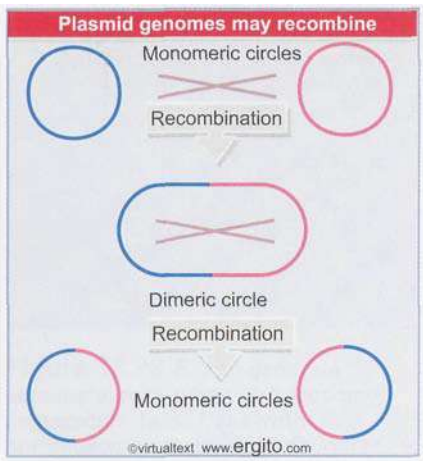


Figure 13.32 Intermolecular recombination merges monomers into dimers, and intramolecular recombination releases individual units from oligomers.

Because the multiple copies of a plasmid in a bacterium consist of the same DNA sequences, they are able to recombine. **Figure 13.32** demonstrates the consequences. A single intermolecular recombination event between two circles generates a dimeric circle; further recombination can generate higher multimeric forms. Such an event reduces the number of physically segregating units. In the extreme case of a single-copy plasmid that has just replicated, formation of a dimer by recombination means that the cell only has one unit to segregate, and the plasmid therefore must inevitably be lost from one daughter cell. To counteract this effect, plasmids often have site-specific recombination systems that act upon particular sequences to sponsor an intramolecular recombination that restores the monomeric condition.

The same types of event can occur with the bacterial chromosome, and **Figure 13.33** shows how they affect its segregation. If no recombination occurs, there is no problem, and the separate daughter chromosomes can segregate to the daughter cells. But a dimer will be produced if homologous recombination occurs between the daughter chromosomes produced

by a replication cycle. If there has been such a recombination event, the daughter chromosomes cannot separate. In this case, a second recombination is required to achieve resolution in the same way as a plasmid dimer.

Most bacteria with circular chromosomes posses the Xer site-specific recombination system. In *E. coli*, this consists of two recombinases, XerC and XerD, which act on a 28 bp target site, called *dif*, that is located in the terminus region of the chromosome. The use of the Xer system is related in an interesting way to cell division. The relevant events are summarized in **Figure 13.34.** XerC can bind to a pair of *dif* sequences and form a Holliday junction between them. The complex may form soon after the replication fork passes over the *dif* sequence, which explains how the two copies of the target sequence can find one another consistently. However, resolution of the junction to give recombinants occurs only in the presence of FtsK, a protein located in the septum that is required for chromosome segregation and cell division. Also, the *dif* target sequence must be located in a region of ~30 kb; if it is moved outside of this region, it cannot support the reaction.

So there is a site-specific recombination available when the terminus sequence of the chromosome is close to the septum. But the bacterium wants to have a recombination only when there has already been a general recombination event to generate a dimer. (Otherwise the site-specific recombination would create the dimer!) How does the system know whether the daughter chromosomes exist as independent monomers or have been recombined into a dimer?

The answer may be that segregation of chromosomes starts soon after replication. If there has been no recombination, the two chromosomes move apart from one another. But the ability of the relevant sequences to move apart from one another may be constrained if a dimer has been formed. This forces them to remain in the vicinity of the septum, where they are exposed to the Xer system.

Bacteria that have the Xer system always have an FtsK homolog, and vice-versa, which suggests that the system has evolved so that resolution is connected to the septum. FtsK is a large transmembrane protein. Its N-terminal domain is associated with the membrane, and causes it to be localized to the septum. Its C-terminal domain has two functions. One is to cause Xer to resolve a dimer into two monomers. It also has an ATPase activity, which it can use to translocate along DNA *in vitro*. This could be used to pump DNA through the septum (in the same way that SpoIIIE transports DNA from the mother compartment into the prespore during sporulation (see next section).

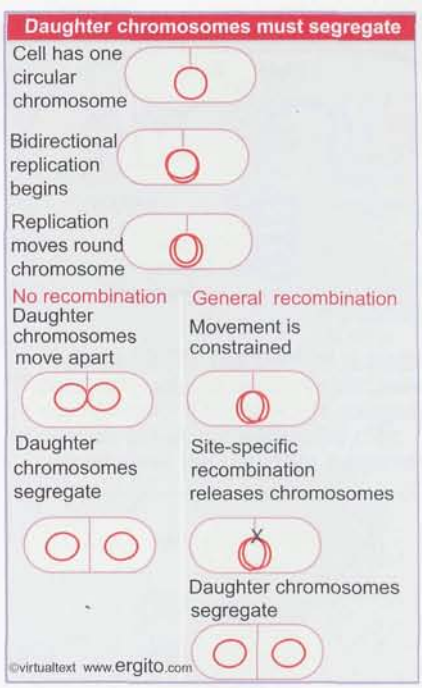


Figure 13.33 A circular chromosome replicates to produce two monomers that segregate to daughter cells. If a dimer is generated by generalized recombination, it can be resolved into two monomers by a site-specific recombination.

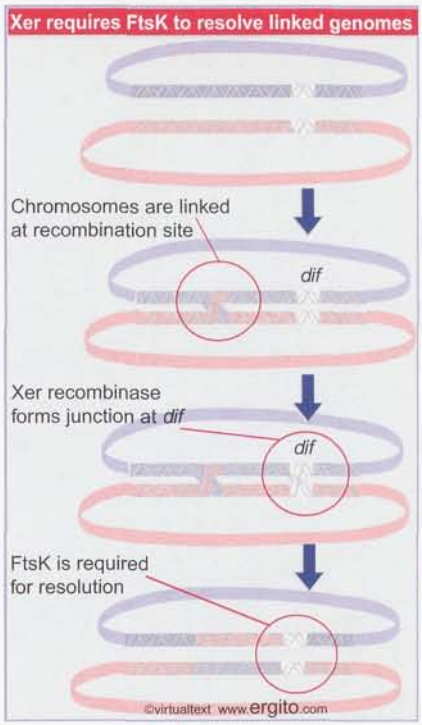


Figure 13.34 A recombination event creates two linked chromosomes. Xer creates a Holliday junction at the *dif* site, but can resolve it only in the presence of FtsK.

13.20 Partitioning involves separation of the chromosomes

Key Concepts

- Replicon origins may be attached to the inner bacterial membrane.
- Chromosomes make abrupt movements from the mid-center to the $^1/_4$ and $^3/_4$ positions.
- Movement may be connected with condensation following replication.

Partitioning is the process by which the two daughter chromosomes find themselves on either side of the position at which the septum forms. Two types of event are required for proper partitioning:

- The two daughter chromosomes must be released from one another so that they can segregate following termination. This requires

disentangling of DNA regions that are coiled around each other in the vicinity of the terminus. Most mutations affecting partitioning map in genes coding for topoisomerases-enzymes with the ability to pass DNA strands through one another. The mutations prevent the daughter chromosomes from segregating, with the result that the DNA is located in a single large mass at midcell. Septum formation then releases an anucleate cell and a cell containing both daughter chromosomes. This tells us that the bacterium must be able to disentangle its chromosomes topologically in order to be able to segregate them into different daughter cells.

- Mutations that affect the partition process itself are rare. We expect to find two classes. *cis-acting* mutations should occur in DNA sequences that are the targets for the partition process. *trans*-acting mutations should occur in genes that code for the protein(s) that cause segregation, which could include proteins that bind to DNA or activities that control the locations on the envelope to which DNA might be attached. Both types of mutation have been found in the systems responsible for partitioning plasmids but only *trans-acting* functions have been found in the bacterial chromosome. In addition, mutations in plasmid site-specific recombination systems increase plasmid loss (because the dividing cell has only one dimer to partition instead of two monomers), and therefore have a phenotype that is similar to partition mutants.

The original form of the model for chromosome segregation shown in Figure 13.26 suggested that the envelope grows by insertion of material between the attachment sites of the two chromosomes, thus pushing them apart. But in fact the cell wall and membrane grow heterogeneously over the whole cell surface. Furthermore, the replicated chromosomes are capable of abrupt movements to their final positions at $^1/_4$ and $^3/_4$ cell length. If protein synthesis is inhibited before the termination of replication, the chromosomes fail to segregate and remain close to the midcell position. But when protein synthesis is allowed to resume, the chromosomes move to the quarter positions in the absence of any further envelope elongation. This suggests that an active process, requiring protein synthesis, may move the chromosomes to specific locations.

Segregation is interrupted by mutations of the *muk* class, which give rise to anucleate progeny at a much increased frequency: both daughter chromosomes remain on the same side of the septum instead of segregating. Mutations in the *muk* genes are not lethal, and may identify components of the apparatus that segregates the chromosomes. The gene *mukA* is identical with the gene for a known outer membrane protein (*tolC*), whose product could be involved with attaching the chromosome to the envelope. The gene *mukB* codes for a large (180 kD) globular protein, which has the same general type of organization as the two groups of SMC proteins that are involved in condensing and in holding together eukaryotic chromosomes (see *23.18 Chromosome condensation is caused by condensins).* SMC-like proteins have also been found in other bacteria. (MukB also has some sequence relationship to the mechanochemical enzyme dynamin, which provides a "motor" for microtubule-associated objects, but it is now thought that this is not significant.)

The insight into the role of MukB was the discovery that some mutations in *mukB* can be suppressed by mutations in *topA*, the gene coding for topoisomerase I. This led to the model that the function of MukBEF proteins is to condense the nucleoid. A defect in this function is the cause of failure to segregate properly. The defect can be compensated by preventing topoisomerases from relaxing negative supercoils; the resulting increase in supercoil density helps to restore the proper state of condensation and thus to allow segregation.

By Book_Crazy [IND]

We still do not understand how genomes are positioned in the cell, but the process may be connected with condensation. **Figure 13.35** shows a current model. The parental genome is centrally positioned. It must be decondensed in order to pass through the replication apparatus. The daughter chromosomes emerge from replication, are disentangled by topoisomerases, and then passed in an uncondensed state to MukBEF, which causes them to form condensed masses at the positions that will become the centers of the daughter cells.

There have been suspicions for years that a physical link exists between bacterial DNA and the membrane, but the evidence remains indirect. Bacterial DNA can be found in membrane fractions, which tend to be enriched in genetic markers near the origin, the replication fork, and the terminus. The proteins present in these membrane fractions may be affected by mutations that interfere with the initiation of replication. The growth site could be a structure on the membrane to which the origin must be attached for initiation.

During sporulation in *B. subtilis*, one daughter chromosome must be segregated into the small forespore compartment (see Figure 9.42). This is an unusual process that involves transfer of the chromosome across the nascent septum. One of the sporulation genes, *spoIIIE*, is required for this process. The SpoIIIE protein is located at the septum and is probably has a translocation function that pumps DNA through.

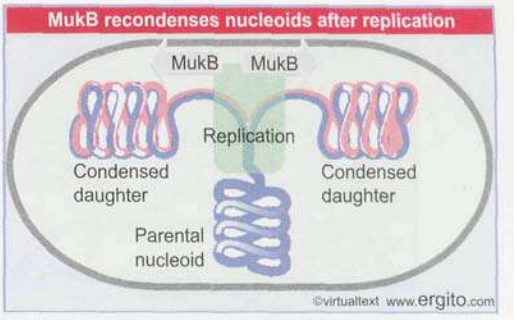


Figure 13.35 The DNA of a single parental nucleoid becomes decondensed during replication. MukB is an essential component of the apparatus that recondenses the daughter nucleoids.

13.21 Single-copy plasmids have a partitioning system

Key Concepts

- Single copy plasmids exist at one plasmid copy per bacterial chromosome origin.
- Multicopy plasmids exist at > 1 plasmid copy per bacterial chromosome origin.
- Homologous recombination between circular plasmids generates dimers and higher multimers.
- Plasmids have site-specific recombination systems that undertake intramolecular recombination to regenerate monomers.
- Partition systems ensure that duplicate plasmids are segregated to different daughter cells produced by a division.

The type of system that a plasmid uses to ensure that it is distributed to both daughter cells at division depends upon its type of replication system. Each type of plasmid is maintained in its bacterial host at a characteristic copy number:

- Single-copy control systems resemble that of the bacterial chromosome and result in one replication per cell division. A single-copy plasmid effectively maintains parity with the bacterial chromosome.
- Multicopy control systems allow multiple initiation events per cell cycle, with the result that there are several copies of the plasmid per bacterium. Multicopy plasmids exist in a characteristic number (typically 10-20) per bacterial chromosome.

Copy number is primarily a consequence of the type of replication control mechanism. The system responsible for initiating replication determines how many origins can be present in the bacterium. Since each plasmid consists of a single replicon, the number of origins is the same as the number of plasmid molecules.

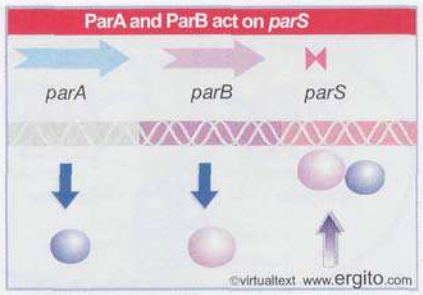


Figure 13.36 A common segregation system consists of genes *parA* and *parB* and the target site *parS*.

Single-copy plasmids have a system for replication control whose consequences are similar to that governing the bacterial chromosome. A single origin can be replicated once; then the daughter origins are segregated to the different daughter cells.

Multicopy plasmids have a replication system that allows a pool of origins to exist. If the number is great enough (in practice > 10 per bacterium), an active segregation system becomes unnecessary, because even a statistical distribution of plasmids to daughter cells will result in the loss of plasmids at frequencies $<10^{-6}$.

Plasmids are maintained in bacterial populations with very low rates of loss ($<10^{-7}$ per cell division is typical, even for a single-copy plasmid). The systems that control plasmid segregation can be identified by mutations that increase the frequency of loss, but that do not act upon replication itself. Several types of mechanism are used to ensure the survival of a plasmid in a bacterial population. It is common for a plasmid to carry several systems, often of different types, all acting independently to ensure its survival. Some of these systems act indirectly, while others are concerned directly with regulating the partition event. However, in terms of evolution, all serve the same purpose: to help ensure perpetuation of the plasmid to the maximum number of progeny bacteria.

Single-copy plasmids require partitioning systems to ensure that the duplicate copies find themselves on opposite sides of the septum at cell division, and are therefore segregated to a different daughter cell. In fact, functions involved in partitioning were first identified in plasmids. The components of a common system are summarized in **Figure 13.36.** Typically there are two *trans-acting* loci (*parA* and *parB)* and a *cis-acting* element *(parS)* located just downstream of the two genes. ParA is an ATPase. It binds to ParB, which binds to the *parS* site on DNA. Deletions of any of the three loci prevent proper partition of the plasmid. Systems of this type have been characterized for the plasmids F, P1, and R1. In spite of their overall similarities, there are no significant sequence homologies between the corresponding genes or *cis*-acting sites.

parS plays a role for the plasmid that is equivalent to the centromere in a eukaryotic cell. Binding of the ParB protein to it creates a structure that segregates the plasmid copies to opposite daughter cells. A bacterial protein, IHF, also binds at this site to form part of the structure. The complex of ParB and IHF with *parS* is called the partition complex. *parS* is a 34 bp sequence containing the IHF-binding site flanked on either side by sequences called *boxA* and *boxB* that are bound by ParB.

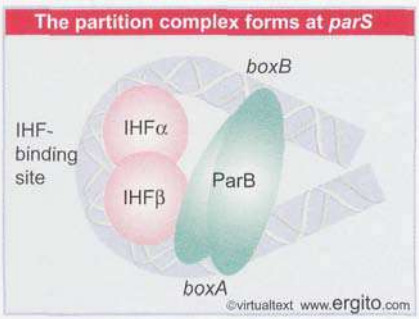


Figure 13.37 The partition complex is formed when IHF binds to DNA at *parS* and bends it so that ParB can bind to sites on either side. The complex is initiated by a heterodimer of IHF and a dimer of ParB, and then more ParB dimers bind.

IHF is the integration host factor, named for the role in which it was first discovered (forming a structure that is involved in the integration of phage lambda DNA into the host chromosome). IHF is a heterodimer with has the capacity to form a large structure in which DNA is wrapped on the surface. The role of IHF is to bend the DNA so that ParB can bind simultaneously to the separated *boxA* and *boxB* sites, as indicated in **Figure 13.37.** Complex formation is initiated when *parS* is bound by a heterodimer of IHF together with a dimer of ParB. This enables further dimers of ParB to bind cooperatively. The interaction of ParA with the partition complex structure is essential but transient.

The protein-DNA complex that assembles on IHF during phage lambda integration binds two DNA molecules to enable them to recombine (see *15.19 Lambda recombination occurs in an intasome).* The role of the partition complex is different: to ensure that two DNA molecules segregate apart from one another. We do not know yet how the formation of the individual complex accomplishes this task. One possibility is that it attaches the DNA to some physical site—for example, on

the membrane—and then the sites of attachment are segregated by growth of the septum.

Proteins related to ParA and ParB are found in several bacteria. In *B. subtilis*, they are called Soj and SpoOJ, respectively. Mutations in these loci prevent sporulation, because of a failure to segregate one daughter chromosome into the forespore (see Figure 9.43). In sporulating cells, SpoOJ localizes at the pole and may be responsible for localizing the origin there. SpoOJ binds to a sequence that is present in multiple copies, dispersed over ~20% of the chromosome in the vicinity of the origin. It is possible that SpoOJ binds both old and newly synthesized origins, maintaining a status equivalent to chromosome pairing, until the chromosomes are segregated to the opposite poles. In *C. crescentus*, ParA and ParB localize to the poles of the bacterium, and ParB binds sequences close to the origin, thus localizing the origin to the pole. These results suggest that a specific apparatus is responsible for localizing the origin to the pole. The next stage of the analysis will be to identify the cellular components with which this apparatus interacts.

The importance to the plasmid of ensuring that all daughter cells gain replica plasmids is emphasized by the existence of multiple, independent systems in individual plasmids that ensure proper partition. Addiction systems, operating on the basis that "we hang together or we hang separately," ensure that a bacterium carrying a plasmid can survive only so long as it retains the plasmid. There are several ways to ensure that a cell dies if it is "cured" of a plasmid, all sharing the principle illustrated in **Figure 13.38** that the plasmid produces both a poison and an antidote. The poison is a killer substance that is relatively stable, whereas the antidote consists of a substance that blocks killer action, but is relatively short lived. When the plasmid is lost, the antidote decays, and then the killer substance causes death of the cell. So bacteria that lose the plasmid inevitably die, and the population is condemned to retain the plasmid indefinitely. These systems take various forms. One specified by the F plasmid consists of killer and blocking proteins. The plasmid R1 has a killer that is the mRNA for a toxic protein, while the antidote is a small antisense RNA that prevents expression of the mRNA.

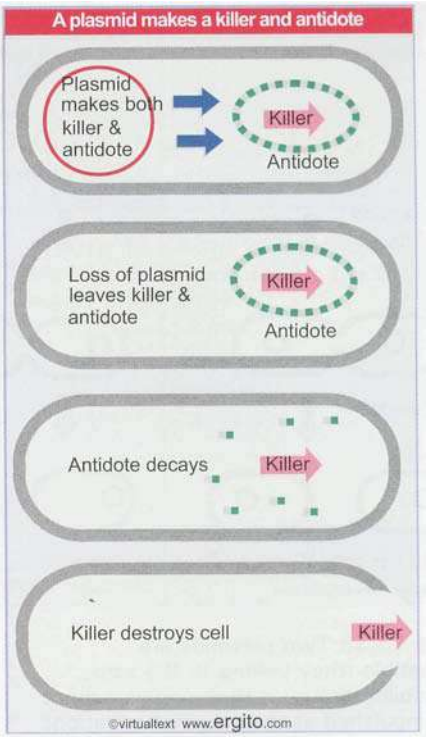


Figure 13.38 Plasmids may ensure that bacteria cannot live without them by synthesizing a long-lived killer and a short-lived antidote.

13.22 Plasmid incompatibility is determined by the replicon

Key Concepts

- Plasmids in a single compatibility group have origins that are regulated by a common control system.

The phenomenon of plasmid incompatibility is related to the regulation of plasmid copy number and segregation. A compatibility group is defined as a set of plasmids whose members are unable to coexist in the same bacterial cell. The reason for their incompatibility is that they cannot be distinguished from one another at some stage that is essential for plasmid maintenance. DNA replication and segregation are stages at which this may apply.

The negative control model for plasmid incompatibility follows the idea that copy number control is achieved by synthesizing a repressor that measures the concentration of origins. (Formally this is the same as the titration model for regulating replication of the bacterial chromosome.)

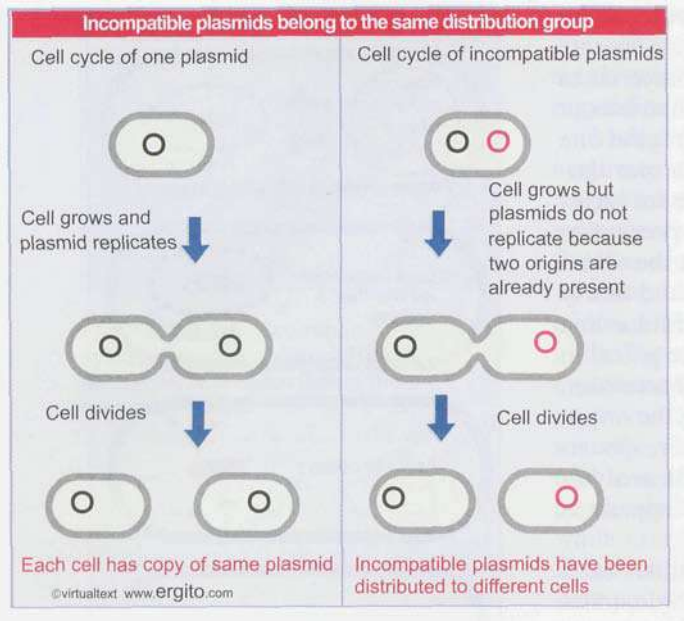


Figure 13.39 Two plasmids are incompatible (they belong to the same compatibility group) if their origins cannot be distinguished at the stage of initiation. The same model could apply to segregation.

The introduction of a new origin in the form of a second plasmid of the same compatibility group mimics the result of replication of the resident plasmid; two origins now are present. So any further replication is prevented until after the two plasmids have been segregated to different cells to create the correct prereplication copy number as illustrated in **Figure 13.39.**

A similar effect would be produced if the system for segregating the products to daughter cells could not distinguish between two plasmids. For example, if two plasmids have the same *cis*-acting partition sites, competition between them would ensure that they would be segregated to different cells, and therefore could not survive in the same line.

The presence of a member of one compatibility group does not directly affect the survival of a plasmid belonging to a different group. Only one replicon of a given compatibility group (of a single-copy plasmid) can be maintained in the bacterium, but it does not interact with replicons of other compatibility groups.

13.23 The ColE1 compatibility system is controlled by an RNA regulator

Key Concepts

- Replication of ColEI requires transcription to pass through the origin, where the transcript is cleaved by RNAaseH to generate a primer end.
- The regulator RNA I is a short antisense RNA that pairs with the transcript and prevents the cleavage that generates the priming end.
- The Rom protein enhances pairing between RNA I and the transcript.

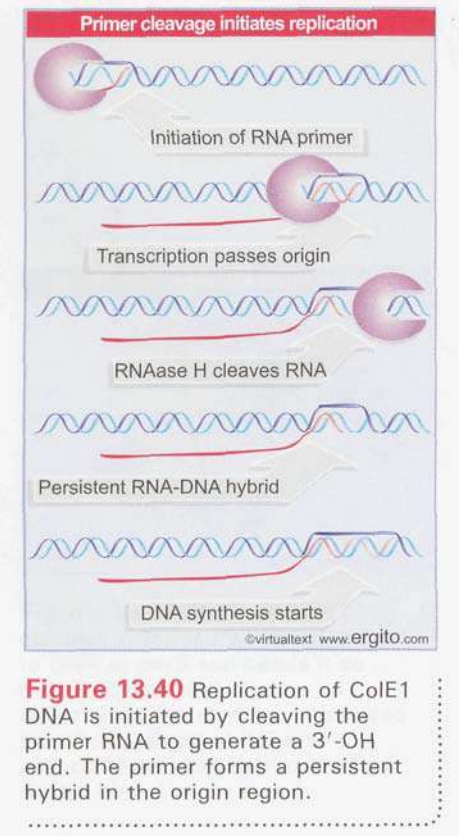


Figure 13.40 Replication of ColE1 DNA is initiated by cleaving the primer RNA to generate a 3′-OH end. The primer forms a persistent hybrid in the origin region.

The best characterized copy number and incompatibility system is that of the plasmid ColEI, a multicopy plasmid that is maintained at a steady level of ~20 copies per *E. coli* cell. The system for maintaining the copy number depends on the mechanism for initiating replication at the ColEI origin, as illustrated in **Figure 13.40.**

Replication starts with the transcription of an RNA that initiates 555 bp upstream of the origin. Transcription continues through the origin. The enzyme RNAase H (whose name reflects its specificity for a substrate of RNA hybridized with DNA) cleaves the transcript at the origin. This generates a 3′-OH end that is used as the "primer" at which DNA synthesis is initiated (the use of primers is discussed in more detail in *14.8 Priming is required to start DNA synthesis).* The primer RNA forms a persistent hybrid with the DNA. Pairing between the RNA and DNA occurs just upstream of the origin (around position -20) and also farther upstream (around position -265).

Two regulatory systems exert their effects on the RNA primer. One involves synthesis of an RNA complementary to the primer; the other involves a protein coded by a nearby locus.

The regulatory species RNA I is a molecule of ~ 108 bases, coded by the opposite strand from that specifying primer RNA. The relationship between the primer RNA and RNA I is illustrated in **Figure 13.41.** The RNA I molecule is initiated within the primer region and terminates

close to the site where the primer RNA initiates. So RNA I is complementary to the 5′-terminal region of the primer RNA. Base pairing between the two RNAs controls the availability of the primer RNA to initiate a cycle of replication.

An RNA molecule such as RNA I that functions by virtue of its complementarity with another RNA coded in the same region is called a countertranscript. This type of mechanism, of course, is another example of the use of antisense RNA (see *11.19 Small RNA molecules can regulate translation).*

Mutations that reduce or eliminate incompatibility between plasmids can be obtained by selecting plasmids of the same group for their ability to coexist. Incompatibility mutations in ColE1 map in the region of overlap between RNA I and primer RNA. Because this region is represented in two different RNAs, either or both might be involved in the effect.

When RNA I is added to a system for replicating ColEl DNA *in vitro,* it inhibits the formation of active primer RNA. But the presence of RNA I does not inhibit the initiation or elongation of primer RNA synthesis. This suggests that RNA I prevents RNAase H from generating the 3' end of the primer RNA. The basis for this effect lies in base pairing between RNA I and primer RNA.

Both RNA molecules have the same potential secondary structure in this region, with three duplex hairpins terminating in single-stranded loops. Mutations reducing incompatibility are located in these loops, which suggests that the initial step in base pairing between RNA I and primer RNA is contact between the unpaired loops.

How does pairing with RNA I prevent cleavage to form primer RNA? A model is illustrated in **Figure 13.42.** In the absence of RNA I, the primer RNA forms its own secondary structure (involving loops and stems). But when RNA I is present, the two molecules pair, and become completely double-stranded for the entire length of RNA I. The new secondary structure prevents the formation of the primer, probably by affecting the ability of the RNA to form the persistent hybrid.

The model resembles the mechanism involved in attenuation of transcription, in which the alternative pairings of an RNA sequence permit or prevent formation of the secondary structure needed for termination by RNA polymerase (see *11.14 Alternative secondary structures control attenuation).* The action of RNA I is exercised by its ability to affect distant regions of the primer precursor.

Formally, the model is equivalent to postulating a control circuit involving two RNA species. A large RNA primer precursor is a positive regulator, needed to initiate replication. The small RNA I is a negative regulator, able to inhibit the action of the positive regulator.

In its ability to act on any plasmid present in the cell, RNA I provides a repressor that prevents newly introduced DNA from functioning, analogous to the role of the lambda lysogenic repressor (see *12.10 The repressor and its operators define the immunity region).* Instead of a repressor protein that binds the new DNA, an RNA binds the newly synthesized precursor to the RNA primer.

Binding between RNA I and primer RNA can be influenced by the Rom protein, coded by a gene located downstream of the origin. Rom enhances binding between RNA I and primer RNA transcripts of >200 bases. The result is to inhibit formation of the primer.

How do mutations in the RNAs affect incompatibility? **Figure 13.43** shows the situation when a cell contains two types of RNA I/primer RNA sequence. The RNA I and primer RNA made from each type of genome can interact, but RNA I from one genome does not interact with primer RNA from the other genome. This situation would arise when a mutation in the region that is common to RNA I and

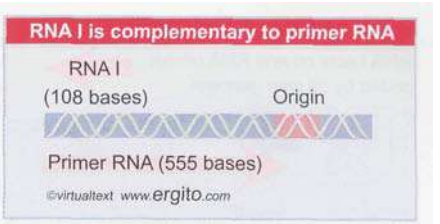


Figure 13.41 The sequence of RNA I is complementary to the 5′ region of primer RNA.

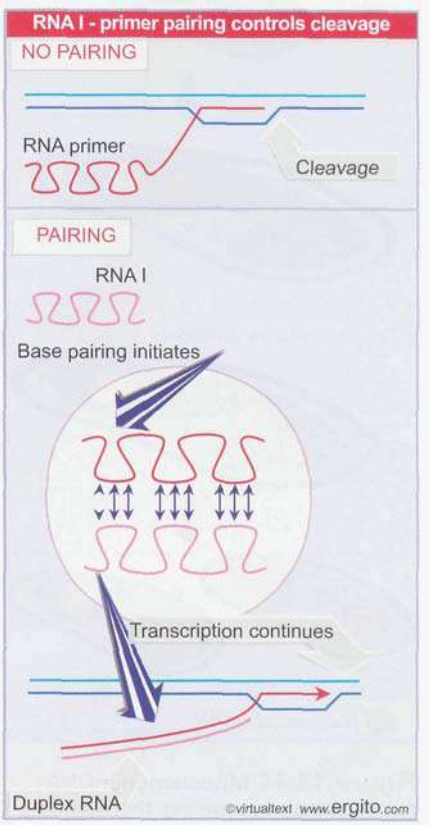


Figure 13.42 Base pairing with RNA I may change the secondary structure of the primer RNA sequence and thus prevent cleavage from generating a 3′-OH end.

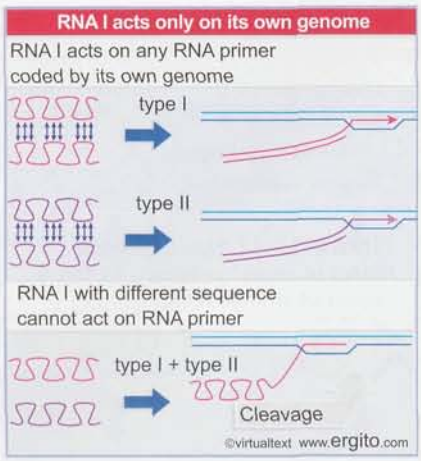


Figure 13.43 Mutations in the region coding for RNA I and the primer precursor need not affect their ability to pair; but they may prevent pairing with the complementary RNA coded by a different plasmid.

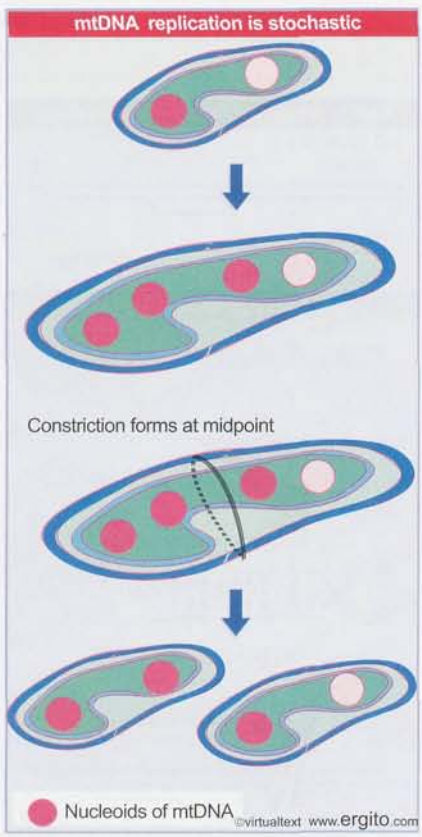


Figure 13.44 Mitochondrial DNA replicates by increasing the number of genomes in proportion to mitochondrial mass, but without ensuring that each genome replicates the same number of times. This can lead to changes in the representation of alleles in the daughter mitochondria.

primer RNA occurred at a location that is involved in the base pairing between them. Each RNA I would continue to pair with the primer RNA coded by the same plasmid, but might be unable to pair with the primer RNA coded by the other plasmid. This would cause the original and the mutant plasmids to behave as members of different compatibility groups.

13.24 How do mitochondria replicate and segregate?

Key Concepts

- mtDNA replication and segregation to daughter mitochondria is stochastic.
- Mitochondrial segregation to daughter cells is also stochastic.

Mitochondria must be duplicated during the cell cycle and segregated to the daughter cells. We understand some of the mechanics of this process, but not its regulation.

At each stage in the duplication of mitochondria—DNA replication, DNA segregation to duplicate mitochondria, organelle segregation to daughter cells—the process appears to be stochastic, governed by a random distribution of each copy. The theory of distribution in this case is analogous that of multicopy bacterial plasmids, with the same conclusion that > 10 copies are required to ensure that each daughter gains at least one copy (see *13.21 Single-copy plasmids have a partitioning system).* When there are mtDNAs with allelic variations (either because of inheritance from different parents or because of mutation), the stochastic distribution may generate cells that have only one of the alleles.

In some situations a mitochondrion has both paternal and maternal alleles. This has two requirements: that both parents provide alleles to the zygote (which of course is not the case when there is maternal inheritance; see *3.18 Organelles have DNA*); and that the parental alleles are found in the same mitochondrion. For this to happen, parental mitochondria must have fused.

The size of the individual mitochondrion may not be precisely defined. Indeed, there is a continuing question as to whether an individual mitochondrion represents a unique and discrete copy of the organelle or whether it is in a dynamic flux in which it can fuse with other mitochondria. We know that mitochondria can fuse in yeast, because recombination between mtDNAs can occur after two haploid yeast strains have mated to produce a diploid strain. This implies that the two mtDNAs must have been exposed to one another in the same mitochondrial compartment. Attempts have been made to test for the occurrence of similar events in animal cells by looking for complementation between alleles after two cells have been fused, but the results are not clear.

13.25 Summary

The entire chromosome is replicated once for every cell division cycle. Initiation of replication commits the cell to a cycle of division; completion of replication may provide a trigger for the actual division process. The bacterial chromosome consists of a single replicon, but a eukaryotic chromosome is divided into many replicons that function over the protracted period of S phase. The problem of replicating the ends of a linear replicon is solved in a variety of ways, most often by converting the replicon to a circular form. Some viruses have special proteins that recognize ends. Eukaryotic chromosomes encounter the problem at their terminal replicons.

Eukaryotic replication is (at least) an order of magnitude slower than bacterial replication. Origins sponsor bidirectional replication, and are probably used in a fixed order during S phase. The only eukaryotic origins identified at the sequence level are those of *S. cerevisiae,* which have a core consensus sequence consisting of 11 base pairs, mostly AT.

The minimal *E. coli* origin consists of ~245 bp and initiates bidirectional replication. Any DNA molecule with this sequence can replicate in *E. coli.* Two replication forks leave the origin and move around the chromosome, apparently until they meet, although *ter* sequences that would cause the forks to terminate after meeting have been identified. Transcription units are organized so that transcription usually proceeds in the same direction as replication.

The rolling circle is an alternative form of replication for circular DNA molecules in which an origin is nicked to provide a priming end. One strand of DNA is synthesized from this end, displacing the original partner strand, which is extruded as a tail. Multiple genomes can be produced by continuing revolutions of the circle.

Rolling circles are used to replicate some phages. The A protein that nicks the ϕX174 origin has the unusual property of *cis*-action. It acts only on the DNA from which it was synthesized. It remains attached to the displaced strand until an entire strand has been synthesized, and then nicks the origin again, releasing the displaced strand and starting another cycle of replication.

Rolling circles also are involved in bacterial conjugation, when an F plasmid is transferred from a donor to a recipient cell, following the initiation of contact between the cells by means of the F-pili. A free F plasmid infects new cells by this means; an integrated F factor creates an Hfr strain that may transfer chromosomal DNA. In the case of conjugation, replication is used to synthesize complements to the single strand remaining in the donor and to the single strand transferred to the recipient, but does not provide the motive power.

A fixed time of 40 minutes is required to replicate the *E. coli* chromosome and a further 20 minutes is required before the cell can divide. When cells divide more rapidly than every 60 minutes, a replication cycle is initiated before the end of the preceding division cycle. This generates multiforked chromosomes. The initiation event depends on titration of cell mass, probably by accumulating an initiator protein. Initiation may occur at the cell membrane, since the origin is associated with the membrane for a short period after initiation.

The septum that divides the cell grows at a location defined by the pre-existing periseptal annulus; a locus of three genes (*minCDE*) codes for products that regulate whether the midcell periseptal annulus or the polar sites derived from previous annuli are used for septum formation. Absence of septum formation generates multinucleated filaments; excess of septum formation generates anucleate minicells.

Many transmembrane proteins interact to form the septum. ZipA is located in the inner bacterial membrane and binds to FtsZ, which is a tubulin-like protein that can polymerize into a filamentous structure called a Z-ring. FtsA is a cytosolic protein that binds to FtsZ. Several other *fts* products, all transmembrane proteins, join the Z-ring in an ordered process that generates a septal ring. The last proteins to bind are the SEDS protein FtsW and the transpeptidase ftsI (PBP3), which together function to produce the peptidoglycans of the septum. Chloroplasts use a related division mechanism that has an FtsZ-like protein, but mitochondria use a different process in which the membrane is constricted by a dynamin-like protein.

Plasmids and bacteria have site-specific recombination systems that regenerate pairs of monomers by resolving dimers created by general recombination. The Xer system acts on a target sequence located in the terminus region of the chromosome. The system is active only in the presence of the FtsK protein of the septum, which may ensure that it acts only when a dimer needs to be resolved.

Partitioning involves the interaction of the ParB protein with the *parS* targetsite to build a structure that includes the IHF protein. This partition complex ensures that replica chromosomes segregate into different daughter cells. The mechanism of segregation may involve movement of DNA, possibly by the action of MukB in condensing chromosomes into masses at different locations as they emerge from replication.

Plasmids have a variety of systems that ensure or assist partition, and an individual plasmid may carry systems of several types. The copy number of a plasmid describes whether it is present at the same level as the bacterial chromosome (one per unit cell) or in greater numbers. Plasmid incompatibility can be a consequence of the mechanisms involved in either replication or partition (for single-copy plasmids). Two plasmids that share the same control system for replication are incompatible because the number of replication events ensures that there is only one plasmid for each bacterial genome.

References

13.1 Introduction

ref Jacob, F., Brenner, S., and Cuzin, F. (1963). On the regulation of DNA replication in bacteria. Cold Spring Harbor Symp. Quant. Biol. 28, 329-348.

13.3 Origins can be mapped by autoradiography and electrophoresis

ref Huberman, J. and Riggs, A. D. (1968). On the mechanism of DNA replication in mammalian chromosomes. J. Mol. Biol. 32, 327-341.

13.4 The bacterial genome is a single circular replicon

rev Brewer, B. J. (1988). When polymerases collide: replication and transcriptional organization of the *E. coli* chromosome. Cell 53, 679-686.

ref Cairns, J. (1963). The bacterial chromosome and its manner of replication as seen by autoradiography. J. Mol. Biol. 6, 208-213.

lismaa, T. P. and Wake, R. G. (1987). The normal replication terminus of the *B. subtilis* chromosome, *terC,* is dispensable for vegetative growth and sporulation. J. Mol. Biol. 195, 299-310.

Liu, B., Wong, M. L., and Alberts, B. (1994). A transcribing RNA polymerase molecule survives DNA replication without aborting its growing RNA chain. Proc. Nat. Acad. Sci. USA 91, 10660-10664.

Steck, T. R. and Drlica, K. (1984). Bacterial chromosome segregation: evidence for DNA gyrase involvement in decatenation. Cell 36, 1081-1088.

By Book_Crazy [IND]

Zyskind, J. W. and Smith, D. W. (1980). Nucleotide sequence of the *S. typhimurium* origin of DNA replication. Proc. Nat. Acad. Sci. USA 77, 2460-2464.

13.5 Each eukaryotic chromosome contains many replicons

rev Fangman, W. L. and Brewer, B. J. (1991). Activation of replication origins within yeast chromosomes. Ann. Rev. Cell Biol. 7, 375-402.

ref Blumenthal, A. B., Kriegstein, H. J., and Hogness, D. S. (1974). The units of DNA replication in *D. melanogaster* chromosomes. Cold Spring Harbor Symp. Quant. Biol. 38, 205-223.

13.6 Replication origins can be isolated in yeast

rev Bell, S. P. and Dutta, A. (2002). DNA replication in eukaryotic cells. Ann. Rev. Biochem. 71, 333-374.

DePamphlis, M. L. (1993). Eukaryotic DNA replication: anatomy of an origin. Ann. Rev. Biochem. 62, 29-63.

ref Chesnokov, I., Remus, D., and Botchan, M. (2001). Functional analysis of mutant and wild-type *Drosophila* origin recognition complex. Proc. Nat. Acad. Sci. USA 98, 1 1997-12002.

Kelly, T. J. and Brown, G. W. (2000). Regulation of chromosome replication. Ann. Rev. Biochem. 69, 829-880.

Marahrens, Y. and Stillman, B. (1992). A yeast chromosomal origin of DNA replication defined by multiple functional elements. Science 255, 817-823.

Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P., and Aparicio, O. M. (2001). Genome-Wide Distribution of ORC and MCM Proteins in *S. cerevisiae:* High-Resolution Mapping of Replication Origins. Science 294, 2357-2360.

13.7 D loops maintain mitochondrial origins

rev Clayton, D., A. (1991). Replication and transcription of vertebrate mitochondrial DNA. Ann. Rev. Cell Biol. 7, 453-478.

Clayton, D. (1982). Replication of animal mitochondrial DNA. Cell 28, 693-705.

Shadel, G. S. and Clayton, D. A. (1997). Mitochondrial DNA maintenance in vertebrates. Ann. Rev. Biochem. 66, 409-435.

13.10 Rolling circles produce multimers of a replicon

ref Gilbert, W. and Dressier, D. (1968). DNA replication: the rolling circle model. Cold Spring Harbor Symp. Quant. Biol. 33, 473-484.

13.12 The F plasmid is transferred by conjugation between bacteria

ref Ihler, G. and Rupp, W. D. (1969). Strand-specific transfer of donor DNA during conjugation in *E. coli.* Proc. Nat. Acad. Sci. USA 63, 138-143.

13.13 Conjugation transfers single-stranded DNA

rev Frost, L. S., Ippen-Ihler, K., and Skurray, R. A. (1994). Analysis of the sequence and gene products of the transfer region of the F sex factor. Microbiol. Rev. 58, 162-210.

Ippen-Ihler, K. A. and Minkley, E. G. (1986). The conjugation system of F, the fertility factor of *E. coli.* Ann. Rev. Genet. 20, 593-624.

Lanka, E. and Wilkins, B. M. (1995). DNA processing reactions in bacterial conjugation. Ann. Rev. Biochem. 64, 141-169.

Willetts, N. and Skurray, R. (1987). Structure and function of the F factor and mechanism of conjugation. In *E. coli and S. typhimurium:* Cellular and Molecular Biology, Ed. F. C. Neidhardt, American Society for Microbiology, Washington DC 1110-1131.

13.14 Replication is connected to the cell cycle

rev Donachie, W. D. (1993). The cell cycle of *E. coli.* Ann. Rev. Immunol. 47, 199-230.

ref Donachie, W. D. and Begg, K. J. (1970). Growth of the bacterial cell. Nature 227, 1220-1224.

Donachie, W. D., Begg, K. J., and Vicente, M. (1976). Cell length, cell growth and cell division. Nature 264, 328-333.

13.15 The septum divides a bacterium into progeny each containing a chromosome

rev de Boer, P. A. J., Cook, W. R., and Rothfield, L. I. (1990). Bacterial cell division. Ann. Rev. Genet. 24, 249-274.

ref Spratt, B. G. (1975). Distinct penicillin binding proteins involved in the division, elongation, and shape of *E. coli* K12. Proc. Nat. Acad. Sci. USA 72, 2999-3003.

13.16 Mutations in division or segregation affect cell shape

ref Adler, H. I. et al. (1967). Miniature *E. coli* cells deficient in DNA. Proc. Nat. Acad. Sci. USA 57, 321-326.

13.17 FtsZ is necessary for septum formation

rev Lutkenhaus, J. and Addinall, S. G. (1997). Bacterial cell division and the Z ring. Ann. Rev. Biochem. 66, 93-116.

Rothfield, L., Justice, S. and Garcia-Lara, J. (1999). Bacterial cell division. Ann. Rev. Genet. 33, 423-438.

ref Bi, E. F. and Lutkenhaus, J. (1991). FtsZ ring structure associated with division in *Escherichia coli.* Nature 354, 161-164.

Hale, C. A. and de Boer, P. A. (1997). Direct binding of FtsZ to ZipA, an essential component of the septal ring structure that mediates cell division in *E. coli.* Cell 88, 175-185.

Mercer, K. L., Mercer, K. L., and Mercer, K. L. (2002). The *E. coli* cell division protein FtsW is required to recruit its cognate transpeptidase, FtsI (PBP3), to the division site. J. Bacteriol. 184, 904-912.

Pichoff, S. and Lutkenhaus, J. (2002). Unique and overlapping roles for ZipA and FtsA in septal ring assembly in Escherichia coli. EMBO J. 21, 685-693.

Stricker, J., Maddox, P., Salmon, E. D., and Erickson, H. P. (2002). Rapid assembly dynamics of the *E. coli* FtsZ-ring demonstrated by fluorescence recovery after photobleaching. Proc. Nat. Acad. Sci. USA 99, 3171-3175.

13.18 *min* genes regulate the location of the septum

ref de Boer, P. A. J. et al. (1989). A division inhibitor and a topological specificity factor coded for by the minicell locus determine proper placement of the division septum in *E. coli.* Cell 56, 641-649.

Hu, Z., Mukherjee, A., Pichoff, S., and Lutkenhaus, J. (1999). The MinC component of the division site selection system in E *coli* interacts with FtsZ to prevent polymerization. Proc. Nat. Acad. Sci. USA 96, 14819-14824.

Pichoff, S. and Lutkenhaus, J. (2001). *Escherichia coli* division inhibitor MinCD blocks septation by preventing Z-ring formation. J. Bacteriol. 183, 6630-6635.

Raskin, D. M. and de Boer, P. A. J. (1997). The MinE ring: an FtsZ-independent cell structure requires for selection of the correct division site in *E. coli*. Cell 91, 685-694.

13.19 Chromosomal segregation may require site-specific recombination

ref Aussel, L., Barre, F. X., Aroyo, M., Stasiak, A., Stasiak, A. Z., and Sherratt, D. (2002). FtsK Is a DNA motor protein that activates chromosome dimer resolution by switching the catalytic state of the XerC and XerD recombinases. Cell 108, 195-205.

Barre, F. X., Aroyo, M., Colloms, S. D., Helfrich, A., Cornet, F., and Sherratt, D. J. (2000). FtsK functions in the processing of a Holliday junction intermediate during bacterial chromosome segregation. Genes Dev. 14, 2976-2988.

Blakely, G., May, G., McCulloch, R., Arciszewska, L. K., Burke, M., Lovett, S. T., and Sherratt, D. J. (1993). Two related recombinases are required for site-specific recombination at dif and cer in *E. coli* K12. Cell 75, 351-361.

13.20 Partitioning involves separation of the chromosomes

rev Draper, G. C. and Gober, J. W. (2002). Bacterial chromosome segregation. Ann. Rev. Microbiol. 56, 567-597.

Errington, J., Bath, J., and Wu, L. J. (2001). DNA transport in bacteria. Nat. Rev. Mol. Cell Biol. 2, 538-545.

Gordon, G. S. and Wright, A. (2000). DNA segregation in bacteria. Ann. Rev. Microbiol. 54, 681-708.

Hiraga, S. (1992). Chromosome and plasmid partition in *E. coli.* Ann. Rev. Biochem. 61, 283-306.

Wake, R. G. and Errington, J. (1995). Chromosome partitioning in bacteria. Ann. Rev. Genet. 29, 41-67.

ref Jacob, F., Ryter, A., and Cuzin, F. (1966). On the association between DNA and the membrane in bacteria. Proc. Roy. Soc. Lond. B Biol. Sci. 164, 267-348.

Sawitzke, J. A. and Austin, S. (2000). Suppression of chromosome segregation defects of *E. coli muk* mutants by mutations in topoisomerase I. Proc. Nat. Acad. Sci. USA 97, 1671-1676.

Wu, L. J. and Errington, J. (1997). Septal localization of the SpollIE chromosome partitioning protein in *B. subtilis.* EMBO J. 16, 2161-2169.

13.21 Single-copy plasmids have a partitioning system

rev Draper, G. C. and Gober, J. W. (2002). Bacterial chromosome segregation. Ann. Rev. Microbiol. 56, 567-597.

Gordon, G. S. and Wright, A. (2000). DNA segregation in bacteria. Ann. Rev. Microbiol. 54, 681-708.

Hiraga, S. (1992). Chromosome and plasmid partition in *E coli* Ann Rev Biochem 61 283-306

ref Davis, M. A. and Austin, S. J. (1988). Recognition of the P1 plasmid centromere analog involves binding of the ParB protein and is modified by a specific host factor EMBOJ 7 1881-1888

Funnell, B. E. (1988). Participation of *E col,* integration host factor in the P1 plasmid partition system. Proc. Nat. Acad. Sci. USA 85, 6657-6661

Mohl D. A. and Gober, J. W. (1997). Cell cycle-dependent polar localization of chromosome partitioning proteins *crescentus.* Cell 88, 675-684.

Surtees, J. A., Surtees, J. A., Surtees, J. A., and Funnell, B. E. (2001). The DNA binding domains of P1 ParB and the architecture of the P1 plasmid partition complex. J. Biol. Chem. 276, 12385-12394.

13.22 Plasmid incompatibility is determined by the replicon

rev Nordstrom, K. and Austin, S. J. (1989). Mechanisms that contribute to the stable segregation of plasmids. Ann. Rev. Genet. 23, 37-69.

Scott, J. R. (1984). Regulation of plasmid replication. Microbiol. Rev. 48, 1-23.

13.23 The ColE1 compatibility system is controlled by an RNA regulator

ref Masukata, H. and Tomizawa, J. (1990). A mechanism of formation of a persistent hybrid between elongating RNA and template DNA. Cell 62, 331-338.

Tomizawa, J.-I. and Itoh, T. (1981). Plasmid ColE1 incompatibility determined by interaction of RNA with primer transcript. Proc. Nat. Acad. Sci. USA 78, 6096-6100.

13.24 How do mitochondria replicate and segregate?

rev Birky, C. W. (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Ann. Rev. Genet. 35, 125-148.

Chapter 14

DNA replication

14.1 Introduction

Replication of duplex DNA is a complex endeavor involving a conglomerate of enzyme activities. Different activities are involved in the stages of initiation, elongation, and termination:

- Initiation involves recognition of an origin by a complex of proteins. Before DNA synthesis begins, the parental strands must be separated and (transiently) stabilized in the single-stranded state. Then synthesis of daughter strands can be initiated at the replication fork.
- Elongation is undertaken by another complex of proteins. The replisome exists only as a protein complex associated with the particular structure that DNA takes at the replication fork. It does not exist as an independent unit (for example, analogous to the ribosome). As the replisome moves along DNA, the parental strands unwind and daughter strands are synthesized.
- At the end of the replicon, joining and/or termination reactions are necessary. Following termination, the duplicate chromosomes must be separated from one another, which requires manipulation of higher-order DNA structure.

Inability to replicate DNA is fatal for a growing cell. Mutants in replication must therefore be obtained as conditional lethals. These are able to accomplish replication under permissive conditions (provided by the normal temperature of incubation), but they are defective under nonpermissive conditions (provided by the higher temperature of 42°C). A comprehensive series of such temperature-sensitive mutants in *E. coli* identifies a set of loci called the *dna* genes. The dna mutants distinguish two stages of replication by their behavior when the temperature is raised:

- The major class of quick-stop mutants cease replication immediately on a temperature rise. They are defective in the components of the replication apparatus, typically in the enzymes needed for elongation (but also include defects in the supply of essential precursors).
- The smaller class of slow-stop mutants complete the current round of replication, but cannot start another. They are defective in the events involved in initiating a cycle of replication at the origin.

An important assay used to identify the components of the replication apparatus is called *in vitro* complementation. An *in vitro* system for

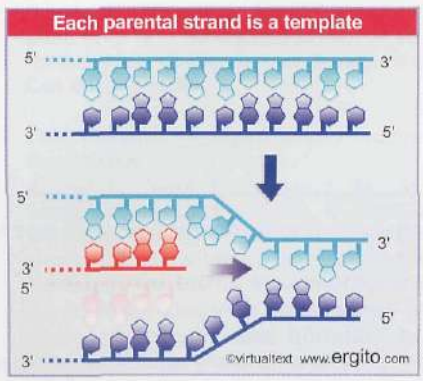


Figure 14.1 Semiconservative replication synthesizes two new strands of DNA.

replication is prepared from a *dna* mutant and operated under conditions in which the mutant gene product is inactive. Extracts from wild-type cells are tested for their ability to restore activity. The protein coded by the *dna* locus can be purified by identifying the active component in the extract.

Each component of the bacterial replication apparatus is now available for study ***in*** *vitro* as a biochemically pure product, and is implicated *in vivo* by mutations in its gene. Eukaryotic replication systems are highly purified, and usually have components analogous to the bacterial proteins, but have not necessarily reached the stage of identification of every single component.

14.2 DNA polymerases are the enzymes that make DNA

Key Concepts

* DNA is synthesized in both semiconservative replication and repair reactions.
• A bacterium or eukaryotic cell has several different DNA polymerase enzymes.
* One bacterial DNA polymerase undertakes semiconservative replication; the others are involved in repair reactions.
• Eukaryotic nuclei, mitochondria, and chloroplasts each have a single unique DNA polymerase required for replication, and other DNA polymerases involved in ancillary or repair activities.

There are two basic types of DNA synthesis.

Figure 14.1 shows the result of semiconservative replication. The two strands of the parental duplex are separated, and each serves as a template for synthesis of a new strand. The parental duplex is replaced with two daughter duplexes, each of which has one parental strand and one newly synthesized strand.

Figure 14.2 shows the consequences of a repair reaction. One strand of DNA has been damaged. It is excised and new material is synthesized to replace it. (Repair synthesis is not the only way to replace damaged DNA; the reactions involved in replacement of damaged sequences are discussed **in** *15 Recombination and repair.)*

An enzyme that can synthesize a new DNA strand on a template strand is called a DNA polymerase. Both prokaryotic and eukaryotic cells contain multiple DNA polymerase activities. Only some of these enzymes actually undertake replication; sometimes they are called DNA replicases. The others are involved in subsidiary roles in replication and/or participate in repair synthesis.

All prokaryotic and eukaryotic DNA polymerases share the same fundamental type of synthetic activity. Each can extend a DNA chain by adding nucleotides one at a time to a 3′-OH end, as illustrated diagrammatically in **Figure 14.3.** The choice of the nucleotide to add to the chain is dictated by base pairing with the template strand.

Some DNA polymerases function as independent enzymes, but others (most notably the **replicases)** are incorporated into large protein assemblies. The DNA-synthesizing subunit is only one of several functions of the replicase, which typically contains many other activities concerned with unwinding DNA, initiating new strand synthesis, and so on.

Figure 14.4 summarizes the DNA polymerases that have been characterized in *E. coli.* DNA polymerase **III,** a multisubunit protein, is the

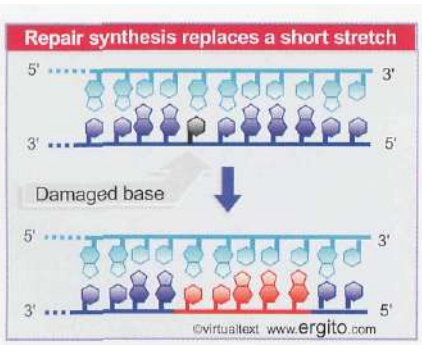


Figure 14.2 Repair synthesis replaces a short stretch of one strand of DNA containing a damaged base.

replicase responsible for *de novo* synthesis of new strands of DNA. DNA polymerase **1** (coded by *polA*) is involved in the repair of damaged DNA and, in a subsidiary role, in semiconservative replication. Other enzymes (DNA polymerases 11, IV and V) are involved in specific repair reactions.

When extracts of *E. coli* are assayed for their ability to synthesize DNA, the predominant enzyme activity is DNA polymerase I. Its activity is so great that it makes it impossible to detect the activities of the enzymes actually responsible for DNA replication! To develop *in vitro* systems in which replication can be followed, extracts are therefore prepared from *polA* mutant cells.

Some phages code for DNA polymerases. They include T4, T5, T7, and SPO1. The enzymes all possess 5'-3' synthetic activities and 3'-5' exonuclease proofreading activities (see next section). In each case, a mutation in the gene that codes for a single phage polypeptide prevents phage development. Each phage polymerase polypeptide associates with other proteins, of either phage or host origin, to make the intact enzyme.

Several classes of eukaryotic DNA polymerases have been identified. DNA polymerases 8 and ε are required for nuclear replication; DNA polymerase a is concerned with "priming" (initiating) replication. Other DNA polymerases are involved in repairing damaged nuclear DNA (β and also ε) or with mitochondrial DNA replication (7).

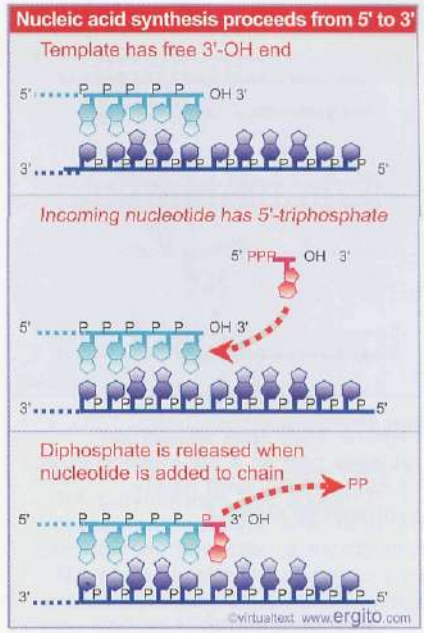


Figure 14.3 DNA synthesis occurs by adding nucleotides to the 3′-OH end of the growing chain, so that the new chain grows in the 5′-3′ direction. The precursor for DNA synthesis is a nucleoside triphosphate, which loses the terminal two phosphate groups in the reaction.

14.3 DNA polymerases have various nuclease activities

Key Concepts

- DNA polymerase 1 has a unique 5'-3' exonuclease activity that can be combined with DNA synthesis to perform nick translation.

Replicases often have nuclease activities as well as the ability to synthesize DNA. A 3'-5' exonuclease activity is typically used to excise bases that have been added to DNA incorrectly. This provides a "proofreading" error-control system (see next section).

The first DNA-synthesizing enzyme to be characterized was DNA polymerase I, which is a single polypeptide of 103 kD. The chain can be cleaved into two parts by proteolytic treatment. The C-terminal two-thirds of the protein contains the polymerase active site, while the N-terminal third contains the proofreading exonuclease.

The larger cleavage product (68 kD) is called the Klenow fragment. It is used in synthetic reactions *in vitro*. It contains the polymerase and the 3'-5' exonuclease activities. The active sites are ~30 A apart in the protein, indicating that there is spatial separation between adding a base and removing one.

The small fragment (35 kD) possesses a 5′-3′ exonucleolytic activity, which excises small groups of nucleotides, up to ~10 bases at a time. This activity is coordinated with the synthetic/proofreading activity. It provides DNA polymerase 1 with a unique ability to start replication *in vitro* at a nick in DNA. (No other DNA polymerase has this ability.) At a point where a phosphodiester bond has been broken in a double-stranded DNA, the enzyme extends the 3′-OH end. As the new segment of DNA is synthesized, it displaces the existing homologous strand in the duplex.

This process of nick translation is illustrated in **Figure 14.5.** The displaced strand is degraded by the *5′-3′* exonucleolytic activity of the

E. coli has 5 DNA polymerases

Enzyme	Gene	Function
I	*polA*	major repair enzyme
II	*polB*	minor repair enzyme
III	*polC*	replicase
IV	*dinB*	SOS repair
V	*umuD′₂C*	SOS repair

©virtualtext www.ergito.com

Figure 14.4 Only one DNA polymerase is the replicase. The others participate in repair of damaged DNA.

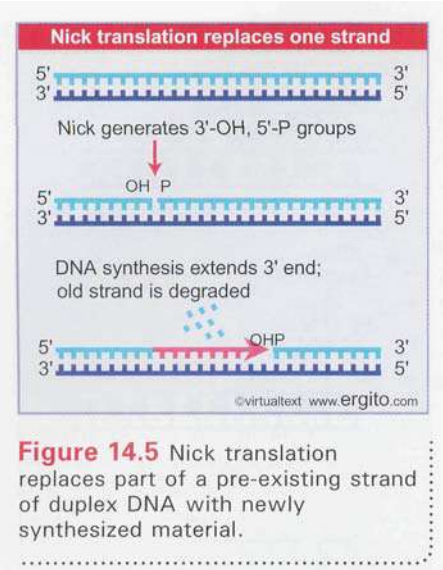


Figure 14.5 Nick translation replaces part of a pre-existing strand of duplex DNA with newly synthesized material.

enzyme. The properties of the DNA are unaltered, except that a segment of one strand has been replaced with newly synthesized material, and the position of the nick has been moved along the duplex. This is of great practical use; nick translation has been a major technique for introducing radioactively labeled nucleotides into DNA *in vitro*.

The 5′-3′ synthetic/3′-5′ exonucleolytic action is probably used *in vivo* mostly for filling in short single-stranded regions in double-stranded DNA. These regions arise during replication, and also when bases that have been damaged are removed from DNA (see *15 Recombination and repair*).

14.4 DNA polymerases control the fidelity of replication

Key Concepts

- DNA polymerases often have a 3′-5′ exonuclease activity that is used to excise incorrectly paired bases.
- The fidelity of replication is improved by proofreading by a factor of ~100.

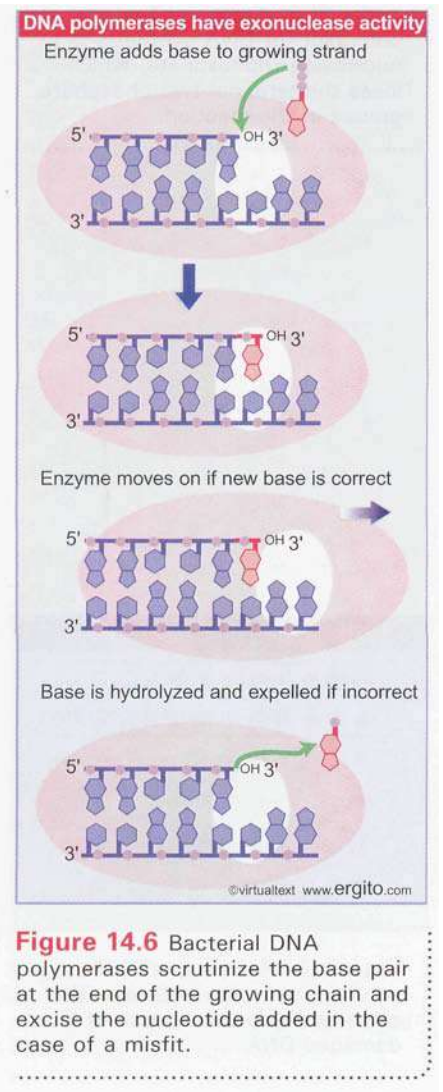


Figure 14.6 Bacterial DNA polymerases scrutinize the base pair at the end of the growing chain and excise the nucleotide added in the case of a misfit.

The fidelity of replication poses the same sort of problem we have encountered already in considering (for example) the accuracy of translation. It relies on the specificity of base pairing. Yet when we consider the interactions involved in base pairing, we would expect errors to occur with a frequency of ~10^{-3} per base pair replicated. The actual rate in bacteria seems to be ~10^{-8}-10^{-10}. This corresponds to ~1 error per genome per 1000 bacterial replication cycles, or ~10^{-6} per gene per generation.

We can divide the errors that DNA polymerase makes during replication into two classes:

- Frameshifts occur when an extra nucleotide is inserted or omitted. Fidelity with regard to frameshifts is affected by the processivity of the enzyme: the tendency to remain on a single template rather than to dissociate and reassociate. This is particularly important for the replication of a homopolymeric stretch, for example, a long sequence of $dT_n : dA_n$, in which "replication slippage" can change the length of the homopolymeric run. As a general rule, increased processivity reduces the likelihood of such events. In multimeric DNA polymerases, processivity is usually increased by a particular subunit that is not needed for catalytic activity *per se*.
- Substitutions occur when the wrong (improperly paired) nucleotide is incorporated. The error level is determined by the efficiency of proofreading, in which the enzyme scrutinizes the newly formed base pair and removes the nucleotide if it is mispaired.

All of the bacterial enzymes possess a 3′-5′ exonucleolytic activity that proceeds in the reverse direction from DNA synthesis. This provides the proofreading function illustrated diagrammatically in **Figure 14.6.** In the chain elongation step, a precursor nucleotide enters the position at the end of the growing chain. A bond is formed. The enzyme moves one base pair farther, ready for the next precursor nucleotide to enter. If a mistake has been made, however, the enzyme uses the exonucleolytic activity to excise the last base that was added.

Different DNA polymerases handle the relationship between the polymerizing and proofreading activities in different ways. In some cases, the activities are part of the same protein subunit, but in others

they are contained in different subunits. Each DNA polymerase has a characteristic error rate that is reduced by its proofreading activity. Proofreading typically decreases the error rate in replication from ~10^{-5} to ~10^{-7} per base pair replicated. Systems that recognize errors and correct them following replication then eliminate some of the errors, bringing the overall rate to $<10^{-9}$ per base pair replicated (see *15.24 Controlling the direction of mismatch repair).*

The replicase activity of DNA polymerase III was originally discovered by a lethal mutation in the *dnaE* locus, which codes for the 130 kD α subunit that possesses the DNA synthetic activity. The 3′-5′ exonucleolytic proofreading activity is found in another subunit, e, coded by *dnaQ*. The basic role of the e subunit in controlling the fidelity of replication *in vivo* is demonstrated by the effect of mutations in *dnaQ:* the frequency with which mutations occur in the bacterial strain is increased by > 10^3-fold.

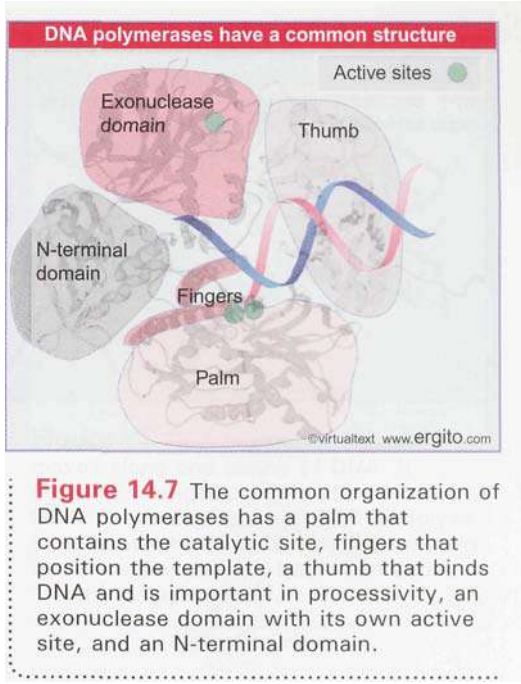


Figure 14.7 The common organization of DNA polymerases has a palm that contains the catalytic site, fingers that position the template, a thumb that binds DNA and is important in processivity, an exonuclease domain with its own active site, and an N-terminal domain.

14.5 DNA polymerases have a common structure

Key Concepts

- Many DNA polymerases have a large cleft composed of three domains that resemble a hand.
- DNA lies across the "palm" in a groove created by the "fingers" and "thumb".

Figure 14.7 shows that all DNA polymerases share some common structural features. The enzyme structure can be divided into several independent domains, which are described by analogy with a human right hand. DNA binds in a large cleft composed of three domains. The "palm" domain has important conserved sequence motifs that provide the catalytic active site. The "fingers" are involved in positioning the template correctly at the active site. The "thumb" binds the DNA as it exits the enzyme, and is important in processivity. The most important conserved regions of each of these three domains converge to form a continuous surface at the catalytic site. The exonuclease activity resides in an independent domain with its own catalytic site. The N-terminal domain extends into the nuclease domain. DNA polymerases fall into five families based on sequence homologies; the palm is well conserved among them, but the thumb and fingers provide analogous secondary structure elements from different sequences.

The catalytic reaction in a DNA polymerase occurs at an active site in which a nucleotide triphosphate pairs with an (unpaired) single strand of DNA. The DNA lies across the palm in a groove that is created by the thumb and fingers. **Figure 14.8** shows the crystal structure of the T7 enzyme complexed with DNA (in the form of a primer annealed to a template strand) and an incoming nucleotide that is about to be added to the primer. The DNA is in the classic B-form duplex up to the last 2 base pairs at the 3' end of the primer, which are in the more open A-form. A sharp turn in the DNA exposes the template base to the incoming nucleotide. The 3' end of the primer (to which bases are added) is anchored by the fingers and palm. The DNA is held in position by contacts that are made principally with the phosphodiester backbone (thus enabling the polymerase to function with DNA of any sequence).

In structures of DNA polymerases of this family complexed only with DNA (that is, lacking the incoming nucleotide), the orientation of the fingers and thumb relative to the palm is more open, with the O

Figure 14.8 The crystal structure of phage T7 DNA polymerase shows that the template strand takes a sharp turn in order to be exposed to the incoming nucleotide. Photograph kindly provided by Charles Richardson and Tom Ellenberger.

helix (O, 01, 02; see Figure 14.8) rotated away from the palm. This suggests that an inward rotation of the O helix occurs to grasp the incoming nucleotide and create the active catalytic site. When a nucleotide binds, the fingers domain rotates 60° toward the palm, with the tops of the fingers moving by 30 A. The thumb domain also rotates toward the palm by 8°. These changes are cyclical: they are reversed when the nucleotide is incorporated into the DNA chain, which then translocates through the enzyme to recreate an empty site.

The exonuclease activity is responsible for removing mispaired bases. But the catalytic site of the exonuclease domain is distant from the active site of the catalytic domain. The enzyme alternates between polymerizing and editing modes, as determined by a competition between the two active sites for the 3' primer end of the DNA. Amino acids in the active site contact the incoming base in such a way that the enzyme structure is affected by a mismatched base. When a mismatched base pair occupies the catalytic site, the fingers cannot rotate toward the palm to bind the incoming nucleotide. This leaves the 3' end free to bind to the active site in the exonuclease domain, which is accomplished by a rotation of the DNA in the enzyme structure.

14.6 DNA synthesis is semidiscontinuous

Key Concepts

- The DNA replicase advances continuously when it synthesizes the leading strand (5′-3′), but synthesizes the lagging strand by making short fragments that are subsequently joined together.

The antiparallel structure of the two strands of duplex DNA poses a problem for replication. As the replication fork advances, daughter strands must be synthesized on both of the exposed parental single strands. The fork moves in the direction from 5′-3′ on one strand, and in the direction from 3′-5′ on the other strand. Yet nucleic acids are synthesized only from a 5' end toward a 3' end. The problem is solved by synthesizing the strand that grows overall from 3′-5′ in a series of short fragments, each actually synthesized in the "backwards" direction, that is, with the customary 5′-3′ polarity.

Consider the region immediately behind the replication fork, as illustrated in **Figure 14.9.** We describe events in terms of the different properties of each of the newly synthesized strands:

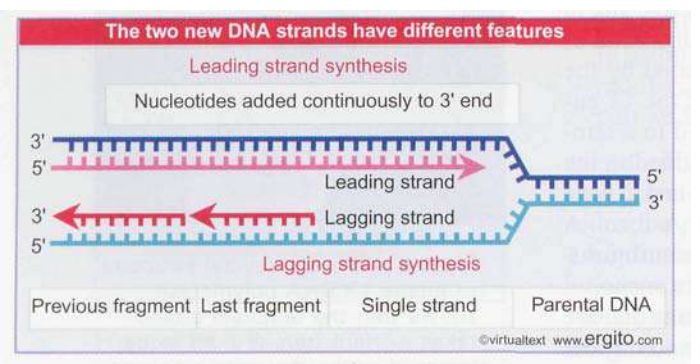


Figure 14.9 The leading strand is synthesized continuously while the lagging strand is synthesized discontinuously.

- On the leading strand DNA synthesis can proceed continuously in the 5' to 3' direction as the parental duplex is unwound.
- On the lagging strand a stretch of single-stranded parental DNA must be exposed, and then a segment is synthesized in the reverse direction (relative to fork movement). A series of these fragments are synthesized, each 5′-3′; then they are joined together to create an intact lagging strand.

Discontinuous replication can be followed by the fate of a very brief label of radioactivity. The label enters newly synthesized DNA in the form of short fragments, sedimenting in the range of 7-11S, corresponding to ~1000-2000 bases in length. These Okazaki fragments are found in replicating DNA in both prokaryotes and eukaryotes. After longer periods of incubation, the label enters larger segments of DNA.

By Book_Crazy [IND]

The transition results from covalent linkages between Okazaki fragments.

(The lagging strand *must* be synthesized in the form of Okazaki fragments. For a long time it was unclear whether the leading strand is synthesized in the same way or is synthesized continuously. All newly synthesized DNA is found as short fragments in *E. coli*. Superficially, this suggests that both strands are synthesized discontinuously. However, it turns out that not all of the fragment population represents *bona fide* Okazaki fragments; some are pseudofragments, generated by breakage in a DNA strand that actually was synthesized as a continuous chain. The source of this breakage is the incorporation of some uracil into DNA in place of thymine. When the uracil is removed by a repair system, the leading strand has breaks until a thymine is inserted.)

So the lagging strand is synthesized discontinuously and the leading strand is synthesized continuously. This is called semidiscontinuous replication.

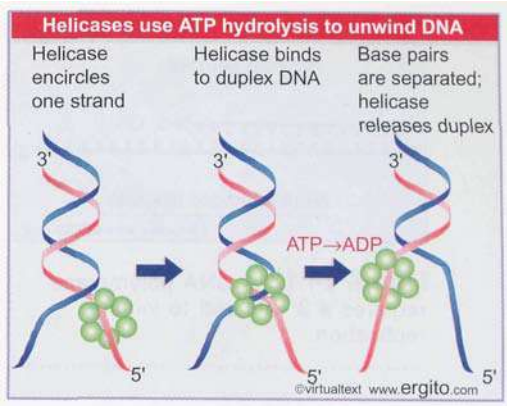


Figure 14.10 A hexameric helicase moves along one strand of DNA. It probably changes conformation when it binds to the duplex, uses ATP hydrolysis to separate the strands, and then returns to the conformation it has when bound only to a single strand.

14.7 The ϕX model system shows how single-stranded DNA is generated for replication

As the replication fork advances, it unwinds the duplex DNA. One of the template strands is rapidly converted to duplex DNA as the leading daughter strand is synthesized. The other remains single-stranded until a sufficient length has been exposed to initiate synthesis of an Okazaki fragment of the lagging strand in the backward direction. The generation and maintenance of single-stranded DNA is therefore a crucial aspect of replication. Two types of function are needed to convert double-stranded DNA to the single-stranded state:

- A helicase is an enzyme that separates the strands of DNA, usually using the hydrolysis of ATP to provide the necessary energy.
- A single-strand binding protein (SSB) binds to the single-stranded DNA, preventing it from reforming the duplex state. The SSB binds as a monomer, but typically in a cooperative manner in which the binding of additional monomers to the existing complex is enhanced.

Helicases separate the strands of a duplex nucleic acid in a variety of situations, ranging from strand separation at the growing point of a replication fork to catalyzing migration of Holliday (recombination) junctions along DNA. There are 12 different helicases in *E. coli*. A helicase is generally multimeric. A common form of helicase is a hexamer. This typically translocates along DNA by using its multimeric structure to provide multiple DNA-binding sites.

Figure 14.10 shows a generalized schematic model for the action of a hexameric helicase. It is likely to have one conformation that binds to duplex DNA and another that binds to single-stranded DNA. Alternation between them drives the motor that melts the duplex, and requires ATP hydrolysis—typically 1 ATP is hydrolyzed for each base pair that is unwound. A helicase usually initiates unwinding at a single-stranded region adjacent to a duplex, and may function with a particular polarity, preferring single-stranded DNA with a 3' end (3′-5′ helicase) or with a 5'end(5'-3' helicase).

The conversion of ϕX174 double-stranded DNA into individual single strands illustrates the features of the strand separation process. **Figure 14.11** shows that a single strand is peeled off the circular strand, resembling the rolling circle described previously in Figure 13.16. The reaction can occur in the absence of DNA synthesis when the appropriate 3 proteins are provided *in vitro*.

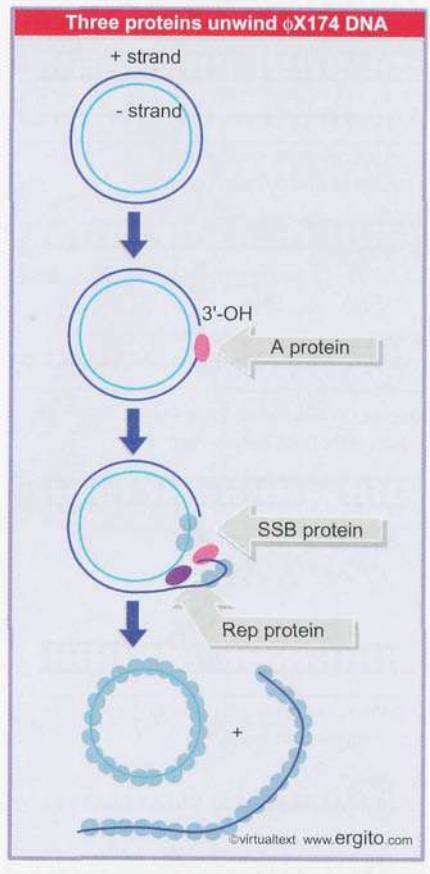


Figure 14.11 ϕX174 DNA can be separated into single strands by the combined effects of 3 functions: nicking with A protein, unwinding by Rep, and single-strand stabilization by SSB.

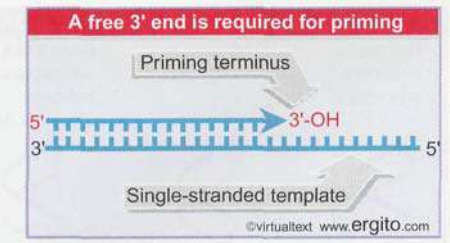


Figure 14.12 A DNA polymerase requires a 3′-OH end to initiate replication.

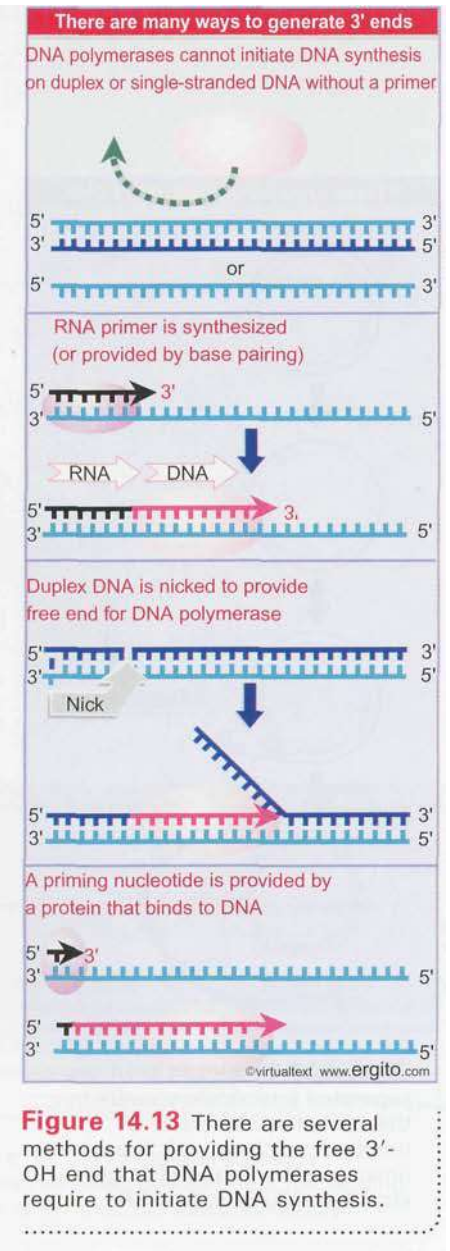


Figure 14.13 There are several methods for providing the free 3′-OH end that DNA polymerases require to initiate DNA synthesis.

The phage A protein nicks the viral (+) strand at the origin of replication. In the presence of 2 host proteins, Rep and SSB, and ATP, the nicked DNA unwinds. The Rep protein provides a helicase that separates the strands; the SSB traps them in single-stranded form. The *E. coli* SSB is a tetramer of 74 kD that binds cooperatively to single-stranded DNA.

The significance of the cooperative mode of binding is that the binding of one protein molecule makes it much easier for another to bind. So once the binding reaction has started on a particular DNA molecule, it is rapidly extended until all of the single-stranded DNA is covered with the SSB protein. Note that this protein is not a DNA-unwinding protein; its function is to stabilize DNA that is already in the single-stranded condition.

Under normal circumstances *in vivo*, the unwinding, coating, and replication reactions proceed in tandem. The SSB binds to DNA as the replication fork advances, keeping the two parental strands separate so that they are in the appropriate condition to act as templates. SSB is needed in stoichiometric amounts at the replication fork. It is required for more than one stage of replication; *ssb* mutants have a quick-stop phenotype, and are defective in repair and recombination as well as in replication. (Some phages use different SSB proteins, notably T4; this shows that there may be specific interactions between components of the replication apparatus and the SSB; see *14.14 Phage T4 provides its own replication apparatus).*

14.8 Priming is required to start DNA synthesis

Key Concepts

- All DNA polymerases require a 3′-OH priming end to initiate DNA synthesis.
- The priming end can be provided by an RNA primer, a nick in DNA, or a priming protein.
- For DNA replication, a special RNA polymerase called a primase synthesizes an RNA chain that provides the priming end.
- *E. coli* has two types of priming reaction, which occur at the bacterial origin *(oriC)* and the ϕX174 origin.
- Priming of replication on double-stranded DNA always requires a replicase, SSB, and primase.
- DnaB is the helicase that unwinds DNA for replication in *E. coli.*

A common feature of all DNA polymerases is that they cannot initiate synthesis of a chain of DNA *de novo*. **Figure 14.12** shows the features required for initiation. Synthesis of the new strand can only start from a pre-existing 3'-OH end; and the template strand must be converted to a single-stranded condition.

The 3'-OH end is called a primer. The primer can take various forms. Types of priming reaction are summarized in **Figure 14.13:**

- A sequence of RNA is synthesized on the template, so that the free 3′-OH end of the RNA chain is extended by the DNA polymerase. This is commonly used in replication of cellular DNA, and by some viruses (see Figure 13.40 in *13.23 The ColE1 compatibility system is controlled by an RNA regulator).*
- A preformed RNA pairs with the template, allowing its 3′-OH end to be used to prime DNA synthesis. This mechanism is used by retroviruses to prime reverse transcription of RNA (see Figure 17.6 in *17.4 Viral DNA is generated by reverse transcription).*

By Book_Crazy [IND]

- A primer terminus is generated within duplex DNA. The most common mechanism is the introduction of a nick, as used to initiate rolling circle replication (see Figure 13.16). In this case, the pre-existing strand is displaced by new synthesis. (Note the difference from nick translation shown in Figure 14.5, in which DNA polymerase I simultaneously synthesizes and degrades DNA from a nick.)
- A protein primes the reaction directly by presenting a nucleotide to the DNA polymerase. This reaction is used by certain viruses (see Figure 13.15 in *13.8 The ends of linear DNA are a problem for replication).*

Priming activity is required to provide 3'-OH ends to start off the DNA chains on both the leading and lagging strands. The leading strand requires only one such initiation event, which occurs at the origin. But there must be a series of initiation events on the lagging strand, since each Okazaki fragment requires its own start *de novo*. Each Okazaki fragment starts with a primer sequence of RNA, ~10 bases long, that provides the 3'-OH end for extension by DNA polymerase.

A primase is required to catalyze the actual priming reaction. This is provided by a special RNA polymerase activity, the product of the *dnaG* gene. The enzyme is a single polypeptide of 60 kD (much smaller than RNA polymerase). The primase is an RNA polymerase that is used only under specific circumstances, that is, to synthesize short stretches of RNA that are used as primers for DNA synthesis. DnaG primase associates transiently with the replication complex, and typically synthesizes an 11-12 base primer. Primers start with the sequence pppAG, opposite the sequence 3'-GTC-5' in the template.

(Some systems use alternatives to the DnaG primase. In the examples of the two phages M13 and G4, which were used for early work on replication, an interesting difference emerged. G4 priming uses DnaG, but M13 priming uses bacterial RNA polymerase. These phages have another unusual feature, which is that the site of priming is indicated by a region of secondary structure.)

There are two types of priming reaction in *E. coli*.

- The *oriC* system, named for the bacterial origin, basically involves the association of the DnaG primase with the protein complex at the replication fork.
- The ϕX system, named for phage ϕX174, requires an initiation complex consisting of additional components, called the primosome (see *14.17 The primosome is needed to restart replication).*

Sometimes replicons are referred to as being of the ϕX or *oriC* type.

The types of activities involved in the initiation reaction are summarized in Figure 14.14. Although other replicons in *E. coli* may have alternatives for some of these particular proteins, the same general types of activity are required in every case. A helicase is required to generate single strands, a single-strand binding protein is required to maintain the single-stranded state, and the primase synthesizes the RNA primer.

DnaB is the central component in both ϕX and *oriC* replicons. It provides the 5′-3′ helicase activity that unwinds DNA. Energy for the reaction is provided by cleavage of ATP. Basically DnaB is the active component of the growing point. In *oriC* replicons, DnaB is initially loaded at the origin as part of a large complex (see *14.15 Creating the replication forks at an origin).* It forms the growing point at which the DNA strands are separated as the replication fork advances. It is part of the DNA polymerase complex and interacts with the DnaG primase to initiate synthesis of each Okazaki fragment on the lagging strand.

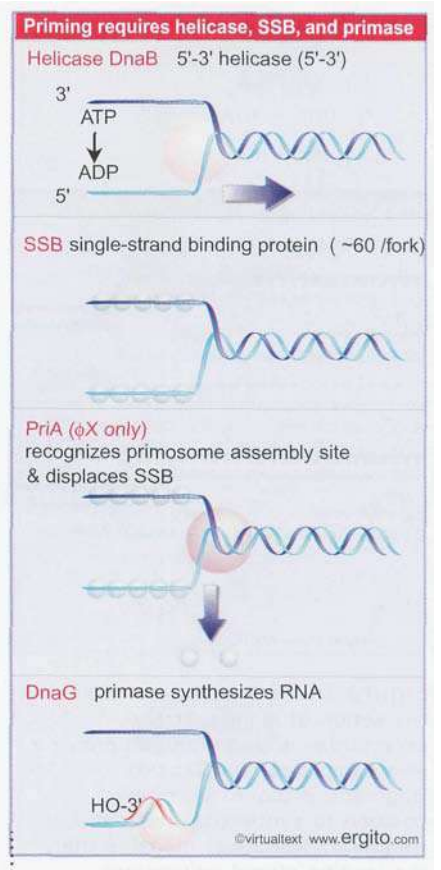


Figure 14.14 Initiation requires several enzymatic activities, including helicases, single-strand binding proteins, and synthesis of the primer.

14.9 Coordinating synthesis of the lagging and leading strands

Key Concepts

- Different enzyme units are required to synthesize the leading and lagging strands.
- In *E. coli* both these units contain the same catalytic subunit (DnaE).
- In other organisms, different catalytic subunits may be required for each strand.

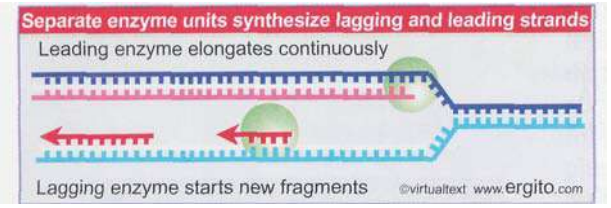


Figure 14.15 Leading and lagging strand polymerases move apart.

Each new DNA strand is synthesized by an individual catalytic unit. **Figure 14.15** shows that the behavior of these two units is different because the new DNA strands are growing in opposite directions. One enzyme unit is moving with the unwinding point and synthesizing the leading strand continuously. The other unit is moving "backwards," relative to the DNA, along the exposed single strand. Only short segments of template are exposed at any one time. When synthesis of one Okazaki fragment is completed, synthesis of the next Okazaki fragment is required to start at a new location approximately in the vicinity of the growing point for the leading strand. This requires a translocation relative to the DNA of the enzyme unit that is synthesizing the lagging strand.

The term "enzyme unit" avoids the issue of whether the DNA polymerase that synthesizes the leading strand is the same type of enzyme as the DNA polymerase that synthesizes the lagging strand. In the case that we know best, *E. coli*, there is only a single type of DNA polymerase catalytic subunit used in replication, the DnaE protein. The active replicase is a dimer, and each half of the dimer contains DnaE as the catalytic subunit, supported by other proteins (which differ between the leading and lagging strands).

The use of a single type of catalytic subunit, however, may be atypical. In the bacterium *B. subtilis*, there are two different catalytic subunits. PolC is the homologue to *E. coli*'s DnaE, and is responsible for synthesizing the leading strand. A related protein, $DnaE_{BS}$, is the catalytic subunit that synthesizes the lagging strand. Eukaryotic DNA polymerases have the same general structure, with different enzyme units synthesizing the leading and lagging strands, but it is not clear whether the same or different types of catalytic subunits are used (see *14.13 Separate eukaryotic DNA polymerases undertake initiation and elongation).*

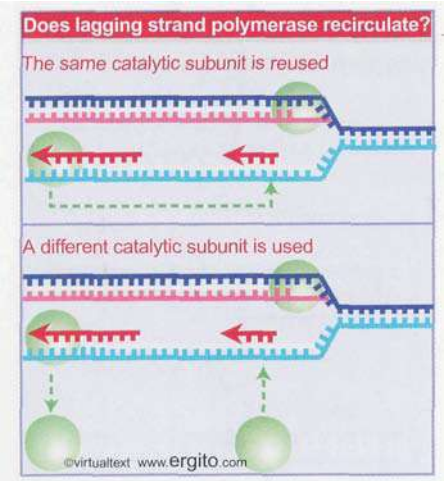


Figure 14.16 The upper model for the action of lagging strand polymerase is that when an enzyme unit completes one Okazaki fragment, it moves to a new position to synthesize the next fragment. The lower model is that the lagging strand polymerase dissociates when it completes an Okazaki fragment, and a new enzyme unit associates with DNA to synthesize the next Okazaki fragment.

A major problem of the semidiscontinuous mode of replication follows from the use of different enzyme units to synthesize each new DNA strand: how is synthesis of the lagging strand coordinated with synthesis of the leading strand? As the replisome moves along DNA, unwinding the parental strands, one enzyme unit elongates the leading strand. Periodically the primosome activity initiates an Okazaki fragment on the lagging strand, and the other enzyme unit must then move in the reverse direction to synthesize DNA. **Figure 14.16** proposes two types of model for what happens to this enzyme unit when it completes synthesis of an Okazaki fragment. The same complex may be reutilized for synthesis of successive Okazaki fragments. Or the complex might dissociate from the template, so that a new complex must be assembled to elongate the next Okazaki fragment. We see in the *14.11 The clamp controls association of core enzyme with DNA* that the first model applies.

By Book_Crazy [IND]

14.10 DNA polymerase holoenzyme has 3 subcomplexes

Key Concepts

- The *E. coli* replicase DNA polymerase III is a 900 kD complex with a dimeric structure.
- Each monomeric unit has a catalytic core, a dimerization subunit, and a processivity component.
- A clamp loader places the processivity subunits on DNA, and they form a circular clamp around the nucleic acid.
- One catalytic core is associated with each template strand.

We can now relate the subunit structure of *E. coli* DNA polymerase III to the activities required for DNA synthesis and propose a model for its action. The holoenzyme is a complex of 900 kD that contains 10 proteins organized into four types of subcomplex:

- There are two copies of the catalytic core. Each catalytic core contains the a subunit (the DNA polymerase activity), e subunit (3′-5′ proofreading exonuclease), and 0 subunit (stimulates exonuclease).
- There are two copies of the dimerizing subunit, τ, which link the two catalytic cores together.
- There are two copies of the processivity subunit, β, which are responsible for holding catalytic cores on to their template strands.
- The γ complex is a group of 5 proteins, the *clamp loader* (γ), that places the processivity subunit on DNA.

A model for the assembly of DNA polymerase III is depicted in Figure 14.17. The holoenzyme assembles on DNA in three stages:

- Aβ dimer plus a 7 complex recognizes the primer-template to form a preinitiation complex. In this reaction, the γ complex cleaves ATP and transfers β subunits to the primed template. The pair of β subunits forms a clamp that binds around the DNA and ensures processivity. The *y* complex uses hydrolysis of ATP to drive the binding of β to DNA.
- Binding to DNA changes the conformation of the site onβ that binds to the γ complex, and as a result it now has a high affinity for the core polymerase. This enables core polymerase to bind, and this is the means by which the core polymerase is brought to DNA. (The processivity of the core by itself is low; but the β clamp ensures that it functions processively on the DNA.)
- Aτ dimer binds to the core polymerase, and provides a dimerization function that binds a second core polymerase (associated with another β clamp). The holoenzyme is asymmetric, because it has only 1 7 complex. The 7 complex is responsible for adding a pair of β dimers to each parental strand of DNA.

Each of the core complexes of the holoenzyme synthesizes one of the new strands of DNA. Because the clamp loader is also needed for unloading the β complex from DNA, the two cores have different abilities to dissociate from DNA. This corresponds to the need to synthesize a continuous leading strand (where polymerase remains associated with the template) and a discontinuous lagging strand (where polymerase repetitively dissociates and reassociates). The clamp loader is associated with the core polymerase that synthesizes the lagging strand, and plays a key role in the ability to synthesize individual Okazaki fragments.

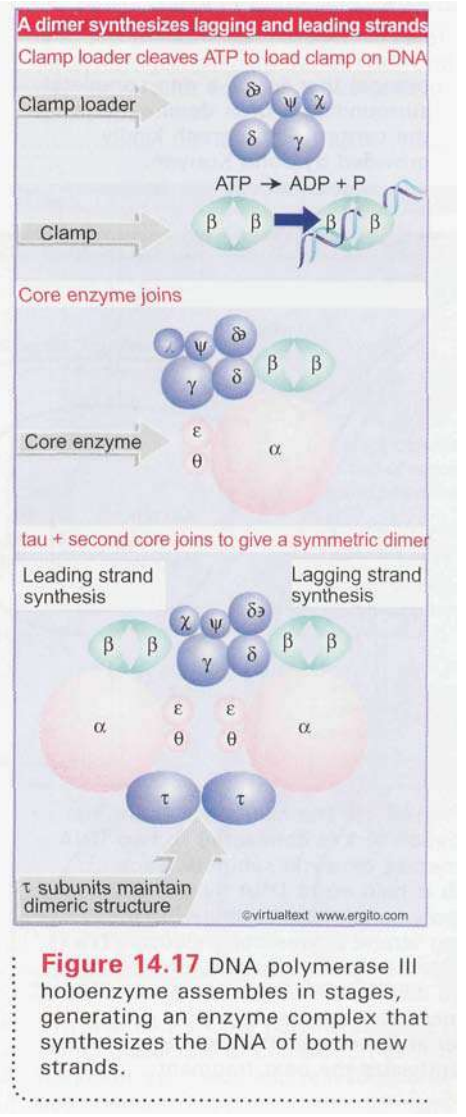


Figure 14.17 DNA polymerase III holoenzyme assembles in stages, generating an enzyme complex that synthesizes the DNA of both new strands.

By Book_Crazy [IND]

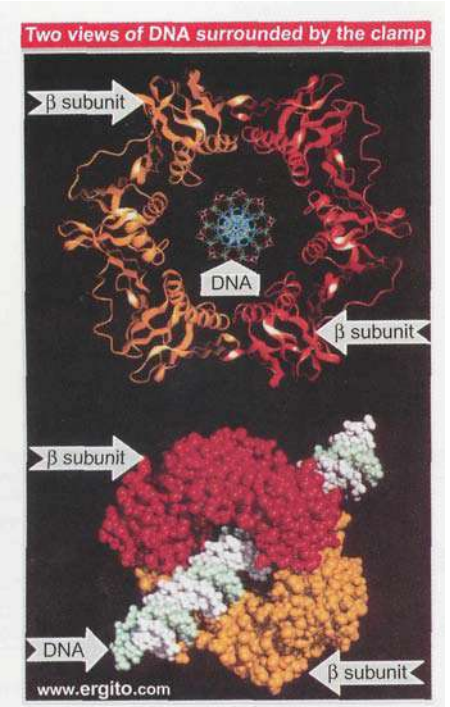


Figure 14.18 The β subunit of DNA polymerase III holoenzyme consists of a head-to-tail dimer (the two subunits are shown in red and orange) that forms a ring completely surrounding a DNA duplex (shown in the center). Photograph kindly provided by John Kuriyan.

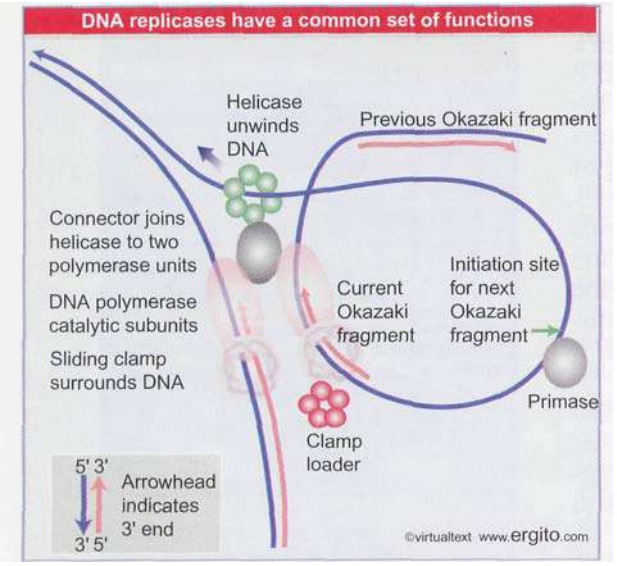


Figure 14.19 The helicase creating the replication fork is connected to two DNA polymerase catalytic subunits, each of which is held on to DNA by a sliding clamp. The polymerase that synthesizes the leading strand moves continuously. The polymerase that synthesizes the lagging strand dissociates at the end of an Okazaki fragment and then reassociates with a primer in the single-stranded template loop to synthesize the next fragment.

14.11 The clamp controls association of core enzyme with DNA

Key Concepts

- The core on the leading strand is processive because its clamp keeps it on the DNA.
- The clamp associated with the core on the lagging strand dissociates at the end of each Okazaki fragment and reassembles for the next fragment.
- The helicase DnaB is responsible for interacting with the primase DnaG to initiate each Okazaki fragment.

The β dimer makes the holoenzyme highly processive. β is strongly bound to DNA, but can slide along a duplex molecule. The crystal structure of β shows that it forms a ring-shaped dimer. The model in **Figure 14.18** shows the β-ring in relationship to a DNA double helix. The ring has an external diameter of 80 A and an internal cavity of 35 A, almost twice the diameter of the DNA double helix (20 A). The space between the protein ring and the DNA is filled by water. Each of the β subunits has three globular domains with similar organization (although their sequences are different). As a result, the dimer has 6-fold symmetry, reflected in 12 α-helices that line the inside of the ring.

The dimer surrounds the duplex, providing the "sliding clamp" that allows the holoenzyme to slide along DNA. The structure explains the high processivity—there is no way for the enzyme to fall off! The α-helices on the inside have some positive charges that may interact with the DNA via the intermediate water molecules. Because the protein clamp does not directly contact the DNA, it may be able to "ice-skate" along the DNA, making and breaking contacts via the water molecules.

How does the clamp get on to the DNA? Because the clamp is a circle of subunits surrounding DNA, its assembly or removal requires the use of an energy-dependent process by the clamp loader. The γ clamp loader is a pentameric circular structure that binds an open form of the β ring preparatory to loading it on to DNA. In effect, the ring is opened at one of the interfaces between the two β subunits by the δ subunit of the clamp loader. The clamp loader uses hydrolysis of ATP to provide the energy to open the ring of the clamp and insert DNA into its central cavity.

The relationship between the β clamp and the γ clamp loader is a paradigm for similar systems used by DNA replicases ranging from bacteriophages to animal cells. The clamp is a heteromer (sometimes a dimer, sometimes a trimer) that forms a ring around DNA with a set of 12 α-helices forming 6-fold symmetry for the structure as a whole. The clamp loader has some subunits that hydrolyze ATP to provide energy for the reaction.

The basic principle that is established by the dimeric polymerase model is that, while one polymerase subunit synthesizes the leading strand continuously, the other cyclically initiates and terminates the Okazaki fragments of the lagging strand within a large single-stranded loop formed by its template strand. **Figure 14.19** draws a generic model for the operation of such a replicase. The replication fork is created by a helicase, typically forming a hexameric ring, that translocates in the 5′-3′ direction on the template for the lagging strand. The helicase is connected to two DNA polymerase catalytic subunits, each of which is associated with a sliding clamp.

By Book_Crazy [IND]

We can describe this model for DNA polymerase III in terms of the individual components of the enzyme complex, as illustrated in **Figure 14.20.** A catalytic core is associated with each template strand of DNA. The holoenzyme moves continuously along the template for the leading strand; the template for the lagging strand is "pulled through," creating a loop in the DNA. DnaB creates the unwinding point, and translocates along the DNA in the "forward" direction.

DnaB contacts the T subunit(s) of the γ complex. This establishes a direct connection between the helicase-primase complex and the catalytic cores. This link has two effects. One is to increase the speed of DNA synthesis by increasing the rate of movement by DNA polymerase core by 10X. The second is to prevent the leading strand polymerase from falling off, that is, to increase its processivity.

Synthesis of the leading strand creates a loop of single-stranded DNA that provides the template for lagging strand synthesis, and this loop becomes larger as the unwinding point advances. After initiation of an Okazaki fragment, the lagging strand core complex pulls the single-stranded template through the *ft* clamp while synthesizing the new strand. The single-stranded template must extend for the length of at least one Okazaki fragment before the lagging polymerase completes one fragment and is ready to begin the next.

What happens to the loop when the Okazaki fragment is completed? When a Pol III holoenzyme meets a nick in DNA, the core complex and clamp dissociate from the *ft* sliding clamp. The core can then reassociate with a new β subunit. **Figure 14.21** suggests that this represents the reaction that occurs at the end of an Okazaki fragment. The core complex dissociates when it completes synthesis of each fragment, releasing the loop. The core complex then associates with a β clamp to initiate the next Okazaki fragment. Probably a new *ft* clamp will already be present at the next initiation site, and the β clamp that has lost its core complex will dissociate from the template (with the assistance of the clamp loader complex) to be used again. So the lagging strand polymerase will probably transfer from one β clamp to the next in each cycle, without dissociating from the replicating complex.

What is responsible for recognizing the sites for initiating synthesis of Okazaki fragments? In *oriC* replicons, the connection between priming and the replication fork is provided by the dual properties of DnaB: it is the helicase that propels the replication fork, and it interacts with the DnaG primase at an appropriate site. Following primer synthesis, the primase is released. The length of the priming RNA is limited to 8-14 bases. Apparently DNA polymerase **III** is responsible for displacing the primase.

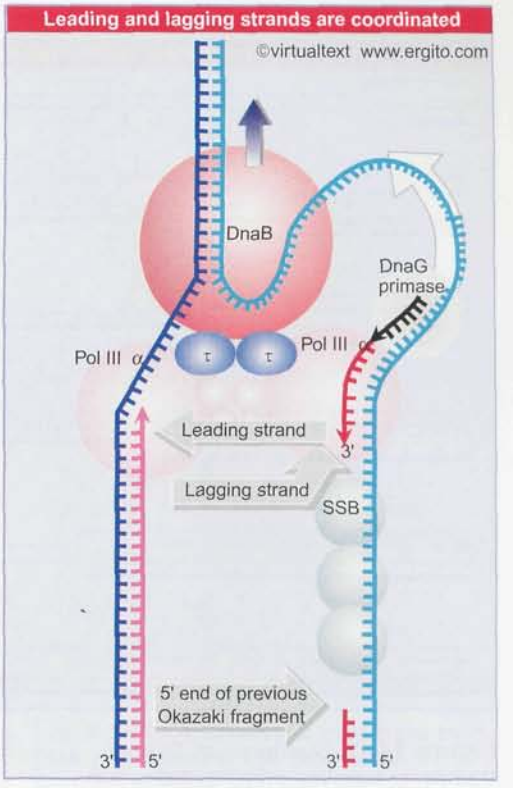


Figure 14.20 Each catalytic core of Pol III synthesizes a daughter strand. DnaB is responsible for forward movement at the replication fork.

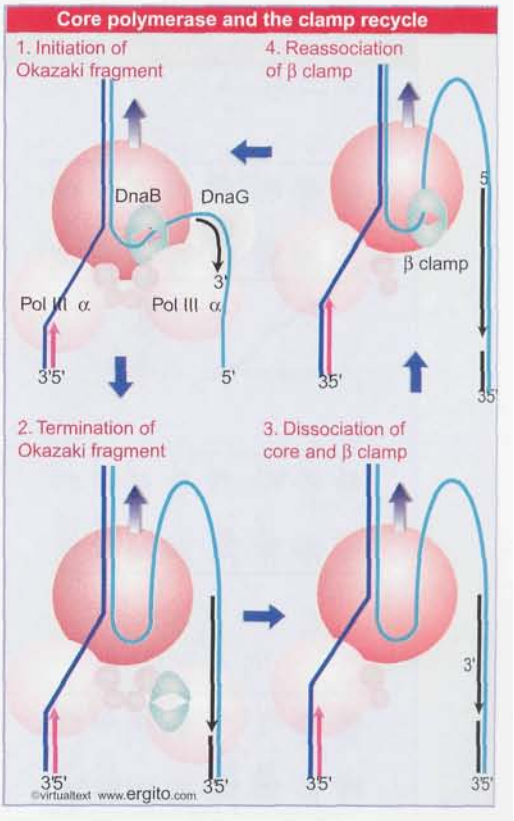


Figure 14.21 Core polymerase and the β clamp dissociate at completion of Okazaki fragment synthesis and reassociate at the beginning.

14.12 Okazaki fragments are linked by ligase

Key Concepts

- Each Okazaki fragment starts with a primer and stops before the next fragment.
- DNA polymerase I removes the primer and replaces it with DNA in an action that resembles nick translation.
- DNA ligase makes the bond that connects the 3' end of one Okazaki fragment to the 5' beginning of the next fragment.

We can now expand our view of the actions involved in joining Okazaki fragments, as illustrated in **Figure 14.22.** The complete order of events is uncertain, but must involve synthesis of RNA

By Book_Crazy [IND]

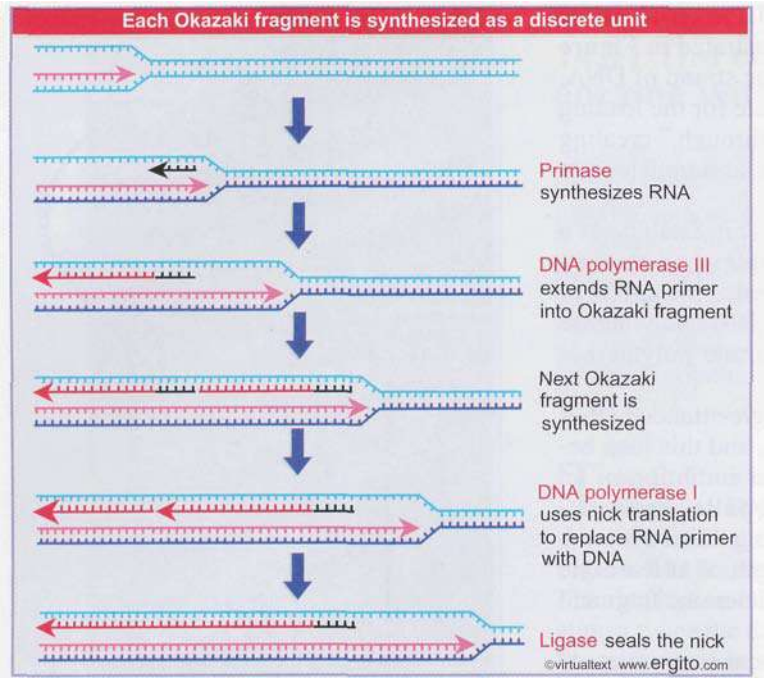


Figure 14.22 Synthesis of Okazaki fragments requires priming, extension, removal of RNA, gap filling, and nick ligation.

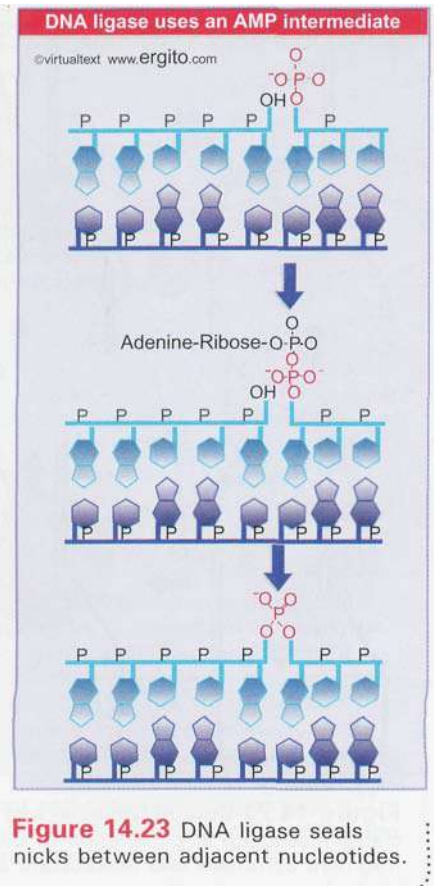


Figure 14.23 DNA ligase seals nicks between adjacent nucleotides.

primer, its extension with DNA, removal of the RNA primer, its replacement by a stretch of DNA, and the covalent linking of adjacent Okazaki fragments.

The figure suggests that synthesis of an Okazaki fragment terminates just before the start of the RNA primer of the preceding fragment. When the primer is removed, there will be a gap. The gap is filled by DNA polymerase I; *polA* mutants fail to join their Okazaki fragments properly. The 5′-3′ exonuclease activity removes the RNA primer while simultaneously replacing it with a DNA sequence extended from the 3′-OH end of the next Okazaki fragment. This is equivalent to nick translation, except that the new DNA replaces a stretch of RNA rather than a segment of DNA. In mammalian systems (where the DNA polymerase does not have a 5′-3′ exonuclease activity), Okazaki fragments are removed by a two-step process/First RNAase HI (an enzyme that is specific for a DNA-RNA hybrid substrate) makes an endonucleolytic cleavage; then a 5′-3′ exonuclease called FEN1 removes the RNA.

Once the RNA has been removed and replaced, the adjacent Okazaki fragments must be linked together. The 3'-OH end of one fragment is adjacent to the 5′-phosphate end of the previous fragment. The responsibility for sealing this nick lies with the enzyme DNA ligase. Ligases are present in both prokaryotes and eukaryotes. Unconnected fragments persist in *lig* mutants, because they fail to join Okazaki fragments together.

The *E. coli* and T4 ligases share the property of sealing nicks that have 3′-OH and 5'-phosphate termini, as illustrated in **Figure 14.23.** Both enzymes undertake a two-step reaction, involving an enzyme-AMP complex. (The *E. coli* and T4 enzymes use different cofactors. The *E. coli* enzyme uses NAD (nicotinamide adenine dinucleotide) as a cofactor, while the T4 enzyme uses ATR) The AMP of the enzyme complex becomes attached to the 5′-phosphate of the nick; and then a phosphodiester bond is formed with the 3'-OH terminus of the nick, releasing the enzyme and the AMP.

14.13 Separate eukaryotic DNA polymerases undertake initiation and elongation

Key Concepts

- A replication fork has 1 complex of DNA polymerase α /primase and 2 complexes of DNA polymerase 8 and/or e.
- The DNA polymerase a /primase complex initiates the synthesis of both DNA strands.
- DNA polymerase 8 elongates the leading strand and a second DNA polymerase 8 or DNA polymerase e elongates the lagging strand.

Eukaryotic cells have a large number of DNA polymerases. They can be broadly divided into those required for semiconservative replication and those involved in synthesizing material to repair damaged DNA. Nuclear DNA replication requires DNA polymerases a, 5, and ε, and mitochondrial replication requires DNA polymerase γ. All the other enzymes are concerned with synthesizing stretches of new DNA to replace

damaged material. **Figure 14.24** shows that all of the nuclear replicases are large heterotetrameric enzymes. In each case, one of the subunits has the responsibility for catalysis, and the others are concerned with ancillary functions, such as priming, processivity, or proofreading. These enzymes all replicate DNA with high fidelity, as does the slightly less complex mitochondrial enzyme. The repair polymerases have much simpler structures, often consisting of a single monomeric subunit (although it may function in the context of a complex of other repair enzymes). Of the enzymes involved in repair, only DNA polymerase β has a fidelity approaching the replicases: all of the others have much greater error rates.

Each of the three nuclear DNA replicases has a different function:

- DNA polymerase a initiates the synthesis of new strands.
- DNA polymerase 8 elongates the leading strand.
- DNA polymerase e may be involved in lagging strand synthesis, but also has other roles.

DNA polymerase a is unusual because it has the ability to initiate a new strand. It is used to initiate both the leading and lagging strands. The enzyme exists as a complex consisting of a 180 kD catalytic subunit, associated with the B subunit that appears necessary for assembly, and two smaller proteins that provide a primase activity. Reflecting its dual capacity to prime and extend chains, it is sometimes called pol a/primase.

The pol a/primase enzyme binds to the initiation complex at the origin and synthesizes a short strand consisting of ~10 bases of RNA followed by 20-30 bases of DNA (sometimes called iDNA). Then it is replaced by an enzyme that will extend the chain. On the leading strand, this is DNA polymerase 8. This event is called the pol switch. It involves interactions among several components of the initiation complex.

DNA polymerase 8 is a highly processive enzyme that continuously synthesizes the leading strand. Its processivity results from its interaction with two other proteins, RF-C and PCNA.

The roles of RF-C and PCNA are analogous to the *E. coli* γ clamp loader and β processivity unit (see *14.11 The clamp controls association of core enzyme with DNA)*. RF-C is a clamp loader that catalyzes the loading of PCNA on to DNA. It binds to the 3' end of the iDNA and uses ATP-hydrolysis to open the ring of PCNA so that it can encircle the DNA. The processivity of DNA polymerase 8 is maintained by PCNA, which tethers DNA polymerase 8 to the template. (PCNA is called proliferating cell nuclear antigen for historical reasons.) The crystal structure of PCNA closely resembles the *E. coli* β subunit: a trimer forms a ring that surrounds the DNA. Although the sequence and subunit organization are different from the dimeric β clamp, the function is likely to be similar.

We are less certain about events on the lagging strand. One possibility is that DNA polymerase 8 also elongates the lagging strand. It has the capability to dimerize, which suggests a model analogous to the behavior of *E. coli* replicase (see *14.10 DNA polymerase holoenzyme has 3 subcomplexes)*. However, there are some indications that DNA polymerase e may elongate the lagging strand, although it also has been identified with other roles.

A general model suggests that a replication fork contains 1 complex of DNA polymerase a/primase and two other DNA polymerase complexes. One is DNA polymerase 8 and the other is either a second DNA polymerase 8 or may possibly be a DNA polymerase e. The two complexes of DNA polymerase 8/e behave in the same way as the two

Eukaryotic DNA polymerases undertake either replication or repair

DNA polymerase	Function	Structure
	High fidelity replicases	
α	Nuclear replication	350 kD tetramer
δ	"	250 kD tetramer
ε	"	350 kD tetramer
γ	Mitochondrial replication	200 kD dimer
	High fidelity repair	
β	Base excision repair	39 kD monomer
	Low fidelity repair	
ζ	Thymine dimer bypass	heteromer
η	Base damage repair	monomer
ι	Required in meiosis	monomer
κ	Deletion and base substitution	monomer

Figure 14.24 Eukaryotic cells have many DNA polymerases. The replicative enzymes operate with high fidelity. Except for the β enzyme, the repair enzymes all have low fidelity. Replicative enzymes have large structures, with separate subunits for different activities. Repair enzymes have much simpler structures.

complexes of DNA polymerase III in the *E. coli* replisome: one synthesizes the leading strand, and the other synthesizes Okazaki fragments on the lagging strand. The exonuclease MF1 removes the RNA primers of Okazaki fragments. The enzyme DNA ligase I is specifically required to seal the nicks between the completed Okazaki fragments.

14.14 Phage T4 provides its own replication apparatus

: Key Concepts

Phage T4 provides its own replication apparatus, which consists of DNA polymerase, the gene *32* SSB, a helicase, a primase, and accessory proteins that increase speed and processivity.

When phage T4 takes over an *E. coli* cell, it provides several functions of its own that either replace or augment the host functions. The phage places little reliance on expression of host functions. The degradation of host DNA is important in releasing nucleotides that are reused in the synthesis of phage DNA. (The phage DNA differs in base composition from cellular DNA in using hydroxymethylcytosine instead of the customary cytosine.)

The phage-coded functions concerned with DNA synthesis in the infected cell can be identified by mutations that impede the production of mature phages. Essential phage functions are identified by conditional lethal mutations, which fall into three phenotypic classes:

- Those in which there is no DNA synthesis at all identify genes whose products either are components of the replication apparatus or are involved in the provision of precursors (especially the hydroxymethylcytosine).
- Those in which the onset of DNA synthesis is delayed are concerned with the initiation of replication.
- Those in which DNA synthesis starts but then is arrested include regulatory functions, the DNA ligase, and some of the enzymes concerned with host DNA degradation.
- There are also nonessential genes concerned with replication; for example, including those involved in glucosylating the hydroxymethylcytosine in the DNA.

Synthesis of T4 DNA is catalyzed by a multienzyme aggregate assembled from the products of a small group of essential genes.

The gene *32* protein (gp32) is a highly cooperative single-strand binding protein, needed in stoichiometric amounts. It was the first example of its type to be characterized. The geometry of the T4 replication fork may specifically require the phage-coded protein, since the *E. coli* SSB cannot substitute. The gp32 forms a complex with the T4 DNA polymerase; this interaction could be important in constructing the replication fork.

The T4 system uses an RNA priming event that is similar to that of its host. With single-stranded T4 DNA as template, the gene *41* and *61* products act together to synthesize short primers. Their behavior is analogous to that of DnaB and DnaG in *E. coli*. The gene *41* protein is the counterpart to DnaB. It is a hexameric helicase that uses hydrolysis of GTP to provide the energy to unwind DNA. The p41/p61 complex moves processively in the 5′-3′ direction in lagging strand synthesis, periodically initiating Okazaki fragments. Another protein, the product of gene *59*, loads the p41/p61 complex on to DNA; it is required to displace the p32 protein in order to allow the helicase to assemble on DNA.

By Book_Crazy [IND]

The gene *61* protein is needed in much smaller amounts than most of the T4 replication proteins. There are as few as 10 copies of gp61 per cell. (This impeded its characterization. It is required in such small amounts that originally it was missed as a necessary component, because enough was present as a contaminant of the gp32 preparation!) Gene *61* protein has the primase activity, analogous to DnaG of *E. coli*. The primase recognizes the template sequence 3'-TTG-5' and synthesizes pentaribonucleotide primers that have the general sequence pppApCpNpNpNp. If the complete replication apparatus is present, these primers are extended into DNA chains.

The gene *43* DNA polymerase has the usual 5′-3′ synthetic activity, associated with a 3′-5′ exonuclease proofreading activity. It catalyzes DNA synthesis and removes the primers. When T4 DNA polymerase uses a single-stranded DNA as template, its rate of progress is uneven. The enzyme moves rapidly through single-stranded regions, but proceeds much more slowly through regions that have a base-paired intrastrand secondary structure. The accessory proteins assist the DNA polymerase in passing these roadblocks, and maintaining its speed.

The remaining three proteins are referred to as "polymerase accessory proteins". They increase the affinity of the DNA polymerase for the DNA, and also its processivity and speed. The gene *45* product is a trimer that acts as a sliding clamp. The structure of the trimer is similar to that of the *E. coli* β dimer, forming a circle around DNA that holds the DNA polymerase subunit more tightly on the template.

The products of genes *44* and *62* form a tight complex, which has ATPase activity. They are the equivalent of the γδ clamp loader complex, and their role is to load p45 on to DNA. Four molecules of ATP are hydrolyzed in loading the p45 clamp and the p43 DNA polymerase on to DNA.

The overall structure of the replisome is similar to that of *E. coli*. It consists of two coupled holoenzyme complexes, one synthesizing the leading strand and the other synthesizing the lagging strand. In this case, the dimerization involves a direct interaction between the p43 DNA polymerase subunits, and p32 plays a role in coordinating the actions of the two DNA polymerase units.

We have dealt with DNA replication so far solely in terms of the progression of the replication fork. The need for other functions is shown by the DNA-delay and DNA-arrest mutants. The four genes of the DNA-delay mutants include *39, 52,* and *60,* which code for the three subunits of T4 topoisomerase II, an activity needed for removing supercoils in the template (see *15.13 Topoisomerases relax or introduce supercoils in DNA)*. The essential role of this enzyme suggests that T4 DNA does not remain in a linear form, but becomes topologically constrained during some stage of replication. The topoisomerase could be needed to allow rotation of DNA ahead of the replication fork.

Comparison of the T4 apparatus with the *E. coli* apparatus suggests that DNA replication poses a set of problems that are solved in analogous ways in different systems. We may now compare the enzymatic and structural activities found at the replication fork in *E. coli,* T4, and HeLa (human) cells. **Figure 14.25** summarizes the functions and assigns them to individual proteins. We can interpret the known properties of replication complex proteins in terms of similar functions, involving the unwinding, priming, catalytic, and sealing reactions. The components of each system interact in restricted ways, as shown by the fact that phage T4 requires its own helicase, primase, clamp, etc., and the bacterial proteins cannot substitute for their phage counterparts.

Replication requires a common set of functions

Function	*E.coli*	HeLa/SV40	Phage T4
Helicase	*DnaB*	T antigen	41
Loading helicase/primase	*DnaC*	T antigen	59
Single strand maintenance	*SSB*	RPA	32
Priming	*DnaG*	Polα/primase	61
Sliding clamp	β	PCNA	45
Clamp loading (ATPase)	γδ complex	RFC	44/62
Catalysis	*Pol III core*	Pol δ	43
Holoenzyme dimerization	τ	?	43
RNA removal	*Pol I*	MF1	43
Ligation	*Ligase*	Ligase I	T4 ligase

Figure 14.25 Similar functions are required at all replication forks.

14.15 Creating the replication forks at an origin

Key Concepts

- Initiation at *oriC* requires the sequential assembly of a large protein complex.
- DnaA binds to short repeated sequences and forms an oligomeric complex that melts DNA.
- 6 DnaC monomers bind each hexamer of DnaB and this complex binds to the origin.
- A hexamer of DnaB forms the replication fork. Gyrase and SSB are also required.

Starting a cycle of replication of duplex DNA requires several successive activities:

- The two strands of DNA must suffer their initial separation. This is in effect a melting reaction over a short region.
- An unwinding point begins to move along the DNA; this marks the generation of the replication fork, which continues to move during elongation.
- The first nucleotides of the new chain must be synthesized into the primer. This action is required once for the leading strand, but is repeated at the start of each Okazaki fragment on the lagging strand.

Some events that are required for initiation therefore occur uniquely at the origin; others recur with the initiation of each Okazaki fragment during the elongation phase.

Plasmids carrying the *E. coli oriC* sequence have been used to develop a cell-free system for replication. Initiation of replication at *oriC in vitro* starts with formation of a complex that requires six proteins: DnaA, DnaB, DnaC, HU, Gyrase, and SSB. Of the six proteins involved in prepriming, DnaA draws our attention as the only one uniquely involved in initiation vis-a-vis elongation. DnaB/DnaC provides the "engine" of initiation at the origin.

The first stage in complex formation is binding to *oriC* by DnaA protein. The reaction involves action at two types of sequences: 9 bp and 13 bp repeats. Together the 9 bp and 13 bp repeats define the limits of the 245 bp minimal origin, as indicated in **Figure 14.26.** An origin is activated by the sequence of events summarized in **Figure 14.27,** in which binding of DnaA is succeeded by association with the other proteins.

The four 9 bp consensus sequences on the right side of *oriC* provide the initial binding sites for DnaA. It binds cooperatively to form a central core around which *oriC* DNA is wrapped. Then DnaA acts at three A-T-rich 13 bp tandem repeats located in the left side of *oriC*. In the presence of ATP, DnaA melts the DNA strands at each of these sites to form an open complex. All three 13 bp repeats must be opened for the reaction to proceed to the next stage.

Altogether, 2-4 monomers of DnaA bind at the origin, and they recruit 2 "prepriming" complexes of DnaB-DnaC to bind, so that there is one for each of the two (bidirectional) replication forks. Each DnaB-DnaC complex consists of 6 DnaC monomers bound to a hexamer of DnaB. Each DnaB-DnaC complex transfers a hexamer of DnaB to an opposite strand of DNA. DnaC hydrolyzes ATP in order to release DnaB.

The prepriming complex generates a protein aggregate of 480 kD, corresponding to a sphere of radius 6 nm. The formation of a complex at *oriC* is detectable in the form of the large protein blob visualized in

Figure 14.26 The minimal origin is defined by the distance between the outside members of the 13-mer and 9-mer repeats.

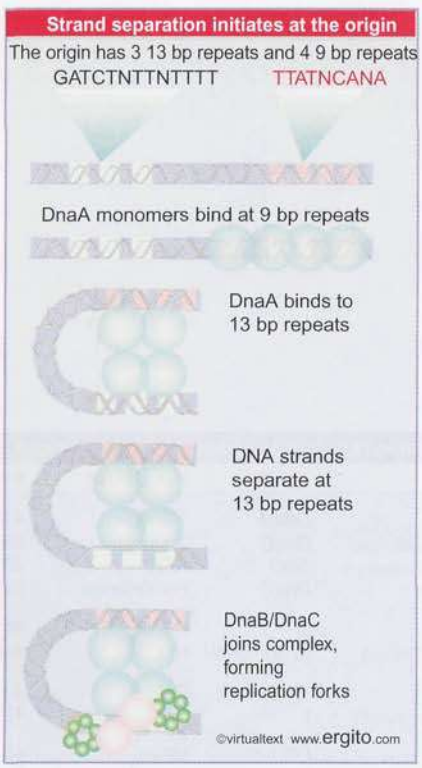


Figure 14.27 Prepriming involves formation of a complex by sequential association of proteins, leading to the separation of DNA strands.

Figure 14.28. When replication begins, a replication bubble becomes visible next to the blob.

The region of strand separation in the open complex is large enough for both DnaB hexamers to bind, initiating the two replication forks. As DnaB binds, it displaces DnaA from the 13 bp repeats, and extends the length of the open region. Then it uses its helicase activity to extend the region of unwinding. Each DnaB activates a DnaG primase, in one case to initiate the leading strand, and in the other to initiate the first Okazaki fragment of the lagging strand.

Two further proteins are required to support the unwinding reaction. Gyrase provides a swivel that allows one strand to rotate around the other (a reaction discussed in more detail in *15.15 Gyrase functions by coil inversion);* without this reaction, unwinding would generate torsional strain in the DNA. The protein SSB stabilizes the single-stranded DNA as it is formed. The length of duplex DNA that usually is unwound to initiate replication is probably <60 bp.

The protein HU is a general DNA-binding protein in *E. coli* (see *18 Rearrangement of DNA).* Its presence is not absolutely required to initiate replication *in vitro,* but it stimulates the reaction. HU has the capacity to bend DNA, and is involved in building the structure that leads to formation of the open complex.

Input of energy in the form of ATP is required at several stages for the prepriming reaction. It is required for unwinding DNA. The helicase action of DnaB depends on ATP hydrolysis; and the swivel action of gyrase requires ATP hydrolysis. ATP is also needed for the action of primase and to activate DNA polymerase **III.**

Following generation of a replication fork as indicated in Figure 14.27, the priming reaction occurs to generate a leading strand. We know that synthesis of RNA is used for the priming event, but the details of the reaction are not known. Some mutations in DnaA can be suppressed by mutations in RNA polymerase, which suggests that DnaA could be involved in an initiation step requiring RNA synthesis *in vivo.*

RNA polymerase could be required to read into the origin from adjacent transcription units; by terminating at sites in the origin, it could provide the 3'-OH ends that prime DNA polymerase **III.** (An example is provided by the use of D loops at mitochondrial origins, as discussed in *13.7 D loops maintain mitochondrial origins.)* Alternatively, the act of transcription could be associated with a structural change that assists initiation. This latter idea is supported by observations that transcription does not have to proceed into the origin; it is effective up to 200 bp away from the origin, and can use either strand of DNA as template *in vitro.* The transcriptional event is inversely related to the requirement for supercoiling *in vitro,* which suggests that it acts by changing the local DNA structure so as to aid melting of DNA.

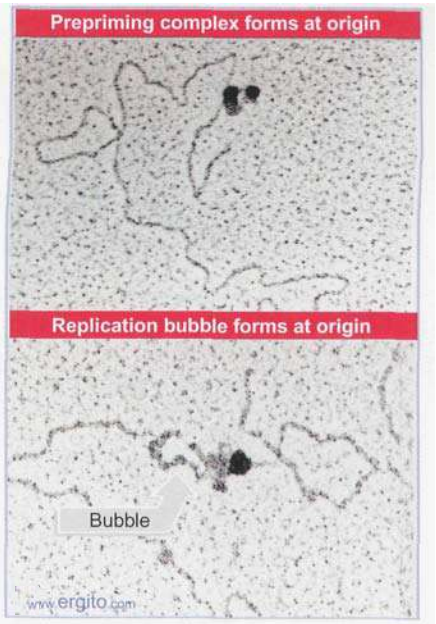


Figure 14.28 The complex at *oriC* can be detected by electron microscopy. Both complexes were visualized with antibodies against DnaB protein. Photographs kindly provided by Barbara Funnell.

14.16 Common events in priming replication at the origin

Key Concepts

- The general principle of bacterial initiation is that the origin is initially recognized by a protein that forms a large complex with DNA.
- A short region of A·T-rich DNA is melted.
- DnaB is bound to the complex and creates the replication fork.

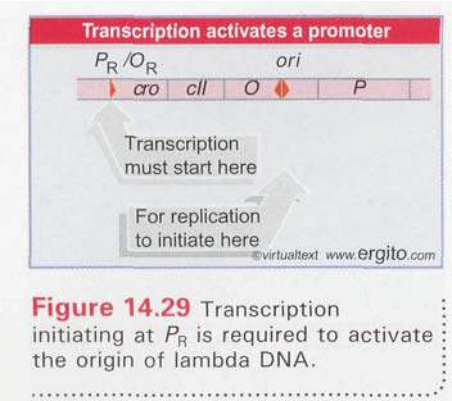


Figure 14.29 Transcription initiating at P_R is required to activate the origin of lambda DNA.

Another system for investigating interactions at the origin is provided by phage lambda. A map of the region is shown in **Figure 14.29.** Initiation of replication at the lambda origin requires "activation" by transcription starting from P_R. As with the events at *oriC*, this does not necessarily imply that the RNA provides a primer for the leading strand. Analogies between the systems suggest that RNA synthesis could be involved in promoting some structural change in the region.

Initiation requires the products of phage genes *O* and *P*, as well as several host functions. The phage O protein binds to the lambda origin; the phage P protein interacts with the O protein and with the bacterial proteins. The origin lies within gene 0, so the protein acts close to its site of synthesis.

Variants of the phage called *Xdv* consist of shorter genomes that carry all the information needed to replicate, but lack infective functions. *Xdv* DNA survives in the bacterium as a plasmid, and can be replicated *in vitro* by a system consisting of the phage-coded proteins O and P together with bacterial replication functions.

Lambda proteins O and P form a complex together with DnaB at the lambda origin, *orik*. The origin consists of two regions; as illustrated in **Figure 14.30,** a series of four binding sites for the O protein is adjacent to an A-T-rich region.

The first stage in initiation is the binding of O to generate a roughly spherical structure of diameter ~11 nm, sometimes called the O-some. The O-some contains ~100 bp or 60 kD of DNA. There are four 18 bp binding sites for O protein, which is ~34 kD. Each site is palindromic, and probably binds a symmetrical O dimer. The DNA sequences of the O-binding sites appear to be bent, and binding of O protein induces further bending.

If the DNA is supercoiled, binding of O protein causes a structural change in the origin. The A-T-rich region immediately adjacent to the O-binding sites becomes susceptible to S1 nuclease, an enzyme that specifically recognizes unpaired DNA. This suggests that a melting reaction occurs next to the complex of O proteins.

The role of the O protein is analogous to that of DnaA at *oriC:* it prepares the origin for binding of DnaB. Lambda provides its own protein, P, which substitutes for DnaC, and brings DnaB to the origin. When lambda P protein and bacterial DnaB proteins are added, the complex becomes larger and asymmetrical. It includes more DNA (a total of ~ 160 bp) as well as extra proteins. The *X* P protein has a special role: it inhibits the helicase action of DnaB. Replication fork movement is triggered when P protein is released from the complex. Priming and DNA synthesis follow.

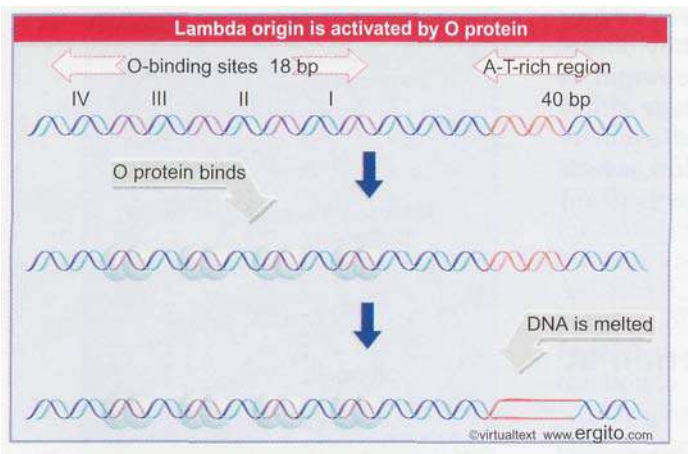


Figure 14.30 The lambda origin for replication comprises two regions. Early events are catalyzed by O protein, which binds to a series of 4 sites; then DNA is melted in the adjacent A-T-rich region. Although the DNA is drawn as a straight duplex, it is actually bent at the origin.

Some proteins are essential for replication without being directly involved in DNA synthesis as such. Interesting examples are provided by the DnaK and DnaJ proteins. DnaK is a chaperone, related to a common stress protein of eukaryotes. Its ability to interact with other proteins in a conformation-dependent manner plays a role in many cellular activities, including replication. The role of DnaK/DnaJ may be to disassemble the pre-priming complex; by causing the release of P protein, they allow replication to begin.

The initiation reactions at *oriC* and oriλ are similar. The same stages are involved, and rely upon overlapping components. The first step is recognition of the origin by a protein that binds to form a complex with the DNA, DnaA for *oriC* and O protein for *oriX*. A short region of A-T-rich DNA is melted. Then DnaB is loaded; this requires different functions at *oriC* and *oriX* (and yet other proteins are required for this stage at other origins). When the helicase DnaB joins

the complex, a replication fork is created. Finally an RNA primer is synthesized, after which replication begins.

The use of *oriC* and *oriλ* provides a general model for activation of origins. A similar series of events occurs at the origin of the virus SV40 in mammalian cells. Two hexamers of T antigen, a protein coded by the virus, bind to a series of repeated sites in DNA. In the presence of ATP, changes in DNA structure occur, culminating in a melting reaction. In the case of SV40, the melted region is rather short and is not A·T-rich, but it has an unusual composition in which one strand consists almost exclusively of pyrimidines and the other of purines. Near this site is another essential region, consisting of A·T base pairs, at which the DNA is bent; it is underwound by the binding of T antigen. An interesting difference from the prokaryotic systems is that T antigen itself possesses the helicase activity needed to extend unwinding, so that an equivalent for DnaB is not needed.

14.17 The primosome is needed to restart replication

Key Concepts

- Initiation of ϕX replication requires the primosome complex to displace SSB from the origin.
- A replication fork stalls when it arrives at damaged DNA.
- After the damage has been repaired, the primosome is required to reinitiate replication.
- The Tus protein binds to *ter* sites and stops DnaB from unwinding DNA, which causes replication to terminate.

Early work on replication made extensive use of phage ϕX174, and led to the discovery of a complex system for priming. ϕX174 DNA is not by itself a substrate for the replication apparatus, because the naked DNA does not provide a suitable template. But once the single-stranded form has been coated with SSB, replication can proceed. A primosome assembles at a unique site on the single-stranded DNA, called the assembly site *(pas)*. *The pas* is the equivalent of an origin for synthesis of the complementary strand of ϕX174. The primosome consists of six proteins: PriA, PriB, PriC, DnaT, DnaB, and DnaC. The key event in localizing the primosome is the ability of PriA to displace SSB from single-stranded DNA.

Although the primosome forms initially at the *pas* on ϕX174 DNA, primers are initiated at a variety of sites. PriA translocates along the DNA, displacing SSB, to reach additional sites at which priming occurs. As in *oriC* replicons, DnaB plays a key role in unwinding and priming in ϕX replicons. The role of PriA is to load DnaB to form a replication fork.

It has always been puzzling that ϕX origins should use a complex structure that is not required to replicate the bacterial chromosome. Why does the bacterium provide this complex?

The answer is provided by the fate of stalled replication forks. Figure 14.31 compares an advancing replication fork with what happens when there is damage to a base in the DNA or a nick in one strand. In either case, DNA synthesis is halted, and the replication fork is either stalled or disrupted. It is not clear whether the components of the fork remain associated with the DNA or disassemble. Replication fork stalling appears to be quite common; estimates for the frequency in *E. coli* suggest that 18-50% of bacteria encounter a problem during a replication cycle.

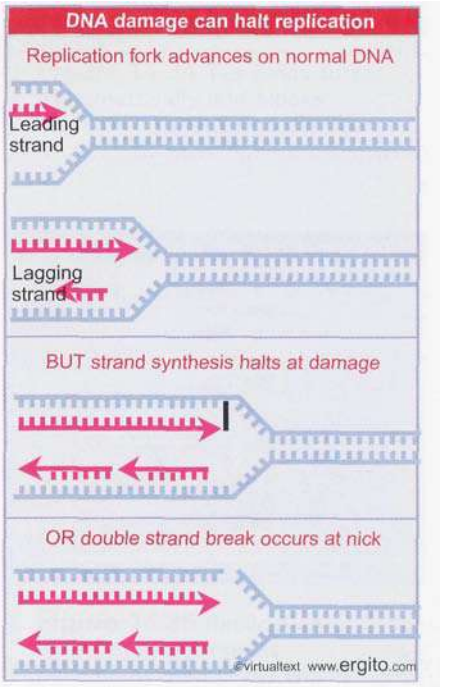


Figure 14.31 Replication is halted by a damaged base or nick in DNA.

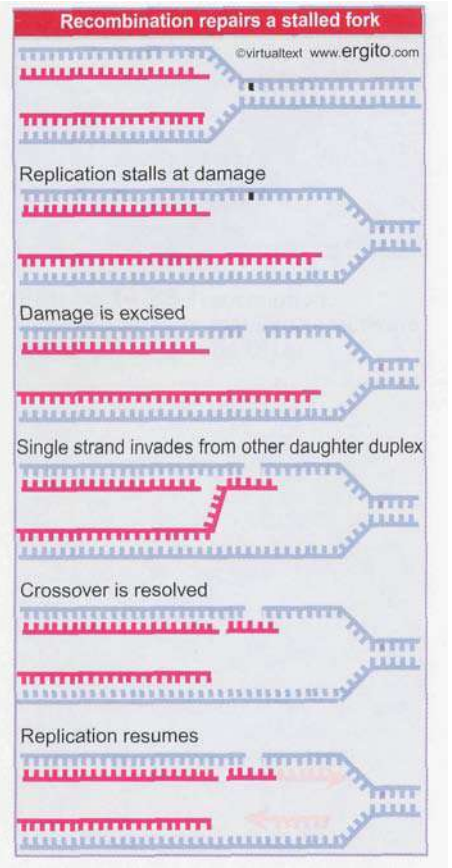


Figure 14.32 When replication halts at damaged DNA, the damaged sequence is excised, and the complementary (newly synthesized) strand of the other daughter duplex crosses over to repair the gap. Replication can now resume, and the gaps are filled in.

The situation is rescued by a recombination event that excises and replaces the damage or provides a new duplex to replace the region containing the double-strand break. The principle of the repair event is to use the built in redundancy of information between the two DNA strands. **Figure 14.32** shows the key events in such a repair event. Basically, information from the undamaged DNA daughter duplex is used to repair the damaged sequence. This creates a typical recombination-junction that is resolved by the same systems that perform homologous recombination. In fact, one view is that the major importance of these systems for the cell is in repairing damaged DNA at stalled replication forks.

After this the damage has been repaired, the replication fork must be restarted. **Figure 14.33** shows that this may be accomplished by assembly of the primosome, which in effect reloads DnaB so that helicase action can continue.

Replication fork reactivation is a common (and therefore important) reaction. It may be required in most chromosomal replication cycles. It is impeded by mutations in either the retrieval systems that replace the damaged DNA or in the components of the primosome.

Replication forks must stop and disassemble at the termination of replication. How is this accomplished?

Sequences that stop movement of replication forks have been identified in the form of the *ter* elements of the *E. coli* chromosome (see Figure 13.7) or equivalent sequences in some plasmids. The common feature of these elements is a 23 bp consensus sequence that provides the binding site for the product of the *tus* gene, a 36 kD protein that is necessary for termination. Tus binds to the consensus sequence, where it provides a contra-helicase activity and stops DnaB from unwinding DNA. The leading strand continues to be synthesized right up to the *ter* element, while the nearest lagging strand is initiated 50-100 bp before reaching *ter*.

The result of this inhibition is to halt movement of the replication fork and (presumably) to cause disassembly of the replication apparatus. **Figure 14.34** reminds us that Tus stops the movement of a replication fork in only one direction. The crystal structure of a Tus-ter complex shows that the Tus protein binds to DNA asymmetrically; α-helices of the protein protrude around the double helix at the end that blocks the replication fork. Presumably a fork proceeding in the opposite direction can displace Tus and thus continue. A difficulty in understanding the function of the system *in vivo* is that it appears to be dispensable, since mutations in the *ter* sites or in *tus* are not lethal.

14.18 Does methylation at the origin regulate initiation?

Key Concepts

- *oriC* contains 11 $\frac{GATC}{CTAG}$ repeats that are methylated on adenine on both strands.
- Replication generates hemimethylated DNA, which cannot initiate replication.
- There is a 13-minute delay before the $\frac{GATC}{CTAG}$ repeats are remethylated.

What feature of a bacterial (or plasmid) origin ensures that it is used to initiate replication only once per cycle? Is initiation associated with some change that marks the origin so that a replicated origin can be distinguished from a nonreplicated origin?

Some sequences that are used for this purpose are included in the origin. *oriC* contains 11 copies of the sequence $_{CTAG}$, which is a target

for methylation at the N^6 position of adenine by the Dam methylase. The reaction is illustrated in **Figure 14.35.**

Before replication, the palindromic target site is methylated on the adenines of each strand. Replication inserts the normal (nonmodified) bases into the daughter strands, generating hemimethylated DNA, in which one strand is methylated and one strand is unmethylated. So the replication event converts Dam target sites from fully methylated to hemimethylated condition.

What is the consequence for replication? The ability of a plasmid relying upon *oriC* to replicate in *dam*⁻ *E. coli* depends on its state of methylation. If the plasmid is methylated, it undergoes a single round of replication, and then the hemimethylated products accumulate, as described in **Figure 14.36.** So a hemimethylated origin cannot be used to initiate a replication cycle.

Two explanations suggest themselves. Initiation may require full methylation of the Dam target sites in the origin. Or initiation may be inhibited by hemimethylation of these sites. The latter seems to be the case, because an origin of nonmethylated DNA can function effectively.

So hemimethylated origins cannot initiate again until the Dam methylase has converted them into fully methylated origins. The GATC sites at the origin remain hemimethylated for ~13 minutes after replication. This long period is unusual, because at typical GATC sites elsewhere in the genome, remethylation begins immediately (<1.5 min) following replication. One other region behaves like *oriC;* the promoter of the *dnaA* gene also shows a delay before remethylation begins.

While it is hemimethylated, the *dnaA* promoter is repressed, which causes a reduction in the level of DnaA protein. So the origin itself is inert, and production of the crucial initiator protein is repressed, during this period.

14.19 Origins may be sequestered after replication

Key Concepts

- SeqA binds to hemimethylated DNA and is required for delaying rereplication.
- SeqA may interact with DnaA.
- While the origins are hemimethylated, they bind to the cell membrane, and may be unavailable to methylases.
- The nature of the connection between the origin and the membrane is still unclear.

What is responsible for the delay in remethylation at *oriC* and *dnaA?* The most likely explanation is that these regions are sequestered in a form in which they are inaccessible to the Dam methylase.

A circuit responsible for controlling reuse of origins is identified by mutations in the gene *seqA.* The mutants reduce the delay in remethylation at both *oriC* and *dnaA.* As a result, they initiate DNA replication too soon, thereby accumulating an excessive number of origins. This suggests that *seqA* is part of a negative regulatory circuit that prevents origins from being remethylated. SeqA binds to hemimethylated DNA more strongly than to fully methylated DNA. It may initiate binding when the DNA becomes hemimethylated, and then its continued presence prevents formation of an open complex at the origin. SeqA does not have specificity for the *oriC* sequence, and it seems likely that this

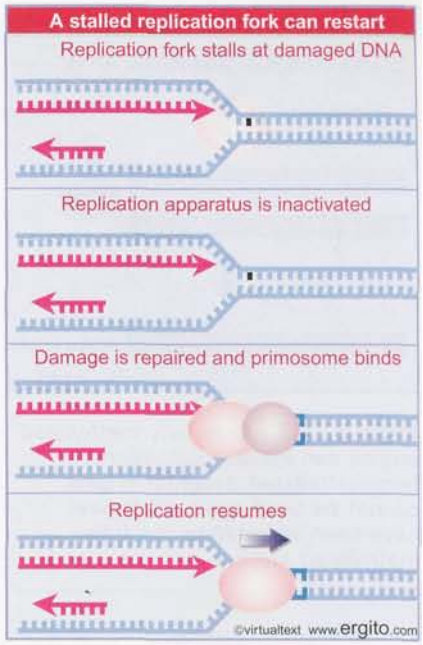


Figure 14.33 The primosome is required to restart a stalled replication fork after the DNA has been repaired.

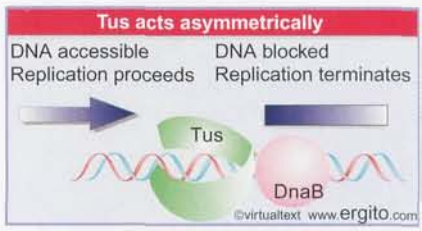


Figure 14.34 Tus binds to *ter* asymmetrically and blocks replication in only one direction.

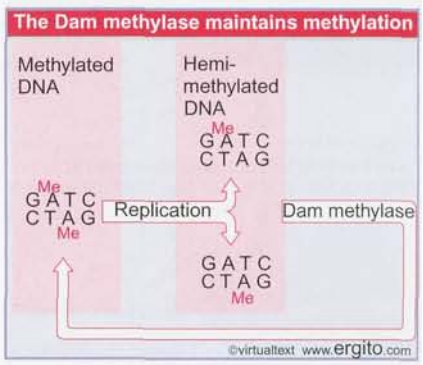


Figure 14.35 Replication of methylated DNA gives hemimethylated DNA, which maintains its state at GATC sites until the Dam methylase restores the fully methylated condition.

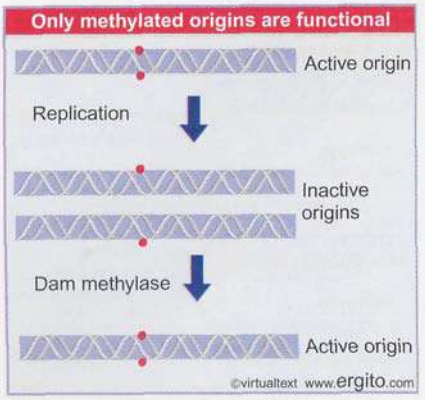


Figure 14.36 Only fully methylated origins can initiate replication; hemimethylated daughter origins cannot be used again until they have been restored to the fully methylated state.

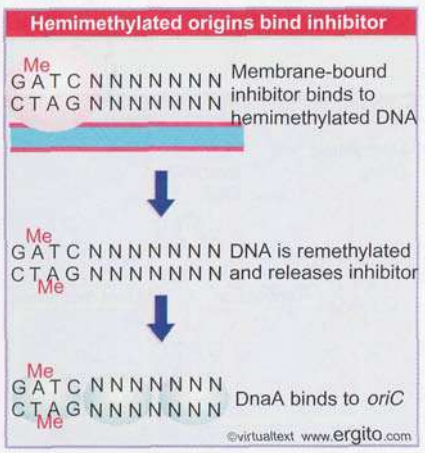


Figure 14.37 A membrane-bound inhibitor binds to hemimethylated DNA at the origin, and may function by preventing the binding of DnaA. It is released when the DNA is remethylated.

is conferred by DnaA protein, which would explain genetic interactions between *seqA* and *dnaA*.

Hemimethylation of the GATC sequences in the origin is required for its association with the cell membrane *in vitro*. Hemimethylated *oriC* DNA binds to the membranes, but DNA that is fully methylated does not bind. One possibility is that membrane association is involved in controlling the activity of the origin. This function could be separate from any role that the membrane plays in segregation (see Figure 13.26). Association with the membrane could prevent reinitiation from occurring prematurely, either indirectly because the origins are sequestered or directly because some component at the membrane inhibits the reaction.

The properties of the membrane fraction suggest that it includes components that regulate replication. An inhibitor is found in this fraction that competes with DnaA protein. This inhibitor can prevent initiation of replication only if it is added to an *in vitro* system before DnaA protein. This suggests the model of Figure 14.37, in which the inhibitor specifically recognizes hemimethylated DNA and prevents DnaA from binding. When the DNA is remethylated, the inhibitor is released, and DnaA now is free to initiate replication. If the inhibitor is associated with the membrane, then association and dissociation of DNA with the membrane may be involved in the control of replication.

The full scope of the system used to control reinitiation is not clear, but several mechanisms may be involved: physical sequestration of the origin; delay in remethylation; inhibition of DnaA binding; repression of DnaA transcription. It is not immediately obvious which of these events cause the others, and whether their effects on initiation are direct or indirect.

We still have to come to grips with the central issue of which feature has the basic responsibility for timing. One possibility is that attachment to the membrane occurs at initiation, and that assembly of some large structure is required to release the DNA. The period of sequestration appears to increase with the length of the cell cycle, which suggests that it directly reflects the clock that controls reinitiation.

As the only member of the replication apparatus uniquely required at the origin, DnaA has attracted much attention. DnaA is a target for several regulatory systems. It may be that no one of these systems by itself is adequate to control frequency of initiation, but the combination achieves the desired result. Some mutations in *dnaA* render replication asynchronous, suggesting that DnaA could be the "titrator" or "clock" that measures the number of origins relative to cell mass. Overproduction of DnaA yields conflicting results, varying from no effect to causing initiation to take place at reduced mass.

It has been difficult to identify the protein component(s) that mediate membrane-attachment. A hint that this is a function of DnaA is provided by its response to phospholipids. Phospholipids promote the exchange of ATP with ADP bound to DnaA. We do not know what role this plays in controlling the activity of DnaA (which requires ATP), but the reaction implies that DnaA is likely to interact with the membrane. This would imply that more than one event is involved in associating with the membrane. Perhaps a hemimethylated origin is bound by the membrane-associated inhibitor, but when the origin becomes fully methylated, the inhibitor is displaced by DnaA associated with the membrane.

If DnaA is the initiator that triggers a replication cycle, the key event will be its accumulation at the origin to a critical level. There are no cyclic variations in the overall concentration or expression of DnaA, which suggests that local events must be responsible. To be active in initiating replication, DnaA must be in the ATP-bound form. DnaA has a weak intrinsic activity that converts the ATP to ADP. This activity is enhanced by the β subunit of DNA polymerase III. When the replicase

is incorporated into the replication complex, this interaction causes hydrolysis of the ATP bound to DnaA, thereby inactivating DnaA, and preventing it from starting another replication cycle. This reaction has been called *RIDA* (regulatory inactivation of DnaA). It is enhanced by a protein called Hda. We do not yet know what controls the timing of the reactivation of DnaA.

Another factor that controls availability of DnaA at the origin is the competition for binding it to other sites on DNA. In particular, a locus called *dat* has a large concentration of DnaA-binding sites. It binds about 8× more DnaA than the origin. Deletion *of dat* causes initiation to occur more frequently. This significantly reduces the amount of DnaA available to the origin, but we do not yet understand exactly what role this may play in controlling the timing of initiation.

14.20 Licensing factor controls eukaryotic rereplication

Key Concepts

- Licensing factor is necessary for initiation of replication at each origin.
- It is present in the nucleus prior to replication, but is inactivated or destroyed by replication.
- Initiation of another replication cycle becomes possible only after licensing factor reenters the nucleus after mitosis.

A eukaryotic genome is divided into multiple replicons, and the origin in each replicon is activated once and only once in a single division cycle. This could be achieved by providing some rate-limiting component that functions only once at an origin or by the presence of a repressor that prevents rereplication at origins that have been used. The critical questions about the nature of this regulatory system are how the system determines whether any particular origin has been replicated, and what protein components are involved.

Insights into the nature of the protein components have been provided by using a system in which a substrate DNA undergoes only one cycle of replication. *Xenopus* eggs have all the components needed to replicate DNA—in the first few hours after fertilization they undertake 11 division cycles without new gene expression—and they can replicate the DNA in a nucleus that is injected into the egg. **Figure 14.38** summarizes the features of this system.

When a sperm or interphase nucleus is injected into the egg, its DNA is replicated only once (this can be followed by use of a density label, just like the original experiment that characterized semiconservative replication, shown previously in Figure 1.12). If protein synthesis is blocked in the egg, the membrane around the injected material remains intact, and the DNA cannot replicate again. However, in the presence of protein synthesis, the nuclear membrane breaks down just as it would for a normal cell division, and in this case subsequent replication cycles can occur. The same result can be achieved by using agents that permeabilize the nuclear membrane. This suggests that the nucleus contains a protein(s) needed for replication that is used up in some way by a replication cycle; although more of the protein is present in the egg cytoplasm, it can only enter the nucleus if the nuclear membrane breaks down. The system can in principle be taken further by developing an *in vitro* extract that supports nuclear replication, thus allowing the components of the extract to be isolated, and the relevant factors identified.

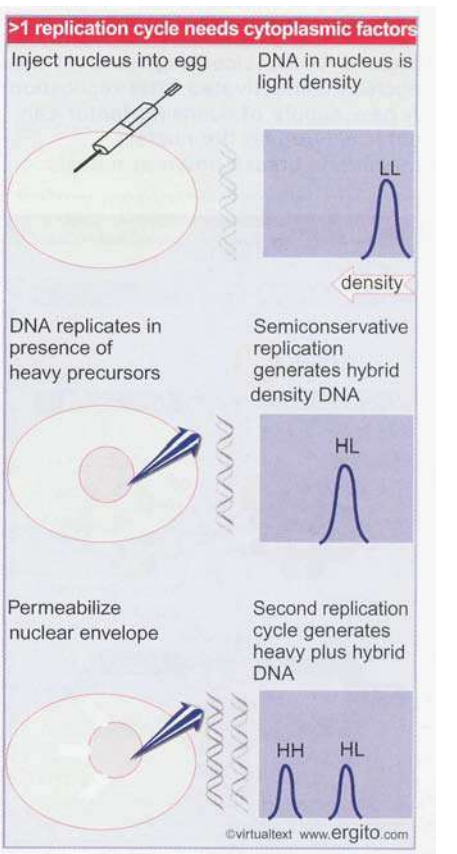


Figure 14.38 A nucleus injected into a *Xenopus* egg can replicate only once unless the nuclear membrane is permeabilized to allow subsequent replication cycles.

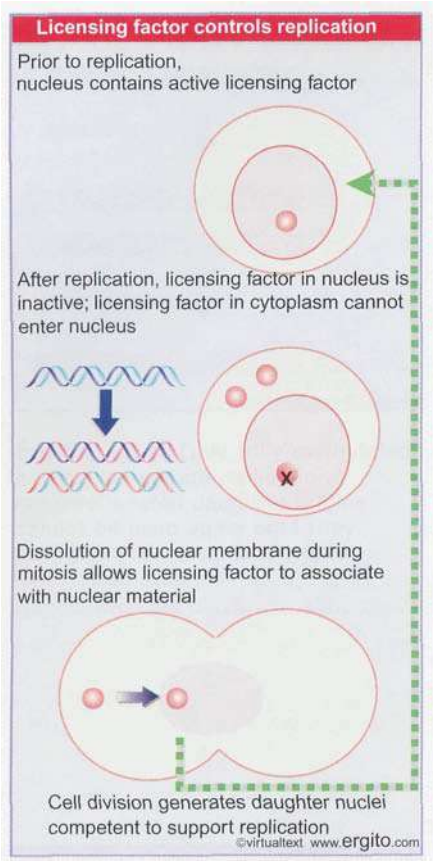


Figure 14.39 Licensing factor in the nucleus is inactivated after replication A new supply of licensing factor can enter only when the nuclear membrane breaks down at mitosis.

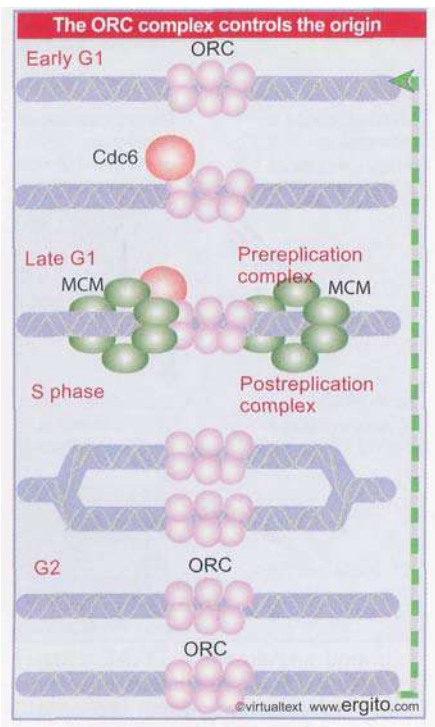


Figure 14.40 Proteins at the origin control susceptibility to initiation.

Figure 14.39 explains the control of reinitiation by proposing that this protein is a licensing factor. It is present in the nucleus prior to replication. One round of replication either inactivates or destroys the factor, and another round cannot occur until further factor is provided. Factor in the cytoplasm can gain access to the nuclear material only at the subsequent mitosis when the nuclear envelope breaks down. This regulatory system achieves two purposes. By removing a necessary component after replication, it prevents more than one cycle of replication from occurring. And it provides a feedback loop that makes the initiation of replication dependent on passing through cell division.

14.21 Licensing factor consists of MCM proteins

Key Concepts

- The ORC is a protein complex that is associated with yeast origins throughout the cell cycle.
- Cdc6 protein is an unstable protein that is synthesized only in G1.
- Cdc6 binds to ORC and allows MCM proteins to bind.
- When replication is initiated, Cdc6 and MCM proteins are displaced. The degradation of Cdc6 prevents reinitiation.
- Some MCM proteins are in the nucleus throughout the cycle, but others may enter only after mitosis.

The key event in controlling replication is the behavior of the ORC complex at the origin. Recall that ORC is a 400 kD complex that binds to the *S. cerevisiae ARS* sequence (see *13.6 Replication origins can be isolated in yeast).* The origin *(ARS)* consists of the A consensus sequence and three B elements (see Figure 13.10). The ORC complex of 6 proteins (all of which are coded by essential genes) binds to the A and adjacent B1 element. ATP is required for the binding, but is not hydrolyzed until some later stage. The transcription factor ABF1 binds to the B3 element; this assists initiation, but it is the events that occur at the A and B1 elements that actually cause initiation. Most origins are localized in regions between genes, which suggests that it may be important for the local chromatin structure to be in a nontranscribed condition.

The striking feature is that ORC remains bound at the origin through the entire cell cycle. However, changes occur in the pattern of protection of DNA as a result of the binding of other proteins to the ORC-origin complex. **Figure 14.40** summarizes the cycle of events at the origin.

At the end of the cell cycle, ORC is bound to A/B1, and generates a pattern of protection *in vivo* that is similar to that found when it binds to free DNA *in vitro.* Basically the region across A-B1 is protected against DNAase, but there is a hypersensitive site in the center of B1.

During G1, this pattern changes, most strikingly by the loss of the hypersensitive site. This is due to the binding of Cdc6 protein to the ORC. In yeast, Cdc6 is a highly unstable protein (half-life <5 minutes). It is synthesized during G1, and typically binds to the ORC between the exit from mitosis and late G1. Its rapid degradation means that no protein is available later in the cycle. In mammalian cells it is controlled differently; it is phosphorylated during S phase, and as a result is exported from the nucleus. Cdc6 provides the connection between ORC and a complex of proteins that is involved in licensing and initiation. Cdc6 has an ATPase activity that is required for it to support initiation.

By Book_Crazy [IND]

An insight into the system that controls availability of licensing factor is provided by certain mutants in yeast. Mutations in the licensing factor itself could prevent initiation of replication. This is how mutations behave in *MCM2, 3, 5*. Mutations in the system that inactivates licensing factor after the start of replication should allow the accumulation of excess quantities of DNA, because the continued presence of licensing factor allows rereplication to occur. Such mutations are found in genes that code for components of the ubiquitination system that is responsible for degrading certain proteins. This suggests that licensing factor may be destroyed after the start of a replication cycle.

The proteins MCM2,3,5 are required for replication and enter the nucleus only during mitosis. Homologues are found in animal cells, where MCM3 is bound to chromosomal material before replication, but is released after replication. The animal cell MCM2,3,5 complex remains in the nucleus throughout the cell cycle, suggesting that it may be one component of the licensing factor. Another component, able to enter only at mitosis, may be necessary for MCM2,3,5 to associate with chromosomal material.

In yeast, the presence of Cdc6 at the origin allows MCM proteins to bind to the complex. Their presence is necessary for initiation to occur at the origin. The origin therefore enters S phase in the condition of a prereplication complex, containing ORC, Cdc6, and MCM proteins. When initiation occurs, Cdc6 and MCM are displaced, returning the origin to the state of the postreplication complex, which contains only ORC. Because Cdc6 is rapidly degraded during S phase, it is not available to support reloading of MCM proteins, and so the origin cannot be used for a second cycle of initiation during the S phase.

If Cdc6 is made available to bind to the origin during G2 (by ectopic expression), MCM proteins do not bind until the following G1, suggesting that there is a secondary mechanism to ensure that they associate with origins only at the right time. This could be another part of licensing control. At least in *S. cerevisiae*, this control does not seem to be exercised at the level of nuclear entry, but this could be a difference between yeasts and animal cells. We discuss how the cell cycle control system regulates initiation (and reinitiation) of replication in *29 Cell cycle and growth regulation.*

The MCM2-7 proteins form a 6-member ring-shaped complex around DNA. Some of the ORC proteins have similarities to replication proteins that load DNA polymerase on to DNA. It is possible that ORC uses hydrolysis of ATP to load the MCM ring on to DNA. In *Xenopus* extracts, replication can be initiated if ORC is removed after it has loaded Cdc6 and MCM proteins. This shows that the major role of ORC is to identify the origin to the Cdc6 and MCM proteins that control initiation and licensing.

The MCM proteins are required for elongation as well as for initiation, and continue to function at the replication fork. Their exact role in elongation is not clear, but one possibility is that they contribute to the helicase activity that unwinds DNA. Another possibility is that they act as an advance guard that acts on chromatin in order to allow a helicase to act on DNA.

14.22 Summary

DNA synthesis occurs by semidiscontinuous replication, in which the leading strand of DNA growing 5'-3' is extended continuously, but the lagging strand that grows overall in the opposite 3'-5' direction is made as short Okazaki fragments, each synthe-

sized 5'-3'. The leading strand and each Okazaki fragment of the lagging strand initiate with an RNA primer that is extended by DNA polymerase. Bacteria and eukaryotes each possess more than one DNA polymerase activity. DNA polymerase III synthesizes both lagging and leading strands in *E. coli*. Many proteins are required for DNA polymerase III action and several constitute part of the replisome within which it functions.

The replisome contains an asymmetric dimer of DNA polymerase III; each new DNA strand is synthesized by a different core complex containing a catalytic (a) subunit. Processivity of the core complex is maintained by the β clamp, which forms a ring round DNA. The clamp is loaded on to DNA by the clamp loader complex. Clamp/clamp loader pairs with similar structural features are widely found in both prokaryotic and eukaryotic replication systems.

The looping model for the replication fork proposes that, as one half of the dimer advances to synthesize the leading strand, the other half of the dimer pulls DNA through as a single loop that provides the template for the lagging strand. The transition from completion of one Okazaki fragment to the start of the next requires the lagging strand catalytic subunit to dissociate from DNA and then to reattach to a β clamp at the priming site for the next Okazaki fragment.

DnaB provides the helicase activity at a replication fork; this depends on ATP cleavage. DnaB may function by itself in *oriC* replicons to provide primosome activity by interacting periodically with DnaG, which provides the primase that synthesizes RNA.

Phage T4 codes for a replication apparatus consisting of 7 proteins: DNA polymerase, helicase, single-strand binding protein, priming activities, and accessory proteins. Similar functions are required in other replication systems, including a HeLa cell system that replicates SV40 DNA. Different enzymes, DNA polymerase a and DNA polymerase 8, initiate and elongate the new strands of DNA.

The ϕX priming event also requires DnaB, DnaC, and DnaT. PriA is the component that defines the primosome assembly site *(pas)* for ϕX replicons; it displaces SSB from DNA in an action that involves cleavage of ATP. PriB and PriC are additional components of the primosome. The importance of the primosome for the bacterial cell is that it is used to restart replication at forks that stall when they encounter damaged DNA.

The common mode of origin activation involves an initial limited melting of the double helix, followed by more general unwinding to create single strands. Several proteins act sequentially at the *E. coli* origin. Replication is initiated at *oriC* in *E. coli* when DnaA binds to a series of 9 bp repeats. This is followed by binding to a series of 13 bp repeats, where it uses hydrolysis of ATP to generate the energy to separate the DNA strands. The pre-priming complex of DnaC-DnaB displaces DnaA. DnaC is released in a reaction that depends on ATP hydrolysis; DnaB is joined by the replicase enzyme, and replication is initiated by two forks that set out in opposite directions. Similar events occur at the lambda origin, where phage proteins 0 and P are the counterparts of bacterial proteins DnaA and DnaC, respectively. In SV40 replication, several of these activities are combined in the functions of T antigen.

The availability of DnaA at the origin is an important component of the system that determines when replication cycles should initiate. Following initiation of replication, DnaA hydrolyzes its ATP under the stimulus of the β sliding clamp, generating an inactive form of the protein. Also, *oriC* must compete with the *dat* site for binding DnaA.

Several sites that are methylated by the Dam methylase are present in the *E. coli* origin, including those of the 13-mer binding sites for DnaA. The origin remains hemimethylated and is in a sequestered state for ~10 minutes following initiation of a replication cycle. During this period it is associated with the membrane, and

reinitiation of replication is repressed. The protein SeqA is involved in sequestration and may interact with DnaA.

After cell division, nuclei of eukaryotic cells have a licensing factor that is needed to initiate replication. Its destruction after initiation of replication prevents further replication cycles from occurring in yeast. Licensing factor cannot be imported into the nucleus from the cytoplasm, and can be replaced only when the nuclear membrane breaks down during mitosis.

The origin is recognized by the ORC proteins, which in yeast remain bound throughout the cell cycle. The protein Cdc6 is available only at S phase. In yeast it is synthesized during S phase and rapidly degraded. In animal cells it is synthesized continuously, but is exported from the nucleus during S phase. The presence of Cdc6 allows the MCM proteins to bind to the origin. The MCM proteins are required for initiation. The action of Cdc6 and the MCM proteins provides the licensing function.

References

14.1 Introduction

ref Hirota, Y., Ryter, A., and Jacob, F. (1968). Thermosensitive mutants of *E. coli* affected in the processes of DNA synthesis and cellular division. Cold Spring Harbor Symp. Quant. Biol. 33, 677-693.

14.5 DNA polymerases have a common structure

rev Johnson, K. A. (1993). Conformational coupling in DNA polymerase fidelity. Ann. Rev. Biochem. 62, 685-713.

Joyce, C. M. and Steitz, T. A. (1994). Function and structure relationships in DNA polymerases. Ann. Rev. Biochem. 63, 777-822.

Hubscher, U., Maga, G., and Spadari, S. (2002). Eukaryotic DNA polymerases. Ann. Rev. Biochem. 71, 133-163.

ref Shamoo, Y. and Steitz, T. A. (1999). Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. Cell 99, 155-166.

14.7 The ϕX model system shows how single-stranded DNA is generated for replication

ref Dillingham, M. S., Wigley, D. B., and Webb, M. R. (2000). Demonstration of unidirectional single-stranded DNA translocation by PcrA helicase: measurement of step size and translocation speed. Biochemistry 39, 205 212.

Singleton, M. R., Sawaya, M. R., Ellenberger, T., and Wigley, D. B. (2000). Crystal structure of T7 gene 4 ring helicase indicates a mechanism for sequential hydrolysis of nucleotides. Cell 101, 589-600.

14.9 Coordinating synthesis of the lagging and leading strands

ref Dervyn, E., Suski, C, Daniel, R., Bruand, C, Chapuis, J., Errington, J., Janniere, L., and Ehrlich, S. D. (2001). Two essential DNA polymerases at the bacterial replication fork. Science 294, 1716-1719.

14.10 DNA polymerase holoenzyme has 3 subcomplexes

ref Studwell-Vaughan, P. S. and O'Donnell, M. (1991). Constitution of the twin polymerase of DNA polymerase III holoenzyme. J. Biol. Chem. 266, 19833-19841.

Stukenberg, P. T., Studwell-Vaughan, P. S., and O'Donnell, M. (1991). Mechanism of the sliding beta-clamp of DNA polymerase III holoenzyme. J. Biol. Chem. 266, 11328-11334.

14.11 The clamp controls association of core enzyme with DNA

rev Benkovic, S. J., Valentine, A. M., and Salinas, F. (2001). Replisome-mediated DNA replication. Ann. Rev. Biochem. 70, 181-208.

ref Davey, M. J., Jeruzalmi, D., Kuriyan, J., and O'Donnell, M. (2002). Motors and switches: AAA + machines within the replisome. Nat. Rev. Mol. Cell Biol. 3, 826-835.

Jeruzalmi, D., O'Donnell, M., and Kuriyan, J. (2001). Crystal structure of the processivity clamp loader gamma (gamma) complex of *E. coli* DNA polymerase III. Cell 106, 429-441.

Kong, X. P., Onrust, R., O'Donnell, M., and Kuriyan, J. (1992). Three-dimensional structure of the beta subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. Cell 69, 425-437.

14.13 Separate eukaryotic DNA polymerases undertake initiation and elongation

rev Goodman, M. F. (2002). Error-prone repair DNA polymerases in prokaryotes and eukaryotes. Ann. Rev. Biochem. 71, 17-50.

Hubscher, U., Maga, G., and Spadari, S. (2002). Eukaryotic DNA polymerases. Ann. Rev. Biochem. 71, 133-163.

ref Karthikeyan, R., Vonarx, E. J., Straffon, A. F., Simon, M., Faye, G., and Kunz, B. A. (2000). Evidence from mutational specificity studies that yeast DNA polymerases delta and epsilon replicate different DNA strands at an intracellular replication fork. J. Mol. Biol. 299, 405-419.

Shiomi, Y., Usukura, J., Masamura, Y., Takeyasu, K., Nakayama, Y., Obuse, C, Yoshikawa, H., and Tsurimoto, T. (2000). ATP-dependent structural change of the eukaryotic clamp-loader protein, replication factor C. Proc. Nat. Acad. Sci. USA 97, 14127-14132.

Waga, S., Masuda, T., Takisawa, H., and Sugino, A. (2001). DNA polymerase epsilon is required for coordinated and efficient chromosomal DNA replication in *Xenopus* egg extracts. Proc. Nat. Acad. Sci. USA 98, 4978-4983.
Zuo, S., Bermudez, V., Zhang, G., Kelman, Z., and Hurwitz, J. (2000). Structure and activity associated with multiple forms of *S. pombe* DNA polymerase delta. J. Biol. Chem. 275, 5153-5162.

14.14 Phage T4 provides its own replication apparatus

ref Ishmael, F. T., Alley, S. C, and Benkovic, S. J. (2002). Assembly of the bacteriophage T4 helicase: architecture and stoichiometry of the gp41-gp59 complex. J. Biol. Chem. 277, 20555-20562.
Salinas, F., and Benkovic, S. J. (2000). Characterization of bacteriophage T4-coordinated leading- and lagging-strand synthesis on a minicircle substrate. Proc. Nat. Acad. Sci. USA 97, 7196-7201.
Schrock, R. D. and Alberts, B. (1996). Processivity of the gene 41 DNA helicase at the bacteriophage T4 DNA replication fork. J. Biol. Chem. 271, 16678-16682.

14.15 Creating the replication forks at an origin

ref Bramhill, D. and Kornberg, A. (1988). Duplex opening by dnaA protein at novel sequences in initiation of replication at the origin of the *E. coli* chromosome. Cell 52, 743-755.
Fuller, R. S., Funnell, B. E., and Kornberg, A. (1984). The dnaA protein complex with the *E. coli* chromosomal replication origin (oriC) and other DNA sites. Cell 38, 889-900.
Funnell, B. E. and Baker, T. A. (1987). *In vitro* assembly of a prepriming complex at the origin of the *E. coli* chromosome. J. Biol. Chem. 262, 10327-10334.
Sekimizu, K, Bramhill, D., and Kornberg, A. (1987). ATP activates dnaA protein in initiating replication of plasmids bearing the origin of the *E. coli* chromosome. Cell 50, 259-265.
Wahle, E., Lasken, R. S., and Kornberg, A. (1989). The dnaB-dnaC replication protein complex of Escherichia coli. II. Role of the complex in mobilizing dnaB functions. J. Biol. Chem. 264, 2469-2475.

14.17 The primosome is needed to restart replication

rev Cox, M. M. (2001). Recombinational DNA repair of damaged replication forks in *E. coli:* questions. Ann. Rev. Genet. 35, 53-82.
ref Cox, M. M., Goodman, M. F., Kreuzer, K. N., Sherratt, D. J., Sandler, S. J., and Marians, K. J. (2000). The importance of repairing stalled replication forks. Nature 404, 37-41.
Kuzminov, A. (1995). Collapse and repair of replication forks in *E. coli.* Mol. Microbiol. 16, 373-384.
McGlynn, P. and Lloyd, R. G. (2002). Recombinational repair and restart of damaged replication forks. Nat. Rev. Mol. Cell Biol. 3, 859-870.
Seigneur, M., Bidnenko, V., Ehrlich, S. D., and Michel, B. (1998). RuvAB acts at arrested replication forks. Cell 95, 419-430.

14.18 Does methylation at the origin regulate initiation?

ref Campbell, J. L. and Kleckner, N. (1990). E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell 62, 967-979.

14.19 Origins may be sequestered after replication

ref Katayama, T., Kurokawa, K., Crooke, E., and Sekimizu, K. (1998). The initiator function of DnaA protein is negatively regulated by the sliding clamp of the *E. coli* chromosomal replicase. Cell 94, 61-71.
Kato, J. and Katayama, T. (2001). Hda, a novel DnaA-related protein, regulates the replication cycle in *Escherichia coli.* EMBO J. 20, 4253-4262.
Kitagawa, R., Ozaki, T., Moriya, S., and Ogawa, T. (1998). Negative control of replication initiation by a novel chromosomal locus exhibiting exceptional affinity for *E. coli* DnaA protein. Genes Dev. 12, 3032-3043.
Kurokawa, K., Nishida, S., Emoto, A., Sekimizu, K., and Katayama, T. (1999). Replication cycle-coordinated change of the adenine nucleotide-bound forms of DnaA protein in *Escherichia coli.* EMBO J. 18, 6642-6652.
Lu, M., Campbell, J. L., Boye, E., and Kleckner, N. (1994). SeqA: a negative modulator of replication initiation in *E. coli.* Cell 77, 413-426.
Slater, S., Wold, S., Lu, M., Boye, E., Skarstad, K., and Kleckner, N. (1995). *E. coli* SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. Cell 82, 927-936.
Wold, S., Boye, S., Wold, S., Slater, S., Kleckner, N., and Skarstad, K. (1998). Effects of purified SeqA protein on *oriC*-dependent DNA replication *in vitro.* EMBO J. 17, 4158-4165.

14.20 Licensing factor controls eukaryotic rereplication

ref Blow, J. J. and Laskey, R. A. (1988). A role for the nuclear envelope in controlling DNA replication within the cell cycle. Nature 332, 546-548.

14.21 Licensing factor consists of MCM proteins

rev Bell, S. P. (2002). The origin recognition complex: from simple origins to complex functions. Genes Dev. 16, 659-672.
Bell, S. P. and Dutta, A. (2002). DNA replication in eukaryotic cells. Ann. Rev. Biochem. 71, 333-374.
Dutta, A. and Bell, S. P. (1997). Initiation of DNA replication in eukaryotic cells. Ann. Rev. Cell Dev. Biol. 13, 293-332.
ref Bell, S. P. and Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature 357, 128-134.
Diffley, J. F., Cocker, J. H., Dowell, S. J., and Rowley, A. (1994). Two steps in the assembly of complexes at yeast replication origins in vivo. Cell 78, 303-316.
Liang, C. and Stillman, B. (1997). Persistent initiation of DNA replication and chromatin-bound MCM proteins during the cell cycle in cdc6 mutants. Genes Dev. 11, 3375-3386.

Perkins, G. and Diffley, J. F. (1998). Nucleotide-dependent prereplicative complex assembly by Cdc6p, a homolog of eukaryotic and prokaryotic clamp-loaders. Mol. Cell 2, 23-32.

Rowles, A., Tada, S., and Blow, J. J. (1999). Changes in association of the *Xenopus* origin recognition complex with chromatin on licensing of replication origins. J. Cell. Sci. 112, 2011-2018.

Weinreich, M., Liang, C, and Stillman, B. (1999). The Cdc6p nucleotide-binding motif is required for loading mcm proteins onto chromatin. Proc. Nat. Acad. Sci. USA 96, 441-446.

Recombination and repair

15.1 Introduction

Evolution could not happen without genetic recombination. If it were not possible to exchange material between (homologous) chromosomes, the content of each individual chromosome would be irretrievably fixed in its particular alleles. When mutations occurred, it would not be possible to separate favorable and unfavorable changes. The length of the target for mutation damage would effectively be increased from the gene to the chromosome. Ultimately a chromosome would accumulate so many deleterious mutations that it would fail to function.

By shuffling the genes, recombination allows favorable and unfavorable mutations to be separated and tested as individual units in new assortments. It provides a means of escape and spreading for favorable alleles, and a means to eliminate an unfavorable allele without bringing down all the other genes with which this allele is associated. This is the basis for natural selection.

Recombination occurs between precisely corresponding sequences, so that not a single base pair is added to or lost from the recombinant chromosomes. Three types of recombination share the feature that the process involves physical exchange of material between duplex DNAs:

- Recombination involving reaction between homologous sequences of DNA is called *generalized* or homologous recombination. In eukaryotes, it occurs at meiosis, usually both in males (during spermatogenesis) and females (during oogenesis). We recall that it happens at the "four strand" stage of meiosis, and involves only two of the four strands (see *1.20 Recombination occurs by physical exchange of DNA).*
- Another type of event sponsors recombination between specific pairs of sequences. This was first characterized in prokaryotes where specialized recombination, also known as site-specific recombination, is

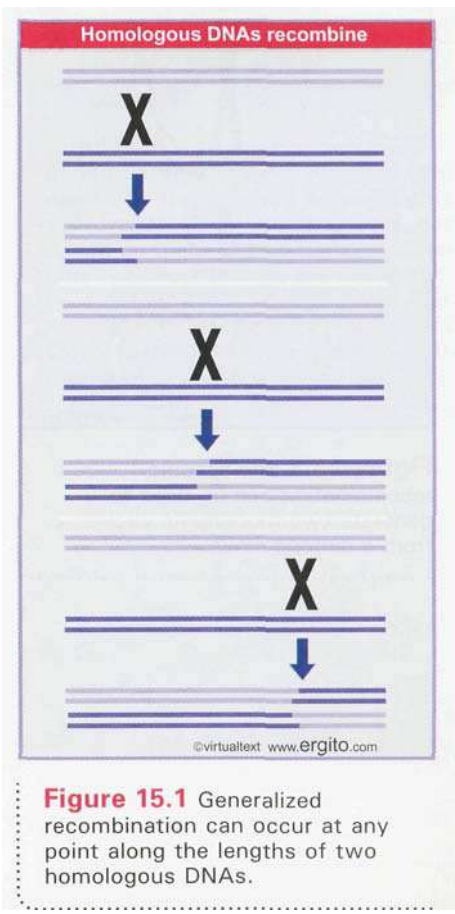


Figure 15.1 Generalized recombination can occur at any point along the lengths of two homologous DNAs.

By Book_Crazy [IND]

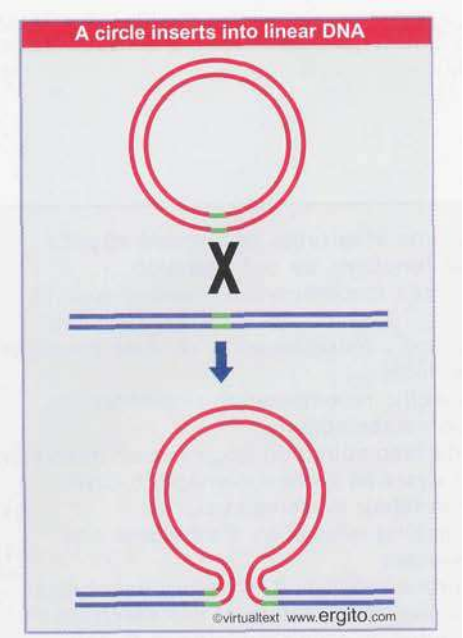


Figure 15.2 Site-specific recombination occurs between two specific sequences (identified in green). The other sequences in the two recombining DNAs are not homologous.

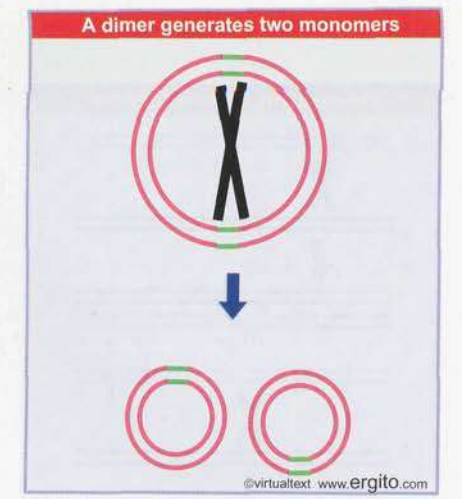


Figure 15.3 Site-specific recombination can be used to generate two monomeric circles from a dimeric circle.

responsible for the integration of phage genomes into the bacterial chromosome. The recombination event involves specific sequences of the phage DNA and bacterial DNA, which include a short stretch of homology. The enzymes involved in this event act only on the particular pair of target sequences in an intermolecular reaction. Some related intramolecular reactions are responsible during bacterial division for regenerating two monomeric circular chromosomes when a dimer has been generated by generalized recombination. Also in this latter class are recombination events that invert specific regions of the bacterial chromosome.

- A different type of event allows one DNA sequence to be inserted into another without relying on sequence homology. Transposition provides a means by which certain elements move from one chromosomal location to another. The mechanisms involved in transposition depend upon breakage and reunion of DNA strands, and thus are related to the processes of recombination (see *16 Transposons* and *17 Retroviruses and retroposons).*
- Another type of recombination is used by RNA viruses, in which the polymerase switches from one template to another while it is synthesizing RNA. As a result, the newly synthesized molecule joins sequence information from two different parents. This type of mechanism for recombination is called copy choice, and is discussed briefly in *17.4 Viral DNA is generated by reverse transcription.*

Let's consider the nature and consequences of the generalized and specialized recombination reactions.

Figure 15.1 makes the point that generalized recombination occurs between two homologous DNA duplexes, and can occur at any point along their length. The two chromosomes are cut at equivalent points, and then each is joined to the other to generate reciprocal recombinants. The crossover (marked by *the X) is the point at which* each becomes joined to the other. There is no change in the overall organization of DNA; the products have the same structure as the parents, and both parents and products are homologous.

Specialized recombination occurs only between specific sites. The results depend on the locations of the two recombining sites. **Figure 15.2** shows that an intermolecular recombination between a circular DNA and a linear DNA inserts the circular DNA into the linear DNA. **Figure 15.3** shows that an intramolecular recombination between two sites on a circular DNA releases two smaller circular DNAs. Specialized recombination is often used to make changes such as these in the organization of DNA. The change in organization is a consequence of the locations of the recombining sites. We have a large amount of information about the enzymes that undertake specialized recombination, which are related to the topoisomerases that act to change the supercoiling of DNA in space.

15.2 Homologous recombination occurs between synapsed chromosomes

Key Concepts

- Chromosomes must synapse (pair) in order for chiasmata to form where crossing-over occurs.
- We can correlate the stages of meiosis with the molecular events that happen to DNA.

Homologous recombination is a reaction between two duplexes of DNA. Its critical feature is that the enzymes responsible can use any pair of homologous sequences as substrates (although some types of sequences may be favored over others). The frequency of recombination is not constant throughout the genome, but is influenced by both global and local effects. The overall frequency may be different in oocytes and in sperm; recombination occurs twice as frequently in female as in male humans. And within the genome its frequency depends upon chromosome structure; for example, crossing-over is suppressed in the vicinity of the condensed and inactive regions of heterochromatin.

Recombination occurs during the protracted prophase of meiosis. **Figure 15.4** compares the visible progress of chromosomes through the five stages of meiotic prophase with the molecular interactions that are involved in exchanging material between duplexes of DNA.

The beginning of meiosis is marked by the point at which individual chromosomes become visible. Each of these chromosomes has replicated previously, and consists of two sister chromatids, each of which contains a duplex DNA. The homologous chromosomes approach one another and begin to pair in one or more regions, forming bivalents. Pairing extends until the entire length of each chromosome is apposed with its homolog. The process is called synapsis or chromosome pairing. When the process is completed, the chromosomes are laterally associated in the form of a synaptonemal complex, which has a characteristic structure in each species, although there is wide variation in the details between species.

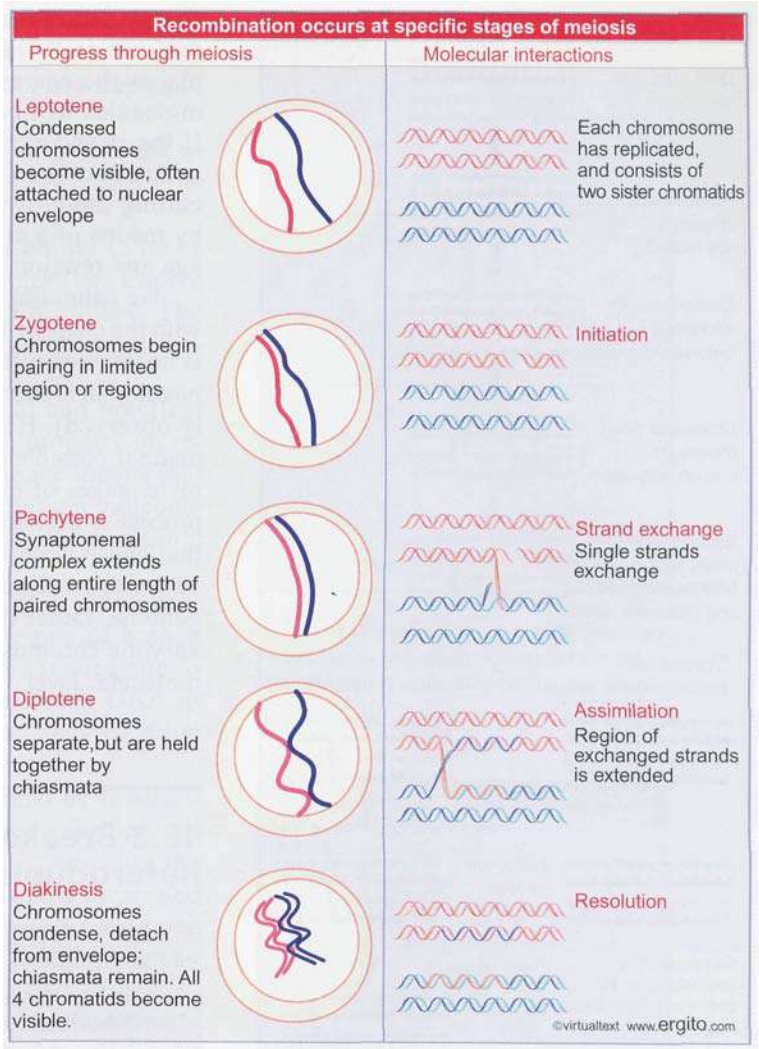


Figure 15.4 Recombination occurs during the first meiotic prophase. The stages of prophase are defined by the appearance of the chromosomes, each of which consists of two replicas (sister chromatids), although the duplicated state becomes visible only at the end. The molecular interactions of any individual crossing-over event involve two of the four duplex DNAs.

Recombination between chromosomes involves a physical exchange of parts, usually represented as a breakage and reunion, in which two nonsister chromatids (each containing a duplex of DNA) have been broken and then linked each with the other. When the chromosomes begin to separate, they can be seen to be held together at discrete sites, the chiasmata. The number and distribution of chiasmata parallel the features of genetic crossing-over. Traditional analysis holds that a chiasma represents the crossing-over event (see Figure 1.32). The chiasmata remain visible when the chromosomes condense and all four chromatids become evident.

What is the molecular basis for these events? Each sister chromatid contains a single DNA duplex, so each bivalent contains 4 duplex molecules of DNA. Recombination requires a mechanism that allows the duplex DNA of one sister chromatid to interact with the duplex DNA of a sister chromatid from the other chromosome. It must be possible for this reaction to occur between any pair of corresponding sequences in the two molecules in a highly specific manner that allows material to be exchanged with precision at the level of the individual base pair.

We know of only one mechanism for nucleic acids to recognize one another on the basis of sequence: complementarity between single strands. The figure shows a general model for the involvement of single strands in recombination. The first step in providing single

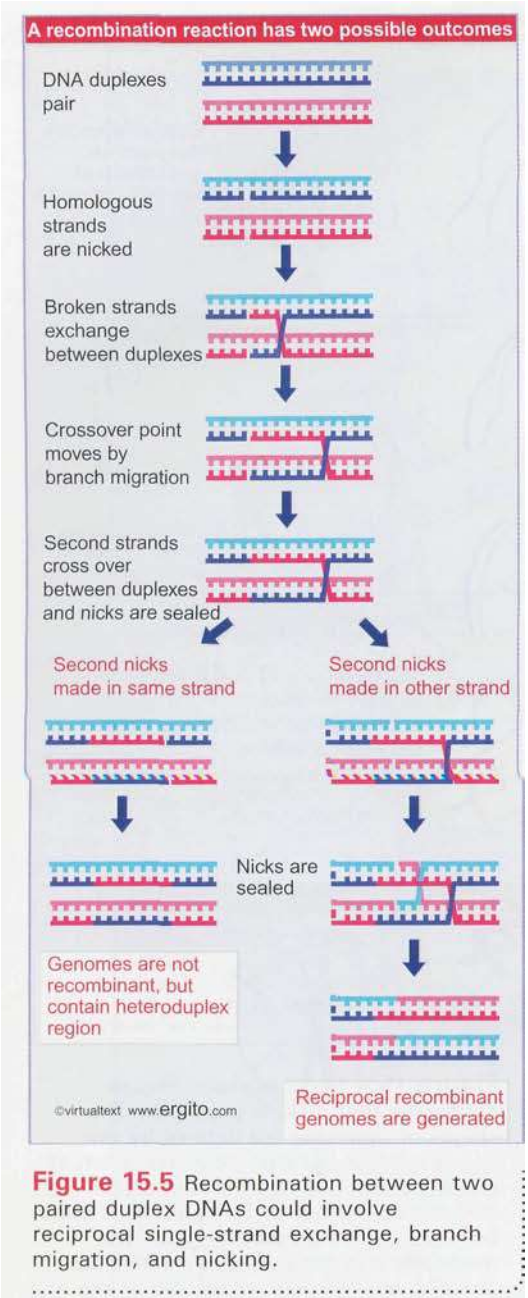


Figure 15.5 Recombination between two paired duplex DNAs could involve reciprocal single-strand exchange, branch migration, and nicking.

strands is to make a break in each DNA duplex. Then one or both of the strands of that duplex can be released. If (at least) one strand displaces the corresponding strand in the other duplex, the two duplex molecules will be specifically connected at corresponding sequences. If the strand exchange is extended, there can be more extensive connection between the duplex. And by exchanging both strands and later cutting them, it is possible to connect the parental duplex molecules by means of a crossover that corresponds to the demands of a breakage and reunion.

We cannot at this juncture relate these molecular events rigorously with the changes that are observed at the level of the chromosomes. There is no detailed information about the molecular events involved in recombination in higher eukaryotic cells (in which meiosis has been most closely observed). However, recently the isolation of mutants in yeast has made it possible to correlate some of the molecular steps with approximate stages of meiosis. Detailed information about the recombination process is available in bacteria, in which molecular activities are known that cause genetic exchange between duplex molecules. However, the bacterial reaction involves interaction between restricted regions of the genome, rather than an entire pairing of genomes. The synapsis of eukaryotic chromosomes remains the most difficult stage to explain at the molecular level.

15.3 Breakage and reunion involves heteroduplex DNA

Key Concepts

- The key event in recombination between two duplex DNA molecules is exchange of single strands.
- When a single strand from one duplex displaces its counterpart in the other duplex, it creates a branched structure.
- The exchange generates a stretch of heteroduplex DNA consisting of one strand from each parent.
- Two (reciprocal) exchanges are necessary to generate a joint molecule.
- The joint molecule is resolved into two separate duplex molecules by nicking two of the connecting strands.
- Whether recombinants are formed depends on whether the strands involved in the original exchange or the other pair of strands are nicked during resolution.

The act of connecting two duplex molecules of DNA is at the heart of the recombination process. Our molecular analysis of recombination therefore starts by expanding our view of the use of base pairing between complementary single strands in recombination. It is useful to imagine the recombination reaction in terms of single-strand exchanges (although we shall see that this is not necessarily how it is actually initiated), because the properties of the molecules created in this way are central to understanding the processes involved in recombination.

Figure 15.5 illustrates a process that starts with breakage at the corresponding points of the homologous strands of two paired DNA duplexes. The breakage allows movement of the free ends created by the nicks. Each strand leaves its partner and crosses over to pair with its complement in the other duplex.

By Book_Crazy [IND]

The reciprocal exchange creates a connection between the two DNA duplexes. The connected pair of duplexes is called a joint molecule. The point at which an individual strand of DNA crosses from one duplex to the other is called the recombinant joint.

At the site of recombination, each duplex has a region consisting of one strand from each of the parental DNA molecules. This region is called hybrid DNA or heteroduplex DNA.

An important feature of a recombinant joint is its ability to move along the duplex. Such mobility is called branch migration. **Figure 15.6** illustrates the migration of a single strand in a duplex. The branching point can migrate in either direction as one strand is displaced by the other.

Branch migration is important for both theoretical and practical reasons. As a matter of principle, it confers a dynamic property on recombining structures. As a practical feature, its existence means that the point of branching cannot be established by examining a molecule *in vitro* (because the branch may have migrated since the molecule was isolated).

Branch migration could allow the point of crossover in the recombination intermediate to move in either direction. The rate of branch migration is uncertain, but as seen *in vitro* is probably inadequate to support the formation of extensive regions of heteroduplex DNA in natural conditions. Any extensive branch migration *in vivo* must therefore be catalyzed by a recombination enzyme.

The joint molecule formed by strand exchange must be *resolved* into two separate duplex molecules. Resolution requires a further pair of nicks. We can most easily visualize the outcome by viewing the joint molecule in one plane as a Holliday structure. This is illustrated in **Figure 15.7,** which represents the structure of Figure 15.5 with one duplex rotated relative to the other. The outcome of the reaction depends on which pair of strands is nicked.

If the nicks are made in the pair of strands that were not originally nicked (the pair that did not initiate the strand exchange), all four of the original strands have been nicked. This releases splice recombinant DNA molecules. The duplex of one DNA parent is covalently linked to the duplex of the other DNA parent, via a stretch of heteroduplex DNA. There has been a conventional recombination event between markers located on either side of the heteroduplex region.

If the same two strands involved in the original nicking are nicked again, the other two strands remain intact. The nicking releases the original parental duplexes, which remain intact except that each has a residuum of the event in the form of a length of heteroduplex DNA. These are called patch recombinants.

These alternative resolutions of the joint molecule establish the principle that *a strand exchange between duplex DNAs always leaves behind a region of heteroduplex DNA, but the exchange may or may not be accompanied by recombination of the flanking regions.*

What is the minimum length of the region required to establish the connection between the recombining duplexes? Experiments in which short homologous sequences carried by plasmids or phages are introduced into bacteria suggest that the rate of recombination is substantially reduced if the homologous region is <75 bp. This distance is appreciably longer than the ~10 bp required for association between complementary single-stranded regions, which suggests that recombination imposes demands beyond annealing of complements as such.

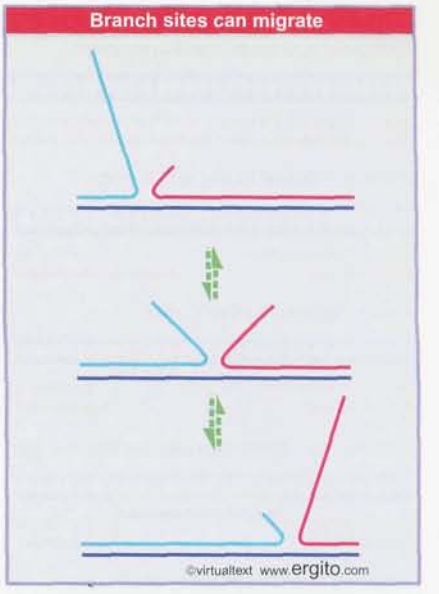


Figure 15.6 Branch migration can occur in either direction when an unpaired single strand displaces a paired strand.

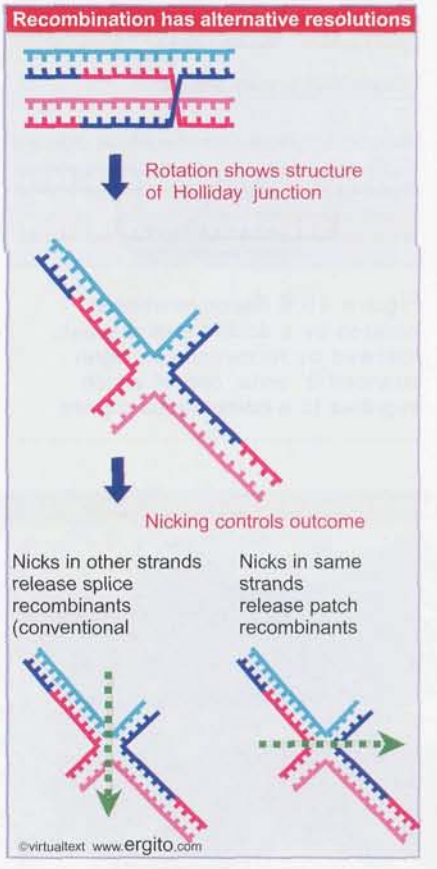


Figure 15.7 Resolution of a Holliday junction can generate parental or recombinant duplexes, depending on which strands are nicked. Both types of product have a region of heteroduplex DNA.

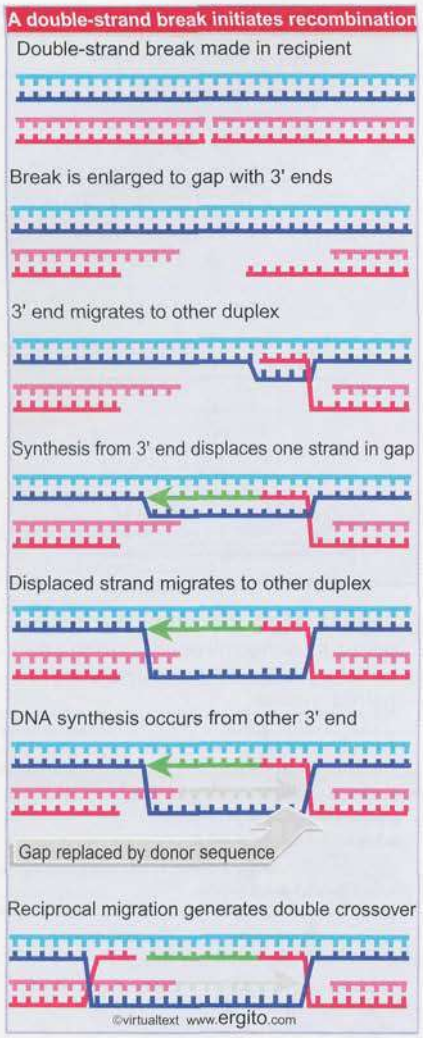


Figure 15.8 Recombination is initiated by a double-strand break, followed by formation of single-stranded 3′ ends, one of which migrates to a homologous duplex.

15.4 Double-strand breaks initiate recombination

Key Concepts

- Recombination is initiated by making a double-strand break in one (recipient) DNA duplex.
- Exonuclease action generates 3′-single-stranded ends that invade the other (donor) duplex.
- New DNA synthesis replaces the material that has been degraded.
- This generates a recombinant joint molecule in which the two DNA duplexes are connected by heteroduplex DNA.

The general model of Figure 15.4 shows that a break must be made in one duplex in order to generate a point from which single strands can unwind to participate in genetic exchange. Both strands of a duplex must be broken to accomplish a genetic exchange. Figure 15.5 shows a model in which individual breaks in single strands occur successively. However, genetic exchange is actually initiated by a double-strand break (DSB). The model is illustrated in **Figure 15.8.**

Recombination is initiated by an endonuclease that cleaves one of the partner DNA duplexes, the "recipient." The cut is enlarged to a gap by exonuclease action. The exonuclease(s) nibble away one strand on either side of the break, generating 3' single-stranded termini. One of the free 3' ends then invades a homologous region in the other, "donor" duplex. This is called single-strand invasion. The formation of heteroduplex DNA generates a D loop, in which one strand of the donor duplex is displaced. The D loop is extended by repair DNA synthesis, using the free 3' end as a primer to generate double-stranded DNA.

Eventually the D loop becomes large enough to correspond to the entire length of the gap on the recipient chromatid. When the extruded single strand reaches the far side of the gap, the complementary single-stranded sequences anneal. Now there is heteroduplex DNA on either side of the gap, and the gap itself is represented by the single-stranded D loop.

The duplex integrity of the gapped region can be restored by repair synthesis using the 3' end on the left side of the gap as a primer. Overall, the gap has been repaired by two individual rounds of single-strand DNA synthesis.

Branch migration converts this structure into a molecule with two recombinant joints. The joints must be resolved by cutting.

If both joints are resolved in the same way, the original noncrossover molecules will be released, each with a region of altered genetic information that is a footprint of the exchange event. If the two joints are resolved in opposite ways a genetic crossover is produced.

The structure of the two-jointed molecule before it is resolved illustrates a critical difference between the double-strand break model and models that invoke only single-strand exchanges:

- Following the double-strand break, heteroduplex DNA has been formed at each end of the region involved in the exchange. Between the two heteroduplex segments is the region corresponding to the gap, which now has the sequence of the donor DNA in both *molecules* (Figure 15.8). So the arrangement of heteroduplex sequences is asymmetric, and part of one molecule has been converted to the sequence of the other (which is why the initiating chromatid is called the recipient).
- Following reciprocal single-strand exchange, each DNA duplex has heteroduplex material covering the region from the initial site of exchange to the migrating branch (Figure 15.5). In variants of the single-strand exchange model in which some DNA is degraded and resynthesized, the initiating chromatid is the donor of genetic information.

By Book_Crazy [IND]

The double-strand break model does not reduce the importance of the formation of heteroduplex DNA, which remains the only plausible means by which two duplex molecules can interact. However, by shifting the responsibility for initiating recombination from single-strand to double-strand breaks, it influences our perspective about the ability of the cell to manipulate DNA.

The involvement of double-strand breaks seems surprising at first sight. Once a break has been made right across a DNA molecule, there is no going back. Compare the events of Figure 15.5 and Figure 15.8. In the single-strand exchange model, at no point has any information been lost. But in the double-strand break model, the initial cleavage is immediately followed by loss of information. Any error in retrieving the information could be fatal. On the other hand, the very ability to retrieve lost information by resynthesizing it from another duplex provides a major safety net for the cell.

15.5 Recombining chromosomes are connected by the synaptonemal complex

Key Concepts

- During the early part of meiosis, homologous chromosomes are paired in the synaptonemal complex.
- The mass of chromatin of each homologue is separated from the other by a proteinaceous complex.

A basic paradox in recombination is that the parental chromosomes never seem to be in close enough contact for recombination of DNA to occur. The chromosomes enter meiosis in the form of replicated (sister chromatid) pairs, visible as a mass of chromatin. They pair to form the synaptonemal complex, and it has been assumed for many years that this represents some stage involved with recombination, possibly a necessary preliminary to exchange of DNA. A more recent view is that the synaptonemal complex is a consequence rather than a cause of recombination. However, we have yet to define how the structure of the synaptonemal complex relates to molecular contacts between DNA molecules.

Synapsis begins when each chromosome (sister chromatid pair) condenses around a structure called the axial element, which is apparently proteinaceous. Then the axial elements of corresponding chromosomes become aligned, and the synaptonemal complex forms as a tripartite structure, in which the axial elements, now called lateral elements, are separated from each other by a central element. **Figure 15.9** shows an example.

Each chromosome at this stage appears as a mass of chromatin bounded by a lateral element. The two lateral elements are separated from each other by a fine but dense central element. The triplet of parallel dense strands lies in a single plane that curves and twists along its axis. The distance between the homologous chromosomes is considerable in molecular terms, more than 200 nm (the diameter of DNA is 2 nm). So a major problem in understanding the role of the complex is that, although it aligns homologous chromosomes, it is far from bringing homologous DNA molecules into contact.

The only visible link between the two sides of the synaptonemal complex is provided by spherical or cylindrical structures observed in fungi and insects. They lie across the complex and are called nodes or recombination nodules; they occur with the same frequency and distribution as the chiasmata. Their name reflects the hope that they may prove to be the sites of recombination.

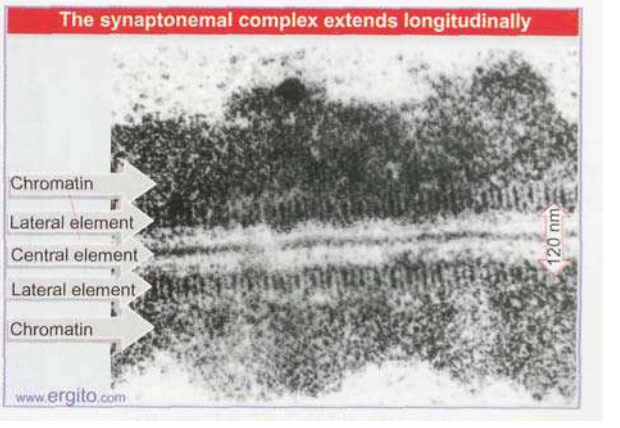


Figure 15.9 The synaptonemal complex brings chromosomes into juxtaposition. This example of *Neotellia* was kindly provided by M. Westergaard and D. Von Wettstein.

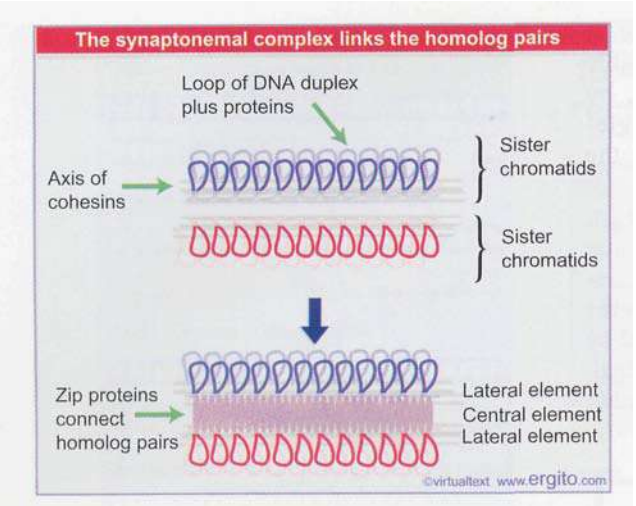


Figure 15.10 Each pair of sister chromatids has an axis made of cohesins. Loops of chromatin project from the axis. The synaptonemal complex is formed by linking together the axes via zip proteins.

From mutations that affect synaptonemal complex formation, we can relate the types of proteins that are involved to its structure. Figure 15.10 presents a molecular view of the synaptonemal complex. Its distinctive structural features are due to two groups of proteins:

- The cohesins form a single linear axis for each pair of sister chromatids from which loops of chromatin extend. This is equivalent to the lateral element of Figure 15.9. (The cohesins belong to a general group of proteins involved in connecting sister chromatids so they segregate properly at mitosis of meiosis; see *29.19 Cohesins hold sister chromatids together).*
- The lateral elements are connected by transverse filaments that are equivalent to the central element of Figure 15.9. These are formed from Zip proteins.

Mutations in proteins that are needed for lateral elements to form are found in the genes coding for cohesins. The cohesins that are used in meiosis include Smc3p (which is also used in mitosis) and Rec8p (which is specific to meiosis and is related to the mitotic cohesin Scc1p). The cohesins appear to bind to specific sites along the chromosomes in both mitosis and meiosis. They are likely to play a structural role in chromosome segregation. At meiosis, the formation of the lateral elements may be necessary for the later stages of recombination, because although these mutations do not prevent the formation of double-strand breaks, they do block formation of recombinants.

The *zip1* mutation allows lateral elements to form and to become aligned, but they do not become closely synapsed. The N-terminal domain of Zip1 protein is localized in the central element, but the C-terminal domain is localized in the lateral elements. Two other proteins, Zip2 and Zip3 are also localized with Zip 1. The group of Zip proteins form transverse filaments that connect the lateral elements of the sister chromatid pairs.

15.6 The synaptonemal complex forms after double-strand breaks

Key Concepts

- Double-strand breaks that initiate recombination occur before the synaptonemal complex forms.
- If recombination is blocked, the synaptonemal complex cannot form.

There is good evidence in yeast that double strand breaks initiate recombination in both homologous and site-specific recombination. Double-strand breaks were initially implicated in the change of mating type, which involves the replacement of one sequence by another (see *18.8 Unidirectional transposition is initiated by the recipient MAT locus).* Double-strand breaks also occur early in meiosis at sites that provide hotspots for recombination. Their locations are not sequence-specific They tend to occur in promoter regions and in general to coincide with more accessible regions of chromatin. The frequency of recombination declines in a gradient on one or both sides of the hotspot. The hotspot identifies the site at which recombination is initiated; and the gradient reflects the probability that the recombination events will spread from it.

We may now interpret the role of double-strand breaks in molecular terms. The flush ends created by the double-strand break are rapidly

converted on both sides into long 3' single-stranded ends, as shown in the model of Figure 15.8. A yeast mutation (*rad50*) that blocks the conversion of the flush end into the single-stranded protrusion is defective in recombination. This suggests that double-strand breaks are necessary for recombination. The gradient is determined by the declining probability that a single-stranded region will be generated as distance increases from the site of the double-strand break.

In *rad50* mutants, the 5' ends of the double-strand breaks are connected to the protein Spo11, which is homologous to the catalytic subunits of a family of type II topoisomerases. This suggests that Spo11 may be a topoisomerase-like enzyme that generates the double-strand breaks. The model for this reaction shown in **Figure 15.11** suggests that Spo11 interacts reversibly with DNA; the break is converted into a permanent structure by an interaction with another protein that dissociates the Spo11 complex. Then removal of Spo11 is followed by nuclease action. At least 9 other proteins are required to process the double-strand breaks. One group of proteins is required to convert the double-strand breaks into protruding 3'-OH single-stranded ends. Another group then enables the single-stranded ends to invade homologous duplex DNA.

The correlation between recombination and synaptonemal complex formation is well established, and recent work has shown that all mutations that abolish chromosome pairing in *Drosophila* or in yeast also prevent recombination. The system for generating the double-strand breaks that initiate recombination is generally conserved. Spo11 homologues have been identified in several higher eukaryotes, and a mutation in the *Drosophila* gene blocks all meiotic recombination.

There are few systems in which it is possible to compare molecular and cytological events at recombination, but recently there has been progress in analyzing meiosis in *S. cerevisiae*. The relative timing of events is summarized in **Figure 15.12.**

Double-strand breaks appear and then disappear over a 60 minute period. The first joint molecules, which are putative recombination intermediates, appear soon after the double-strand breaks disappear. The sequence of events suggests that double-strand breaks, individual pairing reactions, and formation of recombinant structures occur in succession at the same chromosomal site.

Double-strand breaks appear during the period when axial elements form. They disappear during the conversion of the paired chromosomes into synaptonemal complexes. This relative timing of events suggests that formation of the synaptonemal complex results from the initiation of recombination via the introduction of double-strand breaks and their

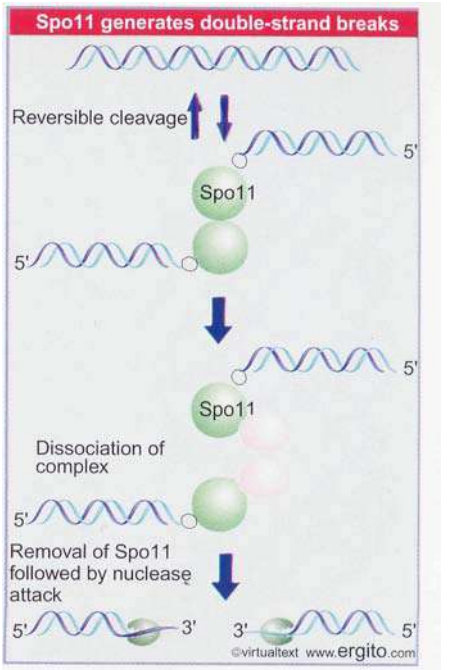


Figure 15.11 Spo11 is covalently joined to the 5′ ends of double-strand breaks.

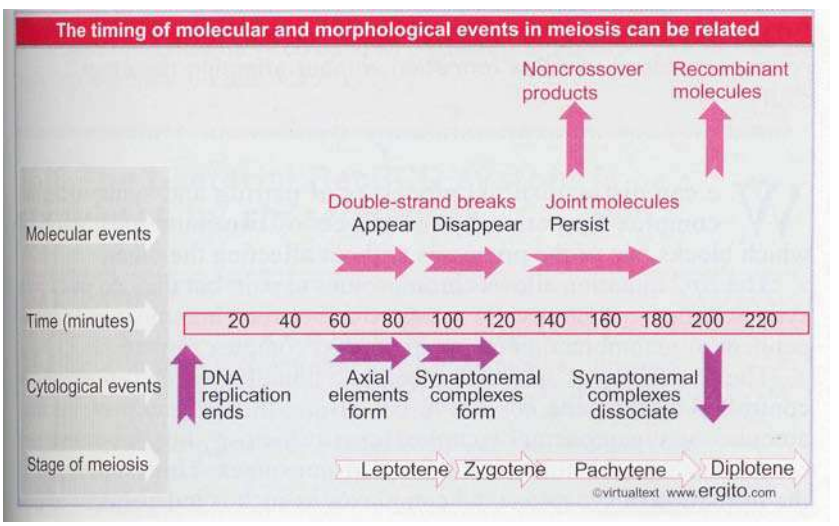


Figure 15.12 Double-strand breaks appear when axial elements form, and disappear during the extension of synaptonemal complexes. Joint molecules appear and persist until DNA recombinants are detected at the end of pachytene.

conversion into later intermediates of recombination. This idea is supported by the observation that the *rad50* mutant cannot convert axial elements into synaptonemal complexes. This refutes the traditional view of meiosis that the synaptonemal complex represents the need for chromosome pairing to precede the molecular events of recombination.

It has been difficult to determine whether recombination occurs at the stage of synapsis, because recombination is assessed by the appearance of recombinants after the completion of meiosis. However, by assessing the appearance of recombinants in yeast directly in terms of the production of DNA molecules containing diagnostic restriction sites, it has been possible to show that recombinants appear at the end of pachytene. This clearly places the completion of the recombination event after the formation of synaptonemal complexes.

So the synaptonemal complex forms after the double-strand breaks that initiate recombination, and it persists until the formation of recombinant molecules. It does not appear to be necessary for recombination as such, because some mutants that lack a normal synaptonemal complex can generate recombinants. Mutations that abolish recombination, however, also fail to develop a synaptonemal complex. This suggests that the synaptonemal complex forms as a consequence of recombination, following chromosome pairing, and is required for later stages of meiosis.

Ever since the model for recombination via a Holliday structure was proposed, it has been assumed that the resolution of this structure gives rise to either noncrossover products (with a residual stretch of hybrid DNA) or to crossovers (recombinants), depending on which strands are involved in resolution (see Figure 15.7). However, recent measurements of the times of production of noncrossover and crossover molecules suggest that this may not be true. Crossovers do not appear until well after the first appearance of joint molecules, but noncrossovers appear almost simultaneously with the joint molecules (see Figure 15.12). If both types of product were produced by the same resolution process, however, we would expect them to appear at the same time. The discrepancy in timing suggests that crossovers are produced as previously thought, by resolution of joint molecules, but that there may be some other route for the production of noncrossovers.

15.7 Pairing and synaptonemal complex formation are independent

Key Concepts

* Mutations can occur in either chromosome pairing or synaptonemal complex formation without affecting the other process.

We can distinguish the processes of pairing and synaptonemal complex formation by the effects of two mutations, each of which blocks one of the processes without affecting the other.

The *zip2* mutation allows chromosomes to pair, but they do not form synaptonemal complexes. So recognition between homologues is independent of recombination or synaptonemal complex formation.

The specificity of association between homologous chromosomes is controlled by the gene *hop2* in *S. cerevisiae.* In *hop2* mutants, normal amounts of synaptonemal complex form at meiosis, but the individual complexes contain nonhomologous chromosomes. This suggests that the formation of synaptonemal complexes as such is independent of ho-

mology (and therefore cannot be based on any extensive comparison of DNA sequences). The usual role of Hop2 is to prevent nonhomologous chromosomes from interacting.

Double-strand breaks form in the mispaired chromosomes in the synaptonemal complexes of *hop2* mutants, but they are not repaired. This suggests that, if formation of the synaptonemal complex requires double-strand breaks, it does not require any extensive reaction of these breaks with homologous DNA.

It is not clear what usually happens during pachytene, before DNA recombinants are observed. It may be that this period is occupied by the subsequent steps of recombination, involving the extension of strand exchange, DNA synthesis, and resolution.

At the next stage of meiosis (diplotene), the chromosomes shed the synaptonemal complex; then the chiasmata become visible as points at which the chromosomes are connected. This has been presumed to indicate the occurrence of a genetic exchange, but the molecular nature of a chiasma is unknown. It is possible that it represents the residuum of a completed exchange, or that it represents a connection between homologous chromosomes where a genetic exchange has not yet been resolved. Later in meiosis, the chiasmata move toward the ends of the chromosomes. This flexibility suggests that they represent some remnant of the recombination event, rather than providing the actual intermediate.

Recombination events occur at discrete points on meiotic chromosomes, but we cannot as yet correlate their occurrence with the discrete structures that have been observed, that is, recombination nodules and chiasmata. However, insights into the molecular basis for the formation of discontinuous structures are provided by the identification of proteins involved in yeast recombination that can be localized to discrete sites. These include MSH4 (which is related to bacterial proteins involved in mismatch-repair), and Dmc1 and Rad51 (which are homologues of the *E. coli* RecA protein). The exact roles of these proteins in recombination remain to be established.

Recombination events are subject to a general control. Only a minority of interactions actually mature as crossovers, but these are distributed in such a way that typically each pair of homologues acquires only 1 -2 crossovers, yet the probability of zero crossovers for a homologue pair is very low (<0.1%). This process is probably the result of a single crossover control, because the nonrandomness of crossovers is generally disrupted in certain mutants. Furthermore, the occurrence of recombination is necessary for progress through meiosis, and a "checkpoint" system (see *29 Cell cycle and growth regulation)* exists to block meiosis if recombination has not occurred. (The block is lifted when recombination has been successfully completed; this system provides a safeguard to ensure that cells do not try to segregate their chromosomes until recombination has occurred.)

15.8 The bacterial RecBCD system is stimulated by *chi* sequences

Key Concepts

- The RecBCD complex has nuclease and helicases activities.
- It binds to DNA downstream of a *chi* sequence, unwinds the duplex, and degrades one strand from 3'-5' as it moves to the *chi* site.
- The *chi* site triggers loss of the RecD subunit and nuclease activity.

The nature of the events involved in exchange of sequences between DNA molecules was first described in bacterial systems. Here the recognition reaction is part and parcel of the recombination mechanism and involves restricted regions of DNA molecules rather than intact chromosomes. But the general order of molecular events is similar: a single strand from a broken molecule interacts with a partner duplex; the region of pairing is extended; and an endonuclease resolves the partner duplexes. Enzymes involved in each stage are known, although they probably represent only some of the components required for recombination.

Bacterial enzymes implicated in recombination have been identified by the occurrence of *rec*⁻ mutations in their genes. The phenotype of Rec⁻ mutants is the inability to undertake generalized recombination. Some 10-20 loci have been identified.

Bacteria do not usually exchange large amounts of duplex DNA, but there may be various routes to initiate recombination in prokaryotes. In some cases, DNA may be available with free single-stranded 3′ ends: DNA may be provided in single-stranded form (as in conjugation; see *13.13 Conjugation transfers single-stranded DNA);* single-stranded gaps may be generated by irradiation damage; or single-stranded tails may be generated by phage genomes undergoing replication by a rolling circle. However, in circumstances involving two duplex molecules (as in recombination at meiosis in eukaryotes), single-stranded regions and 3′ ends must be generated.

One mechanism for generating suitable ends has been discovered as a result of the existence of certain hotspots that stimulate recombination. They were discovered in phage lambda in the form of mutants, called *chi*, that have single base-pair changes creating sequences that stimulate recombination. These sites lead us to the role of other proteins involved in recombination.

These sites share a constant nonsymmetrical sequence of 8 bp:

5′ GCTGGTGG 3′
3′ CGACCACC 5′

The *chi* sequence occurs naturally in *E. coli* DNA about once every 5-10 kb. Its absence from wild-type lambda DNA, and also from other genetic elements, shows that it is not essential for recombination.

A *chi* sequence stimulates recombination in its general vicinity, say within a distance of up to 10 kb from the site. A *chi* site can be activated by a double-strand break made several kb away on one particular side (to the right of the sequence as written above). This dependence on orientation suggests that the recombination apparatus must associate with DNA at a broken end, and then can move along the duplex only in one direction.

chi sites are targets for the action of an enzyme coded by the genes *recBCD*. This complex enzyme exercises several activities. It is a potent nuclease that degrades DNA, originally identified as the activity exonuclease V It has a helicase activity that can unwind duplex DNA in the presence of SSB; and it has an ATPase activity. Its role in recombination may be to provide a single-stranded region with a free 3' end.

Figure 15.13 shows how these reactions are coordinated on a substrate DNA that has a *chi* site. When RecBCD binds DNA on the right site of *chi*, it moves along unwinding the DNA. It degrades the released single strand with the 3' end. When it reaches the *chi* site, it pauses and cleaves one (the top) strand of the DNA at a position between 4 and 6 bases to the right of *chi*. The top strand of the *chi* site is recognized in single-stranded form. Recognition of the *chi* site causes the RecD subunit to dissociate or become inactivated, as a result of which the enzyme loses its nuclease activity. However, it continues to function as a helicase.

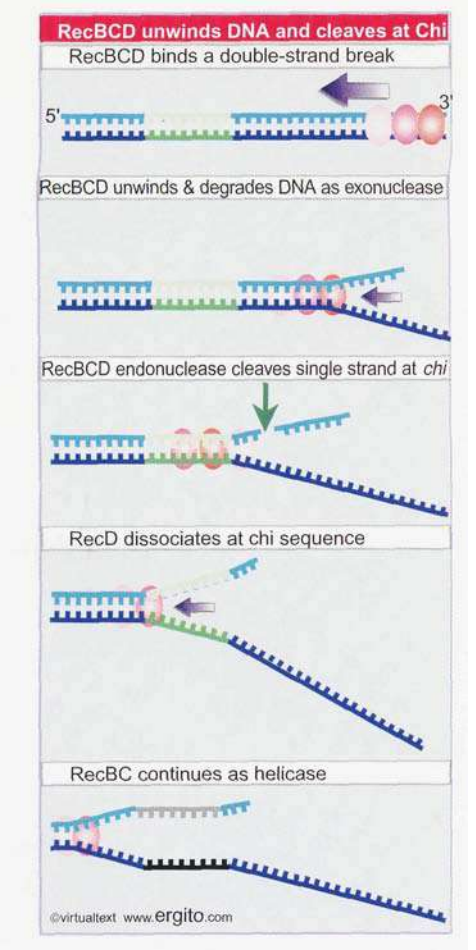


Figure 15.13 RecBCD nuclease approaches a *chi* sequence from one side, degrading DNA as it proceeds; at the *chi* site, it makes an endonucleolytic cut, loses RecD, and retains only the helicase activity.

15.9 Strand-transfer proteins catalyze single-strand assimilation

Key Concepts

- RecA forms filaments with single-stranded or duplex **DNA** and catalyzes the ability of a single-stranded **DNA with** a free 3' to displace its counterpart in a **DNA** duplex.

The *E. coli* protein RecA was the first example to be discovered of a DNA strand-transfer protein. It is the paradigm for a group that includes several other bacterial and archaeal proteins, Rad51 in *S. cerevisiae*, and the higher eukaryotic protein Dmc1. Analysis of yeast *rad51* mutants shows that this class of protein plays a central role in recombination. They accumulate double-strand breaks and fail to form normal synaptonemal complexes. This reinforces the idea that exchange of strands between DNA duplexes is involved in formation of the synaptonemal complex, and raises the possibility that chromosome synapsis is related to the bacterial strand assimilation reaction.

RecA in bacteria has two quite different types of activity: it can stimulate protease activity in the SOS response (see *15.27 RecA triggers the SOS system);* and can promote base pairing between a single strand of DNA and its complement in a duplex molecule. Both activities are activated by single-stranded DNA in the presence of ATR

The DNA-handling activity of RecA enables a single strand to displace its homologue in a duplex in a reaction that is called single-strand uptake or single-strand assimilation. The displacement reaction can occur between DNA molecules in several configurations and has three general conditions:

- One of the DNA molecules must have a single-stranded region.
- One of the molecules must have a free 3′ end.
- The single-stranded region and the 3′ end must be located within a region that is complementary between the molecules.

The reaction is illustrated in **Figure 15.14.** When a linear single strand invades a duplex, it displaces the original partner to its complement. The reaction can be followed most easily by making either the donor or recipient a circular molecule. The reaction proceeds 5′-3′ along the strand whose partner is being displaced and replaced. that is, the reaction involves an exchange in which (at least) one of the exchanging strands has a free 3' end.

Single-strand assimilation is potentially related to the initiation of recombination. All models call for an intermediate in which one or

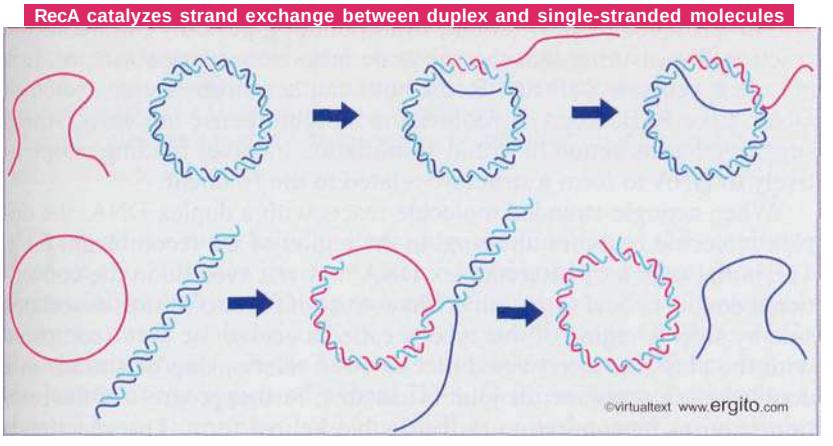


Figure 15.14 RecA promotes the assimilation of invading single strands into duplex DNA so long as one of the reacting strands has a free end.

By Book_Crazy [IND]

both single strands cross over from one duplex to the other (see Figure 15.5 and Figure 15.8). RecA could catalyze this stage of the reaction. In the bacterial context, RecA acts on substrates generated by RecBCD. RecBCD-mediated unwinding and cleavage can be used to generate ends that initiate the formation of heteroduplex joints. RecA can take the single strand with the 3' end that is released when RecBCD cuts at *chi*, and can use it to react with a homologous duplex sequence, thus creating a joint molecule.

All of the bacterial and archaeal proteins in the RecA family can aggregate into long filaments with single-stranded or duplex DNA. (Eukaryotic homologues of RecA do not form filaments, so the mechanics of the reaction are likely to be different in eukaryotes.) There are 6 RecA monomers per turn of the filament, which has a helical structure with a deep groove that contains the DNA. The stoichiometry of binding is 3 nucleotides (or base pairs) per RecA monomer. The DNA is held in a form that is extended 1.5 times relative to duplex B DNA, making a turn every 18.6 nucleotides (or base pairs). When duplex DNA is bound, it contacts RecA via its minor groove, leaving the major groove accessible for possible reaction with a second DNA molecule.

The interaction between two DNA molecules occurs within these filaments. When a single strand is assimilated into a duplex, the first step is for RecA to bind the single strand into a filament. Then the duplex is incorporated, probably forming some sort of triple-stranded structure. In this system, synapsis precedes physical exchange of material, because the pairing reaction can take place even in the absence of free ends, when strand exchange is impossible. A free 3' end is required for strand exchange. The reaction occurs within the filament, and RecA remains bound to the strand that was originally single, so that at the end of the reaction RecA is bound to the duplex molecule.

All of the proteins in this family can promote the basic process of strand exchange without a requirement for energy input. However, RecA augments this activity by using ATP hydrolysis. Large amounts of ATP are hydrolyzed during the reaction. The ATP may act through an allosteric effect on RecA conformation. When bound to ATP, the DNA-binding site of RecA has a high affinity for DNA; this is needed to bind DNA and for the pairing reaction. Hydrolysis of ATP converts the binding site to low affinity, which is needed to release the heteroduplex DNA.

We can divide the reaction that RecA catalyzes between single-stranded and duplex DNA into three phases:

- a slow presynaptic phase in which RecA polymerizes on single-stranded DNA;
- a fast pairing reaction between the single-stranded DNA and its complement in the duplex to produce a heteroduplex joint;
- a slow displacement of one strand from the duplex to produce a long region of heteroduplex DNA.

The presence of SSB (single-strand binding protein) stimulates the reaction, by ensuring that the substrate lacks secondary structure. It is not clear yet how SSB and RecA both can act on the same stretch of DNA. Like SSB, RecA is required in stoichiometric amounts, which suggests that its action in strand assimilation involves binding cooperatively to DNA to form a structure related to the filament.

When a single-stranded molecule reacts with a duplex DNA, the duplex molecule becomes unwound in the region of the recombinant joint. The initial region of heteroduplex DNA may not even lie in the conventional double helical form, but could consist of the two strands associated side by side. A region of this type is called a paranemic joint (compared with the classical intertwined plectonemic relationship of strands in a double helix). A paranemic joint is unstable; further progress of the reaction requires its conversion to the double-helical form. This reaction is

equivalent to removing negative supercoils and may require an enzyme that solves the unwinding/rewinding problem by making transient breaks that allow the strands to rotate about each other.

All of the reactions we have discussed so far represent only a part of the potential recombination event: the invasion of one duplex by a single strand. Two duplex molecules can interact with each other under the sponsorship of RecA, provided that one of them has a single-stranded region of at least 50 bases. The single-stranded region can take the form of a tail on a linear molecule or of a gap in a circular molecule.

The reaction between a partially duplex molecule and an entirely duplex molecule leads to the exchange of strands. An example is illustrated in **Figure 15.15.** Assimilation starts at one end of the linear molecule, where the invading single strand displaces its homologue in the duplex in the customary way. But when the reaction reaches the region that is duplex in both molecules, the invading strand unpairs from its partner, which then pairs with the other displaced strand.

At this stage, the molecule has a structure indistinguishable from the recombinant joint in Figure 15.7. The reaction sponsored *in vitro* by RecA can generate Holliday junctions, which suggests that the enzyme can mediate reciprocal strand transfer. We know less about the geometry of four-strand intermediates bound by RecA, but presumably two duplex molecules can lie side by side in a way consistent with the requirements of the exchange reaction.

The biochemical reactions characterized *in vitro* leave open many possibilities for the functions of strand-transfer proteins *in vivo*. Their involvement is triggered by the availability of a single-stranded 3' end. In bacteria, this is most likely generated when RecBCD processes a double-strand break to generate a single-stranded end. One of the main circumstances in which this is invoked may be when a replication fork stalls at a site of DNA damage (see *15.26 Recombination is an important mechanism to recover from replication errors).* The introduction of DNA during conjugation, when RecA is required for recombination with the host chromosome, is more closely related to conventional recombination. In yeast, double-strand breaks may be generated by DNA damage or as part of the normal process of recombination. In either case, processing of the break to generate a 3' single-stranded end is followed by loading the single strand into a filament with Rad51, followed by a search for matching duplex sequences. This can be used in both repair and recombination reactions.

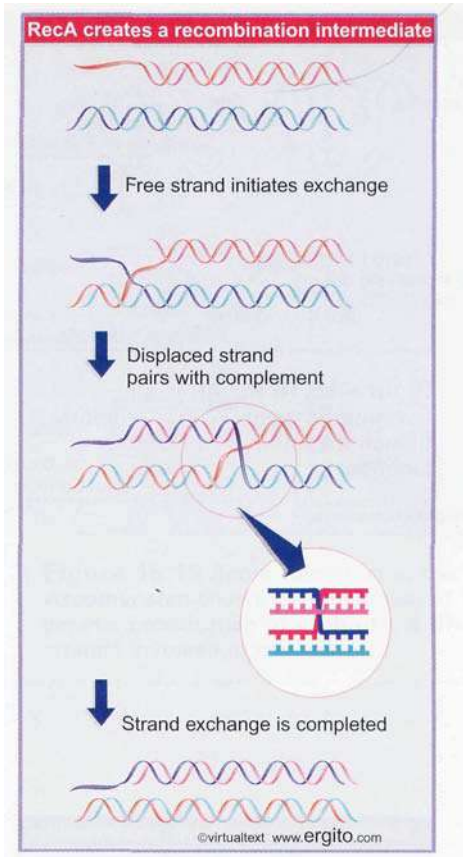


Figure 15.15 RecA-mediated strand exchange between partially duplex and entirely duplex DNA generates a joint molecule with the same structure as a recombination intermediate.

15.10 The Ruv system resolves Holliday junctions

Key Concepts

- The Ruv complex acts on recombinant junctions.
- RuvA recognizes the structure of the junction and RuvB is a helicase that catalyzes branch migration.
- RuvC cleaves junctions to generate recombination intermediates.

One of the most critical steps in recombination is the resolution *of the Holliday junction, which determines whether there is a* reciprocal recombination or a reversal of the structure *that leaves only* a short stretch of hybrid DNA (see Figure 15.5 and Figure 15.7). Branch migration from the exchange site (see Figure 15.6) determines the length of the region of hybrid DNA (with or without recombination). The proteins involved in stabilizing and resolving Holliday

By Book_Crazy [IND]

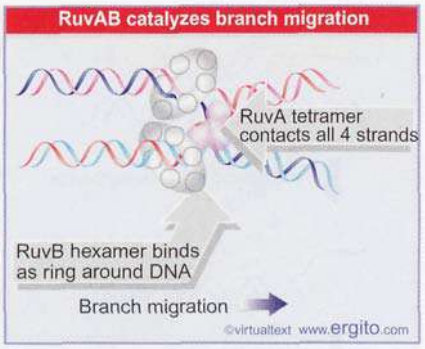


Figure 15.16 RuvAB is an asymmetric complex that promotes branch migration of a Holliday junction.

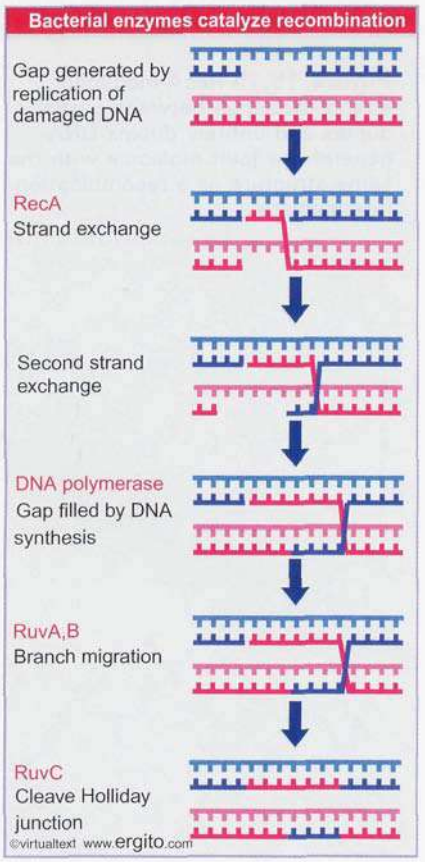


Figure 15.17 Bacterial enzymes can catalyze all stages of recombination in the repair pathway following the production of suitable substrate DNA molecules.

junctions have been identified as the products of the *Ruv* genes in *E. coli*. RuvA and RuvB increase the formation of heteroduplex structures. RuvA recognizes the structure of the Holliday junction. RuvA binds to all four strands of DNA at the crossover point and forms two tetramers that sandwich the DNA. RuvB is a hexameric helicase with an ATPase activity that provides the motor for branch migration. Hexameric rings of RuvB bind around each duplex of DNA upstream of the crossover point. A diagram of the complex is shown in **Figure 15.16.**

The RuvAB complex can cause the branch to migrate as fast as 10-20 bp/sec. A similar activity is provided by another helicase, RecG. RuvAB displaces RecA from DNA during its action. The RuvAB and RecG activities both can act on Holliday junctions, but if both are mutant, *E. coli* is completely defective in recombination activity.

The third gene, *ruvC*, codes for an endonuclease that specifically recognizes Holliday junctions. It can cleave the junctions *in vitro* to resolve recombination intermediates. A common tetranucleotide sequence provides a hotspot for RuvC to resolve the Holliday junction. The tetranucleotide (ATTG) is asymmetric, and thus may direct resolution with regard to which pair of strands is nicked. This determines whether the outcome is patch recombinant formation (no overall recombination) or splice recombinant formation (recombination between flanking markers). Crystal structures of RuvC and other junction-resolving enzymes show that there is a little structural similarity among the group, in spite of their common function.

All of this suggests that recombination uses a "resolvasome" complex that includes enzymes catalyzing branch migration as well as junction-resolving activity. It is possible that mammalian cells contain a similar complex.

We may now account for the stages of recombination in *E. coli* in terms of individual proteins. **Figure 15.17** shows the events that are involved in using recombination to repair a gap in one duplex by retrieving material from the other duplex. The major caveat in applying these conclusions to recombination in eukaryotes is that bacterial recombination generally involves interaction between a fragment of DNA and a whole chromosome. It occurs as a repair reaction that is stimulated by damage to DNA, and this is not entirely equivalent to recombination between genomes at meiosis. Nonetheless, similar molecular activities are involved in manipulating DNA.

Another system of resolvases has been characterized in yeast and mammals. Mutants in *S. cerevisiae mus81* are defective in recombination. Mus81 is a component of an endonuclease that resolves Holliday junctions into duplex structures. The resolvase is important both in meiosis and for restarting stalled replication forks (see *15.26 Recombination is an important mechanism to recover from replication errors).*

15.11 Gene conversion accounts for interallelic recombination

Key Concepts

- Heteroduplex DNA that is created by recombination can have mismatched sequences where the recombining alleles are not identical.
- Repair systems may remove mismatches by changing one of the strands so its sequence is complementary to the other.

By Book_Crazy [IND]

The involvement of heteroduplex DNA explains the characteristics of recombination between alleles; indeed, allelic recombination provided the impetus for the development of the heteroduplex model. When recombination between alleles was discovered, the natural assumption was that it takes place by the same mechanism of reciprocal recombination that applies to more distant loci. That is to say that an individual breakage and reunion event occurs within the locus to generate a reciprocal pair of recombinant chromosomes. However, in the close quarters of a single gene, the formation of heteroduplex DNA itself is usually responsible for the recombination event.

Individual recombination events can be studied in the Ascomycetes fungi, because the products of a single meiosis are held together in a large cell, the ascus, also sometimes called the tetrad. Even better, in some fungi, the four haploid nuclei produced by meiosis are arranged in a linear order. Actually, a mitosis occurs after the production of these four nuclei, giving a linear series of eight haploid nuclei. Figure **15.18** shows that each of these nuclei effectively represents the genetic character of one of the eight strands of the four chromosomes produced by the meiosis.

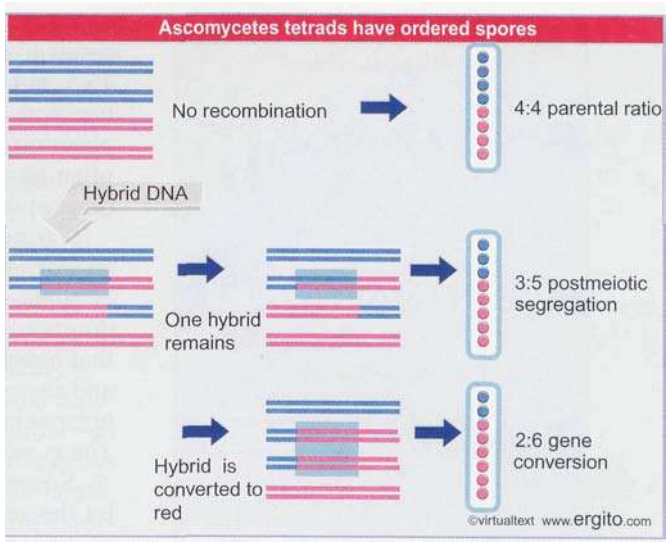


Figure 15.18 Spore formation in the *Ascomycetes* allows determination of the genetic constitution of each of the DNA strands involved in meiosis.

Meiosis in a heterozygote should generate four copies of each allele. This is seen in the majority of spores. But there are some spores with abnormal ratios. They are explained by the formation and correction of heteroduplex DNA in the region in which the alleles differ. The figure illustrates a recombination event in which a length of hybrid DNA occurs on one of the four meiotic chromosomes, a possible outcome of recombination initiated by a double-strand break.

Suppose that two alleles differ by a single point mutation. When a strand exchange occurs to generate heteroduplex DNA, the two strands of the heteroduplex will be mispaired at the site of mutation. So each strand of DNA carries different genetic information. If no change is made in the sequence, the strands separate at the ensuing replication, each giving rise to a duplex that perpetuates its information. This event is called postmeiotic segregation, because it reflects the separation of DNA strands after meiosis. Its importance is that it demonstrates directly the existence of heteroduplex DNA in recombining alleles.

Another effect is seen when examining recombination between alleles: the proportions of the alleles differ from the initial 4:4 ratio. This effect is called gene conversion. It describes a nonreciprocal transfer of information from one chromatid to another.

Gene conversion results from exchange of strands between DNA molecules, and the change in sequence may have either of two causes at the molecular level:

- As indicated by the double-strand break model in Figure 15.8, one DNA duplex may act as a donor of genetic information that directly replaces the corresponding sequences in the recipient duplex by a process of gap generation, strand exchange, and gap filling.
- As part of the exchange process, heteroduplex DNA is generated when a single strand from one duplex pairs with its complement in the other duplex. Repair systems recognize mispaired bases in heteroduplex DNA, and may then excise and replace one of the strands to restore complementarity. Such an event converts the strand of DNA representing one allele into the sequence of the other allele.

Gene conversion does not depend on crossing-over, but is correlated with it. A large proportion of the aberrant asci show genetic recombination between two markers on either side of a site of interallelic gene

conversion. This is exactly what would be predicted if the aberrant ratios result from initiation of the recombination process as shown in Figure 15.5, but with an approximately equal probability of resolving the structure with or without recombination (as indicated in Figure 15.7). The implication is that fungal chromosomes initiate crossing-over about twice as often as would be expected from the measured frequency of recombination between distant genes.

Various biases are seen when recombination is examined at the molecular level. Either direction of gene conversion may be equally likely, or allele-specific effects may create a preference for one direction. Gradients of recombination may fall away from hotspots. We now know that hotspots represent sites at which double-strand breaks are initiated, and the gradient is correlated with the extent to which the gap at the hotspot is enlarged and converted to long single-stranded ends (see *15.6 The synaptonemal complex forms after double-strand breaks).*

Some information about the extent of gene conversion is provided by the sequences of members of gene clusters. Usually, the products of a recombination event will separate and become unavailable for analysis at the level of DNA sequence. However, when a chromosome carries two (non-allelic) genes that are related, they may recombine by an "unequal crossing-over" event (see *4.7 Unequal crossing-over rearranges gene clusters).* All we need to note for now is that a heteroduplex may be formed between the two nonallelic genes. Gene conversion effectively converts one of the nonallelic genes to the sequence of the other.

The presence of more than one gene copy on the same chromosome provides a footprint to trace these events. For example, if heteroduplex formation and gene conversion occurred over part of one gene, this part may have a sequence identical with or very closely related to the other gene, while the remaining part shows more divergence. Available sequences suggest that gene conversion events may extend for considerable distances, up to a few thousand bases.

15.12 Supercoiling affects the structure of DNA

Key Concepts

- Supercoiling occurs only in a closed DNA with no free ends.
- A closed DNA can be a circular DNA molecule or a linear molecule where both ends are anchored in a protein structure.
- Any closed DNA molecule has a linking number, which is the sum of the twisting number and writhing number.
- Turns can be repartitioned between the twisting number and writhing number, so that a change in the structure of the double helix can compensate for a change in its coiling in space.
- The linking number can be changed only by breaking and making bonds in DNA.

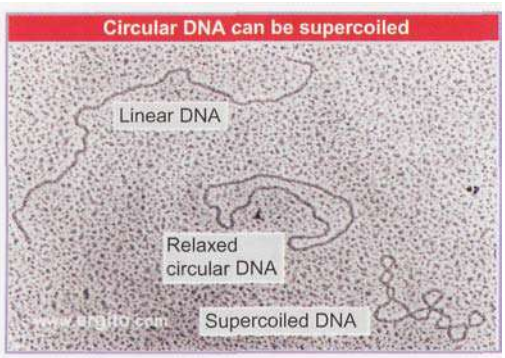


Figure 15.19 Linear DNA is extended, a circular DNA remains extended if it is relaxed (nonsupercoiled), but a supercoiled DNA has a twisted and condensed form.

The winding of the two strands of DNA around each other in the double helical structure makes it possible to change the structure by influencing its conformation in space. If the two ends of a DNA molecule are fixed, the double helix can be wound around itself in space. This is called supercoiling. The effect can be imagined like a rubber band twisted around itself. The simplest example of a DNA with no fixed ends is a circular molecule. The effect of supercoiling can be seen by comparing the nonsupercoiled circular DNA lying flat in Figure 15.19 with the supercoiled circular molecule that forms a twisted and therefore more condensed shape.

By Book_Crazy [IND]

The consequences of supercoiling depend on whether the DNA is twisted around itself in the same sense as the two strands within the double helix (clockwise) or in the opposite sense. Twisting in the same sense produces *positive supercoiling.* This has the effect of causing the DNA strands to wound around one another more tightly, so that there are more base pairs per turn. Twisting in the opposite sense produces *negative supercoiling.* This causes the DNA strands to be twisted around one another less tightly, so there are fewer base pairs per turn. Negative supercoiling can be thought of as creating tension in the DNA that is relieved by unwinding the double helix. The ultimate effect of negative supercoiling is to generate a region in which the two strands of DNA have separated—formally there are zero base pairs per turn.

Topological manipulation of DNA is a central aspect of all its functional activities—recombination, replication, and transcription—as well as of the organization of higher-order structure. All synthetic activities involving double-stranded DNA require the strands to separate. However, the strands do not simply lie side by side; they are intertwined. Their separation therefore requires the strands to rotate about each other in space. Some possibilities for the unwinding reaction are illustrated in **Figure 15.20.**

We might envisage the structure of DNA in terms of a free end that would allow the strands to rotate about the axis of the double helix for unwinding. Given the length of the double helix, however, this would involve the separating strands in a considerable amount of flailing about, which seems unlikely in the confines of the cell.

A similar result is achieved by placing an apparatus to control the rotation at the free end. However, the effect must be transmitted over a considerable distance, again involving the rotation of an unreasonable length of material.

Consider the effects of separating the two strands in a molecule whose ends are not free to rotate. When two intertwined strands are pulled apart from one end, the result is to increase their winding about each other farther along the molecule. The problem can be overcome by introducing a transient nick in one strand. An internal free end allows the nicked strand to rotate about the intact strand, after which the nick can be sealed. Each repetition of the nicking and sealing reaction releases one superhelical turn.

A closed molecule of DNA can be characterized by its linking number, the number of times one strand crosses over the other in space. Closed DNA molecules of identical sequence may have different linking numbers, reflecting different degrees of supercoiling. Molecules of DNA that are the same except for their linking numbers are called topological isomers.

The linking number is made up of two components: the writhing number (W) and the twisting number (T).

The twisting number, T, is a property of the double helical structure itself, representing the rotation of one strand about the other. It represents the total number of turns of the duplex. It is determined by the number of base pairs per turn. For a relaxed closed circular DNA lying flat in a plane, the twist is the total number of base pairs divided by the number of base pairs per turn.

The writhing number, W, represents the turning of the axis of the duplex in space. It corresponds to the intuitive concept of supercoiling, but does not have exactly the same quantitative definition or measurement. For a relaxed molecule, W = 0, and the linking number equals the twist.

We are often concerned with the change in linking number, AL, given by the equation

$$AL = AW + AT$$

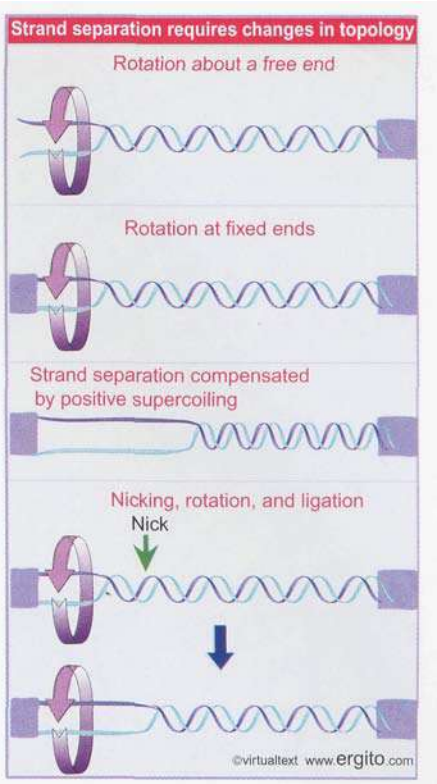


Figure 15.20 Separation of the strands of a DNA double helix could be achieved by several means.

The equation states that any change in the total number of revolutions of one DNA strand about the other can be expressed as the sum of the changes of the coiling of the duplex axis in space (AW) and changes in the screwing of the double helix itself (AT). In a free DNA molecule, W and T are freely adjustable, and any AL (change in linking number) is likely to be expressed by a change in W, that is, by a change in supercoiling.

A decrease in linking number, that is, a change of -AL, corresponds to the introduction of some combination of negative supercoiling and/or underwinding. An increase in linking number, measured as a change of +AL, corresponds to a decrease in negative supercoiling/underwinding.

We can describe the change in state of any DNA by the specific linking difference, CT = $\Delta L/L_0$, where L_0 is the linking number when the DNA is relaxed. If all of the change in linking number is due to change in W (that is, AT = 0), the specific linking difference equals the supercoiling density. In effect, σ as defined in terms of $\Delta L/L_0$ can be assumed to correspond to superhelix density so long as the structure of the double helix itself remains constant.

The critical feature about the use of the linking number is that this parameter is an invariant property of any individual closed DNA molecule. The linking number cannot be changed by any deformation short of one that involves the breaking and rejoining of strands. A circular molecule with a particular linking number can express it in terms of different combinations of T and W, but cannot change their sum so long as the strands are unbroken. (In fact, the partition of L between T and W prevents the assignment of fixed values for the latter parameters for a DNA molecule in solution.)

The linking number is related to the actual enzymatic events by which changes are made in the topology of DNA. The linking number of a particular closed molecule can be changed only by breaking a strand or strands, using the free end to rotate one strand about the other, and rejoining the broken ends. When an enzyme performs such an action, it must change the linking number by an integer; this value can be determined as a characteristic of the reaction. Then we can consider the effects of this change in terms of AW and AT.

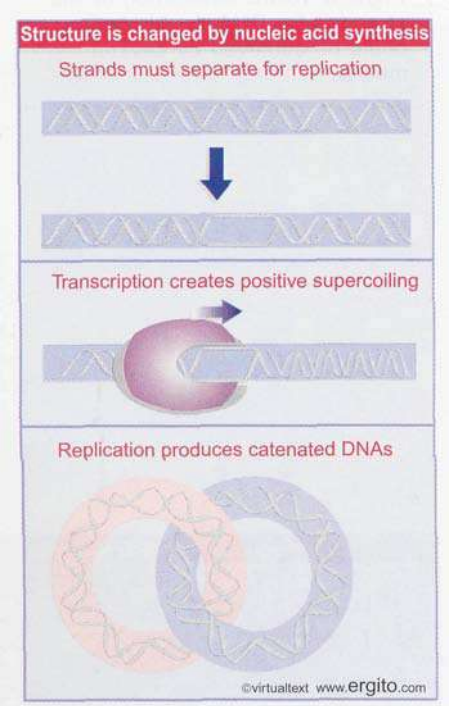


Figure 15.21 The topological structure of DNA is changed during replication and transcription. Strand separation for replication (or transcription) requires a base turn of DNA to be unwound. Transcription creates positive supercoils ahead of the RNA polymerase. Replication of a circular template produces two catenated daughter templates.

15.13 Topoisomerases relax or introduce supercoils in DNA

Key Concepts

- Topoisomerases change the linking number by breaking bonds in DNA, changing the conformation of the double helix in space, and remaking the bonds.
- Type I enzymes act by breaking a single strand of DNA; type II enzymes act by making double-strand breaks.

Changes in the topology of DNA can be caused in several ways. **Figure 15.21** shows some examples. In order to start replication or transcription, the two strands of DNA must be unwound. In the case of replication, the two strands separate permanently and each reforms a duplex with the newly-synthesized daughter strand. In the case of transcription, the movement of RNA polymerase creates a region of positive supercoiling in front and a region of negative supercoiling behind the enzyme. This must be resolved before the positive supercoils impede the movement of the enzyme (see *9.15 Supercoiling is an important feature of transcription).* When a circular DNA molecule is

By Book_Crazy [IND]

replicated, the circular products may be catenated, with one passed through the other. They must be separated in order for the daughter molecules to segregate to separate daughter cells. Yet another situation in which supercoiling is important is the folding of the DNA thread into a chain of nucleosomes in the eukaryotic nucleus (see *20.6 The periodicity of DNA changes on the nucleosome).* All of the situations are resolved by the actions of topoisomerases.

DNA topoisomerases are enzymes that catalyze changes in the topology of DNA by transiently breaking one or both strands of DNA, passing the unbroken strand(s) through the gap, and then resealing the gap. The ends that are generated by the break are never free, but are manipulated exclusively within the confines of the enzyme—in fact, they are covalently linked to the enzyme. Topoisomerases act on DNA irrespective of its sequence, but some enzymes involved in site-specific recombination function in the same way and also fit the definition of topoisomerases (see *15.18 Site-specific recombination resembles topoisomerase activity).*

Topoisomerases are divided into two classes, according to the nature of the mechanisms they employ. Type I topoisomerases act by making a transient break in one strand of DNA. Type II topoisomerases act by introducing a transient double-strand break. Topoisomerases in general vary with regard to the types of topological change they introduce. Some topoisomerases can relax (remove) only negative supercoils from DNA; others can relax both negative and positive supercoils. Enzymes that can introduce negative supercoils are called gyrases; those that can introduce positive supercoils are called reverse gyrases.

There are four topoisomerase enzymes in *E. coli,* called topoisomerase I, III, IV and DNA gyrase. DNA topoisomerase I and III are type I enzymes. Gyrase and DNA topoisomerase IV are type II enzymes. Each of the four enzymes is important in one or more of the situations described in Figure 15.21:

- The overall level of negative supercoiling in the bacterial nucleoid is the result of a *balance* between *the introduction of supercoils by* gyrase ana' their relaxation *by* topoisomerases f and *IV.* This is a crucial aspect of nucleoid structure (see *19.4 The bacterial genome is supercoiled*), and affects initiation of transcription at certain promoters (see *9.15 Supercoiling is an important feature of transcription).*
- The same enzymes are involved in resolving the problems created by transcription; gyrase converts the positive supercoils that are generated ahead of RNA polymerase into negative supercoils, and topoisomerases I and IV remove the negative supercoils that are left behind the enzyme. Similar, but more complicated, effects occur during replication, and the enzymes have similar roles in dealing with them.
- As replication proceeds, the daughter duplexes can become twisted around one another, in a stage known as precatenation. The precatenanes are removed by topoisomerase IV, which also decatenates any catenated genomes that are left at the end of replication. The functions of topoisomerase III partially overlap those of topoisomerase IV.

The enzymes in eukaryotes follow the same principles, although the detailed division of responsibilities may be different. They do not show sequence or structural similarity with the prokaryotic enzymes. Most eukaryotes contain a single topoisomerase I enzyme that is required both for replication fork movement and for relaxing supercoils generated by transcription. A topoisomerase II enzyme(s) is required to unlink chromosomes following replication. Other individual topoisomerases have been *implicated in recombination and repair* activities.

15.14 Topoisomerases break and reseal strands

Key Concepts

- Type I topoisomerases function by forming a covalent bond to one of the broken ends, moving one strand around the other, and then transferring the bound end to the other broken end. Because bonds are conserved, no input of energy is required.

The common action for all topoisomerases is to link one end of each broken strand to a tyrosine residue in the enzyme. A type I enzyme links to the single broken strand; a type II enzyme links to one end of each broken strand. The topoisomerases are further divided into the A and B groups according to whether the linkage is to a 5' phosphate or 3' phosphate. The use of the transient phosphodiester-tyrosine bond suggests a mechanism for the action of the enzyme; it transfers a phosphodiester bond(s) in DNA to the protein, manipulates the structure of one or both DNA strands, and then rejoins the bond(s) in the original strand.

The *E. coli* enzymes are all of type A, using links to 5' phosphate. This is the general pattern for bacteria, where there are almost no type B topoisomerases. All four possible types of topoisomerase (IA, IB, IIA, IIB) are found in eukaryotes.

A model for the action of topoisomerase IA is illustrated in **Figure 15.22**. The enzyme binds to a region in which duplex DNA becomes separated into its single strands; then it breaks one strand, pulls the other strand through the gap, and finally seals the gap. The transfer of bonds from nucleic acid to protein explains how the enzyme can function without requiring any input of energy. There has been no irreversible hydrolysis of bonds; their energy has been conserved through the transfer reactions. The model is supported by the crystal structure of the enzyme.

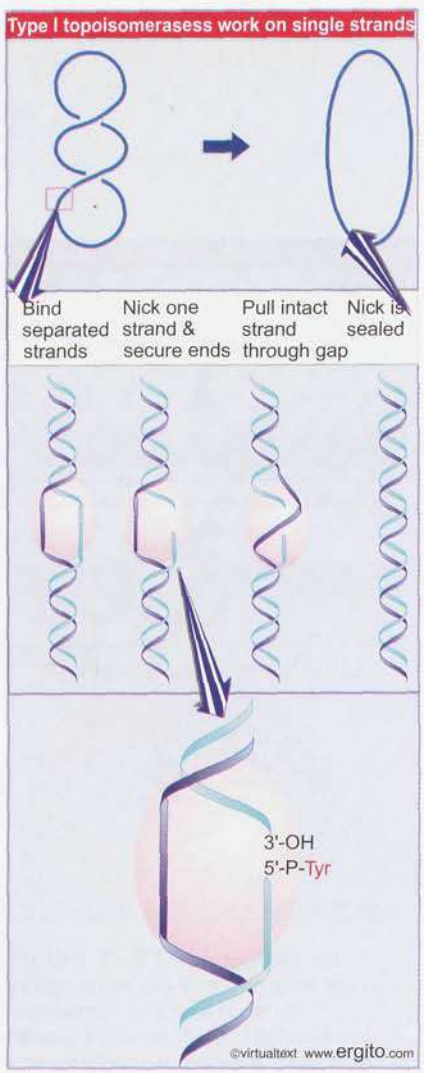


Figure 15.22 Bacterial type I topoisomerases recognize partially unwound segments of DNA and pass one strand through a break made in the other.

The reaction changes the linking number in steps of 1. Each time one strand is passed through the break in the other, there is a AL of +1. The figure illustrates the enzyme activity in terms of moving the individual strands. In a free supercoiled molecule, the interchangeability of W and T should let the change in linking number be taken up by a change of AW = +1, that is, by one less turn of negative supercoiling.

The reaction is equivalent to the rotation illustrated in the bottom part of Figure 15.20, with the restriction that the enzyme limits the reaction to a single-strand passage per event. (By contrast, the introduction of a nick in a supercoiled molecule allows free strand rotation to relieve all the tension by multiple rotations.)

The type I topoisomerase also can pass one segment of a single-stranded DNA through another. This single-strand passage reaction can introduce knots in DNA and can catenate two circular molecules so that they are connected like links on a chain. We do not understand the uses (if any) to which these reactions are put *in vivo*.

Type II topoisomerases generally relax both negative and positive supercoils. The reaction requires ATP, with one ATP hydrolyzed for each catalytic event. As illustrated in **Figure 15.23,** the reaction is mediated by making a double-stranded break in one DNA duplex. The double-strand is cleaved with a 4-base stagger between the ends, and each subunit of the dimeric enzyme attaches to a protruding broken end. Then another duplex region is passed through the break. The ATP is used in the following religation/release step, when the ends are rejoined and the DNA duplexes are released. This is why inhibiting the ATPase activity of the enzyme results in a "cleavable complex" that contains broken DNA.

By Book_Crazy [IND]

A formal consequence of two-strand transfer is that the linking number is always changed in multiples of two. The topoisomerase II activity can be used also to introduce or resolve catenated duplex circles and knotted molecules.

The reaction probably represents a nonspecific recognition of duplex DNA in which the enzyme binds any two double-stranded segments that cross each other. The hydrolysis of ATP may be used to drive the enzyme through conformational changes that provide the force needed to push one DNA duplex through the break made in the other. Because of the topology of supercoiled DNA, the relationship of the crossing segments allows supercoils to be removed from either positively or negatively supercoiled circles.

15.15 Gyrase functions by coil inversion

Key Concepts

- *E. coli* gyrase is a type II topoisomerase that used hydrolysis of ATP to provide energy to introduce negative supercoils into DNA.

Bacterial DNA gyrase is a topoisomerase of type II that is able to introduce negative supercoils into a relaxed closed circular molecule. DNA gyrase binds to a circular DNA duplex and supercoils it processively and catalytically: it continues to introduce supercoils into the same DNA molecule. One molecule of DNA gyrase can introduce ~100 supercoils per minute.

The supercoiled form of DNA has a higher free energy than the relaxed form, and the energy needed to accomplish the conversion is supplied by the hydrolysis of ATP. In the absence of ATP, the gyrase can relax negative but not positive supercoils, although the rate is more than 10X slower than the rate of introducing supercoils.

The *E. coli* DNA gyrase is a tetramer consisting of two types of subunit, each of which is a target for antibiotics (the most often used being nalidixic acid which acts on GyrA, and novobiocin which acts on GyrB). The drugs inhibit replication, which suggests that DNA gyrase is necessary for DNA synthesis to proceed. Mutations that confer resistance to the antibiotics identify the loci that code for the subunits.

Gyrase binds its DNA substrate around the outside of the protein tetramer. Gyrase protects ~140 bp of DNA from digestion by micrococcal nuclease. The sign inversion model for gyrase action is illustrated in **Figure 15.24.** The enzyme binds the DNA in a crossover configuration that is equivalent to a positive supercoil. This induces a compensating negative supercoil in the unbound DNA. Then the enzyme breaks the double strand at the crossover of the positive supercoil, passes the other duplex through, and seals the break.

The reaction directly inverts the sign of the supercoil: it has been converted from a +1 turn to a -1 turn. So the linking number has changed by $\Delta L = -2$, conforming with the demand that all events involving double-strand passage must change the linking number by a multiple of two.

Gyrase then releases one of the crossing segments of the (now negative) bound supercoil; this allows the negative turns to redistribute along DNA (as change in either T or W or both), and the cycle begins again. The same type of topological manipulation is responsible for catenation and knotting.

On releasing the inverted supercoil, the conformation of gyrase changes. For the enzyme to undertake another cycle of supercoiling, its

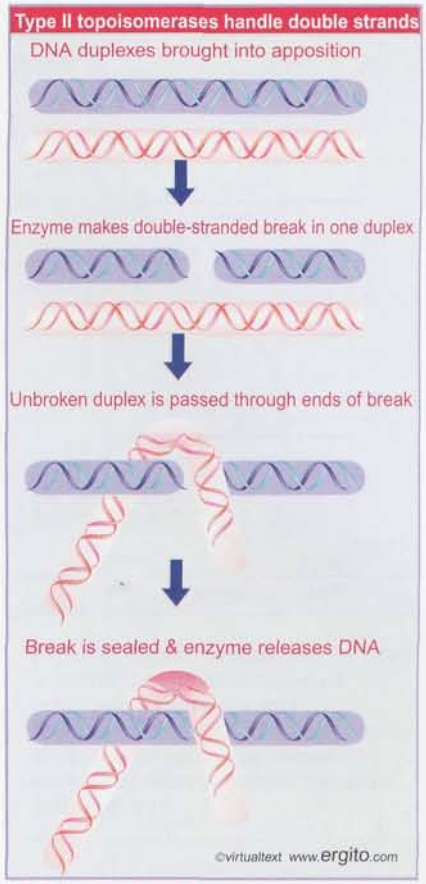


Figure 15.23 Type II topoisomerases can pass a duplex DNA through a double-strand break in another duplex.

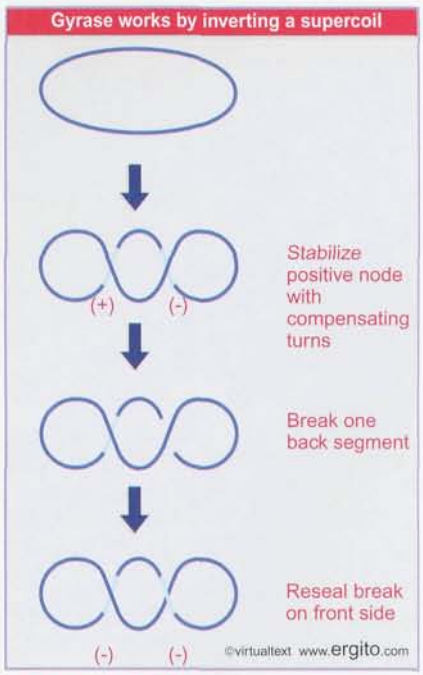


Figure 15.24 DNA gyrase may introduce negative supercoils in duplex DNA by inverting a positive supercoil.

By Book_Crazy [IND]

original conformation must be restored. This process is called enzyme turnover. It is thought to be driven by the hydrolysis of ATP, since the replacement of ATP by an analog that cannot be hydrolyzed allows gyrase to introduce only one inversion (-2 supercoils) per substrate. So the enzyme does not need ATP for the supercoiling reaction, but does need it to undertake a second cycle. Novobiocin interferes with the ATP-dependent reactions of gyrase, by preventing ATP from binding to the B subunit.

The ATP-independent relaxation reaction is inhibited by nalidixic acid. This implicates the A subunit in the breakage and reunion reaction. Treating gyrase with nalidixic acid allows DNA to be recovered in the form of fragments generated by a staggered cleavage across the duplex. The termini all possess a free 3′-OH group and a 4-base 5' single-strand extension covalently linked to the A subunit. The covalent linkage retains the energy of the phosphate bond; this can be used to drive the sealing reaction, explaining why gyrase can undertake relaxation without ATP. The sites of cleavage are fairly specific, occurring about once every 100 bp.

15.16 Specialized recombination involves specific sites

Key Concepts

- Specialized recombination involves reaction between specific sites that are not necessarily homologous.
- Phage lambda integrates into the bacterial chromosome by recombination between a site on the phage and the *att* site on the *E. coli* chromosome.
- The phage is excised from the chromosome by recombination between the sites at the end of the linear prophage.
- Phage lambda *int* codes for an integrase that catalyzes the integration reaction.

Specialized recombination involves a reaction between two specific sites. The target sites are short, typically in a 14-50 bp length range. In some cases the two sites have the same sequence, but in other cases they are nonhomologous. The reaction is used to insert a free phage DNA into the bacterial chromosome or to excise an integrated phage DNA from the chromosome, and in this case the two recombining sequences are different from one another. It is also used before division to regenerate monomeric circular chromosomes from a dimer that has been created by a generalized recombination event (see *13.19 Chromosomal segregation may require site-specific recombination*). In this case the recombining sequences are identical.

The enzymes that catalyze site-specific recombination are generally called recombinases, and >100 of them are now known. Those involved in phage integration or related to these enzymes are also known as the integrase family. Prominent members of the integrase family are the prototype Int from phage lambda, Cre from phage P1, and the yeast FLP enzyme (which catalyzes a chromosomal inversion).

The classic model for site-specific recombination is illustrated by phage lambda. The conversion of lambda DNA between its different life forms involves two types of events. The pattern of gene expression is regulated as described in *12 Phage strategies.* And the physical condition of the DNA is different in the lysogenic and lytic states:

- In the lytic lifestyle, lambda DNA exists as an independent, circular molecule in the infected bacterium.
- In the lysogenic state, the phage DNA is an integral part of the bacterial chromosome (called prophage).

Transition between these states involves site-specific recombination:

- To enter the lysogenic condition, free lambda DNA must be inserted into the host DNA. This is called integration.
- To be released from lysogeny into the lytic cycle, prophage DNA must be released from the chromosome. This is called excision.

Integration and excision occur by recombination at specific loci on the bacterial and phage DNAs called attachment (att) sites. The attachment site on the bacterial chromosome is called att^{λ} in bacterial genetics. The locus is defined by mutations that prevent integration of lambda; it is occupied by prophage λ in lysogenic strains. When the att^{λ} site is deleted from the *E. coli* chromosome, an infecting lambda phage can establish lysogeny by integrating elsewhere, although the efficiency of the reaction is <0.1% of the frequency of integration at att^{λ}. This inefficient integration occurs at secondary attachment sites, which resemble the authentic *att* sequences.

For describing the integration/excision reactions, the bacterial attachment site (att^{λ}) is called *attB*, consisting of the sequence components *BOB'*. The attachment site on the phage, *attP*, consists of the components *POP'*. **Figure 15.25** outlines the recombination reaction between these sites. The sequence O is common to *attB* and *attP*. It is called the core sequence; and the recombination event occurs within it. The flanking regions *B, B'* and *P, P'* are referred to as the arms; each is distinct *in sequence*. Because the phage DNA *is* circular, the recombination event inserts it into the bacterial chromosome as a linear sequence. The prophage is bounded by two new *att* sites, the products of the recombination, called *attL* and *attR*.

An important consequence of the constitution of the *att* sites is that the integration and excision reactions do not involve the same pair of reacting sequences. Integration requires recognition between *attP* and *attB;* while excision requires recognition between *attL* and *attR*. The directional character of site-specific recombination is controlled by the identity of the recombining sites.

Although the recombination event is reversible, different conditions prevail for each direction of the reaction. This is an important feature in the life of the phage, since it offers a means to ensure that an integration event is not immediately reversed by an excision, and vice versa.

The difference in the pairs of sites reacting at integration and excision is reflected by a difference in the proteins that mediate the two reactions:

- Integration *{attB* X *attP)* requires the product of the phage gene *int*, which codes for an integrase enzyme, and a bacterial protein called integration host factor (IHF).
- Excision *{attL* X *attR)* requires the product of phage gene *xis*, in addition to Int and IHF.

So Int and IHF are required for both reactions. Xis plays an important role in controlling the direction; it is required for excision, but inhibits integration.

A similar system, but with somewhat simpler requirements for both sequence and protein components, is found in the bacteriophage P1. The Cre recombinase coded by the phage catalyzes a recombination between two target sequences. Unlike phage lambda, where the recombining sequences are different, in phage P1 they are identical. Each

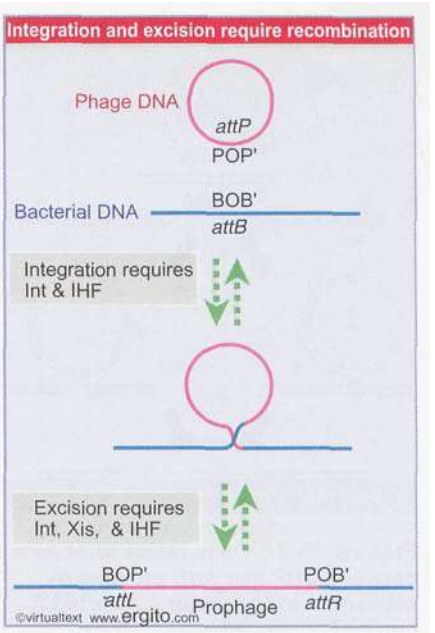


Figure 15.25 Circular phage DNA is converted to an integrated prophage by a reciprocal recombination between *attP* and *attB;* the prophage is excised by reciprocal recombination between *attL* and *attR*.

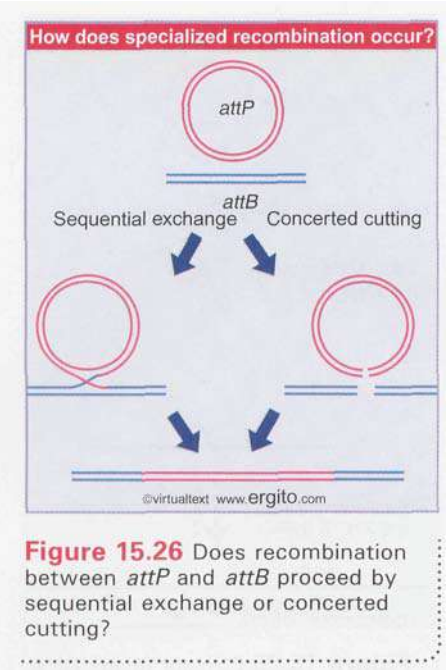


Figure 15.26 Does recombination between *attP* and *attB* proceed by sequential exchange or concerted cutting?

consists of a 34 bp-long sequence called *loxP*. The Cre recombinase is sufficient for the reaction; no accessory proteins are required. Because of its simplicity and its efficiency, what is now known as the Cre/*lox* system has been adapted for use in eukaryotic cells, where it has become one of the standard techniques for undertaking site-specific recombination.

15.17 Site-specific recombination involves breakage and reunion

Key Concepts

- Cleavages staggered by 7 bp are made in both *attB* and *attP* and the ends are joined cross-wise.

The *att* sites have distinct sequence requirements, and *attP* is much larger than *attB*. The function of *attP* requires a stretch of 240 bp, but the function *of attB* can be exercised by the 23 bp fragment extending from –11 to +11, in which there are only 4 bp on either side of the core. The disparity in their sizes suggests that *attP* and *attB* play different roles in the recombination, with *attP* providing additional information necessary to distinguish it from *attB*.

Does the reaction proceed by a concerted mechanism in which the strands in *attP* and *attB* are cut simultaneously and exchanged? Or are the strands exchanged one pair at a time, the first exchange generating a Holliday junction, the second cycle of nicking and ligation occurring to release the structure? The alternatives are depicted in **Figure 15.26.**

The recombination reaction has been halted at intermediate stages by the use of "suicide substrates," in which the core sequence is nicked. The presence of the nick interferes with the recombination process. This makes it possible to identify molecules in which recombination has commenced but has not been completed. The structures of these intermediates suggest that exchanges of single strands take place sequentially.

The model illustrated in **Figure 15.27** shows that if *attP* and *attB* sites each suffer the same staggered cleavage, complementary single-

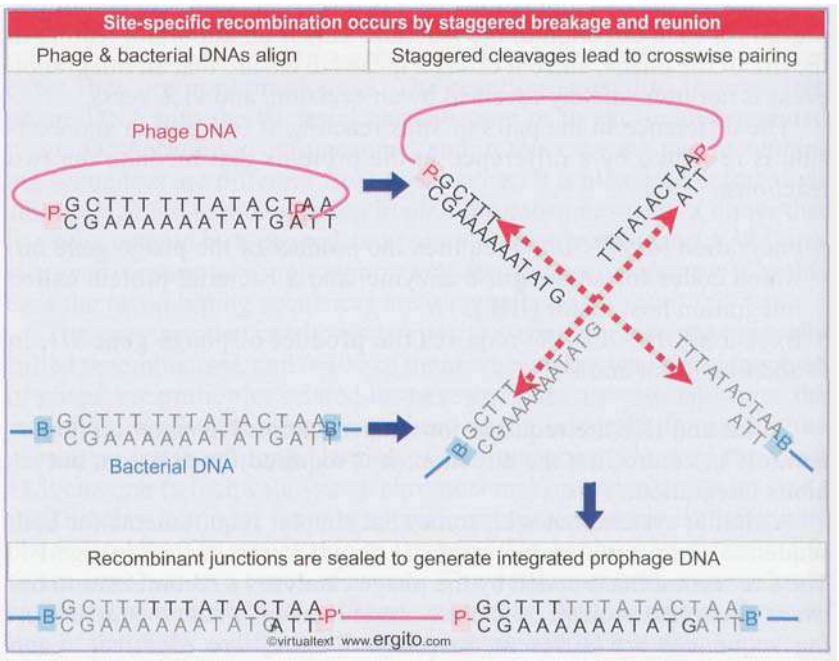


Figure 15.27 Staggered cleavages in the common core sequence of *attP* and *attB* allow crosswise reunion to generate reciprocal recombinant junctions.

By Book_Crazy [IND]

stranded ends could be available for crosswise hybridization. The distance between the lambda crossover points is 7 bp, and the reaction generates 3′-phosphate and 5'-OH ends. The reaction is shown for simplicity as generating overlapping single-stranded ends that anneal, but actually occurs by a process akin to the recombination event of Figure 15.5. The corresponding strands on each duplex are cut at the same position, the free 3′ ends exchange between duplexes, the branch migrates for a distance of 7 bp along the region of homology, and then the structure is resolved by cutting the other pair of corresponding strands.

15.18 Site-specific recombination resembles topoisomerase activity

Key Concepts

- Integrases are related to topoisomerases, and the recombination reaction resembles topoisomerase action except that nicked strands from *different* duplexes are sealed together.
- The reaction conserves energy by using a catalytic tyrosine in the enzyme to break a phosphodiester bond and link to the broken 3' end.
- Two enzyme units bind to each recombination site and the two dimers synapse to form a complex in which the transfer reactions occur.

Integrases use a mechanism similar to that of type 1 topoisomerases, in which a break is made in one DNA strand at a time. The difference is that a recombinase reconnects the ends cross-wise, whereas a topoisomerase makes a break, manipulates the ends, and then rejoins the original ends. The basic principle of the system is that four molecules of the recombinase are required, one to cut each of the four strands of the two duplexes that are recombining.

Figure 15.28 shows the nature of the reaction catalyzed by an integrase. The enzyme is a monomeric protein that has an active site capable of cutting and ligating DNA. The reaction involves an attack by a tyrosine on a phosphodiester bond. The 3' end of the DNA chain is linked through a phosphodiester bond to a tyrosine in the enzyme. This releases a free 5′ hydroxyl end.

Two enzyme units are bound to each of the recombination sites. At each site, only one of the enzyme active sites attacks the DNA. The symmetry of the system ensures that complementary strands are broken in each recombination site. Then the free 5'-OH end in each site attacks the 3'-phosphotyrosine link in the other site. This generates a Holliday junction.

The structure is resolved when the other two enzyme units (which had not been involved in the first cycle of breakage and reunion) act on the other pair of complementary strands.

The successive interactions accomplish a conservative strand exchange, in which there are no deletions or additions of nucleotides at the exchange site, and there is no need for input of energy. The transient 3'-phosphotyrosine link between protein and DNA conserves the energy of the cleaved phosphodiester bond.

Figure 15.29 shows the reaction intermediate, based on the crystal structure. (Trapping the intermediate was made possible by using a "suicide substrate", consisting of a synthetic DNA duplex with a missing phosphodiester bond, so that the attack by the enzyme does not

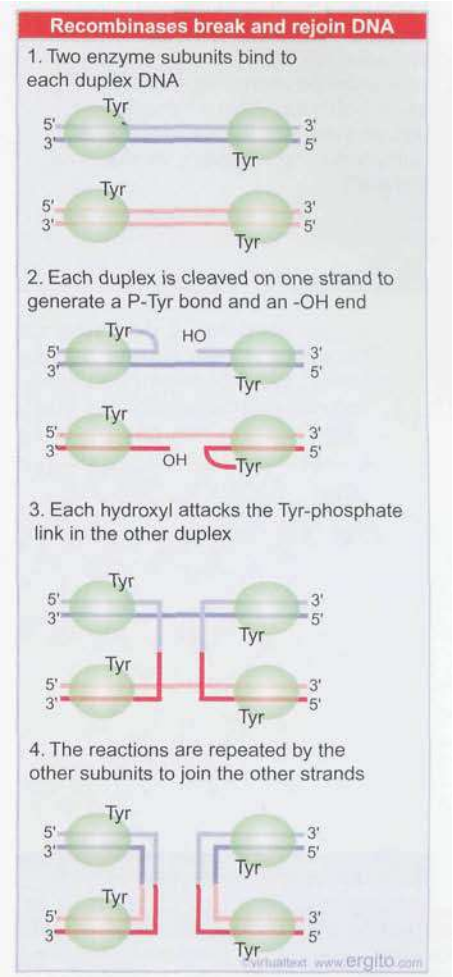


Figure 15.28 Integrases catalyze recombination by a mechanism similar to topoisomerases. Staggered cuts are made in DNA and the 3′-phosphate end is covalently linked to a tyrosine in the enzyme. Then the free hydroxyl group of each strand attacks the P-Tyr link of the other strand. The first exchange shown in the figure generates a Holliday structure. The structure is resolved by repeating the process with the other pair of strands.

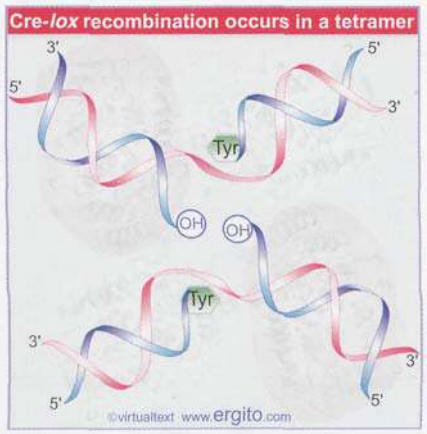


Figure 15.29 A synapsed *loxA* recombination complex has a tetramer of Cre recombinases, with one enzyme monomer bound to each half site. Two of the four active sites are in use, acting on complementary strands of the two DNA sites.

generate a free 5'-OH end.) The structure of the Cre-*lox* complex shows two Cre molecules, each bound to a 15 bp length of DNA. The DNA is bent by ~100° at the center of symmetry. Two of these complexes assemble in an anti-parallel way to form a tetrameric protein structure bound to two synapsed DNA molecules. Strand exchange takes place in a central cavity of the protein structure that contains the central 6 bases of the crossover region.

The tyrosine that is responsible for cleaving DNA in any particular half site is provided by the enzyme subunit that is bound to that half site. This is called *cis* cleavage. This is true also for the Int integrase and XerD recombinase. However, the FLP recombinase cleaves in *trans*, involving a mechanism in which the enzyme subunit that provides the tyrosine is *not* the subunit bound to that half site, but is one of the other subunits.

15.19 Lambda recombination occurs in an intasome

Key Concepts

- Lambda integration takes place in a large complex that also includes the host protein IHF.
- The excision reaction requires Int and Xis and recognizes the ends of the prophage DNA as substrates.

Unlike the Cre/*lox* recombination system, which requires only the enzyme and the two recombining sites, phage lambda recombination occurs in a large structure, and has different components for each direction of the reaction (integration versus excision).

A host protein called IHF is required for both integration and excision. IHF is a 20 kD protein of two different subunits, coded by the genes *him A* and *himD*. IHF is not an essential protein in *E. coli*, and is not required for homologous bacterial recombination. It is one of several proteins with the ability to wrap DNA on a surface. Mutations in the *him* genes prevent lambda site-specific recombination, and can be suppressed by mutations in λ *int*, which suggests that IHF and Int interact. Site-specific recombination can be performed *in vitro* by Int and IHF.

The *in vitro* reaction requires supercoiling in *attP*, but not in *attB*. When the reaction is performed *in vitro* between two supercoiled DNA molecules, almost all of the supercoiling is retained by the products. So there cannot be any free intermediates in which strand rotation could occur. This was one of the early hints that the reaction proceeds through a Holliday junction. We now know that the reaction proceeds by the mechanism typical of this class of enzymes, related to the topoisomerase I mechanism (see previous section).

Int has two different modes of binding. The C-terminal domain behaves like the Cre recombinase. It binds to inverted sites at the core sequence, positioning itself to make the cleavage and ligation reactions on each strand at the positions illustrated in **Figure 15.30.** The N-terminal domain binds to sites in the arms of *attP* that have a different consensus sequence. This binding is responsible for the aggregation of subunits into the intasome. The two domains probably bind DNA simultaneously, thus bringing the arms of *attP* close to the core.

IHF binds to sequences of ~20 bp in *attP*. The IHF binding sites are approximately adjacent to sites where Int binds. Xis binds to two

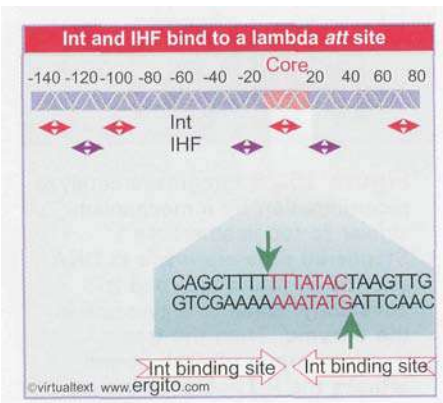


Figure 15.30 Int and IHF bind to different sites in *attP*. The Int recognition sequences in the core region include the sites of cutting.

sites located close to one another in *attP*, so that the protected region extends over 30-40 bp. Together, Int, Xis, and IHF cover virtually all of *attP*. The binding of Xis changes the organization of the DNA so that it becomes inert as a substrate for the integration reaction.

When Int and IHF bind to *attP*, they generate a complex in which all the binding sites are pulled together on the surface of a protein. Supercoiling of *attP* is needed for the formation of this intasome. The only binding sites in *attB* are the two Int sites in the core. But Int does not bind directly to *attB* in the form of free DNA. The intasome is the intermediate that "captures" *attB*, as indicated schematically in **Figure 15.31.**

According to this model, the initial recognition between *attP* and *attB* does not depend directly on DNA homology, but instead is determined by the ability of Int proteins to recognize both *att* sequences. The two *att* sites then are brought together in an orientation predetermined by the structure of the intasome. Sequence homology becomes important at this stage, when it is required for the strand exchange reaction.

The asymmetry of the integration and excision reactions is shown by the fact that Int can form a similar complex with *attR* only if Xis is added. This complex can pair with a condensed complex that Int forms at *attL*. IHF is not needed for this reaction.

Much of the complexity of site-specific recombination may be caused by the need to regulate the reaction so that integration occurs preferentially when the virus is entering the lysogenic state, while excision is preferred when the prophage is entering the lytic cycle. By controlling the amounts of Int and Xis, the appropriate reaction will occur.

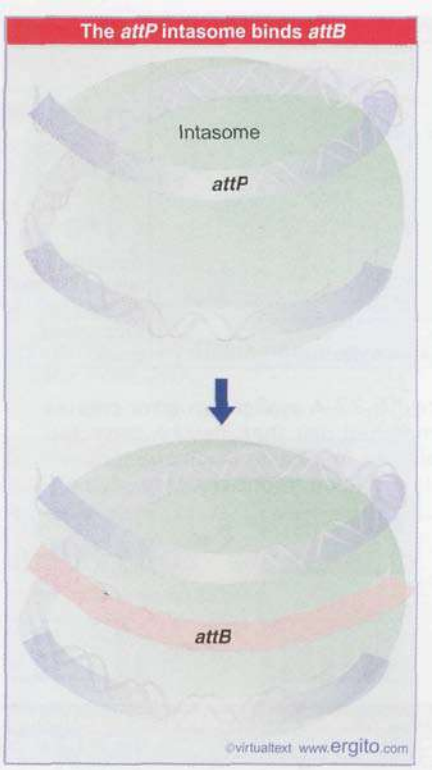


Figure 15.31 Multiple copies of Int protein may organize *attP* into an intasome, which initiates site-specific recombination by recognizing *attB* on free DNA.

15.20 Repair systems correct damage to DNA

Key Concepts

- Repair systems recognize DNA sequences that do not conform to standard base pairs.
- Excision systems remove one strand of DNA at the site of damage and then replace it.
- Recombination-repair systems use recombination to replace the double-stranded region that has been damaged.
- All these systems are prone to introducing errors during the repair process.
- Photoreactivation is a nonmutagenic repair system that acts specifically on pyrimidine dimers.

Any event that introduces a deviation from the usual double-helical structure of DNA is a threat to the genetic constitution of the cell. We can divide such changes into two general classes:

- *Single base changes* affect the sequence but not the overall structure of DNA. They do not affect transcription or replication, when the strands of the DNA duplex are separated. So these changes exert their damaging effects on future generations through the consequences of the change in DNA sequence. The cause of this type of effect is the conversion of one base into another that is not properly paired with the partner base. They may be happen as the result of mutation of a base *in situ* or by replication errors.

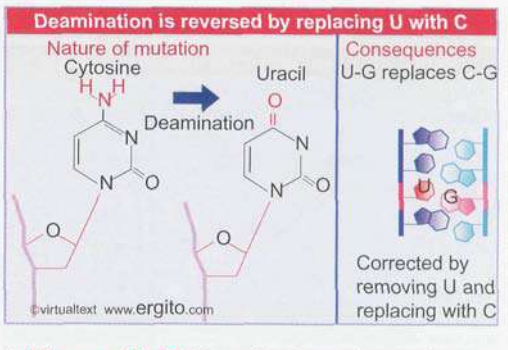


Figure 15.32 Deamination of cytosine creates a U-G base pair. Uracil is preferentially removed from the mismatched pair.

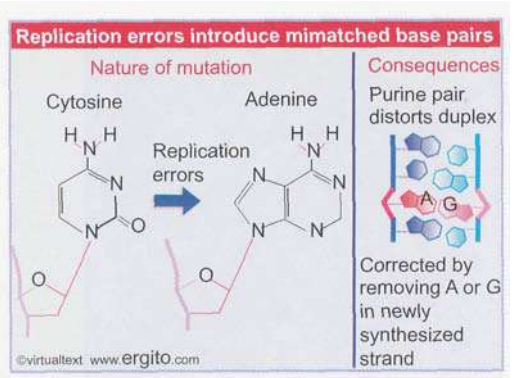


Figure 15.33 A replication error creates a mismatched pair that may be corrected by replacing one base; if uncorrected, a mutation is fixed in one daughter duplex.

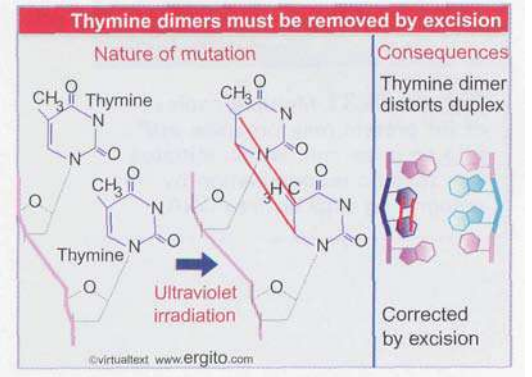


Figure 15.34 Ultraviolet irradiation causes dimer formation between adjacent thymines. The dimer blocks replication and transcription.

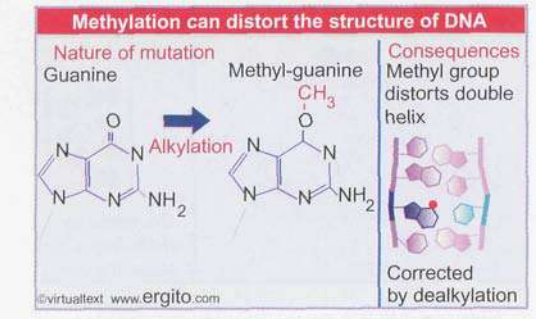


Figure 15.35 Methylation of a base distorts the double helix and causes mispairing at replication.

Figure 15.32 shows that deamination of cytosine to uracil (spontaneously or by chemical mutagen) creates a mismatched U-G pair. **Figure 15.33** shows that a replication error might insert adenine instead of cytosine to create an A-G pair. Similar consequences could result from covalent addition of a small group to a base that modifies its ability to base pair. These changes may result in very minor structural distortion (as in the case of a U-G pair) or quite significant change (as in the case of an A-G pair), but the common feature is that the mismatch persists only until the next replication. So only limited time is available to repair the damage before it is fixed by replication.

- Structural distortions may provide a physical impediment to replication or transcription. Introduction of covalent links between bases on one strand of DNA or between bases on opposite strands inhibits replication and transcription. **Figure 15.34** shows the example of ultraviolet irradiation, which introduces covalent bonds between two adjacent thymine bases, giving an intrastrand pyrimidine dimer. **Figure 15.35** shows that similar consequences could result from addition of a bulky adduct to a base that distorts the structure of the double helix. A single-strand nick or the removal of a base, as shown in **Figure 15.36,** prevents a strand from serving as a proper template for synthesis of RNA or DNA. The common feature in all these changes is that the damaged adduct remains in the DNA, continuing to cause structural problems and/or induce mutations, until it is removed.

Injury to DNA is minimized by systems that recognize and correct the damage. The repair systems are as complex as the replication apparatus itself, which indicates their importance for the survival of the cell. When a repair system reverses a change to DNA, there is no consequence. But a mutation may result when it fails to do so. The measured rate of mutation reflects a balance between the number of damaging events occurring in DNA and the number that have been corrected (or miscorrected).

Repair systems often can recognize a range of distortions in DNA as signals for action, and a cell is likely to have several systems able to deal with DNA damage. The importance of DNA repair in eukaryotes is indicated by the identification of >130 repair genes in the human genome. We may divide them into several general types, as summarized in **Figure 15.37:**

- Some enzymes directly reverse specific sorts of damage to DNA.
- There are pathways for base excision repair, nucleotide excision repair, and mismatch repair, all of which function by removing and replacing material.
- There are systems that function by using recombination to retrieve an undamaged copy that is used to replace a damaged duplex sequence.
- The nonhomologous end-joining pathway rejoins broken double-stranded ends.
- Several different DNA polymerases may be involved in resynthesizing stretches of replacement DNA.

Direct repair is rare and involves the reversal or simple removal of the damage. Photoreactivation of pyrimidine dimers, in which the offending covalent bonds are reversed by a light-dependent enzyme, is a good example. This system is widespread in nature, and appears to be especially important in plants. In *E. coli* it depends on the product of a single gene (*phr*) that codes for an enzyme called photolyase.

Mismatches between the strands of DNA are one of the major targets for repair systems. Mismatch repair is accomplished by scrutiniz-

ing DNA for apposed bases that do not pair properly. Mismatches that arise during replication are corrected by distinguishing between the "new" and "old" strands and preferentially correcting the sequence of the newly synthesized strand. Mismatches also occur when hybrid DNA is created during recombination, and their correction upsets the ratio of parental alleles (see Figure 15.18). Other systems deal with mismatches generated by base conversions, such as the result of deamination. The importance of these systems is emphasized by the fact that cancer is caused in human populations by mutation of genes related to those involved in mismatch repair in yeast.

Mismatches are usually corrected by excision repair, which is initiated by a recognition enzyme that sees an actual damaged base or a change in the spatial path of DNA. There are two types of excision repair system.

- *Base excision repair* systems directly remove the damaged base and replace it in DNA. A good example is DNA uracil glycolase, which removes uracils that are mispaired with guanines (see *15.22 Base flipping is used by methylases and glycosylases).*
- *Nucleotide excision repair* systems excise a sequence that includes the damaged base(s); then a new stretch of DNA is synthesized to replace the excised material. **Figure 15.38** summarizes the main events in the operation of such a system. Such systems are common. Some recognize general damage to DNA. Others act upon specific types of base damage. There are often multiple excision repair systems in a single cell type.

Recombination-repair systems handle situations in which damage remains in a daughter molecule, and replication has been forced to bypass the site, typically creating a gap in the daughter strand. A retrieval system uses recombination to obtain another copy of the sequence from an undamaged source; the copy is then used to repair the gap.

A major feature in recombination and repair is the need to handle double-strand breaks. DSBs initiate crossovers in homologous recombination. They can also be created by problems in replication, when they may trigger the use of recombination-repair systems. When DSBs are created by environmental damage (for example, by radiation damage) or because of the shortening of telomeres (see *30.28 Genetic instability is a key event in cancer),* they can cause mutations. One system for handling DSBs can join together nonhomologous DNA ends.

Mutations that affect the ability of *E. coli* cells to engage in DNA repair fall into groups, which correspond to several repair pathways (not necessarily all independent). The major known pathways are the *uvr* excision repair system, the methyl-directed mismatch-repair system, and the *recB* and *recF* recombination and recombination-repair pathways. The enzyme activities associated with these systems are endonucleases and exonucleases (important in removing damaged DNA), resolvases (endonucleases that act specifically on recombinant junctions), helicases to unwind DNA, and DNA polymerases to synthesize new DNA. Some of these enzyme activities are unique to particular repair pathways, but others participate in multiple pathways.

The replication apparatus devotes a lot of attention to quality control. DNA polymerases use proofreading to check the daughter strand sequence and to remove errors. Some of the repair systems are less accurate when they synthesize DNA to replace damaged material. For this reason, these systems have been known historically as error-prone systems.

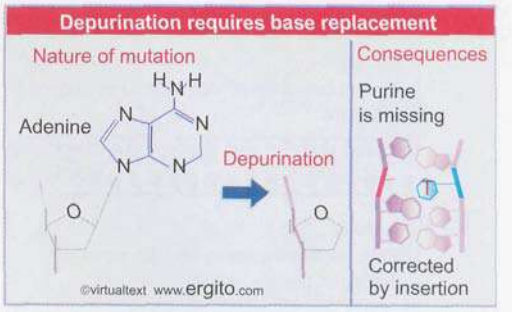


Figure 15.36 Depurination removes a base from DNA.

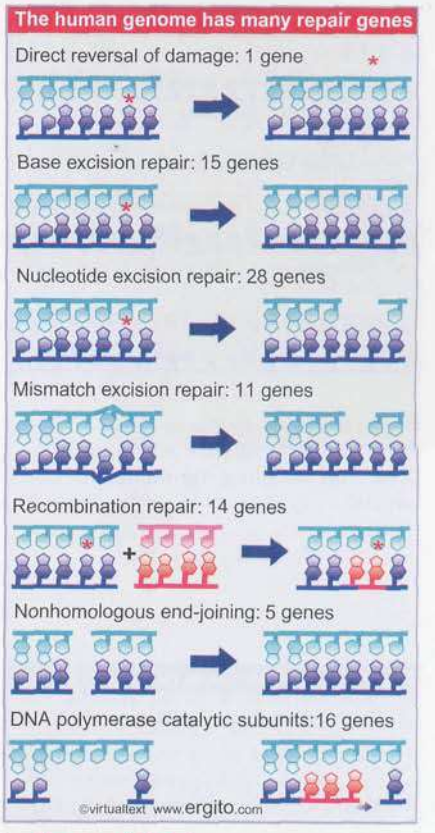


Figure 15.37 Repair genes can be classified into pathways that use different mechanisms to reverse or bypass damage to DNA.

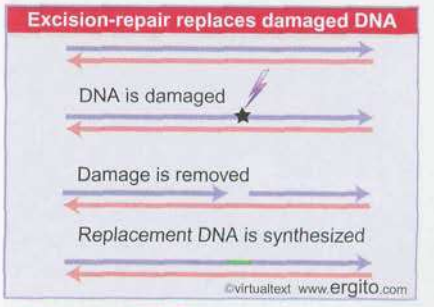


Figure 15.38 Excision-repair directly replaces damaged DNA and then resynthesizes a replacement stretch for the damaged strand.

By Book_Crazy [IND]

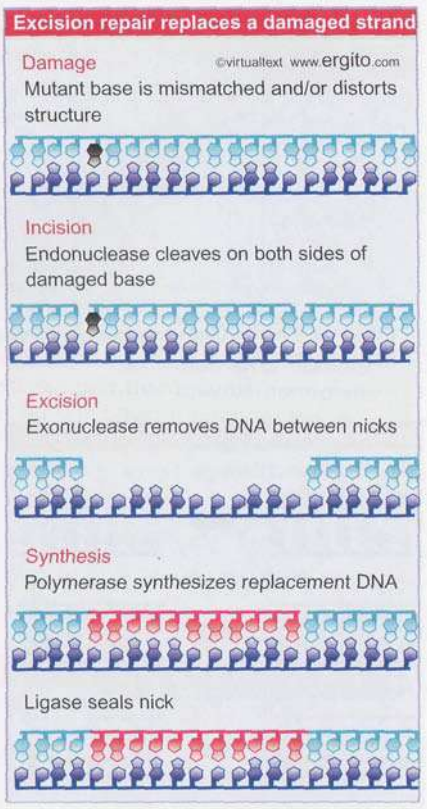


Figure 15.39 Excision-repair removes and replaces a stretch of DNA that includes the damaged base(s).

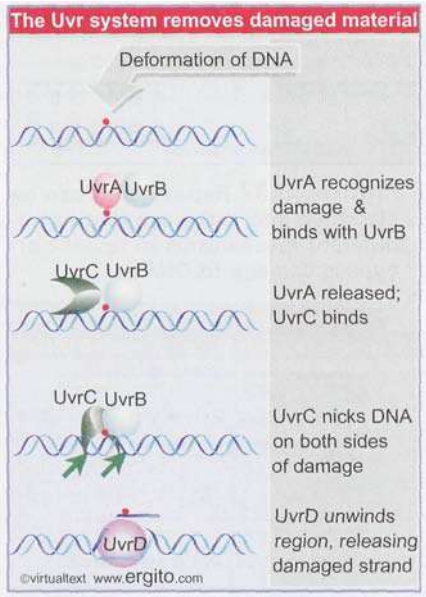


Figure 15.40 The Uvr system operates in stages in which UvrAB recognizes damage, UvrBC nicks the DNA, and UvrD unwinds the marked region.

When the repair systems are eliminated, cells become exceedingly sensitive to ultraviolet irradiation. The introduction of UV-induced damage has been a major test for repair systems, and so in assessing their activities and relative efficiencies, we should remember that the emphasis might be different if another damaged adduct were studied.

15.21 Excision repair systems in *E. coli*

Key Concepts

- The Uvr system makes incisions ~ 12 bases apart on both sides of damaged DNA, removes the DNA between them, and resynthesizes new DNA.

Excision repair systems vary in their specificity, but share the same general features. Each system removes mispaired or damaged bases from DNA and then synthesizes a new stretch of DNA to replace them. The main type of pathway for excision repair is illustrated in **Figure 15.39.**

In the incision step, the damaged structure is recognized by an endonuclease that cleaves the DNA strand on both sides of the damage.

In the excision step, a 5′-3′ exonuclease removes a stretch of the damaged strand.

In the *synthesis* step, the resulting single-stranded region serves as a template for a DNA polymerase to synthesize a replacement for the excised sequence. (Synthesis of the new strand could be associated with removal of the old strand, in one coordinated action.) Finally, DNA ligase covalently links the 3' end of the new material to the old material.

The *uvr* system of excision repair includes three genes, *uvrA,B,C,* that code for the components of a repair endonuclease. It functions in the stages indicated in **Figure 15.40**. First, a UvrAB combination recognizes pyrimidine dimers and other bulky lesions. Then UvrA dissociates (this requires ATP), and UvrC joins UvrB. The UvrBC combination makes an incision on each side, one 7 nucleotides from the 5' side of the damaged site, and the other 3-4 nucleotides away from the 3' side. This also requires ATP. UvrD is a helicase that helps to unwind the DNA to allow release of the single strand between the two cuts. The enzyme that excises the damaged strand is DNA polymerase I. The enzyme involved in the repair synthesis probably also is DNA polymerase I (although DNA polymerases II and III can substitute for it). The events are basically the same, although their order is different, in the eukaryotic repair pathway shown in Figure 15.53.

UvrABC repair accounts for virtually all of the excision repair events in *E. coli.* In almost all (99%) of cases, the average length of replaced DNA is ~12 nucleotides. (For this reason, this is sometimes described as short-patch repair). The remaining 1% involve the replacement of stretches of DNA mostly ~1500 nucleotides long, but extending up to >9000 nucleotides (sometimes called long-patch repair). We do not know why some events trigger the long-patch rather than short-patch mode.

The Uvr complex can be directed to damaged sites by other proteins. Damage to DNA may prevent transcription, but the situation is handled by a protein called Mfd that displaces the RNA polymerase and recruits the Uvr complex (see Figure 21.18 in *21.12 A connection between transcription and repair).*

15.22 Base flipping is used by methylases and glycosylases

Key Concepts

- Uracil and alkylated bases are recognized by glycosylases and removed directly from DNA.
- Pyrimidine dimers are reversed by breaking the covalent bonds between them.
- Methylases add a methyl group to cytosine.
- All these types of enzyme act by flipping the base out of the double helix, where, depending on the reaction, it is either removed or is modified and returned to the helix.

As an alternative to the conventional removal of part of a polynucleotide chain by nuclease activity, glycosylases and lyases can remove bases from the chain. **Figure 15.41** shows that a glycosylase cleaves the bond between the damaged or mismatched base and the deoxyribose. **Figure 15.42** shows that some glycosylases are also lyases that can take the reaction a stage further by using an amino (NH_2) group to attack the deoxyribose ring. Although the results of the glycosylase and lyase reaction appear different, the basic mechanisms of their attack on the DNA are similar.

The interaction of these enzymes with DNA is remarkable. It follows the model first demonstrated for methyltransferases—enzymes that add a methyl group to cytosine in DNA. The methylase flips the target cytosine completely out of the helix. **Figure 15.43** shows that it enters a cavity in the enzyme where it is modified. Then it is returned to its normal position in the helix. All this occurs without input of an external energy source.

One of the most common reactions in which a base is directly removed from DNA is catalyzed by uracil-DNA glycosylase. Uracil occurs in DNA most typically because of a (spontaneous) deamination of cytosine. It is recognized by the glycosylase and removed. The reaction is similar to that of the methylase: the uracil is flipped out of the helix and into the active site in the glycosylase.

Alkylated bases (typically in which a methyl group has been added to a base) are removed by a similar mechanism. A single human enzyme, alkyladenine DNA glycosylase (AAG) recognizes and removes a variety of alkylated substrates, including 3-methyladenine, 7-methylguanine, and hypoxanthine.

By contrast with this mechanism, 1-methyl-adenine is corrected by an enzyme that uses an oxygenating mechanism (coded in *E. coli* by the gene *alkB* which has homologues widely spread through nature, including three genes in man). The methyl group is oxidized to a CH_2OH group, and then the release of the HCHO moiety (formaldehyde) restores the structure of adenine. A very interesting development is the discovery that the bacterial enzyme, and one of the human enzymes, can also repair the same damaged base in RNA. In the case of the human enzyme, the main target may be ribosomal RNA. This is the first known repair event with RNA as a target.

Another enzyme to use base flipping is the photolyase that reverses the bonds between pyrimidine dimers (see Figure 15.34). The pyrimidine dimer is flipped into a cavity in the enzyme. Close to this cavity is an active site that contains an electron donor, which provides the electrons to break the bonds. Energy for the reaction is provided by light in the visible wavelength.

By Book_Crazy [IND]

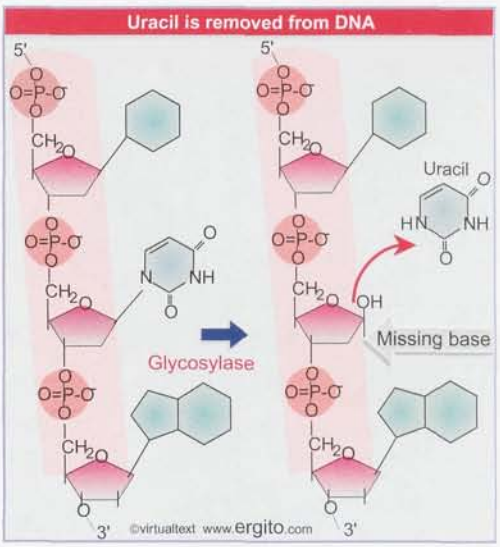


Figure 15.41 A glycosylase removes a base from DNA by cleaving the bond to the deoxyribose.

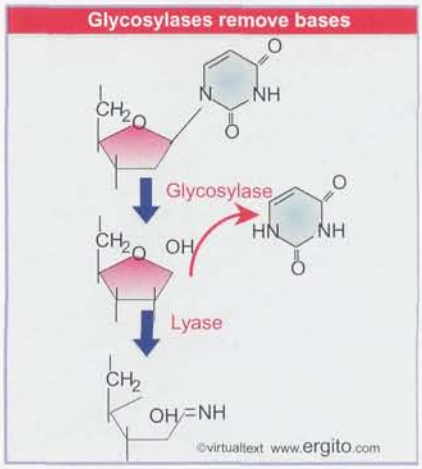


Figure 15.42 A glycosylase hydrolyzes the bond between base and deoxyribose (using H_2O), but a lyase takes the reaction further by opening the sugar ring (using NH_2).

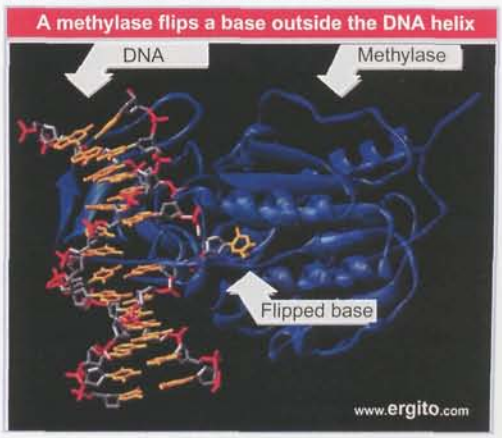


Figure 15.43 A methylase "flips" the target cytosine out of the double helix in order to modify it. Photograph kindly provided by Rich Roberts.

The common feature of these enzymes is the flipping of the target base into the enzyme structure. A variation on this theme is used by T4 endonuclease V, now renamed T4-pdg (pyrimidine dimer glycosylase) to reflect its mode of action. It flips out the adenine base that is *complementary* to the thymine on the 5' side of the pyrimidine dimer. So in this case, the target for the catalytic action of the enzyme remains in the DNA duplex, and the enzyme uses flipping as an indirect mechanism to get access to its target.

When a base is removed from DNA, the reaction is followed by excision of the phosphodiester backbone by an endonuclease, DNA synthesis by a DNA polymerase to fill the gap, and ligation by a ligase to restore the integrity of the polynucleotide chain.

15.23 Error-prone repair and mutator phenotypes

Key Concepts

- Damaged DNA that has not been repaired causes DNA polymerase III to stall during replication.
- DNA polymerase V (coded by *umuCD*), or DNA polymerase IV (coded by *dinB*) can synthesize a complement to the damaged strand.
- The DNA synthesized by the repair DNA polymerase often has errors in its sequence.
- Proteins that affect the fidelity of replication may be identified by mutator genes, in which mutation causes an increased rate of spontaneous mutation.

The existence of repair systems that engage in DNA synthesis raises the question of whether their quality control is comparable with that of DNA replication. So far as we know, most systems, including *uvr*-controlled excision repair, do not differ significantly from DNA replication in the frequency of mistakes. However, error-prone synthesis of DNA occurs in *E. coli* under certain circumstances.

The error-prone feature was first observed when it was found that the repair of damaged λ phage DNA is accompanied by the induction of mutations if the phage is introduced into cells that had previously been irradiated with UV. This suggests that the UV irradiation of the host has activated functions that generate mutations. The mutagenic response also operates on the bacterial host DNA.

What is the actual error-prone activity? It is a DNA polymerase that inserts incorrect bases, which represent mutations, when it passes any site at *which it cannot* insert complementary base pairs in the daughter strand. Functions involved in this error-prone pathway are identified by mutations in the genes *umuD* and *umuC*, which abolish UV-induced mutagenesis. This implies that the UmuC and UmuD proteins cause mutations to occur after UV irradiation. The genes constitute the *umuDC* operon, whose expression is induced by DNA damage. Their products form a complex $UmuD'_2C$, consisting of two subunits of a truncated UmuD protein and one subunit of UmuC. UmuD is cleaved by RecA, which is activated by DNA damage (see *15.27 RecA triggers the SOS system).*

The $UmuD'_2C$ complex has DNA polymerase activity. It is called DNA polymerase V, and is responsible for synthesizing new DNA to replace sequences that have been damaged by UV. This is the only enzyme in *E. coli* that can bypass the classic pyrimidine dimers produced by UV (or other bulky adducts). The polymerase activity is "error-

prone". Mutation of *umuC* or *umuD* inactivate the enzyme, which makes UV irradiation lethal. Some plasmids carry genes called *mucA* and *mucB*, which are homologues of *umuD* and *umuC*, and whose introduction into a bacterium increases resistance to UV killing and susceptibility to mutagenesis.

How does an alternative DNA polymerase get access to the DNA? When the replicase (DNA polymerase III) encounters a block, such as a thymidine dimer, it stalls. Then it is displaced from the replication fork and replaced by DNA polymerase V In fact, DNA polymerase V uses some of the same ancillary proteins as DNA polymerase III. The same situation is true for DNA polymerase IV, the product of *dinB*, which is another enzyme that acts on damaged DNA. DNA polymerases IV and V are part of a larger family, including eukaryotic DNA polymerases, that are involved in repairing damaged DNA (see *15.28 Eukaryotic cells have conserved repair systems).*

15.24 Controlling the direction of mismatch repair

Key Concepts

* The *mut* genes code for a mismatch repair system that deals with mismatched base pairs.
* There is a bias in the selection of which strand to replace at mismatches.
* The strand lacking methylation at a hemimethylated GATC/CTAG is usually replaced.
* This is used to remove errors in a newly synthesized strand of DNA. At G·T and C·T mismatches, the T is preferentially removed.

Genes whose products are involved in controlling the fidelity of DNA synthesis during either replication or repair may be identified by mutations that have a mutator phenotype. A mutator mutant has an increased frequency of spontaneous mutation. If identified originally by the mutator phenotype, a gene is described as *mut;* but often a *mut* gene is later found to be equivalent with a known replication or repair activity.

The general types of activities are identified by *mut* genes fall into two groups:

- The major group consists of components of mismatch-repair systems. Failure to remove a damaged or mispaired base before replication allows it to induce a mutation. Functions in this group include the *dam* methylase that identifies the target for repair, and enzymes that participate directly or indirectly in the removal of particular types of damage *(mutH,S,L,Y).*
- A smaller group, typified by *dnaQ* (which codes for a subunit of DNA polymerase III), is concerned with the accuracy of synthesizing new DNA.

When a structural distortion is removed from DNA, the wild-type sequence is restored. In most cases, the distortion is due to the creation of a base that is not naturally found in DNA, and which is therefore recognized and removed by the repair system.

A problem arises if the target for repair is a mispaired partnership of (normal) bases created when one was mutated. The repair system has no intrinsic means of knowing which is the wild-type base and which is the mutant! All it sees are two improperly paired bases, either of which can provide the target for excision repair.

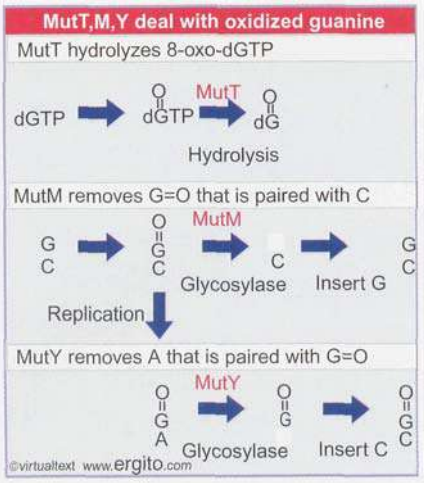


Figure 15.44 Preferential removal of bases in pairs that have oxidized guanine is designed to minimize mutations.

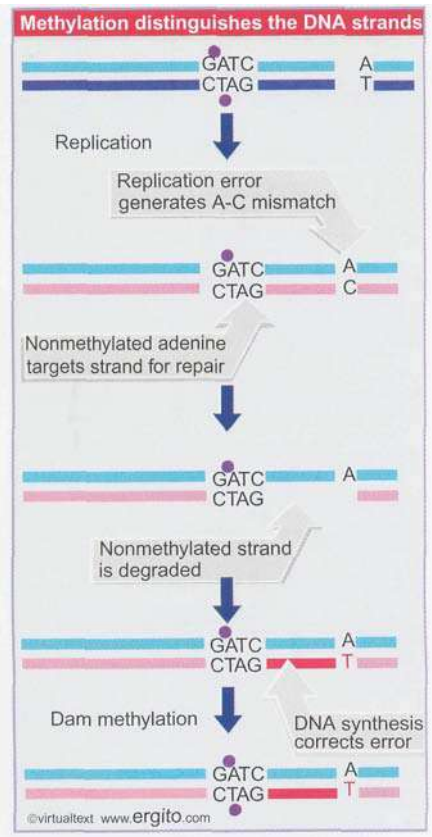


Figure 15.45 GATC sequences are targets for the Dam methylase after replication. During the period before this methylation occurs, the nonmethylated strand is the target for repair of mismatched bases.

If the mutated base is excised, the wild-type sequence is restored. But if it happens to be the original (wild-type) base that is excised, the new (mutant) sequence becomes fixed. Often, however, the direction of excision repair is not random, but is biased in a way that is likely to lead to restoration of the wild-type sequence.

Some precautions are taken to direct repair in the right direction. For example, for cases such as the deamination of 5-methyl-cytosine to thymine, there is a special system to restore the proper sequence (see also *1.14 Many hotspots result from modified bases).* The deamination generates a G·T pair, and the system that acts on such pairs has a bias to correct them to G·C pairs (rather than to A·T pairs). The system that undertakes this reaction includes the *mutL,S* products that remove T from both G-T and C·T mismatches.

The *mutT,M,Y* system handles the consequences of oxidative damage. A major type of chemical damage is caused by oxidation of G to 8-oxo-G. **Figure 15.44** shows that the system operates at three levels. MutT hydrolyzes the damaged precursor (8-oxo-dGTP), which prevents it from being incorporated into DNA. When guanine is oxidized in DNA, its partner is cytosine; and MutM preferentially removes the C from 8-oxo-G·C pairs. Oxidized guanine mispairs with A, and so when 8-oxo-G survives and is replicated, it generates an 8-oxo-G·A pair. MutY removes A from these pairs. MutM and MutY are glycosylases that directly remove a base from DNA. This creates a apurinic site that is recognized by an endonuclease whose action triggers the involvement of the excision repair system.

When mismatch errors occur during replication in *E. coli*, it is possible to distinguish the original strand of DNA. Immediately after replication of methylated DNA, only the original parental strand carries the methyl groups. In the period while the newly synthesized strand awaits the introduction of methyl groups, the two strands can be distinguished.

This provides the basis for a system to correct replication errors. The *dam* gene codes for a methylase whose target is the adenine in the sequence GATC / CTAG (see Figure 14.35). The hemimethylated state is used to distinguish replicated origins from nonreplicated origins. The same target sites are used by a replication-related repair system.

Figure 15.45 shows that DNA containing mismatched base partners is repaired preferentially by excising the strand that lacks the methylation. The excision is quite extensive; mismatches can be repaired preferentially for >1 kb around a GATC site. The result is that the newly synthesized strand is corrected to the sequence of the parental strand.

E. coli dam⁻ mutants show an increased rate of spontaneous mutation. This repair system therefore helps reduce the number of mutations caused by errors in replication. It consists of several proteins, coded by the *mut* genes. MutS binds to the mismatch and is joined by MutL. MutS can use two DNA-binding sites, as illustrated in **Figure 15.46.** The first specifically recognizes mismatches. The second is not specific for sequence or structure, and is used to translocate along DNA until a GATC sequence is encountered. Hydrolysis of ATP is used to drive the translocation. Because MutS is bound to both the mismatch site and to DNA as it translocates, it creates a loop in the DNA.

Recognition of the GATC sequence causes the MutH endonuclease to bind to MutSL. The endonuclease then cleaves the unmethylated strand. This strand is then excised from the GATC site to the mismatch site. The excision can occur in either the 5′-3′ direction (using RecJ or exonuclease VII) or in the 3′-5′ direction (using exonuclease I), assisted by the helicase UvrD. The new DNA strand is synthesized by DNA polymerase III.

The *msh* repair system of *S. cerevisiae* is homologous to the *E. coli mut* system. MSH2 provides a scaffold for the apparatus that recognizes mismatches. MSH3 and MSH6 provide specificity factors. The MSH2-

MSH3 complex binds mismatched loops of 2-4 nucleotides, and the MSH2-MSH6 complex binds to single base mismatches or insertions or deletions. Other proteins are then required for the repair process itself.

Homologues of the MutSL system also are found in higher eukaryotic cells. They are responsible for repairing mismatches that arise as the result of replication slippage. In a region such as a microsatellite where a very short sequence is repeated several times, realignment between the newly synthesized daughter strand and its template can lead to a stuttering in which the DNA polymerase slips backward and synthesizes extra repeating units. These units in the daughter strand are extruded as a single-stranded loop from the double helix (see Figure 4.27). They are repaired by homologues of the MutSL system as shown in **Figure 15.47.**

The importance of the MutSL system for mismatch repair is indicated by the high rate at which it is found to be defective in human cancers. Loss of this system leads to an increased mutation rate (see *30.29 Defects in repair systems cause mutations to accumulate in tumors).*

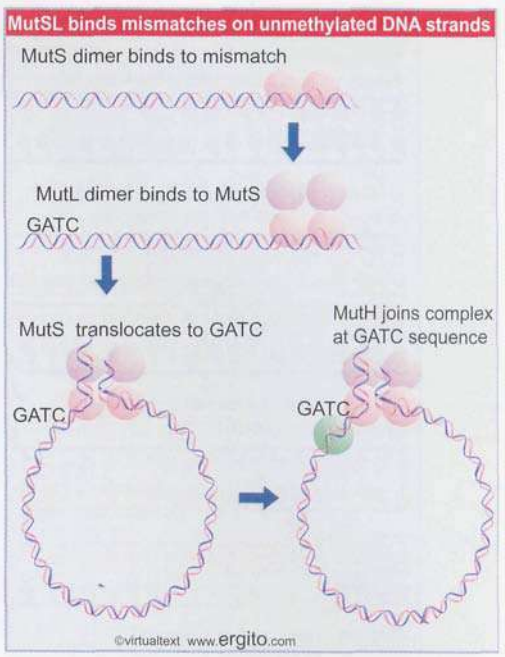


Figure 15.46 MutS recognizes a mismatch and translocates to a GATC site. MutH cleaves the unmethylated strand at the GATC. Endonucleases degrade the strand from the GATC to the mismatch site.

15.25 Recombination-repair systems in *E. coli*

Key Concepts

- The *rec* genes of *E. coli* code for the principal retrieval system.
- It functions when replication leaves a gap in newly synthesized strand opposite a damaged sequence.
- The single strand of another duplex is used to replace the gap.
- Then the damaged sequence is removed and resynthesized.

Recombination-repair systems use activities that overlap with those involved in genetic recombination. They are also sometimes called "post-replication repair," because they function after replication. Such systems are effective in dealing with *the* defects produced in daughter duplexes by replication of a template that contains damaged bases. An example is illustrated in **Figure 15.48.** Restarting stalled replication forks could be the major role of the recombination-repair systems (see *14.17 The primosome is needed to restart replication).*

Consider a structural distortion, such as a pyrimidine dimer, on one strand of a double helix. When the DNA is replicated, the dimer prevents the damaged site from acting as a template. Replication is forced to skip past it.

DNA polymerase probably proceeds up to or close to the pyrimidine dimer. Then the polymerase ceases synthesis of the corresponding daughter strand. Replication restarts some distance farther along. A substantial gap is left in the newly synthesized strand.

The resulting daughter duplexes are different in nature. One has the parental strand containing the damaged adduct, facing a newly synthesized strand with a lengthy gap. The other duplicate has the undamaged parental strand, which has been copied into a normal complementary strand. The retrieval system takes advantage of the normal daughter.

The gap opposite the damaged site in the first duplex is filled by stealing the homologous single strand of DNA from the normal duplex. Following this single-strand exchange, the recipient duplex has a parental (damaged) strand facing a wild-type strand. The donor duplex has a normal parental strand facing a gap; the gap can be filled by repair synthesis in the usual way, generating a normal duplex. So the damage is confined to the original distortion (although the same recombination-repair events must be repeated after every replication cycle unless and until the damage is removed by an excision repair system).

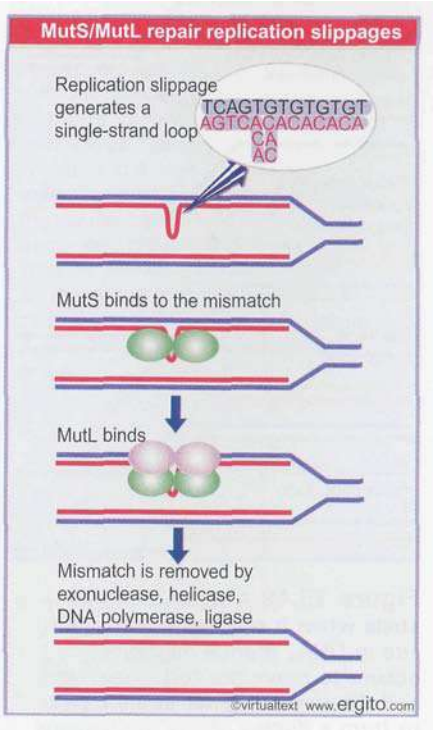


Figure 15.47 The MutS/MutL system initiates repair of mismatches produced by replication slippage.

By Book_Crazy [IND]

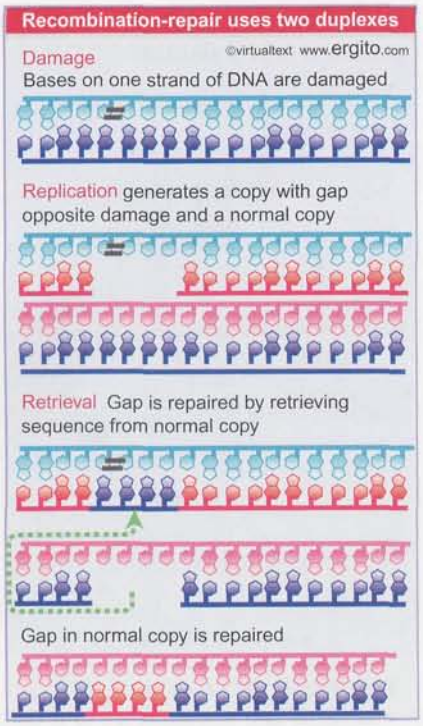


Figure 15.48 An *E. coli* retrieval system uses a normal strand of DNA to replace the gap left in a newly synthesized strand opposite a site of unrepaired damage.

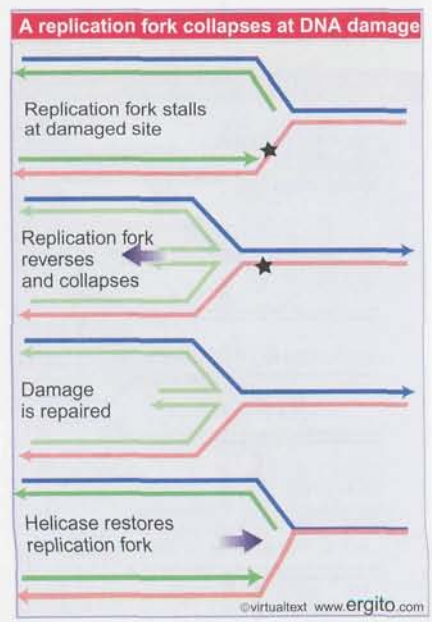


Figure 15.49 A replication fork stalls when it reaches a damaged site in DNA. Branch migration occurs to move the fork backward, and the two daughter strands pair to form a duplex. After the damage has been repaired, a helicase may cause forward branch migration to restore the structure of the fork. Arrowheads indicate 3′ ends.

The principal pathway for recombination-repair in *E. coli* is identified by the *rec* genes (see Figure 15.13, Figure 15.14, Figure 15.15). In *E. coli* deficient in excision repair, mutation in the *recA* gene essentially abolishes all the remaining repair and recovery facilities. Attempts to replicate DNA in *uvr recA* cells produce fragments of DNA whose size corresponds with the expected distance between thymine dimers. This result implies that the dimers provide a lethal obstacle to replication in the absence of RecA function. It explains why the double mutant cannot tolerate > 1 -2 dimers in its genome (compared with the ability of a wild-type bacterium to handle as many as 50).

One *rec* pathway involves the *recBC* genes, and is well characterized; the other involves *recF*, and is not so well defined. They fulfill different functions *in vivo*. The RecBC pathway is involved in restarting stalled replication forks (see next section). The RecF pathway is involved in repairing the gaps in a daughter strand that are left after replicating past a pyrimidine dimer.

The RecBC and RecF pathways both function prior to the action of RecA (although in different ways). They lead to the association of RecA with a single-stranded DNA. The ability of RecA to exchange single strands allows it to perform the retrieval step in Figure 15.48. Nuclease and polymerase activities then complete the repair action.

The RecF pathway contains a group of three genes: *recF*, *recO*, and *recR*. The proteins form two types of complex, RecOR and RecOF. They promote the formation of RecA filaments on single-stranded DNA. One of their functions it to make it possible for the filaments to assemble in spite of the presence of the SSB, which is inhibitory. Although they are thought to function at gaps, the reaction *in vitro* requires a free 5′ end.

The designations of repair and recombination genes are based on the phenotypes of the mutants; but sometimes a mutation isolated in one set of conditions and named as a *uvr* locus turns out to have been isolated in another set of conditions as a *rec* locus. This uncertainty makes an important point. We cannot yet define how many functions belong to each pathway or how the pathways interact. The *uvr* and *rec* pathways are not entirely independent, because *uvr* mutants show reduced efficiency in recombination-repair. We must expect to find a network of nuclease, polymerase, and other activities, constituting repair systems that are partially overlapping (or in which an enzyme usually used to provide some function can be substituted by another from a different pathway).

15.26 Recombination is an important mechanism to recover from replication errors

Key Concepts

- A replication fork may stall when it encounters a damaged site or a nick in DNA.
- A stalled fork may reverse by pairing between the two newly synthesized strands.
- A stalled fork may restart repairing the damage and using a helicase to move the fork forward.
- The structure of the stalled fork is the same as a Holliday junction and may be converted to a duplex and DSB by resolvases.

All cells have many pathways to repair damage in DNA. Which pathway is used will depend upon the type of damage and the

situation. Excision-repair pathways can in principle be used at any time, but recombination-repair can be used only when there is a second duplex with a copy of the damaged sequence, that is, post-replication. A special situation is presented when damaged DNA is replicated, because the replication fork may stall at the site of damage. Recombination-repair pathways are involved in allowing the fork to be restored after the damage has been repaired or to allow it to bypass the damage.

Figure 15.49 shows one possible outcome when a replication fork stalls. The fork stops moving forward when it encounters the damage. The replication apparatus disassembles, at least partially. This allows branch migration to occur, when the fork effectively moves backward, and the new daughter strands pair to form a duplex structure. After the damage has been repaired, a helicase rolls the fork forward to restore its structure. Then the replication apparatus can reassemble, and replication is restarted (see *14.17 The primosome is needed to restart replication).*

Another pathway for handling a stalled repJication fork is provided by recombination-repair. **Figure 15.50** shows that the structure of the stalled fork is essentially the same as a Holliday junction created by recombination between two duplex DNAs. This makes it a target for resolvases. A double-strand break is generated if a resolvase cleaves either pair of complementary strands. In addition, if the damage is in fact a nick, another double-strand break is created at this site.

Stalled replication forks can be rescued by recombination-repair. We don't know the exact sequence of events, but one possible scenario is outlined in **Figure 15.51.** The principle is that a recombination event occurs on either side of the damaged site, allowing an undamaged single strand to pair with the damaged strand. This allows the replication fork to be reconstructed, so that replication can continue, effectively bypassing the damaged site.

In *E. coli,* the RecBC system has an important role in recombination-repair at stalled replication forks (in fact, this may be its major function in the bacterium). RecBC is involved in generating a single strand end on one daughter duplex, which RecA can then cause to pair with the other daughter duplex.

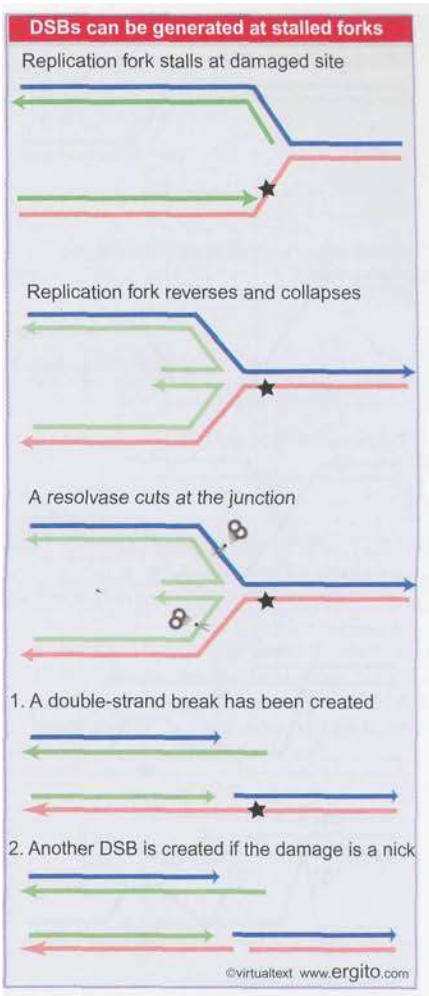


Figure 15.50 The structure of a stalled replication fork resembles a Holliday junction and can be resolved in the same way by resolvases. The results depend on whether the site of damage contains a nick. Result 1 shows that a double-strand break is generated by cutting a pair of strands at the junction. Result 2 shows a second DSB is generated at the site of damage if it contains a nick. Arrowheads indicate 3′ ends.

15.27 RecA triggers the SOS system

Key Concepts

- Damage to DNA causes RecA to trigger the SOS response, consisting of genes coding for many repair enzymes.
- RecA activates the autocleavage activity of LexA.
- LexA represses the SOS system; its autocleavage activates those genes.

The direct involvement of RecA protein in recombination-repair is only one of its activities. This extraordinary protein also has another, quite distinct function. It can be activated by many treatments that damage DNA or inhibit replication in *E. coli.* This causes it to trigger a complex series of phenotypic changes called the SOS response, *which involves the expression of many genes whose products include* repair functions. These dual activities of the RecA protein make *it* difficult to know whether a deficiency in repair in *recA* mutant cells is due to loss of the DNA strand-exchange function of RecA or to some other function whose induction depends on the protease activity.

By Book_Crazy [IND]

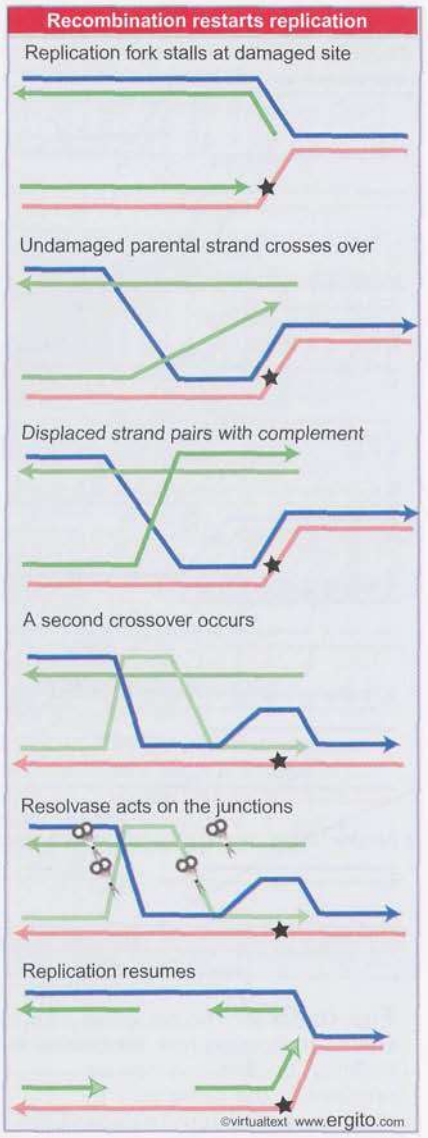


Figure 15.51 When a replication fork stalls, recombination-repair can place an undamaged strand opposite the damaged site. This allows replication to continue.

The inducing damage can take the form of ultraviolet irradiation (the most studied case) or can be caused by crosslinking or alkylating agents. Inhibition of replication by any of several means, including deprivation of thymine, addition of drugs, or mutations in several of the *dna* genes, has the same effect.

The response takes the form of increased capacity to repair damaged DNA, achieved by inducing synthesis of the components of both the long-patch excision repair system and the Rec recombination-repair pathways. In addition, cell division is inhibited. Lysogenic prophages may be induced.

The initial event in the response is the activation of RecA by the damaging treatment. We do not know very much about the relationship between the damaging event and the sudden change in RecA activity. Because a variety of damaging events can induce the SOS response, current work focuses on the idea that RecA is activated by some common intermediate in DNA metabolism.

The inducing signal could consist of a small molecule released from DNA; or it might be some structure formed in the DNA itself. *In vitro*, the activation of RecA requires the presence of single-stranded DNA and ATP. So the activating signal could be the presence of a single-stranded region at a site of damage. Whatever form the signal takes, its interaction with RecA is rapid: the SOS response occurs within a few minutes of the damaging treatment.

Activation of RecA causes proteolytic cleavage of the product of the *lexA* gene. LexA is a small (22 kD) protein that is relatively stable in untreated cells, where it functions as a repressor at many operons. The cleavage reaction is unusual; LexA has a latent protease activity that is activated by RecA. When RecA is activated, it causes LexA to undertake an autocatalytic cleavage; this inactivates the LexA repressor function, and coordinately induces all the operons to which it was bound. The pathway is illustrated in **Figure 15.52.**

The target genes for LexA repression include many repair functions. Some of these SOS genes are active only in treated cells; others are active in untreated cells, but the level of expression is increased by cleavage of LexA. In the case of *uvrB*, which is a component of the excision repair system, the gene has two promoters; one functions independently of LexA, the other is subject to its control. So after cleavage of LexA, the gene can be expressed from the second promoter as well as from the first.

LexA represses its target genes by binding to a 20 bp stretch of DNA called an SOS box, which includes a consensus sequence with 8 absolutely conserved positions. Like other operators, the SOS boxes overlap with the respective promoters. At the *lexA* locus, the subject of autogenous repression, there are two adjacent SOS boxes.

RecA and LexA are mutual targets in the SOS circuit: RecA triggers cleavage of LexA, which represses *recA* and itself. The SOS response therefore causes amplification of both the RecA protein and the LexA repressor. The results are not so contradictory as might at first appear.

The increase in expression of RecA protein is necessary (presumably) for its direct role in the recombination-repair pathways. On induction, the level of RecA is increased from its basal level of ~1200 molecules/cell by up to 50X. The high level in induced cells means there is sufficient RecA to ensure that all the LexA protein is cleaved. This should prevent LexA from reestablishing repression of the target genes.

But the main importance of this circuit for the cell lies in the ability to return rapidly to normalcy. When the inducing signal is removed, the RecA protein loses the ability to destabilize LexA. At this moment, the *lexA* gene is being expressed at a high level; in the absence of activated

RecA, the LexA protein rapidly accumulates in the uncleaved form and turns off the SOS genes. This explains why the SOS response is freely reversible.

RecA also triggers cleavage of other cellular targets, sometimes with more direct consequences. The UmuD protein is cleaved when RecA is activated; the cleavage event activates UmuD and the error-prone repair system. The current model for the reaction is that the $UmuD_2UmuC$ complex binds to a RecA filament near a site of damage, RecA activates the complex by cleaving UmuD to generate UmuD', and the complex then synthesizes a stretch of DNA to replace the damaged material.

Activation of RecA also causes cleavage of some other repressor proteins, including those of several prophages. Among these is the lambda repressor (with which the protease activity was discovered). This explains why lambda is induced by ultraviolet irradiation; the lysogenic repressor is cleaved, releasing the phage to enter the lytic cycle.

This reaction is not a cellular SOS response, but instead represents a recognition by the prophage that the cell is in trouble. Survival is then best assured by entering the lytic cycle to generate progeny phages. In this sense, prophage induction is piggybacking onto the cellular system by responding to the same indicator (activation of RecA).

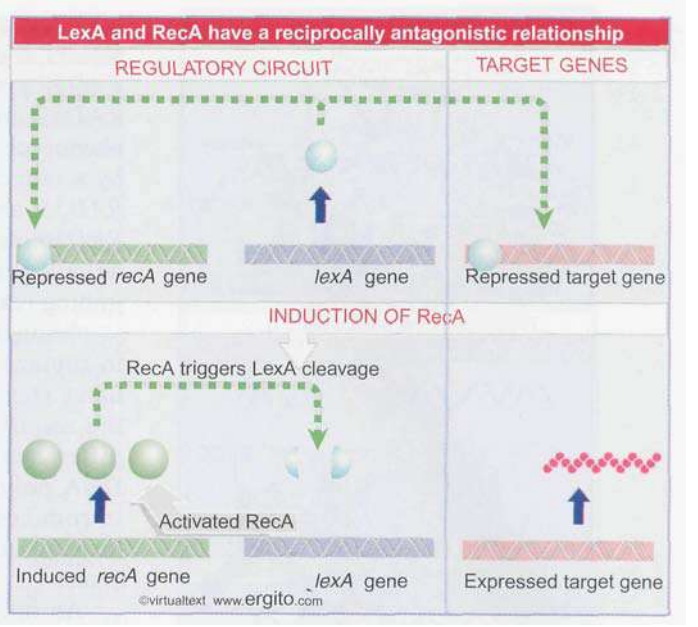


Figure 15.52 The LexA protein represses many genes, including repair functions, *recA* and *lexA*. Activation of RecA leads to proteolytic cleavage of LexA and induces all of these genes.

The two activities of RecA are relatively independent. The *recA441* mutation allows the SOS response to occur without inducing treatment, probably because RecA remains spontaneously in the activated state. Other mutations abolish the ability to be activated. Neither type of mutation affects the ability of RecA to handle DNA. The reverse type of mutation, inactivating the recombination function but leaving intact the ability to induce the SOS response, would be useful in disentangling the direct and indirect effects of RecA in the repair pathways.

15.28 Eukaryotic cells have conserved repair systems

Key Concepts

- The yeast *RAD* mutations, identified by radiation sensitive phenotypes, are in genes that code for repair systems.
- Xeroderma pigmentosum is a human disease caused by mutations in any one of several repair genes.
- Transcriptionally active genes are preferentially repaired.

The types of repair functions recognized in *E. coli* are common to a wide range of organisms. The best characterized eukaryotic systems are in yeast, where Rad51 is the counterpart to RecA. In yeast, the main function of the strand-transfer protein is homologous recombination. Many of the repair systems found in yeast have direct counterparts in higher eukaryotic cells, and in several cases these systems are involved with human diseases (see also *30.29 Defects in repair systems cause mutations to accumulate in tumors).*

Genes involved in repair functions have been characterized genetically in yeast by virtue of their sensitivity to radiation. They are called *RAD* genes. There are three general groups of repair genes in the yeast

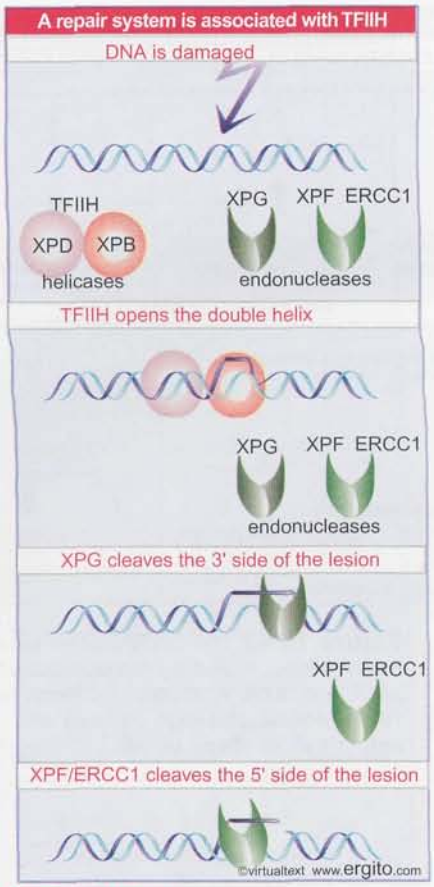


Figure 15.53 A helicase unwinds DNA at a damaged site, endonucleases cut on either side of the lesion, and new DNA is synthesized to replace the excised stretch.

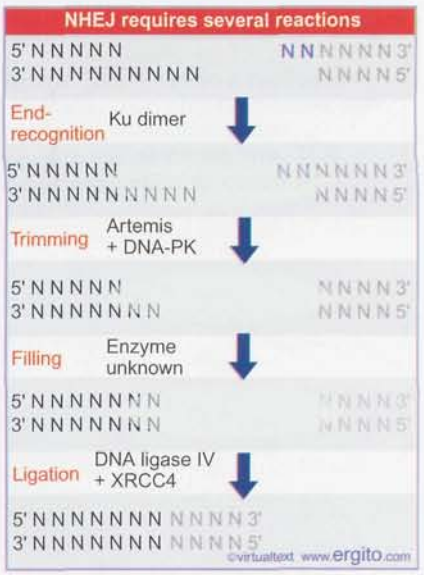


Figure 15.54 Nonhomologous end joining requires recognition of the broken ends, trimming of overhanging ends and/or filling, followed by ligation.

S. cerevisiae, identified by the *RAD3* group (involved in excision repair), the *RAD6* group (required for post-replication repair), and the *RAD52* group (concerned with recombination-like mechanisms). The RAD52 group is divided into two subgroups by a difference in mutant phenotypes. One subgroup affects homologous recombination, as seen by a reduction in mitotic recombination in *RAD50*, *RAD51*, *RAD54*, *RAD55*, and *RAD57*. By contrast, recombination rates are increased in *RAD50*, *MRE11*, and *XRS2* mutants; this subgroup is not deficient in homologous recombination, but is deficient in nonhomologous DNA joining reactions.

A superfamily of DNA polymerases involved in synthesizing DNA to replace material at damaged sites is identified by the *dinB* and *umuCD* genes that code for DNA polymerases IV and V in *E. coli*, and the *rad30* gene coding for DNA polymerase η of *S. cerevisiae*. A difference between the bacterial and yeast enzymes is that the yeast DNA polymerase is not error-prone at thymine dimers: it accurately introduces an A-A pair opposite a T-T dimer. When it replicates through other sites of damage, however, it is more prone to introduce errors.

An interesting feature of repair that has been best characterized in yeast is its connection with transcription. Transcriptionally active genes are preferentially repaired. The consequence is that the transcribed strand is preferentially repaired (removing the impediment to transcription). The cause appears to be a mechanistic connection between the repair apparatus and RNA polymerase. The RAD3 protein, which is a helicase required for the incision step, is a component of a transcription factor associated with RNA polymerase (see *21.12 A connection between transcription and repair*.

Mammalian cells show heterogeneity in the amount of DNA resynthesized at each lesion after damage. However, the patches are always relatively short, <10 bases.

An indication of the existence and importance of the mammalian repair systems is given by certain human hereditary disorders. The best investigated of these is xeroderma pigmentosum (XP), a recessive disease resulting in hypersensitivity to sunlight, in particular to ultraviolet. The deficiency results in skin disorders (and sometimes more severe defects).

The disease is caused by a deficiency in excision repair. Fibroblasts from XP patients cannot excise pyrimidine dimers and other bulky adducts. Mutations fall into 8 genes, called *XP-A* to *XP-G*. They have homologues in the *RAD* genes of yeast, showing that this pathway is a widely used in eukaryotes. Several of the XP products are components of the factor TFIIH, which is involved with the repair of damaged DNA that is encountered by RNA polymerase during transcription. **Figure 15.53** shows its role in this repair pathway. The *XPV gene* is the human homologue of yeast *rad30*. Skin cancers that occur in *XPV* mutants are presumably due to loss of the DNA polymerase η activity that enables replication to occur accurately in spite of T-T dimers.

15.29 A common system repairs double-strand breaks

Key Concepts

- The NHEJ pathway can ligate blunt ends of duplex DNA.
- Mutations in the NHEJ pathway cause human diseases.

Double-strand breaks occur in cells in various circumstances. They initiate the process of homologous recombination and are an intermediate in the recombination of immunoglobulin genes (see *26.9 The RAG proteins catalyze breakage and reunion).* They also occur as the result of damage to DNA, for example, by irradiation. The major mechanism to repair these breaks is called non-homologous end-joining (NHEJ), and consists of ligating the blunt ends together.

The steps involved in NHEJ are summarized in **Figure 15.54.** The same enzyme complex undertakes the process in both NHEJ and immune recombination. The first stage is recognition of the broken ends by a heterodimer consisting of the proteins Ku70 and Ku80. They form a scaffold that holds the ends together and allows other enzymes to act on them. A key component is the DNA-dependent protein kinase (DNA-PKcs), which is activated by DNA to phosphorylate protein targets. One of these targets is the protein Artemis, which in its activated form has both exonuclease and endonuclease activities, and can both trim overhanging ends and cleave the hairpins generated by recombination of immunoglobulin genes. The DNA polymerase activity that fills in any remaining single-stranded protrusions is not known. The actual joining of the double-stranded ends is undertaken by the DNA ligase IV, which functions in conjunction with the protein XRCC4. Mutations in any of these components may render eukaryotic cells more sensitive to radiation. Some of the genes for these proteins are mutated in patients who have diseases due *to deficiencies in* DNA repair.

The Ku heterodimer is the sensor that detects DNA damage by binding to the broken ends. The crystal structure in **Figure 15.55** shows why it binds only to ends. The bulk of the protein extends for about two turns along one face of DNA (lower), but a narrow bridge between the subunits, located in the center of the structure, completely encircles DNA. This means that the heterodimer needs to slip onto a free end.

Ku can bring broken ends together by binding two DNA molecules. The ability of Ku heterodimers to associate with one *another* suggests that the reaction might take place as illustrated in **Figure 15.56.** This would predict that the ligase would act by binding in the region between the bridges on the individual heterodimers. Presumably Ku must change its structure in order to be released from DNA.

Deficiency in DNA repair causes several human diseases. The common feature is that an inability to repair double-strand breaks in DNA leads to chromosomal instability. The instability is revealed by chromosomal aberrations, which are associated with an increased rate of mutation, in turn leading to an increased susceptibility to cancer in patients with the disease. The basic cause can be mutation in pathways that control DNA repair or in the genes that code for enzymes of the repair complexes. The phenotypes can be very similar, as in the case of Ataxia telangiectasia (AT), which is caused by failure of a cell cycle checkpoint pathway, and Nijmegan breakage syndrome (NBS), which is caused by a mutation of a repair enzyme. One of the lessons that we learn from characterizing the repair pathways is that they are conserved in mammals, yeast, and bacteria.

The recessive human disorder of Bloom's syndrome is caused by mutations in a helicase gene (called *BLM)* that is homologous to *recQ* of *E. coli.* The mutation results in an increased frequency of chromosomal breaks and sister *chromatid exchanges. BLM associates with other* repair proteins as part of a large complex. One of the proteins with which it interacts is hMLH1, a mismatch-repair protein that is the human homologue of bacterial *mutL.* The yeast homologues of these two proteins, Sgs1 and MLH1, also associate, identifying these genes as parts of a well-conserved repair pathway.

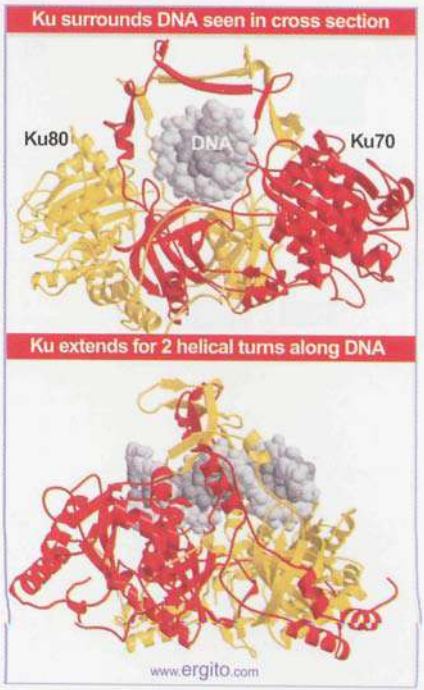


Figure 15.55 The Ku70-Ku80 *heterodimer binds along* two turns of the DNA double helix and surrounds the helix at the center of the binding site. Photograph kindly provided by Jonathan Goldberg.

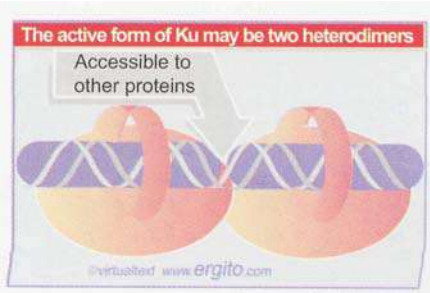


Figure 15.56 If two heterodimers of Ku bind to DNA, the distance *between the two bridges that* encircle DNA is ~12 bp.

By Book_Crazy [IND]

Nijmegan breakage syndrome results from mutations in a gene coding for a protein (variously called Nibrin, p95, or NBS1) that is a component of the Mre11/Rad50 repair complex. Its involvement in repairing double-strand breaks is shown by the formation of foci containing the group of proteins when human cells are irradiated with agents that induce double-strand breaks. After irradiation, the kinase ATMP (coded by the *AT* gene) phosphorylates NBS1; this activates the complex, which localizes at sites of DNA damage. Subsequent steps involve triggering a checkpoint (a mechanism that prevents the cell cycle from proceeding until the damage is repaired; see *15.29 A common system repairs double-strand breaks)*, and recruiting other proteins that are required to repair the damage.

15.30 Summary

Recombination involves the physical exchange of parts between corresponding DNA molecules. This results in a duplex DNA in which two regions of opposite parental origins are connected by a stretch of hybrid (heteroduplex) DNA in which one strand is derived from each parent. Correction events may occur at sites that are mismatched within the hybrid DNA. Hybrid DNA can also be formed without recombination occurring between markers on either side. Gene conversion occurs when an extensive region of hybrid DNA forms during normal recombination (or between nonallelic genes in an aberrant event) and is corrected to the sequence of only one parental strand; then one gene takes on the sequence of the other.

Recombination is initiated by a double-strand break in DNA. The break is enlarged to a gap with a single-stranded end; then the free single-stranded end forms a heteroduplex with the allelic sequence. The DNA in which the break occurs actually incorporates the sequence of the chromosome that it invades, so the initiating DNA is called the recipient. Hot spots for recombination are sites where double strand breaks are initiated. A gradient of gene conversion is determined by the likelihood that a sequence near the free end will be converted to a single strand; this decreases with distance from the break.

Recombination is initiated in yeast by Spo11, a topoisomerase-like enzyme that becomes linked to the free 5' ends of DNA. The DSB is then processed by generating single-stranded DNA that can anneal with its complement in the other chromosome. Yeast mutations that block synaptonemal complex formation show that recombination is required for its formation. Formation of the synaptonemal complex may be initiated by double-strand breaks, and it may persist until recombination is completed. Mutations in components of the synaptonemal complex block its formation but do not prevent chromosome pairing, so homologue recognition is independent of recombination and synaptonemal complex formation.

The full set of reactions required for recombination can be undertaken by the Rec and Ruv proteins of *E. coli*. A single stranded region with a free end is generated by the RecBCD nuclease. The enzyme binds to DNA on one side of a *chi* sequence, and then moves to the *chi* sequence, unwinding DNA as it progresses. A single-strand break is made at the *chi* sequence. *chi* sequences provide hotspots for recombination. The single-strand provides a substrate for RecA, which has the ability to synapse homologous DNA molecules by sponsoring a reaction in which a single strand from one molecule invades a duplex of the other molecule. Heteroduplex DNA is formed by displacing one of the original strands of the duplex. These actions create a recombination junction, which is resolved by the Ruv proteins. RuvA and RuvB act at a heteroduplex, and RuvC cleaves Holliday junctions.

By Book_Crazy [IND]

Recombination, like replication and (probably) transcription, requires topological manipulation of DNA. Topoisomerases may relax (or introduce) supercoils in DNA, and are required to disentangle DNA molecules that have become catenated by recombination or by replication. Type I topoisomerases introduce a break in one strand of a DNA duplex; type II topoisomerases make double-stranded breaks. The enzyme becomes linked to the DNA by a bond from tyrosine to either 5' phosphate (type A enzymes) or 3' phosphate (type B enzymes).

The enzymes involved in site-specific recombination have actions related to those of topoisomerases. Among this general class of recombinases, those concerned with phage integration form the subclass of integrases. The Cre-*lox* system uses two molecules of Cre to bind to each *lox* site, so that the recombining complex is a tetramer. This is one of the standard systems for inserting DNA into a foreign genome. Phage lambda integration requires the phage Int protein and host IHF protein and involves a precise breakage and reunion in the absence of any synthesis of DNA. The reaction involves wrapping of the *attP* sequence of phage DNA into the nucleoprotein structure of the intasome, which contains several copies of Int and IHF; then the host *attB* sequence is bound, and recombination occurs. Reaction in the reverse direction requires the phage protein Xis. Some integrases function by *cis*-cleavage, where the tyrosine that reacts with DNA in a half site is provided by the enzyme subunit bound to that half site; others function by *trans*-cleavage, where a different protein subunit provides the tyrosine.

Bacteria contain systems that maintain the integrity of their DNA sequences in the face of damage or errors of replication and that distinguish the DNA from sequences of a foreign source.

Repair systems can recognize mispaired, altered, or missing bases in DNA, or other structural distortions of the double helix. Excision repair systems cleave DNA near a site of damage, remove one strand, and synthesize a new sequence to replace the excised material. The Uvr system provides the main excision-repair pathway in *E. coli*. The *dam* system is involved in correcting mismatches generated by incorporation of incorrect bases during replication and functions by preferentially removing the base on the strand of DNA that is not methylated at the *dam* target sequence. Eukaryotic homologues of the *E. coli* MutSL system are involved in repairing mismatches that result from replication slippage; mutations in this pathway are common in certain types of cancer.

Recombination-repair systems retrieve information from a DNA duplex and use it to repair a sequence that has been damaged on both strands. The RecBC and RecF pathways both act prior to RecA, whose strand-transfer function is involved in all bacterial recombination. A major use of recombination-repair may be to recover from the situation created when a replication fork stalls.

The other capacity of recA is the ability to induce the SOS response. RecA is activated by damaged DNA in an unknown manner. It triggers cleavage of the LexA repressor protein, thus releasing repression of many loci, and inducing synthesis of the enzymes of both excision repair and recombination-repair pathways. Genes under LexA control possess an operator SOS box. RecA also directly activates some repair activities. Cleavage of repressors of lysogenic phages may induce the phages to enter the lytic cycle.

Repair systems can be connected with transcription in both prokaryotes and eukaryotes. Human diseases are caused by mutations in genes coding for repair activities that are associated with the transcription factor TFIIH. They have homologues in the *RAD* genes of yeast, suggesting that this repair system is widespread.

Nonhomologous end joining (NHEJ) is a general reaction for repairing broken ends in (eukaryotic) DNA. The Ku heterodimer brings the broken ends together so they can be ligated. Several human diseases are caused by mutations in enzymes of this pathway.

By Book_Crazy [IND]

References

15.4 Double-strand breaks initiate recombination

rev Lichten, M. and Goldman, A. S. (1995). Meiotic recombination hotspots. Ann. Rev. Genet. 29, 423-444.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983). The double-strand-break repair model for recombination. Cell 33, 25-35.

ref Hunter, N. and Kleckner, N. (2001). The single-end invasion: an asymmetric intermediate at the double-strand break to double-Holliday junction transition of meiotic recombination. Cell 106, 59-70.

15.5 Recombining chromosomes are connected by the synaptonemal complex

rev Roeder, G. S. (1997). Meiotic chromosomes: it takes two to tango. Genes Dev. 11, 2600-2621.
Zickler, D. and Kleckner, N. (1999). Meiotic chromosomes: integrating structure and function. Ann. Rev. Genet. 33, 603-754.

ref Dong, H. and Roeder, G. S. (2000). Organization of the yeast Zip1 protein within the central region of the synaptonemal complex. J. Cell Biol. 148, 417-426.

Blat, Y. and Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the central region. Cell 98, 249-259.

Klein, F. et al. (1999). A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. Cell 98, 91-103.

Sym, M., Engebrecht, J. A., and Roeder, G. S. (1993). ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. Cell 72, 365-378.

15.6 The synaptonemal complex forms after double-strand breaks

rev McKim, K. S., Jang, J. K., and Manheim, E. A. (2002). Meiotic recombination and chromosome segregation in *Drosophila* females. Ann. Rev. Genet. 36, 205-232.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. 2, 360-369.

ref Allers, T. and Lichten, M. (2001). Differential timing and control of noncrossover and crossover recombination during meiosis. Cell 106, 47-57.
Weiner, B. M. and Kleckner, N. (1994). Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. Cell 77, 977-991.

15.9 Strand-transfer proteins catalyze single-strand assimilation

rev Kowalczykowski, S. C. and Eggleston, A. K. (1994). Homologous pairing and DNA strand-exchange proteins. Ann. Rev. Biochem. 63, 991-1043.
Kowalczykowski, S. C, Dixon, D. A., Eggleston, A. K., Lauder, S. D., and Rehrauer, W. M. (1994). Biochemistry of homologous recombination in Escherichia coli. Microbiol. Rev. 58, 401-465.
Lusetti, S. L. and Cox, M. M. (2002). The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. Ann. Rev. Biochem. 71, 71-100.

15.10 The Ruv system resolves Holliday junctions

rev Lilley, D. M. and White, M. F. (2001). The junction-resolving enzymes. Nat. Rev. Mol. Cell Biol. 2, 433-443.
West, S. C. (1997). Processing of recombination intermediates by the RuvABC proteins. Ann. Rev. Genet. 31, 213-244.

ref Boddy, M. N., Gaillard, P. H., McDonald, W. H., Shanahan, P., Yates, J. R., and Russell, P. (2001). Mus81-Eme1 are essential components of a Holliday junction resolvase. Cell 107, 537-548.

Chen, X. B., Melchionna, R., Denis, C. M., Gaillard, P. H., Blasina, A., Van de Weyer, I., Boddy, M. N., Russell, P., Vialard, J., and McGowan, C. H. (2001). Human Mus81-associated endonuclease cleaves Holliday junctions *in vitro*. Mol. Cell 8, 1117-1127.

Constantinou, A., Davies, A. A., and West, S. C. (2001). Branch migration and Holliday junction resolution catalyzed by activities from mammalian cells. Cell 104, 259-268.

Kaliraman, V., Mullen, J. R., Fricke, W. M., Bastin-Shanower, S. A., and Brill, S. J. (2001). Functional overlap between Sgs1-Top3 and the Mms4-Mus81 endonuclease. Genes Dev. 15, 2730-2740.

15.13 Topoisomerases relax or introduce supercoils in DNA

rev Champoux, J. J. (2001). DNA topoisomerases: structure, function, and mechanism. Ann. Rev. Biochem. 70, 369-413.
Wang, J. C. (2002). Cellular roles of DNA topoisomerases: a molecular perspective. Nat. Rev. Mol. Cell Biol. 3, 430-440.

15.14 Topoisomerases break and reseal strands

rev Champoux, J. J. (2001). DNA topoisomerases: structure, function, and mechanism. Ann. Rev. Biochem. 70, 369-413.

ref Lima, C. D., Wang, J. C, and Mondragon, A. (1994). Three-dimensional structure of the 67K N-termina fragment of *E. coli* DNA topoisomerase I. Nature 367, 138-146.

15.16 Specialized recombination involves specific sites

rev Craig, N. L. (1988). The mechanism of conservative site-specific recombination. Ann. Rev. Genet. 22, 77-105.

ref Metzger, D., Clifford, J., Chiba, H., and Chambon, P. (1995). Conditional site-specific recombination in mammalian cells using a ligand-dependent chimeric Cre recombinase. Proc. Nat. Acad. Sci. USA 92, 6991-6995.

Nunes-Duby, S. E., Kwon, H. J., Tirumalai, R. S., Ellenberger, T., and Landy, A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. Nuc. Acids Res. 26, 391-406.

15.18 Site-specific recombination resembles topoisomerase activity

ref Guo, F., Gopaul, D. N., and van Duyne, G. D. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 389, 40-46.

15.19 Lambda recombination occurs in an intasome

ref Wojciak, J. M., Sarkar, D., Landy, A., and Clubb, R. T. (2002). Arm-site binding by lambda -integrase: solution structure and functional characterization of its amino-terminal domain. Proc. Nat. Acad. Sci. USA 99, 3434-3439.

15.20 Repair systems correct damage to DNA

rev Wood, R. D., Mitchell, M., Sgouros, J., and Lindahl, T. (2001). Human DNA repair genes. Science 291, 1284-1289.

15.22 Base flipping is used by methylases and glycosylases

rev McCullough, A. K., Dodson, M. L., and Lloyd, R. S. (1999). Initiation of base excision repair: glycosylase mechanisms and structures. Ann. Rev. Biochem. 68, 255-285.

ref Aas, P. A., Otterlei, M., Falnes, P. A., Vagbe, C. B., Skorpen, F., Akbari, M., Sundheim, O., Bjoras, M., Slupphaug, G., Seeberg, E., and Krokan, H. E. (2003). Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. Nature 421, 859-863.

Falnes, P. A., Johansen, R. F., and Seeberg, E. (2002). AlkB-mediated oxidative demethylation reverses DNA damage in *E. coli*. Nature 419, 178-182.

Klimasauskas, S., Kumar, S., Roberts, R. J. and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. Cell 76, 357-369.

Lau, A. Y., Scherer, O. D., Samson, L., Verdine, G. L., and Ellenberger, T. (1998). Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: mechanisms for nucleotide flipping and base excision. Cell 95, 249-258.

Lau, A. Y., Glassner, B. J., Samson, L. D., and Ellenberger, T. (2000). Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. Proc. Nat. Acad. Sci. USA 97, 13573-13578.

Mol, D. D. et al. (1995). Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. Cell 80, 869-878.

Park, H. W., Kim, S. T., Sancar, A., and Deisenhofer, J. (1995). Crystal structure of DNA photolyase from *E. coli.* Science 268, 1866-1872.

Savva, R. et al. (1995). The structural basis of specific base-excision repair by uracil-DNA glycosylase. Nature 373, 487-493.

Trewick, S. C, Henshaw, T. F., Hausinger, R. P., Lindahl, T., and Sedgwick, B. (2002). Oxidative demethylation by *E. coli* AlkB directly reverts DNA base damage. Nature 419, 174-178.

Vassylyev, D. G. et al. (1995). Atomic model of a pyrimidine dimer excision repair enzyme complexed with a DNA substrate: structural basis for damaged DNA recognition. Cell 83, 773-782.

15.23 Error-prone repair and mutator phenotypes

ref Friedberg, E. C, Feaver, W. J., and Gerlach, V. L. (2000). The many faces of DNA polymerases: strategies for mutagenesis and for mutational avoidance. Proc. Nat. Acad. Sci. USA 97, 5681-5683.

Goldsmith, M., Sarov-Blat, L., and Livneh, Z. (2000). Plasmid-encoded MucB protein is a DNA polymerase (pol RI) specialized for lesion bypass in the presence of MucA, RecA, and SSB. Proc. Nat. Acad. Sci. USA 97, 11227-11231.

Maor-Shoshani, A., Reuven, N. B., Tomer, G., and Livneh, Z. (2000). Highly mutagenic replication by DNA polymerase V (UmuC) provides a mechanistic basis for SOS untargeted mutagenesis. Proc. Nat. Acad. Sci. USA 97, 565-570.

Wagner, J., Gruz, P., Kim, S. R., Yamada, M., Matsui, K., Fuchs, R. P., and Nohmi, T. (1999). The *dinB* gene encodes a novel *E. coli* DNA polymerase, DNA pol IV, involved in mutagenesis. Mol. Cell 4, 281-286.

15.24 Controlling the direction of mismatch repair

ref Strand, M., Prolla, T. A., Liskay, and Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365, 274-276.

15.25 Recombination-repair systems in *E. coli*

rev West, S. C. (1997). Processing of recombination intermediates by the RuvABC proteins. Ann. Rev. Genet. 31, 213-244.

ref Bork, J. M. and Inman, R. B. (2001). The RecOR proteins modulate RecA protein function at 5' ends of single-stranded DNA. EMBO J. 20, 7313-7322.

15.26 Recombination is an important mechanism to recover from replication errors

rev Michel, B., Viguera, E., Grompone, G., Seigneur, M., and Bidnenko, V. (2001). Rescue of arrested replication forks by homologous recombination. Proc. Nat. Acad. Sci. USA 98, 8181-8188.

ref Cox, M. M., Goodman, M. F., Kreuzer, K. N., Sherratt, D. J., Sandler, S. J., and Marians, K. J. (2000). The importance of repairing stalled replication forks. Nature 404, 37-41.

Kuzminov, A. (2001). Single-strand interruptions in replicating chromosomes cause double-strand breaks. Proc. Nat. Acad. Sci. USA 98, 8241-8246.

McGlynn, P. and Lloyd, R. G. (2002). Recombinational repair and restart of damaged replication forks. Nat. Rev. Mol. Cell Biol. 3, 859-870.

15.27 RecA triggers the SOS system

ref Tang, M. et al. (1999). UmuD′$_2$C is an error-prone DNA polymerase, *E. coli* pol V. Proc. Nat. Acad. Sci. USA 96, 8919-8924.

15.28 Eukaryotic cells have conserved repair systems

rev Prakash, S. and Prakash, L. (2002). Translesion DNA synthesis in eukaryotes: a one- or two-polymerase affair. Genes Dev. 16, 1872-1883.

ref Friedberg, E. C, Feaver, W. J., and Gerlach, V. L. (2000). The many faces of DNA polymerases: strategies for mutagenesis and for mutational avoidance. Proc. Nat. Acad. Sci. USA 97, 5681-5683.

Johnson, R. E., Prakash, S., and Prakash, L. (1999). Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Pol eta. Science 283, 1001-1004.

15.29 A common system repairs double-strand breaks

rev D'Amours, D. and Jackson, S. P. (2002). The Mre11 complex: at the crossroads of DNA repair and checkpoint signalling. Nat. Rev. Mol. Cell Biol. 3, 317-327.

ref Carney, J. P., Maser, R. S., Olivares, H., Davis, E. M., Le Beau, M., Yates, J. R., Hays, L., Morgan, W. F., and Petrini, J. H. (1998). The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. Cell 93, 477-486.

Cary, R. B., Peterson, S. R., Wang, J., Bear, D. G., Bradbury, E. M., and Chen, D. J. (1997). DNA looping by Ku and the DNA-dependent protein kinase. Proc. Nat. Acad. Sci. USA 94, 4267-4272.

Ellis, N. A., Groden, J., Ye, T. Z., Straughen, J., Lennon, D. J., Ciocci, S., Proytcheva, M., and German, J. (1995). The Bloom's syndrome gene product is homologous to RecQ helicases. Cell 83, 655-666.

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M. R. (2002). Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. Cell 108, 781-794.

Ramsden, D. A. and Gellert, M. (1998). Ku protein stimulates DNA end joining by mammalian DNA ligases: a direct role for Ku in repair of DNA double-strand breaks. EMBO J. 17, 609-614.

Varon, R. et al. (1998). Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. Cell 93, 467-476.

Walker, J. R., Corpina, R. A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature 412, 607-614.

Chapter 16

Transposons

16.1 Introduction

Genomes evolve both by acquiring new sequences and by rearranging existing sequences.

The sudden introduction of new sequences results from the ability of vectors to carry information between genomes. Extrachromosomal elements move information horizontally by mediating the transfer of (usually rather short) lengths of genetic material. In bacteria, plasmids move by conjugation (see *13.13 Conjugation transfers single-stranded DNA*), while phages spread by infection (see *12 Phage strategies*). Both plasmids and phages occasionally transfer host genes along with their own replicon. Direct transfer of DNA occurs between some bacteria by means of transformation (see *1.2 DNA is the genetic material of bacteria*). In eukaryotes, some viruses, notably the retroviruses, can transfer genetic information during an infective cycle (see *17.6 Retroviruses may transduce cellular sequences*).

Rearrangements are sponsored by processes internal to the genome. Two of the major causes are summarized in **Figure 16.1**.

Unequal recombination results from mispairing by the cellular systems for homologous recombination. Nonreciprocal recombination results in duplication or rearrangement of loci (see *4.7 Unequal crossing-over rearranges gene clusters*). Duplication of sequences within a genome provides a major source of new sequences. One copy of the sequence can retain its original function, while the other may evolve into a new function. Furthermore, significant differences between individual genomes are found at the molecular level because of polymorphic variations caused by recombination. We saw in *4.14 Minisatellites are useful for genetic mapping* that recombination between minisatellites adjusts their lengths so that every individual genome is distinct.

Another major cause of variation is provided by transposable elements or transposons: these are discrete sequences in the genome that are mobile—they are able to transport themselves to other locations within the genome. The mark of a transposon is that it does not utilize an independent form of the element (such as phage or plasmid DNA), but moves directly from one site in the genome to another. Unlike most other processes involved in genome restructuring, transposition does not rely on any relationship between the sequences at the donor and recipient sites. Transposons are restricted to moving themselves, and sometimes additional sequences, to new sites elsewhere within the same genome; they are therefore an internal counterpart to the vectors that can transport sequences from one genome to another. They may provide the major source of mutations in the genome.

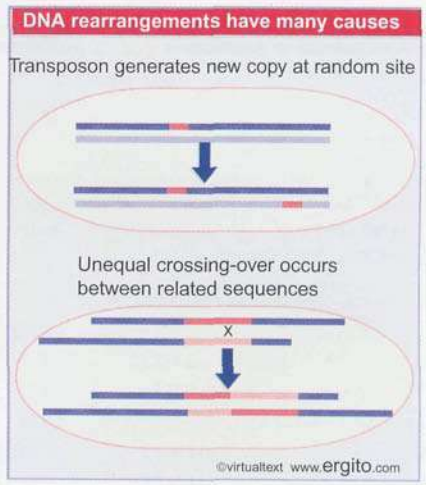


Figure 16.1 A major cause of sequence change within a genome is the movement of a transposon to a new site. This may have direct consequences on gene expression. Unequal crossing-over between related sequences causes rearrangements. Copies of transposons can provide targets for such events.

Transposons fall into two general classes. The groups of transposons reviewed in this chapter exist as sequences of DNA coding for proteins that are able directly to manipulate DNA so as to propagate themselves within the genome. The transposons reviewed in *17 Retroviruses and retroposons* are related to retroviruses, and the source of their mobility is the ability to make DNA copies of their RNA transcripts; the DNA copies then become integrated at new sites in the genome.

Transposons that mobilize via DNA are found in both prokaryotes and eukaryotes. Each bacterial transposon carries gene(s) that code for the enzyme activities required for its own transposition, although it may also require ancillary functions of the genome in which it resides (such as DNA polymerase or DNA gyrase). Comparable systems exist in eukaryotes, although their enzymatic functions are not so well characterized. A genome may contain both functional and nonfunctional (defective) elements. Often the majority of elements in a eukaryotic genome are defective, and have lost the ability to transpose independently, although they may still be recognized as substrates for transposition by the enzymes produced by functional transposons. A eukaryotic genome contains a large number and variety of transposons. The fly genome has >50 types of transposon, with a total of several hundred individual elements.

Transposable elements can promote rearrangements of the genome, directly or indirectly:

- The transposition event itself may cause deletions or inversions or lead to the movement of a host sequence to a new location.
- Transposons serve as substrates for cellular recombination systems by functioning as "portable regions of homology"; two copies of a transposon at different locations (even on different chromosomes) may provide sites for reciprocal recombination. Such exchanges result in deletions, insertions, inversions, or translocations.

The intermittent activities of a transposon seem to provide a somewhat nebulous target for natural selection. This concern has prompted suggestions that (at least some) transposable elements confer neither advantage nor disadvantage on the phenotype, but could constitute "selfish DNA," concerned only with their own propagation. Indeed, in considering transposition as an event that is distinct from other cellular recombination systems, we tacitly accept the view that the transposon is an independent entity that resides in the genome.

Such a relationship of the transposon to the genome would resemble that of a parasite with its host. Presumably the propagation of an element by transposition is balanced by the harm done if a transposition event inactivates a necessary gene, or if the number of transposons becomes a burden on cellular systems. Yet we must remember that any transposition event conferring a selective advantage—for example, a genetic rearrangement—will lead to preferential survival of the genome carrying the active transposon.

16.2 Insertion sequences are simple transposition modules

Key Concepts

- An insertion sequence is a transposon that codes for the enzyme(s) needed for transposition flanked by short inverted terminal repeats.
- The target site at which a transposon is inserted is duplicated during the insertion process to form two repeats in direct orientation at the ends of the transposon.
- The length of the direct repeat is 5-9 bp and is characteristic for any particular transposon.

Transposable elements were first identified at the molecular level in the form of spontaneous insertions in bacterial operons. Such an insertion prevents transcription and/or translation of the gene in which it is inserted. Many different types of transposable elements have now been characterized.

The simplest transposons are called insertion sequences (reflecting the way in which they were detected). Each type is given the prefix IS, followed by a number that identifies the type. (The original classes were numbered IS 1-4; later classes have numbers reflecting the history of their isolation, but not corresponding to the total number of elements so far isolated!)

The IS elements are normal constituents of bacterial chromosomes and plasmids. A standard strain of *E. coli* is likely to contain several (<10) copies of any one of the more common IS elements. To describe an insertion into a particular site, a double colon is used; so λ::IS1 describes an IS1 element inserted into phage lambda.

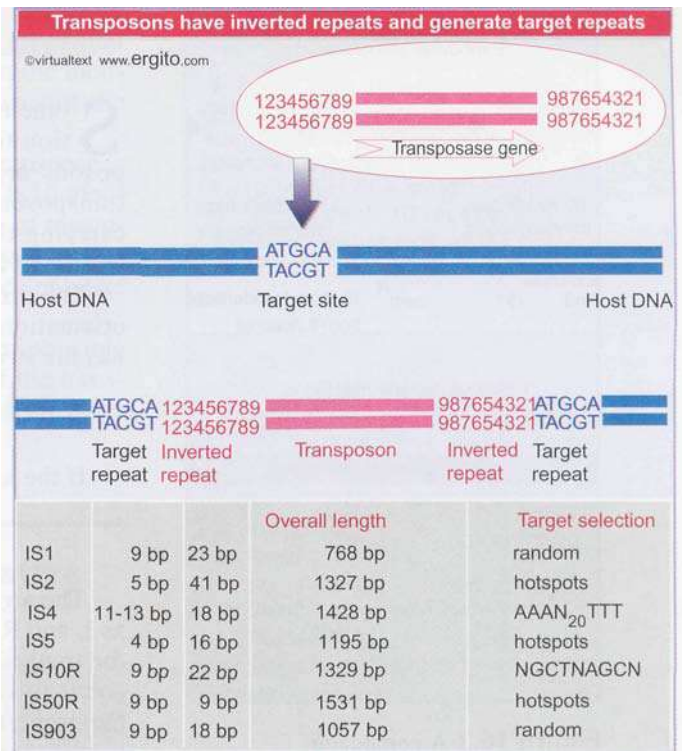
The IS elements are autonomous units, each of which codes only for the proteins needed to sponsor its own transposition. Each IS element is different in sequence, but there are some common features in organization. The structure of a generic transposon before and after insertion at a target site is illustrated in **Figure 16.2,** which also summarizes the details of some common IS elements.

An IS element ends in short inverted terminal repeats; usually the two copies of the repeat are closely related rather than identical. As illustrated in the figure, the presence of the inverted terminal repeats means that the same sequence is encountered proceeding toward the element from the flanking DNA on either side of it.

When an IS element transposes, a sequence of host DNA at the site of insertion is duplicated. The nature of the duplication is revealed by comparing the sequence of the target site before and after an insertion has occurred. Figure 16.2 shows that at the site of insertion, the IS DNA is always flanked by very short direct repeats. (In this context, "direct" indicates that two copies of a sequence are repeated in the same orientation, not that the repeats are adjacent.) But in the original gene (prior to insertion), the target site has the sequence of only one of these repeats. In the figure, the target site consists of the sequence ATGCA/TACGT. After transposition, one copy of this sequence is present on either side of the transposon.

The sequence of the direct repeat varies among individual transposition events undertaken by a transposon, but the length is constant for any particular IS element (a reflection of the mechanism of transposition). The most common length for the direct repeats is 9 bp.

An IS element therefore displays a characteristic structure in which its ends are identified by the inverted terminal repeats, while the adjacent ends of the flanking host DNA are identified by the short direct repeats. When observed in a sequence of DNA, this type of organization is taken to be diagnostic of a transposon, and makes a prima facie case that the sequence originated in a transposition event.



			Overall length	Target selection
IS1	9 bp	23 bp	768 bp	random
IS2	5 bp	41 bp	1327 bp	hotspots
IS4	11-13 bp	18 bp	1428 bp	$AAAN_{20}TTT$
IS5	4 bp	16 bp	1195 bp	hotspots
IS10R	9 bp	22 bp	1329 bp	NGCTNAGCN
IS50R	9 bp	9 bp	1531 bp	hotspots
IS903	9 bp	18 bp	1057 bp	random

Figure 16.2 Transposons have inverted terminal repeats and generate direct repeats of flanking DNA at the target site. In this example, the target is a 5 bp sequence. The ends of the transposon consist of inverted repeats of 9 bp, where the numbers 1 through 9 indicate a sequence of base pairs.

Most IS elements insert at a variety of sites within host DNA. However, some show (varying degrees of) preference for particular hotspots.

The inverted repeats define the ends of a transposon. Recognition of the ends is common to transposition events sponsored by all types of transposon. *Cis*-acting mutations that prevent transposition are located in the ends, which are recognized by a protein(s) responsible for transposition. The protein is called a transposase.

By Book_Crazy [IND]

All the IS elements except IS1 contain a single long coding region, starting just inside the inverted repeat at one end, and terminating just before or within the inverted repeat at the other end. This codes for the transposase. IS1 has a more complex organization, with two separate reading frames; the transposase is produced by making a frameshift during translation to allow both reading frames to be used.

The frequency of transposition varies among different elements. The overall rate of transposition is ~ 10^{-3} 10^{-4} per element per generation. Insertions in individual targets occur at a level comparable with the spontaneous mutation rate, usually ~10^{-5}–10^{-7} per generation. Reversion (by precise excision of the IS element) is usually infrequent, with a range of rates of 10^{-6} to 10^{-10} per generation, ~10^{3} times less frequent than insertion.

16.3 Composite transposons have IS modules

Key Concepts

- Transposons can carry other genes in addition to those coding for transposition.
- Composite transposons have a central region flanked by an IS element at each end.
- Either one or both of the IS elements of a composite transposon may be able to undertake transposition.
- A composite transposon may transpose as a unit, but an active IS element at either end may also transpose independently.

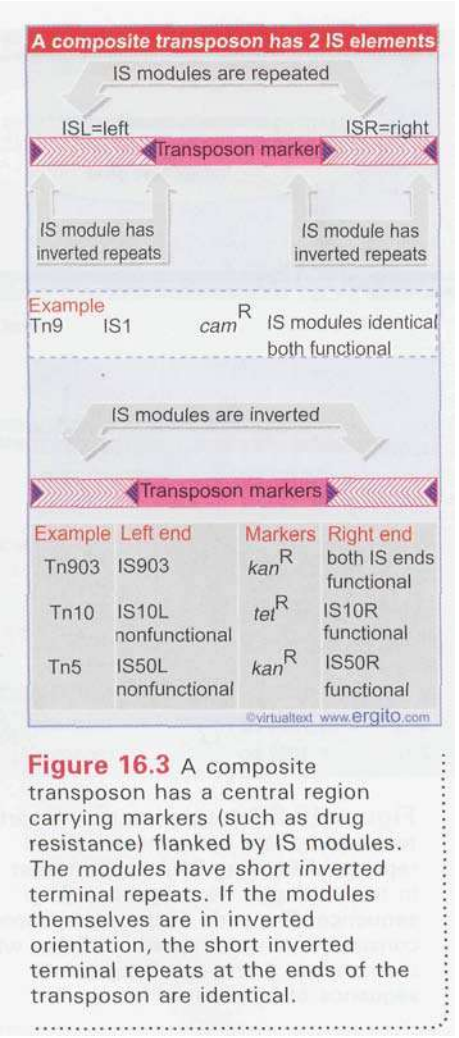
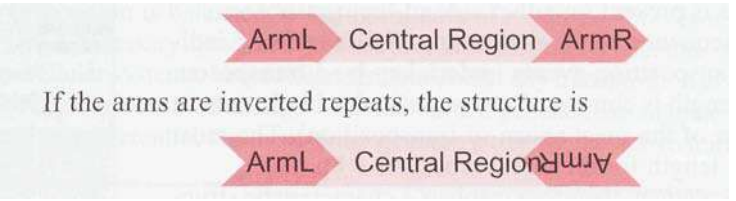


Figure 16.3 A composite transposon has a central region carrying markers (such as drug resistance) flanked by IS modules. The modules have short inverted terminal repeats. If the modules themselves are in inverted orientation, the short inverted terminal repeats at the ends of the transposon are identical.

Some transposons carry drug resistance (or other) markers in addition to the functions concerned with transposition. These transposons are named Tn followed by a number. One class of larger transposons are called composite elements, because a central region carrying the drug marker(s) is flanked on either side by "arms" that consist of IS elements.

The arms may be in either the same or (more commonly) inverted orientation. So a composite transposon with arms that are direct repeats has the structure



The arrows indicate the orientation of the arms, which are identified as L and R according to an (arbitrary) orientation of the genetic map of the transposon from left to right. The structure of a composite transposon is illustrated in more detail in **Figure 16.3,** which also summarizes the properties of some common composite transposons.

Since arms consist of IS modules, and each module has the usual structure ending in inverted repeats, the composite transposon also ends in the same short inverted repeats.

In some cases, the modules of a composite transposon are identical, such as Tn9 (direct repeats of IS1) or Tn903 (inverted repeats of IS903). In other cases, the modules are closely related, but not identical. So we can distinguish the L and R modules in Tn10 or in Tn5.

A functional IS module can transpose either itself or the entire transposon. When the modules of a composite transposon are identi-

cal, presumably either module can sponsor movement of the transposon, as in the case of Tn9 or Tn903. When the modules are different, they may differ in functional ability, so transposition can depend entirely or principally on one of the modules, as in the case of Tn10 or Tn5.

We assume that composite transposons evolved when two originally independent modules associated with the central region. Such a situation could arise when an IS element transposes to a recipient site close to the donor site. The two identical modules may remain identical or diverge. The ability of a single module to transpose the entire composite element explains the lack of selective pressure for both modules to remain active.

What is responsible for transposing a composite transposon instead of just the individual module? This question is especially pressing in cases where both the modules are functional. In the example of Tn9, where the modules are IS 1 elements, presumably each is active in its own right as well as on behalf of the composite transposon. Why is the transposon preserved as a whole, instead of each insertion sequence looking out for itself?

Two IS elements in fact can transpose any sequence residing between them, as well as themselves. **Figure 16.4** shows that if Tn10 resides on a circular replicon, its two modules can be considered to flank either the tet^R gene of the original Tn10 or the sequence in the other part of the circle. So a transposition event can involve either the original Tn10 transposon (marked by the movement of tet^R) or the creation of the new "inside-out" transposon with the alternative central region.

Note that both the original and "inside-out" transposons have inverted modules, but these modules evidently can function in either orientation relative to the central region. The frequency of transposition for composite transposons declines with the distance between the modules. So length dependence is a factor in determining the sizes of the common composite transposons.

A major force supporting the transposition of composite transposons is selection for the marker(s) carried in the central region. An IS 10 module is free to move around on its own, and mobilizes an order of magnitude more frequently than Tn10. But Tn10 is held together by selection for tet^R: so that under selective conditions, the relative frequency of intact TnlO transposition is much increased.

The IS elements code for transposase activities that are responsible both for creating a target site and for recognizing the ends of the transposon. Only the ends are needed for a transposon to serve as a substrate for transposition.

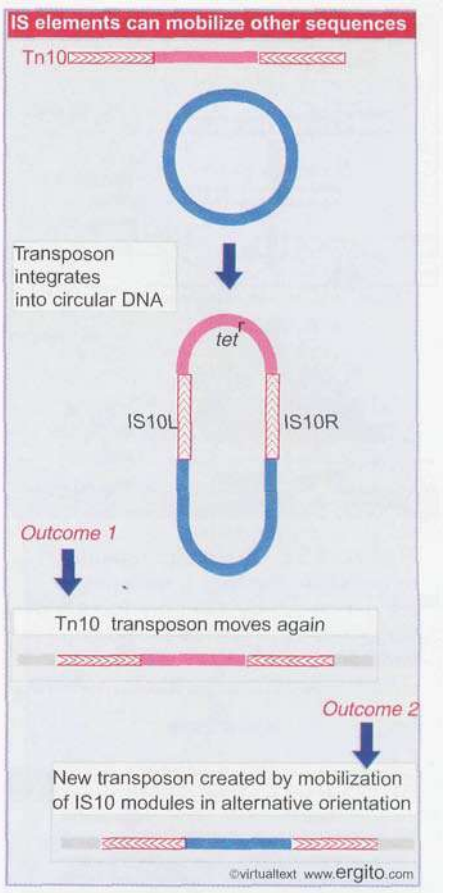


Figure 16.4 Two IS10 modules create a composite transposon that can mobilize any region of DNA that lies between them. When Tn10 is part of a small circular molecule, the IS10 repeats can transpose either side of the circle.

16.4 Transposition occurs by both replicative and nonreplicative mechanisms

Key Concepts

- All transposons use a common mechanism in which staggered nicks are made in target DNA, the transposon is joined to the protruding ends, and the gaps are filled.
- The order of events and exact nature of the connections between transposon and target DNA determine whether transposition is replicative or nonreplicative.

The insertion of a transposon into a new site is illustrated in **Figure 16.5.** It consists of making staggered breaks in the target DNA,

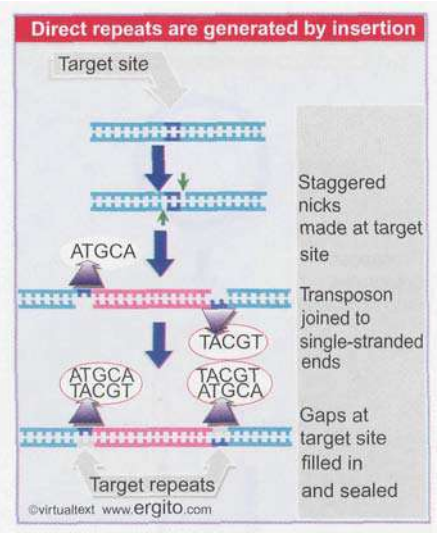


Figure 16.5 The direct repeats of target DNA flanking a transposon are generated by the introduction of staggered cuts whose protruding ends are linked to the transposon.

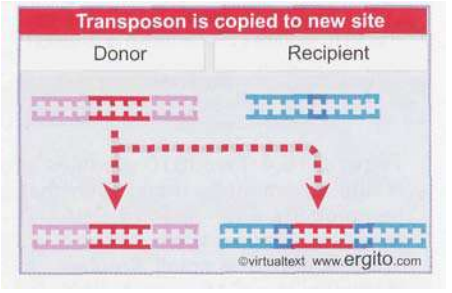


Figure 16.6 Replicative transposition creates a copy of the transposon, which inserts at a recipient site. The donor site remains unchanged, so both donor and recipient have a copy of the transposon.

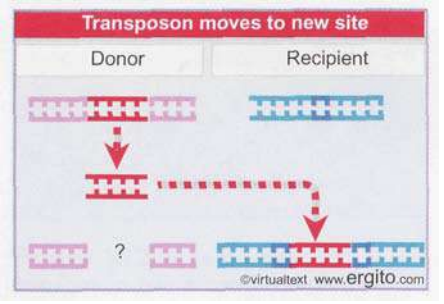


Figure 16.7 Nonreplicative transposition allows a transposon to move as a physical entity from a donor to a recipient site. This leaves a break at the donor site, which is lethal unless it can be repaired.

joining the transposon to the protruding single-stranded ends, and filling in the gaps. The generation and filling of the staggered ends explain the occurrence of the direct repeats of target DNA at the site of insertion. The stagger between the cuts on the two strands determines the length of the direct repeats; so the target repeat characteristic of each transposon reflects the geometry of the enzyme involved in cutting target DNA.

The use of staggered ends is common to all means of transposition, but we can distinguish three different types of mechanism by which a transposon moves:

- In replicative transposition, the element is duplicated during the reaction, so that the transposing entity is a copy of the original element. **Figure 16.6** summarizes the results of such a transposition. The transposon is copied as part of its movement. One copy remains at the original site, while the other inserts at the new site. So transposition is accompanied by an increase in the number of copies of the transposon. Replicative transposition involves two types of enzymatic activity: a transposase that acts on the ends of the original transposon; and a resolvase that acts on the duplicated copies. A group of transposons related to TnA move only by replicative transposition (see *16.7 Replicative transposition proceeds through a cointegrate*).
- In nonreplicative transposition, the transposing element moves as a physical entity directly from one site to another, and is conserved. The insertion sequences and composite transposons Tn10 and Tn5 use the mechanism shown in **Figure 16.7,** which involves the release of the transposon from the flanking donor DNA during transfer. This type of mechanism requires only a transposase. Another mechanism utilizes the connection of donor and target DNA sequences and shares some steps with replicative transposition (see *16.6 Common intermediates for transposition*). Both mechanisms of nonreplicative transposition cause the element to be inserted at the target site and lost from the donor site. What happens to the donor molecule after a nonreplicative transposition? Its survival requires that host repair systems recognize the double-strand break and repair it.
- Conservative transposition describes another sort of nonreplicative event, in which the element is excised from the donor site and inserted into a target site by a series of events in which every nucleotide bond is conserved. **Figure 16.8** summarizes the result of a conservative event. This exactly resembles the mechanism of lambda integration discussed in *15.17 Site-specific recombination involves breakage and reunion,* and the transposases of such elements are related to the *X* integrase family. The elements that use this mechanism are large, and can mediate transfer not only of the element itself but also of donor DNA from one bacterium to another. Although originally classified as transposons, such elements may more properly be regarded as episomes.

Although some transposons use only one type of pathway for transposition, others may be able to use multiple pathways. The elements IS1 and IS903 use both nonreplicative and replicative pathways, and the ability of phage Mu to turn to either type of pathway from a common intermediate has been well characterized (see *16.6 Common intermediates for transposition*).

The same basic types of reaction are involved in all classes of transposition event. The ends of the transposon are disconnected from the donor DNA by cleavage reactions that generate 3′-OH ends. Then the exposed ends are joined to the target DNA by transfer reactions, involving transesterification in which the 3′-OH end directly attacks the target DNA. These reactions take place within a nucleoprotein complex that

contains the necessary enzymes and both ends of the transposon. Transposons differ as to whether the target DNA is recognized before or after the cleavage of the transposon itself.

The choice of target site is in effect made by the transposase. In some cases, the target is chosen virtually at random. In others, there is specificity for a consensus sequence or for some other feature in DNA. The feature can take the form of a structure in DNA, such as bent DNA, or for a protein-DNA complex. In the latter case, the nature of the target complex can cause the transposon to insert at specific promoters (such as Ty1 or Ty3 which select pol III promoters in yeast), inactive regions of the chromosome, or replicating DNA.

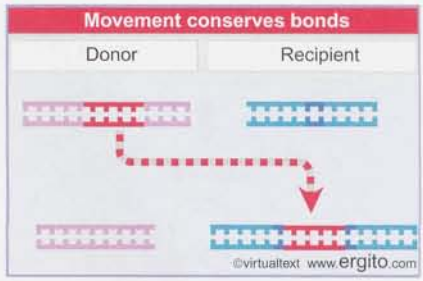


Figure 16.8 Conservative transposition involves direct movement with no loss of nucleotide bonds; compare with lambda integration and excision.

16.5 Transposons cause rearrangement of DNA

Key Concepts

- Homologous recombination between multiple copies of a transposon causes rearrangement of host DNA.
- Homologous recombination between the repeats of a transposon may lead to precise or imprecise excision.

In addition to the "simple" intermolecular transposition that results in insertion at a new site, transposons promote other types of DNA rearrangements. Some of these events are consequences of the relationship between the multiple copies of the transposon. Others represent alternative outcomes of the transposition mechanism, and they leave clues about the nature of the underlying events.

Rearrangements of host DNA may result when a transposon inserts a copy at a second site near its original location. Host systems may undertake reciprocal recombination between the two copies of the transposon; the consequences are determined by whether the repeats are the same or in inverted orientation.

Figure 16.9 illustrates the general rule that recombination between any pair of direct repeats will delete the material between them. The intervening region is excised as a circle of DNA (which is lost from the cell); the chromosome retains a single copy of the direct repeat. A recombination between the directly repeated IS1 modules of the composite transposon Tn9 would replace the transposon with a single IS1 module.

Deletion of sequences adjacent to a transposon could therefore result from a two-stage process; transposition generates a direct repeat of a transposon, and recombination occurs between the repeats. However, the majority of deletions that arise in the vicinity of transposons probably result from a variation in the pathway followed in the transposition event itself.

Figure 16.10 depicts the consequences of a reciprocal recombination between a pair of inverted repeats. The region between the repeats becomes inverted; the repeats themselves remain available to sponsor further inversions. A composite transposon whose modules are inverted is a stable component of the genome, although the direction of the central region with regard to the modules could be inverted by recombination.

Excision is not supported by transposons themselves, but may occur when bacterial enzymes recognize homologous regions in the transposons. This is important because the loss of a transposon may restore function at the site of insertion. Precise excision requires removal of the transposon plus one copy of the duplicated sequence. This is rare; it

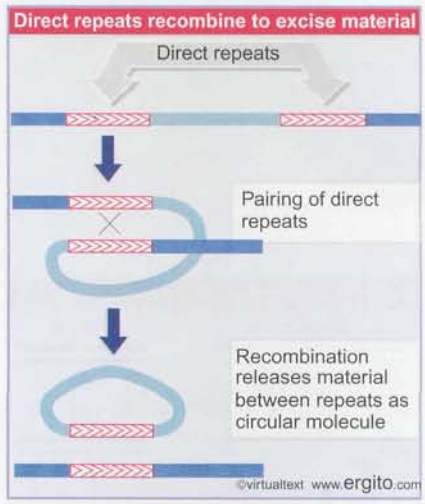


Figure 16.9 Reciprocal recombination between direct repeats excises the material between them; each product of recombination has one copy of the direct repeat.

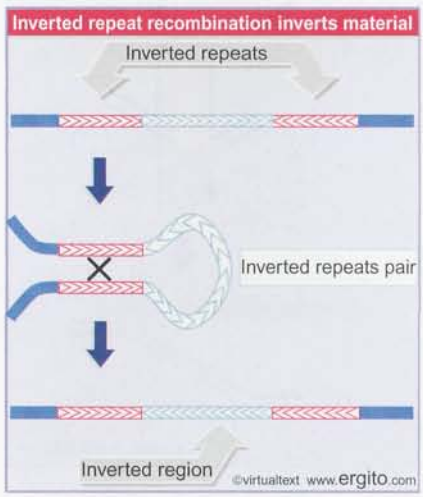


Figure 16.10 Reciprocal recombination between inverted repeats inverts the region between them.

occurs at a frequency of ~10^{-6} for Tn5 and ~10^{-9} for Tn10. It probably involves a recombination between the 9 bp duplicated target sites.

Imprecise excision leaves a remnant of the transposon. Although the remnant may be sufficient to prevent reactivation of the target gene, it may be insufficient to cause polar effects in adjacent genes, so that a change of phenotype occurs. Imprecise excision occurs at a frequency of ~10^{-7} for TnlO. It involves recombination between sequences of 24 bp in the IS 10 modules; these sequences are inverted repeats, but since the IS 10 modules themselves are inverted, they form direct repeats in TnlO.

The greater frequency of imprecise excision compared with precise excision probably reflects the increase in the length of the direct repeats (24 bp as opposed to 9 bp). Neither type of excision relies on transposon-coded functions, but the mechanism is not known. Excision is RecA-independent and could occur by some cellular mechanism that generates spontaneous deletions between closely spaced repeated sequences.

16.6 Common intermediates for transposition

Key Concepts

- Transposition starts by forming a strand transfer complex in which the transposon is connected to the target site through one strand at each end.
- The Mu transposase forms the complex by synapsing the ends of Mu DNA, followed by nicking and transfer reaction.
- Replicative transposition follows if the complex is replicated and nonreplicative transposition follows if it is repaired.

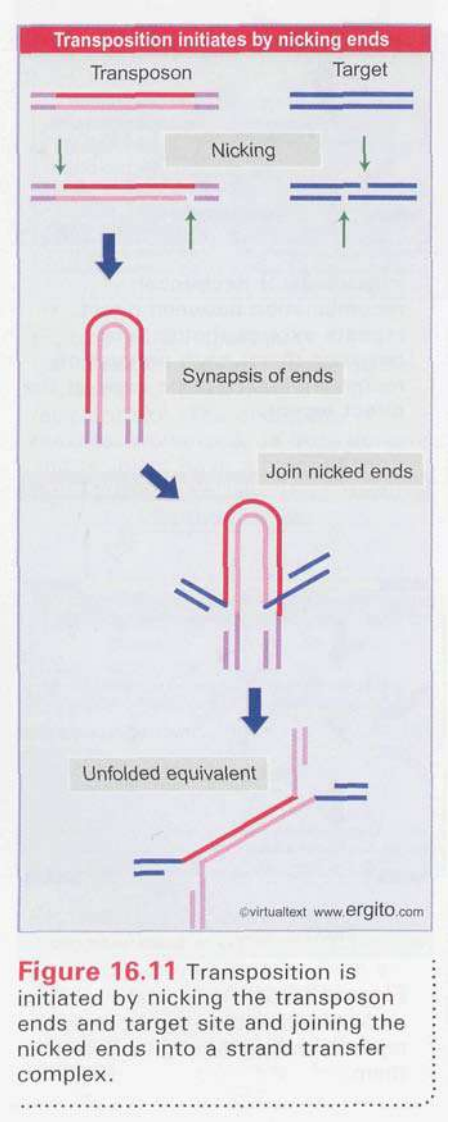


Figure 16.11 Transposition is initiated by nicking the transposon ends and target site and joining the nicked ends into a strand transfer complex.

Many mobile DNA elements transpose from one chromosomal location to another by a fundamentally similar mechanism. They include IS elements, prokaryotic and eukaryotic transposons, and bacteriophage Mu. Insertion of the DNA copy of retroviral RNA uses a similar mechanism (see *17.2 The retrovirus life cycle involves transposition-like events)*. The first stages of immunoglobulin recombination also are similar (see *26.9 The RAG proteins catalyze breakage and reunion)*.

Transposition starts with a common mechanism for joining the transposon to its target. Figure 16.11 shows that the transposon is nicked at both ends, and the target site is nicked on both strands. The nicked ends are joined crosswise to generate a covalent connection between the transposon and the target. The two ends of the transposon are brought together in this process; for simplicity in following the cleavages, the synapsis stage is shown after cleavage, but actually occurs previously.

Much of this pathway was first revealed with phage Mu, which uses the process of transposition in two ways. Upon infecting a host cell, Mu integrates into the genome by nonreplicative transposition; during the ensuing lytic cycle, the number of copies is amplified by replicative transposition. Both types of transposition involve the same type of reaction between the transposon and its target, but the subsequent reactions are different.

The initial manipulations of the phage DNA are performed by the MuA transposase. Three MuA-binding sites with a 22 bp consensus are located at each end of Mu DNA. L1, L2, and L3 are at the left end; R1, R2, and R3 are at the right end. A monomer of MuA can bind to each site. MuA also binds to an internal site in the phage genome. Binding of MuA at both the left and right ends and the internal site forms a complex. The

role of the internal site is not clear; it appears to be necessary for formation of the complex, but not for strand cleavage and subsequent steps.

Joining the Mu transposon DNA to a target site passes through the three stages illustrated in **Figure 16.12.** This involves only the two sites closest to each end of the transposon. MuA subunits bound to these sites form a tetramer. This achieves synapsis of the two ends of the transposon. The tetramer now functions in a way that ensures a coordinated reaction on both ends of Mu DNA. MuA has two sites for manipulating DNA, and their mode of action compels subunits of the transposase to act in *trans.* The consensus-binding site binds to the 22 bp sequences that constitute the L1, L2, R1, and R2 sites. The active site cleaves the Mu DNA strands at positions adjacent to the MuA-binding sites L1 and R1. But the active site cannot cleave the DNA sequence that is adjacent to the consensus sequence in the consensus-binding site. However, it can cleave the appropriate sequence on a different stretch of DNA.

The ends of the transposon are thus cleaved by MuA subunits acting in *trans.* The *trans* mode of action means that the monomers actually bound to L1 and Rl do not cleave the adjacent sites. One of the monomers bound to the left end nicks the site at the right end, and vice versa. (We do not know which monomer is active at this stage of the reaction.) The strand transfer reaction also occurs in *trans;* the monomer at L1 transfers the strand at R1, and vice versa. It could be the case that different monomers catalyze the cleavage and strand transfer reactions for a given end.

A second protein, MuB, assists the reaction. It has an influence on the choice of target sites. Mu has a preference for transposing to a target site > **10-15** kb away from the original insertion. This is called "target immunity." It is demonstrated in an *in vitro* reaction containing donor (Mu-containing) and target (Mu-deficient) plasmids, MuA and MuB proteins, *E. coli* HU protein, and Mg^{2+} and ATP. The presence of MuB and ATP restricts transposition exclusively to the target plasmid. The reason is that when MuB binds to the MuA-Mu DNA complex, MuA causes MuB to hydrolyze ATP, after which MuB is released. However, MuB binds (nonspecifically) to the target DNA, where it stimulates the recombination activity of MuA when a transposition complex forms. In effect, the prior presence of MuA "clears" MuB from the donor, thus giving a preference for transposition to the target.

The product of these reactions is a strand transfer complex in which the transposon is connected to the target site through one strand at each end. The next step of the reaction differs and determines the type of transposition. We see in the next two sections how the common structure can be a substrate for replication (leading to replicative transposition) or used directly for breakage and reunion (leading to nonreplicative transposition).

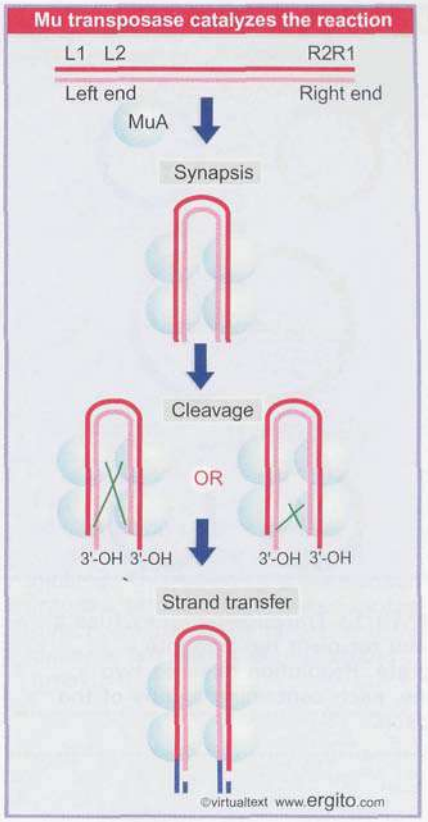


Figure 16.12 Mu transposition passes through three stable stages. MuA transposase forms a tetramer that synapses the ends of phage Mu. Transposase subunits act in *trans* to nick each end of the DNA; then a second *trans* action joins the nicked ends to the target DNA.

16.7 Replicative transposition proceeds through a cointegrate

Key Concepts

- Replication of a strand transfer complex generates a cointegrate, which is a fusion of the donor and target replicons.
- The cointegrate has two copies of the transposon, which lie between the original replicons.
- Recombination between the transposon copies regenerates the original replicons, but the recipient has gained a copy of the transposon.
- The recombination reaction is catalyzed by a resolvase coded by the transposon.

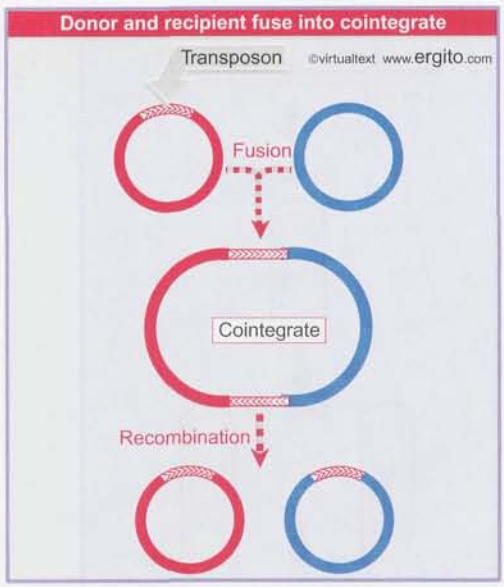


Figure 16.13 Transposition may fuse a donor and recipient replicon into a cointegrate. Resolution releases two replicons, each containing a copy of the transposon.

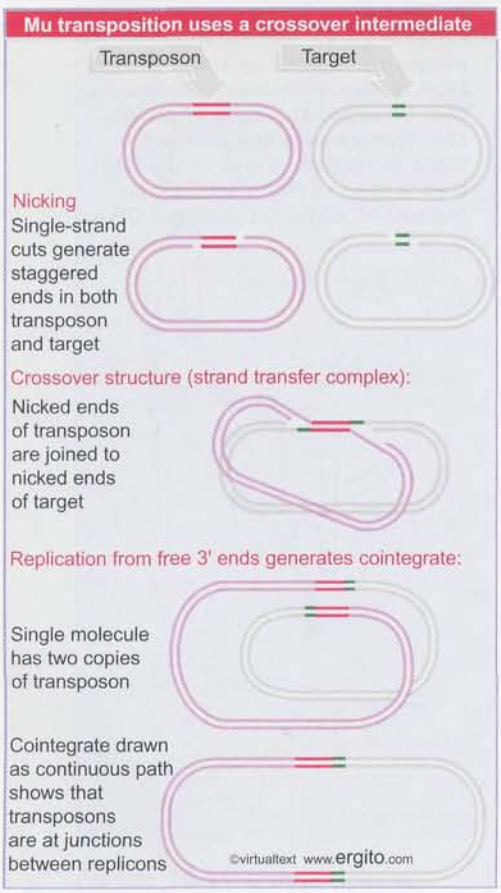


Figure 16.14 Mu transposition generates a crossover structure, which is converted by replication into a cointegrate.

The basic structures involved in replicative transposition are illustrated in **Figure 16.13:**

- The 3′ ends of the strand transfer complex are used as primers for replication. This generates a structure called a cointegrate, which represents a fusion of the two original molecules. The cointegrate has two copies of the transposon, one at each junction between the original replicons, oriented as direct repeats. The crossover is formed by the transposase, as described in the previous section. Its conversion into the cointegrate requires host replication functions.
- A homologous recombination between the two copies of the transposon releases two individual replicons, each of which has a copy of the transposon. One of the replicons is the original donor replicon. The other is a target replicon that has gained a transposon flanked by short direct repeats of the host target sequence. The recombination reaction is called resolution; the enzyme activity responsible is called the resolvase.

The reactions involved in generating a cointegrate have been defined in detail for phage Mu, and are illustrated in **Figure 16.14.** The process starts with the formation of the strand transfer complex (sometimes also called a crossover complex). The donor and target strands are ligated so that each end of the transposon sequence is joined to one of the protruding single strands generated at the target site. The strand transfer complex generates a crossover-shaped structure held together at the duplex transposon. The fate of the crossover structure determines the mode of transposition.

The principle of replicative transposition is that replication through the transposon duplicates it, creating copies at both the target and donor sites. The product is a cointegrate.

The crossover structure contains a single-stranded region at each of the staggered ends. These regions are pseudoreplication forks that provide a template for DNA synthesis. (Use of the ends as primers for replication implies that the strand breakage must occur with a polarity that generates a 3'-OH terminus at this point.)

If replication continues from both the pseudoreplication forks, it will proceed through the transposon, separating its strands, and terminating at its ends. Replication is probably accomplished by host-coded functions. At this juncture, the structure has become a cointegrate, possessing direct repeats of the transposon at the junctions between the replicons (as can be seen by tracing the path around the cointegrate).

16.8 Nonreplicative transposition proceeds by breakage and reunion

Key Concepts

* Nonreplicative transposition results if a crossover structure is nicked on the unbroken pair of donor strands, and the target strands on either side of the transposon are ligated.
- Two pathways for nonreplicative transposition differ according to whether the first pair of transposon strands are joined to the target before the second pair are cut (Tn5), or whether all four strands are cut before joining to the target (Tn10).

The crossover structure can also be used in nonreplicative transposition. The principle of nonreplicative transposition by this mechanism is that a breakage and reunion reaction allows the target to be

reconstructed with the insertion of the transposon; the donor remains broken. No cointegrate is formed.

Figure 16.15 shows the cleavage events that generate nonreplicative transposition of phage Mu. Once the unbroken donor strands have been nicked, the target strands on either side of the transposon can be ligated. The single-stranded regions generated by the staggered cuts must be filled in by repair synthesis. The product of this reaction is a target replicon in which the transposon has been inserted between repeats of the sequence created by the original single-strand nicks. The donor replicon has a double-strand break across the site where the transposon was originally located.

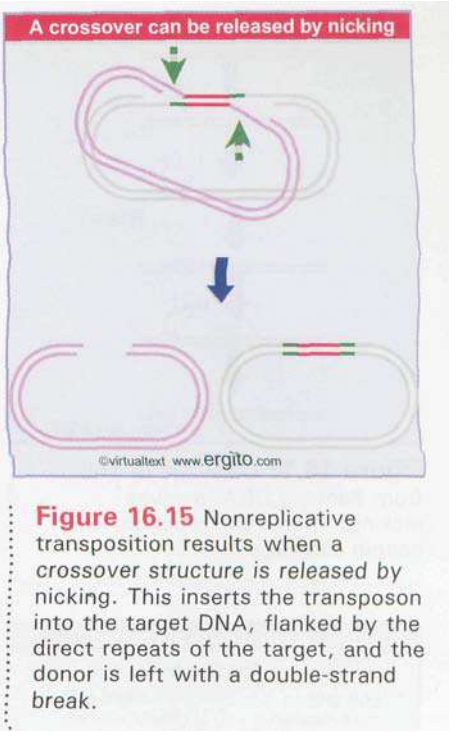


Figure 16.15 Nonreplicative transposition results when a crossover structure is released by nicking. This inserts the transposon into the target DNA, flanked by the direct repeats of the target, and the donor is left with a double-strand break.

Nonreplicative transposition can also occur by an alternative pathway in which nicks are made in target DNA, but a double-strand break is made on either side of the transposon, releasing it entirely from flanking donor sequences (as envisaged in Figure 16.7). This "cut and paste" pathway is used by Tn10, as illustrated in **Figure 16.16**.

A neat experiment to prove that TnlO transposes nonreplicatively made use of an artificially constructed heteroduplex of Tn10 that contained single base mismatches. If transposition involves replication, the transposon at the new site will contain information from only one of the parent Tn10 strands. But if transposition takes place by physical movement of the existing transposon, the mismatches will be conserved at the new site, which proved to be the case.

The basic difference in Figure 16.16 from the model of Figure 16.15 is that both strands of Tn10 are cleaved before any connection is made to the target site. The first step in the reaction is recognition of the transposon ends by the transposase, forming a proteinaceous structure within which the reaction occurs. At each end of the transposon, the strands are cleaved in a specific order—first the transferred strand (the one to be connected to the target site) is cleaved, then the other strand (this is the same order as in the Mu transposition of Figure 16.14 and Figure 16.15).

Tn5 also transposes by nonreplicative transposition, and **Figure 16.17** shows the interesting cleavage reaction that separates the transposon from the flanking sequences. First one DNA strand is nicked. The 3′-OH end that is released then attacks the other strand of DNA. This releases the flanking sequence and joins the two strands of the transposon in a hairpin. Then an activated water molecule attacks the hairpin to generate free ends for each strand of the transposon.

Then the cleaved donor DNA is released, and the transposon is joined to the nicked ends at the target site. The transposon and the target site remain constrained in the proteinaceous structure created by the transposase (and other proteins). The double-strand cleavage at each end of the transposon precludes any replicative-type transposition and forces the reaction to proceed by nonreplicative transposition, thus giving the same outcome as in Figure 16.14, but with the individual cleavage and joining steps occurring in a different order.

The Tn5 and Tn10 transposases both function as dimers. Each subunit in the dimer has an active site that successively catalyzes the double-strand breakage of the two strands at one end of the transposon and then catalyzes staggered cleavage of the target site. **Figure 16.18** illustrates the structure of the Tn5 transposase bound to the cleaved transposon. Each end of the transposon is located in the active site of one subunit. One end of the subunit also contacts the other end of the transposon. This controls the geometry of the transposition reaction. Each of the active sites will cleave one strand of the target DNA. It is the geometry of the complex that determines the distance between these sites on the two target strands *(9* base pairs in *the* case of Tn5).

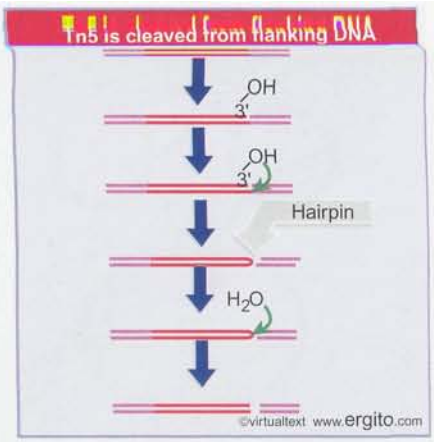


Figure 16.17 Cleavage of Tn5 from flanking DNA involves nicking, interstrand reaction, and hairpin cleavage.

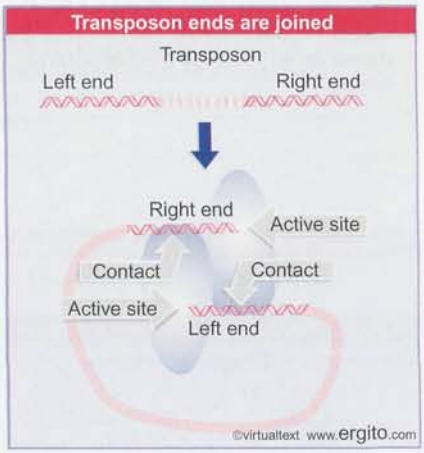


Figure 16.18 Each subunit of the Tn5 transposase has one end of the transposon located in its active site and also makes contact at a different site with the other end of the transposon.

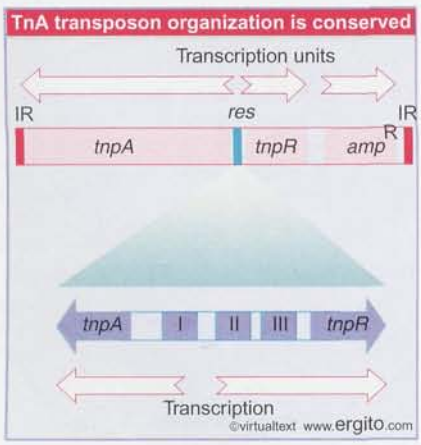


Figure 16.19 Transposons of the TnA family have inverted terminal repeats, an internal *res* site, and three known genes.

16.9 TnA transposition requires transposase and resolvase

Key Concepts

- Replicative transposition of TnA requires a transposase to form the cointegrate structure and a resolvase to release the two replicons.
- The action of the resolvase resembles lambda Int protein and belongs to the general family of topoisomerase-like site-specific recombination reactions, which pass through an intermediate in which the protein is covalently bound to the DNA.

Replicative transposition is the only mode of mobility of the TnA family, which consists of large (~5 kb) transposons. They are not composites relying on IS-type transposition modules, but comprise independent units carrying genes for transposition as well as for features such as drug resistance. The TnA family includes several related transposons, of which Tn3 and TnlOOO (formerly called 78) are the best characterized. They have the usual terminal feature of closely related inverted repeats, generally ~38 bp in length. *Cis*-acting deletions in either repeat prevent transposition of an element. A 5 bp direct repeat is generated at the target site. They carry resistance markers such as amp^r.

The two stages of TnA-mediated transposition are accomplished by the transposase and the resolvase, whose genes, *tnpA* and *tnpR*, are identified by recessive mutations. The transposition stage involves the ends of the element, as it does in IS-type elements. Resolution requires a specific internal site. This feature is unique to the TnA family.

Mutants in *tnpA* cannot transpose. The gene product is a transposase that binds to a sequence of ~25 bp located within the 38 bp of the inverted terminal repeat. A binding site for the *E. coli* protein IHF exists adjacent to the transposase binding site; and transposase and IHF bind cooperatively. The transposase recognizes the ends of the element and also makes the staggered 5 bp breaks in target DNA where the transposon is to be inserted. IHF is a DNA-binding protein that is often involved in assembling large structures in *E. coli;* its role in the transposition reaction may not be essential.

The *tnpR* gene product has dual functions. It acts as a repressor of gene expression and it provides the resolvase function.

Mutations in *tnpR* increase the transposition frequency. The reason is that TnpR represses the transcription of both *tnpA* and its own gene. So inactivation of TnpR protein allows increased synthesis of TnpA, which results in an increased frequency of transposition. This implies that the amount of the TnpA transposase must be a limiting factor in transposition.

The *tnpA* and *tnpR* genes are expressed divergently from an A·T-rich intercistronic control region, indicated in the map of Tn3 given in **Figure 16.19.** Both effects of TnpR are mediated by its binding in this region.

In its capacity as the resolvase, TnpR is involved in recombination between the direct repeats of Tn3 in a cointegrate structure. A cointegrate can in principle be resolved by a homologous recombination between any corresponding pair of points in the two copies of the transposon. But the Tn3 resolution reaction occurs only at a specific site.

The site of resolution is called *res.* It is identified by *cis*-acting deletions that block completion of transposition, causing the accumulation

of cointegrates. In the absence of *res*, the resolution reaction can be substituted by RecA-mediated general recombination, but this is much less efficient.

The sites bound by the TnpR resolvase are summarized in the lower part of Figure 16.19. Binding occurs independently at each of three sites, each 30-40 bp long. The three binding sites share a sequence homology that defines a consensus sequence with dyad symmetry.

Site I includes the region genetically defined as the *res* site; in its absence, the resolution reaction does not proceed at all. However, resolution also involves binding at sites II and III, since the reaction proceeds only poorly if either of these sites is deleted. Site I overlaps with the startpoint for *tnpA* transcription. Site II overlaps with the startpoint for *tnpR* transcription; an operator mutation maps just at the left end of the site.

Do the sites interact? One possibility is that binding at all three sites is required to hold the DNA in an appropriate topology. Binding at a single set of sites may repress *tnpA* and *tnpR* transcription without introducing any change in the DNA.

An *in vitro* resolution assay uses a cointegrate-like DNA molecule as substrate. The substrate must be supercoiled; its resolution produces two catenated circles, each containing one *res* site. The reaction requires large amounts of the TnpR resolvase; no host factors are needed. Resolution occurs in a large nucleoprotein structure. Resolvase binds to each *res* site, and then the bound sites are brought together to form a structure ~10 nm in diameter. Changes in supercoiling occur during the reaction, and DNA is bent at the *res* sites by the binding of transposase.

Resolution occurs by breaking and rejoining bonds without input of energy. The products identify an intermediate stage in cointegrate resolution; they consist of resolvase covalently attached to both 5' ends of double-stranded cuts made at the *res* site. The cleavage occurs symmetrically at a short palindromic region to generate two base extensions. Expanding the view of the crossover region located in site I, we can describe the cutting reaction as:

The reaction resembles the action of lambda Int at the *att* sites (see *15.17 Site-specific recombination involves breakage and reunion).* Indeed, 15 of the 20 bp of the *res* site are identical to the bases at corresponding positions in *att.* This suggests that the site-specific recombination of lambda and resolution of TnA have evolved from a common type of recombination reaction; and indeed, we see in *26.9 The RAG proteins catalyze breakage and reunion* that recombination involving immunoglobulin genes has the same basis. The common feature in all these reactions is the transfer of the broken end to the catalytic protein as an intermediate stage before it is rejoined to another broken end (see *15.18 Site-specific recombination resembles topoisomerase activity).*

The reactions themselves are analogous in terms of manipulation of DNA, although resolution occurs only between intramolecular sites, whereas the recombination between *att* sites is intermolecular and directional (as seen by the differences in *attB* and *attP* sites). However,

the mechanism of protein action is different in each case. Resolvase functions in a manner in which four subunits bind to the recombining *res* sites. Each subunit makes a single-strand cleavage. Then a reorganization of the subunits relative to one another physically moves the DNA strands, placing them in a recombined conformation. This allows the nicks to be sealed, along with the release of resolvase.

16.10 Transposition of Tn10 has multiple controls

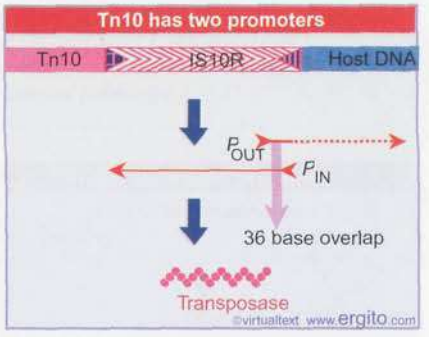


Figure 16.20 Two promoters in opposite orientation lie near the outside boundary of IS10R. The strong promoter P_{OUT} sponsors transcription toward the flanking host DNA. The weaker promoter P_{IN} causes transcription of an RNA that extends the length of IS10R and is translated into the transposase.

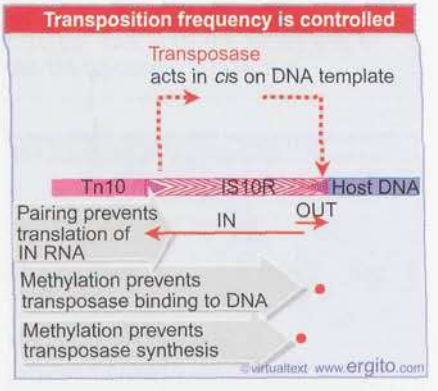


Figure 16.21 Several mechanisms restrain the frequency of Tn10 transposition, by affecting either the synthesis or function of transposase protein. Transposition of an individual transposon is restricted by methylation to occur only after replication. In multicopy situations, *cis*-preference restricts the choice of target, and OUT/IN RNA pairing inhibits synthesis of transposase.

Key Concepts

- Multicopy inhibition reduces the rate of transposition of any one copy of a transposon when other copies of the same transposon are introduced into the genome.
- Multiple mechanisms affect the rate of transposition.

Control of the frequency of transposition is important for the cell. A transposon must be able to maintain a certain minimum frequency of movement in order to survive; but too great a frequency could be damaging to the host cell. Every transposon appears to have mechanisms that control its frequency of transposition. A variety of mechanisms have been characterized for Tn10.

Tn10 is a composite transposon in which the element IS10R provides the active module. The organization of IS 1 OR is summarized in **Figure 16.20.** Two promoters are found close to the outside boundary. The promoter P_{IN} is responsible for transcription of IS 1 OR. The promoter P_{OUT} causes transcription to proceed toward the adjacent flanking DNA. Transcription usually terminates within the transposon, but occasionally continues into the host DNA; sometimes this readthrough transcription is responsible for activating adjacent bacterial genes.

The phenomenon of "multicopy inhibition" reveals that expression of the IS 10R transposase gene is regulated. Transposition of a TnlO element on the bacterial chromosome is reduced when additional copies of 1S10R are introduced via a multicopy plasmid. The inhibition requires the POUT promoter, and is exercised at the level of translation. The basis for the effect lies with the overlap in the 5′ terminal regions of the transcripts from P_{IN} and P_{OUT}. OUT RNA is a transcript of 69 bases. It is present at > 100x the level of IN RNA for two reasons: P_{OUT} is a much stronger promoter than P_{IN}; and OUT RNA is more stable than IN RNA.

OUT RNA functions as an antisense RNA (see *11.19 Small RNA molecules can regulate translation).* The level of OUT RNA has no effect in a single-copy situation, but has a significant effect when >5 copies are present. There are usually ~5 copies of OUT RNA per copy of IS 10 (which corresponds to ~150 copies of OUT RNA in a typical multicopy situation). OUT RNA base pairs with IN RNA; and the excess of OUT RNA ensures that IN RNA is bound rapidly, before a ribosome can attach. So the paired IN RNA cannot be translated.

The quantity of transposase protein is often a critical feature. TnlO, whose transposase is synthesized at the low level of 0.15 molecules per cell per generation, displays several interesting mechanisms. **Figure 16.21** summarizes the various effects that influence transposition frequency.

A continuous reading frame on one strand of IS10R codes for the transposase. The level of the transposase limits the rate of transposition.

By Book_Crazy [IND]

Mutants in this gene can be complemented in *trans* by another, wild-type IS 10 element, but only with some difficulty. This reflects a strong preference of the transposase for *cis*-action; the enzyme functions efficiently only with the DNA template from which it was transcribed and translated. *Cis*-preference is a common feature of transposases coded by IS elements. (Other proteins that display *cis*-preference include the A protein involved in ϕX174 replication; see *13.11 Rolling circles are used to replicate phage genomes.)*

Does *cis*-preference reflect an ability of the transposase to recognize more efficiently those DNA target sequences that lie nearer to the site where the enzyme is synthesized? One possible explanation is that the transposase binds to DNA so tightly after (or even during) protein synthesis that it has a very low probability of diffusing elsewhere. Another possibility is that the enzyme may be unstable when it is not bound to DNA, so that protein molecules failing to bind quickly (and therefore nearby) never have a chance to become active.

Together the results of *cis*-preference and multicopy inhibition ensure that an increase in the number of copies of Tn10 in a bacterial genome does not cause an increased frequency of transposition that could damage the genome.

The effects of methylation provide the most important system of regulation for an individual element. They reduce the frequency of transposition and (more importantly) couple transposition to passage of the replication fork. The ability of IS 10 to transpose is related to the replication cycle by the transposon's response to the state of methylation at two sites. One site is within the inverted repeat at the end of IS10R, where the transposase binds. The other site is in the promoter P_{IN}, from which the transposase gene is transcribed.

Both of these sites are methylated by the *dam* system described in *14.18 Does methylation at the origin regulate initiation?* The Dam methylase modifies the adenine in the sequence GATC on a newly synthesized strand generated by replication. The frequency of Tn10 transposition is increased 1000-fold in *dam*⁻ strains in which the two target sites lack methyl groups.

Passage of a replication fork over these sites generates hemimethylated sequences; this activates the transposon by a combination of transcribing the transposase gene more frequently from P_{IN} and enhancing binding of transposase to the end of IS10R. In a wild-type bacterium, the sites remain hemimethylated for a short period after replication.

Why should it be desirable for transposition to occur soon after replication? The nonreplicative mechanism of Tn10 transposition places the donor DNA at risk of being destroyed (see Figure 16.7). The cell's chances of survival may be increased if replication has just occurred to generate a second copy of the donor sequence. The mechanism is effective because only 1 of the 2 newly replicated copies gives rise to a transposition event (determined by which strand of the transposon is unmethylated at the *dam* sites).

Since a transposon selects its target site at random, there is a reasonable probability that it may land in an active operon. Will transcription from the outside continue through the transposon and thus activate the transposase, whose overproduction may in turn lead to high (perhaps lethal) levels of transposition? TnlO protects itself against such events by two mechanisms. Transcription across the IS10R terminus decreases its activity, presumably by inhibiting its ability to bind transposase. And the mRNA that extends from upstream of the promoter is poorly translated, because it has a secondary structure in which the initiation codon is inaccessible.

By Book_Crazy [IND]

16.11 Controlling elements in maize cause breakage and rearrangements

Key Concepts

- Transposition in maize was discovered because of the effects of the chromosome breaks generated by transposition of "controlling elements".
- The break generates one chromosome that has a centromere and a broken end and one acentric fragment.
- The acentric fragment is lost during mitosis, and this can be detected by the disappearance of dominant alleles in a heterozygote.
- Fusion between the broken ends of the chromosome generates dicentric chromosomes, which undergo further cycles of breakage and fusion.
- The fusion-breakage-bridge cycle is responsible for the occurrence of somatic variegation.

One of the most visible consequences of the existence and mobility of transposons occurs during plant development, when somatic variation occurs. This is due to changes in the location or behavior of controlling elements.

Two features of maize have helped to follow transposition events. Controlling elements often insert near genes that have visible but nonlethal effects on the phenotype. And because maize displays clonal development, the occurrence and timing of a transposition event can be visualized as depicted diagrammatically in **Figure 16.22.**

The nature of the event does not matter: it may be a point mutation, insertion, excision, or chromosome break. What is important is that it occurs in a heterozygote to alter the expression of one allele. Then the descendants of a cell that has suffered the event display a new phenotype, while the descendants of cells not affected by the event continue to display the original phenotype.

Mitotic descendants of a given cell remain in the same location and give rise to a sector of tissue. A change in phenotype during somatic development is called variegation; it is revealed by a sector of the new phenotype residing within the tissue of the original phenotype. The size of the sector depends on the number of divisions in the lineage giving rise to it; so the size of the area of the new phenotype is determined by the timing of the change in genotype. The earlier its occurrence in the cell lineage, the greater the number of descendants and thus the size of patch in the mature tissue. This is seen most vividly in the variation in kernel color, when patches of one color appear within another color.

Insertion of a controlling element may affect the activity of adjacent genes. Deletions, duplications, inversions, and translocations all occur at the sites where controlling elements are present. Chromosome breakage is a common consequence of the presence of some elements. A unique feature of the maize system is that the activities of the controlling elements are regulated during development. The elements transpose and promote genetic rearrangements at characteristic times and frequencies during plant development.

The characteristic behavior of controlling elements in maize is typified by the Ds element, which was originally identified by its ability to provide a site for chromosome breakage. The consequences are illustrated in **Figure 16.23**. Consider a heterozygote in which Ds lies on one homologue between the centromere and a series of dominant markers. The other homologue lacks Ds and has recessive markers (C, *bz*, *wx*)

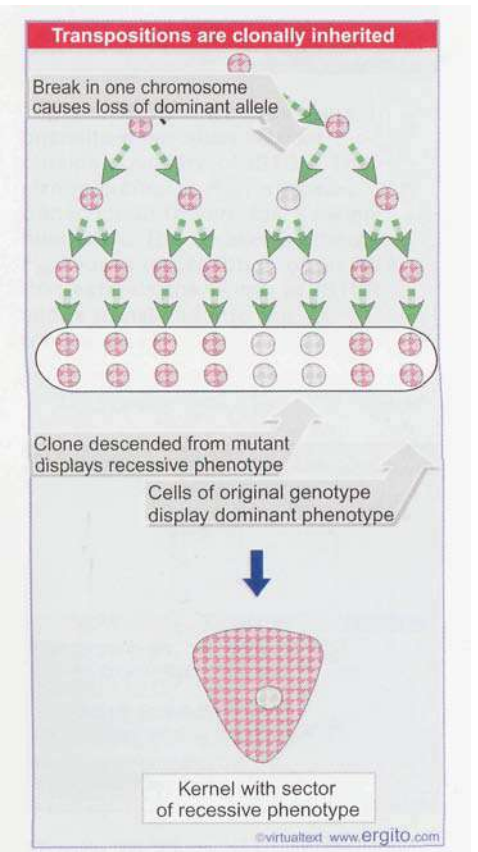


Figure 16.22 Clonal analysis identifies a group of cells descended from a single ancestor in which a transposition- mediated event altered the phenotype. Timing of the event during development is indicated by the number of cells; tissue specificity of the event may be indicated by the location of the cells.

Breakage at Ds generates an acentric fragment carrying the dominant markers. Because of its lack of a centromere, this fragment is lost at mitosis. So the descendant cells have only the recessive markers carried by the intact chromosome. This gives the type of situation whose results are depicted in Figure 16.22.

Figure 16.24 shows that breakage at Ds leads to the formation of two unusual chromosomes. These are generated by joining the broken ends of the products of replication. One is a U-shaped acentric fragment consisting of the joined sister chromatids for the region distal to Ds (on the left as drawn in the figure). The other is a U-shaped dicentric chromosome comprising the sister chromatids proximal to Ds (on its right in the figure). The latter structure leads to the classic breakage-fusion-bridge cycle illustrated in the figure.

Follow the fate of the dicentric chromosome when it attempts to segregate on the mitotic spindle. Each of its two centromeres pulls toward an opposite pole. The tension breaks the chromosome at a random site between the centromeres. In the example of the figure, breakage occurs between loci A and B, with the result that one daughter chromosome has a duplication of A, while the other has a deletion. If A is a dominant marker, the cells with the duplication will retain A phenotype, but cells with the deletion will display a recessive loss of function phenotype.

The breakage-fusion-bridge cycle continues through further cell generations, allowing genetic changes to continue in the descendants. For example, consider the deletion chromosome that has lost A. In the next cycle, a break occurs between B and C, so that the descendants are divided into those with a duplication of B and those with a deletion. Successive losses of dominant markers are revealed by subsectors within sectors.

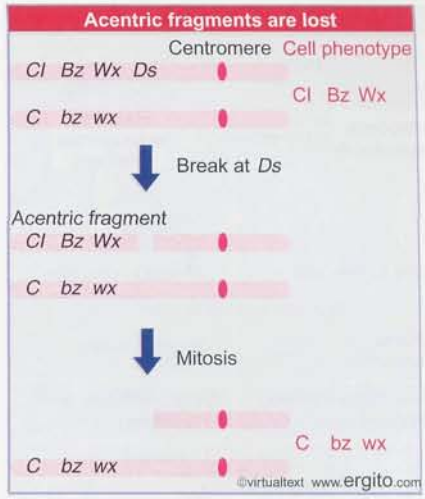


Figure 16.23 A break at a controlling element causes loss of an acentric fragment; if the fragment carries the dominant markers of a heterozygote, its loss changes the phenotype.

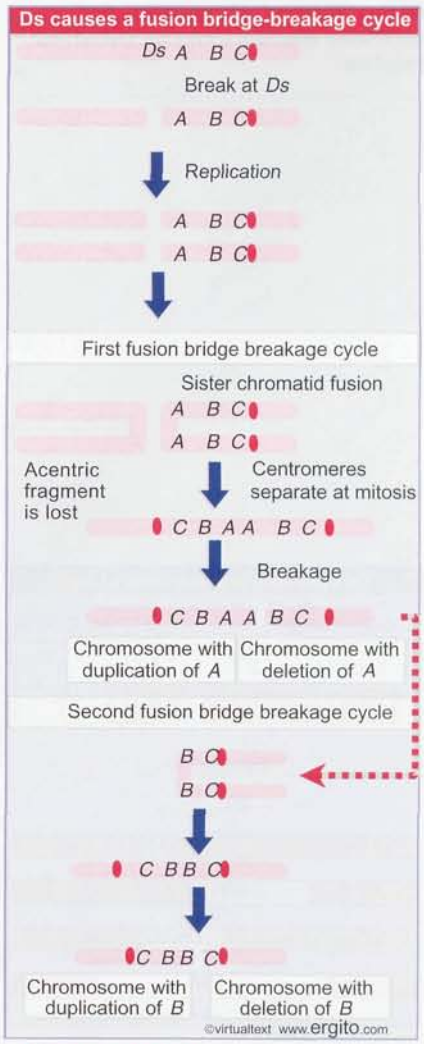


Figure 16.24 *Ds* provides a site to initiate the chromatid fusion-bridge-breakage cycle.

16.12 Controlling elements form families of transposons

Key Concepts

- Each family of transposons in maize has both autonomous and nonautonomous controlling elements.
- Autonomous controlling elements code for proteins that enable them to transpose.
- Nonautonomous controlling elements have mutations that eliminate their capacity to catalyze transposition, but they can transpose when an autonomous element provides the necessary proteins.
- Autonomous controlling elements have changes of phase, when their properties alter as a result of changes in the state of methylation.

The maize genome contains several families of controlling elements. The numbers, types, and locations of the elements are characteristic for each individual maize strain. They may occupy a significant part of the genome. The members of each family are divided into two classes:

- Autonomous controlling elements have the ability to excise and transpose. Because of the continuing activity of an autonomous element, its insertion at any locus creates an unstable or "mutable" allele. Loss of the autonomous element itself, or of its ability to transpose, converts a mutable allele to a stable allele.

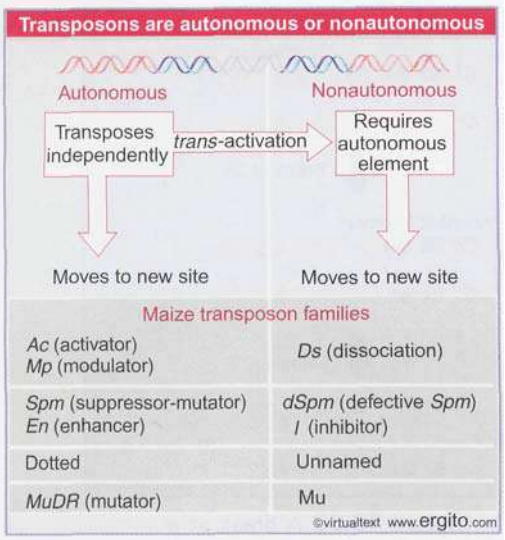


Figure 16.25 Each controlling element family has both autonomous and nonautonomous members. Autonomous elements are capable of transposition. Nonautonomous elements are deficient in transposition. Pairs of autonomous and nonautonomous elements can be classified in >4 families.

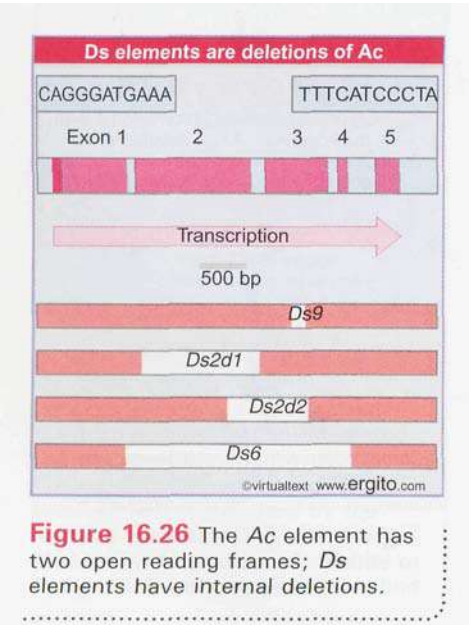


Figure 16.26 The *Ac* element has two open reading frames; *Ds* elements have internal deletions.

- Nonautonomous controlling elements are stable; they do not transpose or suffer other spontaneous changes in condition. They become unstable only when an autonomous member of the same family is present elsewhere in the genome. When complemented in *trans* by an autonomous element, a nonautonomous element displays the usual range of activities associated with autonomous elements, including the ability to transpose to new sites. Nonautonomous elements are derived from autonomous elements by loss of *trans*-acting functions needed for transposition.

Families of controlling elements are defined by the interactions between autonomous and nonautonomous elements. A family consists of a single type of autonomous element accompanied by many varieties of nonautonomous elements. A nonautonomous element is placed in a family by its ability to be activated in *trans* by the autonomous elements. The major families of controlling elements in maize are summarized in **Figure 16.25.**

Characterized at the molecular level, the maize transposons share the usual form of organization—inverted repeats at the ends and short direct repeats in the adjacent target DNA—but otherwise vary in size and coding capacity. All families of transposons share the same type of relationship between the autonomous and nonautonomous elements. The autonomous elements have open reading frames between the terminal repeats, whereas the nonautonomous elements do not code for functional proteins. Sometimes the internal sequences are related to those of autonomous elements; sometimes they have diverged completely.

The Mutator transposon is one of the simplest elements. The autonomous element MuDR codes for the genes *mudrA* (which codes for the MURA transposase) and *mudrB* (which codes for a nonessential accessory protein). The ends of the elements are marked by 200 bp inverted repeats. Nonautonomous elements—basically any unit that has the inverted repeats, which may not have any internal sequence relationship to MuDR—are also mobilized by MURA.

There are typically several members (~10) of each transposon family in a plant genome. By analyzing autonomous and nonautonomous elements of the Ac/Ds family, we have molecular information about many individual examples of these elements. **Figure 16.26** summarizes their structures.

Most of the length of the autonomous Ac element is occupied by a single gene consisting of 5 exons. The product is the transposase. The element itself ends in inverted repeats of 11 bp; and a target sequence of 8 bp is duplicated at the site of insertion.

Ds elements vary in both length and sequence, but are related to Ac. They end in the same 11 bp inverted repeats. They are shorter than Ac, and the length of deletion varies. At one extreme, the element Ds9 has a deletion of only 194 bp. In a more extensive deletion, the Ds6 element retains a length of only 2 kb, representing 1 kb from each end of Ac. A complex double Ds element has one Ds6 sequence inserted in reverse orientation into another.

Nonautonomous elements lack internal sequences, but possess the terminal inverted repeats (and possibly other sequence features). Nonautonomous elements are derived from autonomous elements by deletions (or other changes) that inactivate the *trans*-acting transposase, but leave intact the sites (including the termini) on which the transposase acts. Their structures range from minor (but inactivating) mutations of Ac to sequences that have major deletions or rearrangements.

At another extreme, the Ds1 family members comprise short sequences whose only relationship to Ac lies in the possession of terminal inverted repeats. Elements of this class need not be directly derived

from Ac, but could be derived by any event that generates the inverted repeats. Their existence suggests that the transposase recognizes only the terminal inverted repeats, or possibly the terminal repeats in conjunction with some short internal sequence.

Transposition of Ac/Ds occurs by a nonreplicative mechanism, and is accompanied by its disappearance from the donor location. Clonal analysis suggests that transposition of Ac/Ds almost always occurs soon after the donor element has been replicated. These features resemble transposition of the bacterial element Tn10 (see *16.10 Transposition of Tn10 has multiple controls).* The cause is the same: transposition does not occur when the DNA of the transposon is methylated on both strands (the typical state before methylation), and is activated when the DNA is hemimethylated (the typical state immediately after replication). The recipient site is frequently on the same chromosome as the donor site, and often quite close to it.

Replication generates two copies of a potential Ac/Ds donor, but usually only one copy actually transposes. What happens to the donor site? The rearrangements that are found at sites from which controlling elements have been lost could be explained in terms of the consequences of a chromosome break, as illustrated previously in Figure 16.23.

Autonomous and nonautonomous elements are subject to a variety of changes in their condition. Some of these changes are genetic, others are epigenetic.

The major change is (of course) the conversion of an autonomous element into a nonautonomous element, but further changes may occur in the nonautonomous element. *Cis*-acting defects may render a nonautonomous element impervious to autonomous elements. So a nonautonomous element may become permanently stable because it can no longer be activated to transpose.

Autonomous elements are subject to "changes of phase," heritable but relatively unstable alterations in their properties. These take the form of a reversible inactivation in which the element cycles between an active and inactive condition during plant development.

Phase changes in both the Ac and Mu types of autonomous element result from changes in the methylation of DNA. Comparisons of the susceptibilities of active and inactive elements to restriction enzymes suggest that the inactive form of the element is methylated in the target sequence GTC • There are several target sites in each element, and we do not know which sites control the effect. In the case of MuDR, demethylation of the terminal repeats increases transposase expression, suggesting that the effect may mediated through control of the promoter for the transposase gene. We should like to know what controls the methylation and demethylation of the elements.

The effect of methylation is common generally among transposons in plants. The best demonstration of the effect of methylation on activity comes from observations made with the *Arabidopsis* mutant *ddm1*, which causes a loss of methylation in heterochromatin. Among the targets that lose methyl groups is a family of transposons related to MuDR. Direct analysis of genome sequences shows that the demethylation causes transposition events to occur. Methylation is probably the major mechanism that is used to prevent transposons from damaging the genome by transposing too frequently.

There may be self-regulating controls of transposition, analogous to the immunity effects displayed by bacterial transposons. An increase in the number of Ac elements in the genome decreases the frequency of transposition. The Ac element may code for a repressor of transposition; the activity could be carried by the same protein that provides transposase function.

16.13 Spm elements influence gene expression

Key Concepts

- Spm elements affect gene expression at their sites of insertion, when the TnpA protein binds to its target sites at the ends of the transposon.
- Spm elements are inactivated by methylation.

The Spm and En autonomous elements are virtually identical; they differ at <10 positions. **Figure 16.27** summarizes the structure. The 13 bp inverted terminal repeats are essential for transposition, as indicated by the transposition-defective phenotype of deletions at the termini. Transposons related to Spm are found in other plants, and are defined as members of the same family by their generally similar organization. They all share nearly identical inverted terminal repeats, and generate 3 bp duplications of target DNA upon transposition. Named for the terminal similarities, they are known as the CACTA group of transposons.

A sequence of 8300 bp is transcribed from a promoter in the left end of the element. The 11 exons contained in the transcript are spliced into a 2500 base messenger. The mRNA codes for a protein of 621 amino acids. The gene is called *tnpA*, and the protein binds to a 12 bp consensus sequence present in multiple copies in the terminal regions of the element. Function of *tnpA* is required for excision, but may not be sufficient.

All of the nonautonomous elements of this family (denoted dSpm for defective Spm) are closely related in structure to the Spm element itself. They have deletions that affect the exons of *tnpA*.

Two additional open reading frames (ORF1 and ORF2) are located within the first, long intron of *tnpA*. They are contained in an alternatively spliced 6000 base RNA, which is present at 1% of the level of the *tnpA* mRNA. The function containing ORFs 1 and 2 is called *tnpB*. It may provide the protein that binds to the 13 bp terminal inverted repeats to cleave the termini for transposition.

In addition to the fully active Spm element, there are Spm-w derivatives that show weaker activity in transposition. The example given in Figure 16.27 has a deletion that eliminates both ORF1 and ORF2. This suggests that the need for TnpB in transposition can be bypassed or substituted.

Spm insertions can control the expression of a gene at the site of insertion. A recipient locus may be brought under either negative or positive control. An Spm-suppressible locus suffers inhibition of expression. An Spm-dependent locus is expressed only with the aid of Spm. When the inserted element is a dSpm, suppression or dependence responds to the *trans*-acting function supplied by an autonomous Spm. What is the basis for these opposite effects?

A dSpm-suppressible allele contains an insertion of dSpm within an exon of the gene. This structure raises the immediate question of how a gene with a dSpm insertion in an exon can ever be expressed! The dSpm sequence can be excised from the transcript by using sequences at its termini. The splicing event may leave a change in the sequence of the mRNA, thus explaining a change in the properties of the protein for which it codes. A similar ability to be excised from a transcript has been found for some Ds insertions.

tnpA provides the suppressor function for which the Spm element was originally named. The presence of a defective element may reduce, but not eliminate, expression of a gene in which it resides. However, the

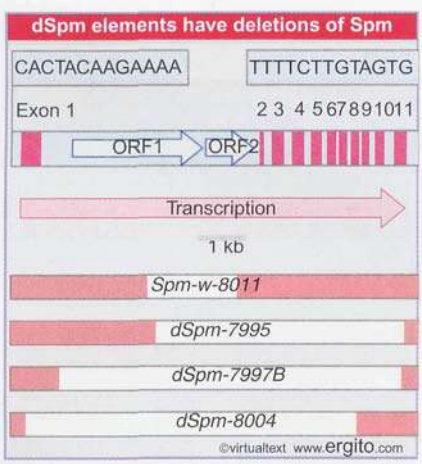


Figure 16.27 *Spm/En* has two genes. *tnpA* consists of 11 exons that are transcribed into a spliced 2500 base mRNA. *tnpB* may consist of a 6000 base mRNA containing ORF1 + ORF2.

introduction of an autonomous element, possessing a functional *tnpA* gene, may suppress expression of the target gene entirely. Suppression is caused by the ability of TnpA to bind to its target sites in the defective element, which blocks transcription from proceeding.

A dSpm-dependent allele contains an insertion near but not within a gene. The insertion appears to provide an enhancer that activates the promoter of the gene at the recipient locus.

Suppression and dependence at dSpm elements appear to rely on the same interaction between the *trans-acting* product of the *tnpA* gene of an autonomous Spm element and the *cis*-acting sites at the ends of the element. So a single interaction between the protein and the ends of the element either suppresses or activates a target locus depending on whether the element is located upstream of or within the recipient gene.

Spm elements exist in a variety of states ranging from fully active to cryptic. A cryptic element is silent and neither transposes itself nor activates dSpm elements. A cryptic element may be reactivated transiently or converted to the active state by interaction with a fully active Spm element. Inactivation is caused by methylation of sequences in the vicinity of the transcription startpoint. The nature of the events that are responsible for inactivating an element by *de novo* methylation or for activating it by demethylation (or preventing methylation) are not yet known.

16.14 The role of transposable elements in hybrid dysgenesis

Key Concepts

- P elements are transposons that are carried in P strains of *D. melanogaster* but not in M strains.
- When a P male is crossed with an M female, transposition is activated.
- The insertion of P elements at new sites in these crosses inactivates many genes and makes the cross infertile.

Certain strains of *D. melanogaster* encounter difficulties in interbreeding. When flies from two of these strains are crossed, the progeny display "dysgenic traits," a series of defects including mutations, chromosomal aberrations, distorted segregation at meiosis, and sterility. The appearance of these correlated defects is called hybrid dysgenesis.

Two systems responsible for hybrid dysgenesis have been identified in *D. melanogaster*. In the first, flies are divided into the types I (inducer) and R (reactive). Reduced fertility is seen in crosses of I males with R females, but not in the reverse direction. In the second system, flies are divided into the two types P (paternal contributing) and M (maternal contributing). **Figure 16.28** illustrates the asymmetry of the system; a cross between a P male and an M female causes dysgenesis, but the reverse cross does not.

Dysgenesis is principally a phenomenon of the germ cells. In crosses involving the P-M system, the F1 hybrid flies have normal somatic tissues. However, their gonads do not develop. The morphological defect in gamete development dates from the stage at which rapid cell divisions commence in the germline.

Any one of the chromosomes of a P male can induce dysgenesis in a cross with an M female. The construction of recombinant chromosomes shows that several regions within each P chromosome are able to cause

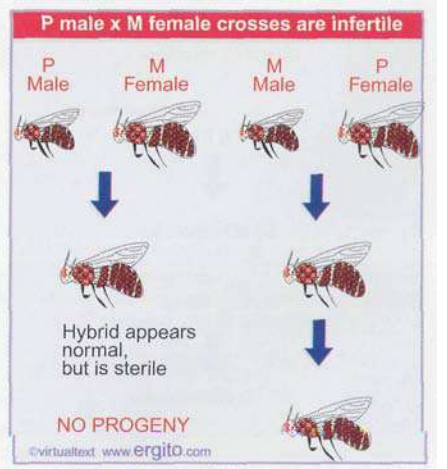


Figure 16.28 Hybrid dysgenesis is asymmetrical; it is induced by P male × M female crosses, but not by M male × P female crosses.

dysgenesis. This suggests that a P male has sequences at many different chromosomal locations that can induce dysgenesis. The locations differ between individual P strains. The P-specific sequences are absent from chromosomes of M flies.

The nature of the P-specific sequences was first identified by mapping the DNA of *w* mutants found among the dysgenic hybrids. All the mutations result from the insertion of DNA into the *w* locus. (The insertion inactivates the gene, causing the white-eye phenotype for which the locus is named.) The inserted sequence is called the P element.

The P element insertions form a classic transposable system. Individual elements vary in length but are homologous in sequence. All P elements possess inverted terminal repeats of 31 bp, and generate direct repeats of target DNA of 8 bp upon transposition. The longest P elements are ~2.9 kb long and have four open reading frames. The shorter elements arise, apparently rather frequently, by internal deletions of a full-length P factor. At least some of the shorter P elements have lost the capacity to produce the transposase, but may be activated in *trans* by the enzyme coded by a complete P element.

A P strain carries 30-50 copies of the P element, about a third of them full length. The elements are absent from M strains. In a P strain, the elements are carried as inert components of the genome. But they become activated to transpose when a P male is crossed with an M female.

Chromosomes from P-M hybrid dysgenic flies have P elements inserted at many new sites. The insertions inactivate the genes in which they are located and often cause chromosomal breaks. The result of the transpositions is therefore to inactivate the genome.

16.15 P elements are activated in the germline

Key Concepts

- P elements are activated in the germline of P male X M female crosses because a tissue-specific splicing event removes one intron, generating the coding sequence for the transposase.
- The P element also produces a repressor of transposition, which is inherited maternally in the cytoplasm.
- The presence of the repressor explains why M male x P female crosses remain fertile.

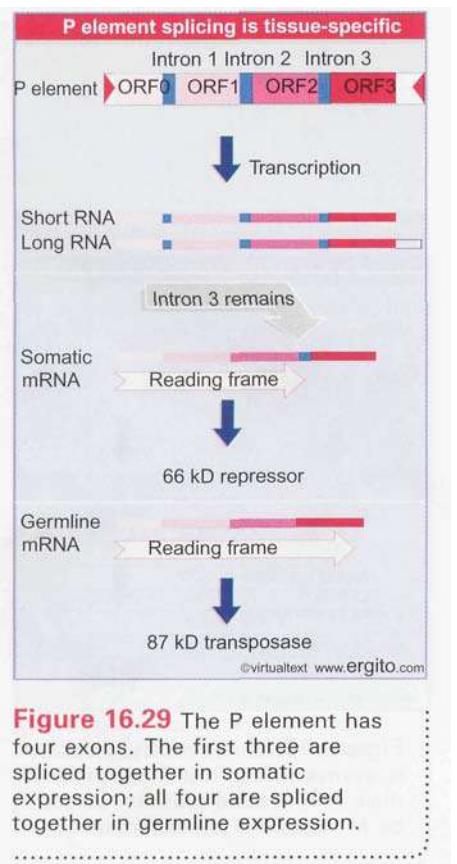


Figure 16.29 The P element has four exons. The first three are spliced together in somatic expression; all four are spliced together in germline expression.

Activation of P elements is tissue-specific, it occurs only in the germline. But P elements are transcribed in both germline and somatic tissues. Tissue-specificity is conferred by a change in the splicing pattern.

Figure 16.29 depicts the organization of the element and its transcripts. The primary transcript extends for 2.5 kb or 3.0 kb, the difference probably reflecting merely the leakiness of the termination site. Two protein products can be produced:

* In somatic tissues, only the first two introns are excised, creating a coding region of ORF0-ORF1-ORF2. Translation of this RNA yields a protein of 66 kD. This protein is a repressor of transposon activity.
- In germline tissues, an additional splicing event occurs to remove intron 3. This connects all four open reading frames into an mRNA that is translated to generate a protein of 87 kD. This protein is the transposase.

Two types of experiment have demonstrated that splicing of the third intron is needed for transposition. First, if the splicing junctions are mutated *in vitro* and the P element is reintroduced into flies, its transposi-

tion activity is abolished. Second, if the third intron is deleted, so that 0RF3 is constitutively included in the mRNA in all tissues, transposition occurs in somatic tissues as well as the germline.

So whenever ORF3 is spliced to the preceding reading frame, the P element becomes active. This is the crucial regulatory event, and usually it occurs only in the germline. What is responsible for the tissue-specific splicing? Somatic cells contain a protein that binds to sequences in exon 3 to prevent splicing of the last intron (see *24.12 Alternative splicing involves differential use of splice junctions*). The absence of this protein in germline cells allows splicing to generate the mRNA that codes for the transposase.

Transposition of a P element requires ~150 bp of terminal DNA. The transposase binds to 10 bp sequences that are adjacent to the 31 bp inverted repeats. Transposition occurs by a nonreplicative "cut and paste" mechanism resembling that of Tn10. (It contributes to hybrid dysgenesis in two ways. Insertion of the transposed element at a new site may cause mutations. And the break that is left at the donor site—see Figure 16.7—has a deleterious effect.)

It is interesting that, in a significant proportion of cases, the break in donor DNA is repaired by using the sequence of the homologous chromosome. If the homologue has a P element, the presence of a P element at the donor site may be restored (so the event resembles the result of a replicative transposition). If the homologue lacks a P element, repair may generate a sequence lacking the P element, thus apparently providing a precise excision (an unusual event in other transposable systems).

The dependence of hybrid dysgenesis on the sexual orientation of a cross shows that the cytoplasm is important as well as the P factors themselves. The contribution of the cytoplasm is described as the cytotype; a line of flies containing P elements has P cytotype, while a line of flies lacking P elements has M cytotype. Hybrid dysgenesis occurs only when chromosomes containing P factors find themselves in M cytotype, that is, when the male parent has P elements and the female parent does *not*.

Cytotype shows an inheritable cytoplasmic effect; when a cross occurs through P cytotype (the female parent has P elements), hybrid dysgenesis is suppressed for several generations of crosses with M female parents. So something in P cytotype, which can be diluted out over some generations, suppresses hybrid dysgenesis.

The effect of cytotype is explained in molecular terms by the model of **Figure 16.30.** It depends on the ability of the 66 kD protein to repress transposition. The protein is provided as a maternal factor in the egg. In a P line, there must be sufficient protein to prevent transposition from occurring, even though the P elements are present. In any cross involving a P female, its presence prevents either synthesis or activity of the transposase. But when the female parent is M type, there is no repressor in the egg, and the introduction of a P element from the male parent results in activity of transposase in the germline. The ability of P cytotype to exert an effect through more than one generation suggests that there must be enough repressor protein in the egg, and it must be stable enough, to be passed on through the adult to be present in the eggs of the next generation.

Strains of *D. melanogaster* descended from flies caught in the wild more than 30 years ago are always M. Strains descended from flies caught in the past 10 years are almost always P. Does this mean that the P element family has invaded wild populations of

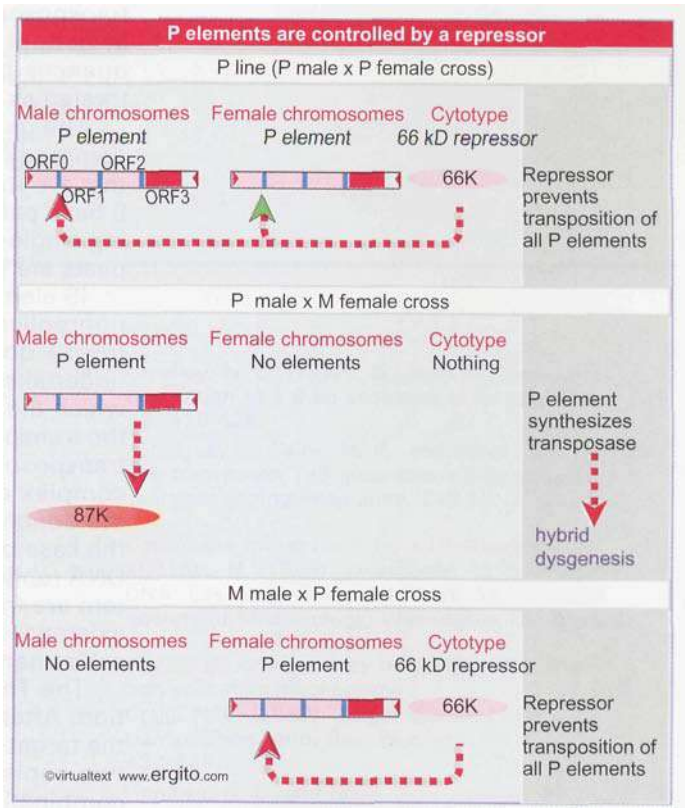


Figure 16.30 Hybrid dysgenesis is determined by the interactions between P elements in the genome and 66 kD repressor in the cytotype.

D. melanogaster in recent years? P elements are indeed highly invasive when introduced into a new population; the source of the invading element would have to be another species.

Because hybrid dysgenesis reduces interbreeding, it is a step on the path to speciation. Suppose that a dysgenic system is created by a transposable element in some geographic location. Another element may create a different system in some other location. Flies in the two areas will be dysgenic for two (or possibly more) systems. If this renders them intersterile and the populations become genetically isolated, further separation may occur. Multiple dysgenic systems therefore lead to inability to mate—and to speciation.

16.16 Summary

Prokaryotic and eukaryotic cells contain a variety of transposons that mobilize by moving or copying DNA sequences. The transposon can be identified only as an entity within the genome; its mobility does not involve an independent form. The transposon could be selfish DNA, concerned only with perpetuating itself within the resident genome; if it conveys any selective advantage upon the genome, this must be indirect. All transposons have systems to limit the extent of transposition, since unbridled transposition is presumably damaging, but the molecular mechanisms are different in each case.

The archetypal transposon has inverted repeats at its termini and generates directs repeats of a short sequence at the site of insertion. The simplest types are the bacterial insertion sequences (IS), which consist essentially of the inverted terminal repeats flanking a coding frame(s) whose product(s) provide transposition activity. Composite transposons have terminal modules that consist of IS elements; one or both of the IS modules provides transposase activity, and the sequences between them (often carrying antibiotic resistance), are treated as passengers.

The generation of target repeats flanking a transposon reflects a common feature of transposition. The target site is cleaved at points that are staggered on each DNA strand by a fixed distance (often 5 or 9 base pairs). The transposon is in effect inserted between protruding single-stranded ends generated by the staggered cuts. Target repeats are generated by filling in the single-stranded regions.

IS elements, composite transposons, and P elements mobilize by nonreplicative transposition, in which the element moves directly from a donor site to a recipient site. A single transposase enzyme undertakes the reaction. It occurs by a "cut and paste" mechanism in which the transposon is separated from flanking DNA. Cleavage of the transposon ends, nicking of the target site, and connection of the transposon ends to the staggered nicks, all occur in a nucleoprotein complex containing the transposase. Loss of the transposon from the *donor creates a double-strand break,* whose fate is not clear. In the case of Tn10, transposition becomes possible immediately after DNA replication, when sites recognized by the *dam* methylation system are transiently hemimethylated. This imposes a demand for the existence of two copies of the donor site, which may enhance the cell's chances for survival.

The TnA family of transposons mobilize by replicative transposition. After the transposon at the donor site becomes connected to the target site, replication generates a cointegrate molecule that has two copies of the transposon. A resolution reaction, involving recombination between two particular sites, then frees the two copies of the transposon, so that one remains at the donor site and one appears at the target site. Two enzymes coded by the transposon are required: transposase recognizes the ends of the transposon and connects them to the target site; and resolvase provides a site-specific recombination function.

Phage Mu undergoes replicative transposition by the same mechanism as TnA. It also can use its cointegrate intermediate to transpose by a nonreplicative mechanism. The difference between this reaction and the nonreplicative transposition of IS elements is that the cleavage events occur in a different order.

The best characterized transposons in plants are the controlling elements of maize, which fall into several families. Each family contains a single type of autonomous element, analogous to bacterial transposons in its ability to mobilize. A family also contains many different nonautonomous elements, derived by mutations (usually deletions) of the autonomous element. The nonautonomous elements lack the ability to transpose, but display transposition activity and other abilities of the autonomous element, when an autonomous element is present to provide the necessary frans-acting functions.

In addition to the direct consequences of insertion and excision, the maize elements may also control the activities of genes at or near the sites where they are inserted; this control may be subject to developmental regulation. Maize elements inserted into genes may be excised from the transcripts, which explains why they do not simply impede gene activity. Control of target gene expression involves a variety of molecular effects, including activation by provision of an enhancer and suppression by interference with post-transcriptional events.

Transposition of maize elements (in particular Ac) is nonreplicative, probably requiring only a single transposase enzyme coded by the element. Transposition occurs preferentially after replication of the element. There are probably mechanisms to limit the frequency of transposition. Advantageous rearrangements of the maize genome may have been connected with the presence of the elements.

P elements in *D. melanogaster* are responsible for hybrid dysgenesis, which could be a forerunner of speciation. A cross between a male carrying P elements and a female lacking them generates hybrids that are sterile. A P element has 4 open reading frames, separated by introns. Splicing of the first 3 ORFs generates a 66 kD repressor, and occurs in all cells. Splicing of all 4 ORFs to generate the 87 kD transposase occurs only in the germline, by a tissue-specific splicing event. P elements mobilize when exposed to cytoplasm lacking the repressor. The burst of transposition events inactivates the genome by random insertions. Only a complete P element can generate transposase, but defective elements can be mobilized in *trans* by the enzyme.

References

16.1 Introduction

rev Campbell, A. (1981). Evolutionary significance of accessory DNA elements in bacteria. Ann. Rev. Immunol. 35, 55-83.

Finnegan, D. J. (1985). Transposable elements in eukaryotes. Int. Rev. Cytol. 93, 281-326.

16.2 Insertion sequences are simple transposition modules

rev Berg, D. E. and Howe, M. (1989). Mobile DNA. American Society for Microbiology, Washington DC.

Calos, M. and Miller, J. H. (1980). Transposable elements. Cell 20, 579-595.

Craig, N. L. (1997). Target site selection in transposition. Ann. Rev. Biochem. 66, 437-474.

Galas, D. J. and Chandler, M. (1989). Bacterial insertion sequences. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 109-162.

Kleckner, N. (1977). Translocatable elements in prokaryotes. Cell 11, 11-23.

Kleckner, N. (1981). Transposable elements in prokaryotes. Ann. Rev. Genet. 15, 341-404.

ref Grindley, N. D. (1978). IS1 insertion generates duplication of a 9 bp sequence at its target site. Cell 13, 419-426.

Johnsrud, L, Calos, M. P., and Miller, J. H. (1978). The transposon Tn9 generates a 9 bp repeated sequence during integration. Cell 15, 1209-1219.

16.3 Composite transposons have IS modules

rev Kleckner, N. (1989). Transposon Tn10. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 227-268.

16.4 Transposition occurs by both replicative and nonreplicative mechanisms

rev Craig, N. L. (1997). Target site selection in transposition. Ann. Rev. Biochem. 66, 437-474.

Grindley, N. D. and Reed, R. R. (1985). Transpositional recombination in prokaryotes. Ann. Rev. Biochem. 54, 863-896.

Haren, L, Ton-Hoang, B., and Chandler, M. (1999). Integrating DNA: transposases and retroviral integrases. Ann. Rev. Microbiol. 53, 245-281.

Scott, J. R. and Churchward, G. G. (1995). Conjugative transposition. Ann. Rev. Immunol. 49, 367-397.

16.6 Common intermediates for transposition

rev Mizuuchi, K. (1992). Transpositional recombination: mechanistic insights from studies of Mu and other elements. Ann. Rev. Biochem. 61, 1011 1051.

Pato, M. L. (1989). Bacteriophage Mu. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 23-52.

ref Aldaz, H., Schuster, E., and Baker, T. A. (1996). The interwoven architecture of the Mu transposase couples DNA synthesis to catalysis. Cell 85, 257-269.

Savilahti, H. and Mizuuchi, K. (1996). Mu transpositional recombination: donor DNA cleavage and strand transfer in *trans* by the Mu transpose. Cell 85, 271-280.

16.8 Nonreplicative transposition proceeds by breakage and reunion

ref Bender, J. and Kleckner, N. (1986). Genetic evidence that Tn10 transposes by a nonreplicative mechanism. Cell 45, 801-815.

Bolland, S. and Kleckner, N. (1996). The three chemical steps of Tn10/IS10 transposition involve repeated utilization of a single active site. Cell 84, 223-233.

Davies, D. R., Goryshin, I. Y., Reznikoff, W. S., Rayment, I., Davies, D. R., Goryshin, I. Y., Reznikoff, W. S., and Rayment, I. (2000). Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science 289, 77-85.

Haniford, D. B., Benjamin, H. W., and Kleckner, N. (1991). Kinetic and structural analysis of a cleaved donor intermediate and a strand transfer intermediate in Tn10 transposition. Cell 64, 171-179.

Kennedy, A. K., Guhathakurta, A., Kleckner, N., and Haniford, D. B. (1998). Tn10 transposition via a DNA hairpin intermediate. Cell 95, 125-134.

16.9 TnA transposition requires transposase and resolvase

rev Sherratt, D. (1989). Tn3 and related transposable elements: site-specific recombination and transposition. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 163-185.

ref Droge, P. et al. (1990). The two functional domains of gamma delta resolvase act on the same recombination site: implications for the mechanism of strand exchange. Proc. Nat. Acad. Sci. USA 87, 5336-5340.

Grindley, N. D. et al. (1982). Transposon-mediated site-specific recombination: identification of three binding sites for resolvase at the *res* sites of γ 5 and Tn3. Cell 30, 19-27.

16.10 Transposition of Tn10 has multiple controls

rev Kleckner, N. (1990). Regulation of transposition in bacteria. Ann. Rev. Cell Biol. 6, 297-327.

Kleckner, N. (1989). Transposon Tn10. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 227-268.

ref Roberts, D. et al. (1985). IS10 transposition is regulated by DNA adenine methylation. Cell 43, 117-130.

16.11 Controlling elements in maize cause breakage and rearrangements

exp Fedoroff, N. (2002). The Discovery of Transposition (www.ergito.com/lookup.jsp?expt = fedoroff)

16.12 Controlling elements form families of transposons

rev Fedoroff, N. (2000). Transposons and genome evolution in plants. Proc. Nat. Acad. Sci. USA 97, 7002-7007.

Fedoroff, N. (1989). Maize transposable elements. Mobile DNA, Eds. Berg, D. E., and Howe, N. American Society of Microbiology, Washington DC 375-412.

Gierl, A., Saedler, H., and Peterson, P. A. (1989). Maize transposable elements. Ann. Rev. Genet. 23, 71-85.

ref Benito, M. I. and Walbot, V. (1997). Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. Mol. Cell Biol. 17, 5165-5175.

Chandler, V. L. and Walbot, V. (1986). DNA modification of a maize transposable element correlates with loss of activity. Proc. Nat. Acad. Sci. USA 83, 1767-1771.

Ros, F. and Kunze, R. (2001). Regulation of activator/dissociation transposition by replication and DNA methylation. Genetics 157, 1723-1733.

Singer, T., Yordan, C, and Martienssen, R. A. (2001). Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev. 15, 591-602.

16.14 The role of transposable elements in hybrid dysgenesis

rev Engels, W. R. '(1383). The P family of transposable elements in *Drosophila.* Ann. Rev. Genet. 17, 315-344.

Engels, W. R. (1989). P elements in *D. melanogaster* In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 437-484.

16.15 P elements are activated in the germline

ref Laski, F. A., Rio, D. C, and Rubin, G. M. (1986). Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing. Cell 44, 7-19.

Chapter 17

Retroviruses and retroposons

17.1 Introduction

Transposition that involves an obligatory intermediate of RNA is unique to eukaryotes, and is provided by the ability of retroviruses to insert DNA copies (proviruses) of an RNA viral genome into the chromosomes of a host cell. Some eukaryotic transposons are related to retroviral proviruses in their general organization, and they transpose through RNA intermediates. As a class, these elements are called retroposons (or sometimes retrotransposons). The very simplest such elements do not themselves have transposition activity, but have sequences that are recognized as substrates for transposition by active elements. So elements that use RNA-dependent transposition range from the retroviruses themselves, able freely to infect host cells, to sequences that transpose via RNA, to those that do not themselves possess the ability to transpose. They share with all transposons the diagnostic feature of generating short direct repeats of target DNA at the site of an insertion.

Even in genomes where active transposons have not been detected, footprints of ancient transposition events are found in the form of direct target repeats flanking dispersed repetitive sequences. The features of these sequences sometimes implicate an RNA sequence as the progenitor of the genomic (DNA) sequence. This suggests that the RNA must have been converted into a duplex DNA copy that was inserted into the genome by a transposition-like event.

Like any other reproductive cycle, the cycle of a retrovirus or retroposon is continuous; it is arbitrary at which point we interrupt it to consider a "beginning." But our perspectives of these elements are biased by the forms in which we usually observe them, indicated in **Figure 17.1**. Retroviruses were first observed as infectious virus particles, capable of transmission between cells, and so the intracellular cycle (involving duplex DNA) is thought of as the means of reproducing the RNA virus. Retroposons were discovered as components of the genome; and the RNA forms have been mostly characterized for their functions as mRNAs. So we think of retroposons as genomic (duplex DNA) sequences that may transpose within a genome; they do not migrate between cells.

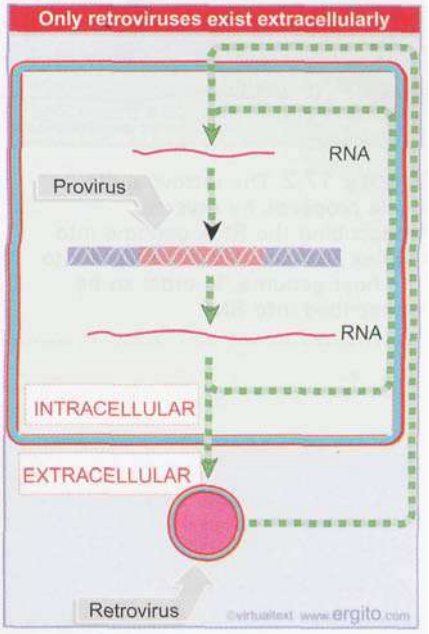


Figure 17.1 The reproductive cycles of retroviruses and retroposons involve alternation of reverse transcription from RNA to DNA with transcription from DNA to RNA. Only retroviruses can generate infectious particles. Retroposons are confined to an intracellular cycle.

17.2 The retrovirus life cycle involves transposition-like events

Key Concepts

- A retrovirus has two copies of its genome of single-stranded RNA.
- An integrated provirus is a double-stranded DNA sequence.
- A retrovirus generates a provirus by reverse transcription of the retroviral genome.

By Book_Crazy [IND]

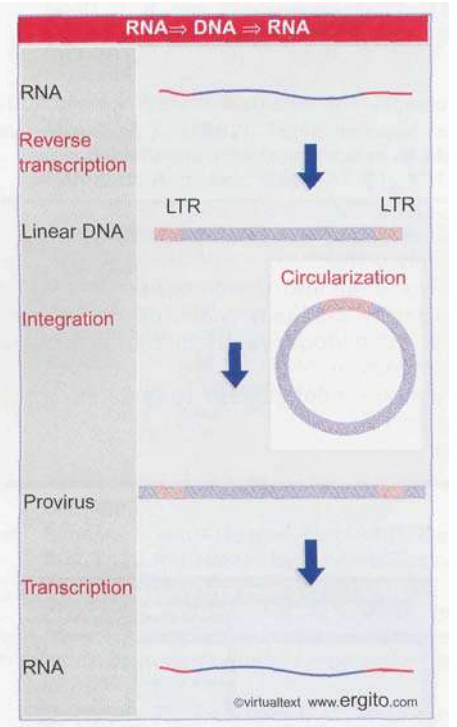


Figure 17.2 The retroviral life cycle proceeds by reverse transcribing the RNA genome into duplex DNA, which is inserted into the host genome, in order to be transcribed into RNA.

Retroviruses have genomes of single-stranded RNA that are replicated through a double-stranded DNA intermediate. The life cycle of the virus involves an obligatory stage in which the double-stranded DNA is inserted into the host genome by a transposition-like event that generates short direct repeats of target DNA.

The significance of this reaction extends beyond the perpetuation of the virus. Some of its consequences are that

- A retroviral sequence that is integrated in the germline remains in the cellular genome as an endogenous provirus. Like a lysogenic bacteriophage, a provirus behaves as part of the genetic material of the organism.
- Cellular sequences occasionally recombine with the retroviral sequence and then are transposed with it; these sequences may be inserted into the genome as duplex sequences in new locations.
- Cellular sequences that are transposed by a retrovirus may change the properties of a cell that becomes infected with the virus.

The particulars of the retroviral life cycle are expanded in **Figure 17.2.** The crucial steps are that the viral RNA is converted into DNA, the DNA becomes integrated into the host genome, and then the DNA provirus is transcribed into RNA.

The enzyme responsible for generating the initial DNA copy of the RNA is reverse transcriptase. The enzyme converts the RNA into a linear duplex of DNA in the cytoplasm of the infected cell. The DNA also is converted into circular forms, but these do not appear to be involved in reproduction.

The linear DNA makes its way to the nucleus. One or more DNA copies become integrated into the host genome. A single enzyme, called integrase, is responsible for integration. The provirus is transcribed by the host machinery to produce viral RNAs, which serve both as mRNAs and as genomes for packaging into virions. Integration is a normal part of the life cycle and is necessary for transcription.

Two copies of the RNA genome are packaged into each virion, making the individual virus particle effectively diploid. When a cell is simultaneously infected by two different but related viruses, it is possible to generate heterozygous virus particles carrying one genome of each type. The diploidy may be important in allowing the virus to acquire cellular sequences. The enzymes reverse transcriptase and integrase are carried with the genome in the viral particle.

17.3 Retroviral genes code for polyproteins

Key Concepts

- A typical retrovirus has three genes, *gag, pol, env.*
- Gag and Pol proteins are translated from a full-length transcript of the genome.
- Translation of Pol requires a frameshift by the ribosome.
- Env is translated from a separate mRNA that is generated by splicing.
- Each of the three protein products is processed by proteases to give multiple proteins.

A typical retroviral sequence contains three or four "genes," the term here identifying coding regions each of which actually gives rise to multiple proteins by processing reactions. A typical retrovirus genome with three genes is organized in the sequence *gag-pol-env* as indicated in **Figure 17.3.**

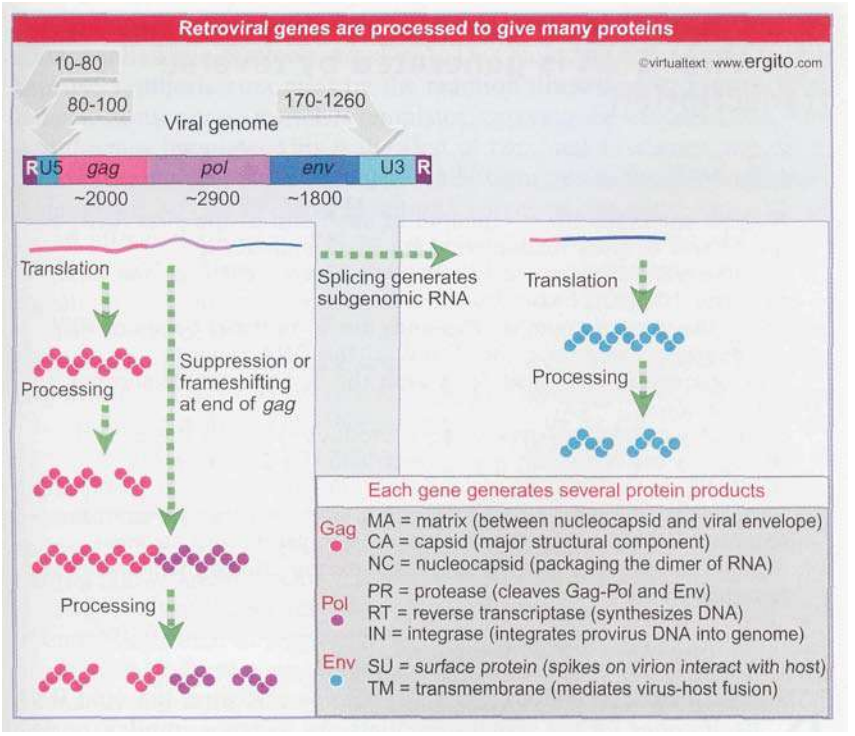


Figure 17.3 The genes of the retrovirus are expressed as polyproteins that are processed into individual products.

Retroviral mRNA has a conventional structure; it is capped at the 5' end and polyadenylated at the 3' end. It is represented in two mRNAs. The full length mRNA is translated to give the Gag and Pol polyproteins. The Gag product is translated by reading from the initiation codon to the first termination codon. This termination codon must be bypassed to express Pol.

Different mechanisms are used in different viruses to proceed beyond the *gag* termination codon, depending on the relationship between the *gag* and *pol* reading frames. When *gag* and *pol* follow continuously, suppression by a glutamyl-tRNA that recognizes the termination codon allows a single protein to be generated. When *gag* and *pol* are in different reading frames, a ribosomal frameshift occurs to generate a single protein. Usually the readthrough is ~5% efficient, so Gag protein outnumbers Gag-Pol protein about 20-fold.

The Env polyprotein is expressed by another means: splicing generates a shorter *subgenomic* messenger that is translated into the Env product.

The *gag* gene gives rise to the protein components of the nucleoprotein core of the virion. The *pol* gene codes for functions concerned with nucleic acid synthesis and recombination. The *env* gene codes for components of the envelope of the particle, which also sequesters components from the cellular cytoplasmic membrane.

Both the Gag or Gag-Pol and the Env products are polyproteins that are cleaved by a protease to release the individual proteins that are found in mature virions. The protease activity is coded by the virus in various forms: it may be part of Gag or Pol, or sometimes takes the form of an additional independent reading frame.

The production of a retroviral particle involves packaging the RNA into a core, surrounding it with capsid proteins, and pinching off a segment of membrane from the host cell. The release of infective particles by such means is shown in **Figure 17.4.** The process is reversed during infection; a virus infects a new host cell by fusing with the plasma membrane and then releasing the contents of the virion.

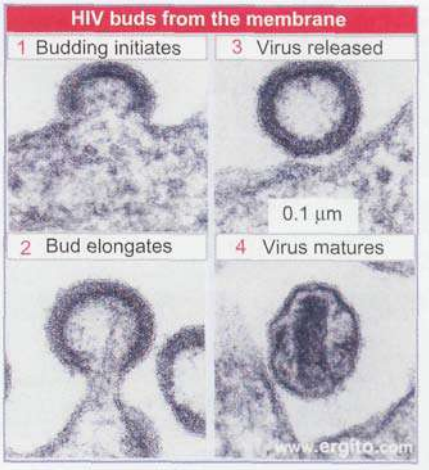


Figure 17.4 Retroviruses (HIV) bud from the plasma membrane of an infected cell. Photograph kindly provided by Matthew Gonda.

17.4 Viral DNA is generated by reverse transcription

Key Concepts

- A short sequence (R) is repeated at each end of the viral RNA, so the 5' and 3' ends respectively are R-U5 and U3-R.
- Reverse transcriptase starts synthesis when a tRNA primer binds to a site 100-200 bases from the 5' end.
- When the enzyme reaches the end, the 5'-terminal bases of RNA are degraded, exposing the 3' end of the DNA product.
- The exposed 3' end base pairs with the 3' terminus of another RNA genome.
- Synthesis continues, generating a product in which the 5' and 3' regions are repeated, giving each end the structure U3-R-U5.
- Similar strand switching events occur when reverse transcriptase uses the DNA product to generate a complementary strand.
- Strand switching is an example of the copy choice mechanism of recombination.

Retroviruses are called plus strand viruses, because the viral RNA itself codes for the protein products. As its name implies, reverse transcriptase is responsible for converting the genome (plus strand RNA) into a complementary DNA strand, which is called the minus strand DNA. Reverse transcriptase also catalyzes subsequent stages in the production of duplex DNA. It has a DNA polymerase activity, which enables it to synthesize a duplex DNA from the single-stranded reverse transcript of the RNA. The second DNA strand in this duplex is called plus strand DNA. And as a necessary adjunct to this activity, the enzyme has an RNAase H activity, which can degrade the RNA part of the RNA-DNA hybrid. All retroviral reverse transcriptases share considerable similarities of amino acid sequence, and homologous sequences can be recognized in some other retroposons.

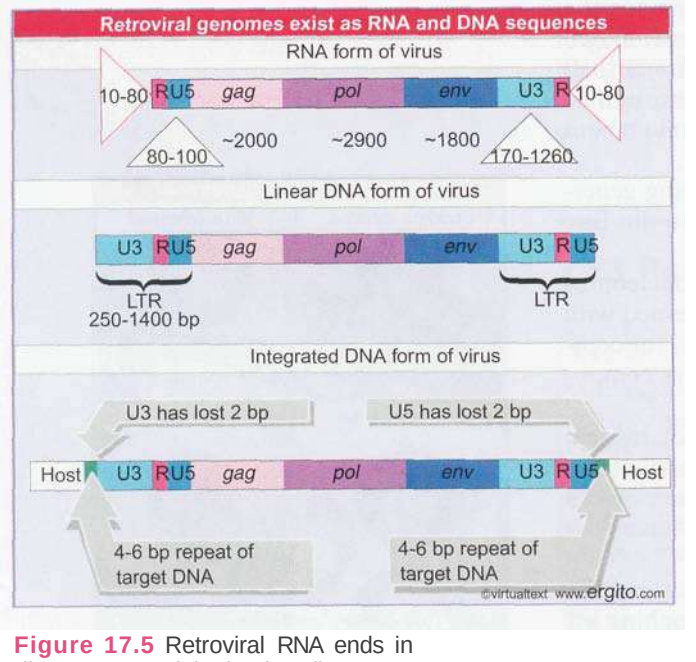


Figure 17.5 Retroviral RNA ends in direct repeats (R), the free linear DNA ends in LTRs, and the provirus ends in LTRs that are shortened by two bases each.

The structures of the DNA forms of the virus are compared with the RNA in **Figure 17.5.** The viral RNA has direct repeats at its ends. These R segments vary in different strains of virus from 10-80 nucleotides. The sequence at the 5' end of the virus is R-U5, and the sequence at the 3' end is U3-R. The R segments are used during the conversion from the RNA to the DNA form to generate the more extensive direct repeats that are found in linear DNA (see Figure 17.6 and Figure 17.7). The shortening of 2 bp at each end in the integrated form is a consequence of the mechanism of integration (see Figure 17.9).

Like other DNA polymerases, reverse transcriptase requires a primer. The native primer is tRNA. An uncharged host tRNA is present in the virion. A sequence of 18 bases at the 3' end of the tRNA is base paired to a site 100-200 bases from the 5' end of one of the viral RNA molecules. The tRNA may also be base paired to another site near the 5' end of the other viral RNA, thus assisting in dimer formation between the viral RNAs.

Here is a dilemma. Reverse transcriptase starts to synthesize DNA at a site only 100-200 bases downstream from the 5' end. How can DNA be generated to represent the intact RNA genome? (This is an extreme variant of the general problem in replicating the ends of any linear nucleic acid; see *13.8 The ends of linear DNA are a problem for replication.)*

By Book_Crazy [IND]

Synthesis *in vitro* proceeds to the end, generating a short DNA sequence called minus strong-stop DNA. This molecule is not found *in vivo* because synthesis continues by the reaction illustrated in **Figure 17.6.** Reverse transcriptase switches templates, carrying the nascent DNA with it to the new template. This is the first of two jumps between templates.

In this reaction, the R region at the 5' terminus of the RNA template is degraded by the RNAase H activity of reverse transcriptase. Its removal allows the R region at a 3' end to base pair with the newly synthesized DNA. Then reverse transcription continues through the U3 region into the body of the RNA.

The source of the R region that pairs with the strong stop minus DNA can be either the 3' end of the same RNA molecule (intramolecular pairing) or the 3' end of a different RNA molecule (intermolecular pairing). The switch to a different RNA template is used in the figure because there is evidence that the sequence of the tRNA primer is not inherited in a retroposon life cycle. (If intramolecular pairing occurred, we would expect the sequence to be inherited, because it would provide the only source for the primer binding sequence in the next cycle. Intermolecular pairing allows another retroviral RNA to provide this sequence.)

The result of the switch and extension is to add a U3 segment to the 5' end. The stretch of sequence U3-R-U5 is called the long terminal repeat (LTR) because a similar series of events adds a U5 segment to the 3' end, giving it the same structure of U5-R-U3. Its length varies from 250-1400 bp (see Figure 17.5).

We now need to generate the plus strand of DNA and to generate the LTR at the other end. The reaction is shown in **Figure 17.7.** Reverse transcriptase primes synthesis of plus strand DNA from a fragment of RNA that is left after degrading the original RNA molecule. A strong stop plus strand DNA is generated when the enzyme reaches the end of the template. This DNA is then transferred to the other end of a minus strand. Probably it is released by a displacement reaction when a second round of DNA synthesis occurs from a primer fragment farther upstream (to its left in the figure). It uses the R region to pair with the 3' end of a minus strand DNA. This double-stranded DNA then requires completion of both strands to generate a duplex LTR at each end.

Each retroviral particle carries two RNA genomes. This makes it possible for recombination to occur during a viral life cycle. In principle this could occur during minus strand synthesis and/or during plus strand synthesis:

- The intermolecular pairing shown in Figure 17.6 allows a recombination to occur between sequences of the two successive RNA templates when minus strand DNA is synthesized. Retroviral recombination is mostly due to strand transfer at this stage, when the nascent DNA strand is transferred from one RNA template to another during reverse transcription.
- Plus strand DNA may be synthesized discontinuously, in a reaction that involves several internal initiations. Strand transfer during this reaction can also occur, but is less common.

The common feature of both events is that recombination results from a change in the template during the act of DNA synthesis. This is a general example of a mechanism for recombination called copy choice. For many years this was regarded as a possible mechanism for general recombination. It is unlikely to be employed by cellular systems, but is a common basis for recombination during infection by RNA viruses, including those that replicate exclusively through RNA forms, such as poliovirus.

Strand switching occurs with a certain frequency during each cycle of reverse transcription, that is, in addition to the transfer reaction that is forced at the end of the template strand. The principle is illustrated

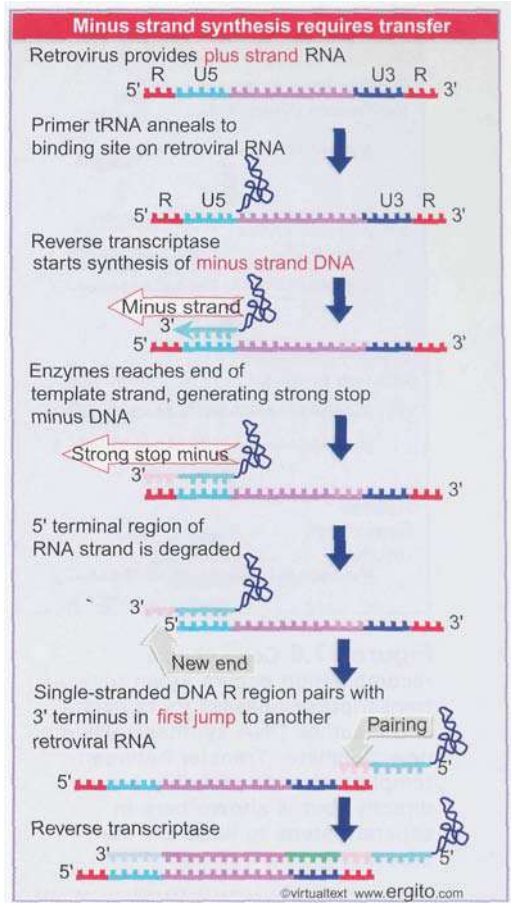


Figure 17.6 Minus strand DNA is generated by switching templates during reverse transcription.

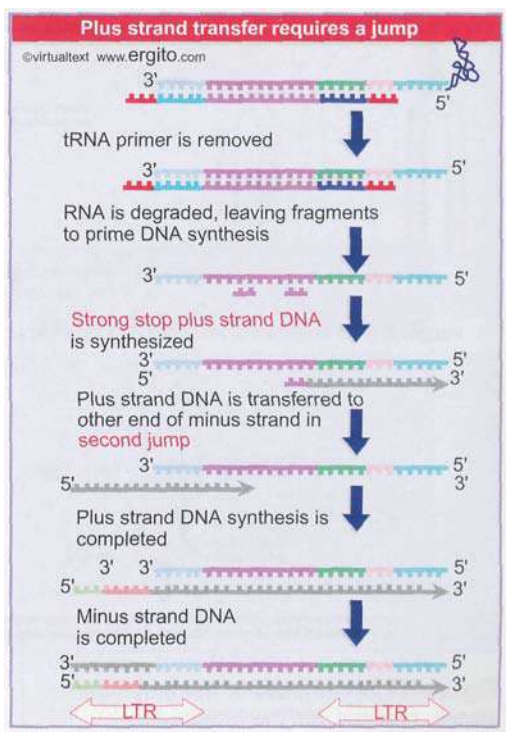


Figure 17.7 Synthesis of plus strand DNA requires a second jump.

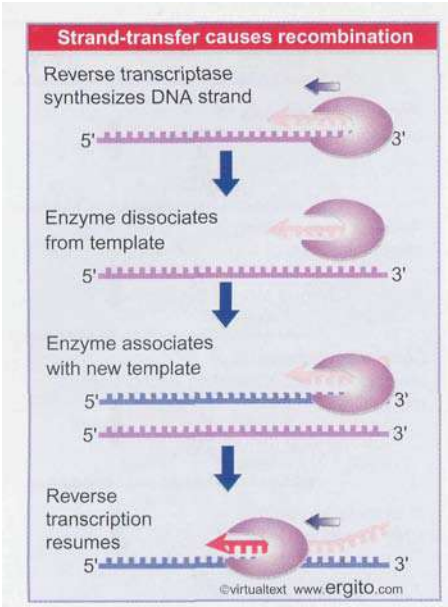


Figure 17.8 Copy choice recombination occurs when reverse transcriptase releases its template and resumes DNA synthesis using a *new template. Transfer between template strands probably occurs* directly, but is shown here in separate steps to illustrate the process.

in **Figure 17.8,** although we do not know much about the mechanism. Reverse transcription *in vivo* occurs in a ribonucleoprotein complex, in which the RNA template strand is bound to virion components, including the major protein of the capsid. In the case of HIV, addition of this protein (NCp7) to an *in vitro* system causes recombination to occur. The effect is probably indirect: NCp7 affects the structure of the RNA template, which in turn affects the likelihood that reverse transcriptase will switch from one template strand to another.

17.5 Viral DNA integrates into the chromosome

Key Concepts

- The organization of proviral DNA in a chromosome is the same as a transposon, with the provirus flanked by short direct repeats of a sequence at the target site.
- Linear DNA is inserted directly into the host chromosome by the retroviral integrase enzyme.
- Two base pairs of DNA are lost from each end of the retroviral sequence during the integration reaction.

The organization of the integrated provirus resembles that of the linear DNA. The LTRs at each end of the provirus are identical. The 3' end of U5 consists of a short inverted repeat relative to the 5' end of U3, so the LTR itself ends in short inverted repeats. The integrated proviral DNA is like a transposon; the proviral sequence ends in inverted repeats and is flanked by short direct repeats of target DNA.

The provirus is generated by directly inserting a linear DNA into a target site. (In addition to linear DNA, there are circular forms of the viral sequences. One has two adjacent LTR sequences generated by joining the linear ends. The other has only one LTR, presumably generated by a recombination event and actually comprising the majority of circles. Although for a long time it appeared that the circle might be an integration intermediate [by analogy with the integration of lambda DNA], we now know that the linear form is used for integration.)

Integration of linear DNA is catalyzed by a single viral product, the integrase. Integrase acts on both the retroviral linear DNA and the target DNA. The reaction is illustrated in **Figure 17.9.**

The ends of the viral DNA are important; as is the case with transposons, mutations in the ends prevent integration. The most conserved feature is the presence of the dinucleotide sequence CA close to the end of each inverted repeat. The integrase brings the ends of the linear DNA together in a ribonucleoprotein complex, and converts the blunt ends into recessed ends by removing the bases beyond the conserved CA; usually this involves \oss of *1* bases.

Target sites are chosen at random with respect to sequence. The integrase makes staggered cuts at a target site. In the example of Figure 17.9, the cuts are separated by 4 bp. The length of the target repeat depends on the particular virus; it may be 4, 5, or 6 bp. Presumably it is determined by the geometry of the reaction of integrase with target DNA.

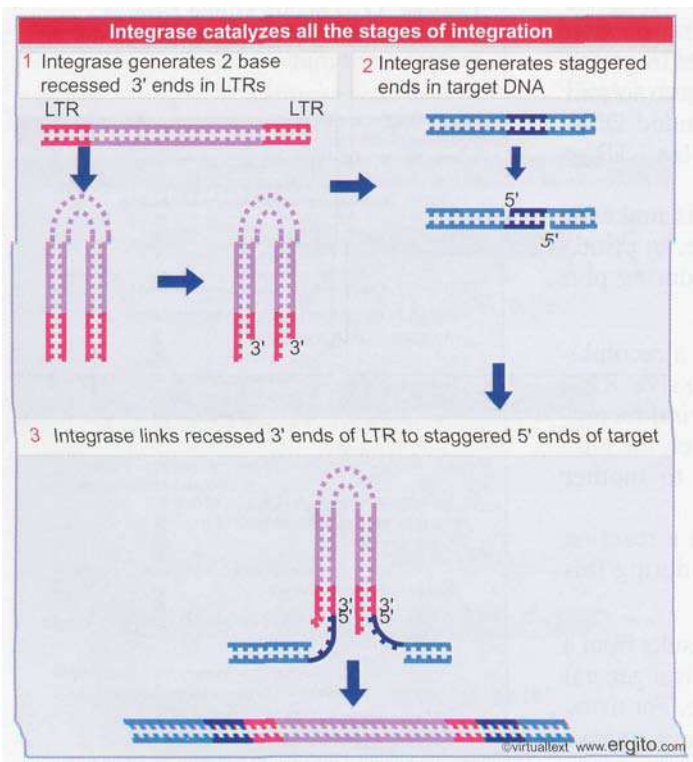


Figure 17.9 Integrase is the only viral protein required for the integration reaction, in which each LTR loses 2 bp and is inserted between 4 bp repeats of target DNA.

The 5′ ends generated by the cleavage of target DNA are covalently joined to the 3' recessed ends of the viral DNA. At this point, both termini of the viral DNA are joined by one strand to the target DNA. The single-stranded region is repaired by enzymes of the host cell, and in the course of this reaction the protruding 2 bases at each 5′ end of the viral DNA are removed. The result is that the integrated viral DNA has lost 2 bp at each LTR; this corresponds to the loss of 2 bp from the left end of the 5' terminal U3 and loss of 2 bp from the right end of the 3' terminal U5. There is a characteristic short direct repeat of target DNA at each end of the integrated retroviral genome.

The viral DNA integrates into the host genome at randomly selected sites. A successfully infected cell gains 1-10 copies of the provirus. (An infectious virus enters the cytoplasm, of course, but the DNA form becomes integrated into the genome in the nucleus. Retroviruses can replicate only in proliferating cells, because entry into the nucleus requires the cell to pass through mitosis, when the viral genome gains access to the nuclear material.)

The U3 region of each LTR carries a promoter. The promoter in the left LTR is responsible for initiating transcription of the provirus. Recall that the generation of proviral DNA is required to place the U3 sequence at the left LTR; so we see that the promoter is in fact generated by the conversion of the RNA into duplex DNA.

Sometimes (probably rather rarely), the promoter in the right LTR sponsors transcription of the host sequences that are adjacent to the site of integration. The LTR also carries an enhancer (a sequence that activates promoters in the vicinity) that can act on cellular as well as viral sequences. Integration of a retrovirus can be responsible for converting a host cell into a tumorigenic state when certain types of genes are activated in this way (see *30.6 Retroviruses activate or incorporate cellular genes).*

Can integrated proviruses be excised from the genome? Homologous recombination could take place between the LTRs of a provirus; solitary LTRs that could be relics of an excision event are present in some cellular genomes.

We have dealt so far with retroviruses in terms of the infective cycle, in which integration is necessary for the production of further copies of the RNA. However, when a viral DNA integrates in a germline cell, it becomes an inherited "endogenous provirus" of the organism. Endogenous viruses usually are not expressed, but sometimes they are activated by external events, such as infection with another virus.

17.6 Retroviruses may transduce cellular sequences

Key Concepts

- Transforming retroviruses are generated by a recombination event in which a cellular RNA sequence replaces part of the retroviral RNA.

An interesting light on the viral life cycle is cast by the occurrence of transducing viruses, variants that have acquired cellular sequences in the form illustrated in **Figure 17.10.** Part of the viral sequence has been replaced by the *v-onc* gene. Protein synthesis generates a Gag-v-Onc protein instead of the usual Gag, Pol, and Env proteins. The resulting virus is replication-defective; it cannot sustain an infective cycle by itself. However, it can be perpetuated in the company of a helper virus that provides the missing viral functions.

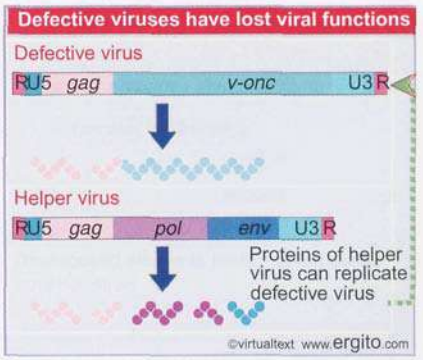


Figure 17.10 Replication-defective transforming viruses have a cellular sequence substituted for part of the viral sequence. The defective virus may replicate with the assistance of a helper virus that carries the wild-type functions.

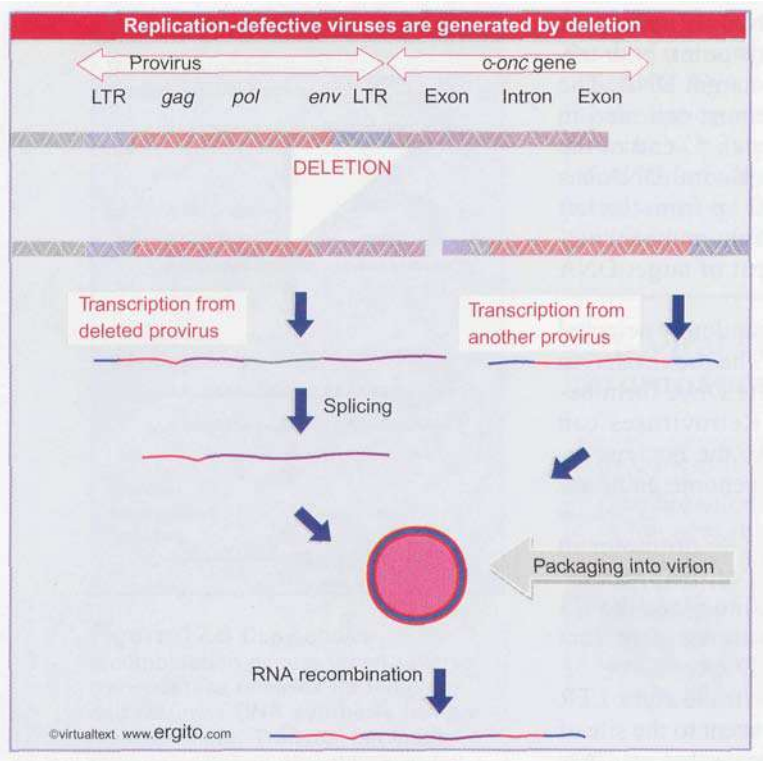


Figure 17.11 Replication-defective viruses may be generated through integration and deletion of a viral genome to generate a fused viral-cellular transcript that is packaged with a normal RNA genome. Nonhomologous recombination is necessary to generate the replication-defective transforming genome.

Onc is an abbreviation for oncogenesis, the ability to *transform* cultured cells so that the usual regulation of growth is released to allow unrestricted division. Both viral and cellular *one* genes may be responsible for creating tumorigenic cells (see *30.7 Retroviral oncogenes have cellular counterparts).*

A *v-onc* gene confers upon a virus the ability to transform a certain type of host cell. Loci with homologous sequences found in the host genome are called *c-onc* genes. How are the *onc* genes acquired by the retroviruses? A revealing feature is the discrepancy in the structures of *c-onc* and *v-onc* genes. The *c-onc* genes usually are interrupted by introns, but the *v-onc* genes are uninterrupted. This suggests that the *v-onc* genes originate from spliced RNA copies of the *c-onc* genes.

A model for the formation of transforming viruses is illustrated in **Figure 17.11**. A retrovirus has integrated near a *c-onc* gene. A deletion occurs to fuse the provirus to the *c-onc* gene; then transcription generates a joint RNA, containing viral sequences at one end and cellular *onc* sequences at the other end. Splicing removes the introns in both the viral and cellular parts of the RNA. The RNA has the appropriate signals for packaging into the virion; virions will be generated if the cell also contains another, intact copy of the provirus. Then some of the diploid virus particles may contain one fused RNA and one viral RNA.

A recombination between these sequences could generate the transforming genome, in which the viral repeats are present at both ends. (Recombination occurs at a high frequency during the retroviral infective cycle, by various means. We do not know anything about its demands for homology in the substrates, but we assume that the nonhomologous reaction between a viral genome and the cellular part of the fused RNA proceeds by the same mechanisms responsible for viral recombination.)

The common features of the entire retroviral class suggest that it may be derived from a single ancestor. Primordial IS elements could have surrounded a host gene for a nucleic acid polymerase; the resulting unit would have the form LTR-pol-LTR. It might evolve into an infectious virus by acquiring more sophisticated abilities to manipulate both DNA and RNA substrates, including the incorporation of genes whose products allowed packaging of the RNA. Other functions, such as transforming genes, might be incorporated later. (There is no reason to suppose that the mechanism involved in acquisition of cellular functions is unique for *onc* genes; but viruses carrying these genes may have a selective advantage because of their stimulatory effect on cell growth.)

17.7 Yeast Ty elements resemble retroviruses

Key Concepts

- Ty transposons have a similar organization to endogenous retroviruses.
- They are retroposons, with a reverse transcriptase activity, that transpose via an RNA intermediate.

By Book_Crazy [IND]

Ty elements comprise a family of dispersed repetitive DNA sequences that are found at different sites in different strains of yeast. Ty is an abbreviation for "transposon yeast." A transposition event creates a characteristic footprint: 5 bp of target DNA are repeated on either side of the inserted *Ty* element. Ty elements are retroposons that transpose by the same mechanism as retroviruses. The frequency of *Ty* transposition is lower than that of most bacterial transposons, ~10^{-7}–10^{-8}.

There is considerable divergence between individual *Ty* elements. Most elements fall into one of two major classes, called *Ty1* and *Ty917*. They have the same general organization illustrated in **Figure 17.12.** Each element is 6.3 kb long; the last 330 bp at each end constitute direct repeats, called δ. Individual *Ty* elements of each type have many changes from the prototype of their class, including base pair substitutions, insertions, and deletions. There are ~30 copies of the *Ty1* type and ~6 of the *Ty917* type in a typical yeast genome. In addition, there are ~100 independent *delta* elements, called solo δs.

The *delta* sequences also show considerable heterogeneity, although the two repeats of an individual *Ty* element are likely to be identical or at least very closely related. The *delta* sequences associated with *Ty* elements show greater conservation of sequence than the solo *delta* elements, which suggests that recognition of the repeats is involved in transposition.

The *Ty* element is transcribed into two poly(A)$^+$ RNA species, which constitute > 5% of the total mRNA of a haploid yeast cell. Both initiate within a promoter in the δ element at the left end. One terminates after 5 kb; the other terminates after 5.7 kb, within the delta sequence at the right end.

The sequence of the *Ty* element has two open reading frames, expressed in the same direction, but read in different phases and overlapping by 13 amino acids. The sequence of *TyA* suggests that it codes for a DNA-binding protein. The sequence of *TyB* contains regions that have homologies with reverse transcriptase, protease, and integrase sequences of retroviruses.

The organization and functions of *TyA* and *TyB* are analogous to the behavior of the retroviral *gag* and *pol* functions. The reading frames *TyA* and *TyB* are expressed in two forms. The TyA protein represents the *TyA* reading frame, and terminates at its end. The *TyB* reading frame, however, is expressed only as part of a joint protein, in which the TyA region is fused to the TyB region by a specific frameshift event that allows the termination codon to be bypassed (analogous to *gag-pol* translation in retroviruses).

Recombination between *Ty* elements seems to occur in bursts; when one event is detected, there is an increased probability of finding others. Gene conversion occurs between *Ty* elements at different locations, with the result that one element is "replaced" by the sequence of the other.

Ty elements can excise by homologous recombination between the directly repeated *delta* sequences. The large number of solo *delta* elements may be footprints of such events. An excision of this nature may be associated with reversion of a mutation caused by the insertion of *Ty;* the level of reversion may depend on the exact *delta* sequences left behind.

A paradox is that both *delta* elements have the same sequence, yet a promoter is active in the *delta* at one end and a terminator is active in the *delta* at the other end. (A similar feature is found in other transposable elements, including the retroviruses.)

Ty elements are classic retroposons, transposing through an RNA intermediate. An ingenious protocol used to detect this event is illustrated in **Figure 17.13.** An intron was inserted into an element to generate a

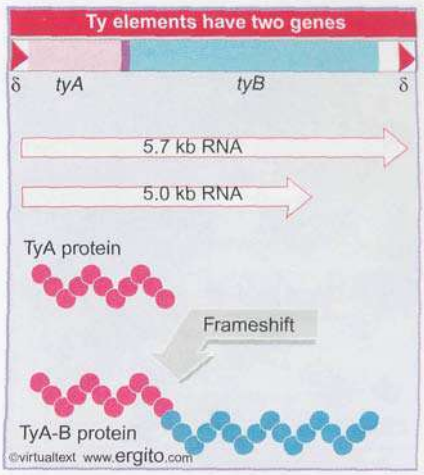


Figure 17.12 Ty elements terminate in short direct repeats and are transcribed into two overlapping RNAs. They have two reading frames, with sequences related to the retroviral *gag* and *pol* genes.

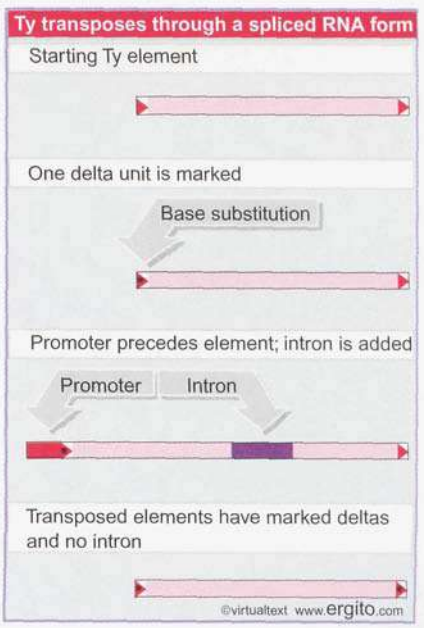


Figure 17.13 A unique Ty element, engineered to contain an intron, transposes to give copies that lack the intron. The copies possess identical terminal repeats, generated from one of the termini of the original Ty element.

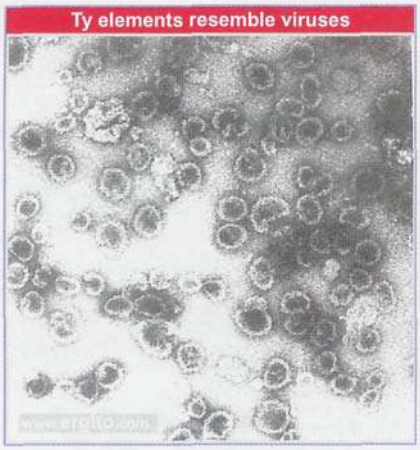


Figure 17.14 Ty elements generate virus-like particles. Photograph kindly provided by Alan Kingsman.

unique *Ty* sequence. This sequence was placed under the control of a *GAL* promoter on a plasmid and introduced into yeast cells. Transposition results in the appearance of multiple copies of the transposon in the yeast genome, but they all lack the intron.

We know of only one way to remove introns: RNA splicing. This suggests that transposition occurs by the same mechanism as with retroviruses. The *Ty* element is transcribed into an RNA that is recognized by the splicing apparatus. The spliced RNA is recognized by a reverse transcriptase and regenerates a duplex DNA copy.

The analogy with retroviruses extends further. The original *Ty* element has a difference in sequence between its two *delta* elements. But the transposed elements possess identical *delta* sequences, derived from the 5' delta of the original element. If we consider the *delta* sequence to be exactly like an LTR, consisting of the regions U3-R-U5, the *Ty* RNA extends from R region to R region. Just as shown for retroviruses in Figure 17.3, Figure 17.4, Figure 17.5, and Figure 17.6, the complete LTR is regenerated by adding a U5 to the 3' end and a U3 to the 5' end.

Transposition is controlled by genes within the *Ty* element. The *GAL* promoter used to control transcription of the marked *Ty* element is inducible; it is turned on by the addition of galactose. Induction of the promoter has two effects. It is necessary to activate transposition of the marked element. And its activation also increases the frequency of transposition of the other *Ty* elements on the yeast chromosome. This implies that the products of the *Ty* element can act in *trans* on other elements (actually on their RNAs).

Although the *Ty* element does not give rise to infectious particles, virus-like particles (VLPs) accumulate within the cells in which transposition has been induced. The particles can be seen in **Figure 17.14.** They contain full-length RNA, double-stranded DNA, reverse transcriptase activity, and a TyB product with integrase activity. The TyA product is cleaved like a *gag* precursor to produce the mature core proteins of the VLR This takes the analogy between the *Ty* transposon and the retrovirus even further. The *Ty* element behaves in short like a retrovirus that has lost its *env* gene and therefore cannot properly package its genome.

Only some of the *Ty* elements in any yeast genome are active: most have lost the ability to transpose (and are analogous to inert endogenous proviruses). Since these "dead" elements retain the 8 repeats, however, they provide targets for transposition in response to the proteins synthesized by an active element.

17.8 Many transposable elements reside in *D. melanogaster*

Key Concepts

- *copia* is a retroposon that is abundant in *D. melanogaster*.

The presence of transposable elements in *D. melanogaster* was first inferred from observations analogous to those that identified the first insertion sequences in *E. coli*. Unstable mutations are found that revert to wild type by deletion, or that generate deletions of the flanking material with an endpoint at the original site of mutation. They are caused by several types of transposable sequence, which are illustrated in **Figure 17.15.** They include the *copia* retroposon, the FB family, and

the P elements discussed previously in *16.14 The role of transposable elements in hybrid dysgenesis.*

The best-characterized family of retroposons is *copia.* Its name reflects the presence of a large number of closely related sequences that code for abundant mRNAs. The *copia* family is taken as a paradigm for several other types of elements whose sequences are unrelated, but whose structure and general behavior appear to be similar.

The number of copies of the *copia* element depends on the strain of fly; usually it is 20-60. The members of the family are widely dispersed. The locations of *copia* elements show a different (although overlapping) spectrum in each strain of *D. melanogaster.*

These differences have developed over evolutionary periods. Comparisons of strains that have diverged recently (over the past 40 years or so) as the result of their propagation in the laboratory reveal few changes. We cannot estimate the rate of change, but the nature of the underlying events is indicated by the result of growing cells in culture. The number of *copia* elements per genome then increases substantially, up to 2-3 times. The additional elements represent insertions of *copia* sequences at new sites. Adaptation to culture in some unknown way transiently increases the rate of transposition to a range of 10^{-3} to 10^{-4} events per generation.

The *copia* element is ~5000 bp long, with identical direct terminal repeats of 276 bp. Each of the direct repeats itself ends in related inverted repeats. A direct repeat of 5 bp of target DNA is generated at the site of insertion. The divergence between individual members of the *copia* family is slight, <5%; variants often contain small deletions. All of these features are common to the other *copia*-like families, although their individual members display greater divergence.

The identity of the two direct repeats of each *copia* element implies either that they interact to permit correction events, or that both are generated from one of the direct repeats of a progenitor element during transposition. As in the similar case of *Ty* elements, this is suggestive of a relationship with retroviruses.

The *copia* elements in the genome are always intact; individual copies of the terminal repeats have not been detected (although we would expect them to be generated if recombination deleted the intervening material). *copia* elements sometimes are found in the form of free circular DNA; like retroviral DNA circles, the longer form has two terminal repeats and the shorter form has only one. Particles containing *copia* RNA have been noticed.

The *copia* sequence contains a single long reading frame of 4227 bp. There are homologies between parts of the *copia* open reading frame and the *gag* and *pol* sequences of retroviruses. A notable absence from the homologies is any relationship with retroviral *env* sequences required for the envelope of the virus, which means that *copia* is unlikely to be able to generate virus-like particles.

Transcripts of *copia* are found as abundant poly(A)$^+$ mRNAs, representing both full-length and part-length transcripts. The mRNAs have a common 5' terminus, resulting from initiation in the middle of one of the terminal repeats. Several proteins are produced, probably involving events such as splicing of RNA and cleavage of polyproteins.

Although we lack direct evidence for *copia*'s mode of transposition, there are so many resemblances with retroviral organization that the conclusion seems inevitable that *copia* must have an origin related to the retroviruses. It is hard to say how many retroviral functions it possesses. We know, of course, that it transposes; but (as is the case with *Ty* elements) there is no evidence for any infectious capacity.

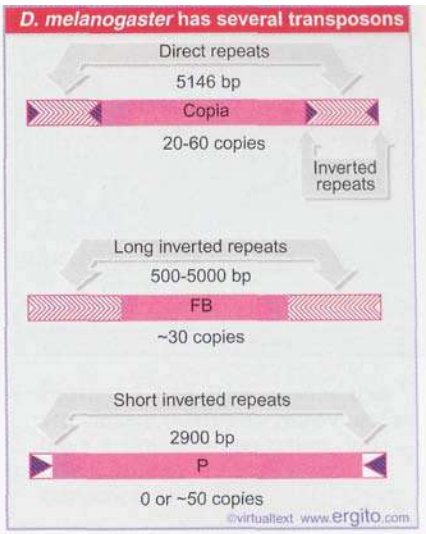


Figure 17.15 Three types of transposable element in *D. melanogaster* have different structures.

17.9 Retroposons fall into three classes

Key Concepts

- Retroposons of the viral superfamily are transposons that mobilize via an RNA that does not form an infectious particle.
- Some directly resemble retroviruses in their use of LTRs, but others do not have LTRs.
- Other elements can be found that were generated by an RNA-mediated transposition event, but do not themselves code for enzymes that can catalyze transposition.
- Transposons and retroposons constitute almost half of the human genome.

Retroposons are defined by their use of mechanisms for transposition that involve reverse transcription of RNA into DNA. Three classes of retroposons are distinguished in **Figure 17.16:**

- Members of the viral superfamily code for reverse transcriptase and/or integrase activities. Like other retroposons, they reproduce like retroviruses but differ from them in not passing through an independent infectious form. They are best characterized in the *Ty* and *copia* elements of yeast and flies.
- The LINES also have reverse transcriptase activity (and may therefore be considered to comprise more distant members of the viral superfamily), but they lack LTRs and use a different mechanism from retroviruses to prime the reverse transcription reaction. They are derived from RNA polymerase II transcripts. A minority of the elements in the genome are fully functional and can transpose autonomously; others have mutations and so can only transpose as the result of the action of a *trans-acting* autonomous element.
- Members of the nonviral superfamily are identified by external and internal features that suggest that they originated in RNA sequences, although in these cases we can only speculate on how a DNA copy was generated. We assume that they were targets for a transposition event by an enzyme system coded elsewhere, that is, they are always nonautonomous. They originated in cellular transcripts. They do not code for proteins that have transposition functions. The most prominent component of this family is called SINES. They are derived from RNA polymerase III transcripts.

Figure 17.17 shows the organization and sequence relationships of elements that code for reverse transcriptase. Like retroviruses, the LTR-containing retroposons can be classified into groups according to the

Mammalian genomes have three types of retroposons

	Viral Superfamily	LINES	Nonviral Superfamily
Common types	*Ty (S. cerevisiae)* *copia* (*D. melanogaster*)	L1 (human) B1, B2 ID, B4 (mouse)	SINES (mammals) Pseudogenes of pol III transcripts
Termini	Long terminal repeats	No repeats	No repeats
Target repeats	4-6 bp	7-21 bp	7-21 bp
Enzyme activities	Reverse transcriptase and/or integrase	Reverse transcriptase /endonuclease	None (or none coding for transposon products)
Organization	May contain introns (removed in subgenomic mRNA)	1 or 2 uninterrupted ORFs	No introns

Figure 17.16 Retroposons can be divided into the viral superfamilies that are retrovirus-like or LINES and the nonviral superfamilies that do not have coding functions.

By Book_Crazy [IND]

number of independent reading frames for *gag*, *pol*, and *int*, and the order of the genes. In spite of these superficial differences of organization, the common feature is the presence of reverse transcriptase and integrase activities. Typical mammalian LINES elements have two reading frames, one coding for a nucleic acid-binding protein, the other for reverse transcriptase and endonuclease activity.

LTR-containing elements can vary from integrated retroviruses to retroposons that have lost the capacity to generate infectious particles. Yeast and fly genomes have the *Ty* and *copia* elements that cannot generate infectious particles. Mammalian genomes have endogenous retroviruses that, when active, can generate infectious particles. The mouse genome has several active endogenous retroviruses which are able to generate particles that propagate horizontal infections. By contrast, almost all endogenous retroviruses lost their activity some 50 million years ago in the human lineage, and the genome now has mostly inactive remnants of the endogenous retroviruses.

LINES and SINES comprise a major part of the animal genome. They were defined originally by the existence of a large number of relatively short sequences that are related to one another (comprising the moderately repetitive DNA described in *3.6 Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences).* The LINES comprise long interspersed sequences, and the SINES comprise short interspersed sequences. (They are described as interspersed sequences or interspersed repeats because of their common occurrence and widespread distribution.)

Plants contain another type of small mobile element, called MITE (for miniature inverted-repeat transposable element). Such elements terminate in inverted repeats, have a 2 or 3 bp target sequence, do not have coding sequences, and are 200-500 bp long. At least 9 such families exist in (for example) the rice genome. They are often found in the regions flanking protein-coding genes. They have no relationship to SINES or LINES.

LINES and SINES comprise a significant part of the repetitive DNA of animal genomes. In many higher eukaryotic genomes, they occupy ~50% of the total DNA. **Figure 17.18** summarizes the distribution of the different types of transposons that constitute almost half of the human genome. Except for the SINES, which are always nonfunctional, the other types of elements all consist of both functional elements and elements that have suffered deletions eliminating parts of the reading frames that code for the protein(s) needed for transposition. The relative proportions of these types of transposons are generally similar in the mouse genome.

A common LINES in mammalian genomes is called L1. The typical member is ~6,500 bp long, terminating in an A-rich tract. The two open reading frames of a full-length element are called ORF1 and ORF2. The number of full-length elements is usually small (~50), and the remainder of the copies are truncated. Transcripts can be found. As implied by its presence in repetitive DNA, the LINES family shows sequence variation among individual members. However, the members of the family

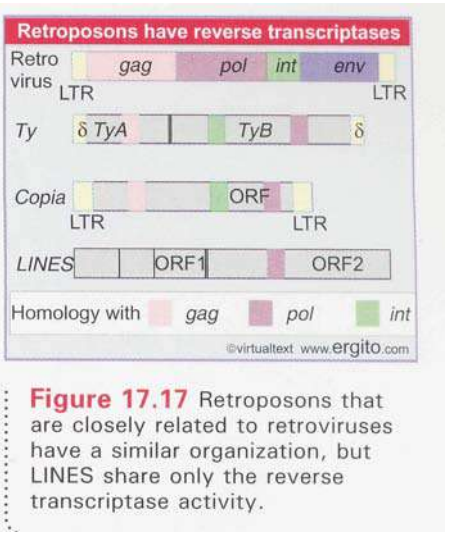


Figure 17.17 Retroposons that are closely related to retroviruses have a similar organization, but LINES share only the reverse transcriptase activity.

Retroviruses and transposons constitute half the human genome

Element	Organization	Length (Kb)	Human genome Number	Fraction
Retrovirus/retroposon	LTR *gag pol (env)* LTR	1-11	450,000	8%
LINES (autonomous) e.g. L1	*ORF1 (pol)* $(A)_n$	6-8	850,000	17%
SINES (nonautonomous) e.g. Alu	$(A)_n$	<0.3	1,500,000	15%
DNA transposon	Transposase	2-3	300,000	3%

©virtualtext www.ergito.com

Figure 17.18 Four types of transposable elements constitute almost half of the human genome.

By Book_Crazy [IND]

within a species are relatively homogeneous compared to the variation shown between species. L1 is the only member of the LINES family that has been active in either the mouse or human lineages, and it seems to have remained highly active in the mouse, but has declined in the human lineage.

Only one SINES has been active in the human lineage; this is the common Alu element. The mouse genome has a counterpart to this element (B1), and also other SINES (B2, ID, B4) that have been active. Human Alu and mouse B1 SINES are probably derived from the 7SL RNA (see next section). The other mouse SINES appear to have originated from reverse transcripts of tRNAs. The transposition of the SINES probably results from their recognition as substrates by an active L1 element.

17.10 The Alu family has many widely dispersed members

Key Concepts

- A major part of repetitive DNA in mammalian genomes consists of repeats of a single family organized like transposons and derived from RNA polymerase III transcripts.

The most prominent SINES comprises members of a single family. Its short length and high degree of repetition make it comparable to simple sequence (satellite) DNA, except that the individual members of the family are dispersed around the genome instead of being confined to tandem clusters. Again there is significant similarity between the members within a species compared with variation between species.

In the human genome, a large part of the moderately repetitive DNA exists as sequences of ~300 bp that are interspersed with nonrepetitive DNA. At least half of the renatured duplex material is cleaved by the restriction enzyme AluI at a single site, located 170 bp along the sequence. The cleaved sequences all are members of a single family, known as the Alu family after the means of its identification. There are ~300,000 members in the haploid genome (equivalent to one member per 6 kb of DNA). The individual Alu sequences are widely dispersed. A related sequence family is present in the mouse (where the 50,000 members are called the B1 family), in the Chinese hamster (where it is called the Alu-equivalent family), and in other mammals.

The individual members of the Alu family are related rather than identical. The human family seems to have originated by a 130 bp tandem duplication, with an unrelated sequence of 31 bp inserted in the right half of the dimer. The two repeats are sometimes called the "left half" and "right half" of the Alu sequence. The individual members of the Alu family have an average identity with the consensus sequence of 87%. The mouse Bl repeating unit is 130 bp long, corresponding to a monomer of the human unit. It has 70-80% homology with the human sequence.

The Alu sequence is related to 7SL RNA, a component of the signal recognition particle (see *8.10 The SRP interacts with the SRP receptor).* The 7SL RNA corresponds to the left half of an Alu sequence with an insertion in the middle. So the 90 5' terminal bases of 7SL RNA are homologous to the left end of Alu, the central 160 bases of 7SL RNA have no homology to Alu, and the 40 3' terminal bases of 7SL RNA are homologous to the right end of Alu. The 7SL RNA is coded by genes that are actively transcribed by RNA polymerase III. It is possible that these genes (or genes related to them) gave rise to the inactive Alu sequences.

The members of the Alu family resemble transposons in being flanked by short direct repeats. However, they display the curious feature that the lengths of the repeats are different for individual members of the family. Because they derive from RNA polymerase III transcripts, it is possible that individual members carry internal active promoters.

A variety of properties have been found for the Alu family, and its ubiquity has prompted many suggestions for its function, but it is not yet possible to discern its true role.

At least some members of the family can be transcribed into independent RNAs. In the Chinese hamster, some (although not all) members of the Alu-equivalent family appear to be transcribed *in vivo*. Transcription units of this sort are found in the vicinity of other transcription units.

Members of the Alu family may be included within structural gene transcription units, as seen by their presence in long nuclear RNA. The presence of multiple copies of the Alu sequence in a single nuclear molecule can generate secondary structure. In fact, the presence of Alu family members in the form of inverted repeats is responsible for most of the secondary structure found in mammalian nuclear RNA.

17.11 Processed pseudogenes originated as substrates for transposition

Key Concepts

- A processed pseudogene is derived from an mRNA sequence by reverse transcription.

When a sequence generated by reverse transcription of an mRNA is inserted into the genome, we can recognize its relationship to the gene from which the mRNA was transcribed. Such a sequence is called a processed pseudogene to reflect the fact that it was processed from RNA and is not active. The characteristic features of a processed pseudogene are compared in **Figure 17.19** with the features of the original gene and the mRNA. The figure shows all the relevant diagnostic features, only some of which are found in any individual example. Any transcript of RNA polymerase II could in principle give rise to such a pseudogene, and there are many examples, including the processed globin pseudogenes that were the first to be discovered (see *4.6 Pseudogenes are dead ends of evolution).*

The pseudogene may start at the point equivalent to the 5' terminus of the RNA, which would be expected only if the DNA had originated from the RNA. Several pseudogenes consist of precisely joined exon sequences; we know of no mechanism to recognize introns in DNA, so this feature argues for an RNA-mediated stage. The pseudogene may end in a short stretch of A·T base pairs, presumably derived from the poly(A) tail of the RNA. On either side of the pseudogene is a short direct repeat, presumed to have been generated by a transposition-like event. Processed pseudogenes reside at locations unrelated to their presumed sites of origin.

The processed pseudogenes do not carry any information that might be used to sponsor a transposition event (or to carry out the preceding

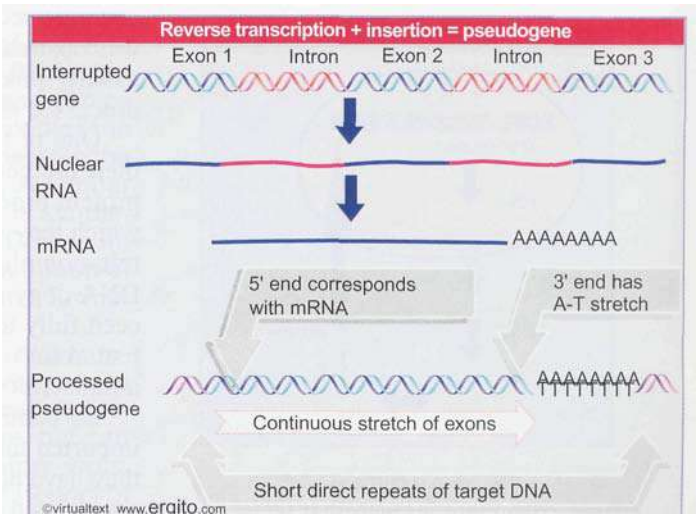


Figure 17.19 Pseudogenes could arise by reverse transcription of RNA to give duplex DNAs that become integrated into the genome.

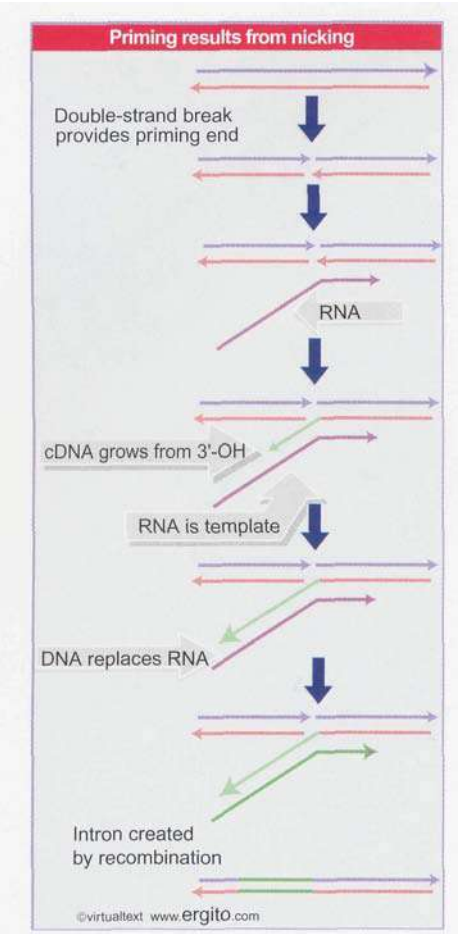


Figure 17.20 Retrotransposition of non-LTR elements occurs by nicking the target to provide a primer for cDNA synthesis on an RNA template. The arrowheads indicate 3′ ends.

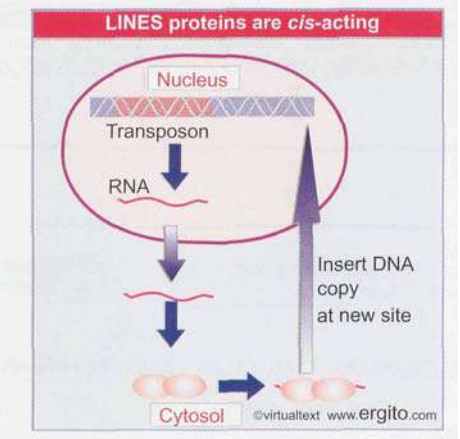


Figure 17.21 A LINES is transcribed into an RNA that is translated into proteins that assemble into a complex with the RNA. The complex translocates to the nucleus, where it inserts a DNA copy into the genome.

reverse transcription of the RNA). This suggests that the RNA was a substrate for another system, coded by a retroposon. In fact, it seems likely that the active LINES elements provide most of the reverse transcriptase activity, and they are responsible not only for their own transposition, but also for acting on the SINES and for generating processed pseudogenes.

17.12 LINES use an endonuclease to generate a priming end

Key Concepts

- LINES do not have LTRs and require the retroposon to code for an endonuclease that generates a nick to prime reverse transcription.

LINES elements, and some others, do not terminate in the LTRs that are typical of retroviral elements. This poses the question: how is reverse transcription primed? It does not involve the typical reaction in which a tRNA primer pairs with the LTR (see Figure 17.6). The open reading frames in these elements lack many of the retroviral functions, such as protease or integrase domains, but typically have reverse transcriptase-like sequences and code for an endonuclease activity. In the human LINES L1, ORF1 is a DNA-binding protein and ORF2 has both reverse transcriptase and endonuclease activities; both products are required for transposition.

Figure 17.20 shows how these activities support transposition. A nick is made in the DNA target site by an endonuclease activity coded by the retroposon. The RNA product of the element associates with the protein bound at the nick. The nick provides a 3'-OH end that primes synthesis of cDNA on the RNA template. A second cleavage event is required to open the other strand of DNA, and the RNA/DNA hybrid is linked to the other end of the gap either at this stage or after it has been converted into a DNA duplex. A similar mechanism is used by some mobile introns (see Figure 25.11).

When elements originate from RNA polymerase II transcripts, the genomic sequences are necessarily inactive: they lack the promoter that was upstream of the original startpoint for transcription. Because they usually possess the features of the mature transcript, they are called processed pseudogenes.

One of the reasons why LINES elements are so effective lies with their method of propagation. When a LINES mRNA is translated, the protein products show a *cis*-preference for binding to the mRNA from which they were translated. **Figure 17.21** shows that the ribonucleoprotein complex then moves to the nucleus, where the proteins insert a DNA copy into the genome. Often reverse transcription does not proceed fully to the end, so the copy is inactive. However, there is the potential for insertion of an active copy, because the proteins are acting on a transcript of the original active element.

By contrast, the proteins produced by the DNA transposons must be imported into the nucleus after being synthesized in the cytoplasm, but they have no means of distinguishing full-length transposons from inactive deleted transposons. **Figure 17.22** shows that instead, they will indiscriminately recognize any element by virtue of the repeats that mark the ends, much reducing their chance of acting on a full-length as opposed to deleted element. The consequence is that inactive elements accumulate and eventually the family dies out because a transposase has such a small chance of finding a target that is a fully functional transposon.

By Book_Crazy [IND]

Are transposition events currently occurring in these genomes or are we seeing only the footprints of ancient systems? This varies with the species. There are few currently active transposons in the human genome, but by contrast several active transposons are known in the mouse genome. This explains the fact that spontaneous mutations caused by LINES insertions occur at a rate of ~3% in mouse, but only 0.1% in man. There appear to be ~ 10-50 active LINES elements in the human genome. Some human diseases can be pinpointed as the result of transposition of L1 into genes, and others result from unequal crossing-over events involving repeated copies of L1. A model system in which LINES transposition occurs in tissue culture cells suggests that a transposition event can introduce several types of collateral damage as well as inserting into a new site; the damage includes chromosomal rearrangements and deletions. Such events may be viewed as agents of genetic change. Neither DNA transposons nor retroviral-like retroposons seem to have been active in the human genome for 40-50 million years, but several active examples of both are found in the mouse.

Note that for transpositions to survive, they must occur in the germline. Presumably similar events occur in somatic cells, but do not survive beyond one generation.

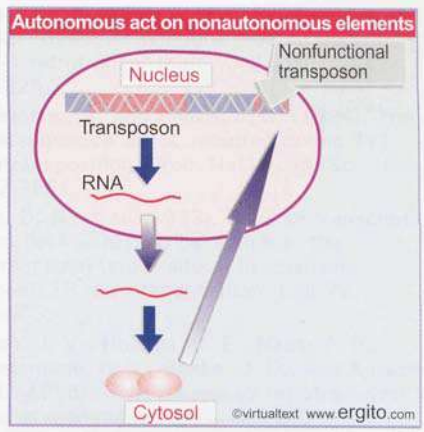


Figure 17.22 A transposon is transcribed into an RNA that is translated into proteins that move independently to the nucleus, where they act on any pair of inverted repeats with the same sequence as the original transposon.

17.13 Summary

Reverse transcription is the unifying mechanism for reproduction of retroviruses and perpetuation of retroposons. The cycle of each type of element is in principle similar, although retroviruses are usually regarded from the perspective of the free viral (RNA) form, while retroposons are regarded from the stance of the genomic (duplex DNA) form.

Retroviruses have genomes of single-stranded RNA that are replicated through a double-stranded DNA intermediate. An individual retrovirus contains two copies of its genome. The genome contains the *gag, pol,* and *env* genes, which are translated into polyproteins, each of which is cleaved into smaller functional proteins. The Gag and Env components are concerned with packing RNA and generating the virion; the Pol components are concerned with nucleic acid synthesis.

Reverse transcriptase is the major component of Pol, and is responsible for synthesizing a DNA (minus strand) copy of the viral (plus strand) RNA. The DNA product is longer than the RNA template; by switching template strands, reverse transcriptase copies the 3' sequence of the RNA to the 5' end of the DNA, and copies the 5' sequence of the RNA to the 3' end of the DNA. This generates the characteristic LTRs (long terminal repeats) of the DNA. A similar switch of templates occurs when the plus strand of DNA is synthesized using the minus strand as template. Linear duplex DNA is inserted into a host genome by the integrase enzyme. Transcription of the integrated DNA from a promoter in the left LTR generates further copies of the RNA sequence.

Switches in template during nucleic acid synthesis allow recombination to occur by copy choice. During an infective cycle, a retrovirus may exchange part of its usual sequence for a cellular sequence; the resulting virus is usually replication-defective, but can be perpetuated in the course of a joint infection with a helper virus. Many of the defective viruses have gained an RNA version *(v-onc)* of a cellular gene *(c-onc).* The *onc* sequence may be any one of a number of genes whose expression in *v-onc* form causes the cell to be transformed into a tumorigenic phenotype.

The integration event generates direct target repeats (like transposons that mobilize via DNA). An inserted provirus therefore has direct terminal repeats of the LTRs, flanked by short repeats of target

DNA. Mammalian and avian genomes have endogenous (inactive) proviruses with such structures. Other elements with this organization have been found in a variety of genomes, most notably in *S. cerevisiae* and *D. melanogaster*. *Ty* elements of yeast and *copia* elements of flies have coding sequences with homology to reverse transcriptase, and mobilize via an RNA form. They may generate particles resembling viruses, but do not have infectious capability. The LINES sequences of mammalian genomes are further removed from the retroviruses, but retain enough similarities to suggest a common origin. They use a different type of priming event to initiate reverse transcription, in which an endonuclease activity associated with the reverse transcriptase makes a nick that provides a 3′-OH end for priming synthesis on an RNA template. The frequency of LINES transposition is increased because its protein products are *cis*-acting; they associate with the mRNA from which they were translated to form a ribonucleoprotein complex that is transported into the nucleus.

Another class of retroposons have the hallmarks of transposition via RNA, but have no coding sequences (or at least none resembling retroviral functions). They may have originated as passengers in a retroviral-like transposition event, in which an RNA was a target for a reverse transcriptase. Processed pseudogenes arise by such events. A particularly prominent family apparently originating from a processing event is the mammalian SINES, including the human Alu family. Some snRNAs, including 7SL snRNA (a component of the SRP) are related to this family.

References

17.2 The retrovirus life cycle involves transposition-like events

rev Varmus, H. and Brown, P. (1989). Retroviruses. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 3-108.

ref Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumor viruses. Nature 226, 1209-1211.

Temin, H. M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226, 121 1-1213.

17.4 Viral DNA is generated by reverse transcription

rev Katz, R. A. and Skalka, A. M. (1994). The retroviral enzymes. Ann. Rev. Biochem. 63, 133-173.
Lai, M. M. C. (1992). RNA recombination in animal and plant viruses. Microbiol. Rev. 56, 61-79.
Hu, W. S. and Temin, H. M. (1990). Retroviral recombination and reverse transcription. Science 250, 1227-1233.

Negroni, M. and Buc, H. (2001). Mechanisms of retroviral recombination. Ann. Rev. Genet. 35, 275-302.

ref Negroni, M. and Buc, H. (2000). Copy-choice recombination by reverse transcriptases: reshuffling of genetic markers mediated by RNA chaperones. Proc. Nat. Acad. Sci. USA 97, 6385-6390.

17.5 Viral DNA integrates into the chromosome

rev Goff, S. P. (1992). Genetics of retroviral integration. Ann. Rev. Genet. 26, 527-544.

ref Craigie, R., Fujiwara, T., and Bushman, F. (1990). The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro*. Cell 62, 829-837.

17.7 Yeast Ty elements resemble retroviruses

ref Boeke, J. D. et al. (1985). Ty elements transpose through an RNA intermediate. Cell 40, 491-500.

17.8 Many transposable elements reside in *D. melanogaster*

ref Mount, S. M. and Rubin, G. M. (1985). Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol. Cell Biol. 5, 1630-1638.

17.9 Retroposons fall into three classes

rev Deininger, P. L. (1989). SINEs: short interspersed repeated DNA elements in higher eukaryotes. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 19-636.

Hutchison, C. A. et al. (1989). LINEs and related retroposons: long interspersed repeated sequences in the eukaryotic genome. In Mobile DNA, Eds. Berg, D. E. and Howe, M. American Society of Microbiology, Washington DC 93-617.
Ostertag, E. M. and Kazazian, H. H. (2001). Biology of mammalian L1 retrotransposons. Ann. Rev. Genet. 35, 501-538.

Weiner, A. M., Deininger, P. L., and Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Ann. Rev. Biochem. 55, 631-661.

ref Waterston et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Bureau, T. E. and Wessler, S. R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4, 1283-1294.

Bureau, T. E., Ronald, P. C, and Wessler, S. R. (1996). A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Nat. Acad. Sci. USA 93, 8524-8529.

Loeb, D. D. et al. (1986). The sequence of a large L1Md element reveals a tandemly repeated 5' end and several features found in retrotransposons. Mol. Cell Biol. 6, 168-182.

Sachidanandam, R. et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. The International SNP Map Working Group. Nature 409, 928-933.

17.12 LINES use an endonuclease to generate a priming end

rev Ostertag, E. M. and Kazazian, H. H. (2001). Biology of mammalian L1 retrotransposons. Ann. Rev. Genet. 35, 501-538.

ref Feng, Q., Moran, J. V., Kazazian, H. H., and Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905-916.

Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell 110, 315-325.

Lauermann, V. and Boeke, J. D. (1994). The primer tRNA sequence is not inherited during Ty1 retrotransposition. Proc. Nat. Acad. Sci. USA 91, 9847-9851.

Luan, D. D. et al. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72, 595-605.

Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., and Kazazian, H. H. (1996). High frequency retrotransposition in cultured mammalian cells. Cell 87, 917-927.

Symer, D. E., Connelly, C, Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G., and Boeke, J. D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110, 327-338.

Chapter 18

Rearrangement of DNA

18.1 Introduction

Although genomic DNA is usually unaltered by somatic development, there are some cases in which sequences are moved within a genome, modified, amplified, or even lost, as a natural event. In this chapter, we discuss a variety of such events in yeast, plants, and lower eukaryotes. Examples of rearrangement or loss of specific sequences are especially common in the lower eukaryotes. Usually these changes involve somatic cells; the germline remains inviolate. (However, there are organisms whose reproductive cycle involves the loss of whole chromosomes or sets of chromosomes.) We also discuss the introduction of new sequences into the genome by experimental means. Reorganization of particular sequences is rare in animals, although an extensive case is represented by the immune system. In *26 Immune diversity*, we discuss the rearrangement and expression of immunoglobulin genes.

There are two types of circumstances in which gene rearrangement is used to control expression:

- Rearrangement may create new genes, needed for expression in particular circumstances, as in the case of the immunoglobulins.
- Rearrangement may be responsible for switching expression from one pre-existing gene to another. This provides a mechanism for regulating gene expression.

Yeast mating type switching and trypanosome antigen variation share a similar type of plan in which gene expression is controlled by manipulation of DNA sequences. Phenotype is determined by the gene copy present at a particular, active locus. But the genome also contains a store of other, alternative sequences, which are silent. A silent copy can be activated only by a rearrangement of sequences in which it replaces the active gene copy. Such a substitution is equivalent to a unidirectional transposition with a specific target site.

The simplest example of this strategy is found in the yeast, *S. cerevisiae*. Haploid *S. cerevisiae* can have either of two mating types. The type is determined by the sequence present at the active mating type locus. But the genome also contains two other, silent loci, one representing each mating type. Transition between mating types is accomplished

by substituting the sequence at the active locus with the sequence from the silent locus carrying the other mating type.

A range of variations is made possible by DNA rearrangement in the African trypanosomes, unicellular parasites that evade the host immune response by varying their surface antigens. The type of surface antigen is determined by the gene sequence at an active locus. This sequence can be changed, however, by substituting a sequence from any one of many silent loci. It seems fitting that the mechanism used to combat the flexibility of the immune apparatus is analogous to that used to generate immune diversity: it relies on physical rearrangements in the genome to change the sequences that are expressed.

Another means of increasing genetic capacity is employed in parasite- or symbiote-host interactions, in which exogenous DNA is introduced from a bacterium into a host cell. The mechanism resembles bacterial conjugation. Expression of the bacterial DNA in its new host changes the phenotype of the cell. In the example of the bacterium *Agrobacterium tumefaciens*, the result is to induce tumor formation by an infected plant cell.

Alterations in the relative proportions of components of the genome during somatic development occur to allow insect larvae to increase the number of copies of certain genes. And the occasional amplification of genes in cultured mammalian cells is indicated by our ability to select variant cells with an increased copy number of some gene. Initiated within the genome, the amplification event can create additional copies of the gene that survive in either intrachromosomal or extrachromosomal form.

When extraneous DNA is introduced into eukaryotic cells, it may give rise to extrachromosomal forms or may be integrated into the genome. The relationship between the extrachromosomal and genomic forms is irregular, depending on chance and to some degree unpredictable events, rather than resembling the regular interchange between free and integrated forms of bacterial plasmids.

Yet, however accomplished, the process may lead to stable change in the genome; following its injection into animal eggs, DNA may even be incorporated into the genome and inherited thereafter as a normal component, sometimes continuing to function. Injected DNA may enter the germline as well as the soma, creating a transgenic animal. The ability to introduce specific genes that function in an appropriate manner could become a major medical technique for curing genetic diseases.

The converse of the introduction of new genes is the ability to disrupt specific endogenous genes. Additional DNA can be introduced within a gene to prevent its expression and to generate a null allele. Breeding from an animal with a null allele can generate a homozygous "knockout," which has no active copy of the gene. This is a powerful method to investigate directly the importance and function of a gene.

Considerable manipulation of DNA sequences therefore is achieved both in authentic situations and by experimental fiat. We are only just beginning to work out the mechanisms that permit the cell to respond to selective pressure by changing its bank of sequences or that allow it to accommodate the intrusion of additional sequences.

18.2 The mating pathway is triggered by pheromone-receptor interactions

- **Key Concepts**

  Yeast of a given mating type secrete a small polypeptide that binds to a receptor on cells of the opposite mating type.

The yeast *S. cerevisiae* can propagate in either the haploid or diploid condition. Conversion between these states takes place by mating (fusion of haploid spores to give a diploid) and by sporulation (meiosis of diploids to give haploid spores). The ability to engage in these activities is determined by the mating type of the strain.

The properties of the two mating types are summarized in **Figure 18.1.** We may view them as resting on the rationale that there is no point in mating unless the haploids are of different genetic types; and sporulation is productive only when the diploid is heterozygous and therefore able to generate recombinants.

The mating type of a (haploid) cell is determined by the genetic information present at the MAT locus. Cells that carry the *MATa* allele at this locus are type a; likewise, cells that carry the MATα allele are type α. Cells of opposite types can mate; cells of the same type cannot.

Recognition of cells of opposite mating type is accomplished by the secretion of pheromones. a cells secrete the small polypeptide α-factor; *a* cells secrete a-factor. The α-factor is a peptide of 13 amino acids; the a-factor is a peptide of 12 amino acids that is modified by addition of a farnesyl (lipid-like) group and carboxymethylation. Each of these peptides is synthesized in the form of a precursor polypeptide that is cleaved to release the mature peptide sequence.

A cell of one mating type carries a surface receptor for the pheromone of the opposite type. When an a cell and an α cell encounter one another, their pheromones act on each other to arrest the cells in the G1 phase of the cell cycle, and various morphological changes occur. In a successful mating, the cell cycle arrest is followed by cell and nuclear fusion to produce an *a/α* diploid cell.

The *a/α* cell carries both the *MATa* and *MATα* alleles and has the ability to sporulate. **Figure 18.2** demonstrates how this design maintains the normal haploid/diploid life cycle. Note that only heterozygous diploids can sporulate; homozygous diploids (either *a/α* or a/a) cannot sporulate.

Much of the information about the yeast mating type pathway was deduced from the properties of mutations that eliminate the ability of a and/or α cells to mate. The genes identified by such mutations are called *STE* (for sterile). Mutations in the genes *STE2* and *STE3* are specific for individual mating types; but mutations in the other *STE* genes eliminate mating in both a and α cells. This situation is explained by the fact that the events that follow the interaction of factor with receptor are identical for both types.

Mating is a symmetrical process that is initiated by the interaction of pheromone secreted by one cell type with the receptor carried by the other cell type. The only genes that are uniquely required for the response pathway in a particular mating type are those coding for the receptors. Either the a factor-receptor or the α factor-receptor interaction switches on the same response pathway. Mutations that eliminate steps in the common pathway have the same effects in both cell types.

Haploids and diploids are different

	MATa	*MATα*	*MATa/MATα*
Cell type	a	α	a/α
Mating	yes	yes	no
Sporulation	no	no	yes
Pheromone	a factor	α factor	none
Receptor	binds α factor	binds a factor	none

Figure 18.1 Mating type controls several activities.

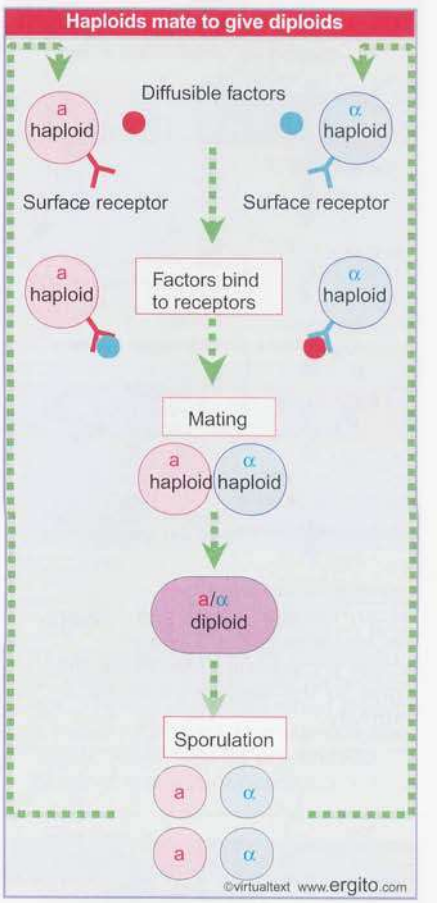


Figure 18.2 The yeast life cycle proceeds through mating of *MATa* and *MATα* haploids to give heterozygous diploids that sporulate to generate haploid spores.

18.3 The mating response activates a G protein

Key Concepts

- Binding of mating type factor to receptor activates a trimeric G protein.
- The βγ dimer is released and activates a signal transduction pathway.

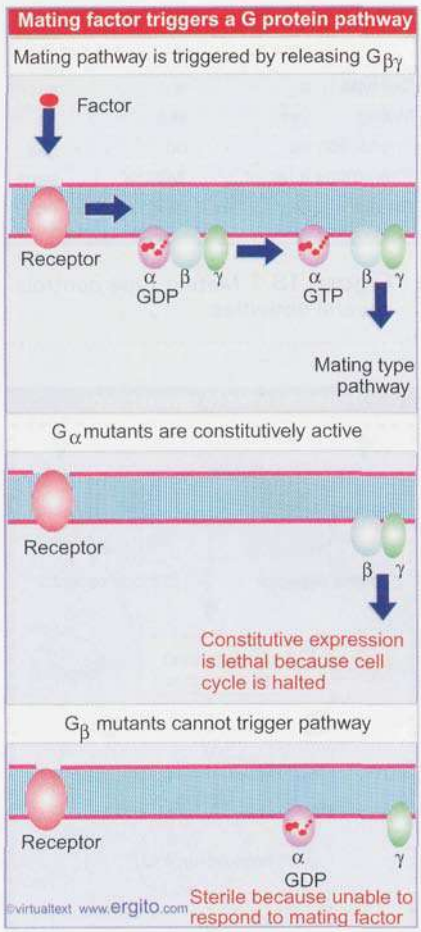


Figure 18.3 Either *a* or α factor/receptor interaction triggers the activation of a G protein, whose βγ subunits transduce the signal to the next stage in the pathway.

The initial steps in the mating-type response are summarized in **Figure 18.3.** The components are similar to those of the "classical" receptor-G protein coupled systems, in which a membrane receptor interacts with a trimeric G protein (see *28.5 G proteins may activate or inhibit target proteins).* Ste2 is the α-receptor in the a cell; Ste3 is the a-receptor in the a cell. When either receptor is activated, it interacts with the same G protein. *This means that the identical pathway is triggered when either type of mating factor interacts with a receptor of the opposite type.*

The trimeric G protein consists of the subunits, a, β, and 7. The a subunit binds a guanine nucleotide. In the intact (trimeric) G protein, the a subunit carries GDP. When the pheromone receptor is activated, it causes the GDP to be displaced by GTP. As a result, the a subunit is released from the βγ dimer. This separation of subunits allows the G protein to activate the next protein in the pathway.

The most common mechanism used in such pathways is for the activated a subunit to interact with the target protein. However, the situation is different in the mating type pathway, where the βγ dimer activates the next stage in the pathway. The component proteins of the G-trimer are identified by mutations in three genes, *SCG1, STE4,* and *STE18,* that affect the response to binding pheromone. Inactivation of *SCG1,* which codes for the G_α protein, causes constitutive expression of the pheromone response pathway (because G_α is unable to maintain $G_{\beta\gamma}$ in the inactive trimeric form). The mutation is lethal, because its effects include arrest of the cell cycle. Inactivation of *STE4* (codes for G_β) or of *STE18* (codes for G_α) create sterility by abolishing the mating-type response.

18.4 The signal is passed to a kinase cascade

Key Concepts

- $G_{\beta\gamma}$ activates the monomeric G protein Cdc42.
- Cdc42 directly controls the structure of the cytoskeleton and activates a kinase cascade.
- In the kinase cascade, the signal passes through a series of kinases.
- The last kinase in the cascade activates a transcription factor and also phosphorylates other targets.
- The effect of the pathway is to repress functions needed for the cell cycle and to activate functions needed for mating.

Figure 18.4 summarizes the main steps of the mating type pathway. (There are also some branches that are not shown in the figure.) When the $G_{\beta\gamma}$ dimer is released, it causes a monomeric G protein to be activated. Actually, the immediate target of $G_{\beta\gamma}$ is Cdc24, which is a nucleotide exchange factor that then activates the monomeric G protein Cdc42. The effect of Cdc24 on Cdc42 is a typical interaction in which the monomeric G protein is activated by replacing its bound GDP with GTP. Cdc42 resembles the Rho family of G proteins that control actin filament structures in higher eukaryotic cells (see *28.15 The activation of Ras is controlled by GTP).*

Cdc42 then activates two pathways:

- one affects the structure of the cell by changing the organization of the cytoskeleton;
- the other is a cascade of kinases that ultimately activates transcription.

Cdc42 acts on many proteins that are involved in modifying the state of assembly of actin filaments, and in this way changes the structure of the cell. It is necessary for budding, which generates the new daughter cell when division occurs. After mating, division stops and budding ceases, although cell growth continues in the direction of the pheromone. It is Cdc42 that is responsible for this change.

By Book_Crazy [IND]

The pathway to regulating transcription is identified by a group of *STE* genes. They form a kinase cascade in which each member activates the next in the pathway by phosphorylating it. This is an example of a MAP kinase cascade (see *28.16 A MAP kinase pathway is a cascade).* Analogous pathways are found in higher organisms, and are compared with the yeast cascade in Figure 28.38. In the case of the yeast pathway, Cdc42 directly activates the kinase STE20. STE20 initiates the kinase cascade by activating STE11, which activates STE7, which activates FUS3, which finally triggers the response by phosphorylating the transcription factor STE 12. This causes STE 12 to migrate to the nucleus, where it activates transcription.

Some components in the signaling pathways do not have catalytic activities but act to assist the other components. Both FAR1 and STE5 in the pathways detailed in Figure 18.4 are scaffold proteins that hold together the other members of the pathway. This may be necessary to ensure specificity in the response, for example, to ensure that each kinase in the MAPK cascade is directed to phosphorylate only the intended next member of the pathway (see *28.17 What determines specificity in signaling?).* Packaging the members of the pathway together also increases the efficiency of the response.

Both FAR1 and STE5 shuttle between the nucleus and cytoplasm. **Figure 18.5** shows that in a dividing cell, before the mating response is triggered, they are concentrated in the nucleus. **Figure 18.6** shows that they both interact with the activated $G_{\beta\gamma}$ dimer, and therefore become concentrated in the cytoplasm as a result of the pheromone-receptor interaction. FAR1 is bound to Cdc24; the result of its relocalization to the cytoplasm is to enable the Cdc24 to activate Cdc42. STE5 binds to the three components of the MAPK cascade; actually, its interaction with the kinases may enhance the ability of STE20 to make the initial phosphorylation of STE11. So we see that, although $G_{\beta\gamma}$ directly activates the effector pathways through its effect on the Cdc24/Cdc42 interaction, it indirectly enhances the pathways by causing the necessary components to accumulate in the cytoplasm.

Branches from the cascade generate additional reactions. STE7 activates two kinases. One is FUS3, whose action is dedicated to carrying forward the mating type pathway, as shown in Figure 18.4. FUS3 acts on earlier components of the pathway, specifically FAR1 and STE5, to enhance their actions, on the transcription factor STE 12, and also on proteins that would otherwise inhibit STE12. The second target for STE7 is KSS1, which has ancillary functions (its main function is in vegetative growth).

Some of the end targets for the cascade are direct substrates of one of the kinases; for example, FUS3 kinase acts on Cln3, which is one of 3 Cln proteins needed for cell cycle progression. Other targets are controlled at the level of gene expression.

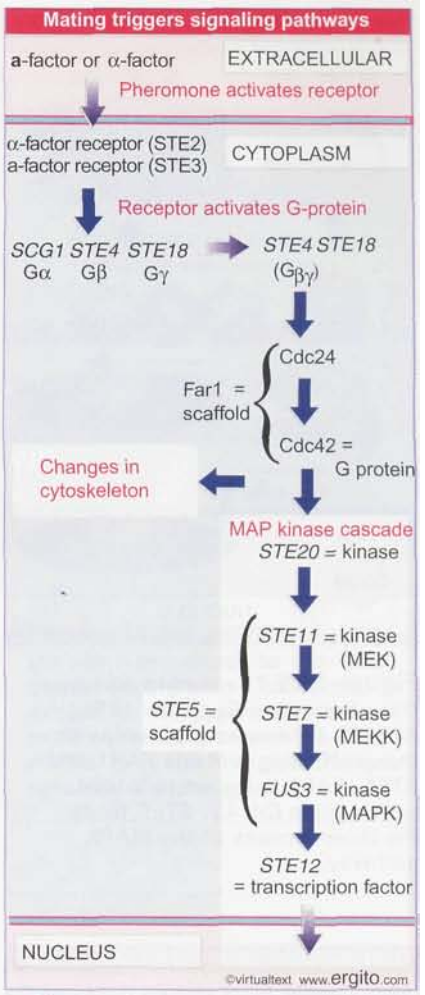


Figure 18.4 The same mating type response is triggered by interaction of either pheromone with its receptor. The signal is transmitted through a series of kinases to a transcription factor; there may be branches to some of the final functions.

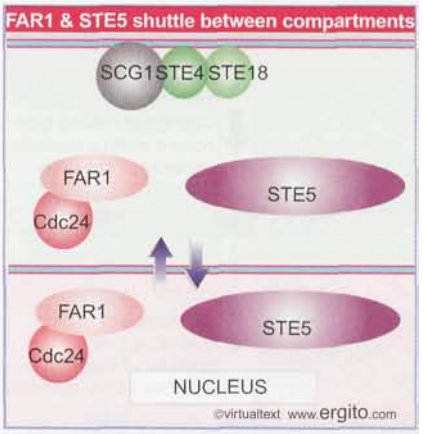


Figure 18.5 In a dividing cell, FAR1 (bound to Cdc24) and STE5 shuttle between the nucleus and cytoplasm. The $G_{\alpha\beta\gamma}$ complex of SCG1, STE4, STE18 is associated with the plasma membrane.

18.5 Yeast can switch silent and active loci for mating type

Key Concepts

- The yeast mating type locus MAT has either the *MATa* or *MATa* genotype.
- Yeast with the dominant allele *HO* switch their mating type at a frequency ~ 10^{-6}.
- The allele at *MAT* is called the active cassette.
- There are also two silent cassettes, *HMLa* and *HMRa*.
- Switching occurs if *MATa* is replaced by *HMLa* or *MATa* is replaced by *HMRa.*

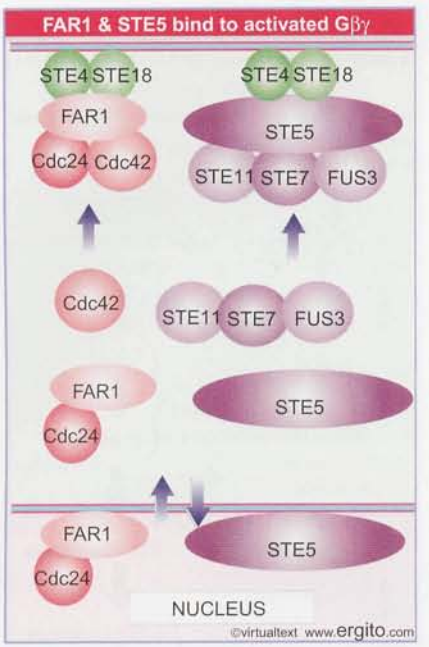


Figure 18.6 The mating pathway frees the active $G_{\beta\gamma}$ dimer (STE4-STE18). It forms complexes with the scaffolding proteins FAR1 and STE5. FAR1 is bound to Cdc24, which binds Cdc42. STE5 binds the three kinases of the MAPK pathway.

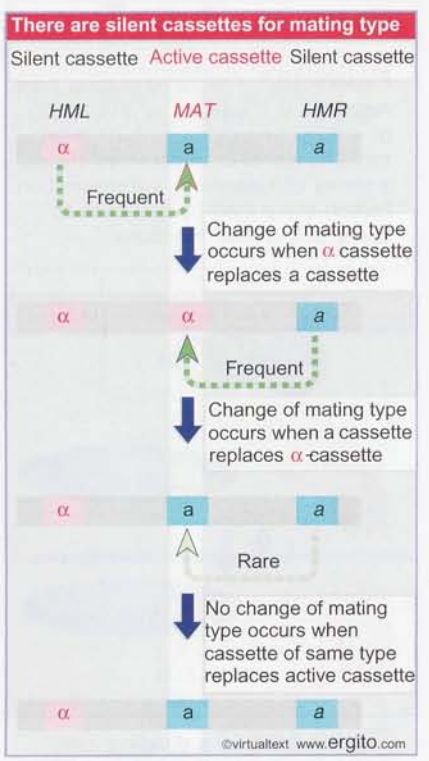


Figure 18.7 Changes of mating type occur when silent cassettes replace active cassettes of opposite genotype.

Yeast mating type is determined by the locus *MAT,* which can have either of the alleles *MATa* or *MATα*. Some yeast strains have the remarkable ability to switch their mating types. These strains carry a dominant allele *HO* and change their mating type frequently, as often as once every generation. Strains with the recessive allele *ho* have a stable mating type, subject to change with a frequency $\sim 10^{-6}$.

The presence of *HO* causes the genotype of a yeast population to change. Irrespective of the initial mating type, in a very few generations there are large numbers of cells of both mating types, leading to the formation of *MATa/'MATa* diploids that take over the population. The production of stable diploids from a haploid population can be viewed as the raison d'être for switching.

The existence of switching suggests that all cells contain the potential information needed to be either *MATa* or *MATa,* but express only one type. Where does the information to change mating types come from? Two additional loci are needed for switching. *HMLa* is needed for switching to give a *MATa* type; *HMRa* is needed for switching to give a *MATa* type. These loci lie on the same chromosome that carries *MAT*. *HML* is far to the left, *HMR* far to the right.

The cassette model for mating type is illustrated in Figure 18.7. It proposes that *MAT* has an *active cassette* of either type a or type *a*. *HML* and *HMR* have *silent cassettes.* Usually *HML* carries an a cassette, while *HMR* carries an *a* cassette. All cassettes carry information that codes for mating type, but only the active cassette at *MAT* is expressed. Mating-type switching occurs when the active cassette is replaced by information from a silent cassette. The newly installed cassette is then expressed.

Switching is nonreciprocal; the copy at *HML* or *HMR* replaces the allele at *MAT*. We know this because a mutation at *MAT* is lost permanently when it is replaced by switching—it does not exchange with the copy that replaces it.

If the silent copy present at *HML* or *HMR* is mutated, switching introduces a mutant allele into the *MAT* locus. The mutant copy at *HML* or *HMR* remains there through an indefinite number of switches. Like replicative transposition, the donor element generates a new copy at the recipient site, while itself remaining inviolate.

Mating-type switching is a directed event, in which there is only one recipient *(MAT),* but two potential donors *(HML* and *HMR)*. Switching usually involves replacement of *MATa* by the copy at *HMLa* or replacement *of MATa* by the copy at *HMRa*. In 80-90% of switches, the *MAT* allele is replaced by one of opposite type. This is determined by the phenotype of the cell. Cells of a phenotype preferentially choose *HML* as donor; cells of a phenotype preferentially choose *HMR*.

Several groups of genes are involved in establishing and switching mating type. As well as the genes that directly determine mating type, they include genes needed to repress the silent cassettes, to switch mating type, or to execute the functions involved in mating.

By comparing the sequences of the two silent cassettes *(HMLa* and *HMRa)* with the sequences of the two types of active cassette *(MATa* and *MATa),* we can delineate the sequences that determine mating type. The organization of the mating type loci is summarized in Figure 18.8. Each cassette contains common sequences that flank a central region that differs in the *a* and *a* types of cassette (called *Ya* or *Yα*). On either side of this region, the flanking sequences are virtually identical, although they are shorter at *HMR*. The active cassette at *MAT is* transcribed from a promoter within the *Y* region.

18.6 The *MAT* locus codes for regulator proteins

Key Concepts

- In α-type haploids, *MATα* turns on genes required to specify α-specific functions required for mating, and turns off genes required for a-mating type.
- In a-type cells, *MATa* is not required.
- In diploids, a1 and a2 products cooperate to repress haploid-specific genes.

The basic function of the *MAT* locus is to control expression of pheromone and receptor genes, and other functions involved in mating. *MATa* codes for two proteins, α1 and a2. *MATα* codes for a single protein, a1. The a and α proteins directly control transcription of various target genes; they function by both positive and negative regulation. They function independently in haploids, and in conjunction in diploids. Their interactions are summarized in the table on the right of Figure 18.9 in terms of three groups of target genes:

- a-specific genes are expressed constitutively in a cells. They are repressed in a cells. The a-specific genes include the a-factor structural gene, and *STE2,* which codes for the α-factor receptor. So the a phenotype is associated with readiness to recognize the pheromone produced by the opposite mating type.
- α-specific functions are induced in a cells, but are not expressed in a cells. They include the α-factor structural gene, and the a-factor receptor gene, *STE3.* Again, the expression of pheromone of one type is associated with expression of receptor for the pheromone of the opposite type.
- Haploid-specific functions include genes that are needed for transcription of pheromone and receptor genes, the *HO* gene involved in switching, and *RME,* a repressor of sporulation. They are expressed constitutively in both types of haploid, but are repressed in a/a diploids. As a result, the a-specific and α-specific functions also remain unexpressed in diploids.

We may now view the functions of the regulators and their targets from the perspective of the *MAT* functions expressed in haploid and

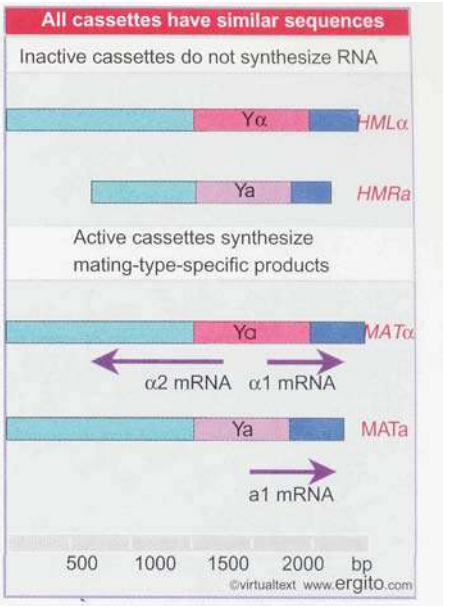


Figure 18.8 Silent cassettes have the same sequences as the corresponding active cassettes, except for the absence of the extreme flanking sequences in *HMRa*. Only the Y region changes between *a* and α types.

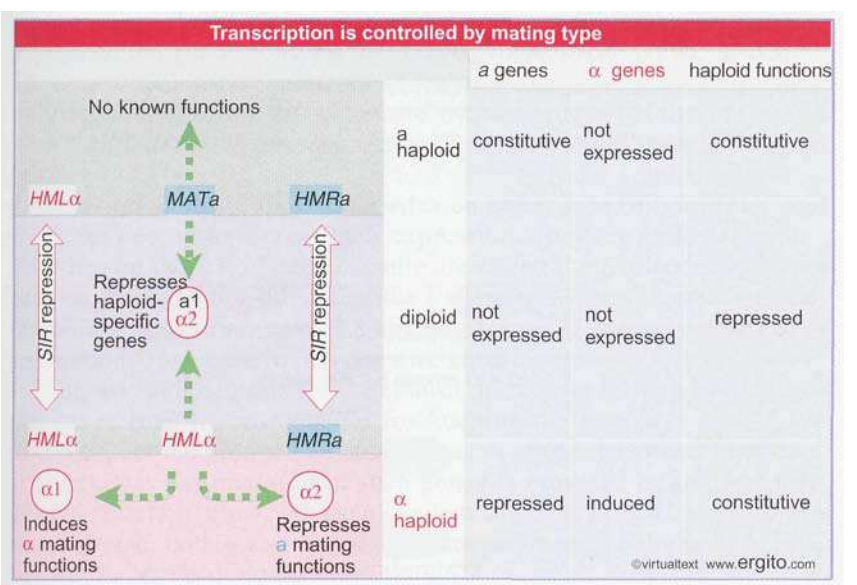


Figure 18.9 In diploids the α1 and α2 proteins cooperate to repress haploid-specific functions. In a haploids, mating functions are constitutive. In α haploids, the a2 protein represses a mating functions, while a1 induces α mating functions.

diploid yeast cells, as outlined in the diagram on the left of Figure 18.9. The *a* and a mating types are regulated by different mechanisms:

- In *a* haploids, a mating functions are expressed constitutively. The functions of the products of *MATa* in the a cell (if any) are unknown. It may be required only to repress haploid functions in diploid cells.
- In a haploids, the a 1 product turns on α-specific genes whose products are needed for a mating type. The a2 product represses the genes responsible for producing a mating type, by binding to an operator sequence located upstream of target genes.
- In diploids, the a1 and a2 products cooperate to repress haploid-specific genes. They combine to recognize an operator sequence different from the target for a2 alone.

The abilities of the a2, a1, and α1 proteins to regulate transcription rely upon some interesting protein-protein interactions between themselves and with other protein(s). The pattern of gene control in *a* cells, a cells, and diploids, is summarized in Figure 18.10.

A protein called PRTF (which is not specific for mating type) is involved in many of these interactions. PRTF binds to a short consensus sequence called the P box. The role of PRTF in gene regulation may be quite extensive, because P boxes are found in a variety of locations. In some of these sites, the P box is required for activation of the gene; but at other loci, PRTF is needed for repression. Its effects may therefore depend on the other proteins that bind at sites adjacent to the P box.

Genes that are a-specific may be activated by PRTF alone. This is adequate to ensure their expression in an a haploid.

The a-specific genes are repressed in an a haploid by the combined action of the a2 protein and PRTF. The a2 protein contains two domains. The C-terminal domain binds to short palindromic elements at the ends of an operator consensus sequence of 32 bp. However, binding of this fragment to DNA does not cause repression. The N-terminal domain is needed for repression and is responsible for making contacts with PRTF. The binding site for PRTF is a P box in the center of the operator. In fact, a and PRTF bind to the operator cooperatively.

Expression of α-specific genes requires the α1 activator. This is another small protein, 175 amino acids long. *Cis*-acting sequences that

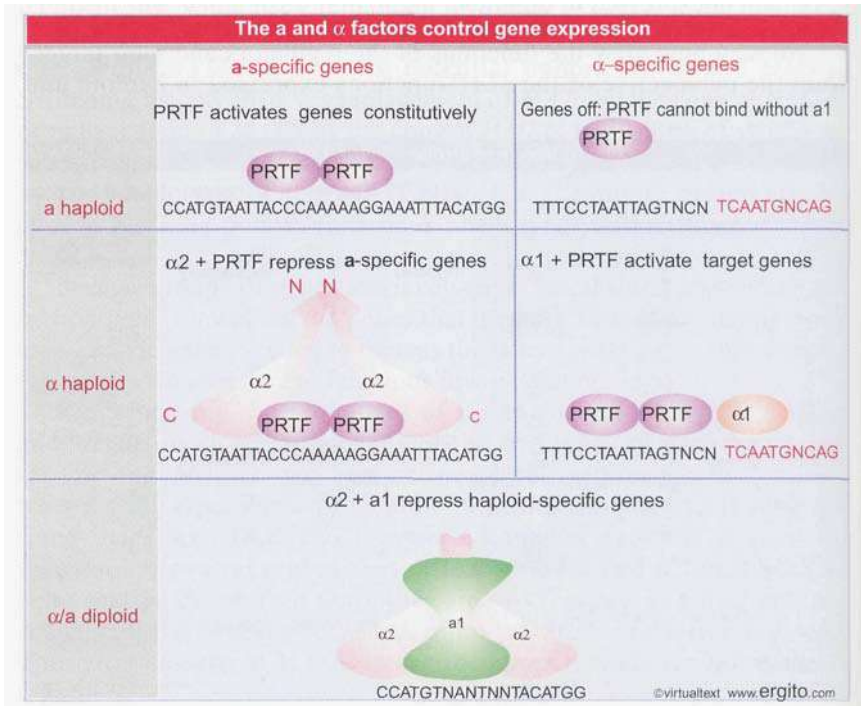


Figure 18.10 Combinations of PRTF, a1, α1 and α2 activate or repress specific groups of genes to correspond with the mating type of the cell.

By Book_Crazy [IND]

confer α-specific transcription are 26 bp long, and can be divided into two parts. The first 16 bp form the P box, where PRTF binds; the adjacent 10 bp sequence forms the binding site for α1. The α1 factor binds only when PRTF is present to bind to the P box. Neither protein alone can bind to its target box, but together they can bind to DNA, presumably as a result of protein-protein interactions.

The α-specific genes are turned off by default in a haploids, because in the absence of a 1 protein, PRTF is unable to bind to activate them.

The a2 protein can also cooperate with the al protein. The combination of these proteins recognizes a different operator. The operator shares the outlying palindromic sequences with the sequence recognized by a2 alone, but is shorter because the sequence between them is different. The α1/α2 combination represses genes with this motif in diploid cells.

The major point to be made from these results is that the phenotype of each type of cell *(a* or a haploid or *a/a* diploid) is determined by the combination of a and a proteins that are expressed. One aspect is the distinction between the haploid and diploid phenotypes; another is the distinction between a and a haploid phenotypes. The latter extends to expression of genes corresponding to the appropriate mating type and to the determination of the direction of switching of mating type (see Figure 18.7). *MATa* cells activate a recombination enhancer on the left arm of chromosome III, which increases recombination over a 40 kb region that includes *HML*. *MATa* cells inactivate the left end of chromosome III.

18.7 Silent cassettes at *HML* and HMR are repressed

Key Concepts

* *HML* and *HMR* are repressed by silencer elements.
* Loci required to maintain silencing include *SIR1-4, RAP1,* and genes for histone H4.
* Binding of ORC (origin recognition complex) at the silencers is necessary for inactivation.

The transcription map in Figure 18.8 reveals an intriguing feature. Transcription of either *MATa* or *MATa* initiates within the *Y* region. Only the *MAT* locus is expressed; yet the same *Y* region is present in the corresponding nontranscribed cassette *(HML* or *HMR).* This implies that regulation of expression is not accomplished by direct recognition of some site overlapping with the promoter. *A site outside the cassettes must distinguish HML and HMR from MAT.*

Deletion analysis shows that sites on either side of both *HML* and *HMR* are needed to repress their expression. They are called silencers. The sites on the left of each cassette are called the E silencers, and the sites on the right side are called the I silencers. These control sites can function at a distance (up to 2.5 kb away from a promoter) and in either orientation. They behave like negative enhancers.

Can we find the basis for the control of cassette activity by identifying genes that are responsible for keeping the cassettes silent? We would expect the products of these genes to act on the silencers. A convenient assay for mutation in such genes is provided by the fact that, when a mutation allows the usually silent cassettes at *HML* and *HMR* to be expressed, both a and a functions are produced, so the cells behave like *MATa/MATa* diploids.

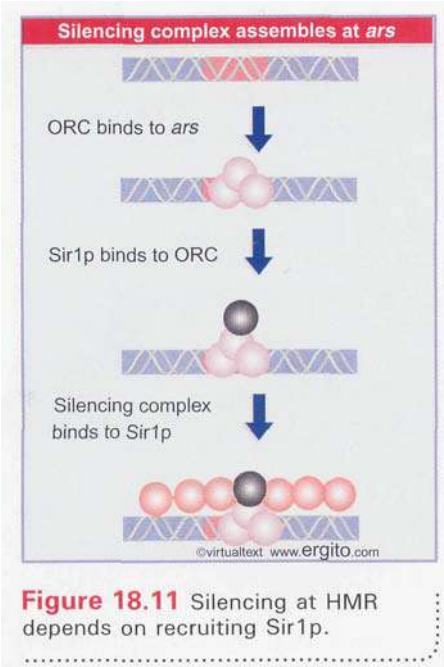


Figure 18.11 Silencing at HMR depends on recruiting Sir1p.

Mutations in several loci abolish silencing and lead to expression of *HML* and *HMR*. The first to be discovered were the four *SIR* loci (silent information regulators). All four wild-type *SIR* loci are needed to maintain *HML* and *HMR* in the repressed state; mutation in any one of these loci to give a *sir*⁻ allele has two effects. Both *HML* and *HMR* can be transcribed. And both the silent cassettes become targets for replacement by switching. So the same regulatory event is involved in repressing a silent cassette and in preventing it from being a recipient for replacement by another cassette.

Other loci required for silencing include *RAP1* (which is also required to maintain telomeric heterochromatin in its inert state) and the genes coding for histone H4. Deletions of the N-terminus of histone H4 or individual point mutations activate the silent cassettes. The effects of these mutations can be overcome either by introducing new mutations in *SIR3* or by over-expressing *SIR1*, which suggests that there is a specific interaction between H4 and the SIR proteins.

The general model suggested by these results is that the SIR proteins act on chromatin structure to prevent expression of the genes. Because mutations in the SIR proteins have the same effects on genes that have been inactivated by the proximity of telomeric heterochromatin, it seems likely that SIR proteins are involved generally in interacting with histones to form heterochromatic (inert) structures (see *23.15 Heterochromatin depends on interactions with histones).*

There is an interesting connection between repression at the silencers and DNA replication. Each silencer contains an *ARS* sequence (an origin of replication). The *ARS* is bound by the ORC (the origin recognition complex) that is involved in initiating replication. Mutations in *ORC* genes prevent silencing, indicating that the binding of ORC protein at the silencer is required for silencing.

There are two separate types of connection between silencing and the replication apparatus:

- The presence of Sir1 is necessary.
- And replication is required.

If a Sirl protein is localized at the silencer (by linkage to another protein that is bound there), the binding of ORC is no longer necessary. This means that the role of ORC is solely to bring in Sirl; it is not required to initiate replication. As illustrated in Figure 18.11, the role of ORC could therefore be to provide an initiating center from which the silencing effect can spread. ORC provides the structure to which Sirlp binds, and Sirlp then recruits the other SIR proteins. This is different from its role in replication.

However, passage through S phase is necessary for silencing to be established. This does not require initiation to occur at the *ARS* in the silencer. The effect could depend on the passage of a replication fork through the silencer, perhaps in order to allow the chromatin structure to be changed.

18.8 Unidirectional transposition is initiated by the recipient *MAT* locus

Key Concepts

- Mating type switching is initiated by a double strand break made at the *MAT* locus by the HO endonuclease.

A switch in mating type is accomplished by a gene conversion in which the recipient site *(MAT)* acquires the sequence of the donor

By Book_Crazy [IND]

type (*HML* or *HMR*). Sites needed for transposition have been identified by mutations at *MAT* that prevent switching. The unidirectional nature of the process is indicated by lack of mutations in *HML* or *HMR*.

The mutations identify a site at the right boundary of *Y* at *MAT* that is crucial for the switching event. The nature of the boundary is shown by analyzing the locations of these point mutations relative to the site of switching (this is done by examining the results of rare switches that occur in spite of the mutation). Some mutations lie within the region that is replaced (and thus disappear from *MAT* after a switch), while others lie just outside the replaced region (and therefore continue to impede switching). So sequences both within and outside the replaced region are needed for the switching event.

Switching is initiated by a double-strand break close to the *Y-Z* boundary that coincides with a site that is sensitive to attack by DNAase. (This is a common feature of chromosomal sites that are involved in initiating transcription or recombination.) It is recognized by an endonuclease coded by the *HO* locus. The HO endonuclease makes a staggered double-strand break just to the right of the *Y* boundary. Cleavage generates the single-stranded ends of 4 bases drawn in **Figure 18.12**. The nuclease does not attack mutant *MAT* loci that cannot switch. Deletion analysis shows that most or all of the sequence of 24 bp surrounding the *Y* junction is required for cleavage *in vitro*. The recognition site is relatively large for a nuclease, and it occurs only at the three mating-type cassettes.

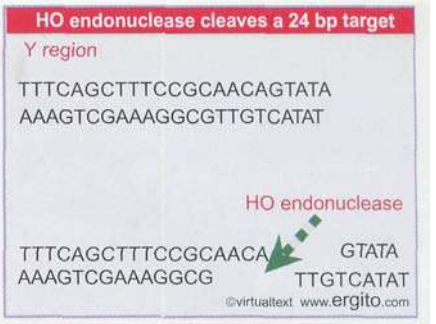


Figure 18.12 HO endonuclease cleaves *MAT* just to the right of the Y region, generating sticky ends with a base overhang.

Only the *MAT* locus and not the *HML* or *HMR* loci are targets for the endonuclease. It seems plausible that the same mechanisms that keep the silent cassettes from being transcribed also keep them inaccessible to the HO endonuclease. This inaccessibility ensures that switching is unidirectional.

The reaction triggered by the cleavage is illustrated schematically in **Figure 18.13** in terms of the general reaction between donor and recipient regions. In terms of the interactions of individual strands of DNA, it follows the scheme for recombination via a double-strand break drawn in Figure 15.8; and the stages following the initial cut require the enzymes involved in general recombination. Mutations in some of these genes prevent switching.

Suppose that the free end of *MAT* invades either the *HML* or *HMR* locus and pairs with the region of homology on the right side. The *Y* region of *MAT* is degraded until a region with homology on the left side is exposed. At this point, *MAT* is paired with *HML* or *HMR* at both the left side and the right side. The *Y* region of *HML* or *HMR* is copied to replace the region lost from *MAT* (which might extend beyond the limits of *Y* itself). The paired loci separate. (The order of events could be different.)

Like the double-strand break model for recombination, the process is initiated by *MAT*, the locus that is to be replaced. In this sense, the description of *HML* and *HMR* as donor loci refers to their ultimate role, but not to the mechanism of the process. Like replicative transposition, the donor site is unaffected, but a change in sequence occurs at the recipient; unlike transposition, the recipient locus suffers a substitution rather than addition of material.

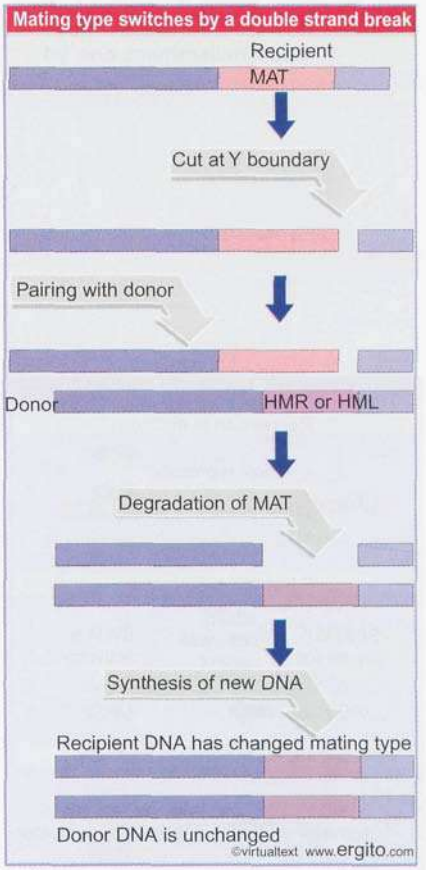


Figure 18.13 Cassette substitution is initiated by a double-strand break in the recipient (*MAT*) locus, and may involve pairing on either side of the Y region with the donor (*HMR* or *HML*) locus.

18.9 Regulation of HO expression controls switching

Key Concepts

- HO endonuclease is synthesized in haploid mother cells, so that a switching event causes both daughters to have the new mating type.

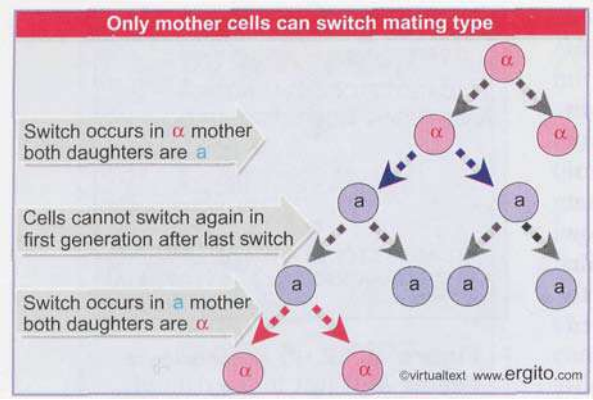


Figure 18.14 Switching occurs only in mother cells; both daughter cells have the new mating type. A daughter cell must pass through an entire cycle before it becomes a mother cell that is able to switch again.

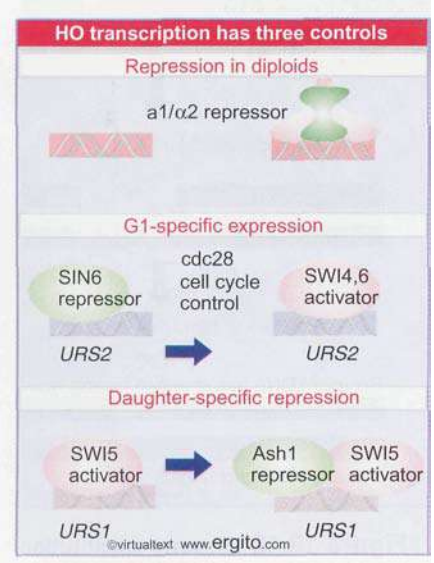


Figure 18.15 Three regulator systems act on transcription of the *HO* gene. Transcription occurs only when all repression is lifted.

Production of the HO endonuclease is regulated at the level of gene transcription. There are three separate control systems:

- *HO* is under mating-type control. It is not synthesized in *MATa*/*MATα* diploids. The reason could be that there is no need for switching when both *MAT* alleles are expressed anyway.
- *HO* is transcribed in mother cells but not in daughter cells.
- *HO* transcription also responds to the cell cycle. The gene is expressed only at the end of the G1 phase of a mother cell.

The timing of nuclease production explains the relationship between switching and cell lineage. **Figure 18.14** shows that switching is detected only in the products of a division; both daughter cells have the same mating type, switched from that of the parent. The reason is that the restriction of *HO* expression to G1 phase ensures that the mating type is switched before the *MAT* locus is replicated, with the result that both progeny have the new mating type.

cis-acting sites that control *HO* transcription reside in the 1500 bp upstream of the gene. The general pattern of control is that repression at any one of many sites, responding to several regulatory circuits, may prevent transcription of *HO*. **Figure 18.15** summarizes the types of sites that are involved.

Mating type control resembles that of other haploid-specific genes. Transcription is prevented (in diploids) by the **a1**/α2 repressor. There are 10 binding sites for the repressor in the upstream region. These sites vary in their conformity to the consensus sequence; we do not know which and how many of them are required for haploid-specific repression.

The control of *HO* transcription involves interplay between a series of activating and repressing events. The genes *SWI1-5* are required for *HO* transcription. They function by preventing products of the genes *SIN1-6* from repressing *HO*. The *SWI* genes were discovered first, as mutants unable to switch; then the *SIN* genes were discovered for their ability to release the blocks caused by particular *SWI* mutations. *SWI-SIN* interactions are involved in both cell-cycle control and the restriction of expression to mother cells.

Some of the *SWI* and *SIN* genes are not specifically concerned with mating type, but are global regulators of transcription, whose functions are needed for expression of many loci. They include the activator complex *SWI1,2,3* and the loci *SIN1-4* that code for chromosomal proteins. Their role in mating type expression is incidental. The "real" regulator is therefore the SWI protein that counteracts the general repression system specifically at the *HO* locus.

Cell-cycle control is conferred by 9 copies of an octanucleotide sequence called *URS2*. A copy of the consensus sequence can confer cell-cycle control on a gene to which it is attached. A gene linked to this sequence is repressed except during a transient period toward the end of G1 phase. SWI4 and SWI6 are the activators that release repression at *URS2*. Their activity depends on the function of the cell-cycle regulator CDC28, which executes the decision that commits the cell to divide (see *29 Cell cycle and growth regulation).*

The target for restricting expression to alternate generations is the activator SWI5 (which antagonizes a general repression system exercised by SIN3,4). In mutants that lack these functions, *HO* is transcribed equally well in mother and daughter cells. This system acts on *URS1* elements in the far upstream region.

SWI5 is not itself the regulator of mother-cell specificity, but is antagonized by Ash1p, a repressor that accumulates preferentially in daughter cells at the end of anaphase. Mutations in *ASH1* allow

daughter cells to switch mating type. The localization of Ash1p is determined by the transport of its mRNA from the mother cell along actin filaments into the daughter bud (*5.16 mRNA can be specifically localized).* Its presence prevents SWI5 from activating the *HO* gene. It works by binding to many copies of a consensus sequence that are distributed throughout the regulatory regions *URS1* and *URS2.* When the daughter cell grows to become a mother cell, the concentration of Ashlp is diluted, and it becomes possible to express the *HO* gene again.

18.10 Trypanosomes switch the VSG frequently during infection

Key Concepts

- The trypanosome life cycle alternates between tsetse fly and mammal.
- The form of the parasite that is transmitted to the mammal has a coat of a VSG (variable surface glycoprotein).
- The VSG is replaced every 1-2 weeks.

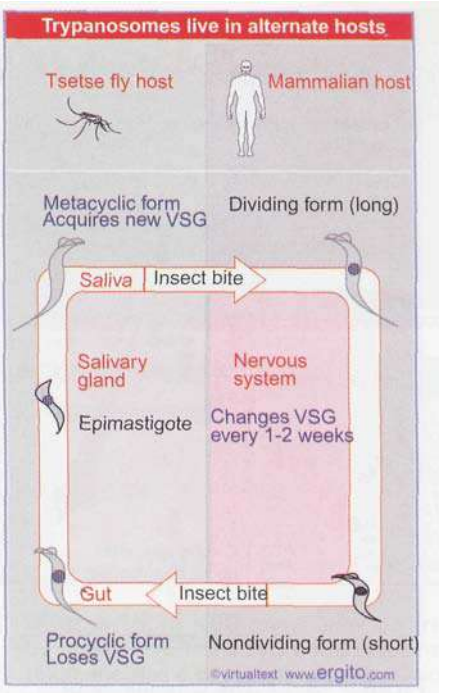


Figure 18.16 A trypanosome passes through several morphological forms when its life cycle alternates between a tsetse fly and mammalian host.

Sleeping sickness in man (and a related disease in cows) is caused by infection with African trypanosomes. The unicellular parasite follows the life cycle illustrated in **Figure 18.16,** in which it alternates between tsetse fly and mammal. The trypanosome may be transferred either to or from the fly when it bites a mammal.

During its life cycle, the parasite undergoes several morphological and biochemical changes. The most significant biochemical change is in the variable surface glycoprotein (VSG), the major component of the surface coat. The coat covers the plasma membrane and consists of a monolayer of 5-10 X 10^6 molecules of a single VSG, which is the only antigenic structure exposed on the surface. A trypanosome expresses only one VSG at any time, and its ability to change the VSG is responsible for its survival through the fly-mammal infective cycle.

Consider the cycle as starting when a fly gains a trypanosome by biting an infected mammal. The trypanosome enters the gut of the fly in the "procyclic form," and loses its VSG. After about three weeks, its progeny differentiate into the "metacyclic form," which re-acquires a VSG coat. This form is transmitted to the mammalian bloodstream during a bite by the fly. The trypanosome multiplies in the mammalian bloodstream. Its progeny continue to express the metacyclic VSG for about a week. Then a new VSG is synthesized, and further transitions occur every 1-2 weeks.

Each of the successive VSG species is immunologically distinct. As a result, the antigen presented to the mammalian immune system is constantly changing. The process of transition is called antigenic variation. The immune response of the organism always lags behind the change in surface antigen, so that the trypanosome evades immune surveillance, and thereby perpetuates itself indefinitely. Each transition of the VSG is accompanied by a new wave of parasitemia, with symptoms of fever, rash, etc.; the parasites eventually invade the central nervous system, after which the mammalian host becomes progressively more lethargic and eventually comatose.

Trypanosomes vary in their host range. The best investigated species is a variety of *Trypanosoma brucei* that grows well on laboratory animals (although not on man). Laboratory strains of *T. brucei*

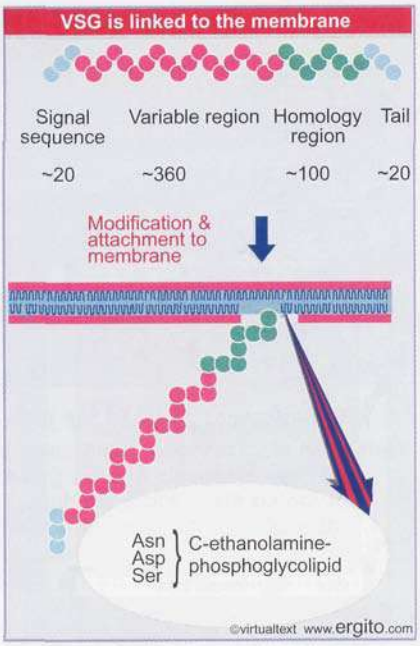


Figure 18.17 The C-terminus of VSG is cleaved and covalently linked to the membrane through a glycolipid.

switch VSGs spontaneously at a rate of 10^{-4} to 10^{-6} per division. Switching occurs independently of the host immune system. In effect, new variants then are selected by the host, because it mounts a response against the old VSG, but fails to recognize and act against the new VSG.

A general view of VSG structure is depicted in Figure 18.17. A nascent VSG is ~500 amino acids long; it has an N-terminal signal sequence, followed by a long variable region that provides the unique antigenic determinant, and a C-terminal homology region ending in a short hydrophobic tail. The nascent VSG is processed at both ends to give the mature form. The signal sequence is cleaved during secretion. The hydrophobic tail is removed before the VSG reaches the outside surface. The new C-terminus is covalently attached to the trypanosome membrane; three types of homology region are distinguished according to the C-terminal amino acid.

The VSG is attached to the membrane via a phosphoglycolipid. As a result, VSG can be released from the membrane by an enzyme that removes fatty acid. This reaction (which is used in purifying the VSG) may be important *in vivo* in allowing one VSG to be replaced by another on the surface of the trypanosome.

18.11 New VSG sequences are generated by gene switching

Key Concepts

- The trypanosome has ~1000 basic copy VSG genes.
- Only a single expression-linked copy located near a telomere is expressed at any given time.
- A basic copy gene is activated by having its sequence copied into the expression site.
- There are only a few potential expression sites.

How many varieties of VSG can be expressed by any one trypanosome? It is not clear that any limit is encountered before death of the host. A single trypanosome can make at least 100 VSGs sufficiently different in sequence that antibodies against any one do not react against the others.

VSG variation is coded in the trypanosome genome. Every individual trypanosome carries the entire VSG repertoire of its strain. Diversity therefore depends on changing expression from one preexisting gene to another.

The trypanosome genome has an unusual organization, consisting of a large number of segregating units. In addition to an unknown number of chromosomes, it contains ~100 "minichromosomes," each containing ~50-150 kb of DNA. Hybridization experiments identify ~1000 VSG genes, scattered among all size classes of chromosomal material.

Each VSG is coded by a basic copy gene. These genes can be divided into two classes according to their chromosomal location:

- Telomeric genes lie within 5-15 kb of a telomere. There could be >200 of these genes if every telomere has one.
- Internal genes reside within chromosomes (more formally, they lie >50 kb from a telomere).

As might be expected of a large family of genes, individual basic copies show varying degrees of relationship, presumably reflecting

their origin by duplication and variation. Genes that are closely related, and which provoke the same antigenic response, are called isogenes.

How is a single VSG gene selected for expression? Only one VSG gene is transcribed in a trypanosome at a given time. The copy of the gene that is active is called the expression-linked copy (ELC). It is said to be located at an expression site. An expression site has a characteristic property: it is located near a telomere.

These features immediately suggest that the route followed to select a gene for expression depends on whether the basic copy is itself telomeric or internal. The two types of event that can create an ELC are summarized in **Figure 18.18:**

- The expression site remains the same, but the ELC is changed. Duplication transfers the sequence of a basic copy to replace the sequence currently occupying the expression site. Either internal or telomeric copies may be activated directly by duplication into the expression site. The substitution of one cassette for another does not interfere with the activity of the site.
- The expression site is changed. Activation *in situ* is available only to a sequence already present at a telomere. When a telomeric site is activated *in situ*, the previous expression site must cease to be active and the new site now becomes the expression site.

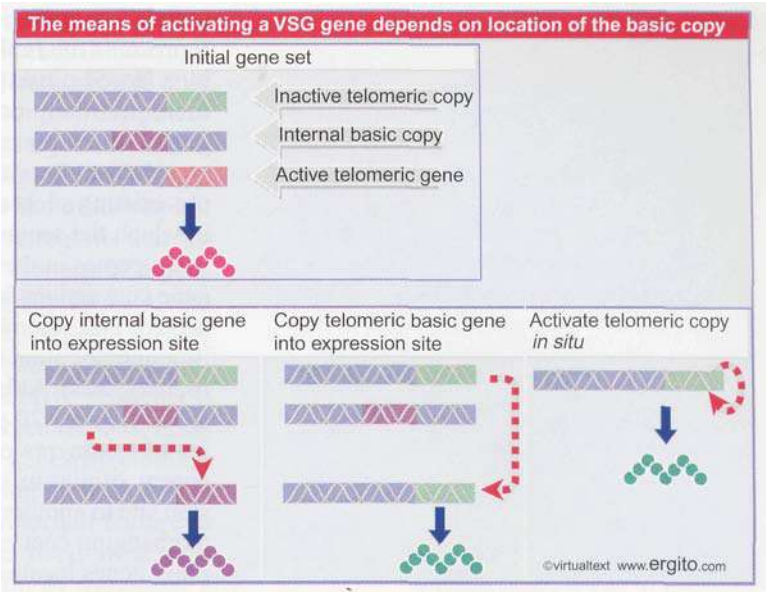


Figure 18.18 VSG genes may be created by duplicative transfer from an internal or telomeric basic copy into an expression site, or by activating a telomeric copy that is already present at a potential expression site.

Internal basic copies probably can be copied into non-expressed telomeric locations as well as into expression sites. So an internal gene could be activated by a two-stage process, in which first it is transposed to a non-expressed telomere, and then this site is activated.

We can follow the fate of genes involved in activation by restriction mapping. A probe representing an expressed sequence can be derived from the mRNA. Then we can determine the status of genes corresponding to the probe. We see different results for internal and telomeric basic copy genes:

- Activation of an internal gene requires generation of new sequences. **Figure 18.19** shows that when an internal gene is activated, a new fragment is found. The original basic-copy gene remains unaltered; the new fragment is generated by the duplication of the gene into a new context (where the sites recognized by the restriction enzyme are in the surrounding sequences and therefore generate a distinct fragment). The new fragment identifies an ELC, located close to a telomere. The ELC appears when the gene is expressed and disappears when the gene is switched off. Duplication into the ELC is the only pathway by which an internal basic copy can be generated.
- Activation of a telomeric gene can occur *in situ*. **Figure 18.20** shows that when a telomeric gene is activated, the gene number need not change. The structure of the gene may be essentially unaffected as detected by restriction mapping. The size of the fragment containing the gene may vary slightly, because the length of the telomere is constantly changing. Telomeric basic copies can also be activated by the same duplication pathway as internal copies; in this case, the basic copy remains at its telomere, while an expression-linked copy appears at another telomere (generating a new fragment as illustrated for internal basic copies in Figure 18.19).

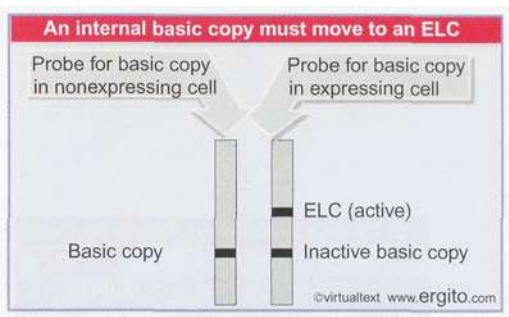


Figure 18.19 Internal basic copies can be activated only by generating a duplication of the gene at an expression-linked site.

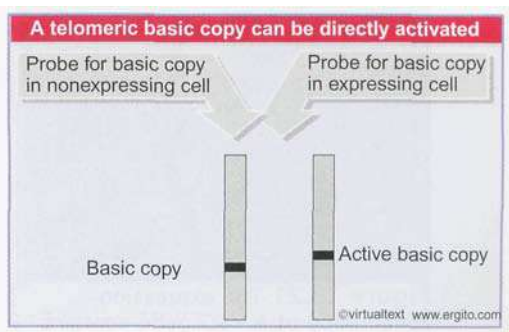


Figure 18.20 Telomeric basic copies can be activated *in situ;* the size of the restriction fragment may change (slightly) when the telomere is extended.

Formation of the ELC occurs by a gene conversion process that requires genetic recombination enzymes—for example, it is greatly

reduced by mutation in *RAD51*. Like the switch in yeast mating type, it represents the replacement of a "cassette" at the active (telomeric) locus by a stored cassette. The VSG system is more versatile in the sense that there are many potential donor cassettes (and also more than a single potential recipient site).

Almost all switches in VSG type involve replacement of the ELC by a pre-existing silent copy. Some exceptional cases have been found, however, in which the sequence of the ELC does not match any of the repertoire of silent copies in the genome. A new sequence may be created by a series of gene conversions in which short stretches of different silent copies are connected. This resembles the mechanism for generating diversity in chicken λ immunoglobulins (see *26.16 Avian immunoglobulins are assembled from pseudogenes*). Although rare, such occurrences extend VSG diversity.

There are ~20 potential expression sites, which means that only a subset of telomeres can function in this capacity. All the expression sites appear similar in sequence and organization. Switching from one expression site to another occurs at a low frequency. This is not a principal means of changing coat expression, but has the effect of switching expression of other genes located within the expression site. Among these genes are two coding for the subunits of a (heterodimeric) transferrin receptor. Changing the transferrin receptor that is expressed by the trypanosome has a strong effect on its affinity for the host transferrin protein.

18.12 VSG genes have an unusual structure

Key Concepts

- The coding region of the VSG gene is flanked by barren regions of repetitive DNA.
- The VSG sequence is transferred into the ELC between the promoter and terminator.
- The 5' end of VSG mRNA is added by a *trans*-splicing reaction to the 5' end that is generated by transcription.
- Activation of an expression site involves a change in the region upstream of the promoter.
- The expression site may be contained in a special extranucleolar body where the VSG is transcribed by RNA polymerase I.

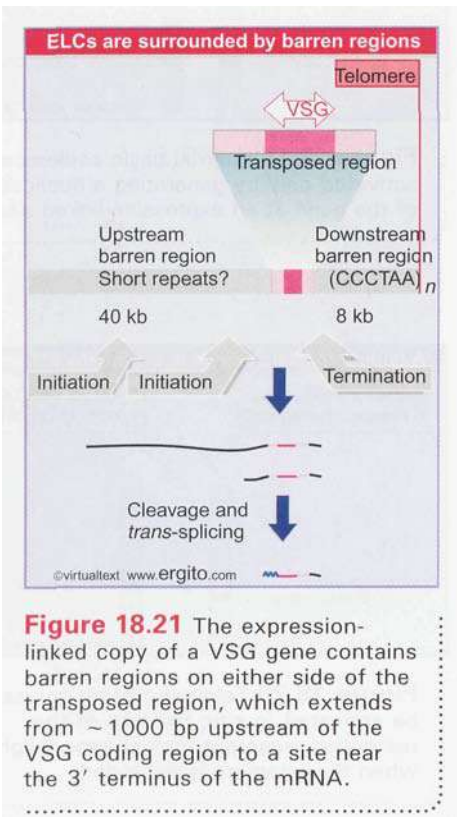


Figure 18.21 The expression-linked copy of a VSG gene contains barren regions on either side of the transposed region, which extends from ~1000 bp upstream of the VSG coding region to a site near the 3′ terminus of the mRNA.

The structure of the VSG gene at the ELC is unusual, as illustrated in Figure 18.21. The length of DNA transferred into the ELC is 2500-3500 bp, somewhat longer than the VSG-coding region of 1500 bp. Most of the additional length is upstream of the gene. The crossover points at which the duplicated sequence joins the ELC do not appear to be precisely determined.

Analysis of events at the 5' end of the VSG mRNA is complicated by the fact that the mature RNA starts with a 35 base sequence coded elsewhere, and added in *trans* to the newly synthesized 5' end (see *24.13 trans-splicing reactions use small RNAs*). The signals for initiating and terminating transcription (and sometimes also the end of the coding region itself) are provided by the sequences flanking the transposed region. In fact, transcription may be initiated several kb upstream of the VSG gene itself. Promoters have been mapped at 4 kb and ~60 kb upstream of the VSG sequence. Use of the more distant promoter generates a transcript that contains other genes as well as the active VSG. The VSG sequence (and other gene sequences) must be released by cleavage from the transcript, after which the 35 base spliced leader is added to the 5' end.

The RNA polymerase that transcribes the expression locus is not the usual RNA polymerase II, but is RNA polymerase I (the enzyme that

usually transcribes rRNA.) The ELC is sequestered in a discrete nuclear body, called the expression site body (ESB). The ESB takes the form of an extranucleolar body containing RNA polymerase I, and is found only in the bloodstream form. This may explain why only one of the potential 20 expression copies is in fact expressed in a given trypanosome. If the ESB is necessary for expression and can only accommodate a single copy, then by default it will prevent expression of all the other copies.

On either side of the transposed region are extensive regions that are not cut by restriction enzymes. These "barren regions" consist of repetitive DNA; they extend some 8 kb downstream and for up to 40 kb upstream of the ELC. Going downstream, the barren region consists largely of repeats of the sequence CCCTAA, and extends to the telomere. Proceeding upstream, it may also consist of repetitive sequences, but their nature is not yet clear. The existence of the barren regions, however, has been an impediment to characterizing ELC genes by cloning.

The order in which VSG genes are expressed during an infection is erratic, but not completely random. This may be an important feature in survival of the trypanosome. If VSG genes were used in a predetermined order, a host could knock out the infection by mounting a reaction against one of the early elements. The need for unpredictability in the production of VSGs may be responsible for the evolution of a system with many donor sequences and multiple recipients.

Antigenic variation is not a unique phenomenon of trypanosomes. The bacterium *Borrelia hermsii* causes relapsing fever in man and analogous diseases in other mammals. The name of the disease reflects its erratic course: periods of illness are spaced by periods of relief. When the fevers occur, spirochetes are found in the blood; they disappear during periods of relief, as the host responds with specific antibodies.

Like the trypanosomes, *Borrelia* survives by altering a surface protein, called the variable major protein (VMP). Changes in the VMP are associated with rearrangements in the genome. The active VMP is located near the telomere of a linear plasmid. We do not yet know the extent of the coded variants or the mechanisms used to alter their expression. It is intriguing, however, that the eukaryote *Trypanosoma* and the prokaryote *Borrelia* should both rely upon antigenic variation as a means for evading immune surveillance.

18.13 The bacterial Ti plasmid causes crown gall disease in plants

Key Concepts

- Infection with the bacterium *A. tumefaciens* can transform plant cells into tumors.
- The infectious agent is a plasmid carried by the bacterium.
- The plasmid also carries genes for synthesizing and metabolizing opines (arginine derivatives) that are used by the tumor cell.

Most events in which DNA is rearranged or amplified occur within a genome, but the interaction between bacteria and certain plants involves the transfer of DNA from the bacterial genome to the plant genome. Crown gall disease, shown in **Figure 18.22,** can be induced in most dicotyledonous plants by the soil bacterium *Agrobacterium tumefaciens*. The bacterium is a parasite that effects a genetic change in the eukaryotic host cell, with consequences for both parasite and host. It improves conditions for survival of the parasite. And it causes the plant cell to grow as a tumor.

Figure 18.22 An *Agrobacterium* carrying a Ti plasmid of the nopaline type induces a teratoma, in which differentiated structures develop. Photograph kindly provided by Jeff Schell.

Ti genes function in bacteria and in plants

Locus	Function	Ti Plasmid
vir	DNA transfer into plant	all
shi	shoot induction	all
roi	root induction	all
nos	nopaline synthesis	nopaline
noc	nopaline catabolism	nopaline
ocs	octopine synthesis	octopine
occ	octopine catabolism	octopine
tra	bacterial transfer genes	all
Inc	incompatibility genes	all
oriV	origin for replication	all

Figure 18.23 Ti plasmids carry genes involved in both plant and bacterial functions.

Agrobacteria are required to induce tumor formation, but the tumor cells do not require the continued presence of bacteria. Like animal tumors, the plant cells have been transformed into a state in which new mechanisms govern growth and differentiation. Transformation is caused by the expression within the plant cell of genetic information transferred from the bacterium.

The tumor-inducing principle of *Agrobacterium* resides in the Ti plasmid, which is perpetuated as an independent replicon within the bacterium. The plasmid carries genes involved in various bacterial and plant cell activities, including those required to generate the transformed state, and a set of genes concerned with synthesis or utilization of opines (novel derivatives of arginine).

Ti plasmids (and thus the *Agrobacteria* in which they reside) can be divided into four groups, according to the types of opine that are made:

- Nopaline plasmids carry genes for synthesizing nopaline in tumors and for utilizing it in bacteria. Nopaline tumors can differentiate into shoots with abnormal structures. They have been called teratomas by analogy with certain mammalian tumors that retain the ability to differentiate into early embryonic structures.
- Octopine plasmids are similar to nopaline plasmids, but the relevant opine is different. However, octopine tumors are usually undifferentiated and do not form teratoma shoots.
- Agropine plasmids carry genes for agropine metabolism; the tumors do not differentiate, develop poorly, and die early.
- Ri plasmids can induce hairy root disease on some plants and crown gall on others. They have agropine type genes, and may have segments derived from both nopaline and octopine plasmids.

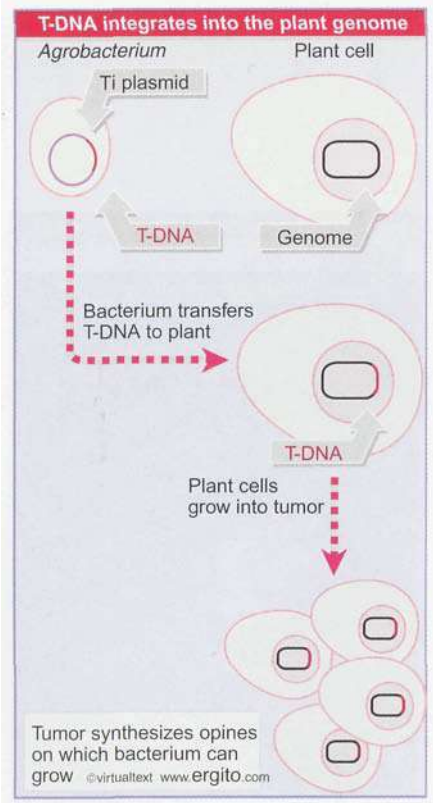


Figure 18.24 T-DNA is transferred from *Agrobacterium* carrying a Ti plasmid into a plant cell, where it becomes integrated into the nuclear genome and expresses functions that transform the host cell.

The types of genes carried by a Ti plasmid are summarized in Figure 18.23. Genes utilized in the bacterium code for plasmid replication and incompatibility, for transfer between bacteria, sensitivity to phages, and for synthesis of other compounds, some of which are toxic to other soil bacteria. Genes used in the plant cell code for transfer of DNA into the plant, for induction of the transformed state, and for shoot and root induction.

The specificity of the opine genes depends on the type of plasmid. Genes needed for opine synthesis are linked to genes whose products catabolize the same opine; thus each strain of *Agrobacterium* causes crown gall tumor cells to synthesize opines that are useful for survival of the parasite. The opines can be used as the sole carbon and/or nitrogen source for the inducing *Agrobacterium* strain. The principle is that the transformed plant cell synthesizes those opines that the bacterium can use.

18.14 T-DNA carries genes required for infection

Key Concepts

- Part of the DNA of the Ti plasmid is transferred to the plant cell nucleus.
- The *vir* genes of the Ti plasmid are located outside the transferred region and are required for the transfer process.
- The *vir* genes are induced by phenolic compounds released by plants in response to wounding.
- The membrane protein VirA is autophosphorylated on histidine when it binds an inducer.
- VirA activates VirG by transferring the phosphate group to it.
- The VirA-VirG is one of several bacterial two component systems that use a phosphohistidine relay.

By Book_Crazy [IND]

The interaction between *Agrobacterium* and a plant cell is illustrated in Figure 18.24. The bacterium does not enter the plant cell, but transfers part of the Ti plasmid to the plant nucleus. The transferred part of the Ti genome is called T-DNA. It becomes integrated into the plant genome, where it expresses the functions needed to synthesize opines and to transform the plant cell.

Transformation of plant cells requires three types of function carried in the *Agrobacterium:*

- Three loci on the *Agrobacterium* chromosome, *chvA, chvB, pscA* are required for the initial stage of binding the bacterium to the plant cell. They are responsible for synthesizing a polysaccharide on the bacterial cell surface.
- The *vir* region carried by the Ti plasmid outside the T-DNA region is required to release and initiate transfer of the T-DNA.
- The T-DNA is required to transform the plant cell.

The organization of the major two types of Ti plasmid is illustrated in Figure 18.25. About 30% of the ~200 kb Ti genome is common to nopaline and octopine plasmids. The common regions include genes involved in all stages of the interaction between *Agrobacterium* and a plant host, but considerable rearrangement of the sequences has occurred between the plasmids.

The T-region occupies ~23 kb. Some 9 kb is the same in the two types of plasmid. The Ti plasmids carry genes for opine synthesis *(Nos* or *Ocs)* within the T-region; corresponding genes for opine catabolism *(Noc* or *Occ)* reside elsewhere on the plasmid. The plasmids code for similar, but not identical, morphogenetic functions, as seen in the induction of characteristic types of tumors.

Functions affecting oncogenicity—the ability to form tumors—are not confined to the T-region. Those genes located outside the T-region must be concerned with establishing the tumorigenic state, but their products are not needed to perpetuate it. They may be concerned with transfer of T-DNA into the plant nucleus or perhaps with subsidiary functions such as the balance of plant hormones in the infected tissue. Some of the mutations are host-specific, preventing tumor formation by some plant species, but not by others.

The virulence genes code for the functions required for the transfer process. Six loci *virA-G* reside in a 40 kb region outside the T-DNA. Their organization is summarized in Figure 18.26. Each locus is transcribed as an individual unit; some contain more than one open reading frame.

We may divide the transforming process into (at least) two stages:

- Agrobacterium contacts a plant cell, and the *vir* genes are induced.
- *vir* gene products cause T-DNA to be transferred to the plant cell nucleus, where it is integrated into the genome.

The *vir* genes fall into two groups, corresponding to these stages. Genes *vir A* and *virG* are regulators that respond to a change in the plant by inducing the other genes. So mutants in *virA* and *virG* are avirulent and cannot express the remaining *vir* genes. Genes *virB,C,D,E* code for proteins involved in the transfer of DNA. Mutants in *virB* and *virD* are avirulent in all plants, but the effects of mutations in *virC* and *virE* vary with the type of host plant.

vir A and *virG* are expressed constitutively (at a rather low level). The signal to which they respond is provided by phenolic compounds generated by plants as a response to wounding. Figure 18.27 presents an example. *N. tabacum* (tobacco) generates the molecules acetosyringone and α-hydroxyacetosyringone. Exposure to these compounds activates *vir A,* which acts on *virG,* which in turn induces the expression de novo of *virB,C,D,E.* This reaction explains why *Agrobacterium* infection succeeds only on wounded plants.

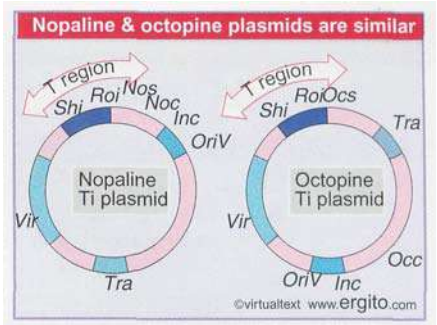


Figure 18.25 Nopaline and octopine Ti plasmids carry a variety of genes, including T-regions that have overlapping functions

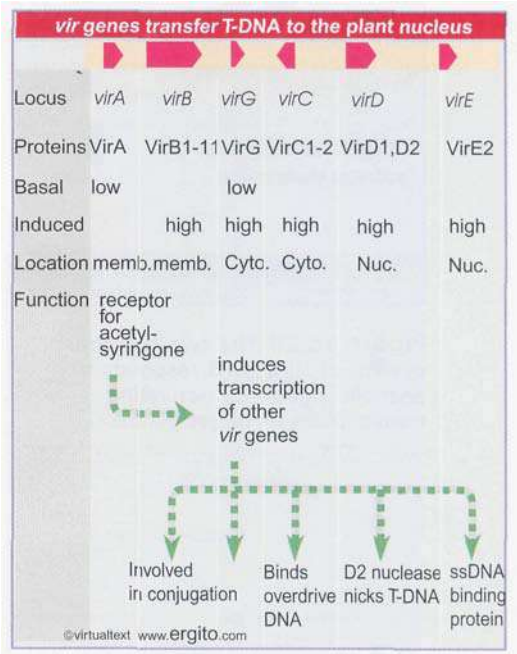


Figure 18.26 The *vir* region of the Ti plasmid has six loci that are responsible for transferring T-DNA to an infected plant.

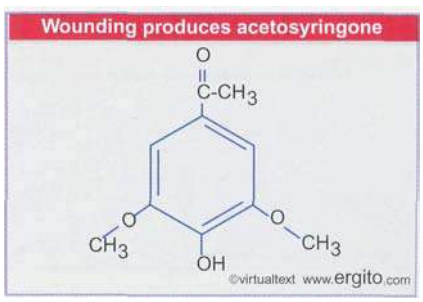


Figure 18.27 Acetosyringone (4-acetyl-2,6-dimethoxyphenol) is produced by *N. tabacum* upon wounding, and induces transfer of T-DNA from *Agrobacterium.*

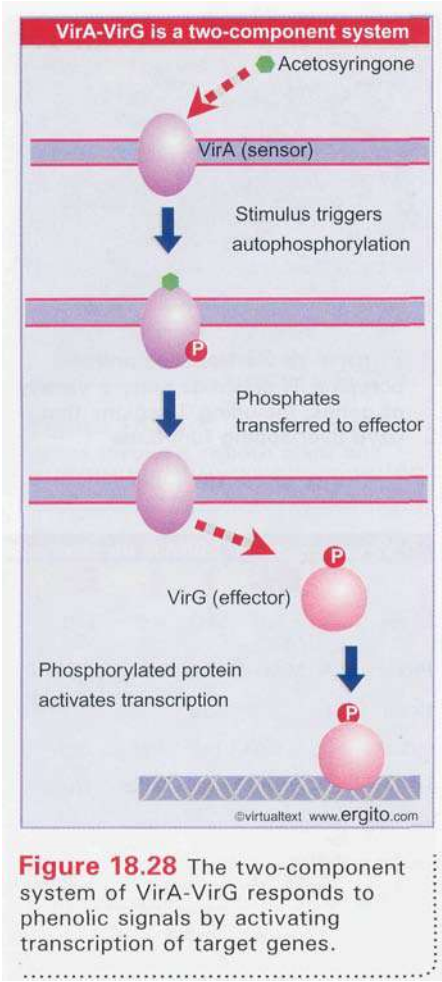


Figure 18.28 The two-component system of VirA-VirG responds to phenolic signals by activating transcription of target genes.

VirA and VirG are an example of a classic type of bacterial system in which stimulation of a sensor protein causes autophosphorylation and transfer of the phosphate to the second protein. The relationship is illustrated in **Figure 18.28.** The VirA-VirG system resembles the EnvZ-OmpR system that responds to osmolarity. The sequence of *virA* is related to *envZ*; and the sequences of *virG* and *ompR* are closely related, suggesting that the effector proteins function in a similar manner.

VirA forms a homodimer that is located in the inner membrane; it may respond to the presence of the phenolic compounds in the periplasmic space. Exposure to these compounds causes VirA to become autophosphorylated on histidine. The phosphate group is then transferred to an Asp residue in VirG. The phosphorylated VirG binds to promoters of the *virB,C,D,E* genes to activate transcription. When *virG* is activated, its transcription is induced from a new startpoint, different from that used for constitutive expression, with the result that the amount of VirG protein is increased.

Of the other *vir* loci, *virD* is the best characterized. The *virD* locus has 4 open reading frames. Two of the proteins coded at *virD*, VirD1 and VirD2, provide an endonuclease that initiates the transfer process by nicking T-DNA at a specific site.

18.15 Transfer of T-DNA resembles bacterial conjugation

Key Concepts

- T-DNA is generated when a nick at the right boundary creates a primer for synthesis of a new DNA strand.
- The preexisting single-strand that is displaced by the new synthesis is transferred to the plant cell nucleus.
- Transfer is terminated when DNA synthesis reaches a nick at the left boundary.
- The T-DNA is transferred as a complex of single-stranded DNA with the VirE2 single strand-binding protein.
- The single stranded T-DNA is converted into double-stranded DNA and integrated into the plant genome.
- The mechanism of integration is not known. T-DNA can be used to transfer genes into a plant nucleus.

The transfer process actually selects the T-region for entry into the plant. **Figure 18.29** shows that the T-DNA of a nopaline plasmid is demarcated from the flanking regions in the Ti plasmid by repeats of 25 bp, which differ at only two positions between the left and right ends. When T-DNA is integrated into a plant genome, it has a well-defined right junction, which retains 1-2 bp of the right repeat. The left junction is variable; the boundary of T-DNA in the plant genome may be located at the 25 bp repeat or at one of a series of sites extending over ~100 bp within the T-DNA. Sometimes multiple tandem copies of T-DNA are integrated at a single site.

A model for transfer is illustrated in **Figure 18.30**. A nick is made at the right 25 bp repeat. It provides a priming end for synthesis of a DNA single strand. Synthesis of the new strand displaces the old strand, which is used in the transfer process. Transfer is terminated when DNA synthesis reaches a nick at the left repeat. This model explains why the right repeat is essential, and it accounts for the polarity of the process. If the left repeat fails to be nicked, transfer could continue farther along the Ti plasmid.

By Book_Crazy [IND]

The transfer process involves production of a single molecule of single-stranded DNA in the infecting bacterium. It is transferred in the form of a DNA-protein complex, sometimes called the T-complex. The DNA is covered by the VirE2 single-strand binding protein, which has a nuclear localization signal and is responsible for transporting T-DNA into the plant cell nucleus. A single molecule of the D2 subunit of the endonuclease remains bound at the 5' end. The *virB* operon codes for 11 products that are involved in the transfer reaction.

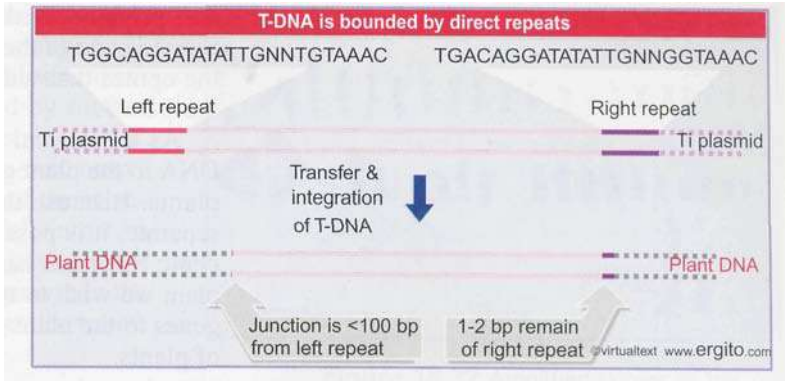


Figure 18.29 T-DNA has almost identical repeats of 25 bp at each end in the Ti plasmid. The right repeat is necessary for transfer and integration to a plant genome. T-DNA that is integrated in a plant genome has a precise junction that retains 1-2 bp of the right repeat, but the left junction varies and may be up to 100 bp short of the left repeat.

Outside T-DNA, but immediately adjacent to the right border, is another short sequence, called *overdrive*, which greatly stimulates the transfer process. Overdrive functions like an enhancer: it must lie on the same molecule of DNA, but enhances the efficiency of taw&fe; even when located several thousand base pairs away from the border. VirC1, and possibly VirC2, may act at the overdrive sequence.

This model for transfer of T-DNA closely resembles the events involved in bacterial conjugation, when the *E. coli* chromosome is transferred from one cell to another in single-stranded form. The genes of the *virB* operon are homologous to the *tra* genes of certain bacterial plasmids that are involved in conjugation (see *13.13 Conjugation transfers single-stranded DNA).* A difference is that the transfer of T-DNA is (usually) limited by the boundary of the left repeat, whereas transfer of bacterial DNA is indefinite.

We do not know how the transferred DNA is integrated into the plant genome. At some stage, the newly generated single strand must be converted into duplex DNA. Circles of T-DNA that are found in infected plant cells appear to be generated by recombination between the left and right 25 bp repeats, but we do not know if they are intermediates. The actual event is likely to involve a nonhomologous recombination, because there is no homology between the T-DNA and the sites of integration.

Is T-DNA integrated into the plant genome as an integral unit? How many copies are integrated? What sites in plant DNA are available for integration? Are genes in T-DNA regulated exclusively by functions on the integrated segment? These questions are central to defining the process by which the Ti plasmid transforms a plant cell into a tumor.

What is the structure of the target site? Sequences flanking the integrated T-DNA tend to be rich in A·T base pairs (a feature displayed in target sites for some transposable elements). The sequence rearrangements that occur at the ends of the integrated T-DNA make it difficult to analyze the structure. We do not know whether the integration process generates new sequences in the target DNA comparable to the target repeats created in transposition.

T-DNA is expressed at its site of integration. The region contains several transcription units, each probably containing a gene expressed from an individual promoter. Their functions are concerned with the state of the plant cell, maintaining its tumorigenic properties, controlling shoot and root formation, and suppressing differentiation into other tissues. None of these genes is needed for T-DNA transfer.

The Ti plasmid presents an interesting organization of functions. Outside the T-region, it carries genes needed to initiate oncogenesis; at least some are concerned with the transfer of T-DNA, and we should like to know whether others function in the plant cell to affect its behavior at this stage. Also outside the T-region are the genes that enable the *Agrobacterium* to catabolize the opine that the transformed plant cell

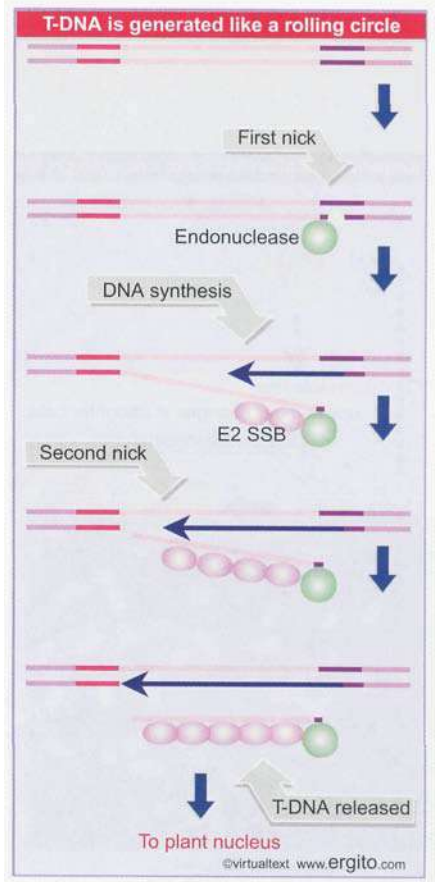


Figure 18.30 T-DNA is generated by displacement when DNA synthesis starts at a nick made at the right repeat. The reaction is terminated by a nick at the left repeat.

will produce. Within the T-region are the genes that control the transformed state of the plant, as well as the genes that cause it to synthesize the opines that will benefit the *Agrobacterium* that originally provided the T-DNA.

As a practical matter, the ability of *Agrobacterium* to transfer T-DNA to the plant genome makes it possible to introduce new genes into plants. Because the transfer/integration and oncogenic functions are separate, it is possible to engineer new Ti plasmids in which the oncogenic functions have been replaced by other genes whose effect on the plant we wish to test. The existence of a natural system for delivering genes to the plant genome should greatly facilitate genetic engineering of plants.

18.16 DNA amplification generates extra gene copies

Key Concepts

- Eukaryotic cells acquire resistance to methotrexate by amplifying the number of *dhfr* genes.
- The initial step of amplification is the generation of extrachromosomal DNA molecules containing tandem repeats of the DHFR locus.
- The extrachromosomal DNA forms double minute chromosomes that are lost frequently at division.
- Stable resistant lines are generated by generation of amplified copies in the chromosome.
- It is not known whether stable lines arise by *de novo* amplification *in situ* or by insertion of extrachromosomal amplified sequences.

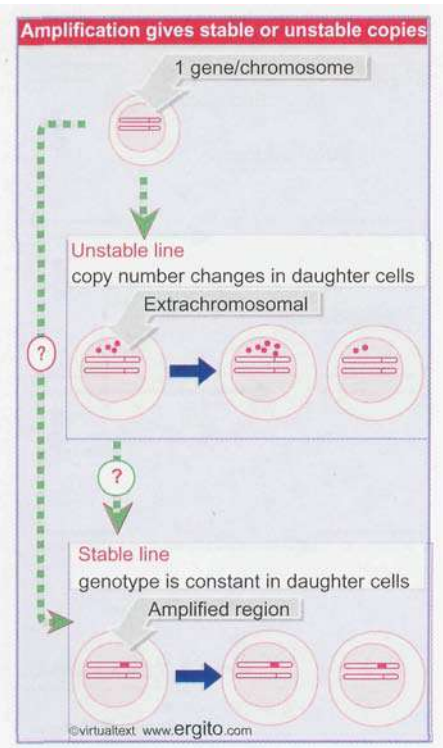


Figure 18.31 The *dhfr* gene can be amplified to give unstable copies that are extrachromosomal (double minutes) or stable (chromosomal). Extrachromosomal copies arise at early times.

The eukaryotic genome has the capacity to accommodate additional sequences of either exogenous or endogenous origin. Endogenous sequences may be produced by amplification of an existing sequence. The additional sequences often take the form of a tandem array, containing many copies of a repeating unit. A gene that is contained within the repeating unit is not necessarily expressed in every copy, but expression tends to increase with the number of copies.

A tandem array of multiple copies may exist in either of two forms in a cell. If it takes the form of an extrachromosomal unit, it is inherited in an irregular manner: there is no equivalent in animal cells to the ability of a bacterial plasmid to be segregated evenly at cell division, so the entire unit is lost at a high frequency. If the array is integrated into the genome, however, it becomes a component of the genotype, and is inherited like any other genomic sequence.

Amplification of endogenous sequences is provoked by selecting cells for resistance to certain agents. The best-characterized example of amplification results from the addition of methotrexate (mtx) to certain cultured cell lines. This reagent blocks folate metabolism. Resistance to it is conferred by mutations that change the activity of the enzyme dihydrofolate reductase (DHFR). As an alternative to change in the enzyme itself, the amount of enzyme may be increased. The cause of this increase is an amplification of the number of *dhfr* structural genes. Amplification occurs at a frequency greater than the spontaneous point mutation rate, generally ranging from 10^{-4}-10^{-6}. Similar events now have been observed in >20 other genes.

A common feature in most of these systems is that highly resistant cells are not obtained in a single step, but instead appear when the cells

are adapted to gradually increasing doses of the toxic reagent. So gene amplification may require several stages. Amplification generally occurs at only one of the two *dhfr* alleles; and increased resistance to methotrexate is accomplished by further increases in the degree of amplification at this locus.

The number *of dhfr* genes in a cell line resistant to methotrexate varies from 40-400, depending on the stringency of the selection and the individual cell line. The mtx^r lines fall into two classes, distinguished by their response when the selective pressure for high levels of DHFR activity is relieved by growth in the absence of methotrexate (the basis for the difference is illustrated in **Figure 18.31).**

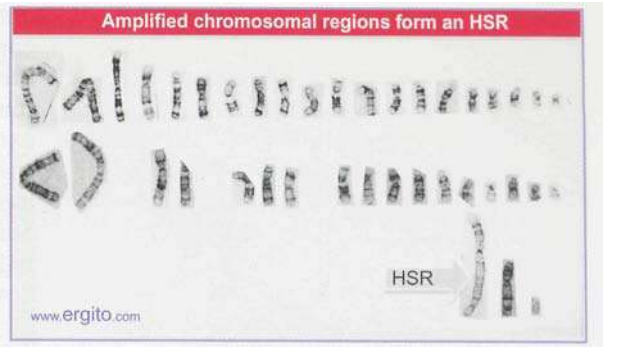


Figure 18.32 Amplified copies of the *dhfr* gene produce a homogeneously staining region *(HSR)* in the chromosome. Photograph kindly provided by Robert Schimke.

- In *stable* lines, the amplified genes are retained, because they reside on the chromosome, at the site usually occupied by the single *dhfr* gene. Usually the other chromosome retains its normal single copy of *dhfr*.
- In *unstable* lines, the amplified genes are at least partially lost when the selective pressure is released, because the amplified genes exist as an extrachromosomal array.

Gene amplification has a visible effect on the chromosomes. In stable lines, the *dhfr* locus can be visualized in the form of a homogeneously staining region (HSR). An example is shown in **Figure 18.32.** The HSR takes its name from the presence of an additional region that lacks any chromosome bands after treatments such as G-banding. This change suggests that some region of the chromosome between bands has undergone an expansion.

In unstable cell lines, no change is seen in the chromosomes carrying *dhfr*. However, large numbers of elements called double-minute chromosomes are visible, as can be seen in **Figure 18.33.** In a typical cell line, each double-minute carries 2-4 *dhfr* genes. The double minutes appear to be self-replicating; but they lack centromeres. As a result, they do not attach to the mitotic spindle and therefore segregate erratically, frequently being lost from the daughter cells. Notwithstanding their name, the actual status of the double minutes is regarded as extrachromosomal.

The irregular inheritance of the double minutes explains the instability of methotrexate resistance in these lines. Double minutes are lost continuously during cell divisions; and in the presence of methotrexate, cells with reduced numbers of *dhfr* genes will die. Only those cells that have retained a sufficient number of double minutes will appear in the surviving population.

The presence of the double minutes reduces the rate at which the cells proliferate. So when the selective pressure is removed, cells lacking the amplified genes have an advantage; they generate progeny more rapidly and soon take over the population. This explains why the amplified state is retained in the cell line only so long as cells are grown in the presence of methotrexate.

Because of the erratic segregation of the double minutes, increases in the copy number can occur relatively quickly as cells are selected at each division for progeny that have gained more than their fair share of the *dhfr* genes. Cells with greater numbers of copies are found in response to increased levels of methotrexate. The behavior of the double-minutes explains the stepwise evolution of the mtx^r condition and the incessant fluctuation in the level of *dhfr* genes in unstable lines.

Both stable and unstable lines are found after long periods of selection for methotrexate resistance. What is the initial step in gene amplification? After short periods of selection, most or all of the resistant cells are unstable. The formation of extrachromosomal copies clearly is a more frequent event than amplification within the

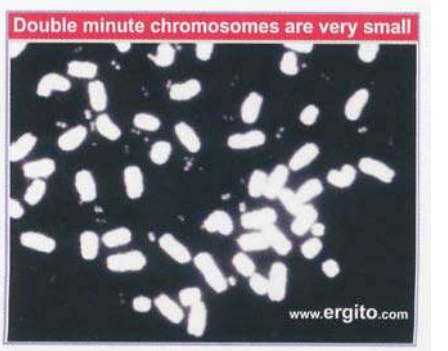


Figure 18.33 Amplified extrachromosomal *dhfr* genes take the form of double-minute chromosomes, as seen in the form of the small white dots. Photograph kindly provided by Robert Schimke.

chromosome. At very early times in the process, amplified *dhfr* genes can be found as (small) extrachromosomal units before double minutes or any change in chromosomes can be detected. This suggests that the acquisition of resistance is most often due to generation of extrachromosomal repeats.

The amplified region is longer than the *dhfr* gene itself. The gene has a length of ~31 kb, but the average length of the repeated unit is 500-1000 kb in the chromosomal HSR. The extent of the amplified region is different in each cell line. The amount of DNA contained in a double minute seems to lie in a range of 100-1000 kb.

How do the extrachromosomal copies arise? We know that their generation occurs without loss of the original chromosomal copy. There are two general possibilities. Additional cycles of replication could be initiated in the vicinity of the *dhfr* gene, followed by nonhomologous recombination between the copies. Or the process could be initiated by nonhomologous recombination between alleles. The extra copies could be released from the chromosome, possibly by some recombination-like event. Depending on the nature of this event, it could generate an extrachromosomal DNA molecule containing one or several copies. If the double minutes contain circular DNA, recombination between them in any case is likely to generate multimeric molecules.

Some information about the events involved in perpetuating the double minutes is given by an unstable cell line whose amplified genes code for a mutant DHFR enzyme. The mutant enzyme is not present in the original (diploid) cell line (so the mutation must have arisen at some point during the amplification process). Despite variations in the number of amplified genes, these cells display only the mutant enzyme. So the wild-type chromosomal genes cannot be continuously generating large numbers of double minutes anew, because these amplified copies would produce normal enzyme.

Once amplified extrachromosomal genes have arisen, therefore, changes in the state of the cell are mediated through these genes and not through the original chromosomal copies. When methotrexate is removed, the cell line loses its double minutes in the usual way. On re-exposure to the reagent, normal genes are amplified to give a new population of double minutes. This shows that none of the extrachromosomal copies of the mutant gene had integrated into the chromosome.

Another striking implication of these results is that the double minutes of the mutant line carried only mutant genes—so if there is more than one *dhfr* gene per double minute, all must be of the mutant type. This suggests that multicopy double minutes can be generated from individual extrachromosomal genes.

A major question has been whether amplified chromosomal copies arise by integration of the extrachromosomal copies or by an independent mechanism. We do not know whether intrachromosomal amplification simply proceeds less often as a *de novo* step or requires extrachromosomal amplification to occur as an intermediate step. The form taken by the amplified genes is influenced by the cell genotype; some cell lines tend to generate double minutes, while others more readily display the HSR configuration.

The type of amplification event also depends upon the particular locus that is involved. Another case of amplification is provided by resistance to an inhibitor of the enzyme transcarbamylase, which occurs by amplification of the CAD gene. (CAD is a protein which has the first three enzymatic activities of the pathway for UMP synthesis.) Amplified CAD DNA is always found within the chromosome. In this case, the amplified genes are found in the form of several dispersed amplified regions, often involving more than one chromosome.

18.17 Transfection introduces exogenous DNA into cells

Key Concepts

- DNA that is transfected into a eukaryotic cells forms a large repeating unit of many head to tail tandem repeats.
- The transfected unit is unstable unless it becomes integrated into a host chromosome.
- Genes carried by the transfected DNA can be expressed.

The procedure for introducing exogenous donor DNA into recipient cells is called transfection. Transfection experiments began with the addition of preparations of metaphase chromosomes to cell suspensions. The chromosomes are taken up rather inefficiently by the cells and give rise to unstable variants at a low frequency. Intact chromosomes rarely survive the procedure; the recipient cell usually gains a fragment of a donor chromosome (which is unstable because it lacks a centromere). Rare cases of stable lines may have resulted from integration of donor material into a resident chromosome.

Similar results are obtained when purified DNA is added to a recipient cell preparation. However, with purified DNA it is possible to add particular sequences instead of relying on random fragmentation of chromosomes. Transfection with DNA yields stable as well as unstable lines, with the former relatively predominant. (These experiments are directly analogous to those performed in bacterial transformation, but are described as transfection because of the historical use of "transformation" to describe changes that allow unrestrained growth of eukaryotic cells.)

Unstable transfectants (sometimes called transient transfectants) reflect the survival of the transfected DNA in extrachromosomal form; stable lines result from integration into the genome. The transfected DNA can be expressed in both cases. However, the low frequencies of transfection make it necessary to use donor markers whose presence in the recipient cells can be selected for. Most transfection experiments have used markers representing readily assayed enzymatic functions, but, in principle, any marker that can be selected can be assayed. This allows the isolation of genes responsible for morphological phenomena. Most notably, transfected cells can be selected for acquisition of the transformed (tumorigenic) phenotype. This type of protocol has led to the isolation of several cellular *onc* genes (see *30.9 Ras oncogenes can be detected in a transfection assay).*

Cotransfection with more than one marker has proved informative about the events involved in transfection and extends the range of questions that we can ask with this technique. A common marker used in such experiments is the *tk* gene, coding for the enzyme thymidine kinase, which catalyzes an essential step in the provision of thymidine triphosphate as a precursor for DNA synthesis.

When *tk* cells are transfected with a DNA preparation containing both a purified rt^+ gene and the $\phi X174$ genome, all the tk^+ transformants have both donor sequences. This is a useful observation, because it allows unselected markers to be introduced routinely by cotransfection with a selected marker.

The arrangement of *tk* and $\phi X174$ sequences is different in each transfected line, but remains the same during propagation of that line. Often multiple copies of the donor sequences are present, the number varying with the individual line. Revertants lose the $\phi X174$ sequences together with *tk* sequences. Amplification of transfected sequences under

selective pressure results in the increase of copy number of all donor sequences *pari passu*. So the two types of donor sequence become physically linked during transfection and suffer the same fate thereafter.

To perform a transfection experiment, the mass of DNA added to the recipient cells is increased by including an excess of "carrier DNA," a preparation of some other DNA (often from salmon sperm). Transfected cells prove to have sequences of the carrier DNA flanking the selected sequences on either side. Transfection therefore appears to be mediated by a large unit, consisting of a linked array of all sequences present in the donor preparation.

Since revertants for the selected marker lose all of this material, it seems likely that the transfected cell gains only a single large unit. The unit is formed by a concatemeric linkage of donor sequences in a reaction that is rapid relative to the other events involved in transfection. This transfecting package is ~1000 kb in length.

Because of the size of the donor unit, we cannot tell from blotting experiments whether it is physically linked to recipient chromosomal DNA (the relevant end fragments are present in too small a relative proportion). It seems plausible that the first stage is the establishment of an unstable extrachromosomal unit, followed by the acquisition of stability via integration.

In situ hybridization can be used to show that transfected cells have donor material integrated into the resident chromosomes. Any given cell line has only a single site of integration; but the site is different in each line. Probably the selection of a site for integration is a random event; sometimes it is associated with a gross chromosomal rearrangement.

The sites at which exogenous material becomes integrated usually do not appear to have any sequence relationship to the transfected DNA. The integration event involves a nonhomologous recombination between the mass of added DNA and a random site in the genome. The recombination event may be provoked by the introduction of a double-strand break into the chromosomal DNA, possibly by the action of DNA repair enzymes that are induced by the free ends of the exogenous DNA. Integrants produced by the integration event are stable, and are therefore more useful than transient transfectants for experiments that rely on the expression of the transfected gene.

18.18 Genes can be injected into animal eggs

Key Concepts

- DNA that is injected into animal eggs can integrate into the genome.
- Usually a large array of tandem repeats integrates at a single site.
- Expression of the DNA is variable and may be affected by the site of integration and other epigenetic effects.

An exciting development of transfection techniques is their application to introduce genes into animals. An animal that gains new genetic information from the addition of foreign DNA is described as transgenic. The approach of directly injecting DNA can be used with mouse eggs, as shown in **Figure 18.34.** Plasmids carrying the gene of interest are injected into the germinal vesicle (nucleus) of the oocyte or into the pronucleus of the fertilized egg. The egg is implanted into a pseudopregnant mouse. After birth, the recipient mouse can be examined to see whether it has gained the foreign DNA, and, if so, whether it is expressed.

By Book_Crazy [IND]

The first questions we ask about any transgenic animal are how many copies it has of the foreign material, where these copies are located, and whether they are present in the germline and inherited in a Mendelian manner. The usual result of such experiments is that a reasonable minority (say ~15%) of the injected mice carry the transfected sequence. Usually, multiple copies of the plasmid appear to have been integrated in a tandem array into a single chromosomal site. The number of copies varies from 1-150. They are inherited by the progeny of the injected mouse as expected of a Mendelian locus.

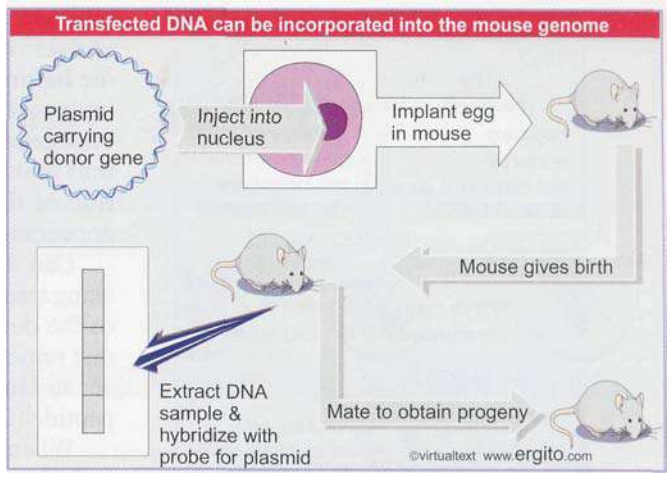


Figure 18.34 Transfection can introduce DNA directly into the germ line of animals.

An important issue that can be addressed by experiments with transgenic animals concerns the independence of genes and the effects of the region within which they reside. If we take a gene, including the flanking sequences that contain its known regulatory elements, can it be expressed independently of its location in the genome? In other words, do the regulatory elements function independently, or is gene expression in addition controlled by other effects, for example, location in an appropriate chromosomal domain?

Are transfected genes expressed with the proper developmental specificity? The general rule now appears to be that there is a reasonable facsimile of proper control: the transfected genes are generally expressed in appropriate cells and at the usual time. There are exceptions, however, in which a transfected gene is expressed in an inappropriate tissue.

In the progeny of the injected mice, expression of the donor gene is extremely variable; it may be extinguished entirely, reduced somewhat, or even increased. Even in the original parents, the level of gene expression does not correlate with the number of tandemly integrated genes. Probably only some of the genes are active. In addition to the question of how many of the gene copies are capable of being activated, a parameter influencing regulation could be the relationship between the gene number and the regulatory proteins: a large number of promoters could dilute out any regulator proteins present in limiting amounts.

What is responsible for the variation in gene expression? One possibility that has often been discussed for transfected genes (and which applies also to integrated retroviral genomes) is that the site of integration is important. Perhaps a gene is expressed if it integrates within an active domain, but not if it integrates in another area of chromatin. Another possibility is the occurrence of epigenetic modification; for example, changes in the pattern of methylation might be responsible for changes in activity. Alternatively, the genes that happened to be active in the parents may have been deleted or amplified in the progeny.

A particularly striking example of the effects of an injected gene is provided by a strain of transgenic mice derived from eggs injected with a fusion consisting of the MT promoter linked to the rat growth hormone structural gene. Growth hormone levels in some of the transgenic mice were several hundred times greater than normal. The mice grew to nearly twice the size of normal mice, as can be seen from **Figure 18.35.**

Figure 18.35 A transgenic mouse with an active rat growth hormone gene (left) is twice the size of a normal mouse (right). Photograph kindly provided by Ralph Brinster.

The introduction of oncogene sequences can lead to tumor formation. Transgenic mice containing the SV40 early coding region and regulatory elements express the viral genes for large T and small t antigens only in some tissues, most often brain, thymus, and kidney. (The T/t antigens are alternatively spliced proteins coded by the early region of the virus; they have the ability to transform cultured cells to a

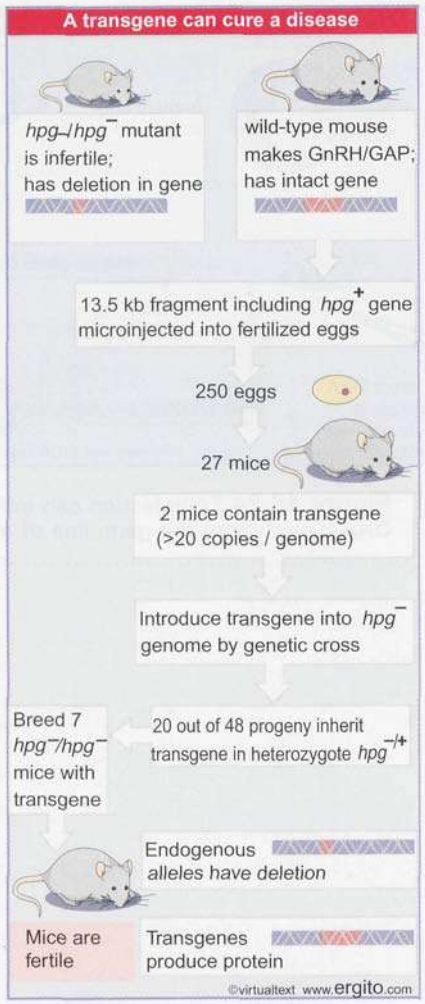


Figure 18.36 Hypogonadism of the *hpg* mouse can be cured by introducing a transgene that has the wild-type sequence.

tumorigenic phenotype; see *30.5 Early genes of DNA transforming viruses have multifunctional oncogenes.)* The transgenic mice usually die before reaching 6 months, as the result of developing tumors in the brain; sometimes tumors are found also in thymus and kidney. Different oncogenes may be used to generate mice developing various cancers, thus making possible a range of model systems. For example, introduction of the *myc* gene under control of an active promoter causes the appearance of adenocarcinomas and other tumors.

Can defective genes be replaced by functional genes in the germline using transgenic techniques? One successful case is represented by a cure of the defect in the hypogonadal mouse. The *hpg* mouse has a deletion that removes the distal part of the gene coding for the polyprotein precursor to GnRH (gonadotropin-releasing hormone) and GnRH-associated peptide (GAP). As a result, the mouse is infertile.

When an intact *hpg* gene is introduced into the mouse by transgenic techniques, it is expressed in the appropriate tissues. **Figure 18.36** summarizes experiments to introduce a transgene into *hpglhpg* homozygous mutant mice. The resulting mice are normal. This provides a striking demonstration that expression of a transgene under normal regulatory control can be indistinguishable from the behavior of the normal allele.

Impediments to using such techniques to cure genetic defects at present are that the transgene must be introduced into the germline of the preceding generation, the ability to express a transgene is not predictable, and an adequate level of expression of a transgene may be obtained in only a small minority of the transgenic animals. Also, the large number of transgenes that may be introduced into the germline, and their erratic expression, could pose problems for the animal in cases in which overexpression of the transgene was harmful.

In the *hpg* murine experiments, for example, only 2 out of 250 eggs mice injected with intact hpg genes gave rise to transgenic mice. Each transgenic animal contained >20 copies of the transgene. Only 20 of the 48 offspring of the transgenic mice retained the transgenic trait. When inherited by their offspring, however, the transgene(s) could substitute for the lack of endogenous *hpg* genes. Gene replacement via a transgene is therefore effective only under restricted conditions.

The disadvantage of direct injection of DNA is the introduction of multiple copies, their variable expression, and often difficulty in cloning the insertion site because sequence rearrangements may have been generated in the host DNA. An alternative procedure is to use a retroviral vector carrying the donor gene. A single proviral copy inserts at a chromosomal site, without inducing any rearrangement of the host DNA. It is possible also to treat cells at different stages of development, and thus to target a particular somatic tissue; however, it is difficult to infect germ cells.

18.19 ES cells can be incorporated into embryonic mice

Key Concepts

- ES (embryonic stem) cells that are injected into a mouse blastocyst generate descendant cells that become part of a chimeric adult mouse.
- When the ES cells contribute to the germline, the next generation of mice may be derived from the ES cell.
- Genes can be added to the mouse germline by transfecting them into ES cells before the cells are added to the blastocyst.

By Book_Crazy [IND]

A powerful technique for making transgenic mice takes advantage of embryonic stem (ES) cells, which are derived from the mouse blastocyst (an early stage of development, which precedes implantation of the egg in the uterus). **Figure 18.37** illustrates the principles of this technique.

ES cells are transfected with DNA in the usual way (most often by microinjection or electroporation). By using a donor that carries an additional sequence such as a drug resistance marker or some particular enzyme, it is possible to select ES cells that have obtained an integrated transgene carrying any particular donor trait. An alternative is to use PCR technology to assay the transfected ES cells for successful integration of the donor DNA. By such means, a population of ES cells is obtained in which there is a high proportion carrying the marker.

These ES cells are then injected into a recipient blastocyst. The ability of the ES cells to participate in normal development of the blastocyst forms the basis of the technique. The blastocyst is implanted into a foster mother, and in due course develops into a *chimeric* mouse. Some of the tissues of the chimeric mice will be derived from the cells of the recipient blastocyst; other tissues will be derived from the injected ES cells. The proportion of tissues in the adult mouse that are derived from cells in the recipient blastocyst and from injected ES cells varies widely in individual progeny; if a visible marker (such as coat color gene) is used, areas of tissue representing each type of cell can be seen.

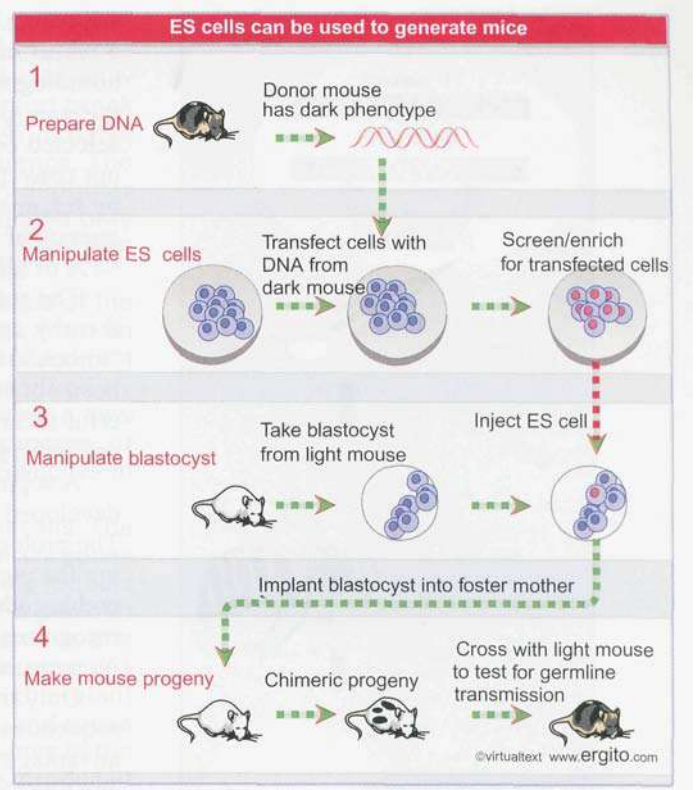


Figure 18.37 ES cells can be used to generate mouse chimeras, which breed true for the transfected DNA when the ES cell contributes to the germ line.

To determine whether the ES cells contributed to the germline, the chimeric mouse is crossed with a mouse that lacks the donor trait. Any progeny that have the trait must be derived from germ cells that have descended from the injected ES cells. By this means, an entire mouse has been generated from an original ES cell!

18.20 Gene targeting allows genes to be replaced or knocked out

Key Concepts

- An endogenous gene can be replaced by a transfected gene using homologous recombination.
- The occurrence of a homologous recombination can be detected by using two selectable markers, one of which is incorporated with the integrated gene, the other of which is lost when recombination occurs.

A further development of these techniques makes it possible to obtain homologous recombinants. A particular use of homologous recombination is to disrupt endogenous genes, as illustrated in **Figure 18.38.** A wild-type gene is modified by interrupting an exon with a marker sequence; most often the *neo* gene that confers resistance to the drug G418 is used. Also, another marker is added on one side of the gene; for example, the TK gene of the herpes virus. When this DNA is introduced into an ES cell, it may be inserted into the genome by either nonhomologous or homologous recombination. A nonhomologous recombination inserts the whole unit, including the flanking TK

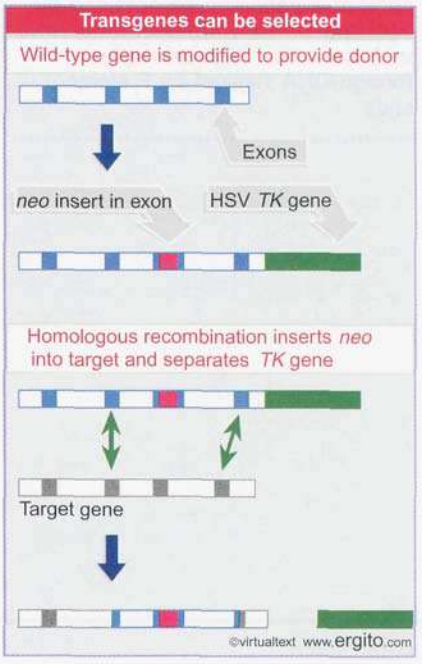


Figure 18.38 A transgene containing *neo* within an exon and TK downstream can be selected by resistance to G418 and loss of TK activity.

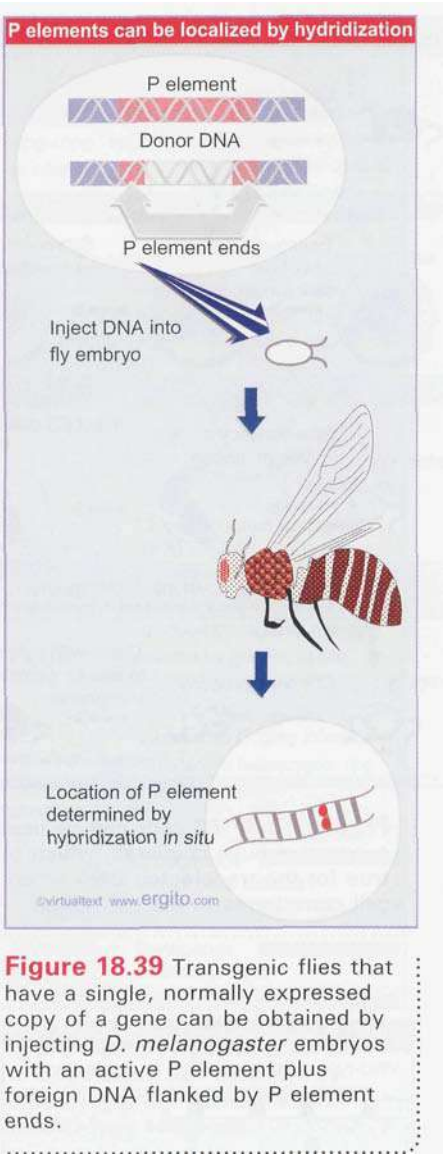


Figure 18.39 Transgenic flies that have a single, normally expressed copy of a gene can be obtained by injecting *D. melanogaster* embryos with an active P element plus foreign DNA flanked by P element ends.

sequence. But a homologous recombination requires two exchanges, as a result of which the flanking TK sequence is lost. Cells in which a homologous recombination has occurred can therefore be selected by the gain of *neo* resistance and absence of TK activity (which can be selected because TK causes sensitivity to the drug gancyclovir). If it is not convenient to use a selectable marker such as TK, cells can simply be screened by PCR assays for the absence of flanking DNA. The frequency of homologous recombination is ~10^{-7}, and probably represents <1% of all recombination events.

The presence of the *neo* gene in an exon disrupts transcription, and thereby creates a null allele. A particular target gene can therefore be "knocked out" by this means; and once a mouse with one null allele has been obtained, it can be bred to generate the homozygote. This is a powerful technique for investigating whether a particular gene is essential, and what functions in the animal are perturbed by its loss.

A sophisticated method for introducing new DNA sequences has been developed with *D. melanogaster* by taking advantage of the P element. The protocol is illustrated in **Figure 18.39.** A defective P element carrying the gene of interest is injected together with an intact P element into preblastoderm embryos. The intact P element provides a transposase that recognizes not only its own ends but also those of the defective element. As a result, either or both elements may be inserted into the genome.

Only the sequences between the ends of the P DNA are inserted; the sequences on either side are not part of the transposable element. An advantage of this technique is that only a single element is inserted in any one event, so the transgenic flies usually carry only one copy of the foreign gene, a great aid in analyzing its behavior.

Several genes that have been introduced in this way all show the same behavior. They are expressed only in the appropriate tissues and at the proper times during development, irrespective of the site of integration. So in *D. melanogaster,* all the information needed to regulate gene expression may be contained within the gene locus itself, and can be relatively impervious to external influence.

With these experiments, we see the possibility of extending from cultured cells to animals the option of examining the regulatory features. The ability to introduce DNA into the genotype allows us to make changes in it, to add new genes that have had particular modifications introduced *in vitro,* or to inactivate existing genes. So it becomes possible to delineate the features responsible for tissue-specific gene expression. Ultimately we may expect routinely to replace defective genes in the genotype in a targeted manner.

18.21 Summary

Yeast mating type is determined by whether the *MAT* locus carries the a or α sequence. Expression in haploid cells of the sequence at *MAT* leads to expression of genes specific for the mating type and to repression of genes specific for the other mating type. Both activation and repression are achieved by control of transcription, and require factors that are not specific for mating type as well as the products of *MAT.* The functions that are activated in either mating type include secretion of the appropriate pheromone and expression on the cell surface of the receptor for the opposite type of pheromone. Interaction between pheromone and receptor on cells of either mating type activates a G protein on the membrane, and sets in train a common pathway that prepares cells for sporulation. Diploid cells do not express mating-type functions.

Additional, silent copies of the mating-type sequences are carried at the loci *HMLα* and *HMRa*. They are repressed by the actions

of the *sir* loci. Cells that carry the HO endonuclease display a unidirectional transfer process in which the sequence at *HMLa* replaces an a sequence at *MAT,* or the sequence at *HMRa* replaces an α sequence at *MAT.* The endonuclease makes a double-strand break at *MAT,* and a free end invades either *HMLa* or *HMRa. MAT* initiates the transfer process, but is the recipient of the new sequence. The HO endonuclease is transcribed in mother cells but not daughter cells, and is under cell-cycle control. So switching is detected only in the products of a division, and the mating type has been switched in both daughter cells.

Trypanosomes carry >1000 sequences coding for varieties of the surface antigen. Only a single VSG is expressed in one cell, from an active site located near a telomere. The active site is localized in a special extranucleolar body, where the VSG gene is transcribed by RNA polymerase I. The VSG may be changed by substituting a new coding sequence at the active site via a gene conversion process, or by switching the site of expression to another telomere. Switches in expression occur every 10^4-10^6 divisions.

Agrobacteria induce tumor formation in wounded plant cells. The wounded cells secrete phenolic compounds that activate *vir* genes carried by the Ti plasmid of the bacterium. The *vir* gene products cause a single strand of DNA from the T-DNA region of the plasmid to be transferred to the plant cell nucleus. Transfer is initiated at one boundary of T-DNA, but ends at variable sites. The single strand is converted into a double strand and integrated into the plant genome. Genes within the T-DNA transform the plant cell, and cause it to produce particular opines (derivatives of arginine). Genes in the Ti plasmid allow *Agrobacterium* to metabolize the opines produced by the transformed plant cell. T-DNA has been used to develop vectors for transferring genes into plant cells.

Endogenous sequences may become amplified in cultured cells. Exposure to methotrexate leads to the accumulation of cells that have additional copies of the *dhfr* gene. The copies may be carried as extrachromosomal arrays in the form of double-minute "chromosomes," or they may be integrated into the genome at the site of one of the *dhfr* alleles. Double-minute chromosomes are unstable, and disappear from the cell line rapidly in the absence of selective pressure. The amplified copies may originate by additional cycles of replication that are associated with recombination events.

New sequences of DNA may be introduced into a cultured cell by transfection or into an animal egg by microinjection. The foreign sequences may become integrated into the genome, often as large tandem arrays. The array appears to be inherited as a unit in a cultured cell. The sites of integration appear to be random. A transgenic animal arises when the integration event occurs into a genome that enters the germ-cell lineage. A transgene or transgenic array is inherited in Mendelian manner, but the copy number and activity of the gene(s) may change in the progeny. Often a transgene responds to tissue- and temporal regulation in a manner that resembles the endogenous gene. Using conditions that promote homologous recombination, an inactive sequence can be used to replace a functional gene, thus creating a null locus. Transgenic mice can be obtained by injecting recipient blastocysts with ES cells that carry transfected DNA.

References

18.2 The mating pathway is triggered by pheromone-receptor interactions

rev Nasmyth, K. (1982). Molecular genetics of yeast mating type. Ann. Rev. Genet. 16. *439-500.*

ref Bender, A. and Sprague, G. F., Jr. (1986). Yeast peptide pheromones, a-factor and alpha-factor, activate a common response mechanism in their target cells. Cell 47, *929-937.*

18.3 The mating response activates a G protein

rev Dohlman, H. G. and Thorner, J. W. (2001). Regulation of G protein-initiated signal transduction in yeast: paradigms and principles. Ann. Rev. Biochem. 70, 703-754.

Kurjan, J. (1992). Pheromone response in yeast. Ann. Rev. Biochem. 61, 1097-1129.

Kurjan, J. (1993). The pheromone response pathway in *S. cerevisiae.* Ann. Rev. Genet. 27, 147-179.

ref Bender, A. and Sprague, G. F., Jr. (1986). Yeast peptide pheromones, a-factor and alpha-factor, activate a common response mechanism in their target cells. Cell 47, 929-937.

18.4 The signal is passed to a kinase cascade

rev Dohlman, H. G. and Thorner, J. W. (2001). Regulation of G protein-initiated signal transduction in yeast: paradigms and principles. Ann. Rev. Biochem. 70, 703-754.

ref Butty, A. C, Pryciak, P. M., Huang, L. S., Herskowitz, I., and Peter, M. (1998). The role of Far1p in linking the heterotrimeric G protein to polarity establishment proteins during yeast mating. Science 282, 1511-1516.

Choi, K.-Y. et al. (1994). Ste5 tethers multiple protein kinases in the MAP kinase cascade required for mating in *S. cerevisiae.* Cell 78, 499-512.

Whiteway, M. S., Wu, C, Leeuw, T., Clark, K., Fourest-Lieuvin, A., Thomas, D. Y., and Leberer, E. (1995). Association of the yeast pheromone response G protein beta gamma subunits with the MAP kinase scaffold Ste5p. Science 269, 1572-1575.

18.5 Yeast can switch silent and active loci for mating type

ref Hicks, J., Strathern, J. N., and Herskowitz, I. (1977). The cassette model of mating type interconversion. In DNA Insertion Elements, Eds. A. Bukhari, J. Shapiro, and S. Adhya, Cold Spring Harbor Laboratory, 457-462.

18.6 The *MAT* locus codes for regulator proteins

rev Nasmyth, K. and Shore, D. (1987). Transcriptional regulation in the yeast life cycle. Science 237, 1162-1170.

18.7 Silent cassettes at HML and HMR are repressed

rev Laurenson, P. and Rine, J. (1992). Silencers, silencing, and heritable transcriptional states. Microbiol. Rev. 56, 543-560.

18.8 Unidirectional transposition is initiated by the recipient MAT locus

ref Strathern, J. N. et al. (1982). Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the MAT locus. Cell 31, 183-192.

18.9 Regulation of HO expression controls switching

ref Bobola, N. et al. (1996). Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. Cell 84, 699-709.

Maxon, M. E. and Herskowitz, I. (2001). Aship is a site-specific DNA-binding protein that actively represses transcription. Proc. Nat. Acad. Sci. USA 98, 1495-1500.

18.10 Trypanosomes switch the VSG frequently during infection

rev Barry, J. D., McCulloch, R., and Barry, R. (2001). Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. Adv. Parasitol. 49, 1-70.

Boothroyd, J. C. (1985). Antigenic variation in African trypanosomes. Ann. Rev. Immunol. 39, 475-502.

Donelson, J. E. and Rice-Ficht, A. C. (1985). Molecular biology of trypanosome antigenic variation. Microbiol. Rev. 49, 107-125.

18.11 New VSG sequences are generated by gene switching

ref McCulloch, R. and Barry, J. D. (1999). A role for RAD51 and homologous recombination in Trypanosoma brucei antigenic variation. Genes Dev. 13, 2875-2888.

18.12 VSG genes have an unusual structure

rev Borst, P. (1986). Discontinuous transcription and antigenic variation in trypanosomes. Ann. Rev. Biochem. 55, 701-732.

ref Navarro, M. and Gull, K. (2001). A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei.* Nature 414 759-763.

18.13 The bacterial Ti plasmid causes crown gall disease in plants

rev Winans, S. C. (1993). Two-way chemical signaling in *Agrobacterium*-plant interactions. Microbiol. Rev. 56, 12-31.

18.15 Transfer of T-DNA resembles bacterial conjugation

rev Zambryski, P. (1988). Basic processes underlying Agrobacterium-mediated DNA transfer to plant cells. Ann. Rev. Genet. 22, 1-30.

Zambryski, P. (1989). Agrobacterium-plant cell DNA transfer. In Mobile DNA, Eds. Berg, D. E. and Howe, M. M., ASM, Washington D. C. 309-334.

18.16 DNA amplification generates extra gene copies

rev Schimke, R. T. (1981). Chromosomal and extrachromosomal localization of amplified DHFR genes in cultured mammalian cells. Cold Spring Harbor Symp. Quant. Biol. 45, 785-797.

Stark, G. R. and Wahl, G. M. (1984). Gene amplification. Ann. Rev. Biochem. 53, 447-491.

18.17 Transfection introduces exogenous DNA into cells

rev Pellicer, A. (1980). Altering genotype and phenotype by DNA-mediated gene transfer. Science 209, 1414-1422.

18.18 Genes can be injected into animal eggs

exp Brinster, R. and Palmiter, R. (2002). Transgenic mice: Expression of Foreign Genes in Animals (www.ergito.com/lookup.jsp?expt= brinster)

18.19 ES cells can be incorporated into embryonic mice

rev Jaenisch, R. (1988). Transgenic animals. Science 240, 1468-1474.

18.20 Gene targeting allows genes to be replaced or knocked out

exp Capecchi, M. (2002). Gene Targeting: Altering the Genome in Mice (www.ergito.com/lookup.jsp?expt = capecchi)

rev Capecchi, M. R. (1989). Altering the genome by homologous recombination. Science 244, 1288-1292.

ref Spradling, A. C. and Rubin, G. M. (1982). Transposition of cloned P elements into *Drosophila* germline chromosomes. Science 218, 341-353.

Chapter 19

Chromosomes

19.1 Introduction

A general principle is evident in the organization of all cellular genetic material. It exists as a compact mass, confined to a limited volume; and its various activities, such as replication and transcription, must be accomplished within this space. The organization of this material must accommodate transitions between inactive and active states.

The condensed state of nucleic acid results from its binding to basic proteins. The positive charges of these proteins neutralize the negative charges of the nucleic acid. The structure of the nucleoprotein complex is determined by the interactions of the proteins with the DNA (or RNA).

A common problem is presented by the packaging of DNA into phages and viruses, into bacterial cells and eukaryotic nuclei. The length of the DNA as an extended molecule would vastly exceed the dimensions of the compartment that contains it. The DNA (or in the case of some viruses, the RNA) must be compressed exceedingly tightly to fit into the space available. *So in contrast with the customary picture of DNA as an extended double helix, structural deformation of DNA to bend or fold it into a more compact form is the rule rather than exception.*

The magnitude of the discrepancy between the length of the nucleic acid and the size of its compartment is evident from the examples summarized in **Figure 19.1.** For bacteriophages and for eukaryotic viruses, the nucleic acid genome, whether single-stranded or double-stranded DNA or RNA, effectively fills the container (which can be rod-like or spherical).

For bacteria or for eukaryotic cell compartments, the discrepancy is hard to calculate exactly, because the DNA is contained in a compact

DNA is highly compressed in all types of genomes

Compartment	Shape	Dimensions	Type of Nucleic Acid	Length
TMV	filament	0.008 x 0.3 mm	1 single-stranded RNA	2 mm = 6.4 kb
Phage fd	filament	0.006 x 0.85 mm	1 single-stranded DNA	2 mm = 6.0 kb
Adenovirus	icosahedron	0.07 mm diameter	1 double-stranded DNA	11 mm = 35.0 kb
Phage T4	icosahedron	0.065 x 0.10 mm	1 double-stranded DNA	55 mm = 170.0 kb
E. coli	cylinder	1.7 x 0.65 mm	1 double-stranded DNA	1.3 mm = 4.2×10^3 kb
Mitochondrion (human)	oblate spheroid	3.0 x 0.5 mm	~10 identical double-stranded DNAs	50 mm = 16.0 kb
Nucleus (human)	spheroid	6 mm diameter	46 chromosomes of double-stranded DNA	1.8 m = 6×10^6 kb

By Book_Crazy [IND]

Figure 19.1 The length of nucleic acid is much greater than the dimensions of the surrounding compartment.

area that occupies only part of the compartment. The genetic material is seen in the form of the nucleoid in bacteria and as the mass of chromatin in eukaryotic nuclei at interphase (between divisions).

The density of DNA in these compartments is high. In a bacterium it is ~ 10 mg/ml, in a eukaryotic nucleus it is ~100 mg/ml, and in the phage T4 head it is >500 mg/ml. Such a concentration in solution would be equivalent to a gel of great viscosity. We do not entirely understand the physiological implications, for example, what effect this has upon the ability of proteins to find their binding sites on DNA.

The packaging of chromatin is flexible; it changes during the eukaryotic cell cycle. At the time of division (mitosis or meiosis), the genetic material becomes even more tightly packaged, and individual chromosomes become recognizable.

The overall compression of the DNA can be described by the packing ratio, the length of the DNA divided by the length of the unit that contains it. For example, the smallest human chromosome contains ~4.6 X 10^7 bp of DNA (~10 times the genome size of the bacterium *E. coli*). This is equivalent to 14,000 μm (= 1.4 cm) of extended DNA. At the most condensed moment of mitosis, the chromosome is ~2 μm long. So the packing ratio of DNA in the chromosome can be as great as 7000.

Packing ratios cannot be established with such certainty for the more amorphous overall structures of the bacterial nucleoid or eukaryotic chromatin. However, the usual reckoning is that mitotic chromosomes are likely to be 5-10X more tightly packaged than interphase chromatin, which therefore has a typical packing ratio of 1000-2000.

A major unanswered question concerns the *specificity* of packaging. Is the DNA folded into *a particular* pattern, or is it different in each individual copy of the genome? How does the pattern of packaging change when a segment of DNA is replicated or transcribed?

19.2 Viral genomes are packaged into their coats

Key Concepts

- The length of DNA that can be incorporated into a virus is limited by the structure of the head shell.
- Nucleic acid within the head shell is extremely condensed.
- Filamentous RNA viruses condense the RNA genome as they assemble the head shell around it.
- Spherical DNA viruses insert the DNA into a preassembled protein shell.

From the perspective of packaging the *individual* sequence, there is an important difference between a cellular genome and a virus. The cellular genome is essentially indefinite in size; the number and location of individual sequences can be changed by duplication, deletion, and rearrangement. So it requires a *generalized* method for packaging its DNA, insensitive to the total content or distribution of sequences. By contrast, two restrictions define the needs of a virus. The amount of nucleic acid to be packaged is *predetermined* by the size of the genome. And it must all fit within a coat assembled from a protein or proteins coded by the viral genes.

A virus particle is deceptively simple in its superficial appearance. The nucleic acid genome is contained within a capsid, a symmetrical or quasi-symmetrical structure assembled from one or only a few proteins.

Attached to the capsid, or incorporated into it, are other structures, assembled from distinct proteins, and necessary for infection of the host cell.

The virus particle is tightly constructed. The internal volume of the capsid is rarely much greater than the volume of the nucleic acid it must hold. The difference is usually less than twofold, and often the internal volume is barely larger than the nucleic acid.

In its most extreme form, the restriction that the capsid must be assembled from proteins coded by the virus means that the entire shell is constructed from a single type of subunit. The rules for assembly of identical subunits into closed structures restrict the capsid to one of two types. The protein subunits stack sequentially in a helical array to form a *filamentous* or rod-like shape. Or they form a pseudospherical shell, a type of structure that conforms to a polyhedron with icosahedral symmetry. Some viral capsids are assembled from more than a single type of protein subunit, but although this extends the exact types of structures that can be formed, viral capsids still all conform to the general classes of quasi-crystalline filaments or icosahedrons.

There are two types of solution to the problem of how to construct a capsid that contains nucleic acid:

- The protein shell can be assembled around the nucleic acid, condensing the DNA or RNA by protein-nucleic acid interactions during the process of assembly.
- Or the capsid can be constructed from its component(s) in the form of an empty shell, into which the nucleic acid must be inserted, being condensed as it enters.

The capsid is assembled around the genome for single-stranded RNA viruses. The principle of assembly is that *the position of the RNA within the capsid is determined directly by its binding to the proteins of the shell.* The best characterized example is TMV (tobacco mosaic virus). Assembly starts at a duplex hairpin that lies within the RNA sequence. From this nucleation center, it proceeds bidirectionally along the RNA, until reaching the ends. The unit of the capsid is a two-layer disk, each layer containing 17 identical protein subunits. The disk is a circular structure, which forms a helix as it interacts with the RNA. At the nucleation center, the RNA hairpin inserts into the central hole in the disk, and the disk changes conformation into a helical structure that surrounds the RNA. Then further disks are added, each disk pulling a new stretch of RNA into its central hole. The RNA becomes coiled in a helical array on the inside of the protein shell, as illustrated in **Figure 19.2.**

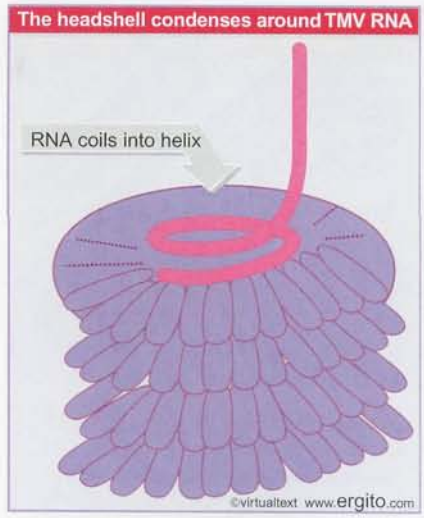


Figure 19.2 A helical path for TMV RNA is created by the stacking of protein subunits in the virion.

The spherical capsids of DNA viruses are assembled in a different way, as best characterized for the phages lambda and T4. In each case, an empty headshell is assembled from a small set of proteins. *Then the duplex genome is inserted into the head,* accompanied by a structural change in the capsid.

Figure 19.3 summarizes the assembly of lambda. It starts with a small headshell that contains a protein "core". This is converted to an empty headshell of more distinct shape. Then DNA packaging begins, the headshell expands in size though remaining the same shape, and finally the full head is sealed by addition of the tail.

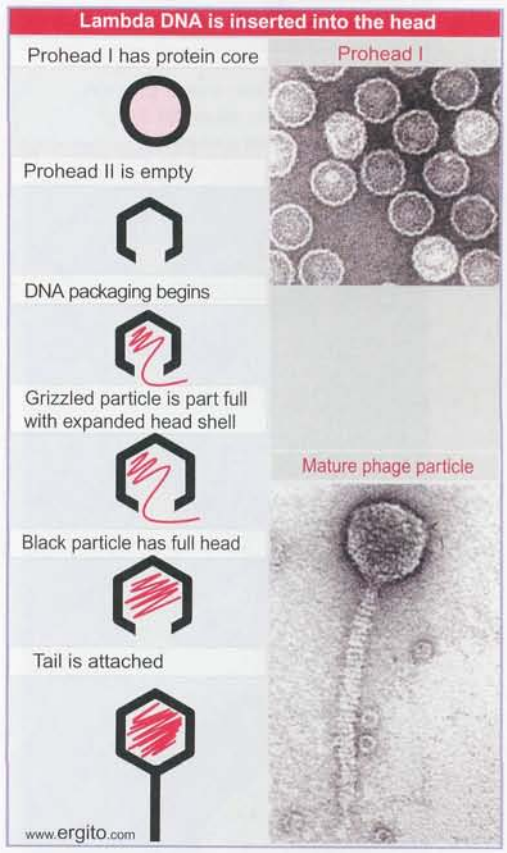


Figure 19.3 Maturation of phage lambda passes through several stages. The empty head changes shape and expands when it becomes filled with DNA. The electron micrographs show the particles at the start and end of the maturation pathway. Photographs kindly provided by A. F. Howatson.

Now a double-stranded DNA considered over short distances is a fairly rigid rod. Yet it must be compressed into a compact structure to fit within the capsid. We should like to know whether packaging involves a smooth coiling of the DNA into the head or requires abrupt bends.

Inserting DNA into a phage head involves two types of reaction: translocation and condensation. Both are energetically unfavorable.

Translocation is an active process in which the DNA is driven into the head by an ATP-dependent mechanism. A common mechanism is used for many viruses that replicate by a rolling circle mechanism to generate long tails that contain multimers of the viral genome. The best characterized

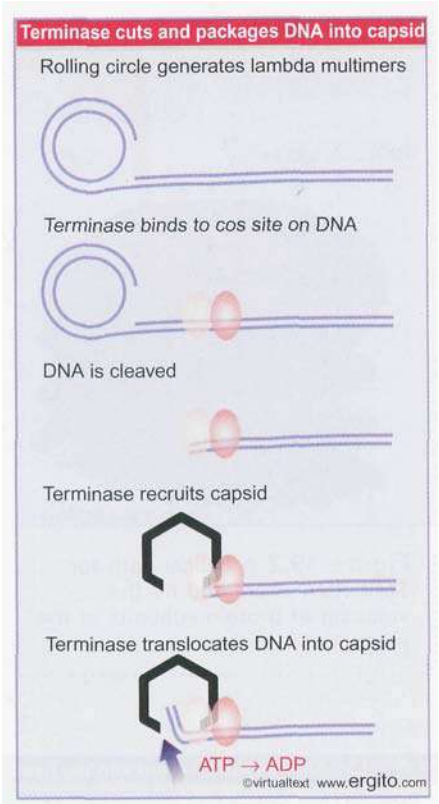


Figure 19.4 Terminase protein binds to specific sites on a multimer of virus genomes generated by rolling circle replication. Terminase cuts the DNA and binds to an empty virus capsid. It uses energy from hydrolysis of ATP to insert the DNA into the capsid.

example is phage lambda. The genome is packaged into the empty capsid by the terminase enzyme. **Figure 19.4** summarizes the process.

The terminase was first recognized for its role in generating the ends of the linear phage DNA by cleaving at *cos* sites. (The name *cos* reflects the fact that it generates cohesive ends that have complementary single-stranded tails.) The phage genome codes two subunits that make up the terminase. One subunit binds to a *cos* site; then it is joined by the other subunit, which cuts the DNA. The terminase assembles into a hetero-oligomer in a complex that also includes IHF (the integration host factor coded by the bacterial genome). It then binds to an empty capsid and uses ATP hydrolysis to power translocation along the DNA. The translocation drives the DNA into the empty capsid.

Another method of packaging uses a structural component of the phage. In the *B. subtilis* phage ϕ29, the motor that inserts the DNA into the phage head is the structure that connects the head to the tail. It functions as a rotary motor, where the motor action effects the linear translocation of the DAN into the phage head. The same motor is used to eject the DNA from the phage head when it infects a bacterium.

Little is known about the mechanism of condensation into an empty capsid, except that the capsid contains "internal proteins" as well as DNA. One possibility is that they provide some sort of "scaffolding" onto which the DNA condenses. (This would be a counterpart to the use of the proteins of the shell in the plant RNA viruses.)

How specific is the packaging? It cannot depend on particular sequences, because deletions, insertions, and substitutions all fail to interfere with the assembly process. The relationship between DNA and the headshell has been investigated directly by determining which regions of the DNA can be chemically crosslinked to the proteins of the capsid. The surprising answer is that all regions of the DNA are more or less equally susceptible. This probably means that when DNA is inserted into the head, it follows a general rule for condensing, but the pattern is not determined by particular sequences.

These varying mechanisms of virus assembly all accomplish the same end: packaging a single DNA or RNA molecule into the capsid. However, some viruses have genomes that consist of multiple nucleic acid molecules. Reovirus contains ten double-stranded RNA segments, all of which must be packaged into the capsid. Specific sorting sequences in the segments may be required to ensure that the assembly process selects one copy of each different molecule in order to collect a complete set of genetic information. In the simpler case of phage ϕ6, which packages three different segments of double-stranded RNA into one capsid, the RNA segments must bind in a specific order; as each is incorporated into the capsid, it triggers a change in the conformation of the capsid that creates binding sites for the next segment.

Some plant viruses are multipartite: their genomes consist of segments, each of which is packaged into a *different* capsid. An example is alfalfa mosaic virus (AMV), which has four different single-stranded RNAs, each packaged independently into a coat comprising the same protein subunit. A successful infection depends on the entry of one of each type into the cell.

The four components of AMV exist as particles of different sizes. This means that the same capsid protein can package each RNA into its own characteristic particle. This is a departure from the packaging of a unique length of nucleic acid into a capsid of fixed shape.

The assembly pathway of viruses whose capsids have only one authentic form may be diverted by mutations that cause the formation of aberrant monster particles in which the head is longer than usual. These mutations show that a capsid protein(s) has an intrinsic ability to assemble into a particular type of structure, but the exact size and shape may vary. Some of the mutations occur in genes that code for assembly factors, which are

needed for head formation, but are not themselves part of the headshell. Such ancillary proteins limit the options of the capsid protein so that it assembles only along the desired pathway. Comparable proteins are employed in the assembly of cellular chromatin (see *20 Nucleosomes).*

19.3 The bacterial genome is a nucleoid

Key Concepts

* The bacterial nucleoid is ~80% DNA by mass and can be unfolded by agents that act on RNA or protein.
* The proteins that are responsible for condensing the DNA have not been identified.

Although bacteria do not display structures with the distinct morphological features of eukaryotic chromosomes, their genomes nonetheless are organized into definite bodies. The genetic material can be seen as a fairly compact clump or series of clumps that occupies about a third of the volume of the cell. **Figure 19.5** displays a thin section through a bacterium in which this nucleoid is evident.

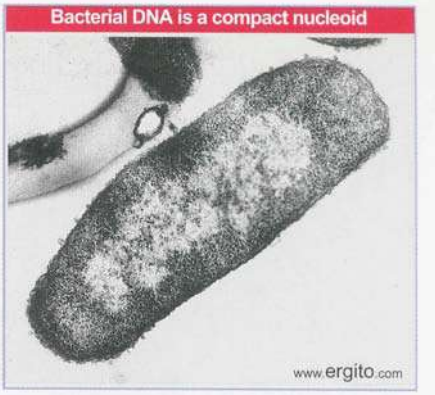


Figure 19.5 A thin section shows the bacterial nucleoid as a compact mass in the center of the cell. Photograph kindly provided by Jack Griffith.

When *E. coli* cells are lysed, fibers are released in the form of loops attached to the broken envelope of the cell. As can be seen from **Figure 19.6,** the DNA of these loops is not found in the extended form of a free duplex, but is compacted by association with proteins.

Several DNA-binding proteins with a superficial resemblance to eukaryotic chromosomal proteins have been isolated in *E. coli.* What criteria should we apply for deciding whether a DNA-binding protein plays a structural role in the nucleoid? It should be present in sufficient quantities to bind throughout the genome. And mutations in its gene should cause some disruption of structure or of functions associated with genome survival (for example, segregation to daughter cells). None of the candidate proteins yet satisfies the genetic conditions.

Figure 19.6 The nucleoid spills out of a lysed *E. coli* cell in the form of loops of a fiber. Photograph kindly provided by Jack Griffith.

Protein HU is a dimer that condenses DNA, possibly wrapping it into a bead-like structure. It is related to IHF (integration host factor), another dimer, which has a structural role in building a protein complex in specialized recombination reactions. Null mutations in either of the genes coding for the subunits of HU (*hupA,B*) have little effect, but loss of both functions causes a cold-sensitive phenotype and some loss of superhelicity in DNA. These results raise the possibility that HU plays some general role in nucleoid condensation.

Protein H1 (also known as H-NS) binds DNA, interacting preferentially with sequences that are bent. Mutations in its gene have turned up in a variety of guises (*osmZ, bglY, pilG*), each identified as an apparent regulator of a different system. These results probably reflect the effect that H1 has on the local topology of DNA, with effects upon gene expression that depend upon the particular promoter.

We might expect that the absence of a protein required for nucleoid structure would have serious effects upon viability. Why then are the effects of deletions in the genes for proteins HU and H1 relatively restricted? One explanation is that these proteins are *redundant,* that any one can substitute for the others, so that deletions of *all* of them would be necessary to interfere seriously with nucleoid structure. Another possibility is that we have yet to identify the proteins responsible for the major features of nucleoid integrity.

The nucleoid can be isolated directly in the form of a very rapidly sedimenting complex, consisting of ~80% DNA by mass. (The analogous

complexes in eukaryotes have ~50% DNA by mass; see next section.) It can be unfolded by treatment with reagents that act on RNA or protein. The possible role of proteins in stabilizing its structure is evident. The role of RNA has been quite refractory to analysis.

19.4 The bacterial genome is supercoiled

Key Concepts

- The nucleoid has ~100 independently negatively supercoiled domains.
- The average density of supercoiling is — 1 turn/100 bp.

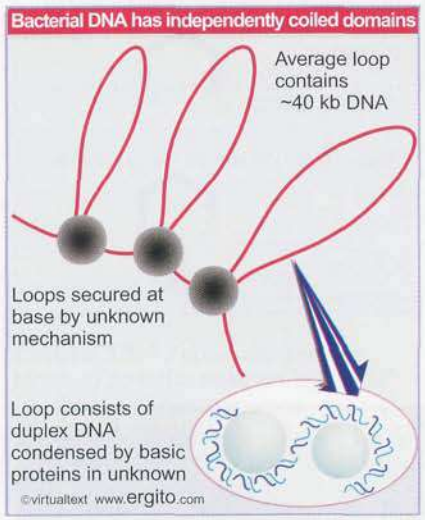


Figure 19.7 The bacterial genome consists of a large number of loops of duplex DNA (in the form of a fiber), each secured at the base to form an independent structural domain.

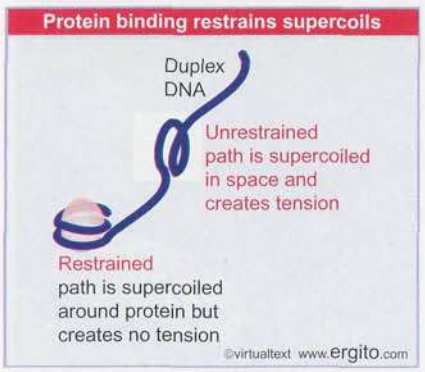


Figure 19.8 An unrestrained supercoil in the DNA path creates tension, but no tension is transmitted along DNA when a supercoil is restrained by protein binding.

The DNA of the compact body isolated *in vitro* behaves as a closed duplex structure, as judged by its response to ethidium bromide. This small molecule intercalates between base pairs to generate *positive* superhelical turns in "closed" circular DNA molecules, that is, molecules in which both strands have covalent integrity. (In "open" circular molecules, which contain a nick in one strand, or with linear molecules, the DNA can rotate freely in response to the intercalation, thus relieving the tension.)

In a natural closed DNA that is *negatively* supercoiled, the intercalation of ethidium bromide first removes the negative supercoils and then introduces positive supercoils. The amount of ethidium bromide needed to achieve zero supercoiling is a measure of the original density of negative supercoils.

Some nicks occur in the compact nucleoid during its isolation; they can also be generated by limited treatment with DNAase. But this does not abolish the ability of ethidium bromide to introduce positive supercoils. This capacity of the genome to retain its response to ethidium bromide in the face of nicking means that it must have many independent chromosomal domains; *the supercoiling in each domain is not affected by events in the other domains.*

This autonomy suggests that the structure of the bacterial chromosome has the general organization depicted diagrammatically in **Figure 19.7.** Each domain consists of a loop of DNA, the ends of which are secured in some (unknown) way that does not allow rotational events to propagate from one domain to another. There are ~100 such domains per genome; each consists of ~40 kb (13 μm) of DNA, organized into some more compact fiber whose structure has yet to be characterized.

The existence of separate domains could permit different degrees of supercoiling to be maintained in different regions of the genome. This could be relevant in considering the different susceptibilities of particular bacterial promoters to supercoiling (see *9.15 Supercoiling is an important feature of transcription).*

Supercoiling in the genome can in principle take either of two forms, as summarized in **Figure 19.8:**

- If a supercoiled DNA is free, its path is *unrestrained,* and negative supercoils generate a state of torsional tension that is transmitted freely along the DNA within a domain. It can be relieved by unwinding the double helix, as described in *15.12 Supercoiling affects the structure of DNA.* The DNA is in a dynamic equilibrium between the states of tension and unwinding.
- Supercoiling can be *restrained* if proteins are bound to the DNA to hold it in a particular three-dimensional configuration. In this case, the supercoils are represented by the path the DNA follows in its

fixed association with the proteins. The energy of interaction between the proteins and the supercoiled DNA stabilizes the nucleic acid, so that no tension is transmitted along the molecule.

Are the supercoils in *E. coli* DNA restrained *in vivo* or is the double helix subject to the torsional tension characteristic of free DNA? Measurements of supercoiling *in vitro* encounter the difficulty that restraining proteins may have been lost during isolation. Various approaches suggest that DNA is under torsional stress *in vivo*.

One approach is to measure the effect of nicking the DNA. Unrestrained supercoils are released by nicking, but restrained supercoils are unaffected. Nicking releases ~50% of the overall supercoiling, suggesting that about half of the supercoiling is transmitted as tension along DNA, but the other half is absorbed by protein binding.

Another approach uses the crosslinking reagent psoralen, which binds more readily to DNA when it is under torsional tension. The reaction of psoralen with *E. coli* DNA *in vivo* corresponds to an average density of one negative superhelical turn / 200 bp (σ = -0.05).

We can also examine the ability of cells to form alternative DNA structures; for example, to generate cruciforms at palindromic sequences. From the change in linking number that is required to drive such reactions, it is possible to calculate the original supercoiling density. This approach suggests an average density of σ = -0.025, or one negative superhelical turn / 100 base pairs.

So supercoils *do* create torsional tension *in vivo*. There may be variation about an average level, and although the precise range of densities is difficult to measure, it is clear that the level is sufficient to exert significant effects on DNA structure, for example, in assisting melting in particular regions such as origins or promoters.

Many of the important features of the structure of the compact nucleoid remain to be established. What is the specificity with which domains are constructed—do the same sequences always lie at the same relative locations, or can the contents of individual domains shift? How is the integrity of the domain maintained? Biochemical analysis by itself is unable to answer these questions fully, but if it is possible to devise suitable selective techniques, the properties of structural mutants should lead to a molecular analysis of nucleoid construction.

19.5 Eukaryotic DNA has loops and domains attached to a scaffold

Key Concepts

- DNA of interphase chromatin is negatively supercoiled into independent domains of ~85 kb.
- Metaphase chromosomes have a protein scaffold to which the loops of supercoiled DNA are attached.

Interphase chromatin is a tangled mass occupying a large part of the nuclear volume, in contrast with the highly organized and reproducible ultrastructure of mitotic chromosomes. What controls the distribution of interphase chromatin within the nucleus?

Some indirect evidence on its nature is provided by the isolation of the genome as a single, compact body. Using the same technique that was developed for isolating the bacterial nucleoid (see previous section), nuclei can be lysed on top of a sucrose gradient. This releases the genome in a form that can be collected by centrifugation. As isolated

Figure 19.9 Histone-depleted chromosomes consist of a protein scaffold to which loops of DNA are anchored. Photograph kindly provided by Ulrich K. Laemmli.

from *D. melanogaster,* it can be visualized as a compactly folded fiber (10 nm in diameter), consisting of DNA bound to proteins.

Supercoiling measured by the response to ethidium bromide corresponds to about one negative supercoil / 200 bp. These supercoils can be removed by nicking with DNAase, although the DNA remains in the form of the 10 nm fiber. This suggests that the supercoiling is caused by the arrangement of the fiber in space, and represents the existing torsion.

Full relaxation of the supercoils requires one nick / 85 kb, identifying the average length of "closed" DNA. This region could comprise a loop or domain similar in nature to those identified in the bacterial genome. Loops can be seen directly when the majority of proteins are extracted from mitotic chromosomes. The resulting complex consists of the DNA associated with ~8% of the original protein content. As seen in **Figure 19.9,** the protein-depleted chromosomes take the form of a central scaffold surrounded by a halo of DNA.

The metaphase scaffold consists of a dense network of fibers. Threads of DNA emanate from the scaffold, apparently as loops of average length 10-30 μm (30-90 kb). The DNA can be digested without affecting the integrity of the scaffold, which consists of a set of specific proteins. This suggests a form of organization in which loops of DNA of ~60 kb are anchored in a central proteinaceous scaffold.

The appearance of the scaffold resembles a mitotic pair of sister chromatids. The sister scaffolds usually are tightly connected, but sometimes are separate, joined only by a few fibers. Could this be the structure responsible for maintaining the shape of the mitotic chromosomes? Could it be generated by bringing together the protein components that usually secure the bases of loops in interphase chromatin?

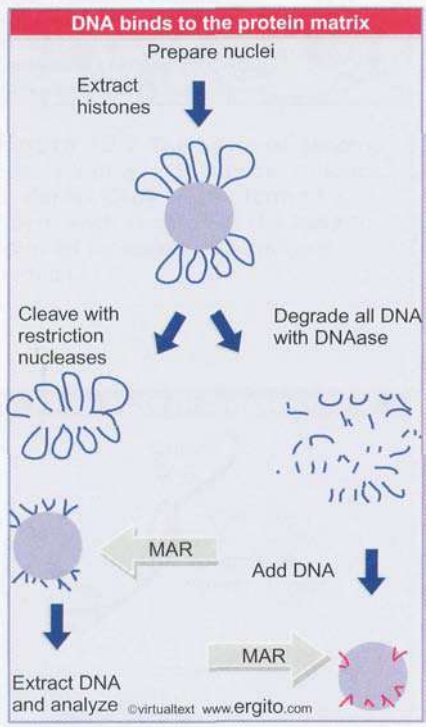


Figure 19.10 Matrix-associated regions may be identified by characterizing the DNA retained by the matrix isolated *in vivo* or by identifying the fragments that can bind to the matrix from which all DNA has been removed.

19.6 Specific sequences attach DNA to an interphase matrix

Key Concepts

- DNA is attached to the nuclear matrix at specific sequences called MARs or SARs.
- The MARs are A·T-rich but do not have any specific consensus sequence.

Is DNA attached to the scaffold via specific sequences? DNA sites attached to proteinaceous structures in interphase nuclei are called MAR (matrix attachment regions); they are sometimes also called *SAR* (scaffold attachment regions). The nature of the structure in interphase cells to which they are connected is not clear. Chromatin often appears to be attached to a matrix, and there have been many suggestions that this attachment is necessary for transcription or replication. When nuclei are depleted of proteins, the DNA extrudes as loops from a residual proteinaceous structure. However, attempts to relate the proteins found in this preparation to structural elements of intact cells have not been successful.

Are particular DNA regions associated with this matrix? *In vivo* and *in vitro* approaches are summarized in **Figure 19.10.** Both start by isolating the matrix as a crude nuclear preparation containing chromatin and nuclear proteins. Different treatments can then be used to characterize DNA in the matrix or to identify DNA able to attach to it.

To analyze the existing MAR, the chromosomal loops can be decondensed by extracting the proteins. Removal of the DNA loops by treatment with restriction nucleases leaves only the (presumptive) *in vivo* MAR sequences attached to the matrix.

The complementary approach is to remove *all* the DNA from the matrix by treatment with DNAase; then isolated fragments of DNA can be tested for their ability to bind to the matrix *in vitro*.

The same sequences should be associated with the matrix *in vivo* or *in vitro*. Once a potential MAR has been identified, the size of the minimal region needed for association *in vitro* can be determined by deletions. We can also then identify proteins that bind to the MAR sequences.

A surprising feature is the lack of conservation of sequence in MAR fragments. They are usually ~70% A·T-rich, but otherwise lack any consensus sequences. However, other interesting sequences often are in the DNA stretch containing the MAR. *Cis*-acting sites that regulate transcription are common. And a recognition site for topoisomerase II is usually present in the MAR. It is therefore possible that an MAR serves more than one function, providing a site for attachment to the matrix, but also containing other sites at which topological changes in DNA are effected.

What is the relationship between the chromosome scaffold of dividing cells and the matrix of interphase cells? Are the same DNA sequences attached to both structures? In several cases, the same DNA fragments that are found with the nuclear matrix *in vivo* can be retrieved from the metaphase scaffold. And fragments that contain MAR sequences can bind to a metaphase scaffold. It therefore seems likely that DNA contains a single type of attachment site, which in interphase cells is connected to the nuclear matrix, and in mitotic cells is connected to the chromosome scaffold.

The nuclear matrix and chromosome scaffold consist of different proteins, although there are some common components. Topoisomerase II is a prominent component of the chromosome scaffold, and is a constituent of the nuclear matrix, suggesting that the control of topology is important in both cases.

19.7 Chromatin is divided into euchromatin and heterochromatin

Key Concepts

- Individual chromosomes can be seen only during mitosis.
- During interphase, the general mass of chromatin is in the form of euchromatin, which is less tightly packed than mitotic chromosomes.
- Regions of heterochromatin remain densely packed throughout interphase.

Each chromosome contains a single, very long duplex of DNA. This explains why chromosome replication is semiconservative like the individual DNA molecule. (This would not necessarily be the case if a chromosome carried many independent molecules of DNA.) The single duplex of DNA is folded into a fiber that runs continuously throughout the chromosome. *So in accounting for interphase chromatin and mitotic chromosome structure, we have to explain the packaging of a single, exceedingly long molecule of DNA into a form in*

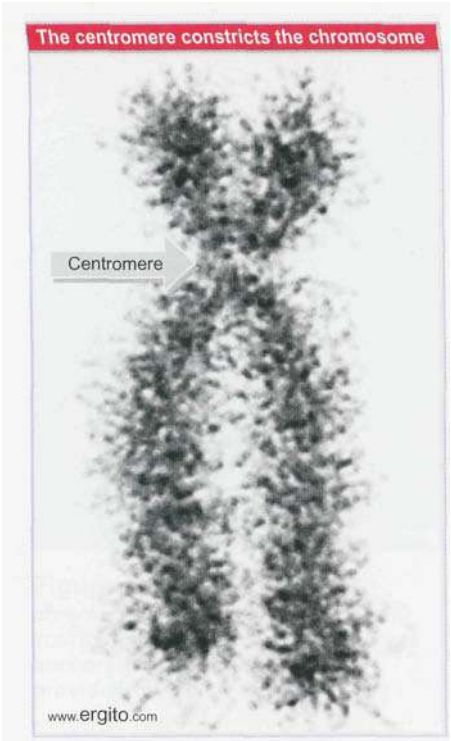


Figure 19.11 The sister chromatids of a mitotic pair each consist of a fiber (~30 nm in diameter) compactly folded into the chromosome. Photograph kindly provided by E. J. DuPraw.

which it can be transcribed and replicated, and can become cyclically more and less compressed.

Individual eukaryotic chromosomes come into the limelight for a brief period, during the act of cell division. Only then can each be seen as a compact unit. **Figure 19.11** is an electron micrograph of a sister chromatid pair, captured at metaphase. (The sister chromatids are daughter chromosomes produced by the previous replication event, still joined together at this stage of mitosis.) Each consists of a fiber with a diameter of ~30 nm and a nubbly appearance. The DNA is 5–1 OX more condensed in chromosomes than in interphase chromatin.

During most of the life cycle of the eukaryotic cell, however, its genetic material occupies an area of the nucleus in which individual chromosomes cannot be distinguished. The structure of the interphase chromatin does not change visibly between divisions. No disruption is evident during the period of replication, when the amount of chromatin doubles. Chromatin is fibrillar, although the overall configuration of the fiber in space is hard to discern in detail. The fiber itself, however, is similar or identical to that of the mitotic chromosomes.

Chromatin can be divided into two types of material, which can be seen in the nuclear section of **Figure 19.12:**

- In most regions, the fibers are much less densely packed than in the mitotic chromosome. This material is called euchromatin. It has a relatively dispersed appearance in the nucleus, and occupies most of the nuclear region in Figure 19.12.
- Some regions of chromatin are very densely packed with fibers, displaying a condition comparable to that of the chromosome at mitosis. This material is called heterochromatin. It is typically found at centromeres, but occurs at other locations also. It passes through the cell cycle with relatively little change in its degree of condensation. It forms a series of discrete clumps in Figure 19.12, but often the various heterochromatic regions aggregate into a densely staining chromocenter. (This description applies to regions that are always heterochromatic, called constitutive heterochromatin; in addition, there is another sort of heterochromatin, called facultative heterochromatin, in which regions of euchromatin are converted to a heterochromatic state).

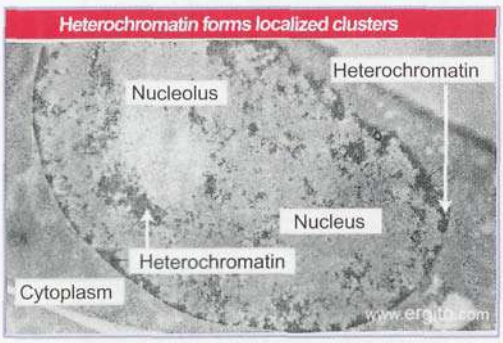


Figure 19.12 A thin section through a nucleus stained with Feulgen shows heterochromatin as compact regions clustered near the nucleolus and nuclear membrane. Photograph kindly provided by Edmund Puvion.

The same fibers run continuously between euchromatin and heterochromatin, which implies that these states represent different degrees of condensation of the genetic material. In the same way, euchromatic regions exist in different states of condensation during interphase and during mitosis. So the genetic material is organized in a manner that permits alternative states to be maintained side by side in chromatin, and allows cyclical changes to occur in the packaging of euchromatin between interphase and division. We discuss the molecular basis for these states in *23 Controlling chromatin structure.*

The structural condition of the genetic material is correlated with its activity. The common features of constitutive heterochromatin are

- It is permanently condensed.
- It often consists of multiple repeats of a few sequences of DNA that are not transcribed.
- The density of genes in this region is very much reduced compared with heterochromatin; and genes that are translocated into or near it are often inactivated.
- Probably resulting from the condensed state, it replicates late in S phase and has a reduced frequency of genetic recombination.

We have some molecular markers for changes in the properties of the DNA and protein components (see *23.15 Heterochromatin depends on interactions with histories*). They include reduced acetylation of his-

tone proteins, increased methylation of one histone protein, and hypermethylation of cytidine bases in DNA. These molecular changes cause the condensation of the material, which is responsible for its inactivity.

Although active genes are contained within euchromatin, only a small minority of the sequences in euchromatin are transcribed at any time. So location in euchromatin is *necessary* for gene expression, but is not *sufficient* for it.

19.8 Chromosomes have banding patterns

Key Concepts

- Certain staining techniques cause the chromosomes to have the appearance of a series of striations called G-bands.
- The bands are lower in G·C content than the interbands.
- Genes are concentrated in the G·C-rich interbands.

Because of the diffuse state of chromatin, we cannot directly determine *the specificity of its organization. But we can ask whether* the structure of the (mitotic) chromosome is ordered. Do particular sequences always lie at particular sites, or is the folding of the fiber into the overall structure a more random event?

At the level of the chromosome, each member of the complement has a different and reproducible ultrastructure. When subjected to certain treatments and then stained with the chemical dye Giemsa, chromosomes generate a series of G-bands. **Figure 19.13** presents an example of the human set.

Until the development of this technique, chromosomes could be distinguished only by their overall size and the relative location of the centromere. Now each chromosome can be identified by its characteristic banding pattern. This pattern allows translocations from one chromosome to another to be identified by comparison with the original diploid set. **Figure 19.14** shows a diagram of the bands of the human X chromosome. The bands are large structures, each ~10^7 bp of DNA, which could include many hundreds of genes.

The banding technique is of enormous practical use, but the mechanism of banding remains a mystery. All that is certain is that the dye stains untreated chromosomes more or less uniformly. So the generation of bands depends on a variety of treatments that change the response of the chromosome (presumably by extracting the component that binds the stain from the nonbanded regions). But similar bands can be generated by a variety of treatments.

The only known feature that distinguishes bands from interbands is that the bands have a lower G-C content than the interbands. This is a peculiar result. If there are ~10 bands on a large chromosome with a total content of ~ 100 Mb, this means that the chromosome is divided into regions of ~5 Mb in length that alternate between low G-C (band) and high G-C (interband) content. There is a tendency for genes (as identified by hybridization with mRNAs) to be located in the interband regions. **All** of this argues for some long-range sequence-dependent organization.

The human genome sequence confirms the basic observation. **Figure 19.15** shows that there are distinct fluctuations in G-C content when the genome is divided into small tranches. The average of 41% G-C is common to mammalian genomes. There are regions as low as 30% or as high as 65%. When longer tranches are examined, there is less variation. The average length of regions with >43% G-C is 200-250 kb. This makes it clear that the band/interband structure does not represent homogeneous segments that alternate in G-C content, although the bands do contain a

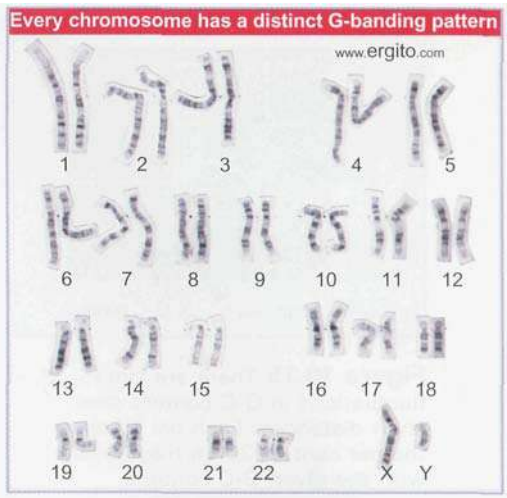


Figure 19.13 G-banding generates a characteristic lateral series of bands in each member of the chromosome set. Photograph kindly provided by Lisa *Shaffer.*

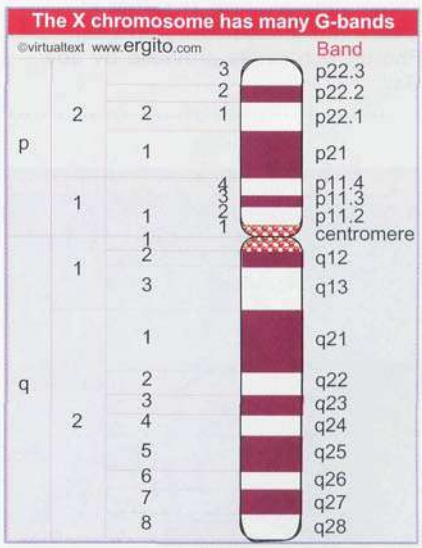


Figure 19.14 The human X chromosome can be divided into distinct regions by its banding pattern. The short arm is *p* and the long arm is *q*; each arm is divided into larger regions that are further subdivided. This map shows a low resolution structure; at higher resolution, some bands are further subdivided into smaller bands and interbands, e.g. *p*21 is divided into *p*21.1, *p*21.2, and *p*21.3.

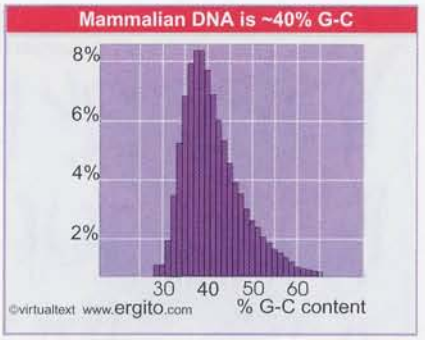


Figure 19.15 There are large fluctuations in G-C content over short distances. Each bar shows the per cent of 20 kb fragments with the given G-C content.

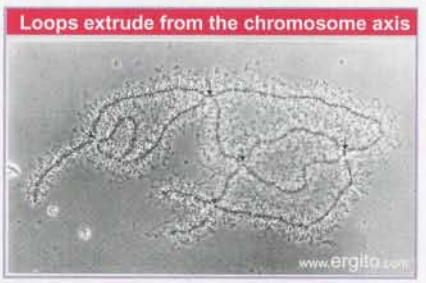


Figure 19.16 A lampbrush chromosome is a meiotic bivalent in which the two pairs of sister chromatids are held together at chiasmata (indicated by arrows). Photograph kindly provided by Joe Gall.

Figure 19.17 A lampbrush chromosome loop is surrounded by a matrix of ribonucleoprotein. Photograph kindly provided by Oscar Miller.

higher content of low GC segments. Genes are concentrated in regions of higher G-C content. We have yet to understand how the G-C content affects chromosome structure.

19.9 Lampbrush chromosomes are extended

Key Concepts

- Sites of gene expression on lampbrush chromosomes show loops that are extended from the chromosomal axis.

It would be extremely useful to visualize gene expression in its natural state, to see what structural changes are associated with transcription. The compression of DNA in chromatin, coupled with the difficulty of identifying particular genes within it, makes it impossible to visualize the transcription of individual active genes.

Gene expression can be visualized directly in certain unusual situations, in which the chromosomes are found in a highly extended form that allows individual loci (or groups of loci) to be distinguished. Lateral differentiation of structure is evident in many chromosomes when they first appear for meiosis. At this stage, the chromosomes resemble a series of beads on a string. The beads are densely staining granules, properly known as chromomeres. However, usually there is little gene expression at meiosis, and it is not practical to use this material to identify the activities of individual genes. But an exceptional situation that allows the material to be examined is presented by lampbrush chromosomes, which have been best characterized in certain amphibians.

Lampbrush chromosomes are formed during an unusually extended meiosis, which can last up to several months! During this period, the chromosomes are held in a stretched-out form in which they can be visualized in the light microscope. Later during meiosis, the chromosomes revert to their usual compact size. So the extended state essentially proffers an unfolded version of the normal condition of the chromosome.

The lampbrush chromosomes are meiotic bivalents, each consisting of two pairs of sister chromatids. **Figure 19.16** shows an example in which the sister chromatid pairs have mostly separated so that they are held together only by chiasmata. Each sister chromatid pair forms a series of ellipsoidal chromomeres, ~1-2 μm in diameter, which are connected by a very fine thread. This thread contains the two sister duplexes of DNA, and runs continuously along the chromosome, through the chromomeres.

The lengths of the individual lampbrush chromosomes in the newt *Notophthalmus viridescens* range from 400-800 μm, compared with the range of 15-20 μm seen later in meiosis. So the lampbrush chromosomes are ~30 times less tightly packed. The total length of the entire lampbrush chromosome set is 5-6 mm, organized into ~5000 chromomeres.

The lampbrush chromosomes take their name from the lateral loops that extrude from the chromomeres at certain positions. (These resemble a lampbrush, an extinct object.) The loops extend in pairs, one from each sister chromatid. The loops are continuous with the axial thread which suggests that they represent chromosomal material extruded from its more compact organization in the chromomere.

The loops are surrounded by a matrix of ribonucleoproteins. These contain nascent RNA chains. Often a transcription unit can be defined by the increase in the length of the RNP moving around the loop. An example is shown in **Figure 19.17.**

So the loop is an extruded segment of DNA that is being actively transcribed. In some cases, loops corresponding to particular genes have been identified. Then the structure of the transcribed gene, and the nature of the product, can be scrutinized *in situ*.

19.10 Polytene chromosomes form bands

Key Concepts

* Polytene chromosomes of Dipterans have a series of bands that can be used as a cytological map.

The interphase nuclei of some tissues of the larvae of Dipteran flies contain chromosomes that are greatly enlarged relative to their usual condition. They possess both increased diameter and greater length. **Figure 19.18** shows an example of a chromosome set from the salivary gland of *D. melanogaster*. They are called polytene chromosomes.

Each member of the polytene set consists of a visible series of bands (more properly, but rarely, described as chromomeres). The bands range in size from the largest with a breadth of ~0.5 μm to the smallest of ~0.05 μm. (The smallest can be distinguished only under the electron microscope.) The bands contain most of the mass of DNA and stain intensely with appropriate reagents. The regions between them stain more lightly and are called interbands. There are ~5000 bands in the *D. melanogaster* set.

The centromeres of all four chromosomes of *D. melanogaster* aggregate to form a chromocenter that consists largely of heterochromatin (in the male it includes the entire Y chromosome). Allowing for this, ~75% of the haploid DNA set is organized into alternating bands and interbands. The length of the chromosome set is ~2000 μm. The DNA in extended form would stretch for ~40,000 μm, so the packing ratio is ~20. This demonstrates vividly the extension of the genetic material relative to the usual states of interphase chromatin or mitotic chromosomes.

What is the structure of these giant chromosomes? Each is produced by the successive replications of a synapsed diploid pair. The replicas do not separate, but remain attached to each other in their extended state. At the start of the process, each synapsed pair has a DNA content of 2C (where C represents the DNA content of the individual chromosome). Then this doubles up to 9 times, at its maximum giving a content of 1024C. The number of doublings is different in the various tissues of the *D. melanogaster* larva.

Each chromosome can be visualized as a large number of parallel fibers running longitudinally, tightly condensed in the bands, less condensed in the interbands. Probably each fiber represents a single (C) haploid chromosome. This gives rise to the name polytene. The degree of polyteny is the number of haploid chromosomes contained in the giant chromosome.

The banding pattern is characteristic for each strain of *Drosophila*. The constant number and linear arrangement of the bands was first noted in the 1930s, when it was realized that they form a *cytological map* of the chromosomes. Rearrangements—such as deletions, inversions, or duplications—result in alterations of the order of bands.

The linear array of bands can be equated with the linear array of genes. So genetic rearrangements, as seen in a linkage map, can be correlated with structural rearrangements of the cytological map. Ultimately, a particular mutation can be located in a particular band. Since

Figure 19.18 The polytene chromosomes of *D. melanogaster* form an alternating series of bands and interbands. Photograph kindly provided by Jose Bonner.

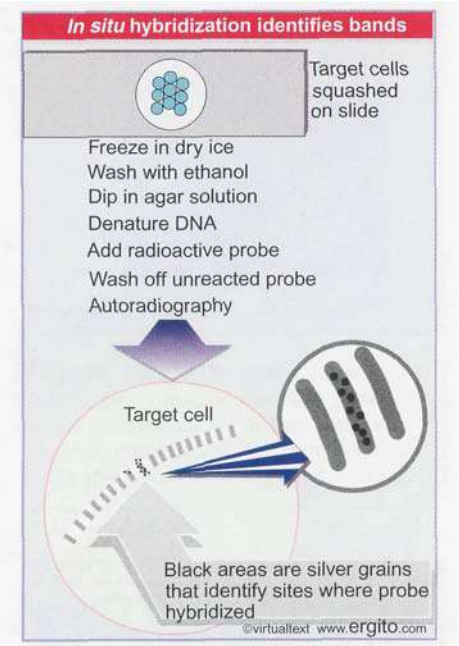


Figure 19.19 Individual bands containing particular genes can be identified by *in situ* hybridization.

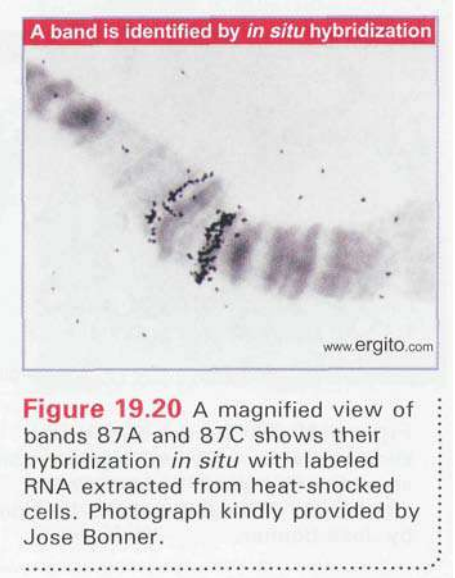


Figure 19.20 A magnified view of bands 87A and 87C shows their hybridization *in situ* with labeled RNA extracted from heat-shocked cells. Photograph kindly provided by Jose Bonner.

the total number of genes in *D. melanogaster* exceeds the number of bands, there are probably multiple genes in most or all bands.

The positions of particular genes on the cytological map can be determined directly by the technique of *in situ* hybridization. The protocol is summarized in **Figure 19.19.** A radioactive probe representing a gene (most often a labeled cDNA clone derived from the mRNA) is hybridized with the denatured DNA of the polytene chromosomes *in situ*. Autoradiography identifies the position or positions of the corresponding genes by the superimposition of grains at a particular band or bands. An example is shown in **Figure 19.20**. With this type of technique at hand, it is possible to determine directly the band within which a particular sequence lies.

19.11 Polytene chromosomes expand at sites of gene expression

Key Concepts

- Bands that are sites of gene expression on polytene chromosomes expand to give "puffs".

One of the intriguing features of the polytene chromosomes is that active sites can be visualized. Some of the bands pass transiently through an expanded state in which they appear like a puff on the chromosome, when chromosomal material is extruded from the axis. An example of some very large puffs (called Balbiani rings) is shown in **Figure 19.21.**

What is the nature of the puff? It consists of a region in which the chromosome fibers unwind from their usual state of packing in the band. The fibers remain continuous with those in the chromosome axis. Puffs usually emanate from single bands, although when they are very large, as typified by the Balbiani rings, the swelling may be so extensive as to obscure the underlying array of bands.

The pattern of puffs is related to gene expression. During larval development, puffs appear and regress in a definite, tissue-specific pattern. A characteristic pattern of puffs is found in each tissue at any given time. Puffs are induced by the hormone ecdysone that controls *Drosophila* development. Some puffs are induced directly by the hormone; others are induced indirectly by the products of earlier puffs.

The puffs are *sites where RNA is being synthesized.* The accepted view of puffing has been that expansion of the band is a consequence of the need to relax its structure in order to synthesize RNA. Puffing has therefore been viewed as a consequence of transcription. A puff can be generated by a single active gene. The sites of puffing differ from ordinary bands in accumulating additional proteins, which include RNA polymerase II and other proteins associated with transcription.

The features displayed by lampbrush and polytene chromosomes suggest a general conclusion. In order to be transcribed, the genetic material is dispersed from its usual more tightly packed state. The question to keep in mind is whether this dispersion at the gross level of the chromosome mimics the events that occur at the molecular level within the mass of ordinary interphase euchromatin.

Do the bands of a polytene chromosome have a functional significance, that is, does each band correspond to some type of genetic unit? You might think that the answer would be immediately evident from the sequence of the fly genome, since by mapping interbands to the sequence it should be possible to determine whether a band has any fixed type of identity. However, so far, no pattern has been found that identifies a functional significance for the bands.

19.12 The eukaryotic chromosome is a segregation device

Key Concepts

- A eukaryotic chromosome is held on the mitotic spindle by the attachment of microtubules to the kinetochore that forms in its centromeric region.
- Centromeres often have heterochromatin that is rich in satellite DNA sequences.

During mitosis, the sister chromatids move to opposite poles of the cell. Their movement depends on the attachment of the chromosome to microtubules, which are connected at their other end to the poles. (The microtubules comprise a cellular filamentous system, reorganized at mitosis so that they connect the chromosomes to the poles of the cell.) The sites in the two regions where microtubule ends are organized—in the vicinity of the centrioles at the poles and at the chromosomes—are called MTOCs (microtubule organizing centers).

Figure 19.22 illustrates the separation of sister chromatids as mitosis proceeds from metaphase to telophase. The region of the chromosome that is responsible for its segregation at mitosis and meiosis is called the centromere. The centromeric region on each sister chromatid is pulled by microtubules to the opposite pole. Opposing this motive force, "glue" proteins called cohesins hold the sister chromatids together. Initially the sister chromatids separate at their centromeres, and then they are released completely from one another during anaphase when the cohesins are degraded (the cohesins are discussed in more detail in *29.19 Cohesins hold sister chromatids together).*

The centromere is pulled toward the pole during mitosis, and the attached chromosome is dragged along behind, as it were. The chromosome therefore provides a device for attaching a large number of genes to the apparatus for division. It contains the site at which the sister chromatids are held together prior to the separation of the individual chromosomes. This shows as a constricted region connecting all four chromosome arms, as in the photograph of Figure 19.11, which shows the sister chromatids at the metaphase stage of mitosis.

The centromere is essential for segregation, as shown by the behavior of chromosomes that have been broken. A single break generates one piece that retains the centromere, and another, an acentric fragment, that lacks it. The acentric fragment does not become attached to the mitotic spindle; and as a result it fails to be included in either of the daughter nuclei.

(When chromosome movement relies on discrete centromeres, there can be *only* one centromere per chromosome. When translocations generate chromosomes with more than one centromere, aberrant structures form at mitosis, since the two centromeres on the *same* sister chromatid can be pulled toward different poles, breaking the chromosome.

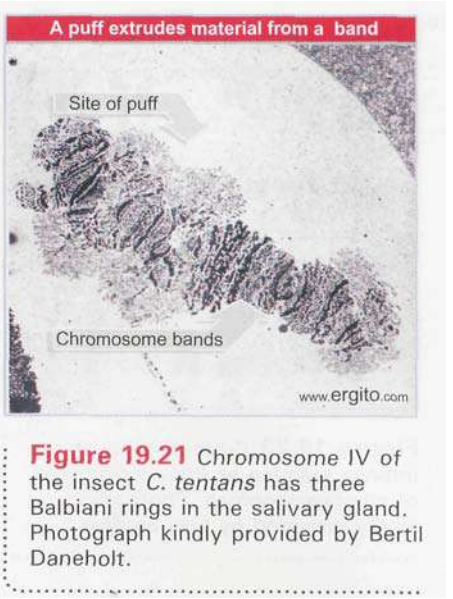


Figure 19.21 Chromosome IV of the insect *C. tentans* has three Balbiani rings in the salivary gland. Photograph kindly provided by Bertil Daneholt.

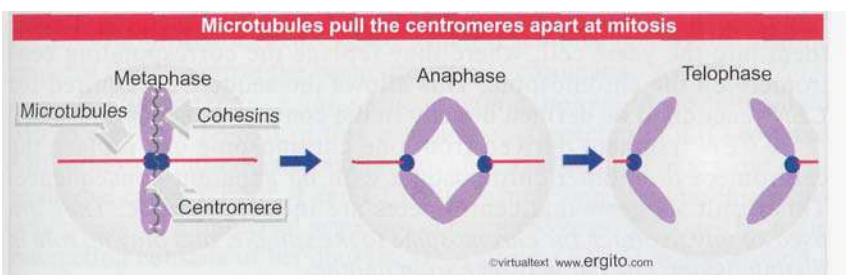


Figure 19.22 Chromosomes are pulled to the poles via microtubules that attach at the centromeres. The sister chromatids are held together until anaphase by glue proteins (cohesins). The centromere is shown here in the middle of the chromosome (metacentric), but can be located anywhere along its length, including close to the end (acrocentric) and at the end (telocentric).

By Book_Crazy [IND]

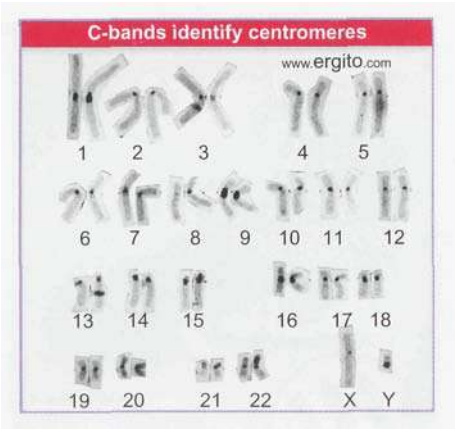


Figure 19.23 C-banding generates intense staining at the centromeres of all chromosomes. Photograph kindly provided by Lisa Shaffer.

However, in some species the centromeres are "diffuse," which creates a different situation. Only discrete centromeres have been analyzed at the molecular level.)

The regions flanking the centromere often are rich in satellite DNA sequences and display a considerable amount of heterochromatin. Because the entire chromosome is condensed, centromeric heterochromatin is not immediately evident in mitotic chromosomes. However, it can be visualized by a technique that generates C-bands. In the example of **Figure 19.23,** all the centromeres show as darkly staining regions. Although it is common, heterochromatin cannot be identified around *every* known centromere, which suggests that it is unlikely to be essential for the division mechanism.

The region of the chromosome at which the centromere forms is defined by DNA sequences (although the sequences have been defined in only a very small number of cases). The centromeric DNA binds specific proteins that are responsible for establishing the structure that attaches the chromosome to the microtubules. This structure is called the kinetochore. It is a darkly staining fibrous object of diameter or length ~400 nm. The kinetochore provides the MTOC on a chromosome. **Figure 19.24** shows the hierarchy of organization that connects centromeric DNA to the microtubules. Proteins bound to the centromeric DNA bind other proteins that bind to microtubules (see *19.14 The centromere binds a protein complex).*

19.13 Centromeres have short DNA sequences in *S. cerevisiae*

Key Concepts

- **CEN** elements are identified in S. *cerevisiae* by the ability to allow a plasmid to segregate accurately at mitosis.
- CEN elements consists of short conserved sequences CDE-I and CDE-III that flank the A·T-rich region CDE-II.

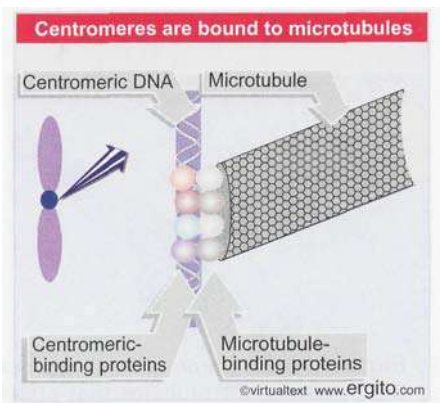


Figure 19.24 The centromere is identified by a DNA sequence that binds specific proteins. These proteins do not themselves bind to microtubules, but establish the site at which the microtubule-binding proteins in turn bind.

If a centromeric sequence of DNA is responsible for segregation, any molecule of DNA possessing this sequence should move properly at cell division, while any DNA lacking it will fail to segregate. This prediction has been used to isolate centromeric DNA in the yeast, *S. cerevisiae*. Yeast chromosomes do not display visible kinetochores comparable to those of higher eukaryotes, but otherwise divide at mitosis and segregate at meiosis by the same mechanisms.

Genetic engineering has produced plasmids of yeast that are replicated like chromosomal sequences (see *13.6 Replication origins can be isolated in yeast).* However, they are unstable at mitosis and meiosis, disappearing from a majority of the cells because they segregate erratically. Fragments of chromosomal DNA containing centromeres have been isolated by their ability to confer mitotic stability on these plasmids.

A *CEN* fragment is identified as the minimal sequence that can confer stability upon such a plasmid. Another way to characterize the function of such sequences is to modify them *in vitro* and then reintroduce them into the yeast cell, where they replace the corresponding centromere on the chromosome. This allows the sequences required for *CEN* function to be defined directly in the context of the chromosome.

A *CEN* fragment derived from one chromosome can replace the centromere of another chromosome with no apparent consequence. This result suggests that centromeres are interchangeable. *They are used simply to attach the chromosome to the spindle, and play no role in distinguishing one chromosome from another.*

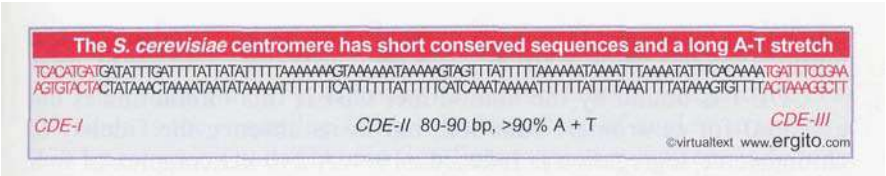


Figure 19.25 Three conserved regions can be identified by the sequence homologies between yeast *CEN* elements.

The sequences required for centromeric function fall within a stretch of ~120 bp. The centromeric region is packaged into a nuclease-resistant structure, and it binds a single microtubule. We may therefore look to the *S. cerevisiae* centromeric region to identify proteins that bind centromeric DNA and proteins that connect the chromosome to the spindle.

Three types of sequence element may be distinguished in the *CEN* region, as summarized in **Figure 19.25:**

- CDE-I is a sequence of 9 bp that is conserved with minor variations at the left boundary of all centromeres.
- CDE-II is a >90% A·T-rich sequence of 80-90 bp found in all centromeres; its function could depend on its length rather than exact sequence. Its constitution is reminiscent of some short tandemly repeated (satellite) DNAs (see *4.12 Arthropod satellites have very short identical repeats).* Its base composition may cause some characteristic distortions of the DNA double helical structure.
- CDE-III is an 11 bp sequence highly conserved at the right boundary of all centromeres. Sequences on either side of the element are less well conserved, and may also be needed for centromeric function. (CDE-III could be longer than 11 bp if it turns out that the flanking sequences are essential.)

Mutations in CDE-I or CDE-II reduce but do not inactivate centromere function, but point mutations in the central CCG of CDE-III completely inactivate the centromere.

19.14 The centromere binds a protein complex

Key Concepts

- A specialized protein complex that is an alternative to the usual chromatin structure is formed at CDE-II.
- The CBF3 protein complex that binds to CDE-III is essential for centromeric function.
- The proteins that connect these two complexes may provide the connection to microtubules.

Can we identify proteins that are necessary for the function of CEN sequences? There are several genes in which mutations affect chromosome segregation, and whose proteins are localized at centromeres. The contributions of these proteins to the centromeric structure are summarized in **Figure 19.26.**

A specialized chromatin structure is built by binding the CDE-II region to a protein called Cse4p, which resembles one of the histone proteins that comprise the basic subunits of chromatin (see *23.15 Heterochromatin depends on interactions with histones).* A protein called Mif2p may also be part of this complex or connected to it. Cse4p and Mif2p have counterparts that are localized at higher eukaryotic centromeres, called CENP-A and CENP-C, which suggests that this interaction may be a universal aspect of centromere construction. The basic interaction consists of bending the DNA of the CDE-II region around a

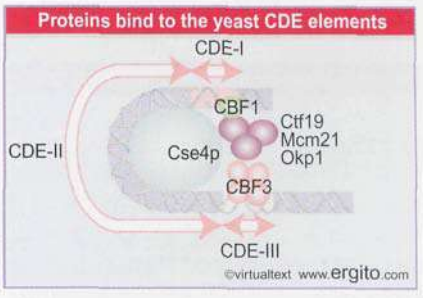


Figure 19.26 The DNA at CDE-II is wound around a protein aggregate including Cse4p, CDE-III is bound to CBF3, and CDE-I is bound to CBF1. These proteins are connected by the group of Ctf19, Mcm21, and Okp1.

protein aggregate; the reaction is probably assisted by the occurrence of intrinsic bending in the CDE-II sequence.

CDE-I is bound by the homodimer CBF1; this interaction is not essential for centromere function, but in its absence the fidelity of chromosome segregation is reduced ~10X. A 240 kD complex of four proteins, called CBF3, binds to CDE-III. This interaction is essential for centromeric function.

The proteins bound at CDE-I and CDE-III are connected to each other and also to the protein structure bound at CDE-II by another group of proteins (Ctf19, Mcm21, Okp1). The connection to the microtubule may be made by this complex.

The overall model suggests that the complex is localized at the centromere by a protein structure that resembles the normal building block of chromatin (the nucleosome). The bending of DNA at this structure allows proteins bound to the flanking elements to become part of a single complex. Some components of the complex (possibly not those that bind directly to DNA) link the centromere to the microtubule. The construction of kinetochores probably follows a similar pattern, and uses related components, in a wide variety of organisms.

19.15 Centromeres may contain repetitious DNA

Key Concepts

- Centromeres in higher eukaryotic chromosomes contain large amounts of repetitious DNA.
- The function of the repetitious DNA is not known.

The length of DNA required for centromeric function is often quite long. (The short, discrete, elements of *S. cerevisiae* may be an exception to the general rule.) In those cases where we can equate specific DNA sequences with the centromeric region, they usually include repetitive sequences.

S. cerevisiae is the only case so far in which centromeric DNA can be identified by its ability to confer stability on plasmids. However, a related approach has been used with the yeast *S. pombe*. This has only 3 chromosomes, and the region containing each centromere has been identified by deleting most of the sequences of each chromosome to create a stable minichromosome. This approach locates the centromeres within regions of 40-100 kb that consist largely or entirely of repetitious DNA. It is not clear how much of each of these rather long regions is required for chromosome segregation at mitosis and meiosis.

Attempts to localize centromeric functions in *Drosophila* chromosomes suggest that they are dispersed in a large region, consisting of 200-600 kb. The large size of this type of centromere suggests that it is likely to contain several separate specialized functions, including sequences required for kinetochore assembly, sister chromatid pairing, etc.

The size of the centromere in *Arabidopsis* is comparable. Each of the 5 chromosomes has a centromeric region in which recombination is very largely suppressed. This region occupies >500 kb. Clearly it includes the centromere, but we have no direct information as to how much of it is required. There are expressed genes within these regions, which casts some doubt on whether the entire region is part of the centromere. At the center of the region is a series of 180 bp repeats; this is

the type of structure generally associated with centromeres. It is too early to say how these structures relate to centromeric function.

The primary motif comprising the heterochromatin of primate centromeres is the a satellite DNA, which consists of tandem arrays of a 170 bp repeating unit. There is significant variation between individual repeats, although those at any centromere tend to be better related to one another than to members of the family in other locations. It is clear that the sequences required for centromeric function reside within the blocks of α satellite DNA, but it is not clear whether the a satellite sequences themselves provide this function, or whether other sequences are embedded within the a satellite arrays.

19.16 Telomeres have simple repeating sequences

Key Concepts

- The telomere is required for the stability of the chromosome end.
- A telomere consists of a simple repeat where a C + A-rich strand has the sequence $C_{>1}(A/T)_{1\text{-}4}$

Another essential feature in all chromosomes is the telomere, which "seals" the end. We know that the telomere must be a special structure, because chromosome ends generated by breakage are "sticky" and tend to react with other chromosomes, whereas natural ends are stable.

We can apply two criteria in identifying a telomeric sequence:

- It must lie at the end of a chromosome (or, at least, at the end of an authentic linear DNA molecule).
- It must confer stability on a linear molecule.

The problem of finding a system that offers an assay for function again has been brought to the molecular level by using yeast. All the plasmids that survive in yeast (by virtue of possessing *ARS* and *CEN* elements) are circular DNA molecules. Linear plasmids are unstable (because they are degraded). Could an authentic telomeric DNA sequence confer stability on a linear plasmid? Fragments from yeast DNA that prove to be located at chromosome ends can be identified by such an assay. And a region from the end of a known natural linear DNA molecule—the extrachromosomal rDNA of *Tetrahymena*—is able to render a yeast plasmid stable in linear form.

Telomeric sequences have been characterized from a wide range of lower and higher eukaryotes. The same type of sequence is found in plants and man, so the construction of the telomere seems to follow a universal principle. Each telomere consists of a long series of short, tandemly repeated sequences. There may be 100-1000 repeats, depending on the organism.

All telomeric sequences can be written in the general form where $\boldsymbol{n} > 1$ and m is 1-4. **Figure 19.27** shows a generic example. One unusual property of the telomeric sequence is the extension of the G-T-rich strand, usually for 14-16 bases as a single strand. The G-tail is probably generated because there is a specific limited degradation of the C-A-rich strand. th ther

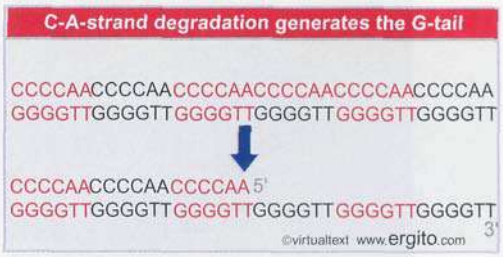


Figure 19.27 A typical telomere has a simple repeating structure with a G-T-rich strand that extends beyond the C-A-rich strand. The G-tail is generated by a limited degradation of the C-A-rich strand.

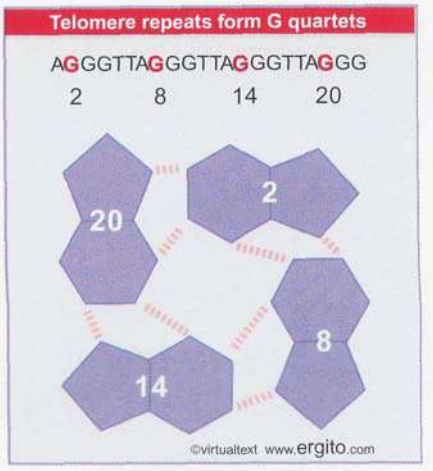


Figure 19.28 The crystal structure of a short repeating sequence from the human telomere forms 3 stacked G quartets. The top quartet contains the first G from each repeating unit. This is stacked above quartets that contains the second G (G3, G9, G15, G21) and the third G (G4, G10, G16, G22).

cell clone is followed, the telomere grows longer by 7-10 bp (1-2 repeats) per generation. Even more revealing is the fate of ciliate telomeres introduced into yeast. After replication in yeast, *yeast telomeric repeats are added onto the ends of the* Tetrahymena *repeats*.

Addition of telomeric repeats to the end of the chromosome in every replication cycle could solve the difficulty of replicating linear DNA molecules discussed in *13.8 The ends of linear DNA are a problem for replication.* The addition of repeats by *de novo* synthesis would counteract the loss of repeats resulting from failure to replicate up to the end of the chromosome. Extension and shortening would be in dynamic equilibrium.

If telomeres are continually being lengthened (and shortened), their exact sequence may be irrelevant. All that is required is for the end to be recognized as a suitable substrate for addition. This explains how the ciliate telomere functions in yeast.

19.17 Telomeres seal the chromosome ends

Key Concepts

- The protein TRF2 catalyzes a reaction in which the 3' repeating unit of the G+T-rich strand forms a loop by displacing its homologue in an upstream region of the telomere.

Isolated telomeric fragments do not behave as though they contain single-stranded DNA; instead they show aberrant electrophoretic mobility and other properties.

Guanine bases have an unusual capacity to associate with one another. The single-stranded G-rich tail of the telomere can form "quartets" of G residues. Each quartet contains 4 guanines that hydrogen bond with one another to form a planar structure. Each guanine comes from the corresponding position in a successive TTAGGG repeating unit. **Figure 19.28** shows an organization based on a recent crystal structure. The quartet that is illustrated represents an association between the first guanine in each repeating unit. It is stacked on top of another quartet that has the same organization, but is formed from the second guanine in each repeating unit. A series of quartets could be stacked like this in a helical manner. While the formation of this structure attests to the unusual properties of the G-rich sequence *in vitro*, it does not of course demonstrate whether the quartet forms *in vivo*.

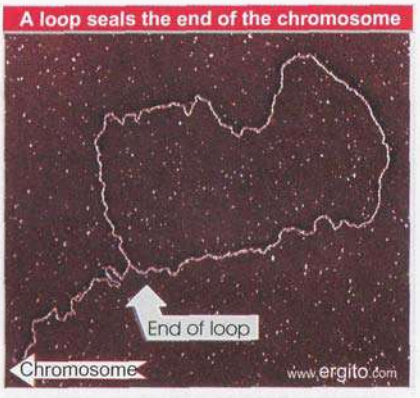


Figure 19.29 A loop forms at the end of chromosomal DNA. Photograph kindly provided by Jack Griffith.

What feature of the telomere is responsible for the stability of the chromosome end? **Figure 19.29** shows that a loop of DNA forms at the telomere. The absence of any free end may be the crucial feature that stabilizes the end of the chromosome. The average length of the loop in animal cells is 5-10 kb.

Figure 19.30 shows that the loop is formed when the 3' single-stranded end of the telomere $(TTAGGG)_n$ displaces the same sequence in an upstream region of the telomere. This converts the duplex region into a structure like a D-loop, where a series of TTAGGG repeats are displaced to form a single-stranded region, and the tail of the telomere is paired with the homologous strand.

The reaction is catalyzed by the telomere-binding protein TRF2, which together with other proteins forms a complex that stabilizes the chromosome ends. Its importance in protecting the ends is indicated by the fact the deletion of TRF2 causes chromosome rearrangements to occur.

19.18 Telomeres are synthesized by a ribonucleoprotein enzyme

Key Concepts

- Telomerase uses the 3′–OH of the G + T telomeric strand to prime synthesis of tandem TTGGGG repeats.
- The RNA component of telomerase has a sequence that pairs with the C + A-rich repeats.
- One of the protein subunits is a reverse transcriptase that uses the RNA as template to synthesis the G+T-rich sequence.

The telomere has two functions:

- One is to protect the chromosome end. Any other DNA end—for example, the end generated by a double strand break—becomes a target for repair systems. The cell has to be able to distinguish the telomere.
- The second is to allow the telomere to be extended. Otherwise it would become shorter with each replication cycle (because replication cannot start at the very end).

Proteins that bind to the telomere provide the solution for both problems. In yeast, different sets of proteins solve each problem, but both are bound to the telomere via the same protein, Cdc13:

- The Stn1 protein protects against degradation (specifically against any extension of the degradation of the C-A-strand that generates the G-tail).
- A telomerase enzyme extends the C-A-rich strand. Its activity is influenced by two proteins that have ancillary roles, such as controlling the length of the extension.

The telomerase uses the 3′–OH of the G + T telomeric strand as a primer for synthesis of tandem TTGGGG repeats. Only dGTP and dTTP are needed for the activity. The telomerase is a large ribonucleoprotein that consists of a templating RNA and a protein with catalytic activity (coded by *EST2*). The short RNA component (159 bases long in *Tetrahymena*, 192 bases long in *Euplotes*) includes a sequence of 15-22 bases that is identical to two repeats of the C-rich repeating sequence. This RNA provides the template for synthesizing the G-rich repeating sequence. The protein component of the telomerase is a catalytic subunit that can act only upon the RNA template provided by the nucleic acid component.

Figure 19.31 shows the action of telomerase. The enzyme progresses discontinuously: the template RNA is positioned on the DNA primer, several nucleotides are added to the primer, and then the enzyme translocates to begin again. The telomerase is a specialized example of a reverse transcriptase, an enzyme that synthesizes a DNA sequence using an RNA template (see *17.4 Viral DNA is generated by reverse transcription*). We do not know how the complementary (C-A-rich) strand of the telomere is assembled, but we may speculate that it could be synthesized by using the 3′–OH of a terminal G-T hairpin as a primer for DNA synthesis.

Telomerase synthesizes the individual repeats that are added to the chromosome ends, but does not itself control the number of repeats. Other proteins are involved in determining the length of the telomere. They can be identified by the *EST1* and *EST3* mutants in yeast that have altered telomere lengths. These proteins may bind telomerase, and influence the length of the telomere by controlling the access of telomerase to its substrate. Proteins that bind telomeres in mammalian cells have been found similarly, but less is known about their functions.

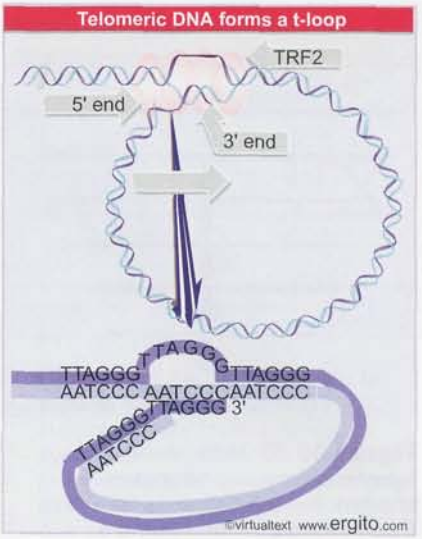


Figure 19.30 The 3′ single-stranded end of the telomere (TTAGGG)n displaces the homologous repeats from duplex DNA to form a t-loop. The reaction is catalyzed by TRF2.

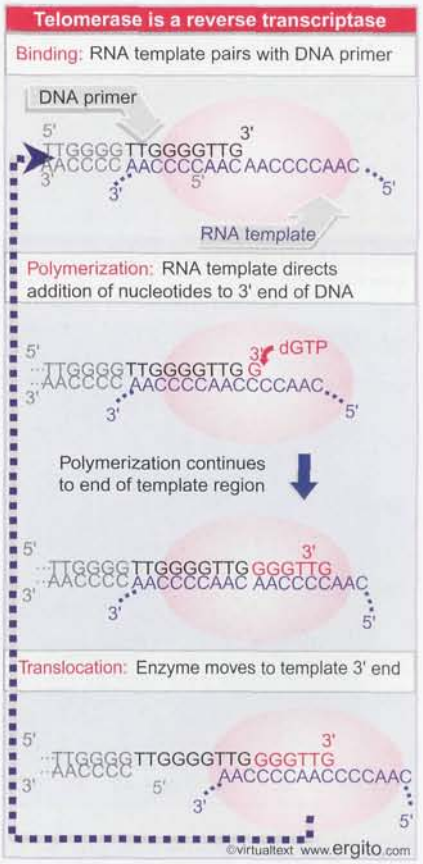


Figure 19.31 Telomerase positions itself by base pairing between the RNA template and the protruding single-stranded DNA primer. It adds G and T bases one at a time to the primer, as directed by the template.

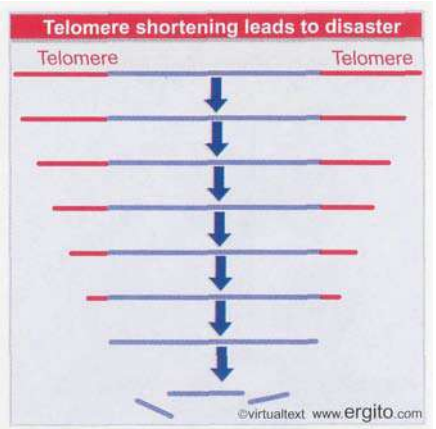


Figure 19.32 Mutation in telomerase causes telomeres to shorten in each cell division. Eventual loss of the telomere causes chromosome breaks and rearrangements.

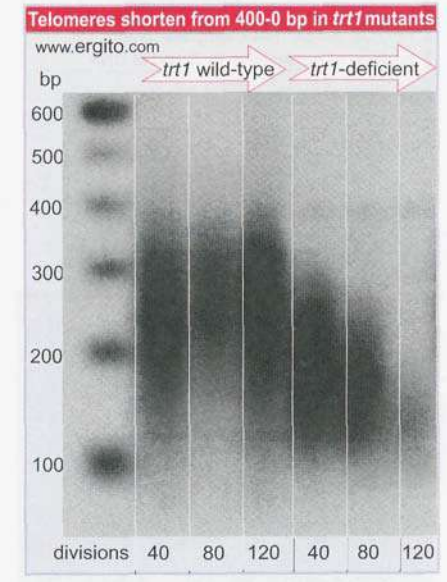


Figure 19.33 Telomere length is maintained at ~350 bp in wild-type yeast, but a mutant in the *trt1* gene coding for the RNA component of telomerase rapidly shortens its telomeres to zero length. Data kindly provided by Tom Cech and Toru Nakamura.

The minimum features required for existence as a chromosome are

- Telomeres to ensure survival.
- A centromere to support segregation.
- An origin to initiate replication.

All of these elements have been put together to construct a yeast artificial chromosome (YAC). This is a useful method for perpetuating foreign sequences. It turns out that the synthetic chromosome is stable only if it is longer than 20-50 kb. We do not know the basis for this effect, but the ability to construct a synthetic chromosome allows us to investigate the nature of the segregation device in a controlled environment.

19.19 Telomeres are essential for survival

Telomerase activity is found in all dividing cells and is generally turned off in terminally differentiated cells that do not divide. **Figure 19.32** shows that if telomerase is mutated in a dividing cell, the telomeres become gradually shorter with each cell division. An example of the effects of such a mutation in yeast are shown in **Figure 19.33,** where the telomere length shortens over ~120 generations from 400 bp to zero.

Loss of telomeres has very bad effects. When the telomere length reaches zero, it becomes difficult for the cells to divide successfully. Attempts to divide typically generate chromosome breaks and translocations. This causes an increased rate of mutation. In yeast this is associated with a loss of viability and the culture becomes predominantly occupied by senescent cells.

But some cells grow out of the senescing culture. They have acquired the ability to extend their telomeres by an alternative to telomerase activity. The survivors fall into groups. One group have circularized their chromosomes—since they now have no telomeres, they have become independent of telomerase. The other group use unequal crossing-over to extend their telomeres (see Figure 19.34). Because the telomere is a repeating structure, it is possible for two telomeres to misalign when chromosomes pair. Recombination between the mispaired regions generates an unequal crossing-over, as shown previously in Figure 4.1, when the length of one recombinant chromosome increases, and the length of the other recombinant chromosome decreases.

Cells usually suppress unequal crossing-over because of its potentially deleterious consequences. Two systems are responsible for suppressing crossing-over between telomeres. One is provided by telomere-binding proteins. In yeast, the frequency of recombination between telomeres is increased by deletion of the gene *taz1*, which codes for a protein that regulates telomerase activity. The second is a general system that undertakes mismatch repair (see *15 Recombination and repair).* In addition to correcting mismatched base pairs that may arise in DNA, this system suppresses recombination between mispaired regions. As shown in **Figure 19.34,** this includes telomeres. When it is mutated, a greater proportion of telomerase-deficient yeast survive the loss of telomeres because recombination between telomeres generates some chromosomes with longer telomeres.

When eukaryotic cells are placed in culture, they usually divide for a fixed number of generations and then enter senescence. The reason appears to be a decline in telomere length because of the absence of telomerase expression (see *30.24 Telomere shortening causes cell senescence).* Cells enter a crisis from which some emerge, but typically with chromosome rearrangements that have resulted from lack of protection of chromosome ends. These rearrangements may cause mutations that contribute to the tumorigenic state. The absence of telomerase

expression in this situation is due to failure to express the gene, and reactivation of telomerase is one of the mechanisms by which these cells then survive continued culture (this of course was not an option in the yeast experiments in which the gene had been deleted).

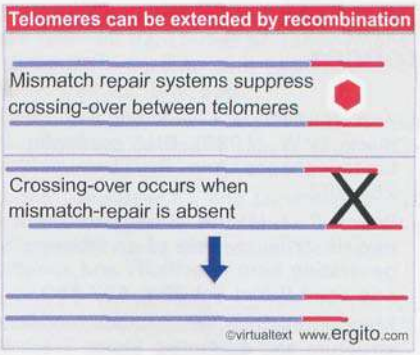


Figure 19.34 Crossing-over in telomeric regions is usually suppressed by mismatch-repair systems, but can occur when they are mutated. An unequal crossing-over event extends the telomere of one of the products, allowing the chromosome to survive in the absence of telomerase.

19.20 Summary

The genetic material of all organisms and viruses takes the form of tightly packaged nucleoprotein. Some virus genomes are inserted into preformed virions, while others assemble a protein coat around the nucleic acid. The bacterial genome forms a dense nucleoid, with ~20% protein by mass, but details of the interaction of the proteins with DNA are not known. The DNA is organized into ~100 domains that maintain independent supercoiling, with a density of unrestrained supercoils corresponding to ~1 / 100-200 bp. Interphase chromatin and metaphase chromosomes both appear to be organized into large loops. Each loop may be an independently supercoiled domain. The bases of the loops are connected to a metaphase scaffold or to the nuclear matrix by specific DNA sites.

Transcriptionally active sequences reside within the euchromatin that comprises the majority of interphase chromatin. The regions of heterochromatin are packaged ~5-10× more compactly, and are transcriptionally inert. All chromatin becomes densely packaged during cell division, when the individual chromosomes can be distinguished. The existence of a reproducible ultrastructure in chromosomes is indicated by the production of G-bands by treatment with Giemsa stain. The bands are very large regions, ~10^7 bp, that can be used to map chromosomal translocations or other large changes in structure.

Lampbrush chromosomes of amphibians and polytene chromosomes of insects have unusually extended structures, with packing ratios <100. Polytene chromosomes of *D. melanogaster* are divided into ~5000 bands, varying in size by an order of magnitude, with an average of ~25 kb. Transcriptionally active regions can be visualized in even more unfolded ("puffed") structures, in which material is extruded from the axis of the chromosome. This may resemble the changes that occur on a smaller scale when a sequence in euchromatin is transcribed.

The centromeric region contains the kinetochore, which is responsible for attaching a chromosome to the mitotic spindle. The centromere often is surrounded by heterochromatin. Centromeric sequences have been identified only in yeast *S. cerevisiae,* where they consist of short conserved elements, CDE-I and CDE-III that binds CBF1 and the CBF3 complex, respectively, and a long AT-rich region called CDE-II that binds Cse4p to form a specialized structure in chromatin. Another group of proteins that binds to this assembly provides the connection to microtubules.

Telomeres make the ends of chromosomes stable. Almost all known telomeres consist of multiple repeats in which one strand has the general sequence $C_n(A/T)_m$, where $n > 1$ and m = 1-4. The other strand, $G_n(T/A)_m$, has a single protruding end that provides a template for addition of individual bases in defined order. The enzyme telomere transferase is a ribonucleoprotein, whose RNA component provides the template for synthesizing the G-rich strand. This overcomes the problem of the inability to replicate at the very end of a duplex. The telomere stabilizes the chromosome end because the overhanging single strand $G_n(T/A)_m$ displaces its homologue in earlier repeating units in the telomere to form a loop, so there are no free ends.

References

19.2 Viral genomes are packaged into their coats

rev Black, L. W. (1989). DNA packaging in dsDNA bacteriophages. Ann. Rev. Immunol. 43, 267-292.

Butler, P. J. (1999). Self-assembly of tobacco mosaic virus: the role of an intermediate aggregate in generating both specificity and speed. Philos Trans R Soc Lond B Biol Sci 354, 537-550.

Klug, A. (1999). The tobacco mosaic virus particle: structure and assembly. Philos Trans R Soc Lond B Biol Sci 354, 531-535.

Mindich, L. (2000). Precise packaging of the three genomic segments of the double-stranded-RNA bacteriophage phi6. Microbiol. Mol. Biol. Rev. 63, 149-160.

ref Caspar, D. L. D. and Klug, A. (1962). Physical principles in the construction of regular viruses. Cold Spring Harbor Symp. Quant. Biol. 27, 1-24.

de Beer, T., Fang, J., Ortega, M., Yang, Q., Maes, L, Duffy, C, Berton, N., Sippy, J., Overduin, M., Feiss, M., and Catalano, C. E. (2002). Insights into specific DNA recognition during the assembly of a viral genome packaging machine. Mol. Cell 9, 981-991.

Dube, P., Tavares, P., Lurz, R., and van Heel, M. (1993). The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. EMBO J. 12, 1303-1309.

Fraenkel-Conrat, H. and Williams, R. C. (1955). Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components. Proc. Nat. Acad. Sci. USA 41, 690-698.

Jiang, Y. J., Aerne, B. L., Smithers, L., Haddon, C, Ish-Horowicz, D., and Lewis, J. (2000). Notch signalling and the synchronization of the somite segmentation clock. Nature 408, 475-479.

Zimmern, D. (1977). The nucleotide sequence at the origin for assembly on tobacco mosaic virus RNA. Cell 11, 463-482.

Zimmern, D. and Butler, P. J. (1977). The isolation of tobacco mosaic virus RNA fragments containing the origin for viral assembly. Cell 11, 455-462.

19.3 The bacterial genome is a nucleoid

rev Brock, T. D. (1988). The bacterial nucleus: a history. Microbiol. Rev. 52, 397-411.

Drlica, K. and Rouviere-Yaniv, J. (1987). Histone-like proteins of bacteria. Microbiol. Rev. 51, 301-319.

19.4 The bacterial genome is supercoiled

rev Hatfield, G. W. and Benham, C. J. (2002). DNA topology-mediated control of global gene expression in Escherichia coli. Ann. Rev. Genet. 36, 175-203.

ref Pettijohn, D. E. and Pfenninger, 0. (1980). Supercoils in prokaryotic DNA restrained in vivo. Proc. Nat. Acad. Sci. USA 77, 1331-1335.

19.8 Chromosomes have banding patterns

ref International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Saccone, S., De Sario, A., Wiegant, J., Raap, A. K., Delia Valle, G., and Bernardi, G. (1993). Correlations between isochores and chromosomal bands in the human genome. Proc. Nat. Acad. Sci. USA 90, 11929-11933.

Venter, J. C. et al. (2001). The sequence of the human genome. Science 291, 1304-1350.

19.12 The eukaryotic chromosome is a segregation device

rev Hyman, A. A. and Sorger, P. K. (1995). Structure and function of kinetochores in budding yeast. Ann. Rev. Cell Dev. Biol. 11, 471-495.

19.13 Centromeres have short DNA sequences in *S. cerevisiae*

rev Blackburn, E. H. and Szostak, J. W. (1984). The molecular structure of centromeres and telomeres. Ann. Rev. Biochem. 53, 163-194.

Clarke, L. and Carbon, J. (1985). The structure and function of yeast centromeres. Ann. Rev. Genet. 19, 29-56.

ref Fitzgerald-Hayes, M., Clarke, L., and Carbon, J. (1982). Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. Cell 29, 235-244.

19.14 The centromere binds a protein complex

rev Kitagawa, K. and Hieter, P. (2001). Evolutionary conservation between budding yeast and human kinetochores. Nat. Rev. Mol. Cell Biol. 2, 678-687.

ref Lechner, J. and Carbon, J. (1991). A 240 kd multisubunit protein complex, CBF3, is a major component of the budding yeast centromere. Cell 64, 717-725.

Meluh, P. B. et al. (1998). Cse4p is a component of the core centromere of *S. cerevisiae.* Cell 94, 607-613.

Meluh, P. B. and Koshland, D. (1997). Budding yeast centromere composition and assembly as revealed by *in vitro* cross-linking. Genes Dev. 11, 3401-3412.

Ortiz, J., Stemmann, 0., Rank, S., and Lechner, J. (1999). A putative protein complex consisting of Ctf19, Mcm21, and Okp1 represents a missing link in the budding yeast kinetochore. Genes Dev. 13, 1140-1155.

19.15 Centromeres may contain repetitious DNA

rev Wiens, G. R. and Sorger, P. K. (1998). Centromeric chromatin and epigenetic effects in kinetochore assembly. Cell 93, 313-316.

ref Copenhaver, G. P. et al. (1999). Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science 286, 2468-2474.

Haaf, T., Warburton, P. E., and Willard, H. F. (1992). Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. Cell 70, 681-696.

Sun, X., Wahlstrom, J., and Karpen, G. (1997). Molecular structure of a functional *Drosophila* centromere. Cell 91, 1007-1019.

19.16 Telomeres have simple repeating sequences

rev Blackburn, E. H. and Szostak, J. W. (1984). The molecular structure of centromeres and telomeres. Ann. Rev. Biochem. 53, 163-194.

Zakian, V. A. (1989). Structure and function of telomeres. Ann. Rev. Genet. 23, 579-604.

ref Wellinger, R. J., Ethier, K., Labrecque, P., and Zakian, V. A. (1996). Evidence for a new step in telomere maintenance. Cell 85, 423-433.

19.17 Telomeres seal the chromosome ends

ref Griffith, J. D. et al. (1999). Mammalian telomeres end in a large duplex loop. Cell 97, 503-514.

Henderson, E., Hardin, C. H., Walk, S. K., Tinoco, I., and Blackburn, E. H. (1987). Telomeric oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. Cell 51, 899-908.
Karlseder, J., Broccoli, D., Dai, Y., Hardy, S., and de Lange, T. (1999). p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. Science 283, 1321-1325.
Parkinson, G. N., Lee, M. P., and Neidle, S. (2002). Crystal structure of parallel quadruplexes from human telomeric DNA. Nature 417, 876-880.
van Steensel, B., Smogorzewska, A., and de Lange, T. (1998). TRF2 protects human telomeres from end-to-end fusions. Cell 92, 401-413.
Williamson, J. R., Raghuraman, K. R., and Cech, T. R. (1989). Monovalent cation-induced structure of telomeric DNA: the G-quartet model. Cell 59, 871-880.

19.18 Telomeres are synthesized by a ribonucleoprotein enzyme

rev Blackburn, E. H. (1991). Structure and function of telomeres. Nature 350, 569-573.
Blackburn, E. H. (1992). Telomerases. Ann. Rev. Biochem. 61, 113-129.
Collins, K. (1999). Ciliate telomerase biochemistry. Ann. Rev. Biochem. 68, 187-218.
Zakian, V. A. (1995). Telomeres: beginning to understand the end. Science 270, 1601-1607.
Zakian, V. A. (1996). Structure, function, and replication of *S. cerevisiae* telomeres. Ann. Rev. Genet. 30, 141-172.
ref Greider, C. and Blackburn, E. H. (1987). The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51, 887-898.
Murray, A., and Szostak, J. W. (1983). Construction of artificial chromosomes in yeast. Nature 305, 189-193.
Pennock, E., Buckley, K., and Lundblad, V. (2001). Cdc13 delivers separate complexes to the telomere for end protection and replication. Cell 104, 387-396.
Shippen-Lentz, D. and Blackburn, E. H. (1990). Functional evidence for an RNA template in telomerase. Science 247, 546-552.

19.19 Telomeres are essential for survival

ref Hackett, J. A., Feldser, D. M., and Greider, C. W. (2001). Telomere dysfunction increases mutation rate and genomic instability. Cell 106, 275-286.
Nakamura, T. M., Morin, G. B., Chapman, K. B., Weinrich, S. L., Andrews, W. H., Lingner, J., Harley, C. B., and Cech, T. R. (1997). Telomerase catalytic subunit homologs from fission yeast and human. Science 277, 955-959.
Nakamura, T. M., Cooper, J. P., and Cech, T. R. (1998). Two modes of survival of fission yeast without telomerase. Science 282, 493-496.
Rizki, A. and Lundblad, V. (2001). Defects in mismatch repair promote telomerase-independent proliferation. Nature 411, 713-716.

Chapter 20

Nucleosomes

20.1 Introduction

Chromatin has a compact organization in which most DNA sequences are structurally inaccessible and functionally inactive. Within this mass is the minority of active sequences. What is the general structure of chromatin, and what is the difference between active and inactive sequences? The high overall packing ratio of the genetic material immediately suggests that DNA cannot be directly packaged into the final structure of chromatin. There must be *hierarchies* of organization.

The fundamental subunit of chromatin has the same type of design in all eukaryotes. The nucleosome contains ~200 bp of DNA, organized by an octamer of small, basic proteins into a bead-like structure. The protein components are histones. They form an interior core; the DNA lies on the surface of the particle. Nucleosomes are an invariant component of euchromatin and heterochromatin in the interphase nucleus, and of mitotic chromosomes. The nucleosome provides the first level of organization, giving a packing ratio of ~6. Its components and structure are well characterized.

The second level of organization is the coiling of the series of nucleosomes into a helical array to constitute the fiber of diameter ~30 nm that is found in both interphase chromatin and mitotic chromosomes (see Figure 19.11). In chromatin this brings the packing ratio of DNA to ~40. The structure of this fiber requires additional proteins, but is not well defined.

The final packing ratio is determined by the third level of organization, the packaging of the 30 nm fiber itself. This gives an overall packing ratio of ~ 1000 in euchromatin, cyclically interchangeable with packing into mitotic chromosomes to achieve an overall ratio of ~10,000. Heterochromatin generally has a packing ratio ~10,000 in both interphase and mitosis.

We need to work through these levels of organization to characterize the events involved in cyclical packaging, replication, and transcription. We assume that association with additional proteins, or modifications of existing chromosomal proteins, are involved in changing the structure of chromatin. We do not know the individual targets for controlling cyclical packaging. Both replication and transcription require unwinding of DNA, and thus must involve an unfolding of the structure that allows the relevant enzymes to manipulate the DNA. This is likely to involve changes in all levels of organization.

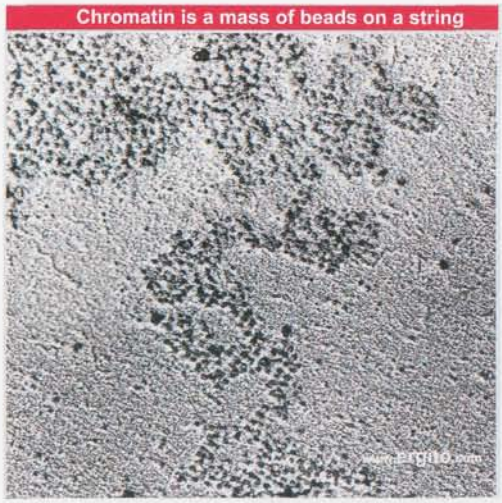


Figure 20.1 Chromatin spilling out of lysed nuclei consists of a compactly organized series of particles. Photograph kindly provided by Pierre Chambon.

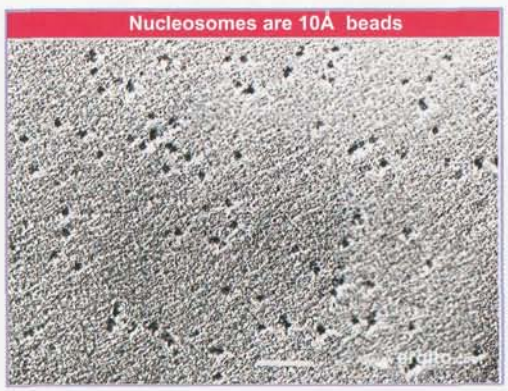


Figure 20.2 Individual nucleosomes are released by digestion of chromatin with micrococcal nuclease. Photograph kindly provided by Pierre Chambon.

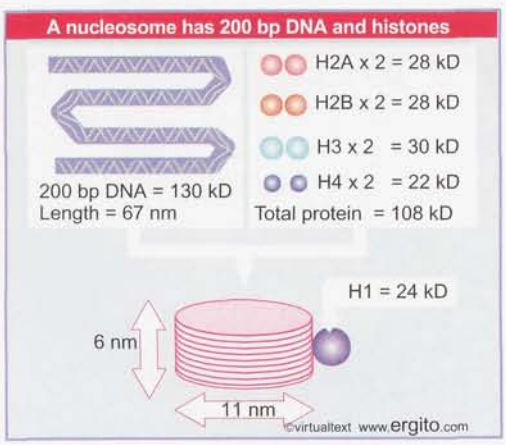


Figure 20.3 The nucleosome consists of approximately equal masses of DNA and histones (including H1). The predicted mass of the nucleosome is 262 kD.

When chromatin is replicated, the nucleosomes must be reproduced on both daughter duplex molecules. As well as asking how the nucleosome itself is assembled, we must inquire what happens to other proteins present in chromatin. Since replication disrupts the structure of chromatin, it both poses a problem for maintaining regions with specific structure and offers an opportunity to change the structure.

The mass of chromatin contains up to twice as much protein as DNA. Approximately half of the protein mass is accounted for by the nucleosomes. The mass of RNA is <10% of the mass of DNA. Much of the RNA consists of nascent transcripts still associated with the template DNA.

The nonhistones include all the proteins of chromatin except the histones. They are more variable between tissues and species, and they comprise a smaller proportion of the mass than the histones. They also comprise a much larger number of proteins, so that any individual protein is present in amounts much smaller than any histone.

The functions of nonhistone proteins include control of gene expression and higher-order structure. So RNA polymerase may be considered to be a prominent nonhistone. The HMG (high-mobility group) proteins comprise a discrete and well-defined subclass of nonhistones (at least some of which are transcription factors). A major problem in working with other nonhistones is that they tend to be contaminated with other nuclear proteins, and so far it has proved difficult to obtain those nonhistone proteins responsible for higher-order structures.

20.2 The nucleosome is the subunit of all chromatin

Key Concepts

- Micrococcal nuclease releases individual nucleosomes from chromatin as 11S particles.
- A nucleosome contains ~200 bp of DNA, 2 copies of each core histone (H2A, H2B, H3, H4), and 1 copy of H1.
- DNA is wrapped around the outside surface of the protein octamer.

When interphase nuclei are suspended in a solution of low ionic strength, they swell and rupture to release fibers of chromatin. **Figure 20.1** shows a lysed nucleus in which fibers are streaming out. In some regions, the fibers consist of tightly packed material, but in regions that have become stretched, they can be seen to consist of discrete particles. These are the nucleosomes. In especially extended regions, individual nucleosomes are connected by a fine thread, a free duplex of DNA. *A continuous duplex thread of DNA runs through the series of particles.*

Individual nucleosomes can be obtained by treating chromatin with the endonuclease micrococcal nuclease. It cuts the DNA thread at the junction between nucleosomes. First, it releases groups of particles; finally, it releases single nucleosomes. Individual nucleosomes can be seen in **Figure 20.2** as compact particles. They sediment at ~11S.

The nucleosome contains ~200 bp of DNA associated with a histone octamer that consists of two copies each of H2A, H2B, H3, and H4. These are known as the core histones. Their association is illustrated diagrammatically in **Figure 20.3.** This model explains the stoichiometry of the core histones in chromatin: H2A, H2B, H3, and H4 are present in equimolar amounts, with 2 molecules of each per ~200 bp of DNA.

Histones H3 and H4 are among the most conserved proteins known. This suggests that their functions are identical in all eukaryotes. The types of H2A and H2B can be recognized in all eukaryotes, but show appreciable species-specific variation in sequence.

Histone H1 comprises a set of closely related proteins that show appreciable variation between tissues and between species (and are absent from yeast). The role of H1 is different from the core histones. It is present in half the amount of a core histone and can be extracted more readily from chromatin (typically with dilute salt [0.5 M] solution). *The H1 can be removed without affecting the structure of the nucleosome, which suggests that its location is external to the particle.*

The shape of the nucleosome corresponds to a flat disk or cylinder, of diameter 11 nm and height 6 nm. The length of the DNA is roughly twice the ~34 nm circumference of the particle. The DNA follows a symmetrical path around the octamer. **Figure 20.4** shows the DNA path diagrammatically as a helical coil that makes two turns around the cylindrical octamer. Note that the DNA "enters" and "leaves" the nucleosome at points close to one another. Histone H1 may be located in this region (see *20.4 Nucleosomes have a common structure).*

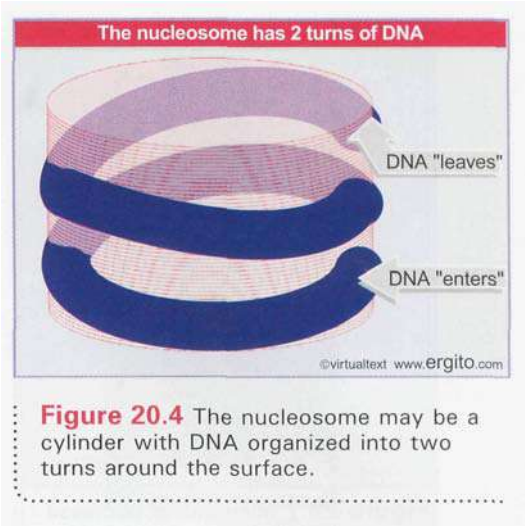


Figure 20.4 The nucleosome may be a cylinder with DNA organized into two turns around the surface.

Considering this model in terms of a cross-section through the nucleosome, in **Figure 20.5** we see that the two circumferences made by the DNA lie close to one another. The height of the cylinder is 6 nm, of which 4 nm is occupied by the two turns of DNA (each of diameter 2 nm).

The pattern of the two turns has a possible functional consequence. Since one turn around the nucleosome takes ~80 bp of DNA, two points separated by 80 bp in the free double helix may actually be close on the nucleosome surface, as illustrated in **Figure 20.6.**

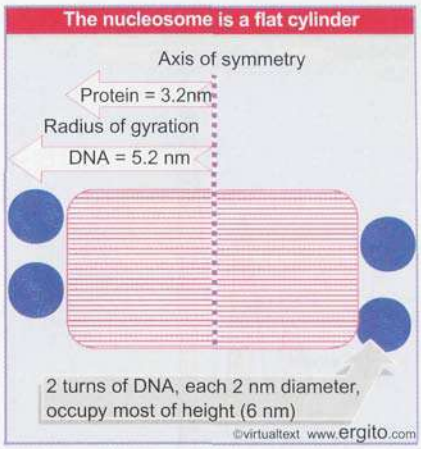


Figure 20.5 The two turns of DNA on the nucleosome lie close together.

20.3 DNA is coiled in arrays of nucleosomes

Key Concepts

- >95% of the DNA is recovered in nucleosomes or multimers when micrococcal nuclease cleaves DNA of chromatin.
- The length of DNA per nucleosome varies for individual tissues in a range from 154-260 bp.

When chromatin is digested with the enzyme micrococcal nuclease, the DNA is cleaved into integral multiples of a unit length. Fractionation by gel electrophoresis reveals the "ladder" presented in **Figure 20.7.** Such ladders extend for ~10 steps, and the unit length, determined by the increments between successive steps, is ~200 bp.

Figure 20.8 shows that the ladder is generated by groups of nucleosomes. When nucleosomes are fractionated on a sucrose gradient, they give a series of discrete peaks that correspond to monomers, dimers, trimers, etc. When the DNA is extracted from the individual fractions and electrophoresed, each fraction yields a band of DNA whose size corresponds with a step on the micrococcal nuclease ladder. The monomeric nucleosome contains DNA of the unit length, the nucleosome dimer contains DNA of twice the unit length, and so on.

So each step on the ladder represents the DNA derived from a discrete number of nucleosomes. *We therefore take the existence of the 200 bp ladder in any chromatin to indicate that the DNA is organized into nucleosomes.* The micrococcal ladder is generated

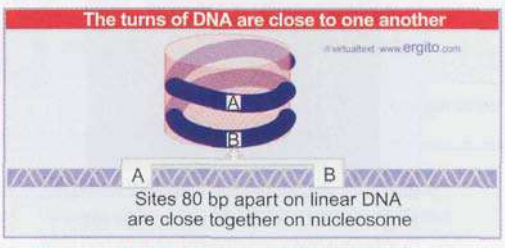


Figure 20.6 Sequences on the DNA that lie on different turns around the nucleosome may be close together.

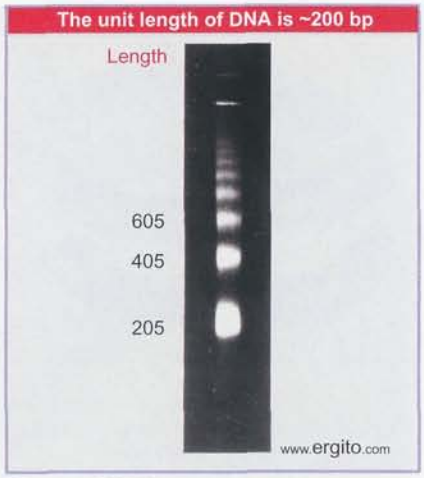


Figure 20.7 Micrococcal nuclease digests chromatin in nuclei into a multimeric series of DNA bands that can be separated by gel electrophoresis. Photograph kindly provided by Markus Noll.

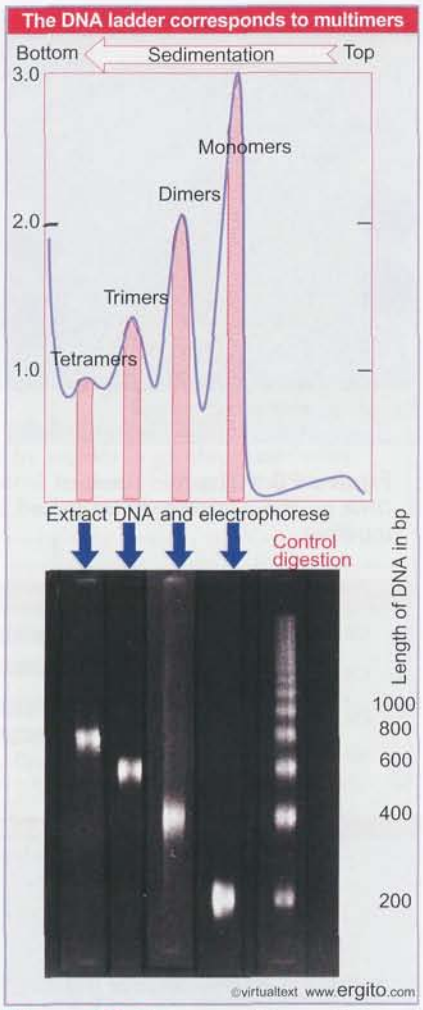


Figure 20.8 Each multimer of nucleosomes contains the appropriate number of unit lengths of DNA. Photograph kindly provided by John Finch.

when only ~2% of the DNA in the nucleus is rendered acid-soluble (degraded to small fragments) by the enzyme. *So a small proportion of the DNA is specifically attacked; it must represent especially susceptible regions.*

When chromatin is spilled out of nuclei, we often see a series of nucleosomes connected by a thread of free DNA (the beads on a string). However, the need for tight packaging of DNA *in vivo* suggests that probably there is usually little (if any) free DNA.

This view is confirmed by the fact that *>95% of the DNA of chromatin can be recovered in the form of the 200 bp ladder*. Almost all DNA must therefore be organized in nucleosomes. In their natural state, nucleosomes are likely to be closely packed, with DNA passing directly from one to the next. Free DNA is probably generated by the loss of some histone octamers during isolation.

The length of DNA present in the nucleosome varies somewhat from the "typical" value of 200 bp. The chromatin of any particular cell type has a characteristic average value (±5 bp). The average most often is between 180 and 200, but there are extremes as low as 154 bp (in a fungus) or as high as 260 bp (in a sea urchin sperm). The average value may be different in individual tissues of the adult organism. And there can be differences between different parts of the genome in a single cell type. Variations from the genome average include tandemly repeated sequences, such as clusters of 5S RNA genes.

20.4 Nucleosomes have a common structure

Key Concepts

- Nucleosomal DNA is divided into the core DNA and linker DNA depending on its susceptibility to micrococcal nuclease.
- The core DNA is the length of 146 bp that is found on the core particles produced by prolonged digestion with micrococcal nuclease.
- Linker DNA is the region of 8-114 bp that is susceptible to early cleavage by the enzyme.
- Changes in the length of linker DNA account for the variation in total length of nucleosomal DNA.
- H1 is associated with linker DNA and may lie at the point where DNA enters and leaves the nucleosome.

A common structure underlies the varying amount of DNA that is contained in nucleosomes of different sources. The association of DNA with the histone octamer forms a core particle containing 146 bp of DNA, irrespective of the total length of DNA in the nucleosome. The variation in total length of DNA per nucleosome is superimposed on this basic core structure.

The core particle is defined by the effects of micrococcal nuclease on the nucleosome monomer. The initial reaction of the enzyme is to cut between nucleosomes, but if it is allowed to continue after monomers have been generated, then it proceeds to digest some of the DNA of the individual nucleosome. This occurs by a reaction in which DNA is "trimmed" from the ends of the nucleosome.

The length of the DNA is reduced in discrete steps, as shown in **Figure 20.9.** With rat liver nuclei, the nucleosome monomers initially have 205 bp of DNA. Then some monomers are found in which the length of DNA has been reduced to ~165 bp. Finally this is reduced to the length of the DNA of the core particle, 146 bp. (The core is reason-

ably stable, but continued digestion generates a "limit digest", in which the longest fragments are the 146 bp DNA of the core, while the shortest are as small as 20 bp.)

This analysis suggests that the nucleosomal DNA can be divided into two regions:

- Core DNA has an invariant length of 146 bp, and is relatively resistant to digestion by nucleases.
- Linker DNA comprises the rest of the repeating unit. Its length varies from as little as 8 bp to as much as 114 bp per nucleosome.

The sharp size of the band of DNA generated by the initial cleavage with micrococcal nuclease suggests that the region immediately available to the enzyme is restricted. It represents only part of each linker. (If the entire linker DNA were susceptible, the band would range from 146 bp to >200 bp.) But once a cut has been made in the linker DNA, the rest of this region becomes susceptible, and it can be removed relatively rapidly by further enzyme action. The connection between nucleosomes is represented in **Figure 20.10.**

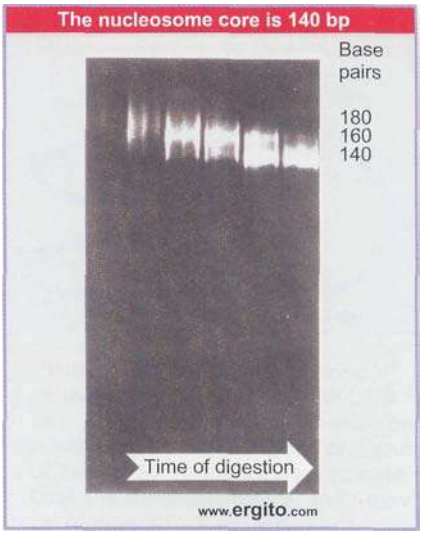


Figure 20.9 Micrococcal nuclease reduces the length of nucleosome monomers in discrete steps. Photograph kindly provided by Roger Kornberg.

Core particles have properties similar to those of the nucleosomes themselves, although they are smaller. Their shape and size are similar to nucleosomes, which suggests that the essential geometry of the particle is established by the interactions between DNA and the protein octamer in the core particle. Because core particles are more readily obtained as a homogeneous population, they are often used for structural studies in preference to nucleosome preparations. (Nucleosomes tend to vary because it is difficult to obtain a preparation in which there has been no end-trimming of the DNA.)

What is the physical nature of the core and the linker regions? *These terms are operational definitions that describe the regions in terms of their relative susceptibility to nuclease treatment.* This description does not make any implication about their actual structure. In fact, the path of DNA on the histone octamer appears to be continuous. It takes 165 bp to make the two turns around the octamer. This is an invariant feature of nucleosomes. The transition from one nucleosome to the next is made within the additional length of DNA, and there could be differences in the path in this region depending on the length of DNA per nucleosome.

The existence of linker DNA depends on factors extraneous to the four core histones. Reconstitution experiments *in vitro* show that histones have an intrinsic ability to organize DNA into core particles, but do not form nucleosomes with the proper unit length. The degree of supercoiling of the DNA is an important factor. Histone H1 and/or nonhistone proteins influence the length of linker DNA associated with the histone octamer in a natural series of nucleosomes. And "assembly proteins" that are not part of the nucleosome structure are involved *in vivo* in constructing nucleosomes from histones and DNA (see *20.10 Reproduction of chromatin requires assembly of nucleosomes).*

Where is histone H1 located? The H1 is lost during the degradation of nucleosome monomers. It can be retained on monomers that still have 165 bp of DNA; but is always lost with the final reduction to the 146 bp core particle. This suggests that H1 could be located in the region of the linker DNA immediately adjacent to the core DNA.

If H1 is located at the linker, it could "seal" the DNA in the nucleosome by binding at the point where the nucleic acid enters and leaves (see Figure 20.4). The idea that H1 lies in the region joining adjacent nucleosomes is consistent with old results that H1 is removed the most readily from chromatin, and that H1-depleted chromatin is more readily "solubilized". And it is easier to obtain a stretched-out fiber of beads on a string when the H1 has been removed.

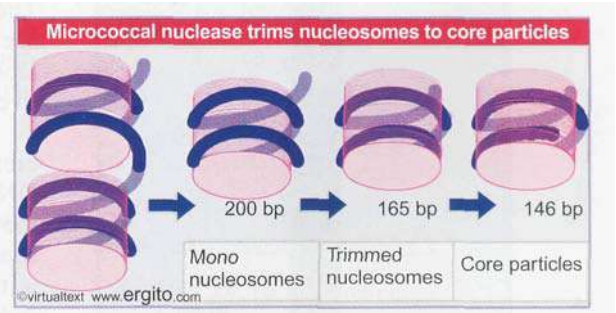


Figure 20.10 Micrococcal nuclease initially cleaves between nucleosomes. Mononucleosomes typically have ~200 bp DNA. End-trimming reduces the length of DNA first to ~165 bp, and then generates core particles with 146 bp.

20.5 DNA structure varies on the nucleosomal surface

Key Concepts

- 1.65 turns of DNA are wound round the histone octamer.
- The structure of the DNA is altered so that it has an increased number of base pairs/turn in the middle, but a decreased number at the ends.

The exposure of DNA on the surface of the nucleosome explains why it is accessible to cleavage by certain nucleases. The reaction with nucleases that attack single strands has been especially informative. The enzymes DNAase I and DNAase II make single-strand nicks in DNA; they cleave a bond in one strand, but the other strand remains intact at this point. So no effect is visible in the double-stranded DNA. But upon denaturation, short fragments are released instead of full-length single strands. If the DNA has been labeled at its ends, the end fragments can be identified by autoradiography as summarized in **Figure 20.11**.

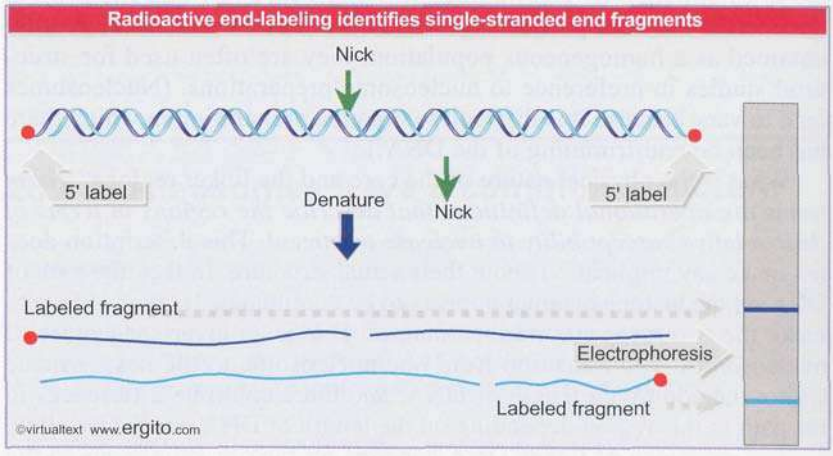


Figure 20.11 Nicks in double-stranded DNA are revealed by fragments when the DNA is denatured to give single strands. If the DNA is labeled at (say) 5' ends, only the 5' fragments are visible by autoradiography. The size of the fragment identifies the distance of the nick from the labeled end.

When DNA is free in solution, it is nicked (relatively) at random. The DNA on nucleosomes also can be nicked by the enzymes, *but only at regular intervals*. When the points of cutting are determined by using radioactively end-labeled DNA and then DNA is denatured and electrophoresed, a ladder of the sort displayed in **Figure 20.12** is obtained.

The interval between successive steps on the ladder is 10-11 bases. The ladder extends for the full distance of core DNA. The cleavage sites are numbered as S1 through S13 (where S1 is ~10 bases from the labeled 5' end. S2 is ~20 bases from it, and so on). Their positions relative to the DNA superhelix are illustrated in **Figure 20.13.**

Not all sites are cut with equal frequency: some are cut rather efficiently, others are cut scarcely at all. The enzymes DNAase I and DNAase II generate the same ladder, although with some differences in the intensities of the bands. This shows that the pattern of cutting represents a unique series of targets in DNA, determined by its organization, with only some slight preference for particular sites imposed by the individual enzyme. The same cutting pattern is obtained by cleaving with a hydroxyl radical, which argues that the pattern reflects the structure of the DNA itself, rather than any sequence preference.

The sensitivity of nucleosomal DNA to nucleases is analogous to a footprinting experiment. So we can assign the lack of reaction at particular target sites to the structure of the nucleosome, in which certain positions on DNA are rendered inaccessible.

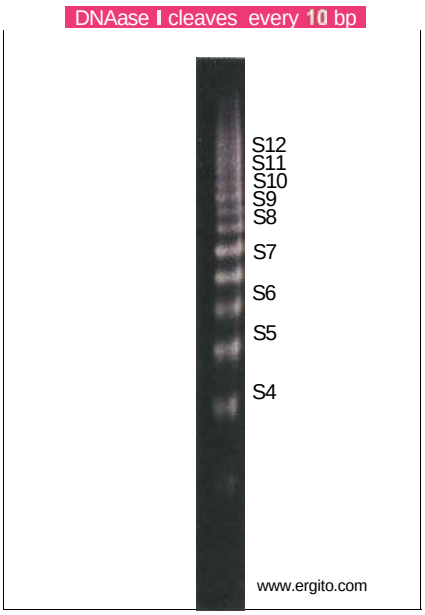


Figure 20.12 Sites for nicking lie at regular intervals along core DNA, as seen in a DNAase I digest of nuclei. Photograph kindly provided by Leonard Lutter.

By Book_Crazy [IND]

Since there are two strands of DNA in the core particle, in an end-labeling experiment both 5' (or both 3') ends are labeled, one on each strand. So the cutting pattern includes fragments derived from both strands. This is implied in Figure 20.11, where each labeled fragment is derived from a different strand. The corollary is that, in an experiment, each labeled band in fact represents two fragments, generated by cutting the *same* distance from *either* of the labeled ends.

How then should we interpret discrete preferences at particular sites? One view is that the path of DNA on the particle is symmetrical (about a horizontal axis through the nucleosome drawn in Figure 20.4). So if (for example) no 80-base fragment is generated by DNAase I, this must mean that the position at 80 bases from the 5' end *of either* strand is not susceptible to the enzyme. The second numbering scheme used in Figure 20.13 reflects this view, and identifies S7 = site 0 as the center of symmetry.

When DNA is immobilized on a flat surface, sites are cut with a regular separation. **Figure 20.14** suggests that this reflects the recurrence of the exposed site with the helical periodicity of B-form DNA. The cutting periodicity (the spacing between cleavage points) coincides with, indeed, is a reflection of, the structural periodicity (the number of base pairs per turn of the double helix). So the distance between the sites corresponds to the number of base pairs per turn. Measurements of this type suggest that the average value for double-helical B-type DNA is 10.5 bp/turn.

What is the nature of the target sites on the nucleosome? **Figure 20.15** shows that each site has 3-4 positions at which cutting occurs; that is, the cutting site is defined ±2 bp. So a cutting site represents a short stretch of bonds on both strands, exposed to nuclease action over 3-4 base pairs. The relative intensities indicate that some sites are preferred to others.

From this pattern, we can calculate the "average" point that is cut. At the ends of the DNA, pairs of sites from S1 to S4 or from S10 to S13 lie apart a distance of 10.0 bases each. In the center of the particle, the separation from sites S4 to S10 is 10.7 bases. (Because this analysis deals with *average* positions, sites need not lie an integral number of bases apart.)

The variation in cutting periodicity along the core DNA (10.0 at the ends, 10.7 in the middle) means that there is variation in the structural periodicity of core DNA. The DNA has more bp/turn than its solution value in the middle, but has fewer bp/turn at the ends. The average periodicity over the nucleosome is less than the 10.5 bp/turn of DNA in solution; it is in the range of 10.2-10.4 bp/turn, depending on the method of measurement.

The crystal structure of the core particle suggests that DNA is organized as a flat superhelix, with 1.65 turns wound around the histone octamer. The pitch of the superhelix varies, with a discontinuity in the middle. Regions of high curvature are arranged symmetrically, and occur at positions ±1 and ±4. These correspond to S6 and S8, and to S3 and S11, which are the sites least sensitive to DNAase I. The high curvature is probably responsible for these changes, but their precise nature remains to be determined at the molecular level.

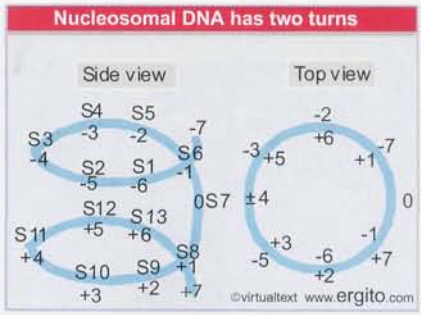


Figure 20.13 Two numbering schemes divide core particle DNA into 10 bp segments. Sites may be numbered S1 to S13 from one end; or taking S7 to identify coordinate 0 of the dyad symmetry, they may be numbered −7 to +7.

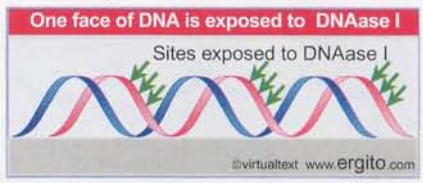


Figure 20.14 The most exposed positions on DNA recur with a periodicity that reflects the structure of the double helix. (For clarity, sites are shown for only one strand.)

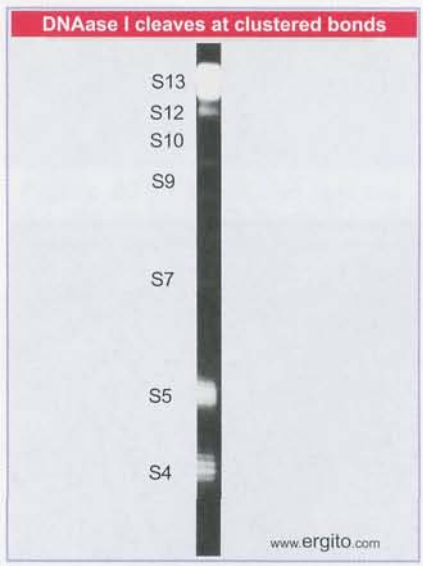


Figure 20.15 High resolution analysis shows that each site for DNAase I consists of several adjacent susceptible phosphodiester bonds as seen in this example of sites S4 and S5 analyzed in end-labeled core particles. Photograph kindly provided by Leonard Lutter.

20.6 The periodicity of DNA changes on the nucleosome

Key Concepts

- ~0.6 negative turns of DNA are absorbed by the change in bp/turn from 10.5 in solution to an average of 10.2 on the nucleosomal surface, explaining the linking number paradox.

By Book_Crazy [IND]

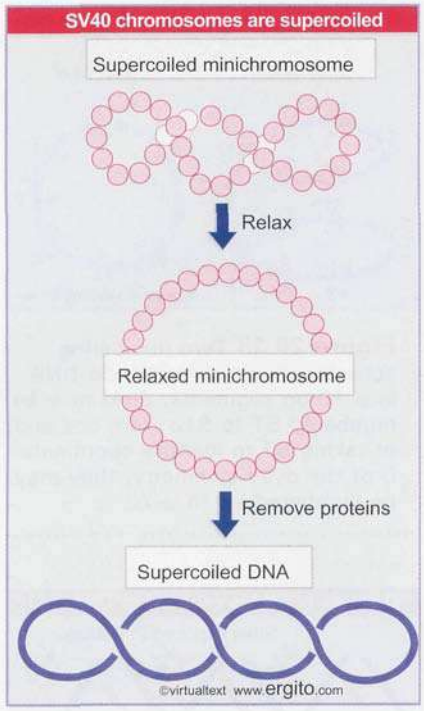


Figure 20.16 The supercoils of the SV40 minichromosome can be relaxed to generate a circular structure, whose loss of histones then generates supercoils in the free DNA.

Some insights into the structure of nucleosomal DNA emerge when we compare predictions for supercoiling in the path that DNA follows with actual measurements of supercoiling of nucleosomal DNA. Much work on the structure of sets of nucleosomes has been carried out with the virus SV40. The DNA of SV40 is a circular molecule of 5200 bp, with a contour length ~1500 nm. In both the virion and infected nucleus, it is packaged into a series of nucleosomes, called a minichromosome.

As usually isolated, the contour length of the minichromosome is ~210 nm, corresponding to a packing ratio of ~7 (essentially the same as the ~6 of the nucleosome itself). Changes in the salt concentration can convert it to a flexible string of beads with a much lower overall packing ratio. This emphasizes the point that nucleosome strings can take more than one form *in vitro*, depending on the conditions.

The degree of supercoiling on the individual nucleosomes of the minichromosome can be measured as illustrated in **Figure 20.16.** First, the free supercoils of the minichromosome itself are relaxed, so that the nucleosomes form a circular string with a superhelical density of 0. Then the histone octamers are extracted. This releases the DNA to follow a free path. Every supercoil that was present but restrained in the minichromosome will appear in the deproteinized DNA as -1 turn. So now the total number of supercoils in the SV40 DNA is measured.

The observed value is close to the number of nucleosomes. The reverse result is seen when nucleosomes are assembled *in vitro* on to a supercoiled SV40 DNA: the formation of each nucleosome removes ~1 negative supercoil.

So the DNA follows a path on the nucleosomal surface that generates ~1 negative supercoiled turn when the restraining protein is removed. But the path that DNA follows on the nucleosome corresponds to -1.65 superhelical turns (see Figure 20.4). This discrepancy is sometimes called the linking number paradox.

The discrepancy is explained by the difference between the 10.2 average bp/turn of nucleosomal DNA and the 10.5 bp/turn of free DNA. In a nucleosome of 200 bp, there are 200/10.2 = 19.6 turns. When DNA is released from the nucleosome, it now has 200/10.5 = 19.0 turns. The path of the less tightly wound DNA on the nucleosome absorbs -0.6 turns, and this explains the discrepancy between the physical path of -1.65 and the measurement of –1.0 superhelical turns. In effect, some of the torsional strain in nucleosomal DNA goes into increasing the number of bp/turn; only the rest is left to be measured as a supercoil.

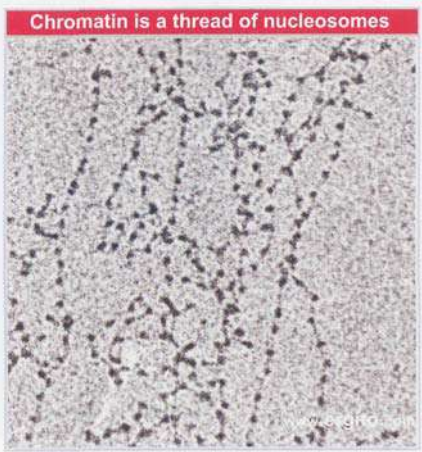


Figure 20.17 The 10 nm fiber in partially unwound state can be seen to consist of a string of nucleosomes. Photograph kindly provided by Barbara Hamkalo.

20.7 The path of nucleosomes in the chromatin fiber

Key Concepts

- 10 nm chromatin fibers are unfolded from 30 nm fibers and consist of a string of nucleosomes.
- 30 nm fibers have 6 nucleosomes/turn, organized into a solenoid.
- Histone H1 is required for formation of the 30 nm fiber.

When chromatin is examined in the electron microscope, two types of fibers are seen: the 10 nm fiber and 30 nm fiber. They are described by the approximate diameter of the thread (that of the 30 nm fiber actually varies from ~25-30 nm).

The 10 nm fiber is essentially a continuous string of nucleosomes. Sometimes, indeed, it runs continuously into a more stretched-out

region in which nucleosomes are seen as a string of beads, as indicated in the example of Figure 20.17. The 10 nm fibril structure is obtained under conditions of low ionic strength and does not require the presence of histone H1. This means that it is a function strictly of the nucleosomes themselves. It may be visualized essentially as a continuous series of nucleosomes, as in Figure 20.18. It is not clear whether such a structure exists *in vivo* or is simply a consequence of unfolding during extraction *in vitro*.

When chromatin is visualized in conditions of greater ionic strength the 30 nm fiber is obtained. An example is given in Figure 20.19. The fiber can be seen to have an underlying coiled structure. It has ~6 nucleosomes for every turn, which corresponds to a packing ratio of 40 (that is, each μm along the axis of the fiber contains 40 μm of DNA). The presence of H1 is required. This fiber is the basic constituent of both interphase chromatin and mitotic chromosomes.

The most likely arrangement for packing nucleosomes into the fiber is a solenoid, illustrated in Figure 20.20. The nucleosomes turn in a helical array, with an angle of ~60° between the faces of adjacent nucleosomes.

The 30 nm and 10 nm fibers can be reversibly converted by changing the ionic strength. This suggests that the linear array of nucleosomes in the 10 nm fiber is coiled into the 30 nm structure at higher ionic strength and in the presence of H1.

Although the presence of H1 is necessary for the formation of the 30 nm fiber, information about its location is conflicting. Its relative ease of extraction from chromatin seems to argue that it is present on the outside of the superhelical fiber axis. But diffraction data, and the fact that it is harder to find in 30 nm fibers than in 10 nm fibers that retain it, would argue for an interior location.

How do we get from the 30 nm fiber to the specific structures displayed in mitotic chromosomes? And is there any further specificity in the arrangement of interphase chromatin; do particular regions of 30 nm fibers bear a fixed relationship to one another or is their arrangement random?

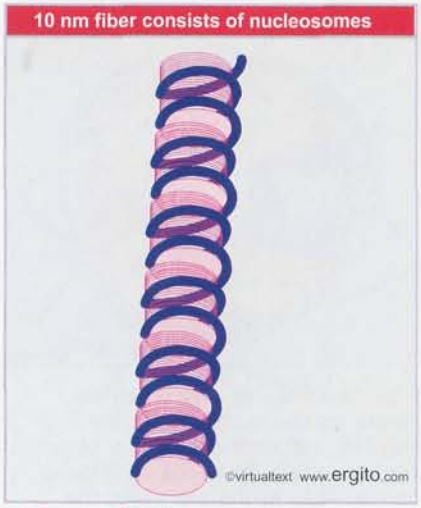


Figure 20.18 The 10 nm fiber is a continuous string of nucleosomes.

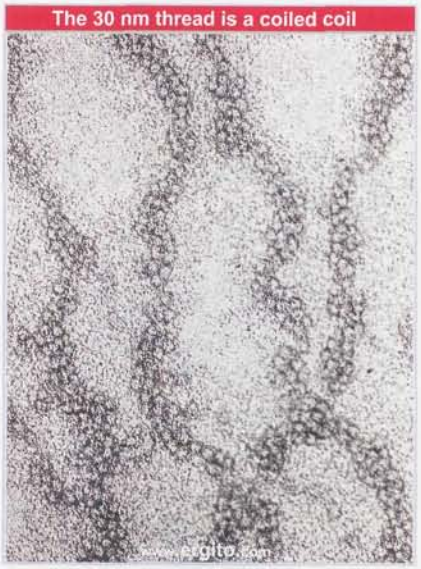


Figure 20.19 The 30 nm fiber has a coiled structure. Photograph kindly provided by Barbara Hamkalo.

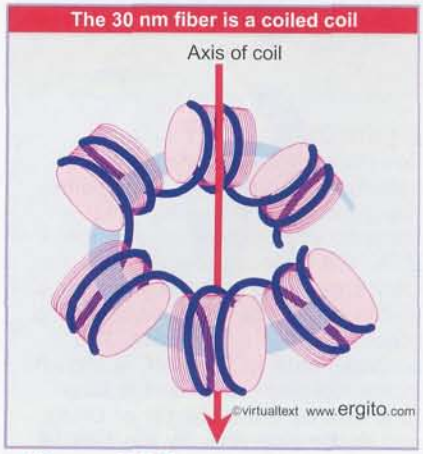


Figure 20.20 The 30 nm fiber may have a helical coil of 6 nucleosomes per turn, organized radially.

20.8 Organization of the histone octamer

Key Concepts

- The histone octamer has a kernel of a $H3_2 \cdot H4_2$ tetramer associated with two H2AH2B dimers.
- Each histone is extensively interdigitated with its partner.
- All core histories have the structural motif of the histone fold. N-terminal tails extend out of the nucleosome.

So far we have considered the construction of the nucleosome from the perspective of how the DNA is organized on the surface. From the perspective of protein, we need to know how the histones interact with each other and with DNA. Do histones react properly only in the presence of DNA, or do they possess an independent ability to form octamers? Most of the evidence about histone-histone interactions is provided by their abilities to form stable complexes, and by crosslinking experiments with the nucleosome.

The core histones form two types of complexes. H3 and H4 form a tetramer ($H3_2 \cdot H4_2$). Various complexes are formed by H2A and H2B, in particular a dimer (H2A·H2B).

Intact histone octamers can be obtained either by extraction from chromatin or (with more difficulty) by letting histones associate *in vitro*

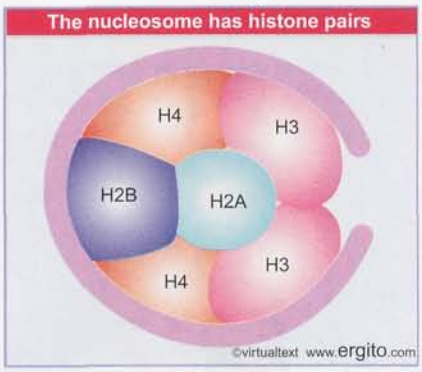


Figure 20.21 In a symmetrical model for the nucleosome, the $H3_2$-$H4_2$ tetramer provides a kernel for the shape. One H2A-H2B dimer can be seen in the top view; the other is underneath.

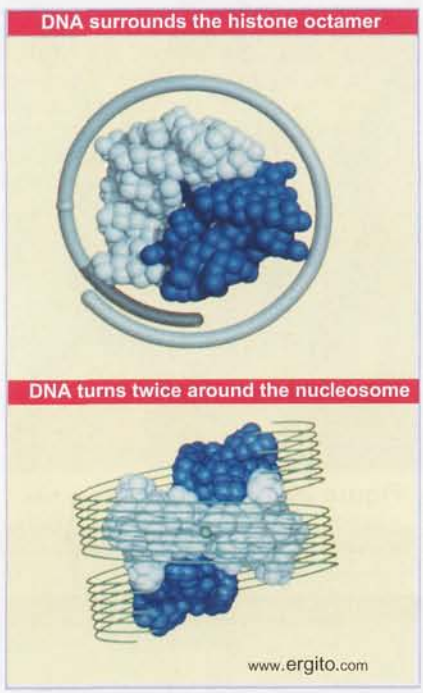


Figure 20.22 The crystal structure of the histone core octamer is represented in a space-filling model with the $H3_2$-$H4_2$ tetramer shown in white and the H2A-H2B dimers shown in blue. Only one of the H2A-H2B dimers is visible in the top view, because the other is hidden underneath. The potential path of the DNA is shown in the top view as a narrow tube (one quarter the diameter of DNA), and in the side view by the parallel lines in a 20 Å wide bundle. Photographs kindly provided by Evangelos Moudrianakis.

under conditions of high-salt and high-protein concentration. The octamer can dissociate to generate a hexamer of histones that has lost an H2A·H2B dimer. Then the other H2A·H2B dimer is lost separately, leaving the $H3_2$·$H4_2$ tetramer. This argues for a form of organization in which the nucleosome has a central "kernel" consisting of the $H3_2$·$H4_2$ tetramer. The tetramer can organize DNA *in vitro* into particles that display some of the properties of the core particle.

Crosslinking studies extend these relationships to show which pairs of histones lie near each other in the nucleosome. (A difficulty with such data is that usually only a small proportion of the proteins becomes crosslinked, so it is necessary to be cautious in deciding whether the results typify the major interactions.) From these data, a model has been constructed for the organization of the nucleosome. It is shown in diagrammatic form in **Figure 20.21.**

Structural studies show that the overall shape of the isolated histone octamer is similar to that of the core particle. This suggests that the histone-histone interactions establish the general structure. The positions of the individual histones have been assigned to regions of the octameric structure on the basis of their interaction behavior and response to crosslinking.

The crystal structure (at 3.1 A resolution) suggests the model for the histone octamer shown in **Figure 20.22.** Tracing the paths of the individual polypeptide backbones in the crystal structure suggests that the histones are not organized as individual globular proteins, but that each is interdigitated with its partner, H3 with H4, and H2A with H2B. So the model distinguishes the $H3_2$·$H4_2$ tetramer (white) from the H2A·H2B dimers (blue), but does not show individual histones.

The top view represents the same perspective that was illustrated schematically in Figure 20.21. The $H3_2$·$H4_2$ tetramer accounts for the diameter of the octamer. It forms the shape of a horseshoe. The H2A·H2B pairs fit in as two dimers, but only one can be seen in this view. The side view represents the same perspective that was illustrated in Figure 20.4. Here the responsibilities of the $H3_2$·$H4_2$ tetramer and of the separate H2A·H2B dimers can be distinguished. The protein forms a sort of spool, with a superhelical path that could correspond to the binding site for DNA, which would be wound in almost two full turns in a nucleosome. The model displays two fold symmetry about an axis that would run perpendicular through the side view.

A more detailed view of the positions of the histones (based on a crystal structure at 2.8 A) is summarized in the next two figures. **Figure 20.23** shows the position of one histone of each type relative to one turn around the nucleosome (numbered from 0 to +7). All four core histones show a similar type of structure in which three α-helices are connected by two loops: this is called the histone fold. These regions interact to form crescent-shaped heterodimers; each heterodimer binds 2.5 turns of the DNA double helix (H2A-H2B binds at **+3.5** - +6; H3-H4 binds at +0.5 - +3 for the circumference that is illustrated). Binding is mostly to the phosphodiester backbones (consistent with the need to package any DNA irrespective of sequence). **Figure 20.24** shows that the $H3_2$·$H4_2$ tetramer is formed by interactions between the two H3 subunits.

Each of the core histones has a globular body that contributes to the central protein mass of the nucleosome. Each histone also has a flexible N-terminal tail, which has sites for modification that may be important in chromatin function. The positions of the tails, which account for about one quarter of the protein mass, are not so well defined. However, the tails of both H3 and H2B can be seen to pass between the turns of the DNA superhelix and extend out of the nucleosome, as seen in **Figure 20.25.** When histone tails are crosslinked to DNA by UV irradiation, more products are obtained with nucleosomes

compared to core particles, which could mean that the tails contact the linker DNA. The tail of H4 appears to contact an H2A-H2B dimer in an adjacent nucleosome; this could be an important feature in the overall structure.

20.9 The N-terminal tails of histones are modified

Key Concepts

- Histones are modified by methylation, acetylation, and phosphorylation.

All of the histones are modified by covalently linking extra moieties to the free groups of certain amino acids. The sites that are modified are concentrated in the N-terminal tails. These modifications have important effects on the structure of chromatin and in controlling gene expression (see *23.5 Histone modification is a key event).*

Acetylation and methylation occur on the free (e) amino group of lysine. As seen in **Figure 20.26,** this removes the positive charge that resides on the NH^+_3 form of the group. Methylation also occurs on arginine and histidine. Phosphorylation occurs on the hydroxyl group of serine and also on histidine. This introduces a negative charge in the form of the phosphate group.

These modifications are transient. Because they change the charge of the protein molecule, they are potentially able to change the functional properties of the octamers. Modification of histones is associated with structural changes that occur in chromatin at replication and transcription. Phosphorylations on specific positions and on different histones may be required for particular processes, for example, the Ser^{10} position of H3 is phosphorylated when chromosomes condense at mitosis.

In synchronized cells in culture, both the pre-existing and newly synthesized core histones appear to be acetylated and methylated during S phase (when DNA is replicated and the histones also are synthesized). During the cell cycle, the modifying groups are later removed.

The coincidence of modification and replication suggests that acetylation (and methylation) could be connected with nucleosome assembly. One speculation has been that the reduction of positive charges on histones might lower their affinity for DNA, allowing the reaction to be better controlled. The idea has lost some ground in view of the observation that nucleosomes can be reconstituted, at least *in vitro,* with unmodified histones. Histone acetylation is essential for nucleosome assembly in yeast, and is probably required for some of the protein-protein interactions that occur during later stages of the reaction (see *23.6 Histone acetylation occurs in two circumstances).*

A cycle of phosphorylation and dephosphorylation occurs with H1, but its timing is different from the modification cycle of the other histones. With cultured mammalian cells, one or two phosphate groups are introduced at S phase. But the major phosphorylation event is the later addition of more groups at mitosis, to bring the total number up to as many as six. All the phosphate groups are removed at the end of the process of division. The phosphorylation of H1 is catalyzed by the M-phase kinase that provides an essential trigger for mitosis (see *29 Cell cycle and growth regulation).* In fact, this enzyme is now often assayed in terms of its H1 kinase activity. Not much is known about phosphatase(s) that remove the groups later.

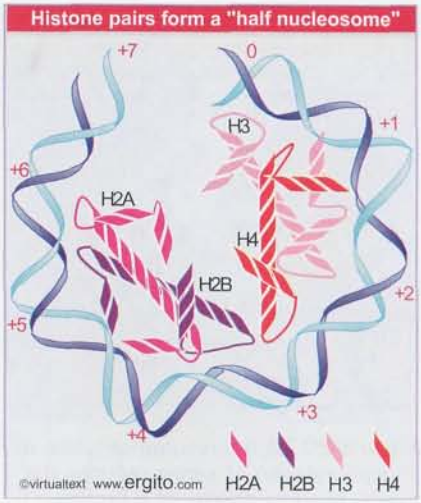


Figure 20.23 Histone positions in a top view show H3-H4 and H2A-H2B pairs in a half nucleosome.

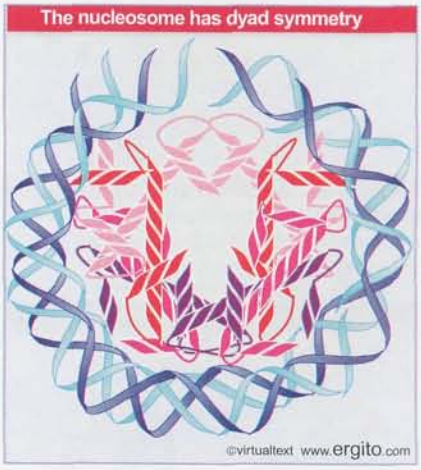


Figure 20.24 Symmetrical organization can be seen by superimposing both half nucleosomes.

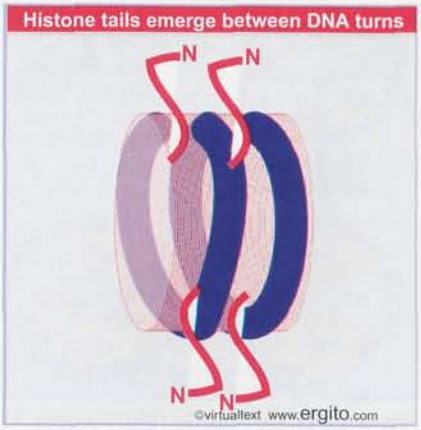


Figure 20.25 The N-terminal histone tails are disordered and exit from the nucleosome between turns of the DNA.

By Book_Crazy [IND]

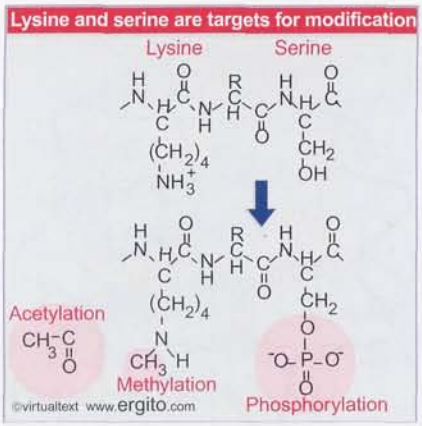


Figure 20.26 Acetylation of lysine or phosphorylation of serine reduces the overall positive charge of a protein.

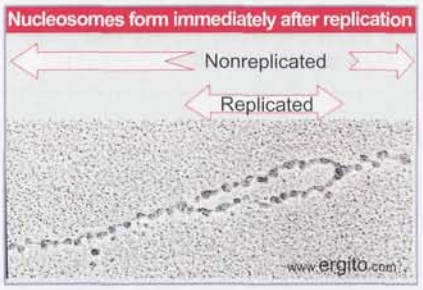


Figure 20.27 Replicated DNA is immediately incorporated into nucleosomes. Photograph kindly provided by S. MacKnight.

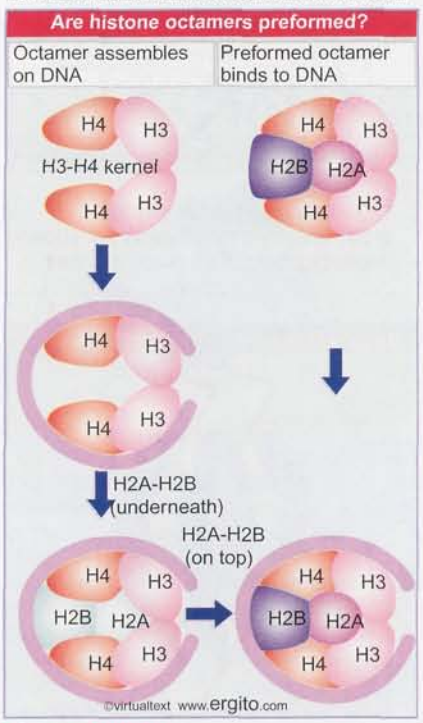


Figure 20.28 *In vitro*, DNA can either interact directly with an intact (crosslinked) histone octamer or can assemble with the $H3_2$-$H4_2$ tetramer, after which two H2A-H2B dimers are added.

The timing of the major H1 phosphorylation has prompted speculation that it is involved in mitotic condensation. However, in *Tetrahymena* (a protozoan) it is possible to delete all the genes for H1 without significantly affecting the overall properties of chromatin. There is a relatively small effect on the ability of chromatin to condense at mitosis. Some genes are activated and others are repressed by this change, suggesting that there are alterations in local structure. Mutations that eliminate sites of phosphorylation in H1 have no effect, but mutations that mimic the effects of phosphorylation produce a phenotype that resembles the deletion. This suggests that the effect of phosphorylating H1 is to eliminate its effects on local chromatin structure.

20.10 Reproduction of chromatin requires assembly of nucleosomes

Key Concepts

- Histone octamers are not conserved during replication, but H2A·H2B dimers and $H3_2$·$H4_2$ tetramers are conserved.
- There are different pathways for the assembly of nucleosomes during replication and independently of replication.
- Accessory proteins are required to assist the assembly of nucleosomes.
- CAF-1 is an assembly protein that is linked to the PCNA subunit of the replisome; it is required for deposition of $H3_2$·$H4_2$ tetramers following replication.
- A different assembly protein and a variant of histone H3 may be used for replication-independent assembly.

Replication separates the strands of DNA and therefore must inevitably disrupt the structure of the nucleosome. The transience of the replication event is a major difficulty in analyzing the structure of a particular region while it is being replicated. The structure of the replication fork is distinctive. It is more resistant to micrococcal nuclease and is digested into bands that differ in size from nucleosomal DNA. The region that shows this altered structure is confined to the immediate vicinity of the replication fork. This suggests that a large protein complex is engaged in replicating the DNA, but the nucleosomes reform more or less immediately behind as it moves along.

Reproduction of chromatin does not involve any protracted period during which the DNA is free of histones. Once DNA has been replicated, nucleosomes are quickly generated on both the duplicates. This point is illustrated by the electron micrograph of **Figure 20.27,** which shows a recently replicated stretch of DNA, already packaged into nucleosomes on both daughter duplex segments.

Both biochemical analysis and visualization of the replication fork therefore suggest that the disruption of nucleosome structure is limited to a short region immediately around the fork. Progress of the fork disrupts nucleosomes, but they form very rapidly on the daughter duplexes as the fork moves forward. In fact, the assembly of nucleosomes is directly linked to the replisome that is replicating DNA.

How do histones associate with DNA to generate nucleosomes? Do the histones *preform* a protein octamer around which the DNA is subsequently wrapped? Or does the histone octamer assemble on DNA from free histones? **Figure 20.28** shows that two pathways can be used *in vitro* to assemble nucleosomes, depending on the conditions that are employed. In one pathway, a preformed octamer binds to DNA. In the other pathway, a tetramer of $H3_2$·$H4_2$ binds first, and then two

$H2A{\cdot}H2B$ dimers are added. Both these pathways are related to reactions that occur *in vivo*. The first reflects the capacity of chromatin to be remodeled by moving histone octamers along DNA (see *23.3 Chromatin remodeling is an active process).* The second represents the pathway that is used in replication.

Accessory proteins are involved in assisting histones to associate with DNA. Candidates for this role can be identified by using extracts that assemble histones and exogenous DNA into nucleosomes. Accessory proteins may act as "molecular chaperones" that bind to the histones in order to release either individual histones or complexes ($H3_2{\cdot}H4_2$ or $H2A{\cdot}H2B$) to the DNA in a controlled manner. This could be necessary because the histones, as basic proteins, have a general high affinity for DNA. *Such interactions allow histones to form nucleosomes without becoming trapped in other kinetic intermediates (that is, other complexes resulting from indiscreet binding of histones to DNA).*

Attempts to produce nucleosomes *in vitro* began by considering a process of assembly between free DNA and histones. But nucleosomes form *in vivo* only when DNA is replicated. A system that mimics this requirement has been developed by using extracts of human cells that replicate SV40 DNA and assemble the products into chromatin. The assembly reaction occurs preferentially on replicating DNA. It requires an ancillary factor, CAF-1, that consists of >5 subunits, with a total mass of 238 kD. CAF-1 is recruited to the replication fork by PCNA, the processivity factor for DNA polymerase. This provides the link between replication and nucleosome assembly, ensuring that nucleosomes are assembled as soon as DNA has been replicated.

CAF-1 acts stoichiometrically, and functions by binding to newly synthesized H3 and H4. This suggests that new nucleosomes form by assembling first the $H3_2{\cdot}H4_2$ tetramer, and then adding the $H2A{\cdot}H2B$ dimers. The nucleosomes that are formed *in vitro* have a repeat length of 200 bp, although they do not have any H1 histone, which suggests that proper spacing can be accomplished without H1.

When chromatin is reproduced, a stretch of DNA *already associated with nucleosomes* is replicated, giving rise to two daughter duplexes. What happens to the pre-existing nucleosomes at this point? Are the histone octamers dissociated into free histones for reuse, or do they remain assembled? The integrity of the octamer can be tested by crosslinking the histones. The next two figures compare the possible outcomes from an experiment in which cells are grown in the presence of heavy amino acids to identify the histones before replication. Then replication is allowed to occur in the presence of light amino acids. At this point the histone octamers are crosslinked and centrifuged on a density gradient. **Figure 20.29** shows that if the original octamers have been conserved, they will be found at a position of high density, and new octamers will occupy a low density position. However, this does not happen. Little material is found at the high density position, which suggests that histone octamers are not conserved. The octamers have an intermediate density, and **Figure 20.30** shows that this is the expected result if the old histones have been released and then reassembled with newly synthesized histones.

The pattern of disassembly and reassembly has been difficult to characterize in detail, but our working model is illustrated in **Figure 20.31.** The replication fork displaces histone octamers, which then dissociate into $H3_2{\cdot}H4_2$ tetramers and $H2A{\cdot}H2B$ dimers. These "old" tetramers and dimers enter a pool that also includes "new" tetramers and dimers, assembled from newly synthesized histones. Nucleosomes assemble ~600 bp behind the replication fork. Assembly is initiated when $H3_2{\cdot}H4_2$ tetramers bind to each of the daughter duplexes, assisted by CAF-1. Then two $H2A{\cdot}H2B$ dimers bind to each $H3_2{\cdot}H4_2$ tetramer to complete the histone octamer. The assembly of tetramers and dimers is random with

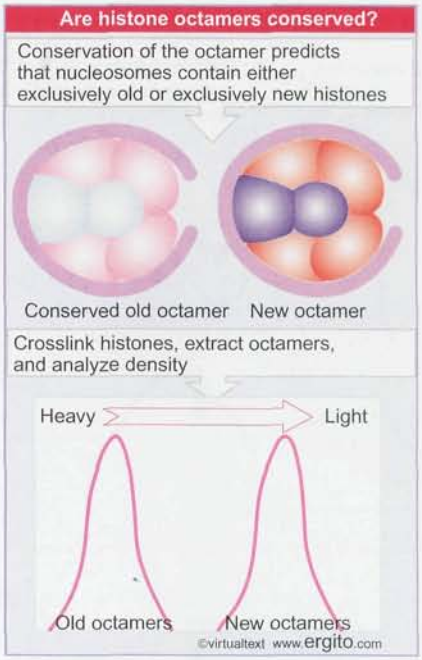


Figure 20.29 If histone octamers were conserved, old and new octamers would band at different densities when replication of heavy octamers occurs in light amino acids.

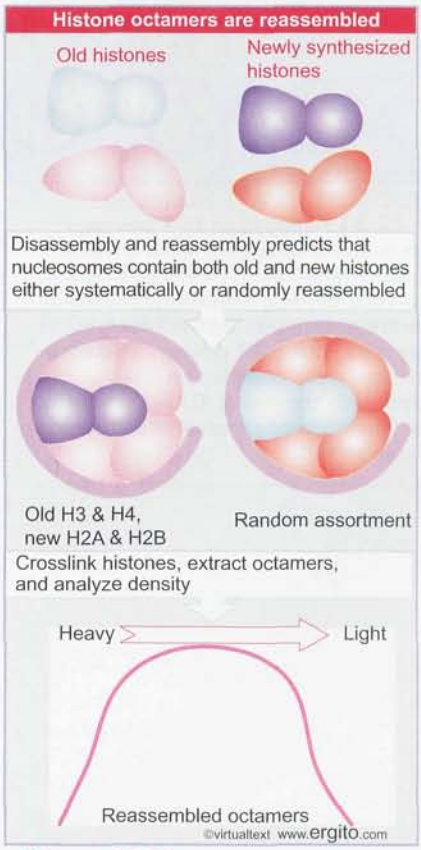


Figure 20.30 When heavy octamers are replicated in light amino acids, the new octamers band diffusely between heavy and light densities, suggesting that disassembly and reassembly has occurred.

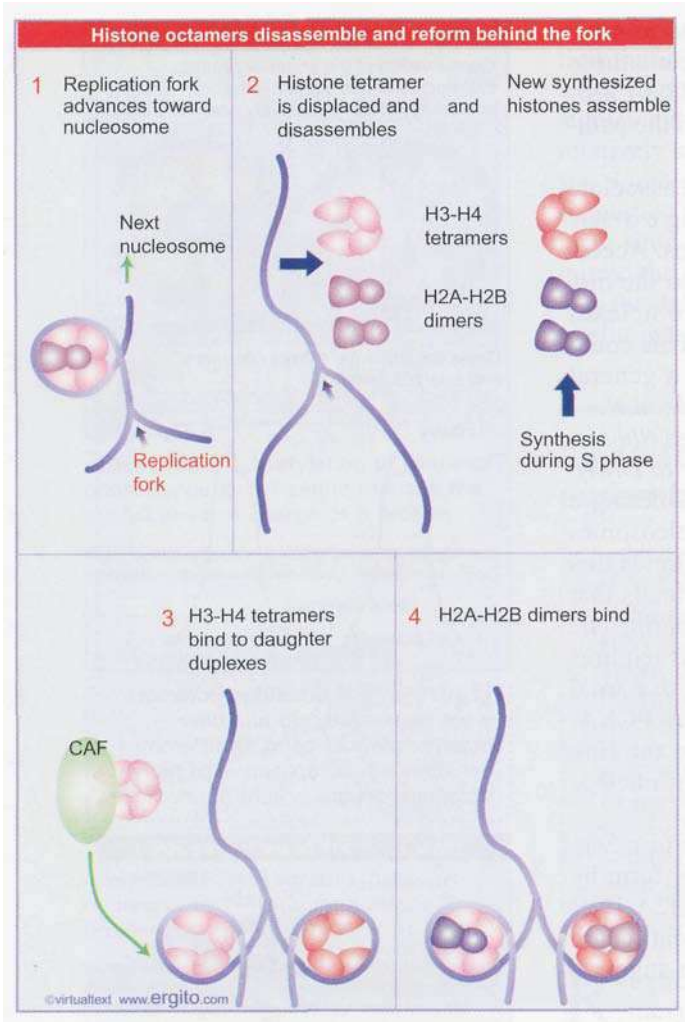


Figure 20.31 Replication fork passage displaces histone octamers from DNA. They disassemble into H3-H4 tetramers and H2A-H2B dimers. Newly synthesized histones are assembled into H3-H4 tetramers and H2A-H2B dimers. The old and new tetramers and dimers are assembled with the aid of CAF-1 at random into new nucleosomes immediately behind the replication fork.

respect to "old" and "new" subunits, explaining the results of Figure 20.30. The "old" $H3_2{\cdot}H4_2$ tetramer could have an ability to be transiently associated with a single strand of DNA during replication; it may in fact have an increased chance of remaining with the leading strand for reuse. It is possible that nucleosomes are disrupted and reassembled in a similar way during transcription (see *20.12 Are transcribed genes organized in nucleosomes?)*.

During S phase (the period of DNA replication) in a eukaryotic cell, the duplication of chromatin requires synthesis of sufficient histone proteins to package an entire genome—basically the same quantity of histones must be synthesized that are already contained in nucleosomes. The synthesis of histone mRNAs is controlled as part of the cell cycle, and increases enormously in S phase. The pathway for assembling chromatin from this equal mix of old and new histones during S phase is called the replication-coupled (RC) pathway.

Another pathway, called the replication-independent (RI) pathway exists for assembling nucleosomes during other phases of cell cycle, when DNA is not being synthesized. This may become necessary as the result of damage to DNA or because nucleosomes are displaced during transcription. The assembly process must necessarily have some differences from the replication-coupled pathway, because it cannot be linked to the replication apparatus. One of the most interesting features of the replication-independent pathway is that it uses different variants of some of the histones from those used during replication.

The histone H3.3 variant differs from the highly conserved H3 histone at 4 amino acid positions. H3.3 slowly replaces H3 in differentiating cells that do not have replication cycles. This happens as the result of assembly of new histone octamers to replace those that have been displaced from DNA for whatever reason. The mechanism that is used to ensure the use of H3.3 in the replication-independent pathway is different in two cases that have been investigated.

In the protozoan *Tetrahymena*, histone usage is determined exclusively by availability. Histone H3 is synthesized only during the cell cycle; the variant replacement histone is synthesized only in nonreplicating cells. In *Drosophila*, however, there is an active pathway that ensures the usage of H3.3 by the replication-independent pathway. New nucleosomes containing H3.3 assemble at sites of transcription, presumably replacing nucleosome that were displaced by RNA polymerase. The assembly process discriminates between H3 and H3.3 on the basis of their sequences, specifically excluding H3 from being utilized. By contrast, replication-coupled assembly uses both types of H3 (although H3.3 is available at much lower levels than H3, and therefore enters only a small proportion of nucleosomes).

CAF-1 is probably not involved in replication-independent assembly. (And there are organisms such as yeast and *Arabidopsis* where its gene is not essential, implying that alternative assembly processes may be used in replication-coupled assembly). A protein that may be involved in replication-independent assembly is called HIRA. Depletion of HIRA from *in vitro* systems for nucleosome assembly inhibits the formation of nucleosomes on nonreplicated DNA, but not on replicating DNA, indicating that the pathways do indeed use different assembly mechanisms.

Assembly of nucleosomes containing an alternative to H3 also occurs at centromeres (see *23.15 Heterochromatin depends on interactions with histones)*. Centromeric DNA replicates early during the

replication phase of the cell cycle (in contrast with the surrounding heterochromatic sequences that replicate later; see 75.5 *Each eukaryotic chromosome contains many replicons).* The incorporation of H3 at the centromeres is inhibited, and instead a protein called CENP-A is incorporated in higher eukaryotic cells (in *Drosophila* it is called Cid, and in yeast it is called Cse4p). This occurs by the replication-independent assembly pathway, apparently because the replication-coupled pathway is inhibited for a brief period of time while centromeric DNA replicates.

20.11 Do nucleosomes lie at specific positions?

Key Concepts

- Nucleosomes may form at specific positions as the result either of the local structure of DNA or of proteins that interact with specific sequences.
- The most common cause of nucleosome positioning is when proteins binding to DNA establish a boundary.
- Positioning may affect which regions of DNA are in the linker and which face of DNA is exposed on the nucleosome surface.

We know that nucleosomes can be reconstituted *in vitro* without regard to DNA sequence, but this does not mean that their formation *in vivo* is independent of sequence. Does a particular DNA sequence always lie in a certain position *in vivo* with regard to the topography of the nucleosome? Or are nucleosomes arranged randomly on DNA, so that a particular sequence may occur at any location, for example, in the core region in one copy of the genome and in the linker region in another?

To investigate this question, it is necessary to use a defined sequence of DNA; more precisely, we need to determine the position relative to the nucleosome of a defined point in the DNA. **Figure 20.32** illustrates the principle of a procedure used to achieve this.

Suppose that the DNA sequence is organized into nucleosomes in only one particular configuration, so that each site on the DNA always is located at a particular position on the nucleosome. This type of organization is called nucleosome positioning (or sometimes nucleosome phasing). In a series of positioned nucleosomes, the linker regions of DNA comprise unique sites.

Consider the consequences for just a single nucleosome. Cleavage with micrococcal nuclease generates a monomeric fragment that constitutes a *specific sequence.* If the DNA is isolated and cleaved with a restriction enzyme that has only one target site in this fragment, it should be cut at a unique point. This produces two fragments, each of unique size.

The products of the micrococcal/restriction double digest are separated by gel electrophoresis. A probe representing the sequence on one side of the restriction site is used to identify the corresponding fragment in the double digest. This technique is called indirect end labeling.

Reversing the argument, the identification of a single sharp band demonstrates that the position of the restriction site is uniquely defined with respect to the end of the nucleosomal DNA (as defined by the micrococcal nuclease cut). So the nucleosome has a unique sequence of DNA.

What happens if the nucleosomes do *not* lie at a single position? Now the linkers consist of *different* DNA sequences in each copy of the genome. So the restriction site lies at a different position each time;

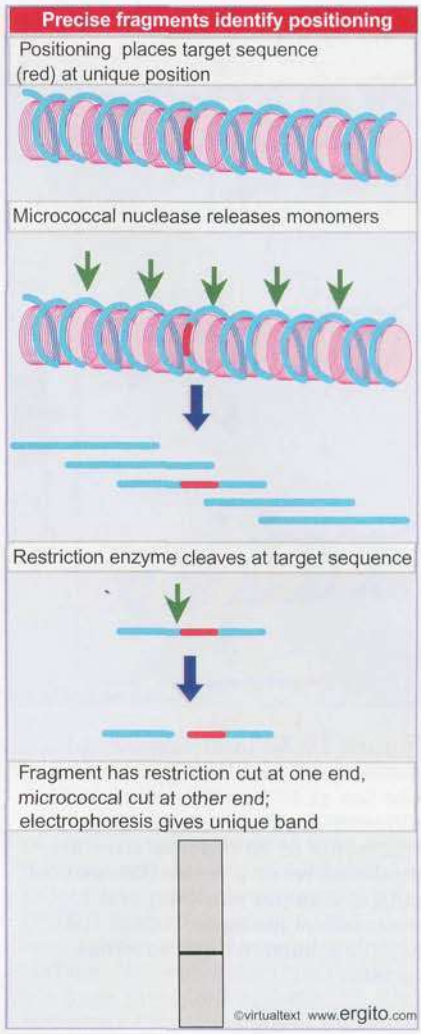


Figure 20.32 Nucleosome positioning places restriction sites at unique positions relative to the linker sites cleaved by micrococcal nuclease.

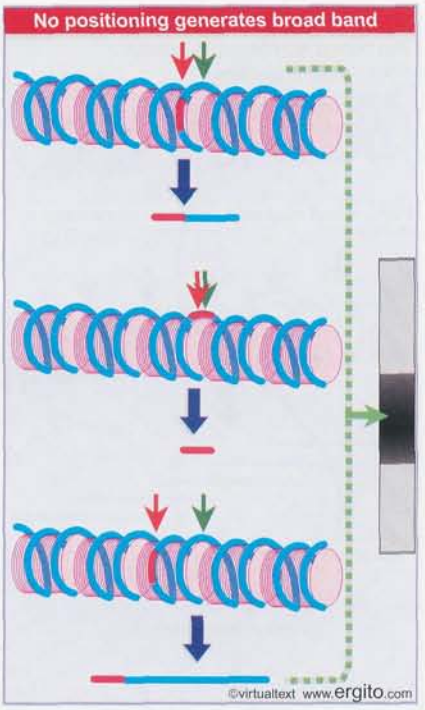


Figure 20.33 In the absence of nucleosome positioning, a restriction site lies at all possible locations in different copies of the genome. Fragments of all possible sizes are produced when a restriction enzyme cuts at a target site (red) and micrococcal nuclease cuts at the junctions between nucleosomes (green).

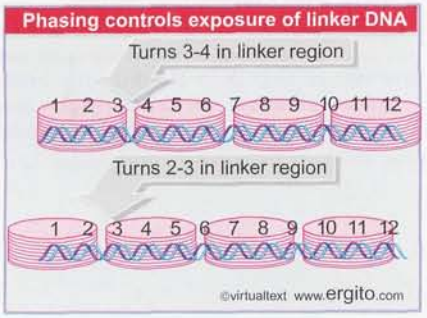


Figure 20.34 Translational positioning describes the linear position of DNA relative to the histone octamer. Displacement of the DNA by 10 bp changes the sequences that are in the more exposed linker regions, but does not alter which face of DNA is protected by the histone surface and which is exposed to the exterior. DNA is really coiled around the nucleosomes, and is shown in linear form only for convenience.

in fact, it lies at all possible locations relative to the ends of the monomeric nucleosomal DNA. **Figure 20.33** shows that the double cleavage then generates a broad smear, ranging from the smallest detectable fragment (~20 bases) to the length of the monomeric DNA.

In discussing these experiments, we have treated micrococcal nuclease as an enzyme that cleaves DNA at the exposed linker regions without any sort of sequence specificity. However, the enzyme actually does have some sequence specificity (biased toward selection of A·T-rich sequences). So we cannot assume that the existence of a specific band in the indirect end-labeling technique represents the distance from a restriction cut to the linker region. It could instead represent the distance from the restriction cut to a preferred micrococcal nuclease cleavage site!

This possibility is controlled by treating the naked DNA in exactly the same way as the chromatin. If there are preferred sites for micrococcal nuclease in the particular region, specific bands are found. Then this pattern of bands can be compared with the pattern generated from chromatin.

A *difference* between the control DNA band pattern and the chromatin pattern provides evidence for nucleosome positioning. Some of the bands present in the control DNA digest may disappear from the nucleosome digest, indicating that preferentially cleaved positions are unavailable. New bands may appear in the nucleosome digest when new sites are rendered preferentially accessible by the nucleosomal organization.

Nucleosome positioning might be accomplished in either of two ways:

- It is intrinsic: *every nucleosome is deposited specifically at a particular DNA sequence.* This modifies our view of the nucleosome as a subunit able to form between any sequence of DNA and a histone octamer.
- It is extrinsic: *the first nucleosome in a region is preferentially assembled at a particular site.* A preferential starting point for nucleosome positioning results from the presence of a region from which nucleosomes are excluded. The excluded region provides a *boundary* that restricts the positions available to the adjacent nucleosome. Then a series of nucleosomes may be assembled sequentially, with a defined repeat length.

It is now clear that the deposition of histone octamers on DNA is not random with regard to sequence. The pattern is intrinsic in some cases, in which it is determined by structural features in DNA. It is extrinsic in other cases, in which it results from the interactions of other proteins with the DNA and/or histones.

Certain structural features of DNA affect placement of histone octamers. DNA has intrinsic tendencies to bend in one direction rather than another; thus A-T-rich regions locate so that the minor groove faces in towards the octamer, whereas G·C-rich regions are arranged so that the minor groove points out. Long runs of dA·dT (>8 bp) avoid positioning in the central superhelical turn of the core. It is not yet possible to sum all of the relevant structural effects and thus entirely to predict the location of a particular DNA sequence with regard to the nucleosome. Sequences that cause DNA to take up more extreme structures may have effects such as the exclusion of nucleosomes, and thus could cause boundary effects.

Positioning of nucleosomes near boundaries is common. If there is some variability in the construction of nucleosomes—for example, if the length of the linker can vary by, say, 10 bp—the specificity of location would decline proceeding away from the first, defined nucleosome

at the boundary. In this case, we might expect the positioning to be maintained rigorously only relatively near the boundary.

The location of DNA on nucleosomes can be described in two ways. **Figure 20.34** shows that translational positioning describes the position of DNA with regard to the boundaries of the nucleosome. In particular, it determines which sequences are found in the linker regions. Shifting the DNA by 10 bp brings the next turn into a linker region. So translational positioning determines which regions are more accessible (at least as judged by sensitivity to micrococcal nuclease).

Because DNA lies on the outside of the histone octamer, one face of any particular sequence is obscured by the histones, but the other face is accessible. Depending upon its positioning with regard to the nucleosome, a site in DNA that must be recognized by a regulator protein could be inaccessible or available. The exact position of the histone octamer with respect to DNA sequence may therefore be important. **Figure 20.35** shows the effect of rotational positioning of the double helix with regard to the octamer surface. If the DNA is moved by a partial number of turns (imagine the DNA as rotating relative to the protein surface), there is a change in the exposure of sequence to the outside.

Both translational and rotational positioning can be important in controlling access to DNA. The best characterized cases of positioning involve the specific placement of nucleosomes at promoters. Translational positioning and/or the exclusion of nucleosomes from a particular sequence may be necessary to allow a transcription complex to form. Some regulatory factors can bind to DNA only if a nucleosome is excluded to make the DNA freely accessible, and this creates a boundary for translational positioning. In other cases, regulatory factors can bind to DNA on the surface of the nucleosome, but rotational positioning is important to ensure that the face of DNA with the appropriate contact points is exposed. We discuss the connection between nucleosomal organization and transcription in *23.4 Nucleosome organization may be changed at the promoter.*

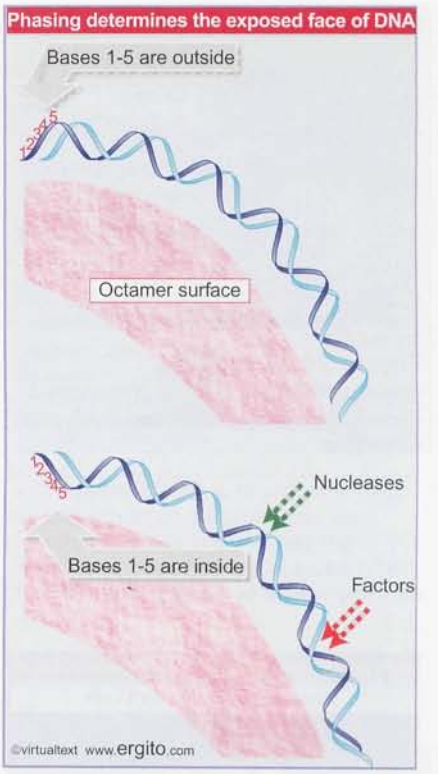


Figure 20.35 Rotational positioning describes the exposure of DNA on the surface of the nucleosome. Any movement that differs from the helical repeat (~10.2 bp/turn) displaces DNA with reference to the histone surface. Nucleotides on the inside are more protected against nucleases than nucleotides on the outside.

20.12 Are transcribed genes organized in nucleosomes?

Key Concepts

- Nucleosomes are found at the same frequency when transcribed genes or nontranscribed genes are digested with micrococcal nuclease.
- Some heavily transcribed genes appear to be exceptional cases that are devoid of nucleosomes.

Attempts to visualize genes during transcription have produced conflicting results. The next two figures show each extreme.

Heavily transcribed chromatin can be seen to be rather extended (too extended to be covered in nucleosomes). In the intensively transcribed genes coding for rRNA, shown in **Figure 20.36,** the extreme packing of RNA polymerases makes it hard to see the DNA. We cannot directly measure the lengths of the rRNA transcripts because the RNA is compacted by proteins, but we know (from the sequence of the rRNA) how long the transcript must be. The length of the transcribed DNA segment, measured by the length of the axis of the "Christmas tree," is ~85% of the length of the rRNA. This means that the DNA is almost completely extended.

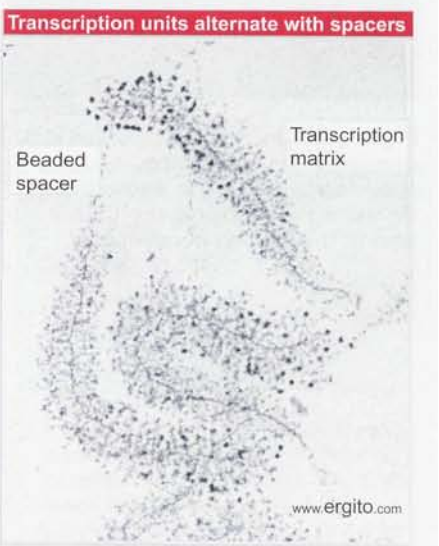


Figure 20.36 The extended axis of an rDNA transcription unit alternates with the only slightly less extended non-transcribed spacer. Photograph kindly provided by Charles Laird.

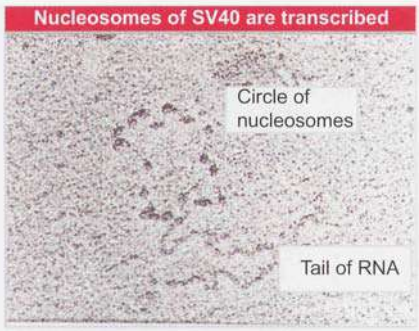


Figure 20.37 An SV40 minichromosome can be transcribed. Photograph kindly provided by Pierre Chambon.

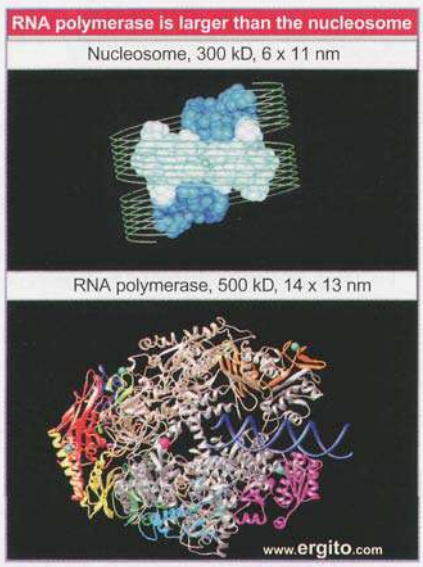


Figure 20.38 RNA polymerase is comparable in size to the nucleosome and might encounter difficulties in following the DNA around the histone octamer.

On the other hand, transcription complexes of SV40 minichromosomes can be extracted from infected cells. They contain the usual complement of histones and display a beaded structure. Chains of RNA can be seen to extend from the minichromosome, as in the example of **Figure 20.37.** This argues that transcription can proceed while the SV40 DNA is organized into nucleosomes. Of course, the SV40 minichromosome is transcribed less intensively than the rRNA genes.

Transcription involves the unwinding of DNA, and may require the fiber to unfold in restricted regions of chromatin. A simple-minded view suggests that some "elbow-room" must be needed for the process. The features of polytene and lampbrush chromosomes described in *19 Chromosomes* offer hints that a more expansive structural organization is associated with gene expression.

In thinking about transcription, we must bear in mind the relative sizes of RNA polymerase and the nucleosome. The eukaryotic enzymes are large multisubunit proteins, typically >500 kD. Compare this with the ~260 kD of the nucleosome. **Figure 20.38** illustrates the approach of RNA polymerase to nucleosomal DNA. Even without detailed knowledge of the interaction, it is evident that it involves the approach of two comparable bodies.

Consider the two turns that DNA makes around the nucleosome. Would RNA polymerase have sufficient access to DNA if the nucleic acid were confined to this path? During transcription, as RNA polymerase moves along the template, it binds tightly to a region of ~50 bp, including a locally unwound segment of ~12 bp. The need to unwind DNA makes it seem unlikely that the segment engaged by RNA polymerase could remain on the surface of the histone octamer.

It therefore seems inevitable that transcription must involve a structural change. So the first question to ask about the structure of active genes is whether DNA being transcribed remains organized in nucleosomes. If the histone octamers are displaced, do they remain attached in some way to the transcribed DNA?

One experimental approach is to digest chromatin with micrococcal nuclease, and then to use a probe to some specific gene or genes to determine whether the corresponding fragments are present in the usual 200 bp ladder at the expected concentration. The conclusions that we can draw from these experiments are limited but important. *Genes that are being transcribed contain nucleosomes at the same frequency as nontranscribed sequences.* So genes do not necessarily enter an alternative form of organization in order to be transcribed.

But since the average transcribed gene probably only has a single RNA polymerase at any given moment, this does not reveal what is happening at sites actually engaged by the enzyme. Perhaps they retain their nucleosomes; more likely the nucleosomes are temporarily displaced as RNA polymerase passes through, but reform immediately afterward.

20.13 Histone octamers are displaced by transcription

Key Concepts

- RNA polymerase displaces histone octamers during transcription in a model system, but octamers reassociate with DNA as soon as the polymerase has passed.
- Nucleosomes are reorganized when transcription passes through a gene.

Experiments to test whether an RNA polymerase can transcribe directly through a nucleosome suggest that the histone octamer is displaced by the act of transcription. **Figure 20.39** shows what happens when the phage T7 RNA polymerase transcribes a short piece of DNA containing a single octamer core *in vitro*. The core remains associated with the DNA, but is found in a different location. The core is most likely to rebind to the same DNA molecule from which it was displaced.

Figure 20.40 shows a model for polymerase progression. DNA is displaced as the polymerase enters the nucleosome, but the polymerase reaches a point at which the DNA loops back and reattaches, forming a closed region. As polymerase advances further, unwinding the DNA, it creates positive supercoils in this loop; the effect could be dramatic, because the closed loop is only ~80 bp, so each base pair through which the polymerase advances makes a significant addition to the supercoiling. In fact, the polymerase progresses easily for the first 30 bp into the nucleosome. Then it proceeds more slowly, as though encountering increasing difficulty in progressing. Pauses occur every 10 bp, suggesting that the structure of the loop imposes a constraint related to rotation around each turn of DNA. When the polymerase reaches the midpoint of the nucleosome (the next bases to be added are essentially at the axis of dyad symmetry), pausing ceases, and the polymerase advances rapidly. This suggests that the midpoint of the nucleosome marks the point at which the octamer is displaced (possibly because positive supercoiling has reached some critical level that expels the octamer from DNA). This releases tension ahead of the polymerase and allows it to proceed. The octamer then binds to the DNA behind the polymerase and no longer presents an obstacle to progress. Probably the octamer changes position without ever completely losing contact with the DNA.

Is the octamer released as an intact unit? Crosslinking the proteins of the octamer does not create an obstacle to transcription. Transcription can continue even when crosslinking is extensive enough to ensure that the central regions of the core histones have been linked. This implies that transcription does not require dissociation of the octamer into its component histones, nor is it likely to require any major unfolding of the central structure. However, addition of histone H1 to this system causes a rapid decline in transcription. This suggests two conclusions: the histone octamer (whether remaining present or displaced) functions as an intact unit; and it may be necessary to remove H1 from active chromatin or to modify its interactions in some way.

So a small RNA polymerase can displace a single nucleosome, which reforms behind it, during transcription. Of course, the situation is more complex in a eukaryotic nucleus. RNA polymerase is very much larger, and the impediment to progress is a string of connected nucleosomes. Overcoming this obstacle requires additional factors that act on chromatin (see *23 Controlling chromatin structure).*

The organization of nucleosomes may be changed by transcription. **Figure 20.41** shows what happens to the yeast *URA3* gene when it transcribed under control of an inducible promoter. Positioning is examined by using micrococcal nuclease to examine cleavage sites relative to a restriction site at the 5' end of the gene. Initially the gene displays a pattern of nucleosomes that are organized from the promoter for a significant distance across the gene; positioning is lost in the 3' regions. When the gene is expressed, a general smear replaces the positioned pattern of nucleosomes. So, nucleosomes are present at the same density but are no longer organized in phase. This suggests that transcription destroys the nucleosomal positioning. When repression is reestablished,

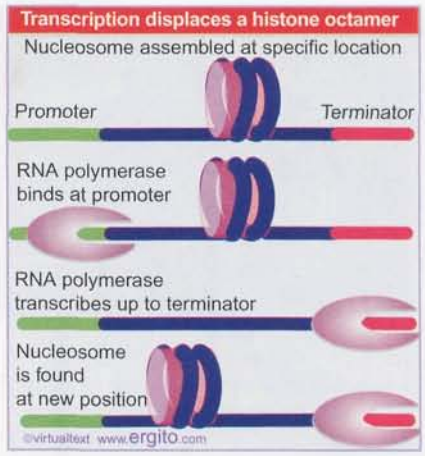


Figure 20.39 A protocol to test the effect of transcription on *nucleosomes* shows that the histone octamer is displaced from DNA and rebinds at a new position.

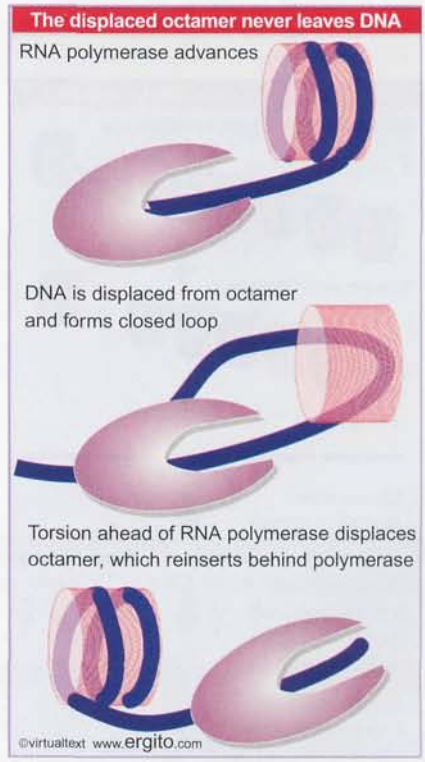


Figure 20.40 RNA polymerase displaces DNA from the histone octamer as it advances. The DNA loops back and attaches (to polymerase or to the octamer) to form a closed loop. As the polymerase proceeds, it generates positive supercoiling ahead. This displaces the octamer, which keeps contact with DNA and/or polymerase, and is inserted behind the RNA polymerase.

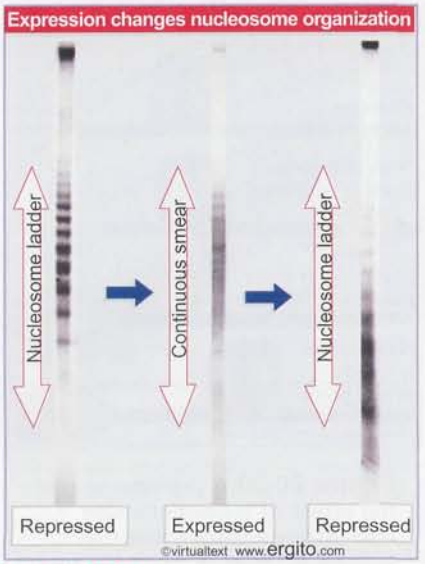


Figure 20.41 The *URA3* gene has translationally positioned nucleosomes before transcription. When transcription is induced, nucleosome positions are randomized. When transcription is repressed, the nucleosomes resume their particular positions. Photograph kindly provided by Fritz Thoma.

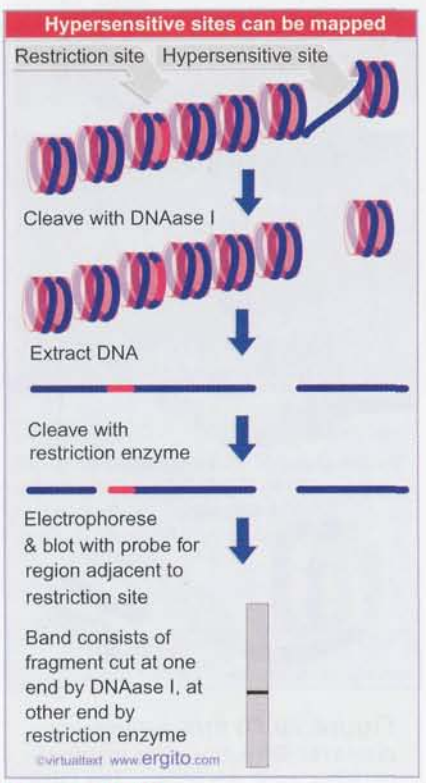


Figure 20.42 Indirect end-labeling identifies the distance of a DNAase hypersensitive site from a restriction cleavage site. The existence of a particular cutting site for DNAase I generates a discrete fragment, whose size indicates the distance of the DNAase I hypersensitive site from the restriction site.

positioning appears within 10 min (although it is not complete). This result makes the interesting point that the positions of the nucleosomes can be adjusted without replication.

The unifying model is to suppose that RNA polymerase displaces histone octamers as it progresses. If the DNA behind the polymerase is available, the octamer reattaches there (possibly or probably never having ever totally lost contact with the DNA. It remains a puzzle how an octamer could retain contact with DNA, without unfolding or losing components, as an object of even larger size than itself proceeds along the DNA. Perhaps the octamer is "passed back" by making contacts with RNA polymerase). If the DNA is not available, for example, because another polymerase continues immediately behind the first, then the octamer may be permanently displaced, and the DNA may persist in an extended form.

20.14 DNAase hypersensitive sites change chromatin structure

Key Concepts

- Hypersensitive sites are found at the promoters of expressed genes.
- They are generated by the binding of transcription factors that displace histone octamers.

In addition to the general changes that occur in active or potentially active regions, structural changes occur at specific sites associated with initiation of transcription or with certain structural features in DNA. These changes were first detected by the effects of digestion with very low concentrations of the enzyme DNAase I.

When chromatin is digested with DNAase I, the first effect is the introduction of breaks in the duplex at specific, hypersensitive sites. Since susceptibility to DNAase I reflects the availability of DNA in chromatin, we take these sites to represent chromatin regions in which the DNA is particularly exposed because it is not organized in the usual nucleosomal structure. A typical hypersensitive site is 100× more sensitive to enzyme attack than bulk chromatin. These sites are also hypersensitive to other nucleases and to chemical agents.

Hypersensitive sites are created by the (tissue-specific) structure of chromatin. Their locations can be determined by the technique of indirect end labeling that we introduced earlier in the context of nucleosome positioning. This application of the technique is recapitulated in **Figure 20.42.** In this case, cleavage at the hypersensitive site by DNAase I is used to generate one end of the fragment, and its distance is measured from the other end that is generated by cleavage with a restriction enzyme.

Many of the hypersensitive sites are related to gene expression. Every active gene has a site, or sometimes more than one site, in the region of the promoter. *Most hypersensitive sites are found only in chromatin of cells in which the associated gene is being expressed;* they do not occur when the gene is inactive. The hypersensitive site(s) appear before transcription begins; and the DNA sequences contained within the hypersensitive sites are required for gene expression, as seen by mutational analysis.

A particularly well-characterized nuclease-sensitive region lies on the SV40 minichromosome. A short segment near the origin of replication, just upstream of the promoter for the late transcription unit, is

cleaved preferentially by DNAase I, micrococcal nuclease, and other nucleases (including restriction enzymes).

The state of the SV40 minichromosome can be visualized by electron microscopy. In up to 20% of the samples, a "gap" is visible in the nucleosomal organization, as evident in **Figure 20.43.** The gap is a region of ~120 nm in length (about 350 bp), surrounded on either side by nucleosomes. The visible gap corresponds with the nuclease-sensitive region. This shows directly that increased sensitivity to nucleases is associated with the exclusion of nucleosomes.

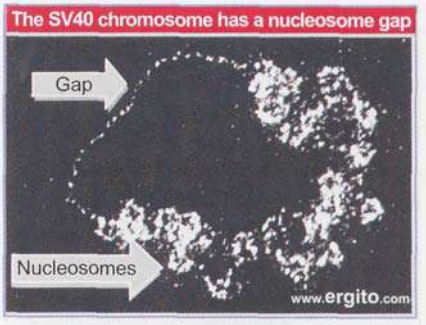


Figure 20.43 The SV40 minichromosome has a nucleosome gap. Photograph kindly provided by Moshe Yaniv.

A hypersensitive site is not necessarily uniformly sensitive to nucleases. **Figure 20.44** shows the maps of two hypersensitive sites.

Within the SV40 gap of ~300 bp, there are two hypersensitive DNAase I sites and a "protected" region. The protected region presumably reflects the association of (nonhistone) protein(s) with the DNA. The gap is associated with the DNA sequence elements that are necessary for promoter function.

The hypersensitive site at the β-globin promoter is preferentially digested by several enzymes, including DNAase I, DNAase II, and micrococcal nuclease. The enzymes have preferred cleavage sites that lie at slightly different points in the same general region. So a region extending from about —70 to —270 is preferentially accessible to nucleases when the gene is transcribable.

What is the structure of the hypersensitive site? Its preferential accessibility to nucleases indicates that it is not protected by histone octamers, but this does not necessarily imply that it is free of protein. A region of free DNA might be vulnerable to damage; and in any case, how would it be able to exclude nucleosomes? We assume that the hypersensitive site results from the binding of specific regulatory proteins that exclude nucleosomes. Indeed, the binding of such proteins is probably the basis for the existence of the protected region within the hypersensitive site.

The proteins that generate hypersensitive sites are likely to be regulatory factors of various types, since hypersensitive sites are found associated with promoters, other elements that regulate transcription, origins of replication, centromeres, and sites with other structural significance. In some cases, they are associated with more extensive organization of chromatin structure. A hypersensitive site may provide a boundary for a series of positioned nucleosomes. Hypersensitive sites associated with transcription may be generated by transcription factors when they bind to the promoter as part of the process that makes it accessible to RNA polymerase (see *23.4 Nucleosome organization may be changed at the promoter)*.

The stability of hypersensitive sites is revealed by the properties of chick fibroblasts transformed with temperature-sensitive tumor viruses. These experiments take advantage of an unusual property: although fibroblasts do not belong to the erythroid lineage, transformation of the cells at the normal temperature leads to activation of the globin genes. The activated genes have hypersensitive sites. If transformation is performed at the higher (nonpermissive) temperature, the globin genes are not activated; and hypersensitive sites do not appear. When the globin genes have been activated by transformation at low temperature, they can be inactivated by raising the temperature. But the hypersensitive sites are retained through at least the next 20 cell doublings.

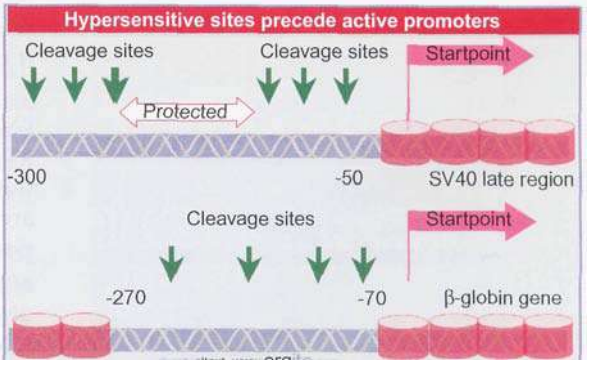


Figure 20.44 The SV40 gap includes hypersensitive sites, sensitive regions, and a protected region of DNA. The hypersensitive site of a chicken β-globin gene comprises a region that is susceptible to several nucleases.

This result demonstrates that acquisition of a hypersensitive site is only one of the features necessary to initiate transcription; and it implies that the events involved in establishing a hypersensitive site are distinct from those concerned with perpetuating it. Once the site has been established, it is perpetuated through replication in the absence of the circumstances needed for induction. Could some specific intervention be needed to abolish a hypersensitive site?

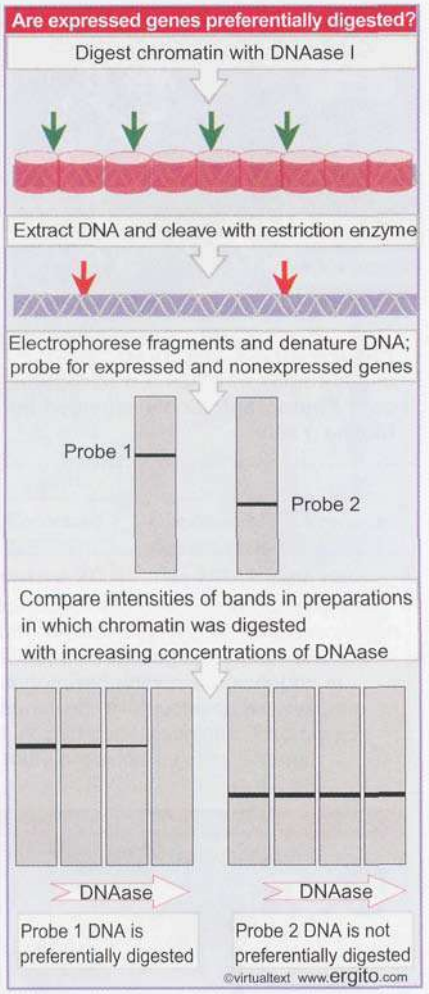


Figure 20.45 Sensitivity to DNAase I can be measured by determining the rate of disappearance of the material hybridizing with a particular probe.

20.15 Domains define regions that contain active genes

Key Concepts

- A domain containing a transcribed gene is defined by increased sensitivity to degradation by DNAase I.

A region of the genome that contains an active gene may have an altered structure. The change in structure precedes, and is different from, the disruption of nucleosome structure that may be caused by the actual passage of RNA polymerase.

One indication of the change in structure of transcribed chromatin is provided by its increased susceptibility to degradation by DNAase I. DNAase I sensitivity defines a chromosomal domain, a region of altered structure including at least one active transcription unit, and sometimes extending farther. (Note that use of the term "domain" does not imply any necessary connection with the structural domains identified by the loops of chromatin or chromosomes.)

When chromatin is digested with DNAase I, it is eventually degraded into acid-soluble material (very small fragments of DNA). The progress of the overall reaction can be followed in terms of the proportion of DNA that is rendered acid soluble. *When only 10% of the total DNA has become acid soluble, more than 50% of the DNA of an active gene has been lost.* This suggests that active genes are preferentially degraded.

The fate of individual genes can be followed by quantitating the amount of DNA that survives to react with a specific probe. The protocol is outlined in **Figure 20.45.** The principle is that the loss of a particular band indicates that the corresponding region of DNA has been degraded by the enzyme.

Figure 20.46 shows what happens to β-globin genes and an ovalbumin gene in chromatin extracted from chicken red blood cells (in which globin genes are expressed and the ovalbumin gene is inactive). The restriction fragments representing the β-globin genes are rapidly lost, while those representing the ovalbumin gene show little degradation. (The ovalbumin gene in fact is digested at the same rate as the bulk of DNA.)

So the bulk of chromatin is relatively resistant to DNAase I and contains nonexpressed genes (as well as other sequences). *A gene becomes relatively susceptible to the enzyme specifically in the tissue(s) in which it is expressed.*

Is preferential susceptibility a characteristic only of rather actively expressed genes, such as globin, or of all active genes? Experiments using probes representing the entire cellular mRNA population suggest that all active genes, whether coding for abundant or for rare mRNAs, are preferentially susceptible to DNAase I. (However, there are variations in the degree of susceptibility.) Since the rarely expressed genes are likely to have very few RNA polymerase molecules actually engaged in transcription at any moment, this implies that the sensitivity to DNAase I does not result from the act of transcription, but is a feature of *genes that are able to be transcribed.*

What is the extent of the preferentially sensitive region? This can be determined by using a series of probes representing the flanking regions as well as the transcription unit itself. The sensitive region always extends over the entire transcribed region; an additional region of several kb on either side may show an intermediate level of sensitivity (probably as the result of spreading effects).

The critical concept implicit in the description of the domain is that a region of high sensitivity to DNAase I extends over a considerable distance. Often we think of regulation as residing in events that occur at a discrete site in DNA—for example, in the ability to initiate transcription at the promoter. Even if this is true, such regulation must determine, or must be accompanied by, a more wide-ranging change in structure. This is a difference between eukaryotes and prokaryotes.

20.16 An LCR may control a domain

Key Concepts

- An LCR is located at the 5' end of the domain and consists of several hypersensitive sites.

Every gene is controlled by its promoter, and some genes also respond to enhancers (containing similar control elements but located farther away) as discussed in *21 Promoters and enhancers*. However, these local controls are not sufficient for all genes. In some cases, a gene lies within a domain of several genes all of which are influenced by regulatory elements that act on the whole domain. The existence of these elements was identified by the inability of a region of DNA including a gene and all its known regulatory elements to be properly expressed when introduced into an animal as a transgene.

The best characterized example of a regulated gene cluster is provided by the mouse β-globin genes. Recall from Figure 4.3 that the α globin and β-globin genes in mammals each exist as clusters of related genes, expressed at different times during embryonic and adult development. These genes are provided with a large number of regulatory elements, which have been analyzed in detail. In the case of the adult human β-globin gene, regulatory sequences are located both 5' and 3' to the gene and include both positive and negative elements in the promoter region, and additional positive elements within and downstream of the gene.

But a human β-globin gene containing all of these control regions is never expressed in a transgenic mouse within an order of magnitude of wild-type levels. Some further regulatory sequence is required. Regions that provide the additional regulatory function are identified by DNAase I hypersensitive sites that are found at the ends of the cluster. The map of **Figure 20.47** shows that the 20 kb upstream of the e-gene contains a group of 5 sites; and there is a single site 30 kb downstream of the β-gene. Transfecting various constructs into mouse erythroleukemia cells shows that sequences between the individual hypersensitive sites in the 5' region can be removed without much effect, but that removal of any of the sites reduces the overall level of expression.

The 5' regulatory sites are the primary regulators, and the cluster of hypersensitive sites is called the LCR (locus control region). We do not know whether the 3' site has any function. The LCR is absolutely required for expression of each of the globin genes in the cluster. Each gene is then further regulated by its own specific controls. Some of these controls are autonomous: expression of the e- and •γ-genes appears intrinsic to those loci in conjunction with the LCR. Other controls appear to rely upon position in the cluster, which provides a suggestion that *gene order* in a cluster is important for regulation.

The entire region containing the globin genes, and extending well beyond them, constitutes a chromosomal domain. It shows increased

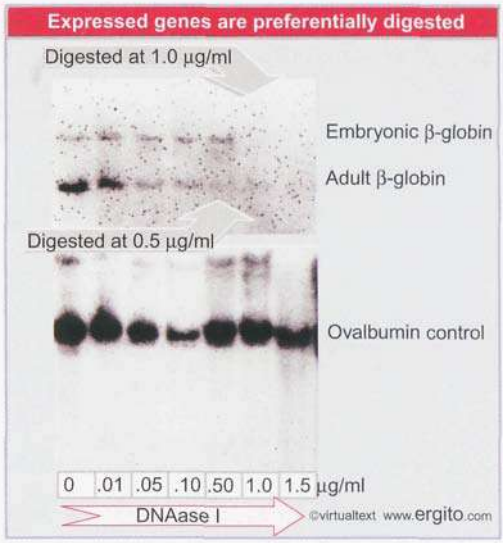


Figure 20.46 In adult erythroid cells, the adult β-globin gene is highly sensitive to DNAase I digestion, the embryonic β-globin gene is partially sensitive (probably due to spreading effects), but ovalbumin is not sensitive. Data kindly provided by Harold Weintraub.

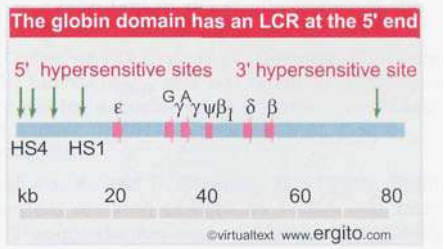


Figure 20.47 A globin domain is marked by hypersensitive sites at either end. The group of sites at the 5′ side constitutes the LCR and is essential for the function of all genes in the cluster.

sensitivity to digestion by DNAase I (see Figure 20.45). Deletion of the 5' LCR restores normal resistance to DNAase over the whole region. Two models for how an LCR works propose that its action is required in order to activate the promoter, or alternatively, to increase the rate of transcription from the promoter. The exact nature of the interactions between the LCR and the individual promoters has not yet been fully defined.

Does this model apply to other gene clusters? The α-globin locus has a similar organization of genes that are expressed at different times, with a group of hypersensitive sites at one end of the cluster, and increased sensitivity to DNAase I throughout the region. Only a small number of other cases are known in which an LCR controls a group of genes.

20.17 Summary

All eukaryotic chromatin consists of nucleosomes. A nucleosome contains a characteristic length of DNA, usually ~200 bp, wrapped around an octamer containing two copies each of histones H2A, H2B, H3, and H4. A single H1 protein is associated with each nucleosome. Virtually all genomic DNA is organized into nucleosomes. Treatment with micrococcal nuclease shows that the DNA packaged into each nucleosome can be divided operationally into two regions. The linker region is digested rapidly by the nuclease; the core region of 146 bp is resistant to digestion. Histones H3 and H4 are the most highly conserved and an $H3_2 \cdot H4_2$ tetramer accounts for the diameter of the particle. The H2A and H2B histones are organized as two H2AH2B dimers. Octamers are assembled by the successive addition of two H2AH2B dimers to the $H3_2 \cdot H4_2$ kernel.

The path of DNA around the histone octamer creates ~1.65 supercoils. The DNA "enters" and "leaves" the nucleosome in the same vicinity, and could be "sealed" by histone H1. Removal of the core histones releases -1.0 supercoils. The difference can be largely explained by a change in the helical pitch of DNA, from an average of 10.2 bp/turn in nucleosomal form to 10.5 bp/turn when free in solution. There is variation in the structure of DNA from a periodicity of 10.0 bp/turn at the nucleosome ends to 10.7 bp/turn in the center. There are kinks in the path of DNA on the nucleosome.

Nucleosomes are organized into a fiber of 30 nm diameter which has 6 nucleosomes per turn and a packing ratio of 40. Removal of H1 allows this fiber to unfold into a 10 nm fiber that consists of a linear string of nucleosomes. The 30 nm fiber probably consists of the 10 nm fiber wound into a solenoid. The 30 nm fiber is the basic constituent of both euchromatin and heterochromatin; nonhistone proteins are responsible for further organization of the fiber into chromatin or chromosome ultrastructure.

There are two pathways for nucleosome assembly. In the replication-coupled pathway, the PCNA processivity subunit of the replisome recruits CAF-1, which is a nucleosome assembly factor. CAF-1 assists the deposition of $H3_2 \cdot H4_2$ tetramers onto the daughter duplexes resulting from replication. The tetramers may be produced either by disruption of existing nucleosomes by the replication fork or as the result of assembly from newly synthesized histones. Similar sources provide the H2A·H2B dimers that then assemble with the $H3_2 \cdot H4_2$ tetramer to complete the nucleosome. Because the $H3_2 \cdot H4_2$ tetramer and the H2A·H2B dimers assemble at random, the new nucleosomes may include both pre-existing and newly synthesized histones.

RNA polymerase displaces histone octamers during transcription. Nucleosomes reform on DNA after the polymerase has passed, unless transcription is very intensive (such as in rDNA) when they may be displaced completely. The replication-independent pathway for nucleosome assembly is responsible for replacing histone

octamers that have been displaced by transcription. It uses the histone variant H3.3 instead of H3. A similar pathway, with another alternative to H3, is used for assembling nucleosomes at centromeric DNA sequences following replication.

Two types of changes in sensitivity to nucleases are associated with gene activity. Chromatin capable of being transcribed has a generally increased sensitivity to DNAase I, reflecting a change in structure over an extensive region that can be defined as a domain containing active or potentially active genes. Hypersensitive sites in DNA occur at discrete locations, and are identified by greatly increased sensitivity to DNAase I. A hypersensitive site consists of a sequence of ~200 bp from which nucleosomes are excluded by the presence of other proteins. A hypersensitive site forms a boundary that may cause adjacent nucleosomes to be restricted in position. Nucleosome positioning may be important in controlling access of regulatory proteins to DNA.

Hypersensitive sites occur at several types of regulators. Those that regulate transcription include promoters, enhancers, and LCRs. Other sites include origins for replication and centromeres. A promoter or enhancer acts on a single gene, but an LCR contains a group of hypersensitive sites and may regulate a domain containing several genes.

References

20.2 The nucleosome is the subunit of all chromatin

rev Kornberg, R. D. (1977). Structure of chromatin. Ann. Rev. Biochem. 46, 931-954.
McGhee, J., D. and Felsenfeld, G. (1980). Nucleosome structure. Ann. Rev. Biochem. 49, 1115-1156.

ref Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science 184, 868-871.
Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D., and Klug, A. (1984). Structure of the nucleosome core particle at 7 A resolution. Nature 311, 532-537.

20.3 DNA is coiled in arrays of nucleosomes

ref Finch, J. T. et al. (1977). Structure of nucleosome core particles of chromatin. Nature 269, 29-36.

20.4 Nucleosomes have a common structure

ref Shen, X. et al. (1995). Linker histones are not essential and affect chromatin condensation in vivo. Cell 82, 47-56.

20.5 DNA structure varies on the nucleosomal surface

rev Wang, J. (1982). The path of DNA in the nucleosome. Cell 29, 724-726.

20.6 The periodicity of DNA changes on the nucleosome

rev Travers, A. A. and Klug, A. (1987). The bending of DNA in nucleosomes and its wider implications. Philos Trans R Soc Lond B Biol Sci 317, 537-561.

20.7 The path of nucleosomes in the chromatin fiber

rev Felsenfeld, G. and McGhee, J. D. (1986). Structure of the 30 nm chromatin fiber. Cell 44, 375-377.

20.8 Organization of the histone octamer

ref Angelov, D., Vitolo, J. M., Mutskov, V., Dimitrov, S., and Hayes, J. J. (2001). Preferential interaction of the core histone tail domains with linker DNA. Proc. Nat. Acad. Sci. USA 98, 6599-6604.
Arents, G., Burlingame, R. W., Wang, B.-C, Love, W. E., and Moudrianakis, E. N. (1991). The nucleosomal core histone octamer at 31 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc. Nat. Acad. Sci. USA 88, 10148-10152.
Luger, K. et al. (1997). Crystal structure of the nucleosome core particle at 28 A resolution. Nature 389, 251-260.

20.10 Reproduction of chromatin requires assembly of nucleosomes

rev Osley, M. A. (1991). The regulation of histone synthesis in the cell cycle. Ann. Rev. Biochem. 60, 827-861.

ref Ahmad, K. and Henikoff, S. (2002). The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol. Cell 9, 1191-1200.
Ahmad, K. and Henikoff, S. (2001). Centromeres are specialized replication domains in heterochromatin. J. Cell Biol. 153, 101-110.
Gruss, C, Wu, J., Koller, T., and Sogo, J. M. (1993). Disruption of the nucleosomes at the replication fork. EMBO J. 12, 4533-4545.
Ray-Gallet, D., Quivy, J. P., Scamps, C, Martini, E. M., Lipinski, M., and Almouzni, G. (2002). HIRA is critical for a nucleosome assembly pathway independent of DNA synthesis. Mol. Cell 9, 1091-1100.
Shibahara, K. and B. Stillman, B. (1999). Replication-dependent marking of DNA by PCNA facilitates CAF-1-coupled inheritance of chromatin. Cell 96, 575-585.
Smith, S. and Stillman, B. (1989). Purification and characterization of CAF-I, a human cell factor required for chromatin assembly during DNA replication *in vitro*. Cell 58, 15-25.
Smith, S. and Stillman, B. (1991). Stepwise assembly of chromatin during DNA replication *in vitro*. EMBO J. 10, 971-980.

Verreault, A. (2000). De novo nucleosome assembly: new pieces in an old puzzle. Genes Dev. 14, 1430-1438.

Yu, L. and Gorovsky, M. A. (1997). Constitutive expression, not a particular primary sequence, is the important feature of the H3 replacement variant hv2 in Tetrahymena thermophila. Mol. Cell. Biol. 17, 6303-6310.

20.12 Are transcribed genes organized in nucleosomes?

rev Kornberg, R. D. and Lorch, Y. (1992). Chromatin structure and transcription. Ann. Rev. Cell Biol. 8, 563-587.

20.13 Histone octamers are displaced by transcription

ref Cavalli, G. and Thoma, F. (1993). Chromatin transitions during activation and repression of galactose-regulated genes in yeast. EMBO J. 12, 4603-4613.

Studitsky, V. M., Clark, D. J., and Felsenfeld, G. (1994). A histone octamer can step around a transcribing polymerase without leaving the template. Cell 76, 371-382.

20.14 DNAase hypersensitive sites change chromatin structure

rev Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. Ann. Rev. Biochem. 57, 159-197.

ref Groudine, M. and Weintraub H. (1982). Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell 30, 131-139.

Moyne, G., Harper, F., Saragosti, S., and Yaniv, M. (1982). Absence of nucleosomes in a histone-containing nucleoprotein complex obtained by dissociation of purified SV40 virions. Cell 30, 123-130.

Scott, W. A. and Wigmore, D. J. (1978). Sites in SV40 chromatin which are preferentially cleaved by endonucleases. Cell 15, 1511-1518.

Varshavsky, A. J., Sundin, O., and Bohn, M. J. (1978). SV40 viral minichromosome: preferential exposure of the origin of replication as probed by restriction endonucleases. Nuc. Acids Res. 5, 3469-3479.

20.15 Domains define regions that contain active genes

ref Stalder, J. et al. (1980). Tissue-specific DNA cleavage in the globin chromatin domain introduced by DNAase I. Cell 20, 451-460.

20.16 An LCR may control a domain

rev Bulger, M. and Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 13, 2465-2477.

Grosveld, F., Antoniou, M., Berry, M., De Boer, E., Dillon, N., Ellis, J., Fraser, P., Hanscombe, O., Hurst, J., and Imam, A. (1993). The regulation of human globin gene switching. Philos Trans R Soc Lond B Biol Sci 339, 183-191.

ref van Assendelft, G. B., Hanscombe, O., Grosveld, F., and Greaves, D. R. (1989). The β-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner. Cell 56, 969-977.

Gribnau, J., de Boer, E., Trimborn, T., Wijgerde, M., Milot, E., Grosveld, F., and Fraser, P. (1998). Chromatin interaction mechanism of transcriptional control in vivo. EMBO J. 17, 6020-6027.

Chapter 21

Promoters and enhancers

21.1 Introduction

Initiation of transcription requires the enzyme RNA polymerase and transcription factors. Any protein that is needed for the initiation of transcription, but which is not itself part of RNA polymerase, is defined as a transcription factor. Many transcription factors act by recognizing *cis*-acting sites on DNA. However, binding to DNA is not the only means of action for a transcription factor. A factor may recognize another factor, or may recognize RNA polymerase, or may be incorporated into an initiation complex only in the presence of several other proteins. The ultimate test for membership of the transcription apparatus is functional: a protein must be needed for transcription to occur at a specific promoter or set of promoters.

A significant difference between the transcription of eukaryotic and prokaryotic mRNAs is that initiation at a eukaryotic promoter involves a large number of factors that bind to a variety of *cis*-acting elements. The promoter is defined as the region containing all these binding sites, that is, which can support transcription at the normal efficiency and with the proper control. So the major feature defining the promoter for a eukaryotic mRNA is the location of binding sites for transcription factors. RNA polymerase itself binds around the startpoint, but does not directly contact the extended upstream region of the promoter. By contrast, the bacterial promoters discussed in *9 Transcription* are largely defined in terms of the binding site for RNA polymerase in the immediate vicinity of the startpoint.

Transcription in eukaryotic cells is divided into three classes. Each class is transcribed by a different RNA polymerase:

- RNA polymerase I transcribes rRNA
- RNA polymerase IT transcribes mRNA
- RNA polymerase III transcribes tRNA and other small RNAs.

Transcription factors are needed for initiation, but are not required subsequently. For the three eukaryotic enzymes, the *factors*, rather than the RNA polymerases themselves, are principally responsible for recognizing the promoter. This is different from bacterial RNA polymerase,

where it is the enzyme that recognizes the promoter sequences. For all eukaryotic RNA polymerases, the factors create a structure at the promoter to provide the target that is recognized by the enzyme. For RNA polymerases I and **III,** these factors are relatively simple, but for RNA polymerase II they form a sizeable group collectively known as the basal factors. The basal factors join with RNA polymerase **II** to form a complex surrounding the startpoint, and they determine the site of initiation. The basal factors together with RNA polymerase constitute the basal transcription apparatus.

The promoters for RNA polymerases I and II are (mostly) upstream of the startpoint, but some promoters for RNA polymerase **III** lie downstream of the startpoint. Each promoter contains characteristic sets of short conserved sequences that are recognized by the appropriate class of factors. RNA polymerases I and **III** each recognize a relatively restricted set of promoters, and rely upon a small number of accessory factors.

Promoters utilized by RNA polymerase **II** show more variation in sequence, and have a modular organization. Short sequence elements that are recognized by transcription factors lie upstream of the startpoint. These *cis-acting* sites usually are spread out over a region of >200 bp. Some of these elements and the factors that recognize them are common: they are found in a variety of promoters and are used constitutively. Others are specific: they identify particular classes of genes and their use is regulated. The elements occur in different combinations in individual promoters.

All RNA polymerase II promoters have sequence elements close to the startpoint that are bound by the basal apparatus and that establish the site of initiation. The sequences farther upstream determine whether the promoter is expressed in all cell types or is specifically regulated. Promoters that are constitutively expressed (their genes are sometimes called housekeeping genes) have upstream sequence elements that are recognized by ubiquitous activators. No element/factor combination is an essential component of the promoter, which suggests that initiation by RNA polymerase **II** may be sponsored in many different ways. Promoters that are expressed only in certain times or places have sequence elements that require activators that are available only at those times or places.

Sequence components of the promoter are defined operationally by the demand that they must be located in the general vicinity of the startpoint and are required for initiation. The enhancer is another type of site involved in initiation. It is identified by sequences that stimulate initiation, but that are located a considerable distance from the startpoint. Enhancer elements are often targets for tissue-specific or temporal regulation. **Figure 21.1** illustrates the general properties of promoters and enhancers.

The components of an enhancer resemble those of the promoter; they consist of a variety of modular elements. However, the elements

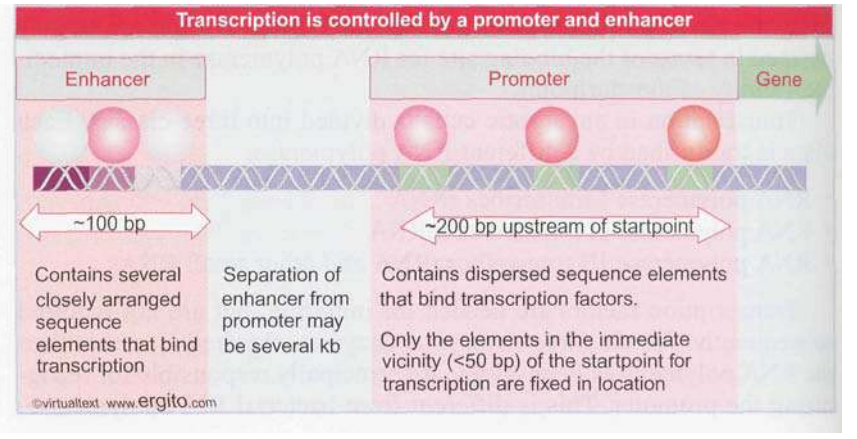


Figure 21.1 A typical gene transcribed by RNA polymerase II has a promoter that extends upstream from the site where transcription is initiated. The promoter contains several short (<10 bp) sequence elements that bind transcription factors, dispersed over >200 bp. An enhancer containing a more closely packed array of elements that also bind transcription factors may be located several kb distant. (DNA may be coiled or otherwise rearranged so that transcription factors at the promoter and at the enhancer interact to form a large protein complex.)

By Book_Crazy [IND]

are organized in a closely packed array. The elements in an enhancer function like those in the promoter, but the enhancer does not need to be near the startpoint. However, proteins bound at enhancer elements interact with proteins bound at promoter elements. The distinction between promoters and enhancers is operational, rather than implying a fundamental difference in mechanism. This view is strengthened by the fact that some types of element are found in both promoters and enhancers.

Eukaryotic transcription is most often under positive regulation: a transcription factor is provided under tissue-specific control to activate a promoter or set of promoters that contain a common target sequence. Regulation by specific repression of a target promoter is less common.

21.2 Eukaryotic RNA polymerases consist of many subunits

Key Concepts

- RNA polymerase I synthesizes rRNA in the nucleolus.
- RNA polymerase II synthesizes mRNA in the nucleoplasm.
- RNA polymerase III synthesizes small RNAs in the nucleoplasm.
- All eukaryotic RNA polymerases have ~ 12 subunits and are aggregates of >500 kD.
- Some subunits are common to all three RNA polymerases.
- The largest subunit in RNA polymerase II has a CTD (carboxy-terminal domain) consisting of multiple repeats of a septamer.

The three eukaryotic RNA polymerases have different locations in the nucleus, corresponding with the genes that they transcribe.

The most prominent activity is the enzyme RNA polymerase I, which resides in the nucleolus and is responsible for transcribing the genes coding for rRNA. It accounts for most cellular RNA synthesis (in terms of quantity).

The other major enzyme is RNA polymerase II, located in the nucleoplasm (the part of the nucleus excluding the nucleolus). It represents most of the remaining cellular activity and is responsible for synthesizing heterogeneous nuclear RNA (hnRNA), the precursor for mRNA.

RNA polymerase III is a minor enzyme activity. This nucleoplasmic enzyme synthesizes tRNAs and other small RNAs.

All eukaryotic RNA polymerases are large proteins, appearing as aggregates of >500 kD. They typically have ~**12** subunits. The purified enzyme can undertake template-dependent transcription of RNA, but is not able to initiate selectively at promoters. The general constitution of a eukaryotic RNA polymerase II enzyme as typified in *S. cerevisiae* is illustrated in **Figure 21.2.** The two largest subunits are homologous to the β and β′ subunits of bacterial RNA polymerase. Three of the remaining subunits are common to all the RNA polymerases, that is, they are also components of RNA polymerases I and **III.**

The largest subunit in RNA polymerase II has a carboxy-terminal domain (CTD), which consists of multiple repeats of a consensus sequence of 7 amino acids. The sequence is unique to RNA polymerase II. There are ~26 repeats in yeast and ~50 in mammals. The number of repeats is important, because deletions that remove (typically) more than half of the repeats are lethal (in yeast). The CTD can be highly phosphorylated on serine or threonine residues; this is involved in the initiation reaction (see *21.11 Initiation is followed by promoter clearance).*

The RNA polymerases of mitochondria and chloroplasts are smaller, and resemble bacterial RNA polymerase rather than any of the nuclear

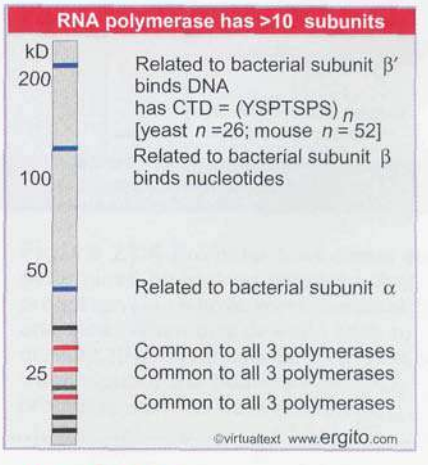


Figure 21.2 Some subunits are common to all classes of eukaryotic RNA polymerases and some are related to bacterial RNA polymerase.

enzymes. Of course, the organelle genomes are much smaller, the resident polymerase needs to transcribe relatively few genes, and the control of transcription is likely to be very much simpler (if existing at all). So these enzymes are analogous to the phage enzymes that do not need the ability to respond to a more complex environment.

A major practical distinction between the eukaryotic enzymes is drawn from their response to the bicyclic octapeptide α-amanitin. In basically all eukaryotic cells the activity of RNA polymerase II is rapidly inhibited by low concentrations of α-amanitin. RNA polymerase I is not inhibited. The response of RNA polymerase III to α-amanitin is less well conserved; in animal cells it is inhibited by high levels, but in yeast and insects it is not inhibited.

21.3 Promoter elements are defined by mutations and footprinting

Key Concepts

- Promoters are defined by their ability to cause transcription of an attached sequence in an appropriate test system *in vitro* or *in vivo.*

The first step in characterizing a promoter is to define the overall length of DNA that contains all the necessary sequence elements. To do this, we need a test system in which the promoter is responsible for the production of an easily assayed product. Historically, several types of systems have been used:

- In the *oocyte system,* a DNA template is injected into the nucleus of the *X. laevis* oocyte. The RNA transcript can be recovered and analyzed. The main limitation of this system is that it is restricted to the conditions that prevail in the oocyte. It allows characterization of DNA sequences, but not of the factors that normally bind them.
- *Transfection systems* allow exogenous DNA to be introduced into a cultured cell and expressed. (The procedure is discussed in *18.17 Transfection introduces exogenous DNA into cells.)* The system is genuinely *in vivo* in the sense that transcription is accomplished by the same apparatus responsible for expressing the cell's own genome. However, it differs from the natural situation because the template consists of a gene that would not usually be transcribed in the host cell. The usefulness of the system may be extended by using a variety of host cells.
- *Transgenic systems* involve the addition of a gene to the germline of an animal. Expression of the transgene can be followed in any or all of the tissues of the animal. Some common limitations apply to transgenic systems and to transfection: the additional gene often is present in multiple copies, and is integrated at a different location from the endogenous gene. Discrepancies between the expression of a gene *in vitro* and its expression as a transgene can yield important information about the role of the genomic context of the gene.
- The *in vitro* system takes the classic approach of purifying all the components and manipulating conditions until faithful initiation is seen. "Faithful" initiation is defined as production of an RNA starting at the site corresponding to the 5' end of mRNA (or rRNA or tRNA precursors). Ultimately this allows us to characterize the individual sequence elements in the promoter and the transcription factors that bind to them.

When a promoter is analyzed, it is important that *only* the promoter sequence changes. **Figure 21.3** shows that the same long upstream sequence is always placed next to the promoter to ensure that it is always in

the same context. Because termination does not occur properly in the *in vitro* systems, the template is cut at some distance from the promoter (usually ~500 bp downstream), to ensure that all polymerases "run off" at the same point, generating an identifiable transcript.

We start with a particular fragment of DNA that can initiate transcription in one of these systems. Then the boundaries of the sequence constituting the promoter can be determined by reducing the length of the fragment from either end, until at some point it ceases to be active, as illustrated in **Figure 21.4.** The boundary upstream can be identified by progressively removing material from this end until promoter function is lost. To test the boundary downstream, it is necessary to reconnect the shortened promoter to the sequence to be transcribed (since otherwise there is no product to assay).

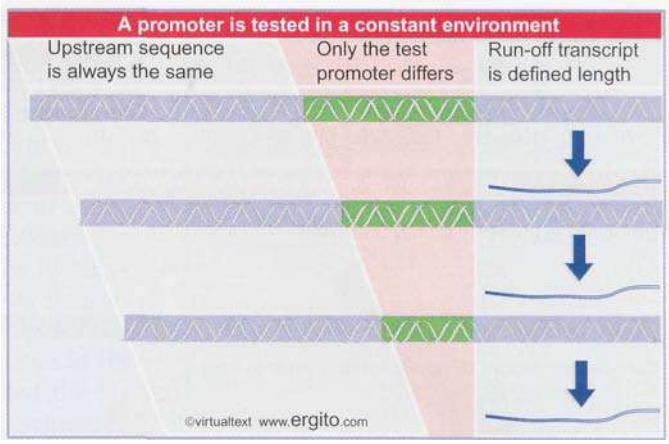


Figure 21.3 A promoter is tested by modifying the sequence that is connected to a constant upstream sequence and a constant downstream transcription unit.

Once the boundaries of the promoter have been defined, the importance of particular bases within it can be determined by introducing point mutations or other rearrangements in the sequence. As with bacterial RNA polymerase, these can be characterized as *up* or *down* mutations. Some of these rearrangements affect only the *rate* of initiation; others influence the *site* at which initiation occurs, as seen in a change of the startpoint. To be sure that we are dealing with comparable products, in each case it is necessary to characterize the 5' end of the RNA.

We can apply several criteria in identifying the sequence components of a promoter (or any other site in DNA):

- Mutations in the site prevent function *in vitro* or *in vivo*. (Many techniques now exist for introducing point mutations at particular base pairs, and in principle every position in a promoter can be mutated, and the mutant sequence tested *in vitro* or *in vivo.)*
- Proteins that act by binding to a site may be footprinted on it. There should be a correlation between the ability of mutations to prevent promoter function and to prevent binding of the factor.
- When a site recognized by a particular factor is present at multiple promoters, it should be possible to derive a consensus sequence that is bound by the factor. A new promoter should become responsive to this factor when an appropriate copy of the element is introduced.

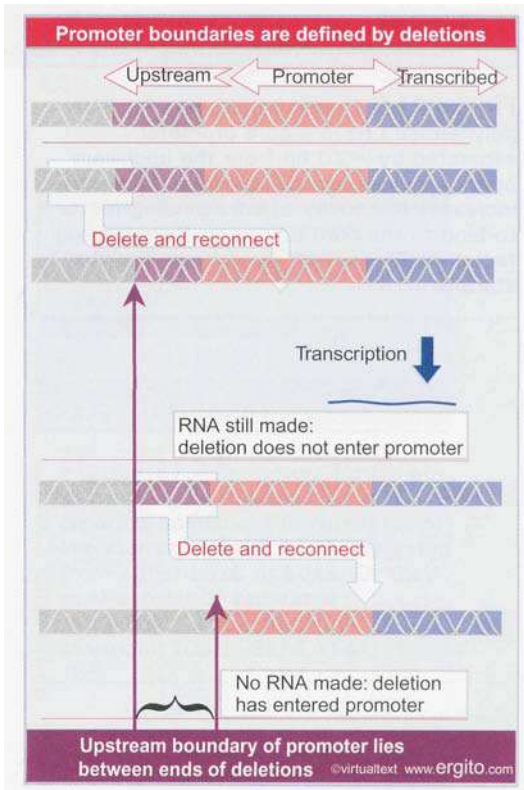


Figure 21.4 Promoter boundaries can be determined by making deletions that progressively remove more material from one side. When one deletion fails to prevent RNA synthesis but the next stops transcription, the boundary of the promoter must lie between them.

21.4 RNA polymerase I has a bipartite promoter

Key Concepts

- The RNA polymerase I promoter consists of a core promoter and an upstream control element.
- The factor UBF1 binds to both regions and enables the factor **SL1** to bind.
- SL1 includes the factor **TBP** that is involved in initiation by all three **RNA** polymerases.
- RNA polymerase binds to the UBF1-SL1 complex at the core promoter.

RNA polymerase I transcribes only the genes for ribosomal RNA, from a single type of promoter. The transcript includes the sequences of both large and small rRNAs, which are later released by cleavages and processing. There are many copies of the transcription unit, alternating with nontranscribed spacers, and organized in a cluster as discussed in *4.8*

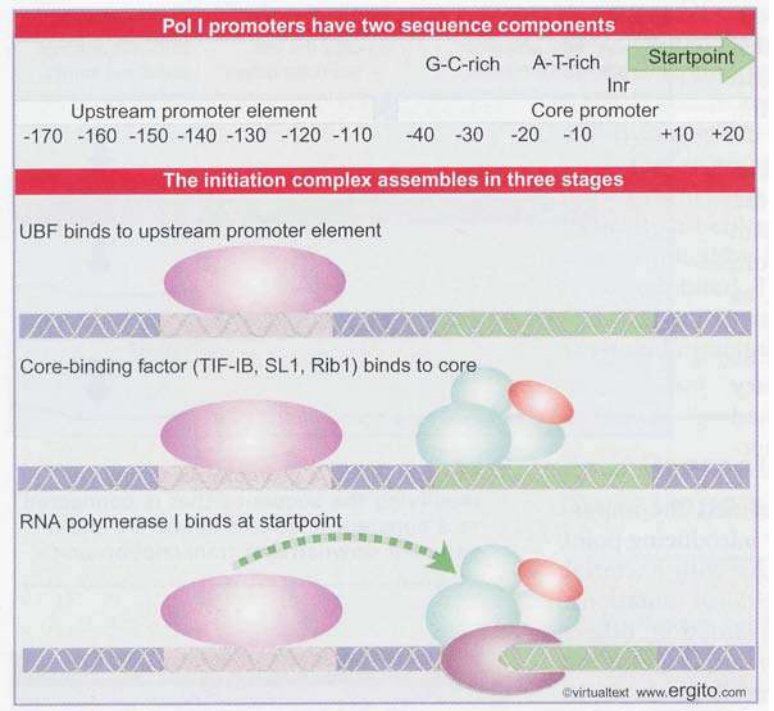


Figure 21.5 Transcription units for RNA polymerase I have a core promoter separated by ~70 bp from the upstream promoter element. UBF binding to the UPE increases the ability of core-binding factor to bind to the core promoter. Core-binding factor positions RNA polymerase I at the startpoint.

Genes for rRNA form tandem repeats. The organization of the promoter, and the events involved in initiation, are illustrated in **Figure 21.5.**

The promoter consists of two separate regions. The core promoter surrounds the startpoint, extending from -45 to +20, and is sufficient for transcription to initiate. It is generally G·C-rich (unusual for a promoter) except for the only conserved sequence element, a short A·T-rich sequence around the startpoint called the Inr. However, its efficiency is very much increased by the upstream promoter element (UPE), another G-C-rich sequence, related to the core promoter sequence, which extends from –180 to –107. This type of organization is common to pol I promoters in many species, although the actual sequences vary widely.

RNA polymerase I requires two ancillary factors. The factor that binds to the core promoter consists of 4 proteins. (It is called SL1, TIF-IB, Rib1 in different species). One of its components, called TBP, is a factor that is required also for initiation by RNA polymerases II and III (see *21.8 TBP is a universal factor).* TBP does not bind directly to G·C-rich DNA, so DNA-binding is probably the responsibility of the other components of the core-binding factor. It is likely that TBP interacts with RNA polymerase, possibly with a common subunit or a feature that has been conserved among polymerases. Core-binding factor enables RNA polymerase I to initiate from the promoter at a low basal frequency.

The core-binding factor has primary responsibility for ensuring that the RNA polymerase is properly localized at the startpoint. We see shortly that a comparable function is provided for RNA polymerases II and III by a factor that consists of TBP associated with other proteins. So a common feature in initiation by all three polymerases is a reliance on a "positioning" factor that consists of TBP associated with proteins that are specific for each type of promoter.

For high frequency initiation, the factor UBF is required. This is a single polypeptide that binds to a G-C-rich element in the upstream promoter element. One indication of how UBF interacts with the core-binding factor is given by the importance of the spacing between the upstream promoter element and the core promoter. This can be changed by distances involving integral numbers of turns of DNA, but not by distances that introduce half turns. This implies that UBF and core-binding factor need to be bound on the same face of DNA in order to interact. In the presence of UBF, core-binding factor binds more efficiently to the core promoter.

21.5 RNA polymerase III uses both downstream and upstream promoters

Key Concepts

- RNA polymerase III has two types of promoters.
- Internal promoters have short consensus sequences located within the transcription unit and cause initiation to occur a fixed distance upstream.
- Upstream promoters contain three short consensus sequences upstream of the startpoint that are bound by transcription factors.

Recognition of promoters by RNA polymerase III strikingly illustrates the relative roles of transcription factors and the polymerase enzyme. The promoters fall into two general classes that are recognized in different ways by different groups of factors. The promoters for 5S and tRNA genes are *internal;* they lie downstream of the startpoint. The promoters for snRNA (small nuclear RNA) genes lie upstream of the startpoint in the more conventional manner of other promoters. In both cases, the individual elements that are necessary for promoter function consist exclusively of sequences recognized by transcription factors, which in turn direct the binding of RNA polymerase.

Before the promoter of 5S RNA genes was identified in *X. laevis*, all attempts to identify promoter sequences assumed that they would lie upstream of the startpoint. But deletion analysis showed that the 5S RNA product continues to be synthesized when the entire sequence upstream of the gene is removed!

When the deletions continue into the gene, a product very similar in size to the usual 5S RNA continues to be synthesized so long as the deletion ends before base +55. **Figure 21.6** shows that the first part of the RNA product corresponds to plasmid DNA; the second part represents the segment remaining of the usual 5S RNA sequence. But when the deletion extends past +55, transcription does not occur. So the promoter lies *downstream of position +55,* but causes RNA polymerase III to initiate transcription a more or less fixed distance upstream.

When deletions extend into the gene from its distal end, transcription is unaffected so long as the first 80 bp remain intact. Once the deletion cuts into this region, transcription ceases. This places the downstream boundary position of the promoter at about position +80.

So the promoter for 5S RNA transcription lies between positions +55 and +80 within the gene. A fragment containing this region can sponsor initiation of any DNA in which it is placed, from a startpoint ~55 bp farther upstream. (The wild-type startpoint is unique; in deletions that lack it, transcription initiates at the purine base nearest to the position 55 bp upstream of the promoter.)

The structures of three types of promoters for RNA polymerase III are summarized in **Figure 21.7.** There are two types of internal promoter. Each contains a bipartite structure, in which two short sequence elements are separated by a variable sequence. Type 1 consists of a boxA sequence separated from a boxC sequence, and type 2 consists of a boxA sequence separated from a boxB sequence. The distance between boxA and boxB in a type 2 promoter can vary quite extensively, but the boxes usually cannot be brought too close together without abolishing function. Type 3 promoters have three sequence elements all located upstream of the startpoint.

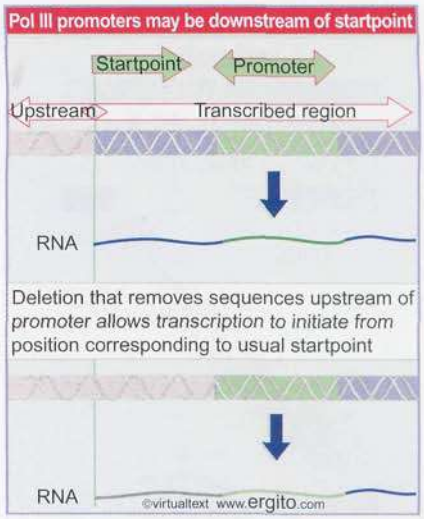


Figure 21.6 Deletion analysis shows that the promoter for 5S RNA genes is internal; initiation occurs a fixed distance (~55 bp) upstream of the promoter.

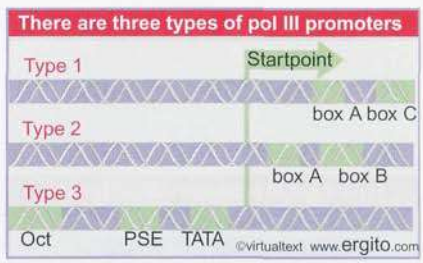


Figure 21.7 Promoters for RNA polymerase III may consist of bipartite sequences downstream of the startpoint, with boxA separated from either boxC or boxB. Or they may consist of separated sequences upstream of the startpoint (Oct, PSE, TATA).

21.6 $TF_{III}B$ is the commitment factor for pol III promoters

Key Concepts

- $TF_{III}A$ and $TF_{III}C$ bind to the consensus sequences and enable $TF_{III}B$ to bind at the startpoint.
- $TF_{III}B$ has TBP as one subunit and enables RNA polymerase to bind.

The detailed interactions are different at the two types of internal promoter, but the principle is the same. $TF_{III}C$ binds downstream of the startpoint, either independently (type 2 promoters) or in

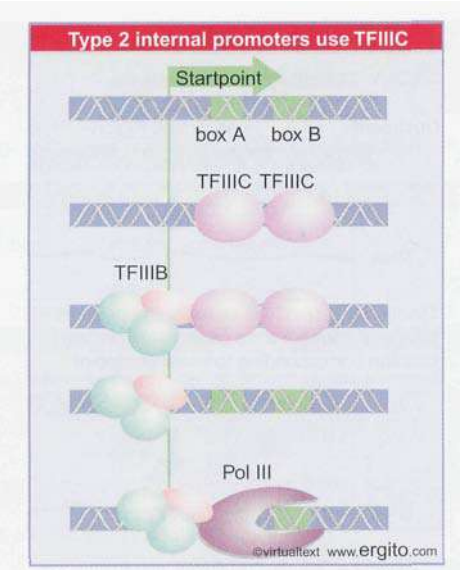


Figure 21.8 Internal type 2 pol III promoters use binding of TFIIIC to boxA and boxB sequences to recruit the positioning factor TFIIIB, which recruits RNA polymerase III.

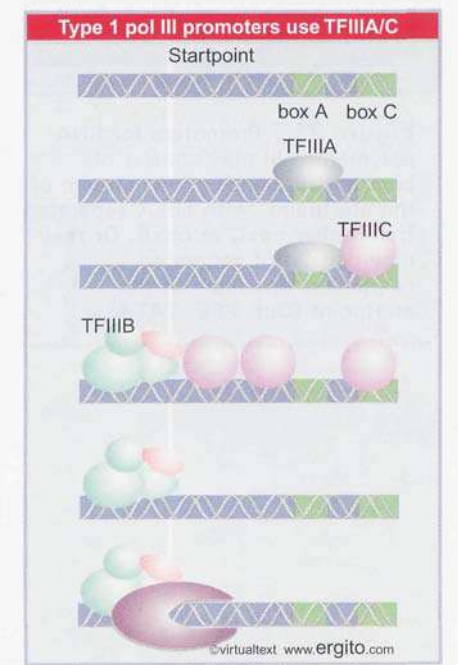


Figure 21.9 Internal type 1 pol III promoters use the assembly factors TFIIIA and TFIIIC, at boxA and boxC, to recruit the positioning factor TFIIIB, which recruits RNA polymerase III.

conjunction with $TF_{III}A$ (type 1 promoters). The presence of $TF_{III}C$ enables the positioning factor $TF_{III}B$ to bind at the startpoint. Then RNA polymerase is recruited.

Figure 21.8 summarizes the stages of reaction at type 2 internal promoters. $TF_{III}C$ binds to both boxA and boxB. This enables $TF_{III}B$ to bind at the startpoint. Then RNA polymerase III can bind.

The difference at type 1 internal promoters is that $TF_{III}A$ must bind at boxA to enable $TF_{III}C$ to bind at boxC. **Figure 21.9** shows that, once $TF_{III}C$ has bound, events follow the same course as at type 2 promoters, with $TF_{III}B$ binding at the startpoint, and RNA polymerase III joining the complex. Type 1 promoters are found only in the genes for 5S rRNA.

$TF_{III}A$ and $TF_{III}C$ are assembly factors, whose sole role is to assist the binding of $TF_{III}B$ at the right location. Once $TF_{III}B$ has bound, $TF_{III}A$ and $TF_{III}C$ can be removed from the promoter (by high salt concentration *in vitro)* without affecting the initiation reaction. *$TF_{III}B$ remains bound in the vicinity of the startpoint and its presence is sufficient to allow RNA polymerase III to bind at the startpoint.* So $TF_{III}B$ is the only true initiation factor required by RNA polymerase **III.** This sequence of events explains how the promoter boxes downstream can cause RNA polymerase to bind at the startpoint, farther upstream. Although the ability to transcribe these genes is conferred by the internal promoter, changes in the region immediately upstream of the startpoint can alter the efficiency of transcription.

$TF_{III}C$ is a large protein complex (>500 kD), comparable in size to RNA polymerase itself, and containing 6 subunits. $TF_{III}A$ is a member of an interesting class of proteins containing a nucleic acid-binding motif called a zinc finger (see *22.9 A zinc finger motif is a DNA-binding domain).* The positioning factor, $TF_{III}IB$, consists of three subunits. It includes the same protein, TBP, that is present in the core-binding factor for pol I promoters, and also in the corresponding transcription factor ($TF_{II}D$) for RNA polymerase II. It also contains Brf, which is related to the factor $TF_{II}B$ that is used by RNA polymerase II. The third subunit is called B"; it is dispensable if the DNA duplex is partially melted, which suggests that its function is to initiate the transcription bubble. The role of B " may be comparable to the role played by sigma factor in bacterial RNA polymerase (see *9.16 Substitution of sigma factors may control initiation).*

The upstream region has a conventional role in the third class of polymerase III promoters. In the example shown in Figure 21.7, there are three upstream elements. These elements are also found in promoters for snRNA genes that are transcribed by RNA polymerase II. (Genes for some snRNAs are transcribed by RNA polymerase II, while others are transcribed by RNA polymerase III.) The upstream elements function in a similar manner in promoters for both polymerases II and **III.**

Initiation at an upstream promoter for RNA polymerase III can occur on a short region that immediately precedes the startpoint and contains only the TATA element. However, efficiency of transcription is much increased by the presence of the PSE and OCT elements. The factors that bind at these elements interact cooperatively. (The PSE element may be essential at promoters used by RNA polymerase II, whereas it is stimulatory in promoters used by RNA polymerase III; its name stands for proximal sequence element.)

The TATA element confers specificity for ftie type of polymerase (II or III) that is recognized by an snRNA promoter. It is bound by a factor that includes the *TBP,* which actually recognizes the sequence in DIVA. The TBP is associated with other proteins, which are specific for the type of promoter. The function of TBP and its associated proteins is to position the RNA polymerase correctly at the startpoint. We discuss this

in more detail for RNA polymerase II (see *21.8 TBP is a universal factor).*

The factors work in the same way for both types of promoters for RNA polymerase III. *The factors bind at the promoter before RNA polymerase itself can bind.* They form a preinitiation complex that directs binding of the RNA polymerase. RNA polymerase III does not itself recognizes the promoter sequence, but binds adjacent to factors that are themselves bound just upstream of the startpoint. For the type 1 and type 2 internal promoters, the assembly factors ensure that $TF_{III}B$ (which includes TBP) is bound just upstream of the startpoint, to provide the positioning information. For the upstream promoters, $TF_{III}B$ binds directly to the region including the TATA box. So irrespective of the location of the promoter sequences, factor(s) are bound close to the startpoint in order to direct binding of RNA polymerase III.

21.7 The startpoint for RNA polymerase II

Key Concepts

- RNA polymerase II requires general transcription factors (called TF||X) to initiate transcription.
- RNA polymerase II promoters have a short conserved sequence Py_2CAPy_5 (the initiator InR) at the startpoint.
- The TATA box is a common component of RNA polymerase II promoters and consists of an A·T-rich octamer located ~25 bp upstream of the startpoint.
- The DPE is a common component of RNA polymerase II promoters that do not contain a TATA box.
- A core promoter for RNA polymerase II includes the InR and either a TATA box or a DPE.

The basic organization of the apparatus for transcribing protein-coding genes was revealed by the discovery that purified RNA polymerase II can catalyze synthesis of mRNA, but cannot initiate transcription unless an additional extract is added. The purification of this extract led to the definition of the general transcription factors—a group of proteins that are needed for initiation by RNA polymerase II at all promoters. RNA polymerase II in conjunction with these factors constitutes the basal transcription apparatus that is needed to transcribe any promoter. The general factors are described as $TF_{II}X$, where "X" is a letter that identifies the individual factor. The subunits of RNA polymerase II and the general transcription factors are conserved among eukaryotes.

Our starting point for considering promoter organization is to define the core promoter as the shortest sequence at which RNA polymerase II can initiate transcription. A core promoter can in principle be expressed in any cell. It comprises the minimum sequence that enables the general transcription factors to assemble at the startpoint. They are involved in the mechanics of binding to DNA and enable RNA polymerase II to initiate transcription. A core promoter functions at only a low efficiency. Other proteins, called activators, are required for a proper level of function (see *21.13 Short sequence elements bind activators).* The activators are not described systematically, but have casual names reflecting their histories of identification.

We may expect any sequence components involved in the binding of RNA polymerase and general transcription factors to be conserved at most or all promoters. As with bacterial promoters, when promoters

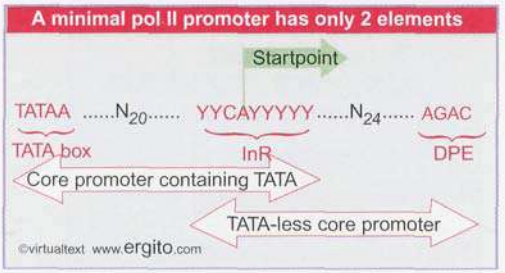


Figure 21.10 The minimal pol II promoter has a TATA box ~25 bp upstream of the InR. The TATA box has the consensus sequence of TATAA. The Inr has pyrimidines (Y) surrounding the CA at the startpoint. The sequence shows the coding strand.

for RNA polymerase II are compared, homologies in the regions near the startpoint are restricted to rather short sequences. These elements correspond with the sequences implicated in promoter function by mutation. **Figure 21.10** shows the construction of a typical pol II core promoter.

At the startpoint, there is no extensive homology of sequence, but there is a tendency for the first base of mRNA to be A, flanked on either side by pyrimidines. (This description is also valid for the CAT start sequence of bacterial promoters.) This region is called the **initiator** (Inr), and may be described in the general form Py_2CAPy_5. The Inr is contained between positions -3 and +5.

Many promoters have a sequence called the TATA box, usually located ~25 bp upstream of the startpoint. It constitutes the only upstream promoter element that has a relatively fixed location with respect to the startpoint. The core sequence is TATAA, usually followed by three more A·T base pairs. The TATA box tends to be surrounded by G·C-rich sequences, which could be a factor in its function. It is almost identical with the -10 sequence found in bacterial promoters; in fact, it could pass for one except for the difference in its location at -25 instead of -10.

Single base substitutions in the TATA box act as strong down mutations. Some mutations reverse the orientation of an A-T pair, so base composition alone is not sufficient for its function. So the TATA box comprises an element whose behavior is analogous to our concept of the bacterial promoter: a short, well-defined sequence just upstream of the startpoint, which is necessary for transcription.

Promoters that do not contain a TATA element are called TATA-less promoters. Surveys of promoter sequences suggest that 50% or more of promoters may be TATA-less. When a promoter does not contain a TATA box, it usually contains another element, the DPE (downstream promoter element) which is located at +28 - +32.

A core promoter can consist either of a TATA box plus InR or of an InR plus DPE.

21.8 TBP is a universal factor

Key Concepts

- TBP is a component of the positioning factor that is required for each type of RNA polymerase to bind its promoter.
- The factor for RNA polymerase II is $TF_{II}D$, which consists of TBP and 11 TAFs, with a total mass ~800 kD.

The first step in complex formation at a promoter containing a TATA box is binding of the factor $TF_{II}D$ to a region that extends upstream from the TATA sequence. $TF_{II}D$ contains two types of component. Recognition of the TATA box is conferred by the TATA-binding protein (TBP), a small protein of ~30 kD. The other subunits are called TAFs (for TBP-associated factors). Some TAFs are stoichiometric with TBP; others are present in lesser amounts. $TF_{II}Ds$ containing different TAFs could recognize different promoters. Some (substoichiometric) TAFs are tissue-specific. The total mass of $TF_{II}D$ typically is ~800 kD, containing TBP and 11 TAFs, varying in mass from 30-250 kD. The TAFs in $TF_{II}D$ are named in the form $TAF_{II}00$, where "00" gives the molecular mass of the subunit.

Positioning factors that consist of TBP associated with a set of TAFs are responsible for identifying all classes of promoters. $TF_{III}B$ (for pol III promoters) and SL1 (for pol I promoters) may both be viewed as

consisting of TBP associated with a particular group of proteins that substitute for the TAFs that are found in $TF_{II}D$. TBP is the key component, and is incorporated at each type of promoter by a different mechanism. In the case of promoters for RNA polymerase II, the key feature in positioning is the fixed distance of the TATA box from the startpoint.

Figure 21.11 shows that the positioning factor recognizes the promoter in a different way in each case. At promoters for RNA polymerase III, $TF_{III}B$ binds adjacent to $TF_{III}C$. At promoters for RNA polymerase I, SL1 binds in conjunction with UBF. $TF_{II}D$ is solely responsible for recognizing promoters for RNA polymerase II. At a promoter that has a TATA element, TBP binds specifically to DNA, but at other promoters it may be incorporated by association with other proteins that bind to DNA. Whatever its means of entry into the initiation complex, it has the common purpose of interaction with the RNA polymerase.

$TF_{II}D$ is ubiquitous, but not unique. All multicellular eukaryotes also express an alternative complex, which has TLF (TBP like factor) instead of TBP. A TLF is typically ~60% similar to TBP. It probably initiates complex formation by the usual set of TF_{II} factors. However, TLF does not bind to the TATA box, and we do not yet know how it works. *Drosophila* also has a third factor, TRF1, which behaves in the same way as TBP and binds its own set of TAFs, to form a complex that functions as an alternative to $TF_{II}D$ at a specific set of promoters.

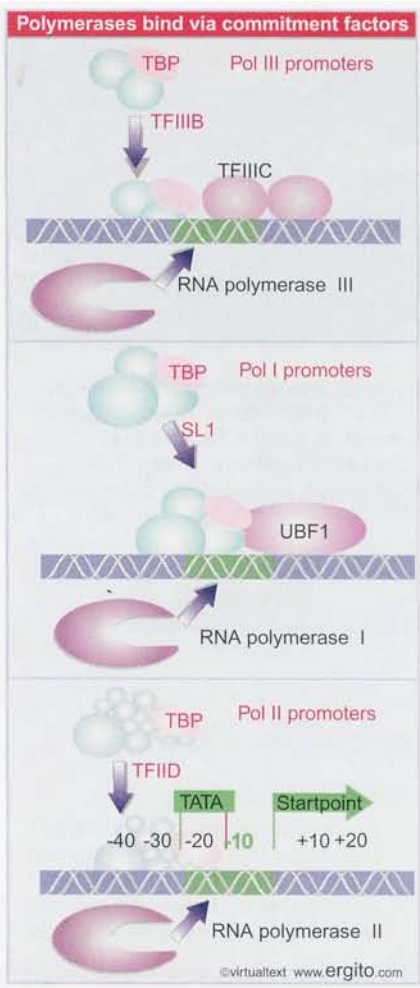


Figure 21.11 RNA polymerases are positioned at all promoters by a factor that contains TBP.

21.9 TBP binds DNA in an unusual way

Key Concepts

- TBP binds to the TATA box in the minor groove of DNA.
- It forms a saddle around the DNA and bends it by ~80°.
- Some of the TAFs resemble histones and may form a structure resembling a histone octamer.

TBP has the unusual property of binding to DNA in the minor groove. (Virtually all known DNA-binding proteins bind in the major groove.) The crystal structure of TBP suggests a detailed model for its binding to DNA. **Figure 21.12** shows that it surrounds one face of DNA, forming a "saddle" around the double helix. In effect, the inner surface of TBP binds to DNA, and the larger outer surface is available to extend contacts to other proteins. The DNA-binding site consists of a C-terminal domain that is conserved between species, while the variable N-terminal tail is exposed to interact with other proteins. It is a measure of the conservation of mechanism in transcriptional initiation that the DNA-binding sequence of TBP is 80% conserved between yeast and Man.

Binding of TBP may be inconsistent with the presence of nucleosomes. Because nucleosomes form preferentially by placing A·T-rich sequences with the minor grooves facing inward, they could prevent binding of TBP. This may explain why the presence of nucleosomes prevents initiation of transcription.

TBP first binds to the minor groove, and then bends the DNA by ~80°, as illustrated in **Figure 21.13.** The TATA box bends towards the major groove, widening the minor groove. The distortion is restricted to the 8 bp of the TATA box; at each end of the sequence, the minor groove has its usual width of ~5 A, but at the center of the sequence the minor groove is >9 A. This is a deformation of the structure, but does not actually separate the strands of DNA, because base pairing is maintained.

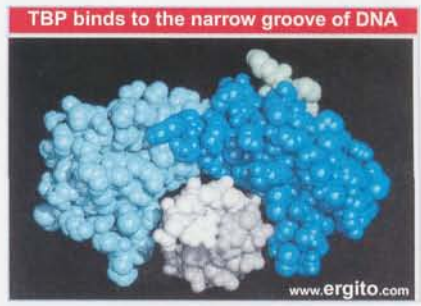


Figure 21.12 A view in cross-section shows that TBP surrounds DNA from the side of the narrow groove. TBP consists of two related (40% identical) conserved domains, which are shown in light and dark blue. The N-terminal region varies extensively and is shown in green. The two strands of the DNA double helix are in light and dark grey. Photograph kindly provided by Stephen Burley.

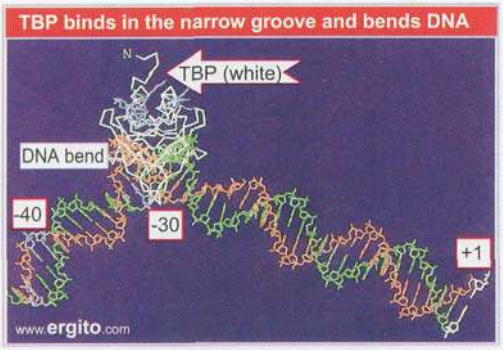


Figure 21.13 The cocrystal structure of TBP with DNA from -40 to the startpoint shows a bend at the TATA box that widens the narrow groove where TBP binds. Photograph provided by Stephen Burley.

This structure has several functional implications. By changing the spatial organization of DNA on either side of the TATA box, it allows the transcription factors and RNA polymerase to form a closer association than would be possible on linear DNA. The bending at the TATA box corresponds to unwinding of about 1/3 of a turn of DNA, and is compensated by a positive writhe. We do not know yet how this relates to the initiation of strand separation.

The presence of TBP in the minor groove, combined with other proteins binding in the major groove, creates a high density of protein-DNA contacts in this region. Binding of purified TBP to DNA *in vitro* protects ~1 turn of the double helix at the TATA box, typically extending from -37 to -25; but binding of the $TF_{II}D$ complex in the initiation reaction regularly protects the region from -45 to -10, and also extends farther upstream beyond the startpoint. TBP is the only general transcription factor that makes sequence-specific contacts with DNA.

Within $TF_{II}D$ as a free protein complex, the factor $TAF_{II}230$ binds to TBP, where it occupies the concave DNA-binding surface. In fact, the structure of the binding site, which lies in the N-terminal domain of $TAF_{II}230$, mimics the surface of the minor groove in DNA. This molecular mimicry allows $TAF_{II}230$ to control the ability of TBP to bind to DNA; the N-terminal domain of $TAF_{II}230$ must be displaced from the DNA-binding surface of TBP in order for $TF_{II}D$ to bind to DNA.

Some TAFs resemble histones; in particular $TAF_{II}42$ and $TAF_{II}62$ appear to be (distant) homologues of histones H3 and H4, and they form a heterodimer using the same motif (the histone fold) that histones use for the interaction. (Histones H3 and H4 form the kernel of the histone octamer—the basic complex that binds DNA in eukaryotic chromatin; see *20.8 Organization of the histone octamer.)* Together with other TAFs, $TAF_{II}42$ and $TAF_{II}62$ may form the basis for a structure resembling a histone octamer; such a structure may be responsible for the nonsequence-specific interactions of $TF_{II}D$ with DNA. Histone folds are also used in pairwise interactions between other $TAF_{II}s$.

Some of the $TAF_{II}s$ may be found in other complexes as well as in $TF_{II}D$. In particular, the histone-like $TAF_{II}s$ are found also in protein complexes that modify the structure of chromatin prior to transcription (see *23.7 Acetylases are associated with activators).*

21.10 The basal apparatus assembles at the promoter

Key Concepts

- Binding of $TF_{II}D$ to the TATA box is the first step in initiation.
- Other transcription factors bind to the complex in a defined order, extending the length of the protected region on DNA.
- When RNA polymerase II binds to the complex, it initiates transcription.

Initiation requires the transcription factors to act in a defined order to build a complex that is joined by RNA polymerase. The series of events can be followed by the increasing size of the protein complex associated with DNA. Footprinting of the DNA regions protected by each complex suggests the model summarized in

Figure 21.14. As each TF1I factor joins the complex, an increasing length of DNA is covered. RNA polymerase is incorporated at a late stage.

Commitment to a promoter is initiated when $TF_{II}D$ binds the TATA box. ($TF_{II}D$ also recognizes the InR sequence at the startpoint.) When $TF_{II}A$ joins the complex, $TF_{II}D$ becomes able to protect a region extending farther upstream. $TF_{II}A$ may activate TBP by relieving the repression that is caused by the $TAF_{II}230$.

Addition of $TF_{II}B$ gives partial protection of the region of the template strand in the vicinity of the startpoint, from —40 to +10. This suggests that $TF_{II}B$ is bound downstream of the TATA box, perhaps loosely associated with DNA and asymmetrically oriented with regard to the two DNA strands. The crystal structure shown in **Figure 21.15** extends this model. $TF_{II}B$ binds adjacent to TBP, extending contacts along one face of DNA. It makes contacts in the minor groove downstream of the TATA box, and contacts the major groove upstream of the TATA box, in a region called the BRE. In archaea, the homologue of $TF_{II}B$ actually makes sequence-specific contacts with the promoter in the BRE region. $TF_{II}B$ may provide the surface that is in turn recognized by RNA polymerase, so that it is responsible for the directionality of the binding of the enzyme.

The factor $TF_{II}F$ is a heterotetramer consisting of two types of subunit. The larger subunit (RAP74) has an ATP-dependent DNA helicase activity that could be involved in melting the DNA at initiation. The smaller subunit (RAP38) has some homology to the regions of bacterial sigma factor that contact the core polymerase; it binds tightly to RNA polymerase II. $TF_{II}F$ may bring RNA polymerase II to the assembling transcription complex and provide the means by which it binds. The complex of TBP and TAFs may interact with the CTD tail of RNA polymerase, and interaction with $TF_{II}B$ may also be important when $TF_{II}F$/polymerase joins the complex.

Polymerase binding extends the sites that are protected downstream to +15 on the template strand and +20 on the nontemplate strand. The enzyme extends the full length of the complex, since additional protection is seen at the upstream boundary.

What happens at TATA-less promoters? The same general transcription factors, including $TF_{II}D$, are needed. The Inr provides the positioning element; $TF_{II}D$ binds to it via an ability of one or more of the TAFs to recognize the Inr directly. Other TAFs in $TF_{II}D$ also recognize the DPE element downstream from the startpoint. The function of TBP at these promoters is more like that at promoters for RNA polymerase I and at internal promoters for RNA polymerase III.

Assembly of the RNA polymerase II initiation complex provides an interesting contrast with prokaryotic transcription. Bacterial RNA polymerase is essentially a coherent aggregate with intrinsic ability to bind DNA; the sigma factor, needed for initiation but not for elongation, becomes part of the enzyme before DNA is bound, although it is later released. But RNA polymerase II can bind to the promoter only after separate transcription factors have bound. The factors play a role analogous to that of bacterial sigma factor—to allow the basic polymerase to recognize DNA specifically at promoter sequences—but have evolved more independence. Indeed, the factors are primarily responsible for the specificity of promoter recognition. Only some of the factors participate in protein-DNA contacts (and only TBP makes sequence-specific contacts); thus protein-protein interactions are important in the assembly of the complex.

When a TATA box is present, it determines the location of the startpoint. Its deletion causes the site of initiation to become erratic, although any overall reduction in transcription is relatively small. Indeed, some TATA-less promoters lack unique startpoints; initiation

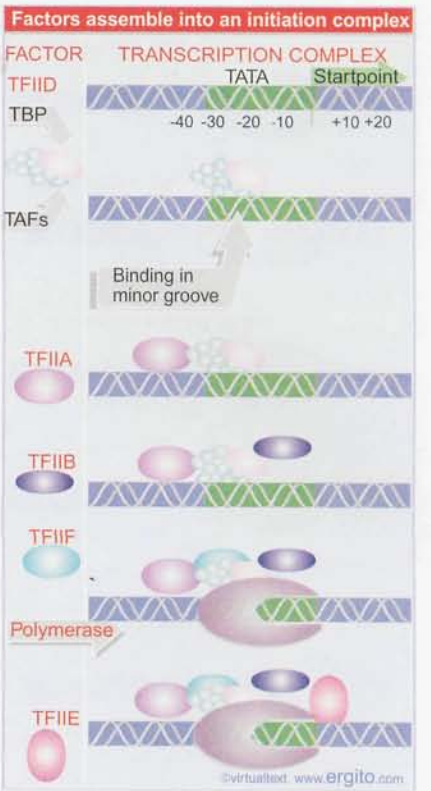


Figure 21.14 An initiation complex assembles at promoters for RNA polymerase II by an ordered sequence of association with transcription factors.

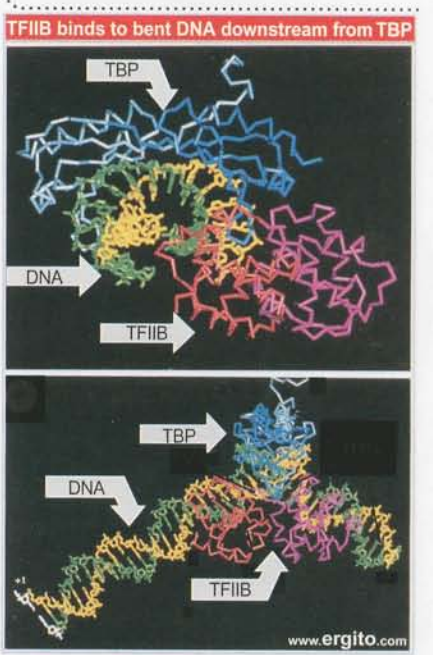


Figure 21.15 Two views of the ternary complex of TFIIB-TBP-DNA show that $TF_{II}B$ binds along the bent face of DNA. The two strands of DNA are green and yellow, TBP is blue, and TFIIB is red and purple. Photograph kindly provided by Stephen Burley.

occurs instead at any one of a cluster of startpoints. The TATA box aligns the RNA polymerase (via the interaction with $TF_{II}D$ and other factors) so that it initiates at the proper site. This explains why its location is fixed with respect to the startpoint. Binding of TBP to TATA is the predominant feature in recognition of the promoter, but two large TAFs ($TAF_{II}250$ and $TAF_{II}150$) also contact DNA in the vicinity of the startpoint and influence the efficiency of the reaction.

Although assembly can take place just at the core promoter *in vitro*, this reaction is not sufficient for transcription *in vivo*, where interactions with activators that recognize the more upstream elements are required. The activators interact with the basal apparatus at various stages during its assembly (see *22.5 Activators interact with the basal apparatus).*

21.11 Initiation is followed by promoter clearance

Key Concepts

- $TF_{II}E$ and $TF_{II}H$ are required to melt DNA to allow polymerase movement.
- Phosphorylation of the CTD may be required for elongation to begin.
- Further phosphorylation of the CTD is required at some promoters to end abortive initiation.
- The CTD may coordinate processing of RNA with transcription.

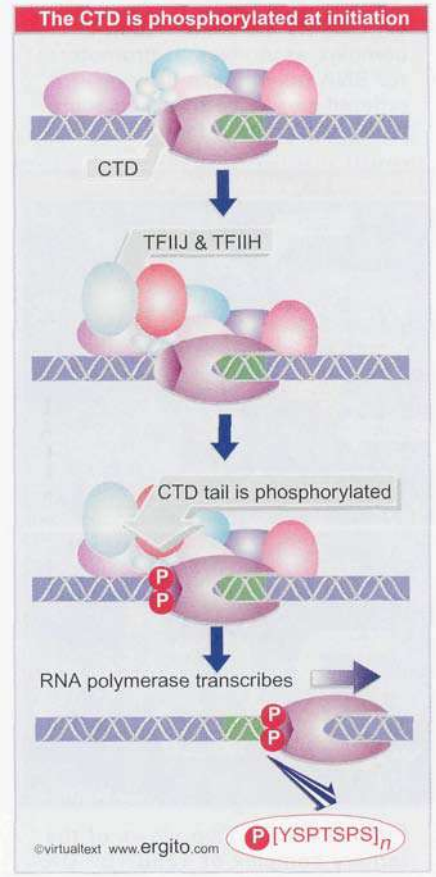


Figure 21.16 Phosphorylation of the CTD by the kinase activity of $TF_{II}H$ may be needed to release RNA polymerase to start transcription.

Most of the transcription factors are required solely to bind RNA polymerase to the promoter, but some act at a later stage. Binding of $TF_{II}E$ causes the boundary of the region protected downstream to be extended by another turn of the double helix, to +30. Two further factors, $TF_{II}H$ and $TF_{II}J$, join the complex after $TF_{II}E$. They do not change the pattern of binding to DNA.

$TF_{II}H$ is the only general transcription factor that has independent enzymatic activities. Its several activities include an ATPase, helicases of both polarities, and a kinase activity that can phosphorylate the CTD tail of RNA polymerase II. $TF_{II}H$ is an exceptional factor that may play a role also in elongation. Its interaction with DNA downstream of the startpoint is required for RNA polymerase to escape from the promoter. $TF_{II}H$ is also involved in repair of damage to DNA (see next section).

The initiation reaction, as defined by formation of the first phosphodiester bond, occurs once RNA polymerase has bound. **Figure 21.16** proposes a model in which phosphorylation of the tail is needed to release RNA polymerase II from the transcription factors so that it can make the transition to the elongating form. Most of the transcription factors are released from the promoter at this stage.

On a linear template, ATP hydrolysis, $TF_{II}E$, and the helicase activity of $TF_{II}H$ (provided by the XPB subunit) are required for polymerase movement. This requirement is bypassed with a supercoiled template. This suggests that $TF_{II}E$ and $TF_{II}H$ are required to melt DNA to allow polymerase movement to begin. The helicase activity of the XPB subunit of $TF_{II}H$ is responsible for the actual melting of DNA.

RNA polymerase II stutters at some genes when it starts transcription. (The result is not dissimilar to the abortive initiation of bacterial RNA polymerase discussed in *9.11 Sigma factor controls binding to DNA,*

although the mechanism is different.) At many genes, RNA polymerase II terminates after a short distance. The short RNA product is degraded rapidly. To extend elongation into the gene, a kinase called P-TEFb is required. This kinase is a member of the cdk family that controls the cell cycle (see *29 Cell cycle and growth regulation)*. P-TEFb acts on the CTD, to phosphorylate it further. We do not yet understand why this effect is required at some promoters but not others or how it is regulated.

The CTD may also be involved, directly or indirectly, in processing RNA after it has been synthesized by RNA polymerase **II. Figure 21.17** summarizes processing reactions in which the CTD may be involved. The capping enzyme (guanylyl transferase), which adds the G residue to the 5' end of newly synthesized mRNA, binds to the phosphorylated CTD: this may be important in enabling it to modify the 5′ end as soon as it is synthesized. A set of proteins called SCAFs bind to the CTD, and they may in turn bind to splicing factors. This may be a means of coordinating transcription and splicing. Some components of the cleavage/polyadenylation apparatus also bind to the CTD. Oddly enough, they do so at the time of initiation, so that RNA polymerase is all ready for the 3′ end processing reactions as soon as it sets out! All of this suggests that the CTD may be a general focus for connecting other processes with transcription. In the cases of capping and splicing, the CTD functions indirectly to promote formation of the protein complexes that undertake the reactions. In the case of 3′ end generation, it may participate directly in the reaction.

The general process of initiation is similar to that catalyzed by bacterial RNA polymerase. Binding of RNA polymerase generates a closed complex, which is converted at a later stage to an open complex in which the DNA strands have been separated. In the bacterial reaction, formation of the open complex completes the necessary structural change to DNA; a difference in the eukaryotic reaction is that further unwinding of the template is needed after this stage.

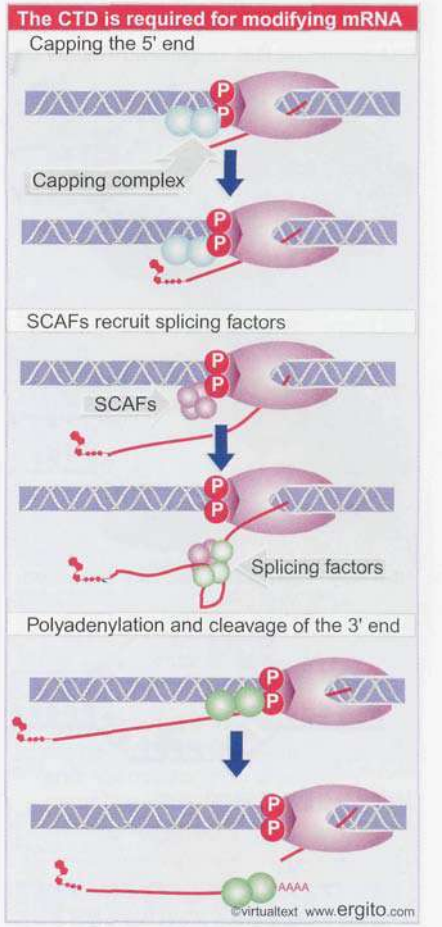


Figure 21.17 The CTD is important in recruiting enzymes that modify RNA.

21.12 A connection between transcription and repair

Key Concepts

- Transcribed genes are preferentially repaired when DNA damage occurs.
- $TF_{II}H$ provides the link to a complex of repair enzymes.
- Mutations in the XPD component of $TF_{II}H$ cause three types of human diseases.

In both bacteria and eukaryotes, there is a direct link from RNA polymerase to the activation of repair. The basic phenomenon was first observed because transcribed genes are preferentially repaired. Then it was discovered that it is only the template strand of DNA that is the target—the nontemplate strand is repaired at the same rate as bulk DNA.

In bacteria, the repair activity is provided by the *uvr* excision-repair system (see *15.21 Excision repair systems in E. coli*). Preferential repair is abolished by mutations in the gene *mfd*, whose product provides the link from RNA polymerase to the Uvr enzymes.

Figure 21.18 shows a model for the link between transcription and repair. When RNA polymerase encounters DNA damage in the template strand, it stalls because it cannot use the damaged sequences as a template to direct complementary base pairing. This explains the

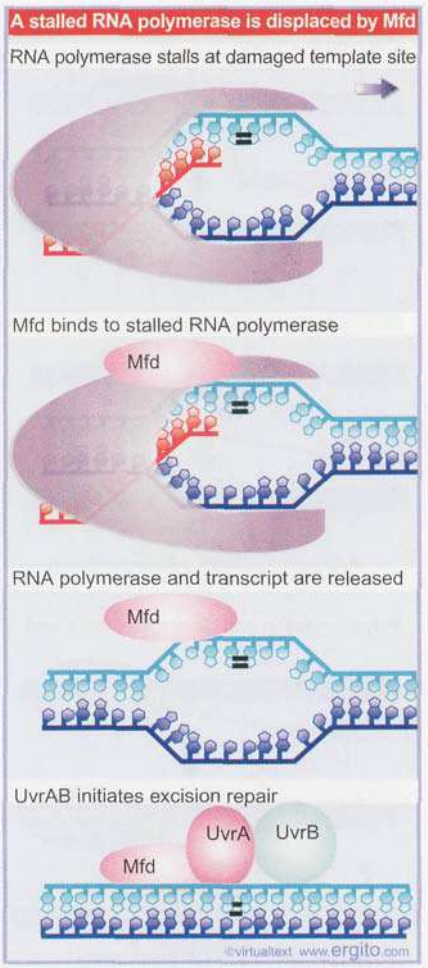


Figure 21.18 Mfd recognizes a stalled RNA polymerase and directs DNA repair to the damaged template strand.

specificity of the effect for the template strand (damage in the nontemplate strand does not impede progress of the RNA polymerase).

The Mfd protein has two roles. First, it displaces the ternary complex of RNA polymerase from DNA. Second, it causes the UvrABC enzyme to bind to the damaged DNA. This leads to repair of DNA by the excision-repair mechanism (see Figure 15.40). After the DNA has been repaired, the next RNA polymerase to traverse the gene is able to produce a normal transcript.

A similar mechanism, although relying on different components, is used in eukaryotes. The template strand of a transcribed gene is preferentially repaired following UV-induced damage. The general transcription factor $TF_{II}H$ is involved. $TF_{II}H$ is found in alternative forms, which consist of a core associated with other subunits.

$TF_{II}H$ has a common function in both initiating transcription and repairing damage. The same helicase subunit (XPD) creates the initial transcription bubble and melts DNA at a damaged site. Its other functions differ between transcription and repair, as provided by the appropriate form of the complex.

Figure 21.19 shows that the basic factor involved in transcription consists of a core (of 5 subunits) associated with other subunits that have a kinase activity. The alternative complex consists of the core associated with a large group of proteins that are coded by repair genes. (The basic model for repair is shown in Figure 15.53.) The repair proteins include a subunit (XPC) that recognizes damaged DNA, which provides the coupling function that enables a template strand to be preferentially repaired when RNA polymerase becomes stalled at damaged DNA. Other proteins associated with the complex include endonucleases (XPG, XPF, ERCC1). Homologous proteins are found in the complexes in yeast (where they are often identified by *rad* mutations that are defective in repair) and in Man (where they are identified by mutations that cause diseases resulting from deficiencies in repairing damaged DNA). (Subunits with the name XP are coded by genes in which mutations cause the disease xeroderma pigmentosum (see *15.28 Eukaryotic cells have conserved repair systems).*

The kinase complex and the repair complex can associate and dissociate reversibly from the core $TF_{II}H$. This suggests a model in which the first form of $TF_{II}H$ is required for initiation, but may be replaced by the other form (perhaps in response to encountering DNA damage). $TF_{II}H$ dissociates from RNA polymerase at an early stage of elongation (after transcription of ~50 bp); its reassociation at a site of damaged DNA may require additional coupling components.

The repair function may require modification or degradation of RNA polymerase. The large subunit of RNA polymerase is degraded when the enzyme stalls at sites of UV damage. We do not yet understand the connection between the transcription/repair apparatus as such and the degradation of RNA polymerase. It is possible that removal of the polymerase is necessary once it has become stalled.

This degradation of RNA polymerase is deficient in cells from patients with Cockayne's syndrome (a repair disorder). Cockayne's syndrome is caused by mutations in either of two genes *(CSA* and *CSB),* both of whose products appear to be part of or bound to $TF_{II}H$. Cockayne's syndrome is also occasionally caused by mutations in XPD.

XPD is a pleiotropic protein, in which different mutations can affect different functions. In fact, XPD is required for the stability of the $TF_{II}H$ complex during transcription, but the helicase activity as such is not needed. Mutations that prevent XPD from stabilizing the complex cause trichothiodystrophy. The helicase activity is required for the repair function. Mutations that affect the helicase activity cause the repair deficiency that results in XP or Cockayne's syndrome.

21.13 Short sequence elements bind activators

Key Concepts

- Short conserved sequence elements are dispersed in the region preceding the startpoint.
- The upstream elements increase the frequency of initiation.
- The factors that bind to them to stimulate transcription are called activators.

A promoter for RNA polymerase II consists of two types of region. The startpoint itself is identified by the Inr and/or by the TATA box close by. In conjunction with the general transcription factors, RNA polymerase II forms an initiation complex surrounding the startpoint, as we have just described. The efficiency and specificity with which a promoter is recognized, however, depend upon short sequences, farther upstream, which are recognized by a different group of factors, usually called activators. Usually the target sequences are ~100 bp upstream of the startpoint, but sometimes they are more distant. Binding of activators at these sites may influence the formation of the initiation complex at (probably) any one of several stages.

An analysis of a typical promoter is summarized in **Figure 21.20.** Individual base substitutions were introduced at almost every position in the 100 bp upstream of the β-globin startpoint. The striking result is that *most mutations do not affect the ability of the promoter to initiate transcription.* Down mutations occur in three locations, corresponding to three short discrete elements. The two upstream elements have a greater effect on the level of transcription than the element closest to the startpoint. Up mutations occur in only one of the elements. We conclude that the three short sequences centered at -30, -75, and -90 constitute the promoter. Each of them corresponds to the consensus sequence for a common type of promoter element.

The TATA box (centered at —30) is the least effective component of the promoter as measured by the reduction in transcription that is caused by mutations. But although initiation is not prevented when a TATA box is mutated, the startpoint varies from its usual precise location. This confirms the role of the TATA box as a crucial positioning component of the core promoter.

The basal elements and the elements upstream of them have different types of functions. The basal elements (the TATA box and Inr) primarily determine the location of the startpoint, but can sponsor initiation only at

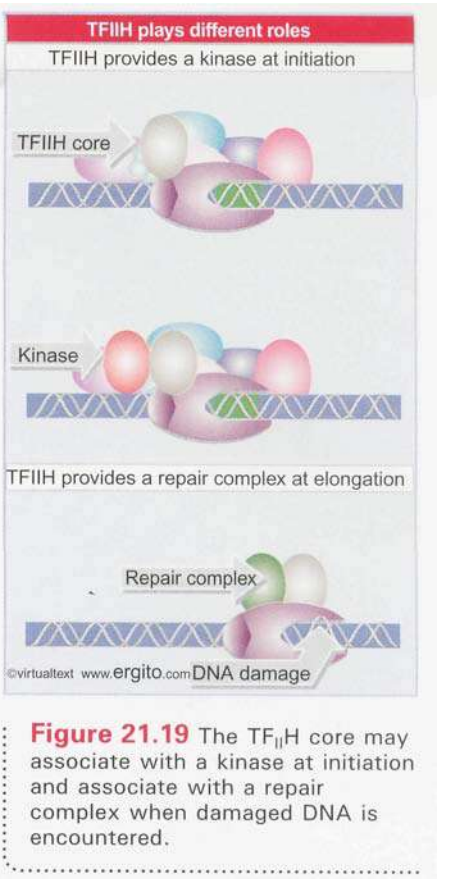


Figure 21.19 The $TF_{II}H$ core may associate with a kinase at initiation and associate with a repair complex when damaged DNA is encountered.

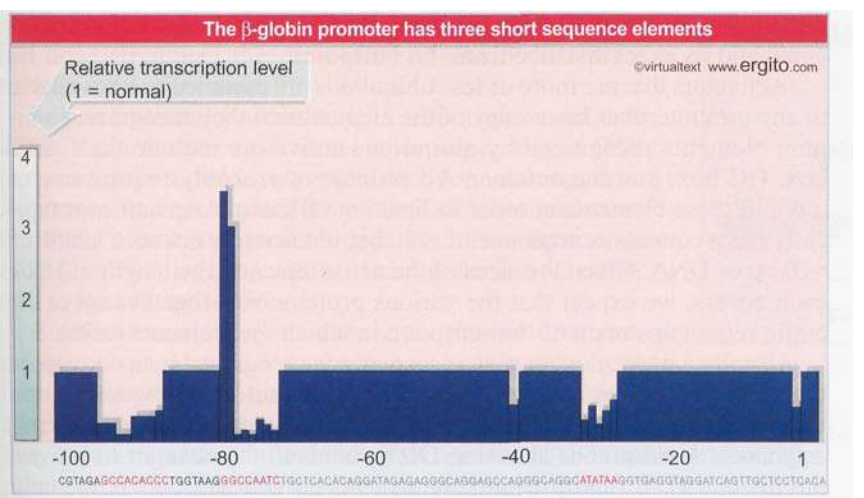


Figure 21.20 Saturation mutagenesis of the upstream region of the β-globin promoter identifies three short regions (centered at -30, -75, and -90) that are needed to initiate transcription. These correspond to the TATA, CAAT, and GC boxes.

By Book_Crazy [IND]

a rather low level. They identify the *location* at which the general transcription factors assemble to form the basal complex. The sequence elements farther upstream influence the *frequency* of initiation, most likely by acting directly on the general transcription factors to enhance the efficiency of assembly into an initiation complex (see *22.5 Activators interact with the basal apparatus).*

The sequence at -75 is the CAAT box. Named for its consensus sequence, it was one of the first common elements to be described. It is often located close to -80, but it can function at distances that vary considerably from the startpoint. It functions in either orientation. Susceptibility to mutations suggests that the CAAT box plays a strong role in determining the efficiency of the promoter, but does not influence its specificity.

The GC box at -90 contains the sequence GGGCGG. Often multiple copies are present in the promoter, and they occur in either orientation. It too is a relatively common promoter component.

21.14 Promoter construction is flexible but context can be important

Key Concepts

- No individual upstream element is essential for promoter function, although one or more elements must be present for efficient initiation.
- Some elements are recognized by multiple factors, and the factor that is used at any particular promoter may be determined by the context of the other factors that are bound.

Promoters are organized on a principle of "mix and match." A variety of elements can contribute to promoter function, but none is essential for all promoters. Some examples are summarized in Figure 21.21. Four types of elements are found altogether in these promoters: TATA, GC boxes, CAAT boxes, and the octamer (an 8 bp element). The elements found in any individual promoter differ in number, location, and orientation. No element is common to all of the promoters. Although the promoter conveys directional information (transcription proceeds only in the downstream direction), the GC and CAAT boxes seem to be able to function in either orientation. This implies that the elements function solely as DNA-binding sites to bring transcription factors into the vicinity of the startpoint; the structure of a factor must be flexible enough to allow it to make protein-protein contacts with the basal apparatus irrespective of the way in which its DNA-binding domain is oriented and its exact distance from the startpoint.

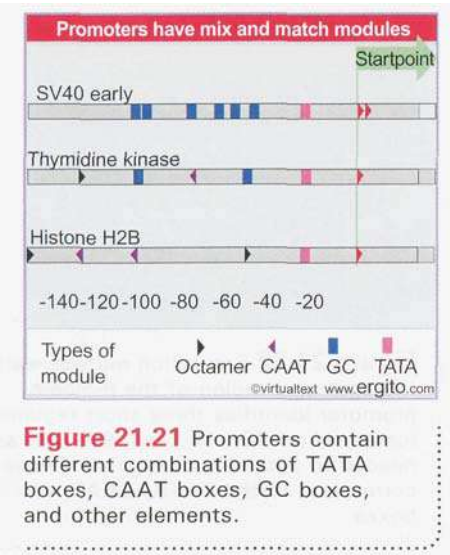


Figure 21.21 Promoters contain different combinations of TATA boxes, CAAT boxes, GC boxes, and other elements.

Activators that are more or less ubiquitous are assumed to be available to any promoter that has a copy of the element that they recognize. Common elements recognized by ubiquitous activators include the CAAT box, GC box, and the octamer. All promoters probably require one or more of these elements in order to function efficiently. An activator typically has a consensus sequence of < 10 bp, but actually covers a length of ~20 bp of DNA. Given the sizes of the activators, and the length of DNA each covers, we expect that the various proteins will together cover the entire region upstream of the startpoint in which the elements reside.

Usually a particular consensus sequence is recognized by a corresponding activator (or by a member of a family of factors). However, sometimes a particular promoter sequence can be recognized by one of several activators. A ubiquitous activator, Oct-1, binds to the octamer to activate the histone H2B (and presumably also other) genes. Oct-1 is the only

octamer-binding factor in nonlymphoid cells. But in lymphoid cells, a different activator, Oct-2, binds to the octamer to activate the immunoglobulin K light gene. So Oct-2 is a tissue-specific activator, while Oct-1 is ubiquitous. The exact details of recognition are not so important to know as the fact that a variety of activators recognize CAAT boxes.

The use of the same octamer in the ubiquitously expressed H2B gene and the lymphoid-specific immunoglobulin genes poses a paradox. Why does the ubiquitous Oct-1 fail to activate the immunoglobulin genes in nonlymphoid tissues? The *context* must be important: Oct-2 rather than Oct-1 may be needed to interact with other proteins that bind at the promoter. These results mean that we cannot predict whether a gene will be activated by a particular activator simply on the basis of the presence of particular elements in its promoter.

21.15 Enhancers contain bidirectional elements that assist initiation

Key Concepts

- An enhancer activates the nearest promoter to it, and can be any distance either upstream or downstream of the promoter.
- A UAS (upstream activator sequence) in yeast behaves like an enhancer but works only upstream of the promoter.
- Similar sequence elements are found in enhancers and promoters.
- Enhancers form complexes of activators that interact directly or indirectly with the promoter.

We have considered the promoter so far as an isolated region responsible for binding RNA polymerase. But eukaryotic promoters do not necessarily function alone. In at least some cases, the activity of a promoter is enormously increased by the presence of an enhancer, which consists of another group of elements, but located at a variable distance from those regarded as comprising part of the promoter itself.

The concept that the enhancer is distinct from the promoter reflects two characteristics. The position of the enhancer relative to the promoter need not be fixed, but can vary substantially. **Figure 21.22** shows that it can be either upstream or downstream. And it can function in either orientation (that is, it can be inverted) relative to the promoter. Manipulations of DNA show that an enhancer can stimulate any promoter placed in its vicinity. In natural genomes, enhancers can be located within genes (that is, just downstream of the promoter) or tens of kilobases away in either direction.

For operational purposes, it is sometimes useful to define the promoter *as a sequence or sequences of DNA that must be in a (relatively) fixed location with regard to the startpoint.* By this definition, the TATA box and other upstream elements are included, but the enhancer is excluded. This is, however, a working definition rather than a rigid classification.

Elements analogous to enhancers, called upstream activator sequences (UAS), are found in yeast. They can function in either orientation, at variable distances upstream of the promoter, but cannot function when located downstream. They have a regulatory role: in several cases the UAS is bound by the regulatory protein(s) that activates the genes downstream.

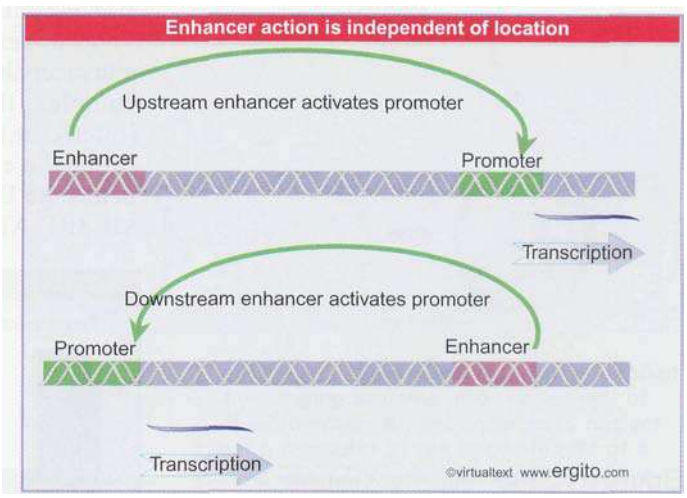


Figure 21.22 An enhancer can activate a promoter from upstream or downstream locations, and its sequence can be inverted relative to the promoter.

Reconstruction experiments in which the enhancer sequence is removed from the DNA and then is inserted elsewhere show that normal transcription can be sustained so long as it is present *anywhere* on the DNA molecule. If a β-globin gene is placed on a DNA molecule that contains an enhancer, its transcription is increased *in vivo* more than 200-fold, even when the enhancer is several kb upstream or downstream of the startpoint, in either orientation. We have yet to discover at what distance the enhancer fails to work.

21.16 Enhancers contain the same elements that are found at promoters

Key Concepts

- Enhancers are made of the same short sequence elements that are found in promoters.
- The density of sequence components is greater in the enhancer than in the promoter.

A difference between the enhancer and a typical promoter is presented by the density of regulatory elements. **Figure 21.23** summarizes the susceptibility of the SV40 enhancer to damage by mutation; and we see that a much greater proportion of its sites directly influences its function than is the case with the promoter analyzed in the same way in Figure 21.20. There is a corresponding increase in the density of protein-binding sites. Many of these sites are common elements in promoters; for example, AP1 and the octamer.

The specificity of transcription may be controlled by either a promoter or an enhancer. A promoter may be specifically regulated, and a nearby enhancer used to increase the efficiency of initiation; or a promoter may lack specific regulation, but become active only when a nearby enhancer is specifically activated. An example is provided by immunoglobulin genes, which carry enhancers *within* the transcription unit. The immunoglobulin enhancers appear to be active only in the B lymphocytes in which the immunoglobulin genes are expressed. Such enhancers provide part of the regulatory network by which gene expression is controlled.

A difference between enhancers and promoters may be that an enhancer shows greater cooperativity between the binding of factors. A complex that assembles at the enhancer that responds to IFN (interferon) *y* assembles cooperatively to form a functional structure called the enhanceosome. Binding of the nonhistone protein HMGI(Y) bends the DNA into a structure that then binds several activators (NF-KB, IRF, ATF-Jun). In contrast with the "mix and match" construction

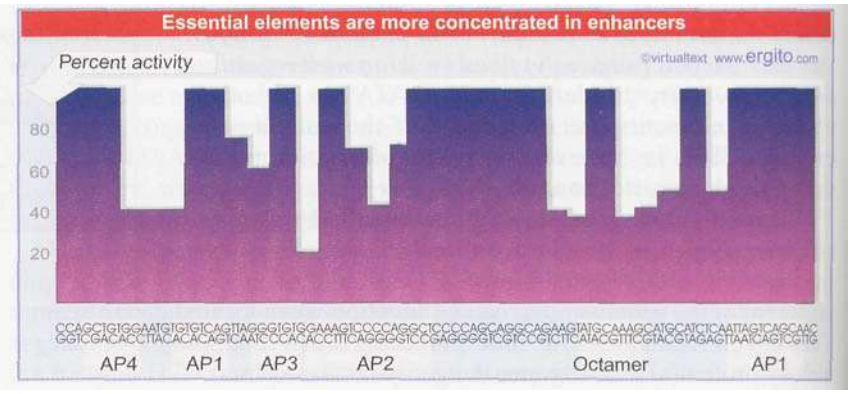


Figure 21.23 An enhancer contains several structural motifs. The histogram plots the effect of all mutations that reduce enhancer function to <75% of wild type. Binding sites for proteins are indicated below the histogram.

By Book_Crazy [IND]

of promoters, all of these components are required to create an active structure at the enhancer. These components do not themselves directly bind to RNA polymerase, but they create a surface that binds a *coactivating complex*. The complex helps the pre-initiation complex of basal transcription factors that is assembling at the promoter to recruit RNA polymerase. We discuss the function of coactivators in more detail in *22.5 Activators interact with the basal apparatus.*

21.17 Enhancers work by increasing the concentration of activators near the promoter

Key Concepts

- Enhancers usually work only in *cis* configuration with a target promoter.
- Enhancers can be made to work in *trans* configuration by linking the DNA that contains the target promoter to the DNA that contains the enhancer via a protein bridge or by catenating the two molecules.
- The principle is that an enhancer works in any situation in which it is constrained to be in the same proximity as the promoter.

How can an enhancer stimulate initiation at a promoter that can be located any distance away on either side of it? When enhancers were first discovered, several possibilities were considered for their action as elements distinctly different from promoters:

- An enhancer could change the overall structure of the template—for example, by influencing the density of supercoiling.
- It could be responsible for locating the template at a particular place within the cell—for example, attaching it to the nuclear matrix.
- An enhancer could provide an "entry site," a point at which RNA polymerase (or some other essential protein) initially associates with chromatin.

Now we take the view that enhancer function involves the same sort of interaction with the basal apparatus as the interactions sponsored by upstream promoter elements. Enhancers are modular, like promoters. Some elements are found in both enhancers and promoters. Some individual elements found in promoters share with enhancers the ability to function at variable distance and in either orientation. So the distinction between enhancers and promoters is blurred: enhancers might be viewed as containing promoter elements that are grouped closely together, with the ability to function at increased distances from the startpoint.

The essential role of the enhancer may be to increase the concentration of activator in the vicinity of the promoter (vicinity in this sense being a relative term). Two types of experiment illustrated in **Figure 21.24** suggest that this is the case.

A fragment of DNA that contains an enhancer at one end and a promoter at the other is not effectively transcribed, but the enhancer can stimulate transcription from the promoter when they are connected by a protein bridge. Since structural effects, such as changes in supercoiling, could not be transmitted across such a bridge, this suggests that the critical feature is bringing the enhancer and promoter into close proximity.

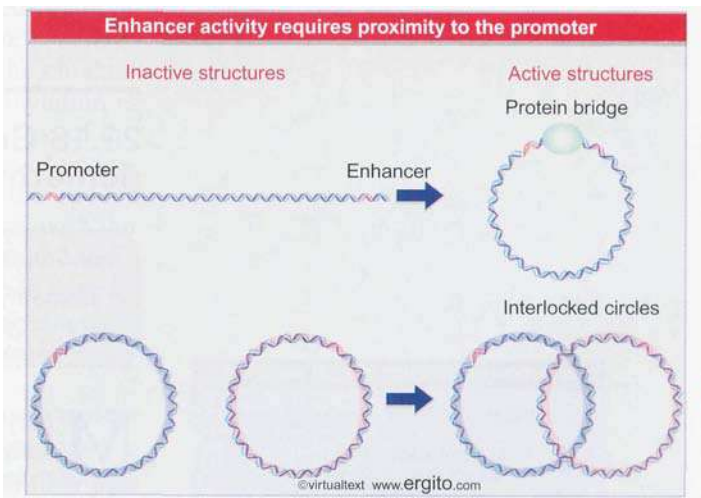


Figure 21.24 An enhancer may function by bringing proteins into the vicinity of the promoter. An enhancer does not act on a promoter at the opposite end of a long linear DNA, but becomes effective when the DNA is joined into a circle by a protein bridge. An enhancer and promoter on separate circular DNAs do not interact, but can interact when the two molecules are catenated.

A bacterial enhancer provides a binding site for the regulator NtrC, which acts upon RNA polymerase using promoters recognized by σ^{54}. When the enhancer is placed upon a circle of DNA that is catenated (interlocked) with a circle that contains the promoter, initiation is almost as effective as when the enhancer and promoter are on the same circular molecule. But there is no initiation when the enhancer and promoter are on separated circles. Again this suggests that the critical feature is localization of the protein bound at the enhancer, to increase its chance of contacting a protein bound at the promoter.

If proteins bound at an enhancer several kb distant from a promoter interact directly with proteins bound in the vicinity of the startpoint, the organization of DNA must be flexible enough to allow the enhancer and promoter to be closely located. This requires the intervening DNA to be extruded as a large "loop." Such loops have been directly observed in the case of the bacterial enhancer.

There is an interesting exception to the rule that enhancers are *cis*-acting in natural situations. This is seen in the phenomenon of transvection. Pairing of somatic chromosomes allows an enhancer on one chromosome to activate a promoter on the partner chromosome. This reinforces the view that enhancers work by proximity.

What limits the activity of an enhancer? Typically it works upon the nearest promoter. There are situations in which an enhancer is located between two promoters, but activates only one of them on the basis of specific protein-protein contacts between the complexes bound at the two elements. The action of an enhancer may be limited by an insulator—an element in DNA that prevents it from acting on promoters beyond (see *21.20 Insulators block the actions of enhancers and heterochromatin).*

The generality of enhancement is not yet clear. We do not know what proportion of cellular promoters require an enhancer to achieve their usual level of expression. Nor do we know how often an enhancer provides a target for regulation. Some enhancers are activated only in the tissues in which their genes function, but others could be active in all cells.

21.18 Gene expression is associated with demethylation

Key Concepts

- Demethylation at the 5' end of the gene is necessary for transcription.

Methylation of DNA is one of the parameters that controls transcription. Methylation in the vicinity of the promoter is associated with the absence of transcription. This is one of several regulatory events that influence the activity of a promoter; like the other regulatory events, typically this will apply to both (allelic) copies of the gene. However, methylation also occurs as an epigenetic event that can distinguish alleles whose sequences are identical. This can result in differences in the expression of the paternal and maternal alleles (see *23.20 DNA methylation is responsible for imprinting).* In this chapter we are concerned with the means by which methylation influences transcription.

The distribution of methyl groups can be examined by taking advantage of restriction enzymes that cleave target sites containing the CG doublet. Two types of restriction activity are compared in **Figure 21.25.**

These isoschizomers are enzymes that cleave the same target sequence in DNA, but have a different response to its state of methylation.

The enzyme HpaII cleaves the sequence CCGG (writing the sequence of only one strand of DNA). But if the second C is methylated, the enzyme can no longer recognize the site. However, the enzyme MspI cleaves the same target site *irrespective* of the state of methylation at this C. So MspI can be used to identify all the CCGG sequences; and HpaII can be used to determine whether or not they are methylated.

With a substrate of nonmethylated DNA, the two enzymes would generate the same restriction bands. But in methylated DNA, the modified positions are not cleaved by HpaII. For every such position, one larger HpaII fragment replaces two MspI fragments. **Figure 21.26** gives an example.

Many genes show a pattern in which the state of methylation is constant at most sites, but varies at others. Some of the sites are methylated in all tissues examined; some sites are unmethylated in all tissues. *A minority of sites are methylated in tissues in which the gene is not expressed, but are not methylated in tissues in which the gene is active.* So an active gene may be described as *undermethylated.*

Experiments with the drug 5-azacytidine produce indirect evidence that demethylation can result in gene expression. The drug is incorporated into DNA in place of cytidine, and cannot be methylated, because the 5' position is blocked. This leads to the appearance of demethylated sites in DNA as the consequence of replication (following the scheme on the right of Figure 14.35).

The phenotypic effects of 5-azacytidine include the induction of changes in the state of cellular differentiation; for example, muscle cells are induced to develop from nonmuscle cell precursors. The drug also activates genes on a silent X chromosome, which raises the possibility that the state of methylation could be connected with chromosomal inactivity.

As well as examining the state of methylation of resident genes, we can compare the results of introducing methylated or nonmethylated DNA into new host cells. Such experiments show a clear correlation: *the methylated gene is inactive, but the nonmethylated gene is active.*

What is the extent of the undermethylated region? In the chicken α-globin gene cluster in adult erythroid cells, the undermethylation is confined to sites that extend from ~500 bp upstream of the first of the two adult a genes to ~500 bp downstream of the second. Sites of undermethylation are present in the entire region, including the spacer between the genes. The region of undermethylation coincides with the region of maximum sensitivity to DNAase I. This argues that undermethylation is a feature of a domain that contains a transcribed gene or genes. As with other changes in chromatin, it seems likely that the absence of methyl groups is associated with the *ability to be transcribed* rather than with the act of transcription itself.

Our problem in interpreting the general association between undermethylation and gene activation is that only a minority (sometimes a small minority) of the methylated sites are involved. It is likely that the state of methylation is critical at specific sites or in a restricted region. It is also possible that a reduction in the level of methylation (or even the complete removal of methyl groups from some stretch of DNA) is part of some structural change needed to permit transcription to proceed.

In particular, demethylation at the promoter may be necessary to make it available for the initiation of transcription. In the γ-globin gene, for example, the presence of methyl groups in the region around the startpoint, between -200 and +90, suppresses transcription. Removal of the 3 methyl groups located upstream of the startpoint or

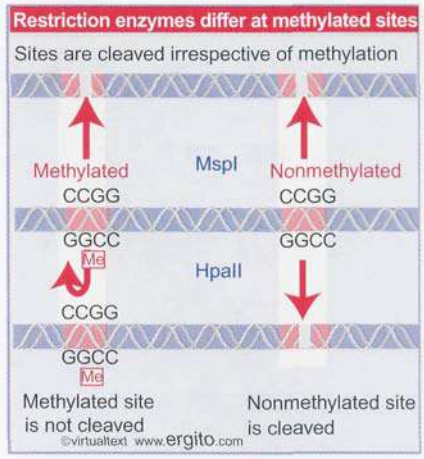


Figure 21.25 The restriction enzyme MspI cleaves all CCGG sequences whether or not they are methylated at the second C, but HpaII cleaves only nonmethylated CCGG tetramers.

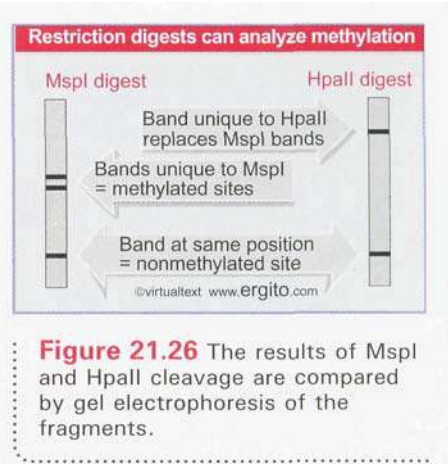


Figure 21.26 The results of MspI and HpaII cleavage are compared by gel electrophoresis of the fragments.

of the 3 methyl groups located downstream does not relieve the suppression. But removal of all methyl groups allows the promoter to function. Transcription may therefore require a methyl-free region at the promoter (see next section). There are exceptions to this general relationship.

Some genes can be expressed even when they are extensively methylated. Any connection between methylation and expression thus is not universal in an organism, but the general rule is that methylation prevents gene expression and demethylation is required for expression.

21.19 CpG islands are regulatory targets

Key Concepts

- CpG islands surround the promoters of constitutively expressed genes where they are unmethylated.
- They are also found at the promoters of some tissue-regulated genes.
- There are ~29,000 CpG islands in the human genome.
- Methylation of a CpG island prevents activation of a promoter within it.
- Repression is caused by proteins that bind to methylated CpG doublets.

The presence of CpG islands in the 5' regions of some genes is connected with the effect of methylation on gene expression. These islands are detected by the presence of an increased density of the dinucleotide sequence, CpG.

The CpG doublet occurs in vertebrate DNA at only ~20% of the frequency that would be expected from the proportion of G·C base pairs. (This may be because CpG doublets are methylated on C, and spontaneous deamination of methyl-C converts it to T, introducing a mutation that removes the doublet.) In certain regions, however, the density of CpG doublets reaches the predicted value; in fact, it is increased by 10X relative to the rest of the genome. The CpG doublets in these regions are unmethylated.

These CpG-rich islands have an average G-C content of ~60%, compared with the 40% average in bulk DNA. They take the form of stretches of DNA typically 1-2 kb long. There are ~45,000 such islands altogether in the human genome. Some of the islands are present in repeated Alu elements, and may just be the consequence of their high G·C-content. The human genome sequence confirms that, excluding these, there are ~29,000 islands. There are fewer in the mouse genome, ~15,500. About 10,000 of the predicted islands in both species appear to reside in a context of sequences that are conserved between the species, suggesting that these may be the islands with regulatory significance. The structure of chromatin in these regions has changes associated with gene expression (see *23.11 Promoter activation involves an ordered series of events);* there is a reduced content of histone H1 (which probably means that the structure is less compact), the other histones are extensively acetylated (a feature that tends to be associated with gene expression), and there are hypersensitive sites (as would be expected of active promoters).

In several cases, CpG-rich islands begin just upstream of a promoter and extend downstream into the transcribed region before petering out. **Figure 21.27** compares the density of CpG doublets in a "general"

region of the genome with a CpG island identified from the DNA sequence. The CpG island surrounds the 5' region of the APRT gene, which is constitutively expressed.

All of the "housekeeping" genes that are constitutively expressed have CpG islands; this accounts for about half of the islands altogether. The other half of the islands occur at the promoters of tissue-regulated genes; only a minority (<40%) of these genes have islands. In these cases, the islands are unmethylated irrespective of the state of expression of the gene. The presence of unmethylated CpG-rich islands may be necessary, but therefore is not sufficient, for transcription. So the presence of unmethylated CpG islands may be taken as an indication that a gene is potentially active, rather than inevitably transcribed. Many islands that are nonmethylated in the animal become methylated in cell lines in tissue culture, and this could be connected with the inability of these lines to express all of the functions typical of the tissue from which they were derived.

Methylation of a CpG island can affect transcription. Two mechanisms can be involved:

- Methylation of a binding site for some factor may prevent it from binding. This happens in a case of binding to a regulatory site other than the promoter (see *23.21 Oppositely imprinted genes can be controlled by a single center).*
- Or methylation may cause specific repressors to bind to the DNA.

Repression is caused by either of two types of protein that bind to methylated CpG sequences. The protein MeCP1 requires the presence of several methyl groups to bind to DNA, while MeCP2 and a family of related proteins can bind to a single methylated CpG base pair. This explains why a methylation-free zone is required for initiation of transcription. Binding of proteins of either type prevents transcription *in vitro* by a nuclear extract.

MeCP2, which directly represses transcription by interacting with complexes at the promoter, is bound also to the Sin3 repressor complex, which contains histone deacetylase activities (see Figure 23.15). This observation provides a direct connection between two types of repressive modifications: methylation of DNA and acetylation of histones.

The absence of methyl groups is associated with gene expression. However, there are some difficulties in supposing that the state of methylation provides a general means for controlling gene expression. In the case of *D. melanogaster* (and other Dipteran insects), there is very little methylation of DNA (although there is gene potentially coding a methyltransferase), and in the nematode *C. elegans* there is no methylation of DNA. The other differences between inactive and active chromatin appear to be the same as in species that display methylation. So in these organisms, any role that methylation has in vertebrates is replaced by some other mechanism.

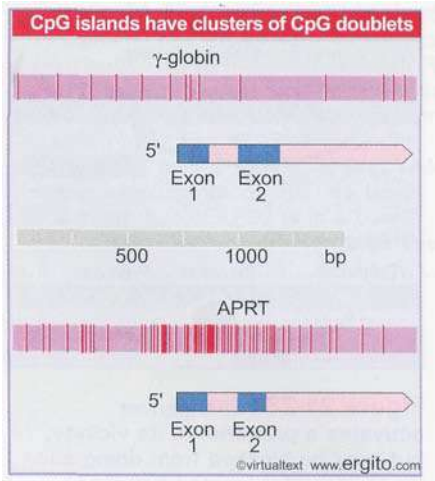


Figure 21.27 The typical density of CpG doublets in mammalian DNA is ~1/100 bp, as seen for a γ-globin gene. In a CpG-rich island, the density is increased to >10 doublets/100 bp. The island in the APRT gene starts ~100 bp upstream of the promoter and extends ~400 bp into the gene. Each vertical line represents a CpG doublet.

21.20 Insulators block the actions of enhancers and heterochromatin

Key Concepts

- Insulators are able to block passage of any activating or inactivating effects from enhancers, silencers, or LCRs.
- Insulators may provide barriers against the spread of heterochromatin.

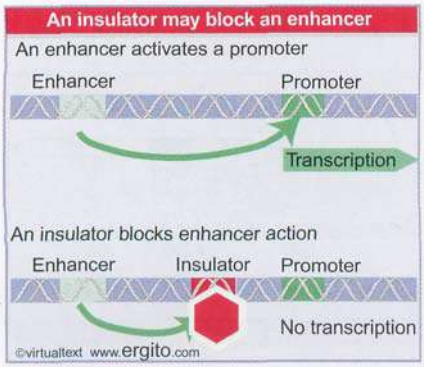


Figure 21.28 An enhancer activates a promoter in its vicinity, but may be blocked from doing so by an insulator located between them.

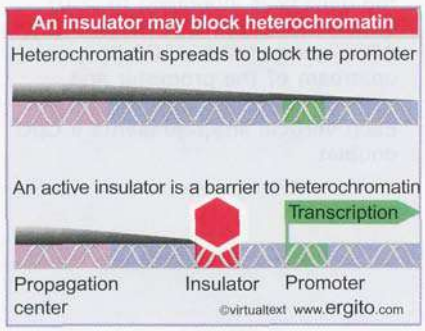


Figure 21.29 Heterochromatin may spread from a center and then blocks any promoters that it covers. An insulator may be a barrier to propagation of heterochromatin that allows the promoter to remain active.

Elements that prevent the passage of activating or inactivating effects are called insulators. They have either or both of two key properties:

- When an insulator is placed between an enhancer and a promoter, *it prevents the enhancer from activating the promoter*. The blocking effect is shown in **Figure 21.28.** This may explain how the action of an enhancer is limited to a particular promoter.
- When an insulator is placed between an active gene and heterochromatin, *it provides a barrier that protects the gene against the inactivating effect that spreads from the heterochromatin.* (Heterochromatin is a region of chromatin that is inactive as the result of its higher order structure; see *23.13 Heterochromatin propagates from a nucleation event.)* The barrier effect is shown in **Figure 21.29.**

Some insulators possess both these properties, but others have only one, or the blocking and barrier functions can be separated. Although both actions are likely to be mediated by changing chromatin structure, they may involve different effects. In either case, however, the insulator defines a limit for long-range effects.

What is the purpose of an insulator? A major function may be to counteract the indiscriminate actions of enhancers on promoters. Most enhancers will work with any promoter in the vicinity. An insulator can restrict an enhancer by blocking the effects from passing beyond a certain point, so that it can act only on a specific promoter. Similarly, when a gene is located near heterochromatin, an insulator can prevent it from being inadvertently inactivated by the spread of the heterochromatin. Insulators therefore function as elements for increasing the precision of gene regulation.

21.21 Insulators can define a domain

Key Concepts

- Insulators are specialized chromatin structures that have hypersensitive sites. Two insulators can protect the region between them from all external effects.

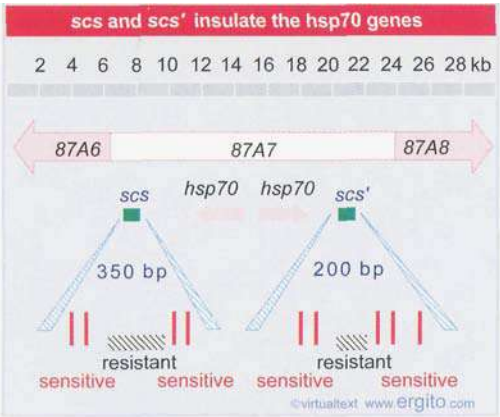


Figure 21.30 Specialized chromatin structures that include hypersensitive sites mark the ends of a domain in the *D. melanogaster* genome and insulate genes between them from the effects of surrounding sequences.

Insulators were discovered during the analysis of the region of the *D. melanogaster* genome summarized in **Figure 21.30.** Two genes for the protein Hsp70 lie within an 18 kb region that constitutes band 87A7. Special structures, called *scs* and *scs'* (specialized chromatin structures), are found at the ends of the band. Each consists of a region that is highly resistant to degradation by DNAase I, flanked on either side by hypersensitive sites, spaced at about 100 bp. The cleavage pattern at these sites is altered when the genes are turned on by heat shock.

The *scs* elements insulate the *hsp 70* genes from the effects of surrounding regions. If we take *scs* units and place them on either side of a *white* gene, the gene can function anywhere it is placed in the genome, even in sites where it would normally be repressed by context, for example, in heterochromatic regions.

The *scs* and *scs'* units do not seem to play either positive or negative roles in controlling gene expression, but just restrict effects from passing from one region to the next. If adjacent regions have repressive effects, however, the *scs* elements might be needed to block the spread of such effects, and therefore could be essential for gene expression. In this case, deletion of such elements could eliminate the expression of the adjacent gene(s).

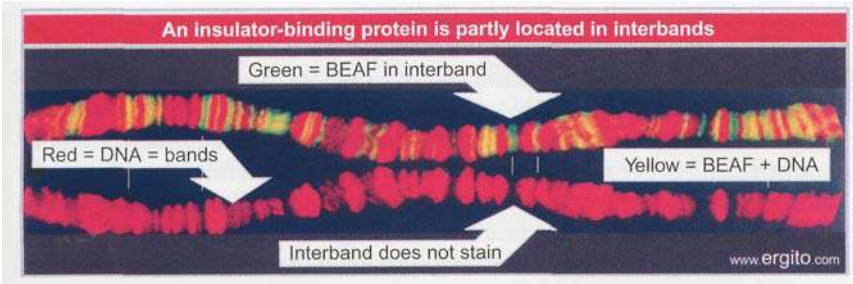


Figure 21.31 A protein that binds to the insulator *scs'* is localized at interbands in *Drosophila* polytene chromosomes. Red staining identifies the DNA (the bands) on both the upper and lower samples; green staining identifies BEAF32 (often at interbands) on the upper sample. Yellow shows coincidence of the two labels (meaning that BEAF32 is in a band). Photograph kindly provided by Uli Laemmli.

scs and *scs'* have different structures and each appears to have a different basis for its insulator activity. The key sequence in the *scs* element is a stretch of 24 bp that binds the product of the *zw5* gene. The insulator property of *scs'* resides in a series of CGATA repeats. The repeats bind a group of related proteins called BEAF-32. The protein shows discrete localization within the nucleus, but the most remarkable data derive from its localization on polytene chromosomes. **Figure 21.31** shows that an anti-BEAF-32 antibody stains ~50% of the interbands of the polytene chromosomes. This suggests that there are many insulators in the genome, and that BEAF-32 is a common part of the insulating apparatus. It would imply that the band is a functional unit, and that interbands often have insulators that block the propagation of activating or inactivating effects.

Another example of an insulator that defines a domain is found in the chick β-globin LCR (the group of hypersensitive sites that controls expression of all β-globin genes; see *20.16 An LCR may control a domain)*. The leftmost hypersensitive site of the chick β-globin LCR (HS4) is an insulator that marks the 5' end of the functional domain. This restricts the LCR to acting only on the globin genes in the domain.

A gene that is surrounded by insulators is usually protected against the propagation of inactivating effects from the surrounding regions. The test is to inserted DNA into a genome at random locations by transfection. The expression of a gene in the inserted sequence is often erratic; in some instances it is properly expressed, but in others it is extinguished (see *18.18 Genes can be injected into animal eggs)*. However, when insulators that have a barrier function are placed on either side of the gene in the inserted DNA, its expression typically is uniform in every case.

21.22 Insulators may act in one direction

Key Concepts

- Some insulators have directionality, and may stop passage of effects in one direction but not the other.

Insulators may have directional properties. Insertions of the transposon *gypsy* into the *yellow (y)* locus of *D. melanogaster* cause loss of gene function in some tissues, but not in others. The reason is that the *y* locus is regulated by four enhancers, as shown in **Figure 21.32.** Wherever *gypsy* is inserted, it blocks expression of all enhancers that it separates from the promoter, but not those that lie on the other side. The sequence responsible for this effect is an insulator that lies at one end of the transposon. The insulator works irrespective of its orientation of insertion.

Some of the enhancers are upstream of the promoter and others are downstream, so the effect cannot depend on position with regard to the

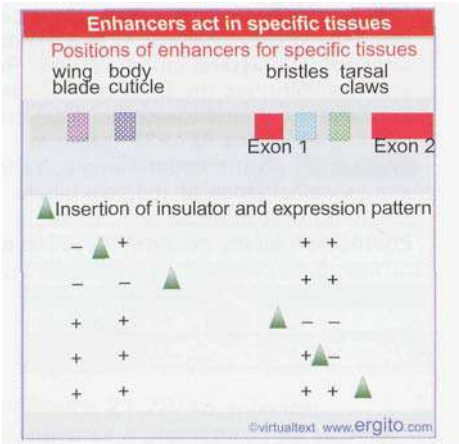


Figure 21.32 The insulator of the gypsy transposon blocks the action of an enhancer when it is placed between the enhancer and the promoter.

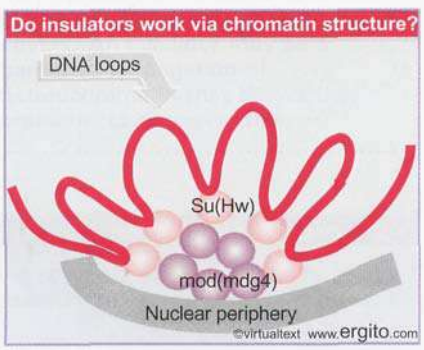


Figure 21.33 Su(Hw)/mod(mdg4) complexes are found at the nuclear periphery. They may organize DNA into loops that limit enhancer-promoter interactions.

promoter, nor can it require transcription to occur through the insulator. This is difficult to explain in terms of looping models for enhancer-promoter interaction, which essentially predict the irrelevance of the intervening DNA. The obvious model to invoke is a tracking mechanism, in which some component must move unidirectionally from the enhancer to the promoter, but this is difficult to reconcile with previous characterizations of the independence of enhancers from such effects.

Proteins that act upon the insulator have been identified through the existence of two other loci that affect insulator function in a *trans-acting* manner. Mutations in *su(Hw)* abolish insulation: *y* is expressed in all tissues in spite of the presence of the insulator. This suggests that *su(Hw)* codes for a protein that recognizes the insulator and is necessary for its action. Su(Hw) has a zinc finger DNA-motif; mapping to polytene chromosomes shows that it is bound at a large number of sites. The insulator contains 12 copies of a 26 bp sequence that is bound by Su(Hw). Manipulations show that the strength of the insulator is determined by the number of copies of the binding sequence.

The second locus is *mod(mdg4)*, in which mutations have the opposite effect. This is observed by the loss of directionality. These mutations increase the effectiveness of the insulator by extending its effects so that it blocks utilization of enhancers on both sides. *su(Hw)* is epistatic to *mod(mdg4)*; this means that in a double mutant we see only the effect of *su(Hw)*. This implies that mod(mdg4) acts through su(Hw). The basic role of the wild-type protein from the *mod(mdg4)* locus is therefore to impose directionality on the ability of su(Hw) to insulate promoters from the boundary.

Binding of su(Hw) to DNA, followed by binding of mod(mdg4) to su(Hw), therefore creates a unidirectional block to activation of a promoter. This suggests that the insulator bound by su(Hw) can spread inactivity in both directions, but mod(mdg4) stops the effect from spreading in one direction. Perhaps there is some intrinsic directionality to chromatin, which results ultimately in the incorporation of su(Hw), mod(mdg4), or some other component in one orientation, presumably by virtue of an interaction with some component of chromatin that is itself preferentially oriented. Any such directionality would need to reverse at the promoter.

It is likely that insulators act by making changes in chromatin structure. One model is prompted by the observation that Su(Hw) and mod(mdg4) binding sites are present at >500 locations in the Drosophila genome. But visualization of the sites where the proteins are bound in the nucleus shows that they are colocalized at ~25 discrete sites around the nuclear periphery. This suggests the model of **Figure 21.33** in which Su(Hw) proteins bound at different sites on DNA are brought together by binding to mod(mdg4). The Su(Hw)/mod(mdg4) complex is localized at the nuclear periphery. The DNA bound to it is organized into loops. An average complex might have ~20 such loops. Enhancer-promoter actions can occur only within a loop, and cannot propagate between them.

21.23 Insulators can vary in strength

Key Concepts

- Insulators can differ in how effectively they block passage of an activating signal.

Sometimes elements with different *cis*-acting properties are combined to generate regions with complex regulatory effects. The *Fab-7* region is defined by deletions in the *bithorax* locus of *Drosophila*. This locus contains a series of *cis*-acting regulatory elements that control the activities of three transcription units (see Figure 31.36). The relevant part of the locus is drawn in **Figure 21.34.** The regulatory elements *iab-6* and *iab-7* control expression of the adjacent gene *Abd-B* in successive regions of the embryo (segments A6 and A7). A deletion of *Fab-7* causes A6 to develop like A7, instead of in the usual way. This is a dominant effect, which suggests that *iab-7* has taken over control from *iab-6*. We can interpret this in molecular terms by supposing that *Fab-7* provides a boundary that prevents *iab-7* from acting when *iab-6* is usually active.

Like other boundary elements, *Fab-7* contains a distinctive chromatin structure that is marked by a series of hypersensitive sites. The region can be divided into two types of elements by smaller deletions and by testing fragments for their ability to provide a boundary. A sequence of ~3.3 kb behaves as an insulator when it is placed in other constructs. A sequence of ~0.8 kb behaves as a repressor that acts on *iab-7*. The presence of these two elements explains the complicated genetic behavior of *Fab-7* (which we have not described in detail).

An insight into the action of the boundary element is provided by the effects of substituting other insulators for *Fab-7*. The effect of *Fab-7* is simply to prevent interaction between *iab-6* and *iab-7*. But when *Fab-7* is replaced by a different insulator [in fact a binding site for the protein Su(Hw)], a stronger effect is seen: *iab-5* takes over from *iab-7*. And when an *scs* element is used, the effect extends to *iab-4*. This suggests a scheme in which stronger elements can block the actions of regulatory sequences that lie farther away.

This conclusion introduces a difficulty for explaining the action of boundary elements. They cannot be functioning in this instance simply by preventing the transmission of effects past the boundary. This argues against models based on simple tracking or inhibiting the linear propagation of structural effects. It suggests that there may be some sort of competitive effect, in which the strength of the element determines how far its effect can stretch.

The situation is further complicated by the existence of anti-insulator elements, which allow an enhancer to overcome the blocking effects of an insulator. This again suggests that these effects are mediated by some sort of control over local chromatin structure.

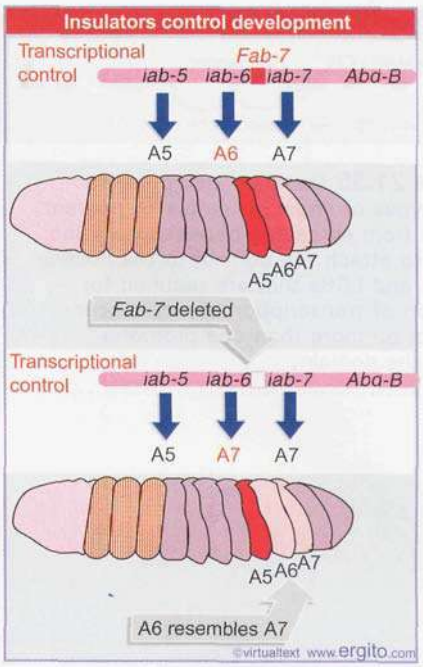


Figure 21.34 *Fab-7* is a boundary element that is necessary for the independence of regulatory elements *iab-6* and *iab-7*.

21.24 What constitutes a regulatory domain?

Key Concepts

- A domain has an insulator, an LCR, a matrix attachment site, and transcription unit(s).

If we now put together the various types of structures that have been found in different systems, we can think about the possible nature of a chromosomal domain. The basic feature of a regulatory domain is that regulatory elements can act only on transcription units within the same domain. A domain might contain more than one transcription unit and/or enhancer.

Figure 21.35 summarizes the structures that might be involved in defining a domain.

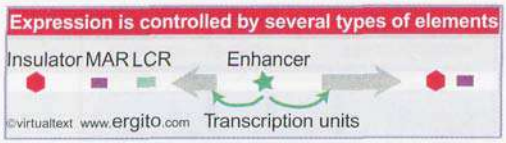


Figure 21.35 Domains may possess three types of sites: insulators to prevent effects from spreading between domains; MARs to attach the domain to the nuclear matrix; and LCRs that are required for initiation of transcription. An enhancer may act on more than one promoter within the domain.

An insulator stops activating or repressing effects from passing. In its simplest form, an insulator blocks either type of effect from passing across it, but there can be more complex relationships in which the insulator blocks only one type of effect and/or acts directionally. We assume that insulators act by affecting higher order chromatin structure, but we do not know the details and varieties of such effects.

A matrix attachment site (MAR) may be responsible for attaching chromatin to a site on the nuclear periphery (see *19.6 Specific sequences attach DNA to an interphase matrix).* These are likely to be responsible for creating physical domains of DNA that take the form of loops extending out from the attachment sites. This looks like one model for insulator action. In fact, some MAR elements behave as insulators in assays *in vitro,* but it seems that the their ability to attach DNA to the matrix can be separated from the insulator function, so there is not a simple cause and effect. It would not be surprising if insulator and MAR elements were associated to maintain a relationship between regulatory effects and physical structure.

An LCR functions at a distance and may be required for any and all genes in a domain to be expressed (see *20.16 An LCR may control a domain).* When a domain has an LCR, its function is essential for all genes in the domain, but LCRs do not seem to be common. Several types of *cis*-acting structures could be required for function. As defined originally, the property of the LCR rests with an enhancer-like hypersensitive site that is needed for the full activity of promoter(s) within the domain.

The organization of domains may help to explain the large size of the genome. A certain amount of space could be required for such a structure to operate, for example, to allow chromatin to become decondensed and to become accessible. Although the exact sequences of much of the unit might be irrelevant, there might be selection for the overall amount of DNA within it, or at least selection might prevent the various transcription units from becoming too closely spaced.

21.25 Summary

Of the three eukaryotic RNA polymerases, RNA polymerase I transcribes rDNA and accounts for the majority of activity, RNA polymerase II transcribes structural genes for mRNA and has the greatest diversity of products, and RNA polymerase III transcribes small RNAs. The enzymes have similar structures, with two large subunits and many smaller subunits; there are some common subunits among the enzymes.

None of the three RNA polymerases recognize their promoters directly. A unifying principle is that transcription factors have primary responsibility for recognizing the characteristic sequence elements of any particular promoter, and they serve in turn to bind the RNA polymerase and to position it correctly at the startpoint. At each type of promoter, the initiation complex is assembled by a series of reactions in which individual factors join (or leave) the complex. The factor TBP is required for initiation by all three RNA polymerases. In each case it provides one subunit of a transcription factor that binds in the vicinity of the startpoint.

A promoter consists of a number of short sequence elements in the region upstream of the startpoint. Each element is bound by a transcription factor. The basal apparatus, which consists of the TF$_{II}$ factors, assembles at the startpoint and enables RNA polymerase to bind. The TATA box (if there is one) near the startpoint, and the initiator region immediately at the startpoint, are responsible for selection of the exact startpoint at promoters for RNA polymerase II. TBP binds directly to the TATA box when there is one; in TATA-less promoters it is located near the startpoint by binding to the DPE down-

stream. After binding of $TF_{II}D$, the other general transcription factors for RNA polymerase II assemble the basal transcription apparatus at the promoter. Other elements in the promoter, located upstream of the TATA box, bind activators that interact with the basal apparatus. The activators and basal factors are released when RNA polymerase begins elongation.

The CTD of RNA polymerase II is phosphorylated during the initiation reaction. $TF_{II}D$ and SRB proteins both may interact with the CTD. It may also provide a point of contact for proteins that modify the RNA transcript, including the 5' capping enzyme, splicing factors, and the 3' processing complex.

Promoters may be stimulated by enhancers, sequences that can act at great distances and in either orientation on either side of a gene. Enhancers also consist of sets of elements, although they are more compactly organized. Some elements are found in both promoters and enhancers. Enhancers probably function by assembling a protein complex that interacts with the proteins bound at the promoter, requiring that DNA between is "looped out."

An insulator blocks the transmission of activating or inactivating effects in chromatin. An insulator that is located between an enhancer and a promoter prevents the enhancer from activating the promoter. Two insulators define the region between them as a regulatory domain; regulatory interactions within the domain are limited to it, and the domain is insulated from outside effects. Most insulators block regulatory effects from passing in either direction, but some are directional. Insulators usually can block both activating effects (enhancer-promoter interactions) and inactivating effects (mediated by spread of heterochromatin), but some are limited to one or the other. Insulators are thought to act via changing higher order chromatin structure, but the details are not certain.

CpG islands contain concentrations of CpG doublets and often surround the promoters of constitutively expressed genes, although they are also found at the promoters of regulated genes. The island including a promoter must be unmethylated for that promoter to be able to initiate transcription. A specific protein binds to the methylated CpG doublets and prevents initiation of transcription.

References

21.2 Eukaryotic RNA polymerases consist of many subunits

rev Doi, R. H. and Wang, L.-F. (1986). Multiple prokaryotic RNA polymerase sigma factors. Microbiol. Rev. 50, 227-243.

Young, R. A. (1991). RNA polymerase II. Ann. Rev. Biochem. 60, 689-715.

21.4 RNA polymerase I has a bipartite promoter

rev Paule, M. R. and White, R. J. (2000). Survey and summary: transcription by RNA polymerases I and III Nuc. Acids Res. 28, 1283-1298.

ref Bell, S. P., Learned, R. M., Jantzen, H. M., and Tjian, R. (1988). Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. Science 241, 1192-1197.

21.5 RNA polymerase III uses both downstream and upstream promoters

ref Bogenhagen, D. F., Sakonju, S., and Brown, D. D. (1980). A control region in the center of the 5S RNA gene directs specific initiation of transcription: II the 3' border of the region. Cell 19, 27-35.

Galli, G., Hofstetter, H., and Birnstiel, M. L. (1981). Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature 294, 626-631.

Kunkel, G. R. and Pederson, T. (1988). Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used. Genes Dev. 2, 196-204.

Pieler, T., Hamm, J., and Roeder, R. G. (1987). The 5S gene internal control region is composed of three distinct sequence elements, organized as two functional domains with variable spacing. Cell 48, 91-100.

Sakonju, S., Bogenhagen, D. F., and Brown, D. D. (1980). A control region in the center of the 5S RNA gene directs specific initiation of transcription: I the 5' border of the region. Cell 19, 13-25.

21.6 $TF_{III}B$ is the commitment factor for pol III promoters

rev Geiduschek, E. P. and Tocchini-Valentini, G. P. (1988). Transcription by RNA polymerase III. Ann. Rev. Biochem. 57, 873-914.

Schramm, L. and Hernandez, N. (2002). Recruitment of RNA polymerase III to its target promoters. Genes Dev. 16, 2593-2620.

ref Kassavatis, G. A., Braun, B. R., Nguyen, L. H., and Geiduschek, E. P. (1990). S. *cerevisiae* TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors. Cell 60, 235-245.

Kassavetis, G. A., Joazeiro, C. A., Pisano, M., Geiduschek, E. P., Colbert, T., Hahn, S., and Blanco, J. A. (1992). The role of the TATA-binding protein in the assembly and function of the multisubunit yeast RNA polymerase III transcription factor, TFIIIB. Cell 71, 1055-1064.

Kassavetis, G. A., Letts, G. A., and Geiduschek, E. P. (1999). A minimal RNA polymerase III transcription system. EMBO J. 18, 5042-5051.

21.7 The startpoint for RNA polymerase II

rev Butler, J. E. and Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. 16, 2583-2592.

Smale, S. T., Jain, A., Kaufmann, J., Emami, K. H., Lo, K., and Garraway, I. P. (1998). The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harb Symp Quant Biol 63, 21-31.

Woychik, N. A. and Hampsey, M. (2002). The RNA polymerase II machinery: structure illuminates function. Cell 108, 453-463.

ref Burke, T. W. and Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 10, 711-724.

Singer, V. L., Wobbe, C. R., and Struhl, K. (1990). A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. Genes Dev. 4, 636-645.

Smale, S. T. and Baltimore, D. (1989). The "initiator" as a transcription control element. Cell 57, 103-113.

Weil, P. A., Luse, D. S., Segall, J., and Roeder, R. G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18, 469-484.

21.8 TBP is a universal factor

rev Berk, A. J. (2000). TBP-like factors come into focus. Cell 103, 5-8.

Hernandez, N. (1993). TBP, a universal eukaryotic transcription factor? Genes Dev. 7, 1291-1308.

Lee, T. I. and Young, R. A. (1998). Regulation of gene expression by TBP-associated proteins. Genes Dev. 12, 1398-1408.

ref Crowley, T. E., Hoey, T., Liu, J. K., Jan, Y. N., Jan, L. Y., and Tjian, R. (1993). A new factor related to TATA-binding protein has highly restricted expression patterns in *Drosophila*. Nature 361, 557-561.

21.9 TBP binds DNA in an unusual way

rev Burley, S. K. and Roeder, R. G. (1996). Biochemistry and structural biology of TFIID. Ann. Rev. Biochem. 65, 769-799.

Lee, T. I. and Young, R. A. (1998). Regulation of gene expression by TBP-associated proteins. Genes Dev. 12, 1398-1408.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev. 10, 2657-2683.

ref Horikoshi, M. et al. (1988). Transcription factor ATD interacts with a TATA factor to facilitate establishment of a preinitiation complex. Cell 54, 1033-1042.

Kim, Y. et al. (1993). Crystal structure of a yeast TBP/TATA box complex. Nature 365, 512-520.

Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993). Cocrystal structure of TBP recognizing the minor groove of a TATA element. Nature 365, 520-527.

Liu, D. et al. (1998). Solution structure of a TBP-TAFII230 complex: protein mimicry of the minor groove surface of the TATA box unwound by TBP. Cell 94, 573-583.

Martinez, E. et al. (1994). TATA-binding protein-associated factors in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J. 13, 3115-3126.

Nikolov, D. B. et al. (1992). Crystal structure of TFIID TATA-box binding protein. Nature 360, 40-46.

Ogryzko, V. V. et al. (1998). Histone-like TAFs within the PCAF histone acetylase complex. Cell 94, 35-44.

Verrijzer, C. P. et al. (1995). Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. Cell 81, 1115-1125.

Zhao, X. and Herr, W. (2002). A regulated two-step mechanism of TBP binding to DNA: a solvent-exposed surface of TBP inhibits TATA box recognition. Cell 108, 615-627.

21.10 The basal apparatus assembles at the promoter

rev Nikolov, D. B. and Burley, S. K. (1997). RNA polymerase II transcription initiation: a structural view. Proc. Nat. Acad. Sci. USA 94, 15-22.

Zawel, L. and Reinberg, D. (1993). Initiation of transcription by RNA polymerase II: a multi-step process. Prog. Nucleic Acid Res. Mol. Biol. 44, 67-108.

ref Buratowski, S., Hahn, S., Guarente, L., and Sharp, P. A. (1989). Five intermediate complexes in transcription initiation by RNA polymerase II. Cell 56, 549-561.

Burke, T. W. and Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 10, 711-724.

Littlefield, O., Korkhin, Y., and Sigler, P. B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. Proc. Nat. Acad. Sci. USA 96, 13668-13673.

Nikolov, D. B. et al. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature 377, 119-128.

21.11 Initiation is followed by promoter clearance

rev Price, D. H. (2000). P-TEFb, a cyclin dependent kinase controlling elongation by RNA polymerase II. Mol. Cell Biol. 20, 2629-2634.

Proudfoot, N. J., Furger, A., and Dye, M. J. (2002). Integrating mRNA processing with transcription. Cell 108, 501-512.

Woychik, N. A. and Hampsey, M. (2002). The RNA polymerase II machinery: structure illuminates function. Cell 108, 453-463.

ref Douziech, M., Coin, F., Chipoulet, J. M., Arai, Y., Ohkuma, Y., Egly, J. M., and Coulombe, B. (2000). Mechanism of promoter melting by the xeroderma pigmentosum complementation group B helicase of transcription factor MH revealed by protein-DNA photo-cross-linking. Mol. Cell Biol. 20, 8168-8177.

Fong, N. and Bentley, D. L. (2001). Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes Dev. 15, 1783-1795.

Goodrich, J. A. and Tjian, R. (1994). Transcription factors HE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. Cell 77, 145-156.

Hirose, Y. and Manley, J. L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev. 14, 1415-1429.

Holstege, F. C, van der Vliet, P. C, and Timmers, H. T. (1996). Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors HE and IIH. EMBO J. 15, 1666-1677.

Kim, T. K., Ebright, R. H., and Reinberg, D. (2000). Mechanism of ATP-dependent promoter melting by transcription factor IIH. Science 288, 1418-1422.

Spangler, L., Wang, X., Conaway, J. W., Conaway, R. C, and Dvir, A. (2001). TFIIH action in transcription initiation and promoter escape requires distinct regions of downstream promoter DNA. Proc. Nat. Acad. Sci. USA 98, 5544-5549.

21.12 A connection between transcription and repair

rev Selby, C. P. and Sancar, A. (1994). Mechanisms of transcription-repair coupling and mutation frequency decline. Microbiol. Rev. 58, 317-329.

ref Bregman, D. et al. (1996). UV-induced ubiquitination of RNA polymerase II: a novel modification deficient in Cockayne syndrome cells. Proc. Nat. Acad. Sci. USA 93, 11586-11590.

Lehmann, A. R. (2001). The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. Genes Dev. 15, 15-23.

Schaeffer, L. et al. (1993). DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor. Science 260, 58-63.

Selby, C. P. and Sancar, A. (1993). Molecular mechanism of transcription-repair coupling. Science 260, 53-58.

Svejstrup, J. Q. et al. (1995). Different forms of TFIIH for transcription and DNA repair: holo-TFIIH and a nucleotide excision repairosome. Cell 80, 21-28.

21.15 Enhancers contain bidirectional elements that assist initiation

rev Muller, M. M., Gerster, T., and Schaffner, W. (1988). Enhancer sequences and the regulation of gene transcription. Eur. J. Biochem. 176, 485-495.

ref Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308.

21.16 Enhancers contain the same elements that are found at promoters

ref Maniatis, T., Falvo, J. V., Kim, T. H., Kim, T. K., Lin, C. H., Parekh, B. S., and Wathelet, M. G. (1998). Structure and function of the interferon-beta enhanceosome. Cold Spring Harbor Symp. Quant. Biol. 63, 609-620.

Munshi, N., Yie, Y., Merika, M., Senger, K., Lomvardas, S., Agalioti, T., and Thanos, D. (1999). The IFN-beta enhancer: a paradigm for understanding activation and repression of inducible gene expression. Cold Spring Harbor Symp. Quant. Biol. 64, 149-159.

21.17 Enhancers work by increasing the concentration of activators near the promoter

rev Blackwood, E. M. and Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. Science 281, 60-63.

ref Mueller-Storm, H. P., Sogo, J. M., and Schaffner, W. (1989). An enhancer stimulates transcription in *trans* when attached to the promoter via a protein bridge. Cell 58, 767-777.

Zenke, M. et al. (1986). Multiple sequence motifs are involved in SV40 enhancer function. EMBO J. 5, 387-397.

21.19 CpG islands are regulatory targets

rev Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. 16, 6-21.

ref Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc. Nat. Acad. Sci. USA 90, 11995-11999.

Bird, A. et al. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, Cp-G-rich DNA. Cell 40, 91-99.

Boyes, J. and Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell 64, 1123-1134.

21.20 Insulators block the actions of enhancers and heterochromatin

rev Gerasimova, T. I. and Corces, V. G. (2001). Chromatin insulators and boundaries: effects on transcription and nuclear organization. Ann. Rev. Genet. 35, 193-208.

West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. Genes Dev. 16, 271-288.

21.21 Insulators can define a domain

ref Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5' element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila.* Cell 74, 505-514.

Cuvier, 0., Hart, C. M., and Laemmli, U. K. (1998). Identification of a class of chromatin boundary elements. Mol. Cell Biol. 18, 7478-7486.

Gaszner, M., Vazquez, J., and Schedl, P. (1999). The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev. 13, 2098-2107.

Kellum, R. and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell 64, 941-950.

Pikaart, M. J., Recillas-Targa, F., and Felsenfeld, G. (1998). Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. Genes Dev. 12, 2852-2862.

Zhao, K., Hart, C. M., and Laemmli, U. K. (1995). Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell 81, 879-889.

21.22 Insulators may act in one direction

ref Gerasimova, T. I., Byrd, K., and Corces, V. G. (2000). A chromatin insulator determines the nuclear localization of DNA. Mol. Cell 6, 1025-1035.

Harrison, D. A., Gdula, D. A., Cyne, R. S., and Corces, V. G. (1993). A leucine zipper domain of the suppressor of hairy-wing protein mediates its repressive effect on enhancer function. Genes Dev. 7, 1966-1978.

Roseman, R. R., Pirrotta, V., and Geyer, P. K. (1993). The su(Hw) protein insulates expression of the *D. melanogasterwhite* gene from chromosomal position-effects. EMBO J. 12, 435-442.

21.23 Insulators can vary in strength

ref Hagstrom, K., Muller, M., and Schedl, P. (1996). Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the *Drosophila* bithorax complex. Genes Dev. 10, 3202-3215.

Mihaly, J. et al. (1997). *In situ* dissection of the Fab-7 region of the bithorax complex into a chromatin domain boundary and a Polycomb-response element. Development 124, 1809-1820.

Zhou, J. and Levine, M. (1999). A novel *cis*-regulatory element, the PTS, mediates an anti-insulator activity in the *Drosophila* embryo. Cell 99, 567-575.

21.24 What constitutes a regulatory domain?

rev West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. Genes Dev. 16, 271-288.

Chapter 22

Activating transcription

22.1 Introduction

Key Concepts

- Eukaryotic gene expression is usually controlled at the level of initiation of transcription.

The phenotypic differences that distinguish the various kinds of cells in a higher eukaryote are largely due to differences in the expression of genes that code for proteins, that is, those transcribed by RNA polymerase II. In principle, the expression of these genes might be regulated at any one of several stages. We can distinguish (at least) five potential control points, forming the series:

Activation of gene structure
i
Initiation of transcription
↓
Processing the transcript
i
Transport to cytoplasm
i
Translation of mRNA

As we see in **Figure 22.1,** gene expression in eukaryotes is largely controlled at the initiation of transcription. For most genes, this is the major control point in their expression. It involves changes in the structure of chromatin at the promoter (see *23.11 Promoter activation involves an ordered series of events),* accompanied by the binding of the basal transcription apparatus (including RNA polymerase II) to the promoter. (Regulation at subsequent stages of transcription is rare in eukaryotic cells. Premature termination occurs at some genes, and is counteracted by a kinase, P-TEFb, but otherwise anti-termination does not seem to be employed.)

The primary transcript is modified by capping at the 5' end, and usually also by polyadenylation at the 3' end. Introns must be excised from the transcripts of interrupted genes. The mature RNA must be exported from the nucleus to the cytoplasm. Regulation of gene expression by selection of sequences at the level of nuclear RNA might involve any or all of these stages, but the one for which we have most

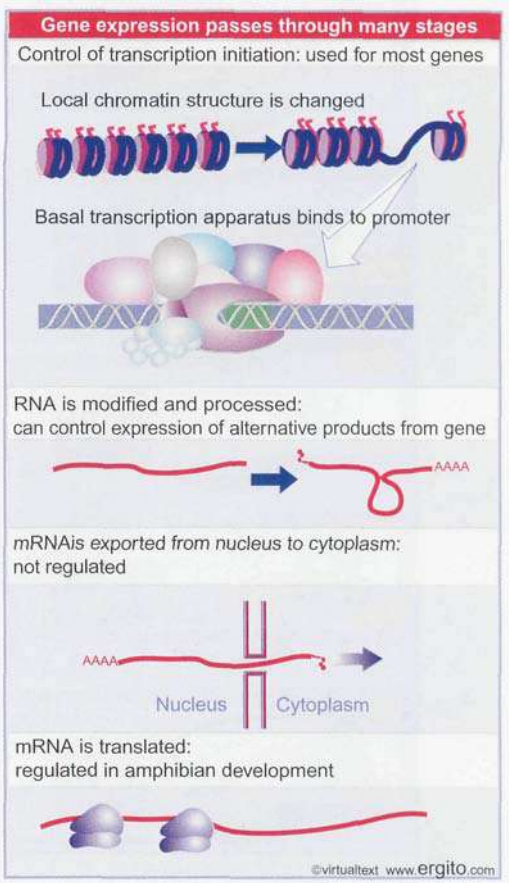


Figure 22.1 Gene expression is controlled principally at the initiation of transcription, and it is rare for the subsequent stages of gene expression to be used to determine whether a gene is expressed, although control of processing may be used to determine which form of a gene is represented in mRNA.

By Book_Crazy [IND]

evidence concerns changes in splicing; some genes are expressed by means of alternative splicing patterns whose regulation controls the type of protein product (see *24.12 Alternative splicing involves differential use of splice junctions).*

Finally, the translation of an mRNA in the cytoplasm can be specifically controlled. There is little evidence for the employment of this mechanism in adult somatic cells, but it occurs in some embryonic situations. This can involve localization of the mRNA to specific sites where it is expressed and/or the blocking of initiation of translation by specific protein factors (see *31.7 How are mRNAs and proteins transported and localized?).*

Regulation of tissue-specific gene transcription lies at the heart of eukaryotic differentiation; indeed, we see examples in *31 Gradients, cascades, and signaling pathways* in which proteins that regulate embryonic development prove to be transcription factors. A regulatory transcription factor serves to provide common control of a large number of target genes, and we seek to answer two questions about this mode of regulation: how does the transcription factor identify its group of target genes; and how is the activity of the transcription factor itself regulated in response to intrinsic or extrinsic signals?

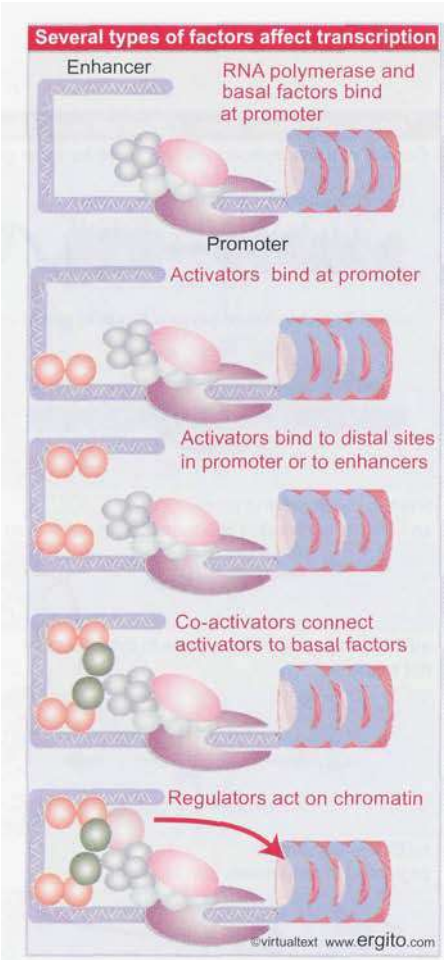


Figure 22.2 Factors involved in gene expression include RNA polymerase and the basal apparatus, activators that bind directly to DNA at the promoter or at enhancers, co-activators that bind to both activators and the basal apparatus, and regulators that act on chromatin structure.

22.2 There are several types of transcription factors

Key Concepts

- The basal apparatus determines the startpoint for transcription.
- Activators determine the frequency of transcription.
- Activators work by making protein-protein contacts with the basal factors.
- Activators may work via coactivators.
- Some components of the transcriptional apparatus work by changing chromatin structure.

Initiation of transcription involves many protein-protein interactions among transcription factors bound at the promoter or at an enhancer as well as with RNA polymerase. We can divide the factors required for transcription into several classes. **Figure 22.2** summarizes their properties:

- Basal factors, together with RNA polymerase, bind at the startpoint and TATA box (see *21.10 The basal apparatus assembles at the promoter).*
- Activators are transcription factors that recognize specific short consensus elements. They bind to sites in the promoter or in enhancers (see *21.13 Short sequence elements bind activators).* They act by increasing the efficiency with which the basal apparatus binds to the promoter. They therefore increase the frequency of transcription, and are required for a promoter to function at an adequate level. Some activators act constitutively (they are ubiquitous), but others have a regulatory role, and are synthesized or activated at specific times or in specific tissues. These factors are therefore responsible for the control of transcription patterns in time and space. The sequences that they bind are called response elements.
- Another group of factors necessary for efficient transcription do not themselves bind DNA. Coactivators provide a connection between activators and the basal apparatus (see *22.5 Activators interact with the basal apparatus).* They work by protein-protein interactions, forming bridges between activators and the basal transcription apparatus.

- Some regulators act to make changes in chromatin (see *23.7 Acetylases are associated with activators*).

The diversity of elements from which a functional promoter may be constructed, and the variations in their locations relative to the startpoint, argues that the activators have an ability to interact with one another by protein-protein interactions in multiple ways. There appear to be no constraints on the potential relationships between the elements. The modular nature of the promoter is illustrated by experiments in which equivalent regions of different promoters have been exchanged. Hybrid promoters, for example, between the thymidine kinase and β-globin genes, work well. This suggests that the main purpose of the elements is to bring the activators they bind into the vicinity of the initiation complex, where protein-protein interactions determine the efficiency of the initiation reaction.

The organization of RNA polymerase II promoters contrasts with that of bacterial promoters, where all the transcription factors must interact directly with RNA polymerase. In the eukaryotic system, only the basal factors interact directly with the enzyme. Activators may interact with the basal factors, or may interact with coactivators that in turn interact with the basal factors. The construction of the apparatus through layers of interactions explains the flexibility with which elements may be arranged, and the distance over which they can be dispersed.

22.3 Independent domains bind DNA and activate transcription

Key Concepts

- DNA-binding activity and transcription-activation are carried by independent domains of an activator.
- The role of the DNA-binding domain is to bring the transcription-activation domain into the vicinity of the promoter.

Activators and other regulatory proteins require two types of ability:

- They recognize specific target sequences located in enhancers, promoters, or other regulatory elements that affect a particular target gene.
- Having bound to DNA, an activator exercises its function by binding to other components of the transcription apparatus.

Can we characterize domains in the activator that are responsible for these activities? Often an activator has separate domains that bind DNA and activate transcription. Each domain behaves as a separate module that functions independently when it is linked to a domain of the other type. The geometry of the overall transcription complex must allow the activating domain to contact the basal apparatus irrespective of the exact location and orientation of the DNA-binding domain.

Upstream promoter elements may be an appreciable distance from the startpoint, and in many cases may be oriented in either direction. Enhancers may be even farther away and always show orientation independence. This organization has implications for both the DNA and proteins. The DNA may be looped or condensed in some way to allow the formation of the transcription complex. And the domains of the activator may be connected in a flexible way, as illustrated diagrammatically in

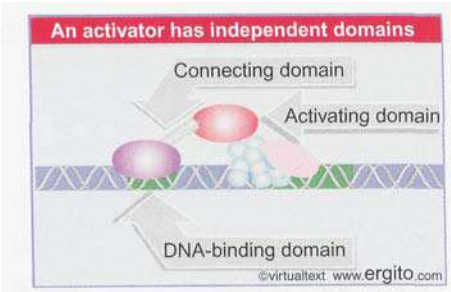


Figure 22.3 DNA-binding and activating functions in a transcription factor may comprise independent domains of the protein.

Figure 22.3. The main point here is that the DNA-binding and activating domains are independent, and connected in a way that allows the activating domain to interact with the basal apparatus irrespective of the orientation and exact location of the DNA-binding domain.

Binding to DNA is necessary for activating transcription. But does activation depend on the *particular* DNA-binding domain?

Figure 22.4 illustrates an experiment to answer this question. The activator GAL4 has a DNA-binding domain that recognizes a *UAS*, and an activating domain that stimulates initiation at the target promoter. The bacterial repressor LexA has an N-terminal DNA-binding domain that recognizes a specific operator. When LexA binds to this operator, it represses the adjacent promoter. In a "swap" experiment, the DNA-binding domain of LexA can be substituted for the DNA-binding domain of GAL4. The hybrid gene can then be introduced into yeast together with a target gene that contains either the *UAS* or a LexA operator.

An authentic GAL4 protein can activate a target gene only if it has a *UAS*. The LexA repressor by itself of course lacks the ability to activate either sort of target. The LexA-GAL4 hybrid can no longer activate a gene with a *UAS*, but it can now activate a gene that has a LexA operator!

This result fits the modular view of transcription activators. The DNA-binding domain serves to bring the protein into the right location. Precisely how or where it is bound to DNA is irrelevant, but, once it is there, the transcription-activating domain can play its role. According to this view, it does not matter whether the transcription-activating domain is brought to the vicinity of the promoter by recognition of a *UAS* via the DNA-binding domain of GAL4 or by recognition of a LexA operator via the LexA specificity module. The ability of the two types of module to function in hybrid proteins suggests that each domain of the protein folds independently into an active structure that is not influenced by the rest of the protein.

The idea that activators have independent domains that bind DNA and that activate transcription is reinforced by the ability of the tat protein of HIV to stimulate initiation without binding DNA at all. The tat protein binds to a region of secondary structure in the RNA product; the part of the RNA required for tat action is called the *tar* sequence. A model for the role of the *tat-tar* interaction in stimulating transcription is shown in **Figure 22.5.**

The *tar* sequence is located just downstream of the startpoint, so that when tat binds to *tar*, it is brought into the vicinity of the initiation complex. This is sufficient to ensure that its activation domain is in close enough proximity to the initiation complex. The activation domain inter-

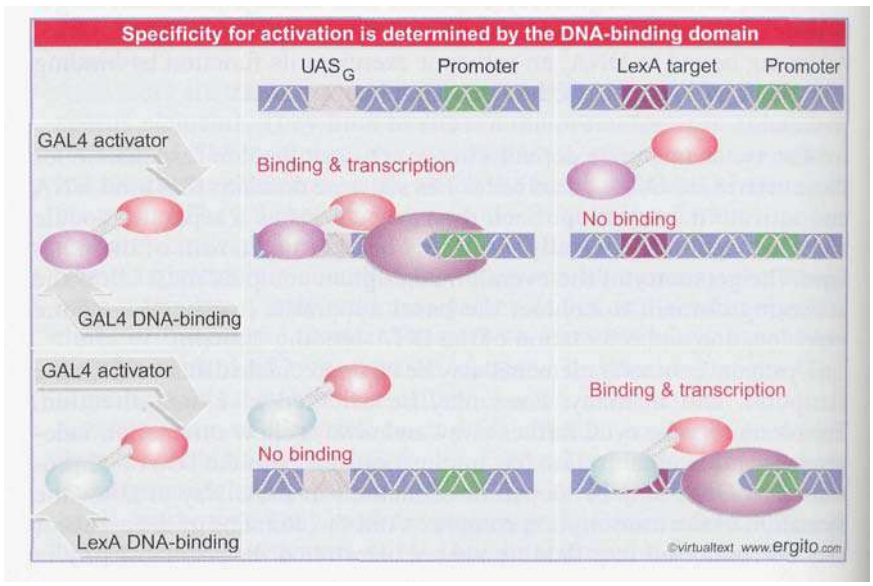


Figure 22.4 The ability of GAL4 to activate transcription is independent of its specificity for binding DNA. When the GAL4 DNA-binding domain is replaced by the LexA DNA-binding domain, the hybrid protein can activate transcription when a LexA operator is placed near a promoter.

By Book_Crazy [IND]

acts with one or more of the transcription factors bound at the complex in the same way as an activator. (Of course, the first transcript must be made in the absence of tat in order to provide the binding site.)

An extreme demonstration of the independence of the localizing and activating domains is indicated by some constructs in which tat was engineered so that the activating domain was connected to a DNA-binding domain instead of to the usual tor-binding sequence. When an appropriate target site was placed into the promoter, the tat activating-domain could activate transcription. This suggests that we should think of the DNA-binding (or in this case the RNA-binding) domain as providing a "tethering" function, *whose main purpose is to ensure that the activating domain is in the vicinity of the initiation complex.*

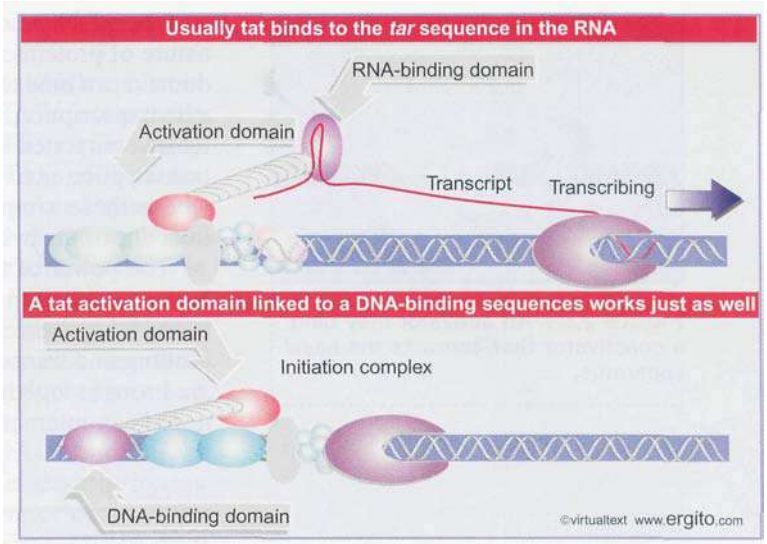


Figure 22.5 The activating domain of the tat protein of HIV can stimulate transcription if it is tethered in the vicinity by binding to the RNA product of a previous round of transcription. Activation is independent of the means of tethering, as shown by the substitution of a DNA-binding domain for the RNA-binding domain.

The notion of tethering is a more specific example of the general idea that initiation requires a high concentration of transcription factors in the vicinity of the promoter. This may be achieved when activators bind to enhancers in the general vicinity, when activators bind to upstream promoter components, or in an extreme case by tethering to the RNA product. The common requirement of all these situations is flexibility in the exact three dimensional arrangement of DNA and proteins. The principle of independent domains is common in transcriptional activators.

We might view the function of the DNA-binding domain as *bringing the activating domain into the vicinity of the startpoint.* This explains why the exact locations of DNA-binding sites can vary within the promoter.

22.4 The two hybrid assay detects protein-protein interactions

Key Concepts

- The two hybrid assay works by requiring an interaction between two proteins where one has a DNA-binding domain and the other has a transcription-activation domain.

The model of domain independence is the basis for an extremely useful assay for detecting protein interactions. In effect, we replace the connecting domain in Figure 22.3 with a protein-protein interaction. The principle is illustrated in **Figure 22.6.** We fuse one of the proteins to be tested to a DNA-binding domain. We fuse the other protein to a transcription-activating domain. (This is done by linking the appropriate coding sequences in each case and making synthetic proteins by expressing each hybrid gene.)

If the two proteins that are being tested can interact with one another, the two hybrid proteins will interact. This is reflected in the name of the technique: the two hybrid assay. The protein with the DNA-binding domain binds to a reporter gene that has a simple promoter containing its target site. But it cannot activate the gene by itself. Activation occurs only if the second hybrid binds to the first hybrid to bring the activation domain to the promoter. Any reporter gene can be used where the product is readily assayed, and this technique has given rise to several automated procedures for rapidly testing protein-protein interactions.

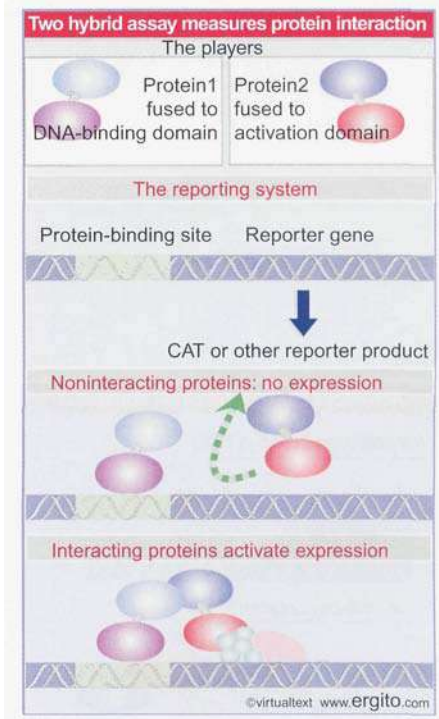


Figure 22.6 The two hybrid technique tests the ability of two proteins to interact by incorporating them into hybrid proteins where one has a DNA-binding domain and the other has a transcription-activating domain.

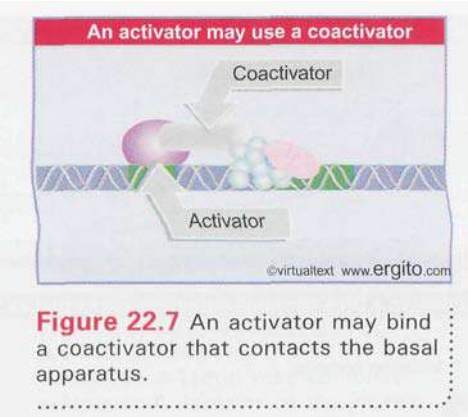


Figure 22.7 An activator may bind a coactivator that contacts the basal apparatus.

The effectiveness of the technique dramatically illustrates the modular nature of proteins. Even when fused to another protein, the DNA-binding domain can bind to DNA and the transcription-activating domain can activate transcription. Correspondingly, the interaction ability of the two proteins being tested is not inhibited by the attachment of the DNA-binding or transcription-activating domains. (Of course, there are some exceptions where these simple rules do not apply and interference between the domains of the hybrid protein prevents the technique from working.)

The power of this assay is that it requires only that the two proteins being tested can interact with each other. They need not have anything to do with transcription. Because of the independence of the DNA-binding and transcription-activating domains, all we require is that they are brought together. This will happen so long as the two proteins being tested can interact in the environment of the nucleus.

22.5 Activators interact with the basal apparatus

Key Concepts

- The principle that governs the function of all activators is that a DNA-binding domain determines specificity for the target promoter or enhancer.
- The DNA-binding domain is responsible for localizing a transcription-activating domain in the proximity of the basal apparatus.
- An activator that works directly has a DNA-binding domain and an activating domain.
- An activator that does not have an activating domain may work by binding a coactivator that has an activating domain.
- Several factors in the basal apparatus are targets with which activators or coactivators interact.
- RNA polymerase may be associated with various alternative sets of transcription factors in the form of a holoenzyme complex.

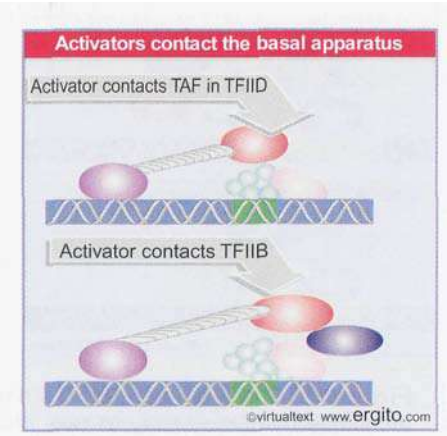


Figure 22.8 Activators may work at different stages of initiation, by contacting the TAFs of TFIID or contacting TFIIB.

An activator may work directly when it consists of a DNA-binding domain linked to a transcription-activating domain, as illustrated in Figure 22.3. In other cases, the activator does not itself have a transcription-activating domain, but binds another protein—a coactivator—that has the transcription-activating domain. **Figure 22.7** shows the action of such an activator. We may regard coactivators as transcription factors whose specificity is conferred by the ability to bind to DNA-binding transcription factors instead of directly to DNA. A particular activator may require a specific coactivator.

But although the protein components are organized differently, the mechanism is the same. An activator that contacts the basal apparatus directly has an activation domain covalently connected to the DNA-binding domain. When an activator works through a coactivator, the connections involve noncovalent binding between protein subunits (compare Figure 22.3 and Figure 22.7). The same interactions are responsible for activation, irrespective of whether the various domains are present in the same protein subunit or divided into multiple protein subunits.

A transcription-activating domain works by making protein-protein contacts with general transcription factors that promote assembly of the basal apparatus. Contact with the basal apparatus may be made with any one of several basal factors, typically $TF_{II}D$, $TF_{II}B$, or $TF_{II}A$. All of these factors participate in early stages of assembly of the basal apparatus (see Figure 21.14). **Figure 22.8** illustrates the situation when such a

contact is made. The major effect of the activators is to influence the assembly of the basal apparatus.

$TF_{II}D$ may be the most common target for activators, which may contact any one of several TAFs. In fact, a major role of the TAFs is to provide the connection from the basal apparatus to activators. This explains why TBP alone can support basal-level transcription, but the TAFs of $TF_{II}D$ are required for the higher levels of transcription that are stimulated by activators. Different TAFs in $TF_{II}D$ may provide surfaces that interact with different activators. Some activators interact only with individual TAFs; others interact with multiple TAFs. We assume that the interaction either assists binding of $TF_{II}D$ to the TATA box or assists the binding of other activators around the $TF_{II}D$-TATA box complex. In either case, the interaction stabilizes the basal transcription complex; this speeds the process of initiation, and thereby increases use of the promoter.

The activating domains of the yeast activators GAL4 and GCN4 have multiple negative charges, giving rise to their description as "acidic activators." Another particularly effective activator of this type is carried by the VP16 protein of the Herpes Simplex Virus. (VP16 does not itself have a DNA-binding domain, but interacts with the transcription apparatus via an intermediary protein.) Experiments to characterize acidic activator functions have often made use of the VP16 activating region linked to a DNA-binding motif.

Acidic activators function by enhancing the ability of $TF_{II}B$ to join the basal initiation complex. Experiments *in vitro* show that binding of $TF_{II}B$ to an initiation complex at an adenovirus promoter is stimulated by the presence of GAL4 or VP16 acid activators; and the VP16 activator can bind directly to $TF_{II}B$. Assembly of $TF_{II}B$ into the complex at this promoter is therefore a rate-limiting step that is stimulated by the presence of an acidic activator.

The resilience of an RNA polymerase II promoter to the rearrangement of elements, and its indifference even to the particular elements present, suggests that the events by which it is activated are relatively general in nature. Any activators whose activating region is brought within range of the basal initiation complex may be able to stimulate its formation. Some striking illustrations of such versatility have been accomplished by constructing promoters consisting of new combinations of elements. For example, when a yeast UAS_G element is inserted near the promoter of a higher eukaryotic gene, this gene can be activated by GAL4 in a mammalian cultured cell. Whatever means GAL4 uses to activate the promoter seems therefore to have been conserved between yeast and higher eukaryotes. The GAL4 protein must recognize some feature of the mammalian transcription apparatus that resembles its normal contacts in yeast.

How does an activator stimulate transcription? We can imagine two general types of model:

- The recruitment model argues that its sole effect is to increase the binding of RNA polymerase to the promoter.
- An alternative model is to suppose that it induces some change in the transcriptional complex, for example, in the conformation of the enzyme, which increases its efficiency.

A test of these models in one case in yeast showed that recruitment can account for activation. When the concentration of RNA polymerase was increased sufficiently, the activator failed to produce any increase in transcription, suggesting that its sole effect is to increase the effective concentration of RNA polymerase at the promoter.

Adding up all the components required for efficient transcription—basal factors, RNA polymerase, activators, coactivators—we get a very large apparatus, consisting of >40 proteins. Is it feasible for this apparatus to assemble step by step at the promoter? Some activators, coactivators, and basal factors may assemble stepwise at the promoter, but

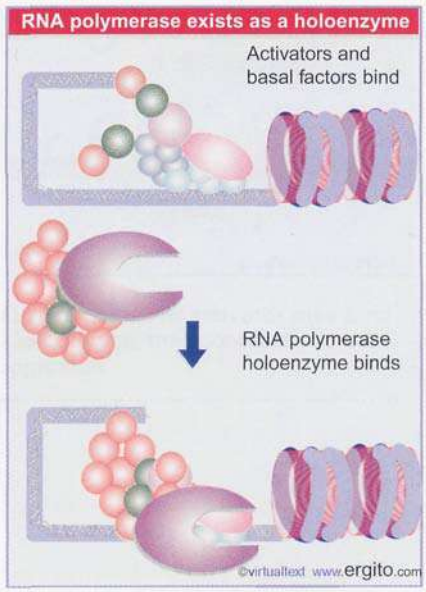


Figure 22.9 RNA polymerase exists as a holoenzyme containing many activators.

then may be joined by a very large complex consisting of RNA polymerase preassembled with further activators and coactivators, as illustrated in **Figure 22.9.**

Several forms of RNA polymerase have been found in which the enzyme is associated with various transcription factors. The most prominent "holoenzyme complex" in yeast (defined as being capable of initiating transcription without additional components) consists of RNA polymerase associated with a 20-subunit complex called mediator. The mediator includes products of several genes in which mutations block transcription, including some *SRB* loci (so named because many of their genes were originally identified as suppressors of mutations in RNA polymerase *B.)* The name was suggested by its ability to mediate the effects of activators. Mediator is necessary for transcription of most yeast genes. Homologous complexes are required for the transcription of most higher eukaryotic genes. Mediator undergoes a conformational change when it interacts with the CTD domain of RNA polymerase. It can transmit either activating or repressing effects from upstream components to the RNA polymerase. It is probably released when a polymerase starts elongation. Some transcription factors influence transcription directly by interacting with RNA polymerase or the basal apparatus, but others work by manipulating structure of chromatin (see *23.3 Chromatin remodeling is an active process).*

22.6 Some promoter-binding proteins are repressors

Key Concepts

- Repression is usually achieved by affecting chromatin structure, but there are repressors that act by binding to specific promoters.

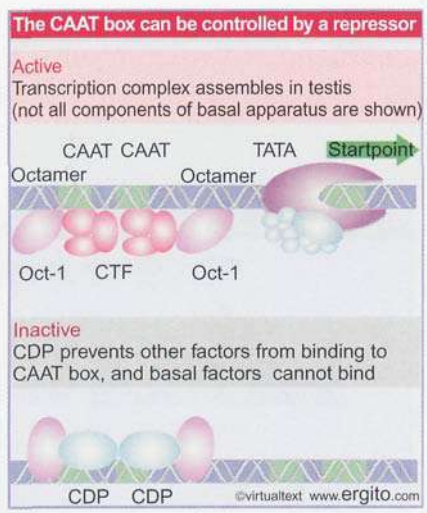


Figure 22.10 A transcription complex involves recognition of several elements in the sea urchin H2B promoter in testis. Binding of the CAAT displacement factor in embryo prevents the CAAT-binding factor from binding, so an active complex cannot form.

Repression of transcription in eukaryotes is generally accomplished at the level of influencing chromatin structure; regulator proteins that function like *trans-acting* bacterial repressors to block transcription are relatively rare, but some examples are known. One case is the global repressor NC2/Drl/DRAPl, a heterodimer that binds to TBP to prevent it from interacting with other components of the basal apparatus. The importance of this interaction is suggested by the lethality of null mutations in the genes that code for the repressor in yeast. Repressors that work in this way have an active role in inhibiting basal apparatus function.

In a more specific case, the CAAT sequence is a target for regulation. Two copies of this element are found in the promoter of a gene for histone H2B (see Figure 21.21) that is expressed only during spermatogenesis in a sea urchin. CAAT-binding factors can be extracted from testis tissue and also from embryonic tissues, but only the former can bind to the CAAT box. In the embryonic tissues, another protein, called the CAAT-displacement protein (CDP), binds to the CAAT boxes, *preventing the activator from recognizing them.*

Figure 22.10 illustrates the consequences for gene expression. In testis, the promoter is bound by transcription factors at the TATA box, CAAT boxes, and octamer sequences. In embryonic tissue, the exclusion of the CAAT-binding factor from the promoter prevents a transcription complex from being assembled. The analogy with the effect of a bacterial repressor in preventing RNA polymerase from initiating at the promoter

is obvious. These results also make the point that the function of a protein in binding to a known promoter element cannot be assumed: it may be an activator, a repressor, or even irrelevant to gene transcription.

22.7 Response elements are recognized by activators

Key Concepts

- Response elements may be located in promoters or enhancers.
- Each response element is recognized by a specific activator.
- A promoter may have many response elements, which may activate transcription independently or in certain combinations.

The principle that emerges from characterizing groups of genes under common control is that *they share a promoter (or enhancer) element that is recognized by an activator.* An element that causes a gene to respond to such a factor is called a response element; examples are the HSE (heat shock response element), GRE (glucocorticoid response element), SRE (serum response element). Response elements contain short consensus sequences; copies of the response elements found in different genes are closely related, but not necessarily identical. The region bound by the factor extends for a short distance on either side of the consensus sequence. In promoters, the elements are not present at fixed distances from the startpoint, but are usually <200 bp upstream of it. The presence of a single element usually is sufficient to confer the regulatory response, but sometimes there are multiple copies.

Response elements may be located in promoters or in enhancers. Some types of elements are typically found in one rather than the other: usually an HSE is found in a promoter, while a GRE is found in an enhancer. We assume that all response elements function by the same general principle. Binding of an activator to the response element is required to allow RNA polymerase to initiate transcription. The difference from constitutively active activators is that the protein is either available or is active only under certain conditions, which determine when the gene is to be expressed.

An example of a situation in which many genes are controlled by a single factor is provided by the heat shock response. This is common to a wide range of prokaryotes and eukaryotes and involves multiple controls of gene expression: an increase in temperature turns off transcription of some genes, turns on transcription of the heat shock genes, and causes changes in the translation of mRNAs. The control of the heat shock genes illustrates the differences between prokaryotic and eukaryotic modes of control. In bacteria, a new sigma factor is synthesized that directs RNA polymerase holoenzyme to recognize an alternative –10 sequence common to the promoters of heat shock genes (see *9.16 Substitution of sigma factors may control initiation*). In eukaryotes, the heat shock genes also possess a common consensus sequence (HSE), but it is located at various positions relative to the startpoint, and is recognized by an independent activator, HSTF. The activation of this factor therefore provides a means to initiate transcription at the specific group of ~20 genes that contains the appropriate target sequence at its promoter.

All the heat shock genes of *D. melanogaster* contain multiple copies of the HSE. The HSTF binds cooperatively to adjacent response elements. Both the HSE and HSTF have been conserved in evolution, and it is striking that a heat shock gene from *D. melanogaster* can be activated in species as distant as mammals or sea urchins. The HSTF proteins of fruit fly and yeast appear similar, and show the same footprint pattern on DNA containing

HSE sequences. Yeast HSTF becomes phosphorylated when cells are heat-shocked; this modification is responsible for activating the protein.

The metallothionein (MT) gene provides an example of how a single gene may be regulated by many different circuits. The metallothionein protein protects the cell against excess concentrations of heavy metals, by binding the metal and removing it from the cell. The gene is expressed at a basal level, but is induced to greater levels of expression by heavy metal ions (such as cadmium) or by glucocorticoids. The control region combines several different kinds of regulatory elements.

The organization of the promoter for a MT gene is summarized in Figure 22.11. A major feature of this map is the high density of elements that can activate transcription. The TATA and GC boxes are located at their usual positions fairly close to the startpoint. Also needed for the basal level of expression are the two basal level elements (BLE), which fit the formal description of enhancers. Although located near the startpoint, they can be moved elsewhere without loss of effect. They contain sequences related to those found in other enhancers, and are bound by proteins that bind the SV40 enhancer.

The TRE is a consensus sequence that is present in several enhancers, including one BLE of metallothionein and the 72 bp repeats of the virus SV40. The TRE has a binding site for factor AP1; this interaction is part of the mechanism for constitutive expression, for which AP1 is an activator. However, AP1 binding also has a second function. The TRE confers a response to phorbol esters such as TPA (an agent that promotes tumors), and this response is mediated by the interaction of AP1 with the TRE. This binding reaction is one (not necessarily the sole) means by which phorbol esters trigger a series of transcriptional changes.

The inductive response to metals is conferred by the multiple MRE sequences, which function as promoter elements. The presence of one MRE confers the ability to respond to heavy metal; a greater level of induction is achieved by the inclusion of multiple elements. The factor MTF1 binds to the MRE in response to the presence of metal ions.

The response to steroid hormones is governed by a GRE, located 250 bp upstream of the startpoint, which behaves as an enhancer. Deletion of this region does not affect the basal level of expression or the level induced by metal ions. But it is absolutely needed for the response to steroids.

The regulation of metallothionein illustrates the general principle that *any one of several different elements, located in either an enhancer or promoter, can independently activate the gene.* The absence of an element needed for one mode of activation does not affect activation in other modes. The variety of elements, their independence of action, and the apparently unlimited flexibility of their relative arrangements, suggest that a factor binding to any one element is able independently to increase the efficiency of initiation by the basal transcription apparatus, probably by virtue of protein-protein interactions that stabilize or otherwise assist formation of the initiation complex.

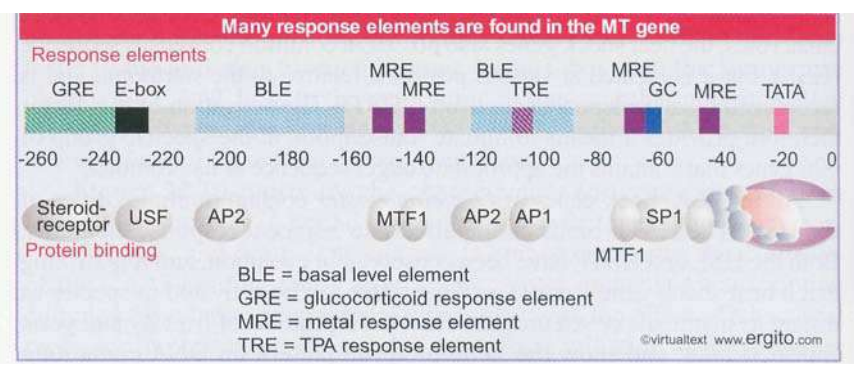


Figure 22.11 The regulatory region of a human metallothionein gene contains regulator elements in both its promoter and enhancer. The promoter has elements for metal induction; an enhancer has an element for response to glucocorticoid. Promoter elements are shown above the map, and proteins that bind them are indicated below.

By Book_Crazy [IND]

22.8 There are many types of DNA-binding domains

Key Concepts

- Activators are classified according to the type of DNA-binding domain.
- Members of the same group have sequence variations of a specific motif that confer specificity for individual target sites.

It is common for an activator to have a modular structure in which different domains are responsible for binding to DNA and for activating transcription. Factors are often classified according to the type of DNA-binding domain. Typically a relatively short motif in this domain is responsible for binding to DNA:

- The zinc finger motif comprises a DNA-binding domain. It was originally recognized in factor $TF_{III}A$, which is required for RNA polymerase III to transcribe 5S rRNA genes. It has since been identified in several other transcription factors (and presumed transcription factors). A distinct form of the motif is found also in the steroid receptors.
- The steroid receptors are defined as a group by a functional relationship: each receptor is activated by binding a particular steroid. The glucocorticoid receptor is the most fully analyzed. Together with other receptors, such as the thyroid hormone receptor or the retinoic acid receptor, the steroid receptors are members of the superfamily of ligand-activated activators with the same general *modus operandi: the protein factor is inactive until it binds a small ligand.*
- The helix-turn-helix motif was originally identified as the DNA-binding domain of phage repressors. One α-helix lies in the major groove of DNA; the other lies at an angle across DNA. A related form of the motif is present in the homeodomain, a sequence first characterized in several proteins coded by genes concerned with developmental regulation in *Drosophila*. It is also present in genes for mammalian transcription factors.
- The amphipathic helix-loop-helix (HLH) motif has been identified in some developmental regulators and in genes coding for eukaryotic DNA-binding proteins. Each amphipathic helix presents a face of hydrophobic residues on one side and charged residues on the other side. The length of the connecting loop varies from 12-28 amino acids. The motif enables proteins to dimerize, and a basic region near this motif contacts DNA.
- Leucine zippers consist of a stretch of amino acids with a leucine residue in every seventh position. A leucine zipper in one polypeptide interacts with a zipper in another polypeptide to form a dimer. Adjacent to each zipper is a stretch of positively charged residues that is involved in binding to DNA.

The activity of an inducible activator may be regulated in any one of several ways, as illustrated schematically in **Figure 22.12:**

- A factor is tissue-specific because it is synthesized only in a particular type of cell. This is typical of factors that regulate development, such as homeodomain proteins.
- The activity of a factor may be directly controlled by modification. HSTF is converted to the active form by phosphorylation. AP1 (a heterodimer between the subunits Jun and Fos) is converted to the active form by phosphorylating the Jun subunit.
- A factor is activated or inactivated by binding a ligand. The steroid receptors are prime examples. Ligand binding may influence the

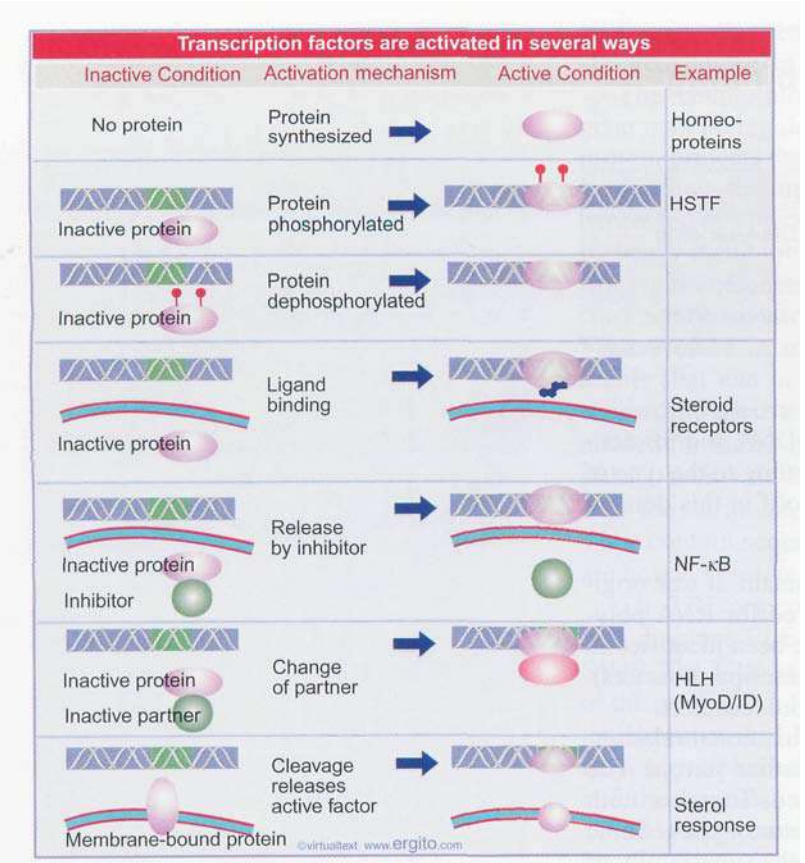


Figure 22.12 The activity of a regulatory transcription factor may be controlled by synthesis of protein, covalent modification of protein, ligand binding, or binding of inhibitors that sequester the protein or affect its ability to bind to DNA.

localization of the protein (causing transport from cytoplasm to nucleus), as well as determining its ability to bind to DNA.

Availability of a factor may vary; for example, the factor NF-KB (which activates immunoglobulin K genes in B lymphocytes) is present in many cell types. But it is sequestered in the cytoplasm by the inhibitory protein I-κB. In B lymphocytes, NF-KB is released from I-κB and moves to the nucleus, where it activates transcription.

An extreme example of control of availability is found when a factor is actually part of a cytoplasmic structure, and is released from that structure to translocate to the nucleus.

A dimeric factor may have alternative partners. One partner may cause it to be inactive; synthesis of the active partner may displace the inactive partner. Such situations may be amplified into networks in which various alternative partners pair with one another, especially among the HLH proteins.

The factor may be cleaved from an inactive precursor. One activator is produced as a protein bound to the nuclear envelope and endoplasmic reticulum. The absence of sterols (such as cholesterol) causes the cytosolic domain to be cleaved; it then translocates to the nucleus and provides the active form of the activator.

(We note *en passant* that mutations of the transcription factors in some of these classes give rise to factors that inappropriately activate, or prevent activation, of transcription; their roles in generating tumors are discussed in *30.18 Oncoproteins may regulate gene expression,* and Figure 30.26 should be compared with Figure 22.12.)

We now discuss in more detail the DNA-binding and activation reactions that are sponsored by some of these classes of proteins.

22.9 A zinc finger motif is a DNA-binding domain

Key Concepts

- A zinc finger is a loop of ~ 23 amino acids that protrudes from a zinc-binding site formed by His and Cys amino acids.
- A zinc finger protein usually has multiple zinc fingers.
- The C-terminal part of each finger forms an α-helix that bind one turn of the major groove of DNA.
- Some zinc finger proteins bind RNA instead of or as well as DNA.

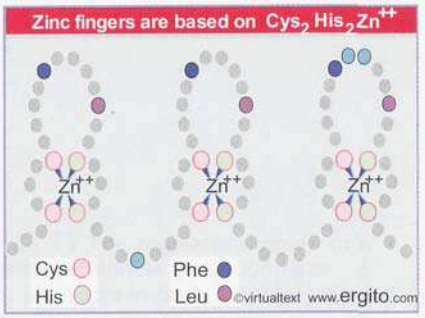


Figure 22.13 Transcription factor SP1 has a series of three zinc fingers, each with a characteristic pattern of cysteine and histidine residues that constitute the zinc-binding site.

Zinc fingers take their name from the structure illustrated in Figure 22.13, in which a small group of conserved amino acids binds a zinc ion to form an independent domain in the protein. Two types of DNA-binding proteins have structures of this type: the classic "zinc finger" proteins; and the steroid receptors.

A "finger protein" typically has a series of zinc fingers, as depicted in the figure. The consensus sequence of a single finger is:

$$Cys\text{-}X_{2\text{-}4}\text{-}Cys\text{-}X_3\text{-}Phe\text{-}X_5\text{-}Leu\text{-}X_2\text{-}His\text{-}X_3\text{-}His$$

The motif takes its name from the loop of amino acids that protrudes from the zinc-binding site and is described as the Cys_2/His_2 finger. The zinc is held in a tetrahedral structure formed by the conserved Cys and His residues. The finger itself comprises ~23 amino acids, and the linker between fingers is usually 7-8 amino acids.

Zinc fingers are a common motif in DNA-binding proteins. The fingers usually are organized as a single series of tandem repeats; occasionally there is more than one group of fingers. The stretch of fingers ranges from 9 repeats that occupy almost the entire protein (as in $TF_{III}A$) to providing just one small domain consisting of 2 fingers (as in the *Drosophila* regulator ADR1). The activator Sp1 has a DNA-binding domain that consists of 3 zinc fingers.

The crystal structure of DNA bound by a protein with three fingers suggests the structure illustrated schematically in **Figure 22.14.** The C-terminal part of each finger forms α-helices that bind DNA; the N-terminal part forms a β-sheet. (For simplicity, the β-sheet and the location of the zinc ion are not shown in the lower part of the figure.) The three α-helical stretches fit into one turn of the major groove; each α-helix (and thus each finger) makes two sequence-specific contacts with DNA (indicated by the arrows). We expect that the nonconserved amino acids in the C-terminal side of each finger are responsible for recognizing specific target sites.

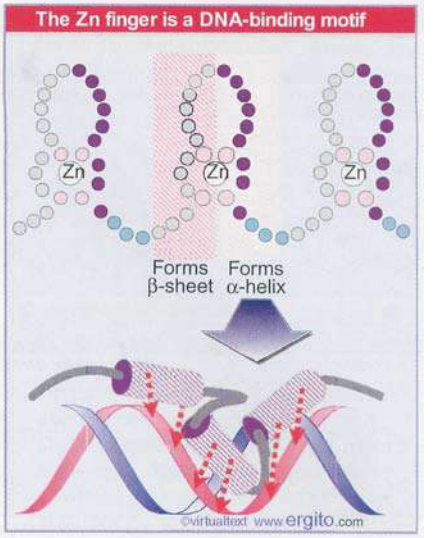


Figure 22.14 Zinc fingers may form α-helices that insert into the major groove, associated with β-sheets on the other side.

Knowing that zinc fingers are found in authentic activators that assist both RNA polymerases II and III, we may view finger proteins from the reverse perspective. When a protein is found to have multiple zinc fingers, there is at least a *prima facie* case for investigating a possible role as a transcription factor. Such an identification has suggested that several loci involved in embryonic development of *D. melanogaster* are regulators of transcription.

However, it is necessary to be cautious about interpreting the presence of (putative) zinc fingers, especially when the protein contains only a single finger motif. Fingers may be involved in binding RNA rather than DNA or even unconnected with any nucleic acid binding activity. For example, the prototype zinc finger protein, $TF_{III}A$, binds both to the 5S gene and to the product, 5S rRNA. A translation initiation factor, eIF2β has a zinc finger; and mutations in the finger influence the recognition of initiation codons. Retroviral capsid proteins have a motif related to the finger that may be involved in binding the viral RNA.

22.10 Steroid receptors are activators

Key Concepts

- Steroid receptors are examples of ligand-responsive activators that are activated by binding a steroid (or other related molecules).
- There are separate DNA-binding and ligand-binding domains.

Steroid hormones are synthesized in response to a variety of neuroendocrine activities, and exert major effects on growth, tissue development, and body homeostasis in the animal world. The major groups of steroids and some other compounds with related (molecular) activities are classified in **Figure 22.15.**

The adrenal gland secretes >30 steroids, the two major groups being the glucocorticoids and mineralocorticoids. Steroids provide the reproductive hormones (androgen male sex hormones and estrogen female sex hormones). Vitamin D is required for bone development.

Other hormones, with unrelated structures and physiological purposes, function at the molecular level in a similar way to the steroid hormones. Thyroid hormones, based on iodinated forms of tyrosine,

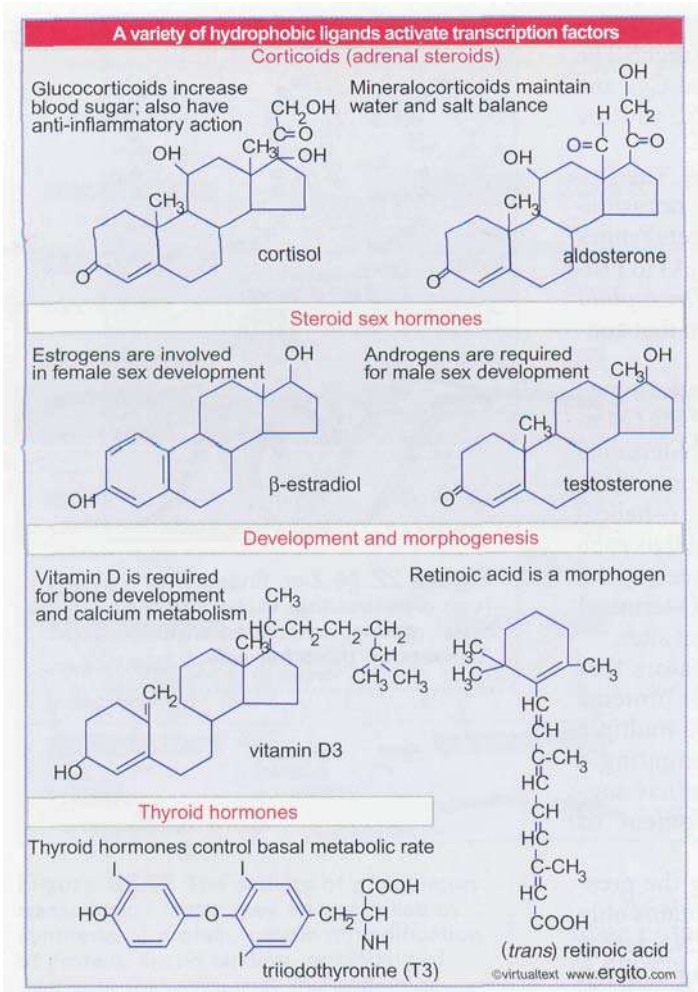


Figure 22.15 Several types of hydrophobic small molecules activate transcription factors.

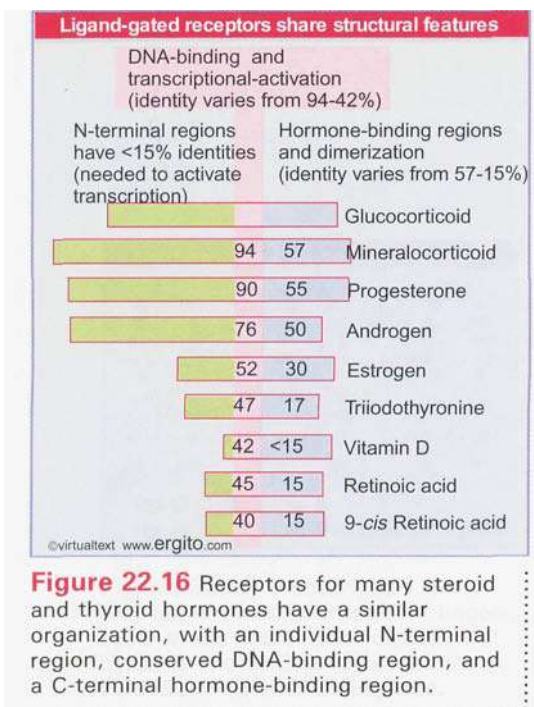


Figure 22.16 Receptors for many steroid and thyroid hormones have a similar organization, with an individual N-terminal region, conserved DNA-binding region, and a C-terminal hormone-binding region.

control basal metabolic rate in animals. Steroid and thyroid hormones also may be important in metamorphosis (ecdysteroids in insects, and thyroid hormones in frogs).

Retinoic acid (vitamin A) is a morphogen responsible for development of the anterior-posterior axis in the developing chick limb bud. Its metabolite, *9-cis* retinoic acid, is found in tissues that are major sites for storage and metabolism of vitamin A.

We may account for these various actions in terms of pathways for regulating gene expression. These diverse compounds share a common mode of action: *each is a small molecule that binds to a specific receptor that activates gene transcription.* ("Receptor" may be a misnomer: the protein is a receptor for steroid or thyroid hormone in the same sense that *lac* repressor is a receptor for a β-galactoside: it is not a receptor in the sense of comprising a membrane-bound protein that is exposed to the cell surface.)

Receptors for the diverse groups of steroid hormones, thyroid hormones, and retinoic acid represent a new "superfamily" of gene regulators, the ligand-responsive activators. All the receptors have independent domains for DNA-binding and hormone binding, in the same relative locations. Their general organization is summarized in **Figure 22.16.**

The central part of the protein is the DNA-binding domain. These regions are closely related for the various steroid receptors (from the most closely related pair with 94% sequence identity to the least well related pair at 42% identity). The act of binding DNA cannot be disconnected from the ability to activate transcription, because mutations in this domain affect both activities.

The N-terminal regions of the receptors show the least conservation of sequence. They include other regions that are needed to activate transcription.

The C-terminal domains bind the hormones. Those in the steroid receptor family show identities ranging from 30-57%, reflecting specificity for individual hormones. Their relationships with the other receptors are minimal, reflecting specificity for a variety of compounds—thyroid hormones, vitamin D, retinoic acid, etc. This domain also has the motifs responsible for dimerization and a region involved in transcriptional activation.

Some ligands have multiple receptors that are closely related, such as the 3 retinoic acid receptors (RARα, β, 7) and the three receptors for 9-*cis*-retinoic acid (RXRα, β, 7).

22.11 Steroid receptors have zinc fingers

Key Concepts

- The DNA binding domain of a steroid receptor is a type of zinc finger that has Cys but not His residues.
- Glucocorticoid and estrogen receptors each have two zinc fingers, the first of which determines the DNA target sequence.
- Steroid receptors bind to DNA as dimers.

Steroid receptors (and some other proteins) have another type of zinc finger that is different from Cys_2/His_2 fingers. The structure is based on a sequence with the zinc-binding consensus:

$Cys\text{-}X_2\text{-}Cys\text{-}X_{13}\text{-}Cys\text{-}X_2\text{-}Cys$

These are called Cys_2/Cys_2 fingers. Proteins with Cys_2/Cys_2 fingers often have nonrepetitive fingers, in contrast with the tandem repetition of the Cys_2/His_2 type. Binding sites in DNA (where known) are short and palindromic.

The glucocorticoid and estrogen receptors each have two fingers, each with a zinc atom at the center of a tetrahedron of cysteines. The two fingers form α-helices that fold together to form a large globular domain. The aromatic sides of the α-helices form a hydrophobic center together with a β-sheet that connects the two helices. One side of the N-terminal helix makes contacts in the major groove of DNA. Two glucocorticoid receptors dimerize upon binding to DNA, and each engages a successive turn of the major groove. This fits with the palindromic nature of the response element (see *22.13 Steroid receptors recognize response elements by a combinatorial code)*.

Each finger controls one important property of the receptor. Figure 22.17 identifies the relevant amino acids. Those on the right side of the first finger determine the sequence of the target in DNA; those on the left side of the second finger control the spacing between the target sites recognized by each subunit in the dimer (see *22.13 Steroid receptors recognize response elements by a combinatorial code)*.

Direct evidence that the first finger binds DNA was obtained by a "specificity swap" experiment. The finger of the estrogen receptor was deleted and replaced by the sequence of the glucocorticoid receptor. The new protein recognized the GRE sequence (the usual target of the glucocorticoid receptor) instead of the ERE (the usual target of the estrogen receptor). This region therefore establishes the specificity with which DNA is recognized.

The differences between the sequences of the glucocorticoid receptor and estrogen receptor fingers lie mostly at the base of the finger. The substitution at two positions shown in Figure 22.18 allows the glucocorticoid receptor to bind at an ERE instead of a GRE.

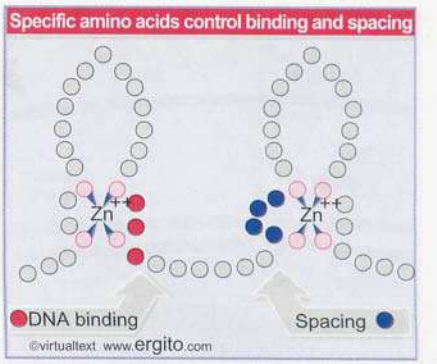


Figure 22.17 The first finger of a steroid receptor controls which DNA sequence is bound (positions shown in red); the second finger controls spacing between the sequences (positions shown in blue).

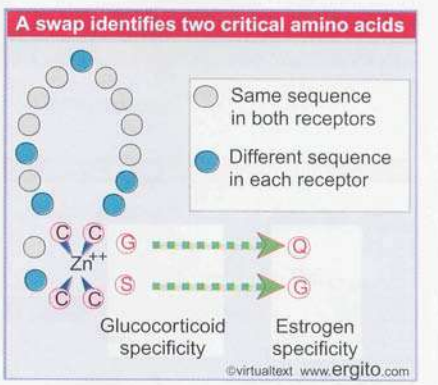


Figure 22.18 Discrimination between GRE and ERE target sequences is determined by two amino acids at the base of the first zinc finger in the receptor.

22.12 Binding to the response element is activated by ligand-binding

Key Concepts

* Binding of ligand to the C-terminal domain increases the affinity of the DNA-binding domain for its specific target site in DNA.

We know most about the interaction of glucocorticoids with their receptor, whose action is illustrated in Figure 22.19. A steroid hormone can pass through the cell membrane to enter the cell by simple diffusion. Within the cell, a glucocorticoid binds the glucocorticoid receptor. (Work on the glucocorticoid receptor has relied on the synthetic steroid hormone, dexamethasone.) The localization of free receptors is not entirely clear; they may be in equilibrium between the nucleus and cytoplasm. But when hormone binds to the receptor, the protein is converted into an activated form that has an increased affinity for DNA, so the hormone-receptor complex is always localized in the nucleus.

The activated receptor recognizes a specific consensus sequence that identifies the GRE, the glucocorticoid response element. The GRE is typically located in an enhancer that may be several kb upstream or downstream of the promoter. When the steroid-receptor complex binds to the enhancer, the nearby promoter is activated, and transcription

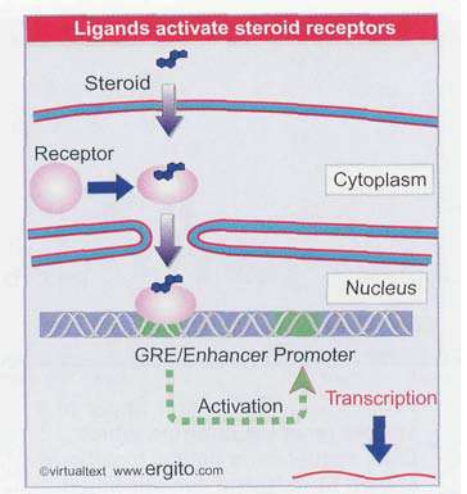


Figure 22.19 Glucocorticoids regulate gene transcription by causing their receptor to bind to an enhancer whose action is needed for promoter function.

initiates there. Enhancer activation provides the general mechanism by which steroids regulate a wide set of target genes.

The C-terminal region regulates the activity of the receptor in a way that varies for the individual receptor. If the C-terminal domain of the glucocorticoid receptor is deleted, the remaining N-terminal protein is constitutively active: it no longer requires steroids for activity. This suggests that, in the absence of steroid, the steroid-binding domain prevents the receptor from recognizing the GRE; it functions as an internal negative regulator. The addition of steroid inactivates the inhibition, releasing the receptor's ability to bind the GRE and activate transcription. The basis for the repression could be internal, relying on interactions with another part of the receptor. Or it could result from an interaction with some other protein, which is displaced when steroid binds.

The interaction between the domains is different in the estrogen receptor. If the hormone-binding domain is deleted, the protein is unable to activate transcription, although it continues to bind to the ERE. This region is therefore required to activate rather than to repress activity.

22.13 Steroid receptors recognize response elements by a combinatorial code

Key Concepts

- A steroid response element consists of two short half sites that may be palindromic or directly repeated.
- There are only two types of half sites.
- A receptor recognizes its response element by the orientation and spacing of the half sites.
- The sequence of the half site is recognized by the first zinc finger.
- The second zinc finger is responsible for dimerization, which determines the distance between the subunits.

Subunit separation in the receptor determines the recognition of spacing in the response element.

Some steroid receptors function as homodimers but others form heterodimers.

Homodimers recognize palindromic response elements; heterodimers recognize response elements with directly repeated half sites.

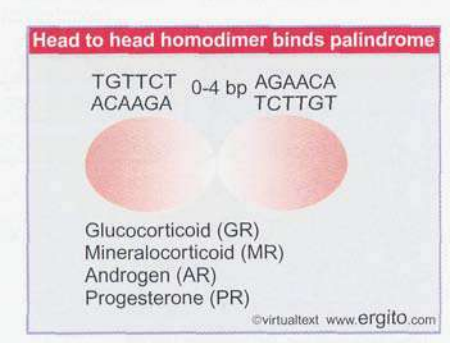


Figure 22.20 Response elements formed from the palindromic half site TGTTCT are recognized by several different receptors depending on the spacing between the half sites.

Each receptor recognizes a response element that consists of two short repeats (or half sites). This immediately suggests that the receptor binds as a dimer, so that each half of the consensus is contacted by one subunit (reminiscent of the λ operator-repressor interaction described in *12.12 Repressor uses a helix-turn-helix motif to bind DNA*).

The half sites may be arranged either as palindromes or as repeats in the same orientation. They are separated by 0-4 base pairs whose sequence is irrelevant. Only two types of half site are used by the various receptors. Their orientation and spacing determine which receptor recognizes the response element. This behavior allows response elements that have restricted consensus sequences to be recognized specifically by a variety of receptors. The rules that govern recognition are not absolute, but may be modified by context, and there are also cases in which palindromic response elements are recognized permissively by more than one receptor.

The receptors fall into two groups:

- Glucocorticoid (GR), mineralocorticoid (MR), androgen (AR), and progesterone (PR) receptors all form homodimers. They recognize response elements whose half sites have the consensus sequence TGTTCT. **Figure 22.20** shows that the half sites are arranged as

palindromes, and the spacing between the sites determines the type of element. The estrogen (ER) receptor functions in the same way, but has the half site sequence TGACCT.

- The 9-*cis*-retinoic acid (RXR) receptor forms homodimers and also forms heterodimers with ~15 other receptors, including thyroid (T3R), vitamin D (VDR), and retinoic acid (RAR). **Figure 22.21** shows that the dimers recognize half elements with the sequence TGACCT. The half sites are arranged as direct repeats, and recognition is controlled by spacing between them. Some of the heterodimeric receptors are activated when the ligand binds to the partner for RXR; others can be activated by ligand binding either to this subunit or to the RXR subunit. These receptors can also form homodimers, which recognize palindromic sequences.

Now we are in a position to understand the basis for specificity of recognition. Recall that Figure 22.17 shows how recognition of the sequence of the half site is conferred by the amino acid sequence in the first finger. Specificity for the spacing between half sites is carried by amino acids in the second finger. The structure of the dimer determines the distance between the subunits that sit in successive turns of the major groove, and thus controls the response to the spacing of half sites. The exact positions of the residues responsible for dimerization differ in individual pairwise combinations.

How do the steroid receptors activate transcription? They do not act directly on the basal apparatus, but function via a coactivating complex. The coactivator includes various activities, including the common component CBP/p300, one of whose functions is to modify the structure of chromatin by acetylating histones (see Figure 23.13).

All receptors in the superfamily are ligand-dependent activators of transcription. However, some are also able to repress transcription. The TR and RAR receptors, in the form of heterodimers with RXR, bind to certain loci in the *absence* of ligand and repress transcription by means of their ability to interact with a corepressor protein. The corepressor functions by the reverse of the mechanism used by coactivators: it inhibits the function of the basal transcription apparatus, one of its actions being the deacetylation of histones (see Figure 23.15). We do not know the relative importance of the repressor activity *vis-à-vis* the ligand-dependent activation in the physiological response to hormone.

The effect of ligand binding on the receptor is to convert it from a repressing complex to an activating complex, as shown in **Figure 22.22.** In the absence of ligand, the receptor is bound to a corepressor complex. The component of the corepressor that binds to the receptor is SMRT. Binding of ligand causes a conformational change that displaces SMRT. This allows the coactivator to bind.

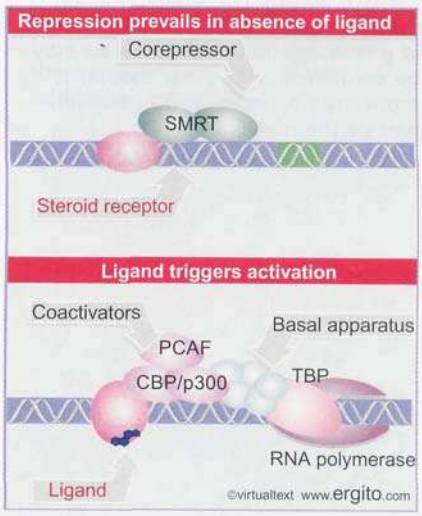


Figure 22.22 TR and RAR bind the SMRT corepressor in the absence of ligand. The promoter is not expressed. When SMRT is displaced by binding of ligand, the receptor binds a coactivator complex. This leads to activation of transcription by the basal apparatus.

22.14 Homeodomains bind related targets in DNA

Key Concepts

- The homeodomain is a DNA-binding domain of 60 amino acids that has three α-helices.
- The C-terminal α-helix-3 is 17 amino acids and binds in the major groove of DNA.
- The N-terminal arm of the homeodomain projects into the minor groove of DNA.
- Proteins containing homeodomains may be either activators or repressors of transcription.

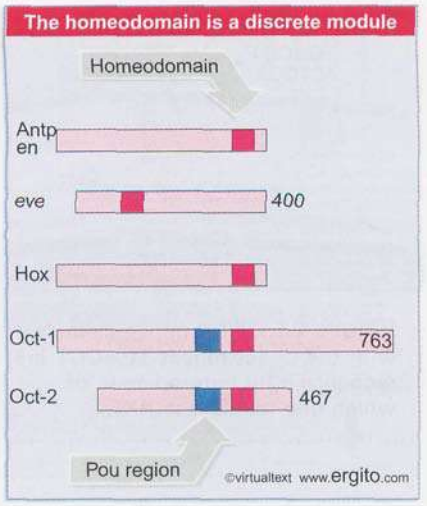


Figure 22.23 The homeodomain may be the sole DNA-binding motif in a transcriptional regulator or may be combined with other motifs. It represents a discrete (60 residue) part of the protein.

The homeobox is a sequence that codes for a domain of 60 amino acids present in proteins of many or even all eukaryotes. Its name derives from its original identification in *Drosophila* homeotic loci (whose genes determine the identity of body structures). It is present in many of the genes that regulate early development in *Drosophila,* and a related motif is found in genes in a wide range of higher eukaryotes. The homeodomain is found in many genes concerned with developmental regulation (see *31.22 The homeobox is a common coding motif in homeotic genes).* Sequences related to the homeodomain are found in several types of animal transcription factors.

In *Drosophila* homeotic genes, the homeodomain often (but not always) occurs close to the C-terminal end. Some examples of genes containing homeoboxes are summarized in **Figure 22.23.** Often the genes have little conservation of sequence except in the homeobox. The conservation of the homeobox sequence varies. A major group of homeobox-containing genes in *Drosophila* has a well conserved sequence, with 80-90% similarity in pairwise comparisons. Other genes have less closely related homeoboxes. The homeodomain is sometimes combined with other motifs in animal transcription factors. One example is presented by the Oct (octamer-binding) proteins, in which a conserved stretch of 75 amino acids called the Pou region is located close to a region resembling the homeodomain. The homeoboxes of the Pou group of proteins are the least closely related to the original group, and thus comprise the farthest extension of the family.

The homeodomain is responsible for binding to DNA, and experiments to swap homeodomains between proteins suggest that the specificity of DNA recognition lies within the homeodomain, but (like the situation with phage repressors) no simple code relating protein and DNA sequences can be deduced. The C-terminal region of the homeodomain shows homology with the helix-turn-helix motif of prokaryotic repressors. We recall from *12.12 Repressor uses a helix-turn-helix motif to bind DNA* that the λ repressor has a "recognition helix" (α-helix-3) that makes contacts in the major groove of DNA, while the other helix (α-helix-2) lies at an angle across the DNA. The homeodomain can be organized into three potential helical regions; the sequences of three examples are compared in **Figure 22.24.** The best conserved part of the sequence lies in the third helix. The difference between these structures and the prokaryotic repressor structures lies in the length of the helix that recognizes DNA, helix-3, which is 17 amino acids long in the homeodomain, compared to 9 residues long in the λ repressor.

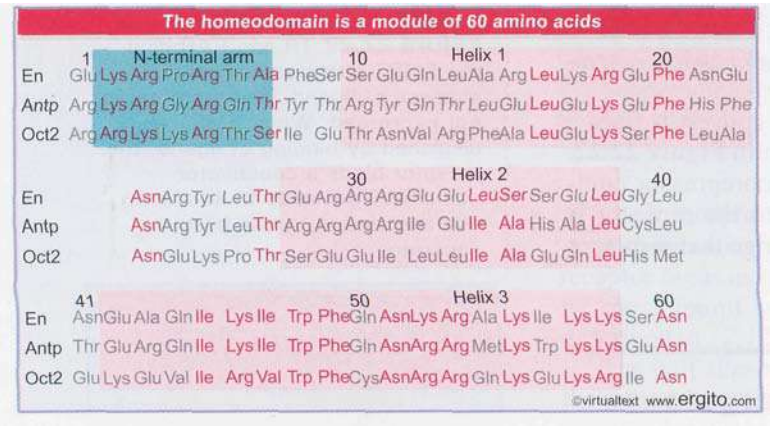


Figure 22.24 The homeodomain of the *Antennapedia* gene represents the major group of genes containing homeoboxes in Drosophila; *engrailed (en)* represents another type of homeotic gene; and the mammalian factor Oct-2 represents a distantly related group of transcription factors. The homeodomain is conventionally numbered from 1 to 60. It starts with the N-terminal arm, and the three helical regions occupy residues 10-22, 28-38, and 42-58. Amino acids in red are conserved in all three examples.

The structure of the homeodomain of the *D. melanogaster* engrailed protein is represented schematically in **Figure 22.25.** Helix 3 binds in the major groove of DNA and makes the majority of the contacts between protein and nucleic acid. Many of the contacts that orient the helix in the major groove are made with the phosphate backbone, so they are not specific for DNA sequence. They lie largely on one face of the double helix, and flank the bases with which specific contacts are made. The remaining contacts are made by the N-terminal arm of the homeodomain, the sequence that just precedes the first helix. It projects into the minor groove. So the N-terminal and C-terminal regions of the homeodomain are primarily responsible for contacting DNA.

A striking demonstration of the generality of this model derives from a comparison of the crystal structure of the homeodomain of engrailed with that of the α2 mating protein of yeast. The DNA-binding domain of this protein resembles a homeodomain, and can form three similar

helices: its structure in the DNA groove can be superimposed almost exactly on that of the engrailed homeodomain. These similarities suggest that all homeodomains bind to DNA in the same manner. This means that a relatively small number of residues in helix-3 and in the N-terminal arm are responsible for specificity of contacts with DNA.

One group of homeodomain-containing proteins is the set of Hox proteins (see Figure 31.39). They bind to DNA with rather low sequence specificity, and it has been puzzling how these proteins can have different specificities. It turns out that Hox proteins often bind to DNA as heterodimers with a partner (called Exd in flies and Pbx in vertebrates). The *heterodimer* has a more restricted specificity *in vitro* than an individual Hox protein; typically it binds the 10 bp sequence TGATNNATNN. Still this is not enough to account *for the* differences *in the specificities of* Hox proteins. A third protein, Hth, which is necessary to localize Exd in the nucleus, also forms part of the complex that binds DNA, and may restrict the binding sites further. But since the same partners (Exd and Hth) are present together with each Hox protein in the trimeric complex, it remains puzzling how each Hox protein has sufficient specificity.

Homeodomain proteins can be either transcriptional activators or repressors. The nature of the factor depends on the other domain(s)—the homeodomain is responsible solely for binding to DNA. The activator or repressor domains both act by influencing the basal apparatus. Activator domains may interact with coactivators that in turn bind to components of the basal apparatus. Repressor domains also interact with the transcription apparatus (that is, they do not act by blocking access to DNA as such). The repressor Eve, for example, interacts directly with $TF_{II}D$.

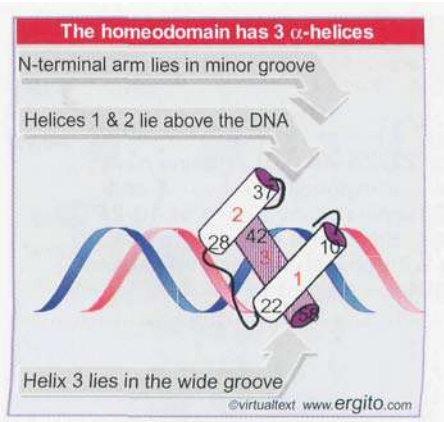


Figure 22.25 Helix 3 of the homeodomain binds in the major groove of DNA, with helices 1 and 2 lying outside the double helix. Helix 3 contacts both the phosphate backbone and specific bases. The N-terminal arm lies in the minor groove, and makes additional contacts.

22.15 Helix-loop-helix proteins interact by combinatorial association

Key Concepts

- Helix-loop-helix proteins have a motif of 40-50 amino acids that comprises two amphipathic α-helices of 15-16 residues separated by a loop.
- The helices are responsible for dimer formation.
- bHLH proteins have a basic sequence adjacent to the HLH motif that is responsible for binding to DNA.
- Class AbHLH proteins are ubiquitously expressed. Class B bHLH proteins are tissue specific.
- A class B protein usually forms a heterodimer with a class A protein.
- HLH proteins that lack the basic region prevent a bHLH partner in a heterodimer from binding to DNA.
- HLH proteins form combinatorial associations that may be changed during development by the addition or removal of specific proteins.

Two common features in DNA-binding proteins are the presence of helical regions that bind DNA, and the ability of the protein to dimerize. Both features are represented in the group of helix-loop-helix proteins that share a common type of sequence motif: a stretch of 40-50 amino acids contains two amphipathic α-helices separated by a linker region (the loop) of varying length. (An amphipathic helix forms two faces, one presenting hydrophobic amino acids, the other presenting charged amino acids.) The proteins in this group form both homodimers and heterodimers by means of interactions between the hydrophobic residues on the corresponding faces of the two helices. The helical regions are 15-16 amino acids long, and each contains several conserved

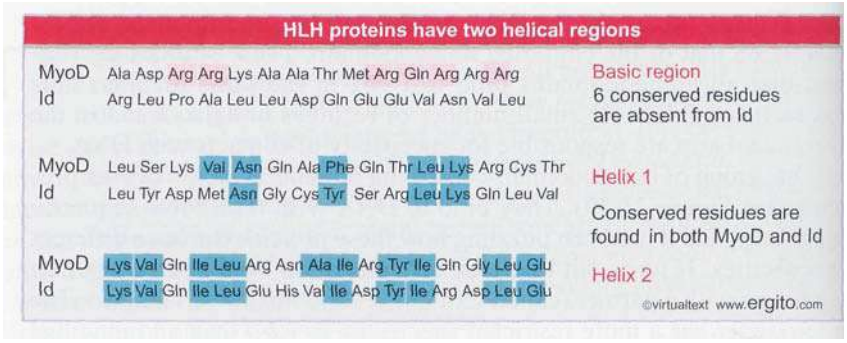


Figure 22.26 All HLH proteins have regions corresponding to helix 1 and helix 2, separated by a loop of 10-24 residues. Basic HLH proteins have a region with conserved positive charges immediately adjacent to helix 1.

residues. Two examples are compared in **Figure 22.26.** The ability to form dimers resides with these amphipathic helices, and is common to all HLH proteins. The loop is probably important only for allowing the freedom for the two helical regions to interact independently of one another.

Most HLH proteins contain a region adjacent to the HLH motif itself that is highly basic, and which is needed for binding to DNA. There are ~6 conserved residues in a stretch of 15 amino acids (see Figure 22.26). Members of the group with such a region are called bHLH proteins. A dimer in which both subunits have the basic region can bind to DNA. The HLH domains probably correctly orient the two basic regions contributed by the individual subunits.

The bHLH proteins fall into two general groups. Class A consists of proteins that are ubiquitously expressed, including mammalian E12/E47. Class B consists of proteins that are expressed in a tissue-specific manner, including mammalian MyoD, myogenin, and Myf-5 (a group of activators that are involved in myogenesis [muscle formation]). A common *modus operandi* for a tissue-specific bHLH protein is to form a heterodimer with a ubiquitous partner. There is also a group of gene products that specify development of the nervous system in *D. melanogaster* (where *Ac-S* is the tissue-specific component and *da* is the ubiquitous component). The Myc proteins (which are the cellular counterparts of oncogene products and are involved in growth regulation) form a separate class of bHLH proteins, whose partners and targets are different.

Dimers formed from bHLH proteins differ in their abilities to bind to DNA. For example, E47 homodimers, E12-E47 heterodimers, and MyoD-E47 heterodimers all form efficiently and bind strongly to DNA; E12 homodimerizes well but binds DNA poorly, while MyoD homodimerizes only poorly. So both dimer formation and DNA binding may represent important regulatory points. At this juncture, it is possible to define groups of HLH proteins whose members form various pairwise combinations, but not to predict from the sequences the strengths of dimer formation or DNA binding. All of the dimers in this group that bind DNA recognize the same consensus sequence, but we do not know yet whether different homodimers and heterodimers have preferences for slightly different target sites that are related to their functions.

Differences in DNA-binding result from properties of the region in or close to the HLH motif; for example, E12 differs from E47 in possessing an inhibitory region just by the basic region, which prevents DNA binding by homodimers. Some HLH proteins lack the basic region and/or contain proline residues that appear to disrupt its function. The example of the protein Id is shown in Figure 22.26. Proteins of this type have the same capacity to dimerize as bHLH proteins, but a dimer that contains one subunit of this type can no longer bind to DNA specifically. This is a forceful demonstration of the importance of doubling the DNA-binding motif in DNA-binding proteins.

The importance of the distinction between the nonbasic HLH and bHLH proteins is suggested by the properties of two pairs of HLH proteins: the *da-Ac-S/emc* pair and the MyoD/Id pair. A model for their functions in forming a regulatory network is illustrated in **Figure 22.27.**

In *D. melanogaster*, the gene *emc (extramacrochaetae)* is required to establish the normal spatial pattern of adult sensory organs. It functions by *suppressing* the functions of several genes, including *da {daughterless)* and the *achaete-scute* complex *(Ac-S). Ac-S* and *da* are genes of the bHLH type. The suppressor *emc* codes for an HLH protein that lacks the basic region. We suppose that, in the absence of *emc* function, the *da* and *Ac-S* proteins form dimers that activate transcription of appropriate target genes, but the production of *emc* protein causes the formation of heterodimers that cannot bind to DNA. So production of *emc* protein in the appropriate cells is necessary to suppress the function of *Ac-S/da*.

The formation of muscle cells is triggered by a change in the transcriptional program that requires several bHLH proteins, including MyoD. MyoD is produced specifically in myogenic cells; and, indeed, overexpression of MyoD in certain other cells can induce them to commence a myogenic program. The trigger for muscle differentiation is probably a heterodimer consisting of MyoD-E12 or MyoD-E47, rather than a MyoD homodimer. Before myogenesis begins, a member of the nonbasic HLH type, the Id protein, may bind to MyoD and/or E12 and E47 to form heterodimers that cannot bind to DNA. It binds to E12/E47 better than to MyoD, and so might function by sequestering the ubiquitous bHLH partner. Overexpression of Id can prevent myogenesis. So the removal of Id could be the trigger that releases MyoD to initiate myogenesis.

A bHLH activator such as MyoD can be controlled in several ways. It is prevented from binding to DNA when it is sequestered by an HLH partner such as Id. It can activate transcription when bound to bHLH partner such as E12 or E47. It can also act as a site-specific repressor when bound to another partner; the bHLH protein MyoR forms a MyoD-MyoR dimer in proliferating myoblasts that represses transcription (at the same target loci at which MyoD-E12/E47 activate transcription).

The behavior of the HLH proteins therefore illustrates two general principles of transcriptional regulation. A small number of proteins form combinatorial associations. Particular combinations have different functions with regard to DNA binding and transcriptional regulation. Differentiation may depend either on the presence or on the removal of particular partners.

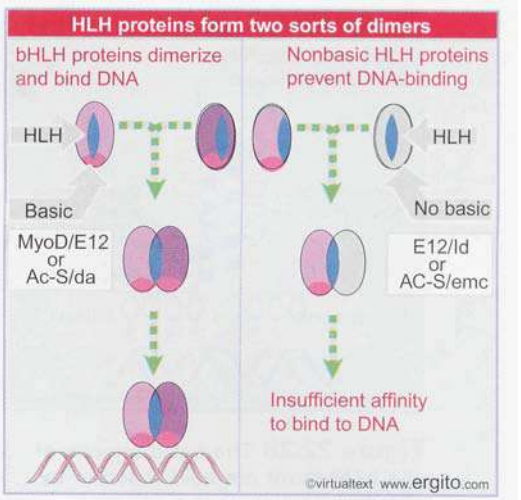


Figure 22.27 An HLH dimer in which both subunits are of the bHLH type can bind DNA, but a dimer in which one subunit lacks the basic region cannot bind DNA.

22.16 Leucine zippers are involved in dimer formation

Key Concepts

- The leucine zipper is an amphipathic helix that dimerizes.
- The zipper is adjacent to a basic region that binds DNA.
- Dimerization forms the bZIP motif in which the two basic regions symmetrically bind inverted repeats in DNA.

Interactions between proteins are a common theme in building a transcription complex, and a motif found in several activators (and other proteins) is involved in both homo- and heteromeric interactions. The leucine zipper is a stretch of amino acids rich in leucine residues that provide a dimerization motif. Dimer formation itself has emerged as a common principle in the action of proteins that recognize specific DNA sequences, and in the case of the leucine zipper, its relationship to DNA binding is especially clear, because we can see how dimerization juxtaposes the DNA-binding regions of each subunit. The reaction is depicted diagrammatically in **Figure 22.28.**

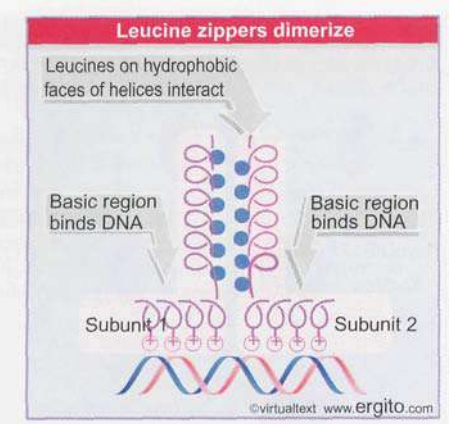


Figure 22.28 The basic regions of the bZIP motif are held together by the dimerization at the adjacent zipper region when the hydrophobic faces of two leucine zippers interact in parallel orientation.

An amphipathic α-helix has a structure in which the hydrophobic groups (including leucine) face one side, while charged groups face the other side. A leucine zipper forms an amphipathic helix in which the leucines of the zipper on one protein could protrude from the α-helix and interdigitate with the leucines of the zipper of another protein in parallel to form a coiled coil. The two right-handed helices wind around each other, with 3.5 residues per turn, so the pattern repeats integrally every 7 residues.

How is this structure related to DNA binding? The region adjacent to the leucine repeats is highly basic in each of the zipper proteins, and could comprise a DNA-binding site. The two leucine zippers in effect form a Y-shaped structure, in which the zippers comprise the stem, and the two basic regions stick out to form the arms that bind to DNA. This is known as the bZIP structural motif. It explains why the target sequences for such proteins are inverted repeats with no separation.

Zippers may be used to sponsor formation of homodimers or heterodimers. They are lengthy motifs. Leucine (or another hydrophobic amino acid) occupies every seventh residue in the potential zipper. There are 4 repeats of the zipper ($Leu\text{-}X_6$) in the protein C/EBP (a factor that binds as a dimer to both the CAAT box and the SV40 core enhancer), and 5 repeats in the factors Jun and Fos (which form the heterodimeric activator, AP1).

AP1 was originally identified by its binding to a DNA sequence in the SV40 enhancer (see Figure 21.23). The active preparation of AP1 includes several polypeptides. A major component is Jun, the product of the gene *c-jun,* which was identified by its relationship with the oncogene *V'jun* carried by an avian sarcoma virus (see *30.18 Oncoproteins may regulate gene expression)*. The mouse genome contains a family of related genes, *c-jun* (the original isolate) and *junB* and *junD* (identified by sequence homology with *jun)*. There are considerable sequence similarities in the three Jun proteins; they have leucine zippers that can interact to form homodimers or heterodimers.

The other major component of AP1 is the product of another gene with an oncogenic counterpart. The *c-fos* gene is the cellular homologue to the oncogene *v-fos* carried by a murine sarcoma virus. Expression of *c-fos* activates genes whose promoters or enhancers possess an AP1 target site. The *c-fos* product is a nuclear phosphoprotein that is one of a group of proteins. The others are described as Fos-related antigens (FRA); they constitute a family of Fos-like proteins.

Fos also has a leucine zipper. Fos cannot form homodimers, but can form a heterodimer with Jun. A leucine zipper in each protein is required for the reaction. The ability to form dimers is a crucial part of the interaction of these factors with DNA. Fos cannot by itself bind to DNA, possibly because of its failure to form a dimer. But the Jun-Fos heterodimer can bind to DNA with same target specificity as the Jun-Jun dimer; and this heterodimer binds to the AP1 site with an affinity ~10× that of the Jun homodimer.

22.17 Summary

Transcription factors include basal factors, activators, and coactivators. Basal factors interact with RNA polymerase at the startpoint. Activators bind specific short response elements (REs) located in promoters or enhancers. Activators function by making protein-protein interactions with the basal apparatus. Some activators interact directly with the basal apparatus; others require coactivators to mediate the interaction. The targets in the

basal apparatus are the TAFs of TF||D, or $TF_{II}B$ or $TF_{II}A$. The interaction stimulates assembly of the basal apparatus. Activators often have a modular construction, in which there are independent domains responsible for binding to DNA and for activating transcription. The main function of the DNA-binding domain may be to tether the activating domain in the vicinity of the initiation complex. Some response elements are present in many genes and are recognized by ubiquitous factors; others are present in a few genes and are recognized by tissue-specific factors.

Several groups of transcription factors have been identified by sequence homologies. The homeodomain is a 60 residue sequence found in genes that regulate development in insects and worms and in mammalian transcription factors. It is related to the prokaryotic helix-turn-helix motif and provides the motif by which the factors bind to DNA.

Another motif involved in DNA-binding is the zinc finger, which is found in proteins that bind DNA or RNA (or sometimes both). A finger has cysteine residues that bind zinc. One type of finger is found in multiple repeats in some transcription factors; another is found in single or double repeats in others.

Steroid receptors were the first members identified of a group of transcription factors in which the protein is activated by binding a small hydrophobic hormone. The activated factor becomes localized in the nucleus, and binds to its specific response element, where it activates transcription. The DNA-binding domain has zinc fingers. The receptors are homodimers or heterodimers. The homodimers all recognize palindromic response elements with the same consensus sequence; the difference between the response elements is the spacing between the inverted repeats. The heterodimers recognize direct repeats, again being distinguished by the spacing between the repeats. The DNA-binding motif of these receptors includes two zinc fingers; the first determines which consensus sequence is recognized, and the second responds to the spacing between the repeats.

The leucine zipper contains a stretch of amino acids rich in leucine that are involved in dimerization of transcription factors. An adjacent basic region is responsible for binding to DNA.

HLH (helix-loop-helix) proteins have amphipathic helices that are responsible for dimerization, adjacent to basic regions that bind to DNA. bHLH proteins have a basic region that binds to DNA, and fall into two groups: ubiquitously expressed and tissue-specific. An active protein is usually a heterodimer between two subunits, one from each group. When a *dimer* has one subunit that does not have the basic region, it fails to bind DNA, so such subunits can prevent gene expression. Combinatorial associations of subunits form regulatory networks.

Many transcription factors function as dimers, and it is common for there to be multiple members of a family that form homodimers and heterodimers. This creates the potential for complex combinations to govern gene expression. In some cases, a family includes inhibitory members, whose participation in dimer formation prevents the partner from activating transcription.

References

22.2 There are several types of transcription factors

rev Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. Ann. Rev. Genet. 34, 77-137.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14, 2551-2569.

22.3 Independent domains bind DNA and activate transcription

rev Guarente, L. (1987). Regulatory proteins in yeast. Ann. Rev. Genet. 21, 425-452.

Ptashne, M. (1988). How eukaryotic transcriptional activators work. Nature 335, 683-689.

22.4 The two hybrid assay detects protein-protein interactions

ref Fields, S., and Song, 0. (1989). A novel genetic system to detect protein-protein interactions. Nature 340, 245-246.

22.5 Activators interact with the basal apparatus

rev Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. Science 236, 1237-1245.

Mitchell. P. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA-binding proteins. Science 245, 371-378.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14, 2551-2569.

Myers, L. C. and Kornberg, R. D. (2000). Mediator of transcriptional regulation. Ann. Rev. Biochem. 69, 729-749.

ref Asturias, F. J., Jiang, Y. W., Myers, L. C, Gustafsson, C. M., and Kornberg, R. D. (1999). Conserved structures of mediator and RNA polymerase II holoenzyme. Science 283, 985-987.

Chen, J.-L. et al. (1994). Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators. Cell 79, 93-105.

Dotson, M. R., Yuan, C. X., Roeder, R. G., Myers, L. C, Gustafsson, C. M., Jiang, Y. W., Li, Y., Kornberg, R. D., and Asturias, F. J. (2000). Structural organization of yeast and mammalian mediator complexes. Proc. Nat. Acad. Sci. USA 97, 14307-14310.

Dynlacht, B. D., Hoey, T., and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell 66, 563-576.

Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H., and Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell 77, 599-608.

Ma, J. and Ptashne, M. (1987). A new class of yeast transcriptional activators. Cell 51, 113-119.

Pugh, B. F. and Tjian, R. (1990). Mechanism of transcriptional activation by Sp1: evidence for coactivators. Cell 61, 1187-1197.

22.6 Some promoter-binding proteins are repressors

ref Goppelt, A., Stelzer, G., Lottspeich, F., and Meisterernst, M. (1996). A mechanism for repression of class II gene transcription through specific binding of NC2 to TBP-promoter complexes via heterodimeric histone fold domains. EMBO J. 15, 3105-31 16.

Inostroza, J. A., Mermelstein, F. H., Ha, I., Lane, W. S., and Reinberg, D. (1992). Dr1, a TATA-binding protein-associated phosphoprotein and inhibitor of class II gene transcription. Cell 70, 477-489.

Kim, T. K., Kim, T. K., Zhao, Y., Ge, H., Bernstein, R., and Roeder, R. G. (1995). TATA-binding protein residues implicated in a functional interplay between negative cofactor NC2 (Dr1) and general factors TFIIA and TFIIB. J. Biol. Chem. 270, 10976-10981.

22.8 There are many types of DNA-binding domains

rev Harrison, S. C. (1991). A structural taxonomy of DNA-binding proteins. Nature 353, 715-719.

Pabo, C. T. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. Ann. Rev. Biochem. 61, 1053-1095.

ref Miller, J. et al. (1985). Repetitive zinc binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. 4, 1609-1614.

Murre, C, McCaw, P. S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell 56, 777-783.

22.9 A zinc finger motif is a DNA-binding domain

ref Kadonaga, J. et al. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell 51, 1079-1090.

Miller, J. et al. (1985). Repetitive zinc binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. 4, 1609-1614.

Pavletich, N. P. and Pabo, C. 0. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 21 A. Science 252, 809-817.

22.10 Steroid receptors are activators

rev Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. Science 240, 889-895.

ref Mangelsdorf, D. J. and Evans, R. (1995). The RXR heterodimers and orphan receptors. Cell 83, 841-850.

22.11 Steroid receptors have zinc fingers

rev Tsai, M.-J. and O'Malley, B. W. (1994). Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Ann. Rev. Biochem. 63, 451-486.

ref Umesono, K. and Evans, R. M. (1989). Determinants of target gene specificity for steroid/thyroid hormone receptors. Cell 57, 1139-1146.

22.13 Steroid receptors recognize response elements by a combinatorial code

rev Yamamoto, K. R. (1985). Steroid receptor regulated transcription of specific genes and gene networks. Ann. Rev. Genet. 19, 209-252.

ref Hurlein, A. J. et al. (1995). Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor corepressor. Nature 377, 397-404.

Mangelsdorf, D. J. and Evans, R. (1995). The RXR heterodimers and orphan receptors. Cell 83, 841-850.

Rastinejad, F., Perlmann, T., Evans, R. M., and Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. Nature 375, 203-211.

Umesono, K., Murakami, K. K., Thompson, C. C, and Evans, R. M. (1991). Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. Cell 65, 1255-1266.

22.14 Homeodomains bind related targets in DNA

rev Gehring, W. J. et al. (1994). Homeodomain-DNA recognition. Cell 78, 211-223.

ref Han, K., Levine, M. S., and Manley, J. L. (1989). Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. Cell 56, 573-583.

Wolberger, C. et al. (1991). Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell 67, 517-528.

22.15 Helix-loop-helix proteins interact by combinatorial association

rev Weintraub, H. (1991). The MyoD gene family: nodal point during specification of the muscle cell lineage. Science 251, 761-766.

ref Benezra, R. et al. (1990). The protein Id: a negative regulator of helix-loop-helix DNA-binding proteins. Cell 61, 49-59.

Davis, R. L et al. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell 51, 987-1000.

Davis, R. L. et al. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell 60, 733-746.

Lassar, A. B. et al. (1991). Functional activity of myogenic HLH proteins requires hetero-oligomerization with E12/E47-like proteins in vivo. Cell 66, 305-315.

Murre, C, McCaw, P. S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell 56, 777-783.

22.16 Leucine zippers are involved in dimer formation

ref Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science 240, 1759-1764.

Vinson, C. R., Sigler, P. B., and McKnight, S. L. (1989). Scissors-grip model for DNA recognition by a family of leucine zipper proteins. Science 246, 911-916.

Chapter 23

Controlling chromatin structure

23.1 Introduction

When transcription is treated in terms of interactions involving DNA and individual transcription factors and RNA polymerases, we get an accurate description of the events that occur *in vitro*, but this lacks an important feature of transcription *in vivo*. The cellular genome is organized as nucleosomes, but initiation of transcription generally is prevented if the promoter region is packaged into nucleosomes. In this sense, histones function as generalized repressors of transcription (a rather old idea), although we see in this Chapter that they are also involved in more specific interactions. Activation of a gene requires changes in the state of chromatin: the essential issue is how the transcription factors gain access to the promoter DNA.

Local chromatin structure is an integral part of controlling gene expression. Genes may exist in either of two structural conditions. Genes are found in an "active" state only in the cells in which they are expressed. The change of structure precedes the act of transcription, and indicates that the gene is "transcribable." This suggests that acquisition of the "active" structure must be the first step in gene expression. Active genes are found in domains of euchromatin with a preferential susceptibility to nucleases (see *20.15 Domains define regions that contain active genes).* Hypersensitive sites are created at promoters before a gene is activated (see *20.14 DNAase hypersensitive sites change chromatin structure).*

More recently it has turned out that there is an intimate and continuing connection between initiation of transcription and chromatin structure. Some activators of gene transcription directly modify histones; in particular, acetylation of histones is associated with gene activation. Conversely, some repressors of transcription function by deacetylating histones. So a reversible change in histone structure in the vicinity of the promoter is involved in the control of gene expression. This may be part of the mechanism by which a gene is maintained in an active or inactive state.

The mechanisms by which local regions of chromatin are maintained in an inactive (silent) state are related to the means by which an

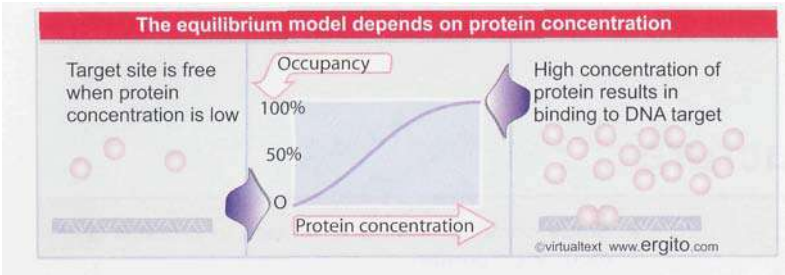


Figure 23.1 In an equilibrium model, the state of a binding site on DNA depends on the concentration of the protein that binds to it.

individual promoter is repressed. The proteins involved in the formation of heterochromatin act on chromatin via the histones, and modifications of the histones may be an important feature in the interaction. Once established, such changes in chromatin may persist through cell divisions, creating an epigenetic state in which the properties of a gene are determined by the self-perpetuating structure of chromatin. The name epigenetic reflects the fact that a gene may have an inherited condition (it may be active or may be inactive) which does not depend on its sequence. Yet a further insight into epigenetic properties is given by the self-perpetuating structures of prions (proteinaceous infectious agents).

23.2 Chromatin can have alternative states

Key Concepts

- Chromatin structure is stable and cannot be changed by altering the equilibrium of transcription factors and histones.

Two types of model have been proposed to explain how the state of expression of DNA is changed: equilibrium and discontinuous change-of-state.

Figure 23.1 shows the equilibrium model. Here the only pertinent factor is the concentration of the repressor or activator protein, which drives an equilibrium between free form and DNA-bound form. When the concentration of the protein is high enough, its DNA-binding site is occupied, and the state of expression of the DNA is affected. (Binding might either repress or activate any particular target sequence.) This type of model explains the regulation of transcription in bacterial cells, where gene expression is determined exclusively by the actions of individual repressor and activator proteins (see *10 The Operon)*. Whether a bacterial gene is transcribed can be predicted from the sum of the concentrations of the various factors that either activate or repress the individual gene. Changes in these concentrations *at any time* will change the state of expression accordingly. In most cases, the protein binding is cooperative, so that once the concentration becomes high enough, there is a rapid association with DNA, resulting in a switch in gene expression.

A different situation applies with eukaryotic chromatin. Early *in vitro* experiments showed that either an active or inactive state can be established, but this is not affected by the subsequent addition of other components. The transcription factor $TF_{III}A$, required for RNA polymerase III to transcribe 5S rRNA genes, cannot activate its target genes *in vitro* if they are complexed with histones. However, if the factor is presented with free DNA, it forms a transcription complex, and then the addition of histones does not prevent the gene from remaining active. Once the factor has bound, it remains at the site, allowing a succession of RNA polymerase molecules to initiate transcription. Whether the factor or histones get to the control site first may be the critical factor.

Figure 23.2 illustrates the two types of condition that can exist at a eukaryotic promoter. In the inactive state, nucleosomes are present, and they prevent basal factors and RNA polymerase from binding. In the active state, the basal apparatus occupies the promoter, and histone octamers cannot bind to it. Each type of state is stable.

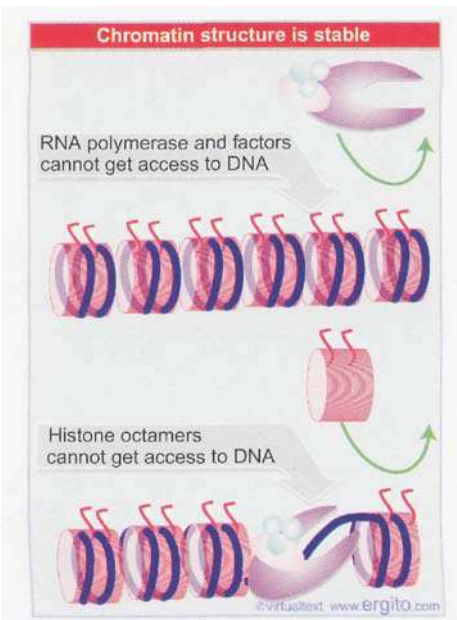


Figure 23.2 If nucleosomes form at a promoter, transcription factors (and RNA polymerase) cannot bind. If transcription factors (and RNA polymerase) bind to the promoter to establish a stable complex for initiation, histones are excluded.

By Book_Crazy [IND]

A similar situation is seen with the $TF_{II}D$ complex at promoters for RNA polymerase II. A plasmid containing an adenovirus promoter can be transcribed *in vitro* by RNA polymerase II in a reaction that requires $TF_{II}D$ and other transcription factors. The template can be assembled into nucleosomes by the addition of histones. If the histones are added *before* the $TF_{II}D$, transcription cannot be initiated. But if the $TF_{II}D$ is added first, the template still can be transcribed in its chromatin form. So $TF_{II}D$ can recognize free DNA, but either cannot recognize or cannot function on nucleosomal DNA. Only the $TF_{II}D$ must be added before the histones; the other transcription factors and RNA polymerase can be added later. This suggests that binding of $TF_{II}D$ to the promoter creates a structure to which the other components of the transcription apparatus can bind.

It is important to note that these *in vitro* systems use disproportionate quantities of components, which may create unnatural situations. The major importance of these results, therefore, is not that they demonstrate the mechanism used *in vivo*, but that they establish the principle that *transcription factors or nucleosomes may form stable structures that cannot be changed merely by changing the equilibrium with free components.*

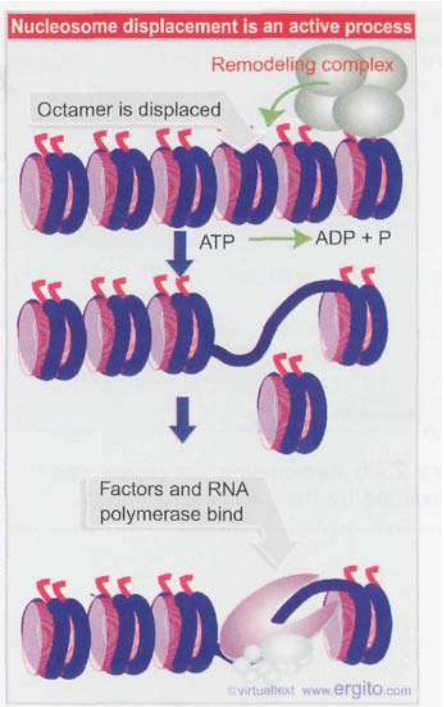


Figure 23.3 The dynamic model for transcription of chromatin relies upon factors that can use energy provided by hydrolysis of ATP to displace nucleosomes from specific DNA sequences.

23.3 Chromatin remodeling is an active process

Key Concepts

- There are several chromatin remodeling complexes that use energy provided by hydrolysis of ATP.
- The SWI/SNF, RSC, and NURF complexes all are very large; there are some common subunits.
- A remodeling complex does not itself have specificity for any particular target site, but must be recruited by a component of the transcription apparatus.

The general process of inducing changes in chromatin structure is called chromatin remodeling. This consists of mechanisms for displacing histones that depend on the input of energy. Many protein-protein and protein-DNA contacts need to be disrupted to release histones from chromatin. There is no free ride: the energy must be provided to disrupt these contacts. **Figure 23.3** illustrates the principle of a *dynamic model* by a factor that hydrolyzes ATP. When the histone octamer is released from DNA, other proteins (in this case transcription factors and RNA polymerase) can bind.

Figure 23.4 summarizes the types of remodeling changes in chromatin that can be characterized *in vitro:*

- Histone octamers may slide along DNA, changing the relationship between the nucleic acid and protein. This alters the position of a particular sequence on the nucleosomal surface.
- The spacing between histone octamers may be changed, again with the result that the positions of individual sequences are altered relative to protein.
- And the most extensive change is that an octamer(s) may be displaced entirely from DNA to generate a nucleosome-free gap.

The most common use of chromatin remodeling is to change the organization of nucleosomes at the promoter of a gene that is to be transcribed. This is required to allow the transcription apparatus to gain access to the promoter. The remodeling most often takes the form of

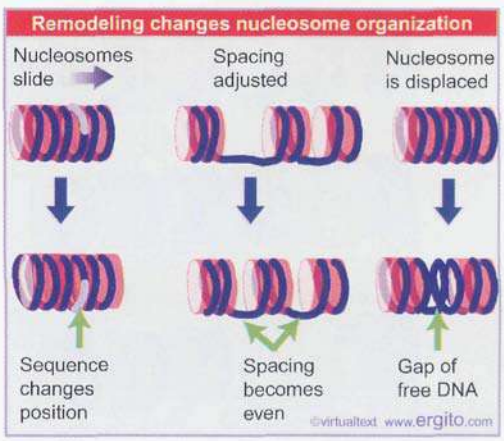


Figure 23.4 Remodeling complexes can cause nucleosomes to slide along DNA, can displace nucleosomes from DNA, or can reorganize the spacing between nucleosomes.

There are several types of remodeling complexes

Type of complex	SWI/SNF	ISW	Other
Yeast	SWI/SNF RSC	ISW1 ISW2	
Fly	dSWI/SNF (Brahma)	NURF CHRAC ACF	
Human	hSWI/SNF	RSF hACF/WCFR hCHRAC	NuRD
Frog			Mi-2

©virtualtext www.ergito.com

Figure 23.5 Remodeling complexes can be classified by their ATPase subunits.

displacing one or more histone octamers. This can be detected by a change in the micrococcal nuclease ladder where protection against cleavage has been lost. It often results in the creation of a site that is hypersensitive to cleavage with DNAase I (see *20.14 DNAase hypersensitive sites change chromatin structure).* Sometimes there are less dramatic changes, for example, involving a change in rotational positioning of a single nucleosome; this may be detected by loss of the DNAaseI 10 base ladder. So changes in chromatin structure may extend from altering the positions of nucleosomes to removing them altogether.

Chromatin remodeling is undertaken by large complexes that use ATP hydrolysis to provide the energy for remodeling. The heart of the remodeling complex is its ATPase subunit. Remodeling complexes are usually classified according to the type of ATPase subunit—those with related ATPase subunits are considered to belong to the same family (usually some other subunits are common also). **Figure 23.5** keeps the names straight. The two major types of complex are SWI/SNF and ISW (ISW stands for imitation SWI). Yeast has two complexes of each type. Complexes of both types are also found in fly and in Man. Each type of complex may undertake a different range of remodeling activities.

SWI/SNF was the first remodeling complex to be identified. Its name reflects the fact that many of its subunits are coded by genes originally identified by *SWI or SNF* mutations in *S. cerevisiae.* Mutations in these loci are pleiotropic, and the range of defects is similar to those shown by mutants that have lost the CTD tail of RNA polymerase II. These mutations also show genetic interactions with mutations in genes that code for components of chromatin, in particular *SIN1,* which codes for a nonhistone protein, and *SIN2,* which codes for histone H3. The *SWI* and *SNF* genes are required for expression of a variety of individual loci (~120 or 2% of *S. cerevisiae* genes are affected). Expression of these loci may require the SWI/SNF complex to remodel chromatin at their promoters.

SWI/SNF acts catalytically *in vitro,* and there are only ~150 complexes per yeast cell. All of the genes encoding the SWI/SNF subunits are nonessential, which implies that yeast must also have other ways of remodeling chromatin. The RSC complex is more abundant and also is essential. It acts at ~700 target loci.

SWI/SNF complexes can remodel chromatin *in vitro* without overall loss of histones or can displace histone octamers. Both types of reaction may pass through the same intermediate in which the structure of the target nucleosome is altered, leading either to reformation of a (remodeled) nucleosome on the original DNA or to displacement of the histone octamer to a different DNA molecule. The SWI/SNF complex alters nucleosomal sensitivity to DNAase I at the target site, and induces changes in protein-DNA contacts that persist after it has been released from the nucleosomes. The SWI2 subunit is the ATPase that provides the energy for remodeling by SWI/SNF.

There are many contacts between DNA and a histone octamer—14 are identified in the crystal structure. All of these contacts must be broken for an octamer to be released or for it to move to a new position. How is this achieved? Some obvious mechanisms can be excluded because we know that single-stranded DNA is not generated during remodeling (and there are no helicase activities associated with the complexes). Present thinking is that remodeling complexes in the SWI and ISW classes use the hydrolysis of ATP to twist DNA on the nucleosomal surface. Indirect evidence suggests that this creates a mechanical force that allows a small region of DNA to be released from the surface and then repositioned.

One important reaction catalyzed by remodeling complexes involves nucleosome sliding. It was first observed that the ISW family affects

nucleosome positioning without displacing octamers. This is achieved by a sliding reaction, in which the octamer moves along DNA. Sliding is prevented if the N-terminal tail of histone H4 is removed, but we do not know exactly how the tail functions in this regard. SWI/SNF complexes have the same capacity; the reaction is prevented by the introduction of a barrier in the DNA, which suggests that a sliding reaction is involved, in which the histone octamer moves more or less continuously along DNA without ever losing contact with it.

One puzzle about the action of the SWI/SNF complex is its sheer size. It has 11 subunits with a combined molecular weight ~2 X 10^6. It dwarfs RNA polymerase and the nucleosome, making it difficult to understand how all of these components could interact with DNA retained on the nucleosomal surface. However, a transcription complex with full activity, called RNA polymerase II holoenzyme, can be found that contains the RNA polymerase itself, all the TF_{II} factors except TBP and $TF_{II}A$, and the SWI/SNF complex, which is associated with the CTD tail of the polymerase. In fact, virtually all of the SWI/SNF complex may be present in holoenzyme preparations. This suggests that the remodeling of chromatin and recognition of promoters is undertaken in a coordinated manner by a single complex.

23.4 Nucleosome organization may be changed at the promoter

Key Concepts

- Remodeling complexes are recruited to promoters by sequence-specific activators.
- The factor may be released once the remodeling complex has bound.
- The MMTV promoter requires a change in rotational positioning of a nucleosome to allow an activator to bind to DNA on the nucleosome.

How are remodeling complexes targeted to specific sites on chromatin? They do not themselves contain subunits that bind specific DNA sequences. This suggests the model shown in **Figure 23.6** in which they are recruited by activators or (sometimes) by repressors.

The interaction between transcription factors and remodeling complexes gives a key insight into their modus operandi. The transcription factor Swi5p activates the *HO* locus in yeast. (Note that Swi5p is not a member of the SWI/SNF complex.) Swi5p enters nuclei toward the end of mitosis and binds to the *HO* promoter. It then recruits SWI/SNF to the promoter. Then Swi5p is released, leaving SWI/SNF at the promoter. This means that a transcription factor can activate a promoter by a "hit and run" mechanism, in which its function is fulfilled once the remodeling complex has bound.

The involvement of remodeling complexes in gene activation was discovered because the complexes are necessary for the ability of certain transcription factors to activate their target genes. One of the first examples was the GAGA factor, which activates the *hsp70 Drosophila* promoter *in vitro*. Binding of GAGA to four $(CT)_n$-rich sites on the promoter disrupts the nucleosomes, creates a hypersensitive region, and causes the adjacent nucleosomes to be rearranged so that they occupy preferential instead of random positions. Disruption is an energy-dependent process that requires the NURF remodeling complex. The organization of nucleosomes is altered so as to create a boundary that determines the positions of the adjacent nucleosomes. During this

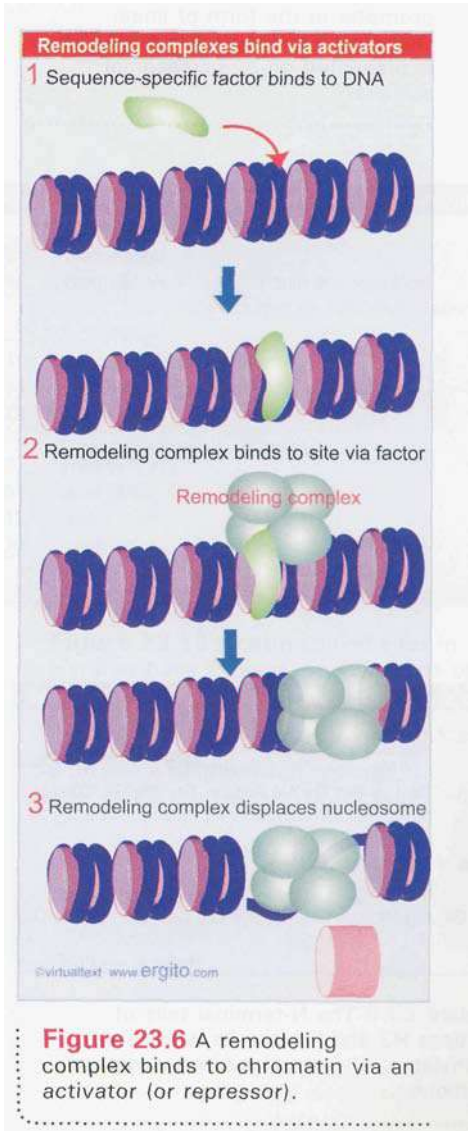


Figure 23.6 A remodeling complex binds to chromatin via an activator (or repressor).

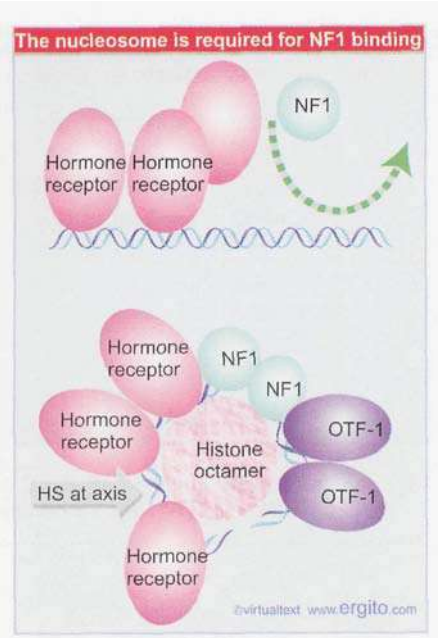


Figure 23.7 Hormone receptor and NF1 cannot bind simultaneously to the MMTV promoter in the form of linear DNA, but can bind when the DNA is presented on a nucleosomal surface.

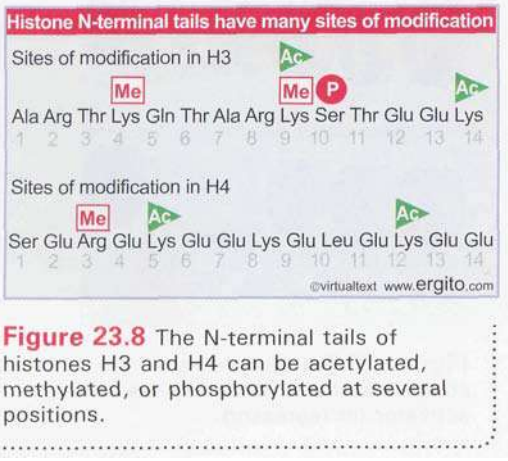


Figure 23.8 The N-terminal tails of histones H3 and H4 can be acetylated, methylated, or phosphorylated at several positions.

process, GAGA binds to its target sites and DNA, and its presence fixes the remodeled state.

The *PHO* system was one of the first in which it was shown that a change in nucleosome organization is involved in gene activation. At the *PHO5* promoter, the bHLH regulator PHO4 responds to phosphate starvation by inducing the disruption of four precisely positioned nucleosomes. This event is independent of transcription (it occurs in a TATA⁻ mutant) and independent of replication. There are two binding sites for PHO4 at the promoter, one located between nucleosomes, which can be bound by the isolated DNA-binding domain of PHO4, and the other within a nucleosome, which cannot be recognized. Disruption of the nucleosome to allow DNA binding at the second site is necessary for gene activation. This action requires the presence of the transcription-activating domain. The activator sequence of VP16 can substitute for the PHO4 activator sequence in nucleosome disruption. This suggests that disruption occurs by protein-protein interactions that involve the same region that makes protein-protein contacts to activate transcription. In this case, it is not known which remodeling complex is involved in executing the effects.

It is not always the case, however, that nucleosomes must be excluded in order to permit initiation of transcription. Some activators can bind to DNA on a nucleosomal surface. Nucleosomes appear to be precisely positioned at some steroid hormone response elements in such a way that receptors can bind. Receptor binding may alter the interaction of DNA with histones, and even lead to exposure of new binding sites. The exact positioning of nucleosomes could be required either because the nucleosome "presents" DNA in a particular rotational phase or because there are protein-protein interactions between the activators and histones or other components of chromatin. So we have now moved some way from viewing chromatin exclusively as a repressive structure to considering which interactions between activators and chromatin can be required for activation.

The MMTV promoter presents an example of the need for specific nucleosomal organization. It contains an array of 6 partly palindromic sites, each bound by one dimer of hormone receptor (HR), which constitute the HRE. It also has a single binding site for the factor NF1, and two adjacent sites for the factor OTF. HR and NF1 cannot bind simultaneously to their sites in free DNA. **Figure 23.7** shows how the nucleosomal structure controls binding of the factors.

The HR protects its binding sites at the promoter when hormone is added, but does not affect the micrococcal nuclease-sensitive sites that mark either side of the nucleosome. This suggests that HR is binding to the DNA on the nucleosomal surface. However, the rotational positioning of DNA on the nucleosome prior to hormone addition allows access to only two of the four sites. Binding to the other two sites requires a change in rotational positioning on the nucleosome. This can be detected by the appearance of a sensitive site at the axis of dyad symmetry (which is in the center of the binding sites that constitute the HRE). NF1 can be footprinted on the nucleosome after hormone induction, so these structural changes may be necessary to allow NF1 to bind, perhaps because they expose DNA and abolish the steric hindrance by which HR blocks NF1 binding to free DNA.

23.5 Histone modification is a key event

Whether a gene is expressed depends on the structure of chromatin both locally (at the promoter) and in the surrounding domain. Chromatin structure correspondingly can be regulated by individual activation events or by changes that affect a wide chromosomal

region. The most localized events concern an individual target gene, where changes in nucleosomal structure and organization occur in the immediate vicinity of the promoter. More general changes may affect regions as large as a whole chromosome.

Changes that affect large regions control the potential of a gene to be expressed. The term silencing is used to refer to repression of gene activity in a local chromosomal region. The term heterochromatin is used to describe chromosomal regions that are large enough to be seen to have a physically more compact structure in the microscope. The basis for both types of change is the same: additional proteins bind to chromatin and either directly or indirectly prevent transcription factors and RNA polymerase from activating promoters in the region.

Changes at an individual promoter control whether transcription is initiated for a particular gene. These changes may be either activating or repressing.

All of these events depend on interactions with histones. Changes in chromatin structure are initiated by modifying the N-terminal tails of the histones, especially H3 and H4. The histone tails consist of the N-terminal 20 amino acids, and extend from the nucleosome between the turns of DNA (see Figure 20.25 in *20.8 Organization of the histone octamer).* **Figure 23.8** shows that they can be modified at several sites, by methylation, acetylation, or phosphorylation (see *20.9 The N-terminal tails of histones are modified).* The modifications reduce positive charge. The histone modifications may directly affect nucleosome structure or create binding sites for the attachment of nonhistone proteins that change the properties of chromatin.

The range of nucleosomes that is targeted for modification can vary. Modification can be a local event, for example, restricted to nucleosomes at the promoter. Or it can be a general event, extending for example to an entire chromosome. **Figure 23.9** shows that there is a general correlation in which acetylation is associated with active chromatin while methylation is associated with inactive chromatin. However, this is not a simple rule, and the particular sites that are modified, as well as combinations of specific modifications may be important, so there are certainly exceptions in which (for example) histones methylated at a certain position are found in active chromatin. Mutations in one of the histone acetylase complexes of yeast have the opposite effect from usual (they prevent silencing of some genes), emphasizing the lack of a uniform effect of acetylation.

The specificity of the modifications is indicated by the fact that many of the modifying enzymes have individual target sites in specific histones. **Figure 23.10** summarizes the effects of some of the modifications. Most modified sites are subject to only a single type of modification. In some cases, modification of one site may activate or inhibit modification of another site. The idea that combinations of signals may be used to define chromatin types has sometimes been called the *histone code.*

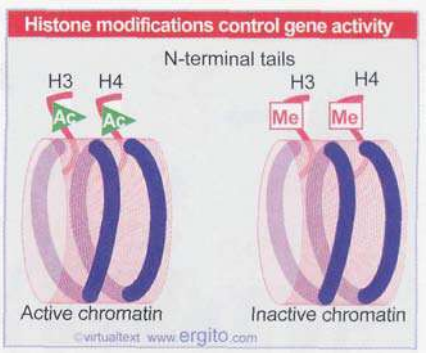


Figure 23.9 Acetylation of H3 and H4 is associated with active chromatin, while methylation is associated with inactive chromatin.

Histone modification affects the structure and function of chromatin

Histone	Site	Modification	Function
H3	Lys-4	methylation	
H3	Lys-9	methylation	chromatin condensation
	"	"	required for DNA methylation
	"	acetylation	
H3	Ser-10	phosphorylation	
H3	Lys-14	acetylation	prevents methylation at Lys-9
H3	Lys-79	methylation	*telomeric silencing*
H4	Arg-3	methylation	
H4	Lys-5	acetylation	
H4	Lys-12	acetylation	
H4	Lys-16	acetylation	nucleosome assembly
	"	"	fly X activation

©virtualtext www.ergito.com

Figure 23.10 Most modified sites in histones have a single, specific type of modification, but some sites can have more than one type of modification. Individual functions can be associated with some of the modifications.

23.6 Histone acetylation occurs in two circumstances

Key Concepts

- Histone acetylation occurs transiently at replication.
- Histone acetylation is associated with activation of gene expression.

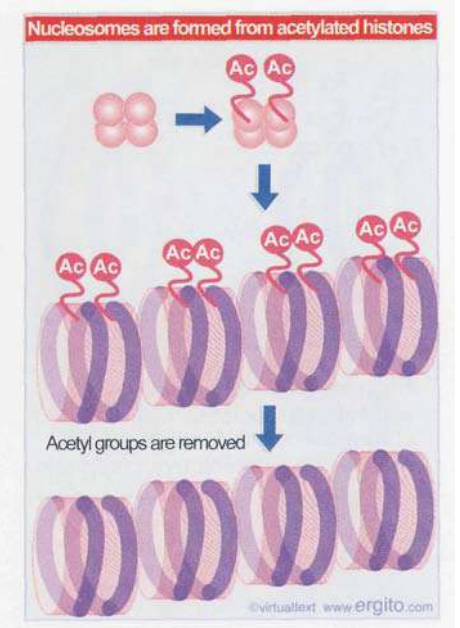


Figure 23.11 Acetylation at replication occurs on histones before they are incorporated into nucleosomes.

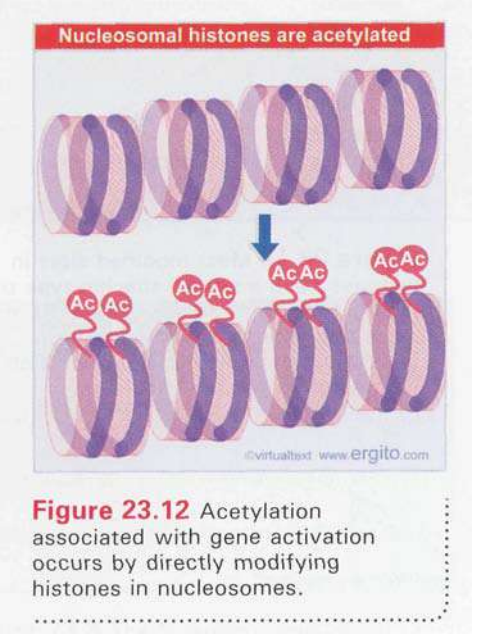


Figure 23.12 Acetylation associated with gene activation occurs by directly modifying histones in nucleosomes.

All the core histones can be acetylated. The major targets for acetylation are lysines in the N-terminal tails of histones H3 and H4. Acetylation occurs in two different circumstances:

- during DNA replication;
- and when genes are activated.

When chromosomes are replicated, during the S phase of the cell cycle, histones are transiently acetylated. **Figure 23.11** shows that this acetylation occurs before the histones are incorporated into nucleosomes. We know that histones H4 and H3 are acetylated at the stage when they are associated with one another in the $H3_2 \cdot H4_2$ tetramer. The tetramer is then incorporated into nucleosomes. Quite soon after, the acetyl groups are removed.

The importance of the acetylation is indicated by the fact that preventing acetylation of both histones H3 and H4 during replication causes loss of viability in yeast. The two histones are redundant as substrates, since yeast can manage perfectly well so long as they can acetylate either one of these histones during S phase. There are two possible roles for the acetylation: it could be needed for the histones to be recognized by factors that incorporate them into nucleosomes; or it could be required for the assembly and/or structure of the new nucleosome.

The factors that are known to be involved in chromatin assembly do not distinguish between acetylated and nonacetylated histones, suggesting that the modification is more likely to be required for subsequent interactions. It has been thought for a long time that acetylation might be needed to help control protein-protein interactions that occur as histones are incorporated into nucleosomes. Some evidence for such a role is that the yeast SAS histone acetylase complex binds to chromatin assembly complexes at the replication fork, where it acetylates ^{16}Lys of histone H4. This may be part of the system that establishes the histone acetylation patterns after replication.

Outside of S phase, acetylation of histones in chromatin is generally correlated with the state of gene expression. The correlation was first noticed because histone acetylation is increased in a domain containing active genes, and acetylated chromatin is more sensitive to DNAase I and (possibly) to micrococcal nuclease. **Figure 23.12** shows that this involves the acetylation of histone tails in nucleosomes. We now know that this occurs largely because of acetylation of the nucleosomes in the vicinity of the promoter when a gene is activated.

In addition to events at individual promoters, widescale changes in acetylation occur on sex chromosomes. This is part of the mechanism by which the activities of genes on the X chromosome are altered to compensate for the presence of two X chromosomes in one species but only one X chromosome (in addition to the Y chromosome) in the other species (see *23.17 X chromosomes undergo global changes*). The inactive X chromosome in female mammals has underacetylated H4. The superactive X chromosome in *Drosophila* males has increased acetylation of H4. This suggests that the presence of acetyl groups may be a prerequisite for a less condensed, active structure. In male *Drosophila*, the X chromosome is acetylated specifically at ^{16}Lys of histone H4. The HAT that is responsible is an enzyme called MOF that is recruited to the chromosome as part of a large protein complex. This "dosage compensation" complex is responsible for introducing general changes in the X chromosome that enable it to be more highly expressed. The increased acetylation is only one of its activities.

23.7 Acetylases are associated with activators

Key Concepts

- Deacetylated chromatin may have a more condensed structure.
- Transcription activators are associated with histone acetylase activities in large complexes.
- Histone acetylases vary in their target specificity.
- Acetylation could affect transcription in a quantitative or qualitative way.

Acetylation is reversible. Each direction of the reaction is catalyzed by a specific type of enzyme. Enzymes that can acetylate histories are called histone acetyltransferases or HATs; the acetyl groups are removed by histone deacetylases or HDACs. There are two groups of HAT enzymes: group A describes those that are involved with transcription; group B describes those involved with nucleosome assembly.

Two inhibitors have been useful in analyzing acetylation. Trichostatin and butyric acid inhibit histone deacetylases, and cause acetylated nucleosomes to accumulate. The use of these inhibitors has supported the general view that acetylation is associated with gene expression; in fact, the ability of butyric acid to cause changes in chromatin resembling those found upon gene activation was one of the first indications of the connection between acetylation and gene activity.

The breakthrough in analyzing the role of histone acetylation was provided by the characterization of the acetylating and deacetylating enzymes, and their association with other proteins that are involved in specific events of activation and repression. A basic change in our view of histone acetylation was caused by the discovery that HATs are not necessarily dedicated enzymes associated with chromatin: rather it turns out that known activators of transcription have HAT activity.

The connection was established when the catalytic subunit of a group A HAT was identified as a homologue of the yeast regulator protein GCN5. Then it was shown that GCN5 itself has HAT activity (with histones H3 and H4 as substrates). GCN5 is part of an adaptor complex that is necessary for the interaction between certain enhancers and their target promoters. Its HAT activity is required for activation of the target gene.

This enables us to redraw our picture for the action of coactivators as shown in **Figure 23.13,** where RNA polymerase is bound at a hypersensitive site and coactivators are acetylating histones on the nucleosomes in the vicinity. Many examples are now known of interactions of this type.

GCN5 leads us into one of the most important acetylase complexes. In yeast, GCN5 is part of the 1.8 MDa SAGA complex, which contains several proteins that are involved in transcription. Among these proteins are several TAF_{II}s. Also, the $TAF_{II}145$ subunit of $TF_{II}D$ is an acetylase. There are some functional overlaps between $TF_{II}D$ and SAGA, most notably that yeast can manage with either $TAF_{II}145$ or GCN5, but is damaged by the deletion of both. This suggests that an acetylase activity is essential for gene expression, but can be provided by either $TF_{II}D$ or SAGA. As might be expected from the size of the SAGA complex, acetylation is only one of its functions, although its other functions in gene activation are less well characterized.

One of the first general activators to be characterized as an HAT was p300/CBP. (Actually, p300 and CBP are different proteins, but they are so closely related that they are often referred to as a single type of activity.) p300/CBP is a coactivator that links an activator to the basal apparatus (see Figure 22.7). p300/CBP interacts with various activators, including hormone receptors, AP-1 (c-Jun and c-Fos), and MyoD. The

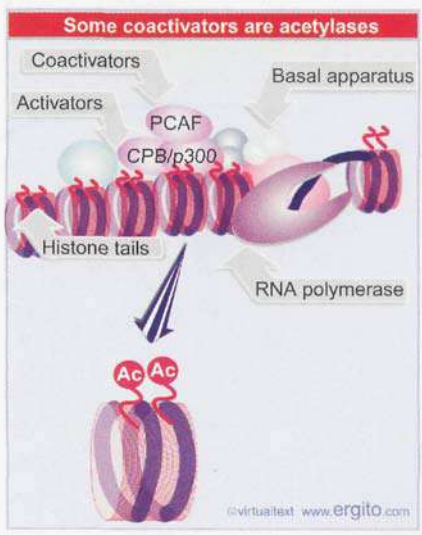


Figure 23.13 Coactivators may have HAT activities that acetylate the tails of nucleosomal histones.

By Book_Crazy [IND]

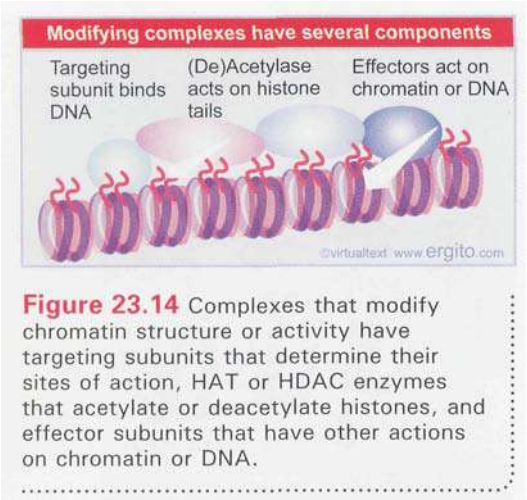


Figure 23.14 Complexes that modify chromatin structure or activity have targeting subunits that determine their sites of action, HAT or HDAC enzymes that acetylate or deacetylate histones, and effector subunits that have other actions on chromatin or DNA.

interaction is inhibited by the viral regulator proteins adenovirus E1A and SV40 T antigen, which bind to p300/CBP to prevent the interaction with transcription factors; this explains how these viral proteins inhibit cellular transcription. (This inhibition is important for the ability of the viral proteins to contribute to the tumorigenic state; see *30.18 Oncoproteins may regulate gene expression).*

p300/CBP acetylates the N-terminal tails of H4 in nucleosomes. Another coactivator, called PCAF, preferentially acetylates H3 in nucleosomes. p300/CBP and PCAF form a complex that functions in transcriptional activation. In some cases yet another HAT is involved: the coactivator ACTR, which functions with hormone receptors, is itself an HAT that acts on H3 and H4, and also recruits both p300/CBP and PCAF to form a coactivating complex. One explanation for the presence of multiple HAT activities in a coactivating complex is that each HAT has a different specificity, and that multiple different acetylation events are required for activation.

A general feature of acetylation is that an HAT is part of a large complex. **Figure 23.14** shows a simplified model for their behavior. Typically the complex will contain a targeting subunit(s) that determines the binding sites on DNA. This determines the target for the HAT. The complex also contains effector subunits that affect chromatin structure or act directly on transcription. Probably at least some of the effectors require the acetylation event in order to act. Deacetylation, catalyzed by an HDAC, may work in a similar way.

Acetylation occurs at both replication (when it is transient) and at transcription (when it is maintained while the gene is active). Is it playing the same role in each case? One possibility is that the important effect is on nucleosome structure. Acetylation may be necessary to "loosen" the nucleosome core. At replication, acetylation of histones could be necessary to allow them to be incorporated into new cores more easily. At transcription, a similar effect could be necessary to allow a related change in structure, possibly even to allow the histone core to be displaced from DNA. Alternatively, acetylation could generate binding sites for other proteins that are required for transcription. In either case, deacetylation would reverse the effect.

Is the effect of acetylation quantitative or qualitative? One possibility is that a certain number of acetyl groups are required to have an effect, and the exact positions at which they occur are largely irrelevant. An alternative is that individual acetylation events have specific effects. We might interpret the existence of complexes containing multiple HAT activities in either way—if individual enzymes have different specificities, we may need multiple activities either to acetylate a sufficient number of different positions or because the individual events are necessary for different effects upon transcription. At replication, it appears, at least with respect to histone H4, that acetylation at any two of three available positions is adequate, favoring a quantitative model in this case. Where chromatin structure is changed to affect transcription, acetylation at specific positions may be important (see *23.15 Heterochromatin depends on interactions with histones).*

23.8 Deacetylases are associated with repressors

Key Concepts

- Deacetylation is associated with repression of gene activity.
- Deacetylases are present in complexes with repressor activity.

By Book_Crazy [IND]

In yeast, mutations in *SIN3* and *Rpd3* behave as though these loci repress a variety of genes. The proteins form a complex with the DNA-binding protein Ume6, which binds to the *URS1* element. The complex represses transcription at the promoters containing *URS1*, as illustrated in **Figure 23.15.** Rpd3 has histone deacetylase activity; we do not know whether the function of Sin3 is just to bring Rpd3 to the promoter or whether it has an additional role in repression.

A similar system for repression is found in mammalian cells. The bHLH family of transcription regulators includes activators that function as heterodimers, including MyoD (see *22.15 Helix-loop-helix proteins interact by combinatorial association).* It also includes repressors, in particular the heterodimer Mad : Max, where Mad can be any one of a group of closely related proteins. The Mad : Max heterodimer (which binds to specific DNA sites) interacts with a homologue of Sin3 (called mSin3 in mouse and hSin3 in man). mSin3 is part of a repressive complex that includes histone binding proteins and the histone deacetylases HDAC1 and HDAC2. Deacetylase activity is required for repression. The modular nature of this system is emphasized by other means of employment: a corepressor (SMRT), which enables retinoid hormone receptors to repress certain target genes, functions by binding mSin3, which in turns brings the HDAC activities to the site. Another means of bringing HDAC activities to the site may be a connection with MeCP2, a protein that binds to methylated cytosines (see *21.19 CpG islands are regulatory targets).*

Absence of histone acetylation is also a feature of heterochromatin. This is true of both constitutive heterochromatin (typically involving regions of centromeres or telomeres) and facultative heterochromatin (regions that are inactivated in one cell although they may be active in another). Typically the N-terminal tails of histones H3 and H4 are not acetylated in heterochromatic regions.

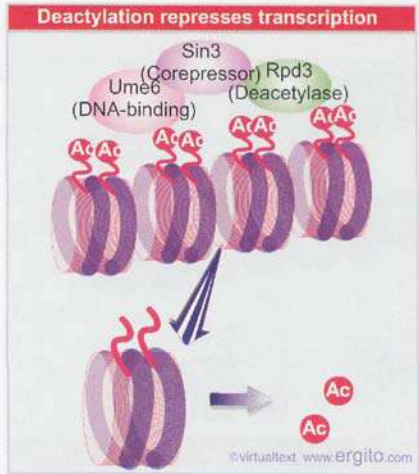


Figure 23.15 A repressor complex contains three components: a DNA binding subunit, a corepressor, and a histone deacetylase.

23.9 Methylation of histones and DNA is connected

Key Concepts

- Methylation of both DNA and histones is a feature of inactive chromatin.
- The two types of methylation event may be connected.

Methylation of both histones and DNA is associated with inactivity. Sites that are methylated in histones include two lysines in the tail of H3 and an arginine in the tail of H4.

Methylation of H3 ^{9}Lys is a feature of condensed regions of chromatin, including heterochromatin as seen in bulk and also smaller regions that are known not to be expressed. The histone methyltransferase enzyme that targets this lysine is called SUV39H1. (We see the origin of this peculiar name in *23.14 Some common motifs are found in proteins that modify chromatin).* Its catalytic site has a region called the SET domain. Other histone methyltransferases act on arginine. In addition, methylation may occur on ^{79}Lys in the globular core region of H3; this may be necessary for the formation of heterochromatin at telomeres.

Most of the methylation sites in DNA are CpG islands (see *21.19 CpG islands are regulatory targets).* CpG sequences in heterochromatin are usually methylated. Conversely, it is necessary for the CpG islands located in promoter regions to be unmethylated in order for a gene to be expressed (see *21.18 Gene expression is associated with demethylation).*

By Book_Crazy [IND]

Methylation of DNA and methylation of histories may be connected. Some histone methyltransferase enzymes contain potential binding sites for the methylated CpG doublet, raising the possibility that a methylated DNA sequence may cause a histone methyltransferase to bind. A possible connection in the opposite direction is indicated by the fact that in the fungus *Neurospora*, the methylation of DNA is prevented by a mutation in a gene coding for a histone methylase that acts on ^{9}Lys of histone H3. This suggests that methylation of the histone is a signal involved in recruiting the DNA methylase to chromatin. The important point is not the detailed order of events—which remains to be worked out—but the fact that one type of modification can be the trigger for another.

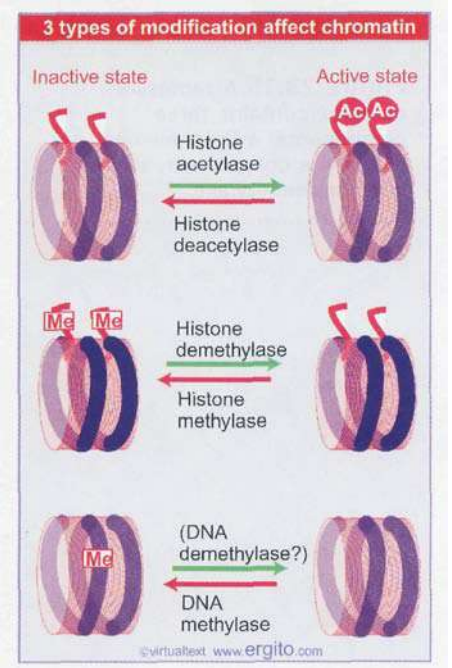


Figure 23.16 Acetylation of histones activates chromatin, and methylation of DNA and histones

23.10 Chromatin states are interconverted by modification

Key Concepts

- Acetylation of histones is associated with gene activation.
- Methylation of DNA and of histones is associated with heterochromatin.

Figure 23.16 summarizes three types of differences that are found between active chromatin and inactive chromatin:

- Active chromatin is acetylated on the tails of histones H3 and H4.
- Inactive chromatin is methylated on ^{9}Lys of histone H3.
- Inactive chromatin is methylated on cytosines of CpG doublets.

The reverse types of events occur if we compare the activation of a promoter with the generation of heterochromatin. The actions of the enzymes that modify chromatin ensure that activating events are mutually exclusive with inactivating events. Methylation of H3 ^{9}Lys and acetylation of H3 ^{14}Lys are mutually antagonistic.

Acetylases and deacetylases may trigger the initiating events. Deacetylation allows methylation to occur, which causes formation of a heterochromatic complex (see *23.15 Heterochromatin depends on interactions with histones*). Acetylation marks a region as active (see next section).

23.11 Promoter activation involves an ordered series of events

Key Concepts

- The remodeling complex may recruit the acetylating complex.
- Acetylation of histones may be the event that maintains the complex in the activated state.

How are acetylases (or deacetylases) recruited to their specific targets? As we have seen with remodeling complexes, the process is likely to be indirect. A sequence-specific activator (or repressor) may interact with a component of the acetylase (or deacetylase) complex to recruit it to a promoter.

There may also be direct interactions between remodeling complexes and histone-modifying complexes. Binding by the SWI/SNF remodeling complex may lead in turn to binding by the SAGA acetylase

complex. Acetylation of histories may then in fact stabilize the association with the SWI/SNF complex, making a mutual reinforcement of the changes in the components at the promoter.

We can connect all of the events at the promoter into the series summarized in **Figure 23.17**. The initiating event is binding of a sequence-specific component (which is able to find its target DNA sequence in the context of chromatin). This recruits a remodeling complex. Changes occur in nucleosome structure. An acetylating complex binds, and the acetylation of target histones provides a covalent mark that the locus has been activated.

Modification of DNA also occurs at the promoter. Methylation of cytosine at CpG doublets is associated with gene inactivity (see *21.18 Gene expression is associated with demethylation*). The basis for recognition of DNA as a target for methylation is not very well established (see *23.20 DNA methylation is responsible for imprinting*).

It is clear that chromatin remodeling at the promoter requires a variety of changes that affect nucleosomes, including acetylation, but what changes are required within the gene to allow an RNA polymerase to traverse it? We know that RNA polymerase can transcribe DNA *in vitro* at rates comparable to the *in vivo* rate (~25 nucleotides per second) only with template of free DNA. Several proteins have been characterized for their abilities to improve the speed with which RNA polymerase transcribes chromatin *in vivo*. The common feature is that they act on chromatin. A current model for their action is that they associate with RNA polymerase and travel with it along the template, modifying nucleosome structure by acting on histones. Among these factors are histone acetylases. One possibility is that the first RNA polymerase to transcribe a gene is a pioneer polymerase carrying factors that change the structure of the transcription unit so as to make it easier for subsequent polymerases.

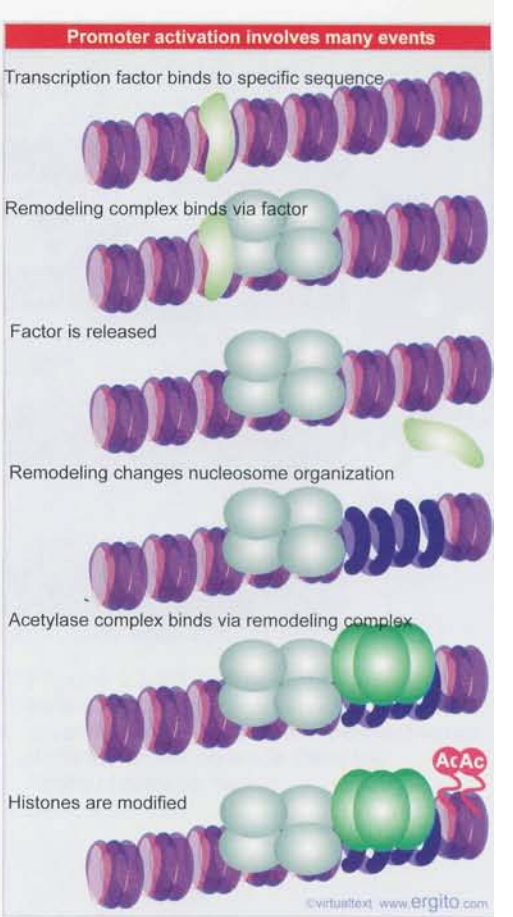


Figure 23.17 Promoter activation involves binding of a sequence-specific activator, recruitment and action of a remodeling complex, and recruitment and action of an acetylating complex.

23.12 Histone phosphorylation affects chromatin structure

Key Concepts

- At least two histones are targets for phosphorylation, possibly with opposing effects.

Histones are phosphorylated in two circumstances:

- cyclically during the cell cycle;
- and in association with chromatin remodeling.

It is has been known for a very long time that histone H1 is phosphorylated at mitosis, and more recently it was discovered that H1 is an extremely good substrate for the Cdc2 kinase that controls cell division. This led to speculations that the phosphorylation might be connected with the condensation of chromatin, but so far no direct effect of this phosphorylation event has been demonstrated, and we do not know whether it plays a role in cell division (see *29.7 Protein phosphorylation and dephosphorylation control the cell cycle*).

Loss of a kinase that phosphorylates histone H3 on ^{10}Ser has devastating effects on chromatin structure. **Figure 23.18** compares the usual extended structure of the polytene chromosome set of *D. melanogaster* (upper photograph) with the structure that is found in a null mutant that has no J1L-1 kinase (lower photograph). The absence of JIL-1 is lethal, but the chromosomes can be visualized in the larvae before they die.

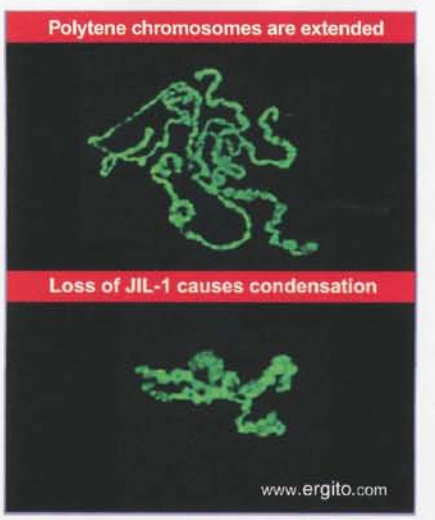


Figure 23.18 Polytene chromosomes of flies that have no JIL-1 kinase have abnormal polytene chromosomes that are condensed instead of extended. Photograph kindly provided by Kristen M. Johansen.

By Book_Crazy [IND]

The cause of the disruption of structure is most likely the failure to phosphorylate histone H3 (of course, JIL-1 may also have other targets). This suggests that H3 phosphorylation is required to generate the more extended chromosome structure of euchromatic regions. Evidence supporting the idea that JIL-1 acts directly on chromatin is that it associates with the complex of proteins that binds to the X chromosome to increase its gene expression in males (see *23.17 X chromosomes undergo global changes).*

This leaves us with somewhat conflicting impressions of the roles of histone phosphorylation. If it is important in the cell cycle, it is likely to be as a signal for condensation. Its effect in chromatin remodeling appears to be the opposite. It is of course possible that phosphorylation of different histones, or even of different amino acid residues in one histone, has opposite effects on chromatin structure.

23.13 Heterochromatin propagates from a nucleation event

Key Concepts

- Heterochromatin is nucleated at a specific sequence and the inactive structure propagates along the chromatin fiber.
- Genes within regions of heterochromatin are inactivated.
- Because the length of the inactive region varies from cell to cell, inactivation of genes in this vicinity causes position effect variegation.
- Similar spreading effects occur at telomeres and at the silent cassettes in yeast mating type.

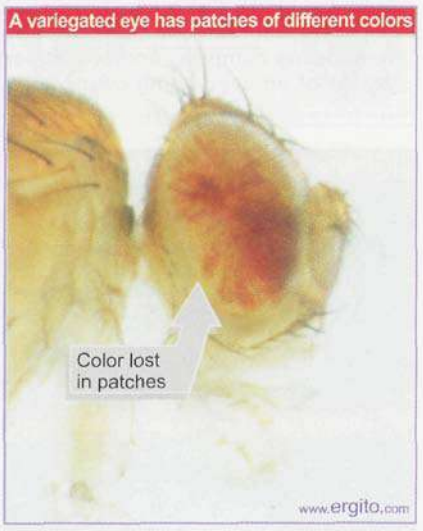


Figure 23.19 Position effect variegation in eye color results when the *white* gene is integrated near heterochromatin. Cells in which *white* is inactive give patches of white eye, while cells in which *white* is active give red patches. The severity of the effect is determined by the closeness of the integrated gene to heterochromatin. Photograph kindly provided by Steve Henikoff.

An interphase nucleus contains both euchromatin and heterochromatin. The condensation state of heterochromatin is close to that of mitotic chromosomes. Heterochromatin is inert. It remains condensed in interphase, is transcriptionally repressed, replicates late in S phase, and may be localized to the nuclear periphery. Centromeric heterochromatin typically consists of satellite DNAs. However, the formation of heterochromatin is not rigorously defined by sequence. When a gene is transferred, either by a chromosomal translocation or by transfection and integration, into a position adjacent to heterochromatin, it may become inactive as the result of its new location, implying that it has become heterochromatic.

Such inactivation is the result of an epigenetic effect (see *23.22 Epigenetic effects can be inherited).* It may differ between individual cells in an animal, and results in the phenomenon of position effect variegation (PEV), in which genetically identical cells have different phenotypes. This has been well characterized in *Drosophila.* **Figure 23.19** shows an example of position effect variegation in the fly eye, in which some regions lack color while others are red, because the *white* gene is inactivated by adjacent heterochromatin in some cells, while it remained active in other cells.

The explanation for this effect is shown in **Figure 23.20.** Inactivation spreads from heterochromatin into the adjacent region for a variable distance. In some cells it goes far enough to inactivate a nearby gene, but in others it does not. This happens at a certain point in embryonic development, and after that point the state of the gene is inherited by all the progeny cells. Cells descended from an ancestor in which the gene was inactivated form patches corresponding to the phenotype of loss-of-function (in the case *of white,* absence of color).

The closer a gene lies to heterochromatin, the higher the probability that it will be inactivated. This suggests that the formation of heterochromatin may be a two-stage process: a *nucleation* event occurs at a specific sequence; and then the inactive structure *propagates* along the chromatin

fiber. The distance for which the inactive structure extends is not precisely determined, and may be stochastic, being influenced by parameters such as the quantities of limiting protein components. One factor that may affect the spreading process is the activation of promoters in the region; an active promoter may inhibit spreading.

Genes that are closer to heterochromatin are more likely to be inactivated, and will therefore be inactive in a greater proportion of cells. On this model, the boundaries of a heterochromatic region might be terminated by exhausting the supply of one of the proteins that is required.

The effect of telomeric silencing in yeast is analogous to position effect variegation in *Drosophila;* genes translocated to a telomeric location show the same sort of variable loss of activity. This results from a spreading effect that propagates from the telomeres.

A second form of silencing occurs in yeast. Yeast mating type is determined by the activity of a single active locus *(MAT),* but the genome contains two other copies of the mating type sequences *{HML* and *HMR),* which are maintained in an inactive form. The silent loci *HML* and *HMR* share many properties with heterochromatin, and could be regarded as constituting regions of heterochromatin in miniature (see *18.7 Silent cassettes at HML and HMR are repressed).*

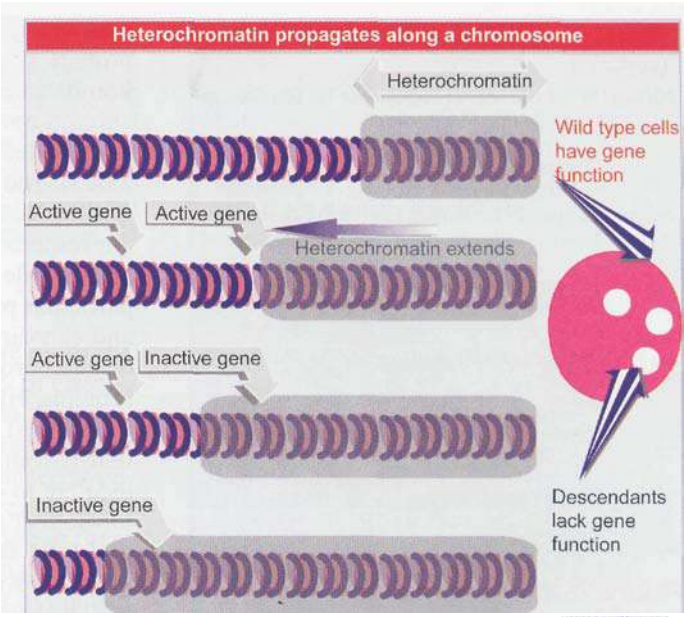


Figure 23.20 Extension of heterochromatin inactivates genes. The probability that a gene will be inactivated depends on its distance from the heterochromatin region.

23.14 Some common motifs are found in proteins that modify chromatin

Key Concepts

- The chromo domain is found in several chromatin proteins that have either activating or repressing effects on gene expression.
- *The SET domain is part of the catalytic site of protein* methyltransferases.

Our insights into the molecular mechanisms for controlling the structure of chromatin start with mutants that affect position effect variegation. Some 30 genes have been identified in *Drosophila.* They are named systematically as *Su(var)* for genes whose products act to suppress variegation and *E(var)* for genes whose products enhance variegation. Remember that the genes were named for the behavior of the mutant loci. Mutations that suppress variegation lie in genes whose products are needed for the formation of heterochromatin. They include enzymes that act on chromatin, such as histone deacetylases, and proteins that are localized to heterochromatin. Mutations that enhance variegation lie in genes whose products are needed to activate gene expression. They include members of the SWI/SNF complex. We see immediately from these properties that modification of chromatin structure is important for controlling the formation of heterochromatin. The universality of these mechanisms is indicated by the fact that many of these loci have homologues in yeast that display analogous properties. Some of the homologues in *S. pombe* are *clr* (cryptic loci regulator) genes, in which mutations affect silencing.

Many of the Su(var) and E(var) proteins have a common protein motif of 60 amino acids called the chromo domain. The fact that this domain is found in proteins of both groups suggests that it represents a motif that participates in protein-protein interactions with targets in chromatin.

Among the Su(var) proteins is HP1 (heterochromatin protein 1). This was originally identified as a protein that is localized to heterochromatin

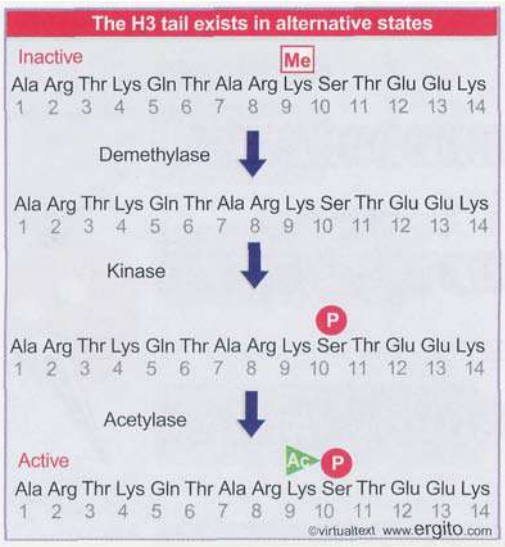


Figure 23.21 Multiple modifications in the H3 tail affect chromatin activity.

by staining polytene chromosomes with an antibody directed against the protein. It was later shown to be the product of the gene *Su(var)2-5*. Its homologue in the yeast *S. pombe* is coded by *swi6*. HP1 contains a chromo domain near the N-terminus, and another domain that is related to it, called the chromo-shadow domain, at the C-terminus (see Figure 23.23). The importance of the chromo domain is indicated by the fact that it is the location of many of the mutations in HP1. The chromo domain(s) are responsible for targeting the protein to heterochromatin. They play a similar role in other proteins, although the individual chromo domains in particular proteins may have different detailed specificities for targeting, and can direct proteins to either heterochromatin or euchromatin. The original protein identified as HP1 is now called HP1α, since two related proteins, HP1β and HP1γ, have since been found.

Su(var)3-9 has a chromo domain and also a SET domain, a motif that is found in several Su(var) proteins. Its mammalian homologues localize to centromeric heterochromatin. It is the histone methyltransferase that acts on ^{9}Lys of histone H3 (see *23.9 Methylation of histories and DNA is connected)*. The SET domain is part of the active site, and in fact is a marker for the methylase activity.

The bromo domain is found in a variety of proteins that interact with chromatin, including histone acetylases. The crystal structure shows that it has a binding site for acetylated lysine. The bromo domain itself recognizes only a very short sequence of 4 amino acids including the acetylated lysine, so specificity for target recognition must depend on interactions involving other regions. Besides the acetylases, the bromo domain is found in a range of proteins that interact with chromatin, including components of the transcription apparatus. This implies that it is used to recognize acetylated histones, which means that it is likely to be found in proteins that are involved with gene activation.

Although there is a general correlation in which active chromatin is acetylated while inactive chromatin is methylated on histones, there are some exceptions to the rule. The best characterized is that acetylation of ^{12}Lys of H4 is associated with heterochromatin.

Multiple modifications may occur on the same histone tail, and one modification may influence another. Phosphorylation of a lysine at one position may be necessary for acetylation of a lysine at another position. **Figure 23.21** shows the situation in the tail of H3, which can exist in either of two alternative states. The inactive state has Methyl-^{9}Lys. The active state has Acetyl-^{9}Lys and Phospho-^{10}Ser. These states can be maintained over extended regions of chromatin. The phosphorylation of ^{10}Ser and the methylation of ^{9}Lys are mutually inhibitory, suggesting the order of events shown in the figure. This situation may cause the tail to flip between the active and active states.

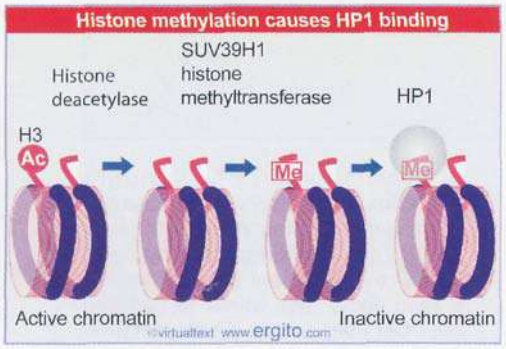


Figure 23.22 SUV39H1 is a histone methyltransferase that acts on ^{9}Lys of histone H3. HP1 binds to the methylated histone.

23.15 Heterochromatin depends on interactions with histones

i Key Concepts

- HP1 is the key protein in forming mammalian heterochromatin, and acts by binding to methylated H3 histone.
- RAP1 initiates formation of heterochromatin in yeast by binding to specific target sequences in DNA.
- The targets of RAP1 include telomeric repeats and silencers at *HML* and *HMR*.
- RAP1 recruits SIR3/SIR4, which interact with the N-terminal tails of H3 and H4.

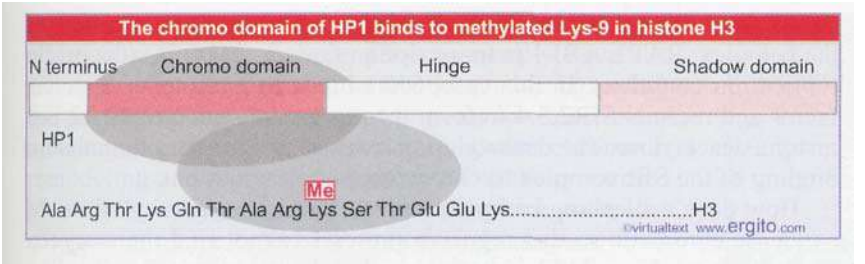


Figure 23.23 Methylation of histone H creates a binding site for HP1.

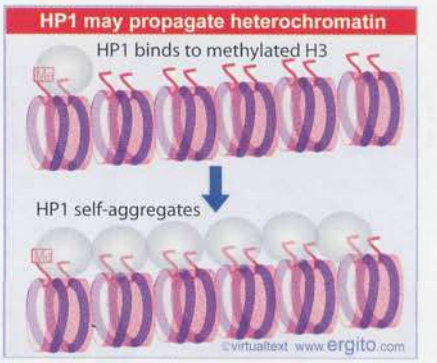


Figure 23.24 Binding of HP1 to methylated histone H3 forms a trigger for silencing because further molecules of HP1 aggregate on the nucleosome chain.

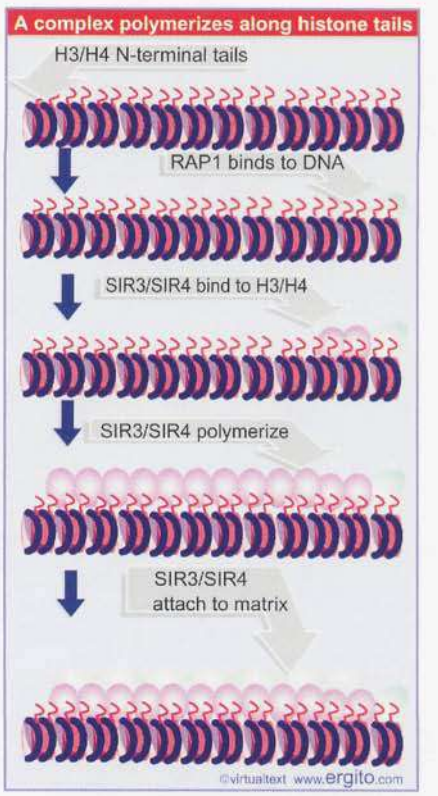


Figure 23.25 Formation of heterochromatin is initiated when RAP1 binds to DNA. SIR3/4 bind to RAP1 and also to histones H3/H4. The complex polymerizes along chromatin and may connect telomeres to the nuclear matrix.

Inactivation of chromatin occurs by the addition of proteins to the nucleosomal fiber. The inactivation may be due to a variety of effects, including condensation of chromatin to make it inaccessible to the apparatus needed for gene expression, addition of proteins that directly block access to regulatory sites, or proteins that directly inhibit transcription.

Two systems that have been characterized at the molecular level involve HP 1 in mammals and the SIR complex in yeast. Although there are no detailed similarities between the proteins involved in each system, the general mechanism of reaction is similar: the points of contact in chromatin are the N-terminal tails of the hi stones.

Mutation of a deacetylase that acts on the H3 Ac-^{14}Lys prevents the methylation at ^{9}Lys. H3 that is methylated at ^{9}Lysbinds the protein HP1 via the chromo domain. This suggests the model for initiating formation of heterochromatin shown in **Figure 23.22.** First the deacetylase acts to remove the modification at ^{14}Lys. Then the SUV39H1 methylase acts on the histone H3 tail to create the methylated signal to which HP1 will bind. **Figure 23.23** expands the reaction to show that the interaction occurs between the chromo domain and the methylated lysine. This is a trigger for forming inactive chromatin. **Figure 23.24** shows that the inactive region may then be extended by the ability of further HP1 molecules to interact with one another.

The existence of a common basis for silencing in yeast is suggested by its reliance on a common set of genetic loci. Mutations in any one of a number of genes cause *HML* and *HMR* to become activated, and also relieve the inactivation of genes that have been integrated near telomeric heterochromatin. The products of these loci therefore function to maintain the inactive state of both types of heterochromatin.

Figure 23.25 proposes a model for actions of these proteins. Only one of them is a sequence-specific DNA-binding protein. This is RAP1, which binds to the $C_{1-3}A$ repeats at the telomeres, and also binds to the ris-acting silencer elements that are needed for repression of *HML* and *HMR*. The proteins SIR3 and SIR4 interact with RAP1 and also with one another (they may function as a heteromultimer). SIR3/SIR4 interact with the N-terminal tails of the histones H3 and H4. (In fact, the first evidence that histones might be involved directly in formation of heterochromatin was provided by the discovery that mutations abolishing silencing at *HML/HMR* map to genes coding for H3 and H4).

RAP 1 has the crucial role of identifying the DNA sequences at which heterochromatin forms. It recruits SIR3/SIR4, and they interact directly with the histones H3/H4. Once SIR3/SIR4 have bound to histones H3/H4, the complex may polymerize further, and spread along the chromatin fiber. This may inactivate the region, either because coating with SIR3/SIR4 itself has an inhibitory effect, or because binding to histones H3/H4 induces some further change in structure. We do not know what limits the spreading of the complex. The C-terminus of SIR3 has a similarity to nuclear lamin proteins (constituents of the nuclear matrix) and may be responsible for tethering heterochromatin to the nuclear periphery.

A similar series of events forms the silenced regions at HMR and HML (see also *18.7 Silent cassettes at HML and HMR are repressed).*

Three sequence-specific factors are involved in triggering formation of the complex: RAP1, ABF1 (a transcription factor), and ORC (the origin replication complex). In this case, SIR1 binds to a sequence-specific factor and recruits SIR2,3,4 to form the repressive structure. SIR2 is a histone deacetylase. The deacetylation reaction is necessary to maintain binding of the SIR complex to chromatin.

How does a silencing complex repress chromatin activity? It could condense chromatin so that regulator proteins cannot find their targets. The simplest case would be to suppose that the presence of a silencing complex is mutually incompatible with the presence of transcription factors and RNA polymerase. The cause could be that silencing complexes block remodeling (and thus indirectly prevent factors from binding) or that they directly obscure the binding sites on DNA for the transcription factors. However, the situation may not be this simple, because transcription factors and RNA polymerase can be found at promoters in silenced chromatin. This could mean that the silencing complex prevents the factors from working rather than from binding as such. In fact, there may be competition between gene activators and the repressing effects of chromatin, so that activation of a promoter inhibits spread of the silencing complex.

Another specialized chromatin structure forms at the centromere. Its nature is suggested by the properties of an *S. cerevisiae* mutation, *cse4*, that disrupts the structure of the centromere. Cse4p is a protein that is related to histone H3. A mammalian centromeric protein, CENP-A, has a related sequence. Genetic interactions between *cse4* and CDE-II, and between *cse4* and a mutation in the H4 histone gene, suggest that a histone octamer may form around a core of Cse4p-H4, and then the centromeric complexes CBF1 and CBF3 may attach to form the centromere.

The centromere may then be associated with the formation of heterochromatin in the region. In human cells, the centromere-specific protein CENP-B is required to initiate modifications of histone H3 (deacetylation of ^{9}Lys and ^{14}Lys, followed by methylation of ^{9}Lys) that trigger an association with the protein Swi6 that leads to the formation of heterochromatin in the region.

23.16 Polycomb and trithorax are antagonistic repressors and activators

Key Concepts

- Polycomb group proteins (Pc-G) perpetuate a state of repression through cell divisions.
- The PRE is a DNA sequence that is required for the action of Pc-G.
- The PRE provides a nucleation center from which Pc-G proteins propagate an inactive structure.
- No individual Pc-G protein has yet been found that can bind the PRE.
- Trithorax group proteins antagonize the actions of the Pc-G.

Heterochromatin provides one example of the specific repression of chromatin. Another is provided by the genetics of homeotic genes in *Drosophila*, which have led to the identification of a protein complex that may maintain certain genes in a repressed state. *Pc* mutants show transformations of cell type that are equivalent to gain-of-function mutations in the genes *Antennapedia* (*Antp*) or *Ultrabithorax*, because these genes are expressed in tissues in which usually they are repressed. This implicates *Pc* in regulating transcription. Furthermore,

Pc is the prototype for a class of loci called the *Pc* group (*Pc-G*); mutations in these genes generally have the same result of derepressing homeotic genes, suggesting the possibility that the group of proteins has some common regulatory role. A connection between chromatin remodeling and repression is indicated by the properties of *brahma*, a fly counterpart to *SWI2*, which codes for component of the SWI/SNF remodeling complex. Loss of *brahma* function suppresses mutations in *Polycomb.*

Consistent with the pleiotropy of *Pc* mutations, Pc is a nuclear protein that can be visualized at ~80 sites on polytene chromosomes. These sites include the *Antp* gene. Another member of the *Pc-G, polyhomeotic,* is visualized at a set of polytene chromosome bands that are identical with those bound by Pc. The two proteins coimmunoprecipitate in a complex of ~2.5 X 10^6 D that contains 10-15 polypeptides. The relationship between these proteins and the products of the ~30 *Pc-G* genes remains to be established; one possibility is that many of these gene products form a general repressive complex that is modified by some of the others for specific loci.

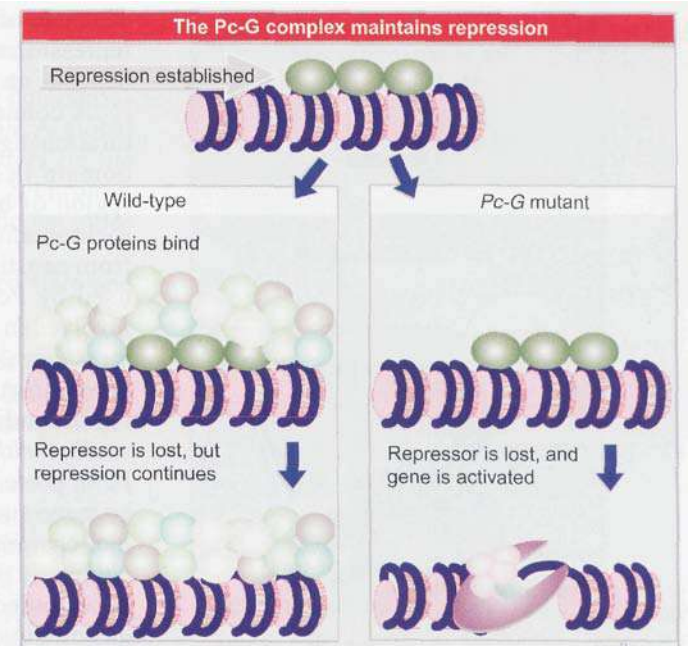


Figure 23.26 Pc-G proteins do not initiate repression, but are responsible for maintaining it.

The Pc-G proteins are not conventional repressors. They are not responsible for determining the initial pattern of expression of the genes on which they act. In the absence of Pc-G proteins, these genes are initially repressed as usual, but later in development the repression is lost without Pc-G group functions. This suggests that *the Pc-G proteins in some way recognize the state of repression when it is established, and they then act to perpetuate it through cell division of the daughter cells.* **Figure 23.26** shows a model in which Pc-G proteins bind in conjunction with a repressor, but the Pc-G proteins remain bound after the repressor is no longer available. This is necessary to maintain repression, so that if Pc-G proteins are absent, the gene becomes activated.

A region of DNA that is sufficient to enable the response to the *Pc-G* genes is called a PRE *{Polycomb* response element). It can be defined operationally by the property that it maintains repression in its vicinity throughout development. The assay for a PRE is to insert it close to a reporter gene that is controlled by an enhancer that is repressed in early development, and then to determine whether the reporter becomes expressed subsequently in the descendants. An effective PRE will prevent such re-expression.

The PRE is a complex structure, ~10 kb. No individual member of the Pc-G proteins has yet been shown to bind to specific sequences in the PRE, so the basis for the assembly of the complex is still unknown. When a locus is repressed by Pc-G proteins, however, the proteins appear to be present over a much larger length of DNA than the PRE itself. Polycomb is found locally over a few kilobases of DNA surrounding a PRE.

This suggests that the PRE may provide a nucleation center, from which a structural state depending on Pc-G proteins may propagate. This model is supported by the observation of effects related to position effect variegation (see Figure 23.20), that is, a gene near to a locus whose repression is maintained by Pc-G may become heritably inactivated in some cells but not others. In one typical situation, crosslinking experiments *in vivo* showed that Pc protein is found over large regions of the *bithorax* complex that are inactive, but the protein is excluded from regions that contain active genes. The idea that this could be due to cooperative interactions within a multimeric complex is supported by the existence of mutations in *Pc* that change its nuclear distribution and abolish the ability of other *Pc-G* members to localize in the nucleus.

The role of Pc-G proteins in maintaining, as opposed to establishing, repression must mean that the formation of the complex at the PRE also depends on the local state of gene expression.

A connection between the Pc-G complex and more general structural changes in chromatin is suggested by the inclusion of a chromo domain in Pc. (In fact, the chromo domain was first identified as a region of homology between Pc and the protein HP1 found in heterochromatin.) Since variegation is caused by the spreading of inactivity from constitutive heterochromatin, it is likely that the chromo domain is used by Pc and HP1 to interact with common components that are involved in inducing the formation of heterochromatic or inactive structures (see *23.14 Some common motifs are found in proteins that modify chromatin).* This model implies that similar mechanisms are used to repress individual loci or to create heterochromatin.

The *trithorax* group *(trxG)* of proteins have the opposite effect to the Pc-G proteins: they act to maintain genes effect

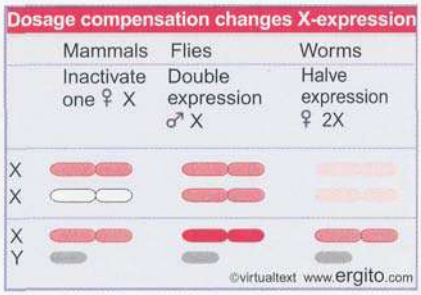


Figure 23.27 Different means of dosage compensation are used to equalize X chromosome expression in male and female.

- In *C. elegans*, the expression of each female X chromosome is halved relative to the expression of the single male X chromosome.

The common feature in all these mechanisms of dosage compensation is that *the entire chromosome is the target for regulation.* A global change occurs that quantitatively affects all of the promoters on the chromosome. We know most about the inactivation of the X chromosome in mammalian females, where the entire chromosome becomes heterochromatic.

The twin properties of heterochromatin are its condensed state and associated inactivity. It can be divided into two types:

- Constitutive heterochromatin contains specific sequences that have no coding function. Typically these include satellite DNAs, and are often found at the centromeres. These regions are invariably heterochromatic because of their intrinsic nature.
- Facultative heterochromatin takes the form of entire chromosomes that are inactive in one cell lineage, although they can be expressed in other lineages. The example *par excellence* is the mammalian X chromosome. The inactive X chromosome is perpetuated in a heterochromatic state, while the active X chromosome is part of the euchromatin. So *identical DNA sequences are involved in both states.* Once the inactive state has been established, it is inherited by descendant cells. This is an example of epigenetic inheritance, because it does not depend on the DNA sequence.

Our basic view of the situation of the female mammalian X chromosomes was formed by the single X hypothesis in 1961. Female mice that are heterozygous for X-linked coat color mutations have a variegated phenotype in which some areas of the coat are wild-type, but others are mutant. **Figure 23.28** shows that this can be explained *if one of the two X chromosomes is inactivated at random in each cell of a small precursor population.* Cells in which the X chromosome carrying the wild-type gene is inactivated give rise to progeny that express only the mutant allele on the active chromosome. Cells derived from a precursor where the other chromosome was inactivated have an active wild-type gene. In the case of coat color, cells descended from a particular precursor stay together and thus form a patch of the same color, creating the pattern of visible variegation. In other cases, individual cells in a population will express one or the other of X-linked alleles; for example, in heterozygotes for the X-linked locus G6PD, any particular red blood cell will express only one of the two allelic forms. (Random inactivation of one X chromosome occurs in eutherian mammals. In marsupials, the choice is directed: it is always the X chromosome inherited from the father that is inactivated.)

Inactivation of the X chromosome in females is governed by the n-1 rule: however many X chromosomes are present, all but one will be inactivated. In normal females there are of course 2 X chromosomes, but in rare cases where nondisjunction has generated a 3X or greater genotype, only one X chromosome remains active. This suggests a general model in which a specific event is limited to one X chromosome and protects it from an inactivation mechanism that applies to all the others.

A single locus on the X chromosome is sufficient for inactivation. When a translocation occurs between the X chromosome and an autosome, this locus is present on only one of the reciprocal products, and only that product can be inactivated. By comparing different translocations, it is possible to map this locus, which is called the *Xic* (X-inactivation center). A cloned region of 450 kb contains all the properties of the *Xic*. When this sequence is inserted as a transgene on to an autosome, the autosome becomes subject to inactivation (in a cell culture system).

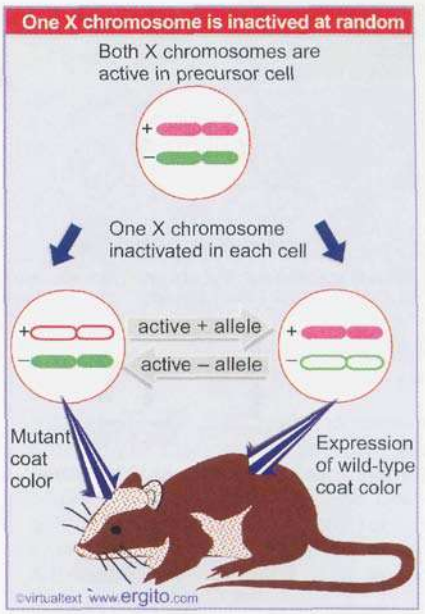


Figure 23.28 X-linked variegation is caused by the random inactivation of one X chromosome in each precursor cell. Cells in which the + allele is on the active chromosome have wild phenotype; but cells in which the – allele is on the active chromosome have mutant phenotype.

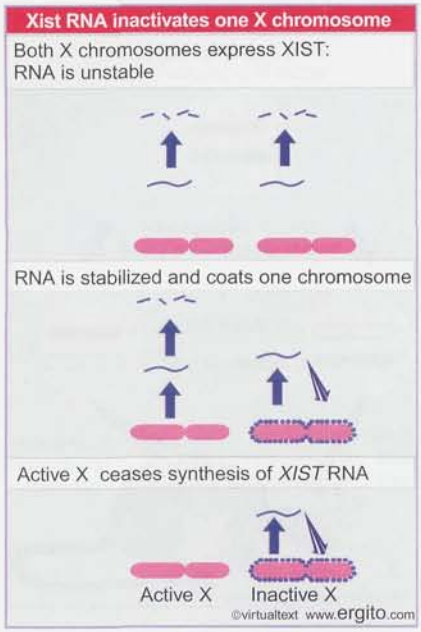


Figure 23.29 X-inactivation involves stabilization of *Xist* RNA, which coats the inactive chromosome.

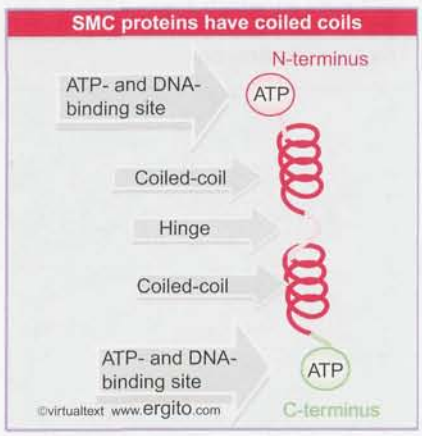


Figure 23.30 An SMP protein has a "Walker module" with an ATP-binding motif and DNA-binding site at each end, connected by coiled coils that are linked by a hinge region.

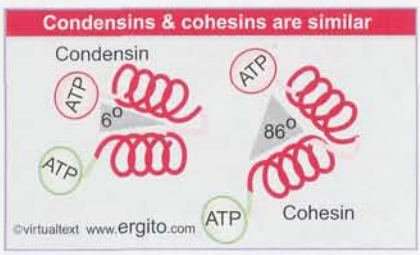


Figure 23.31 The two halves of a condensin are folded back at an angle of 6°. Cohesins have a more open conformation with an angle of 86° between the two halves.

Xic is a *cis*-acting locus that contains the information necessary to count X chromosomes and inactivate all copies but one. Inactivation spreads from *Xic* along the entire X chromosome. When *Xic* is present on an X chromosome-autosome translocation, inactivation spreads into the autosomal regions (although the effect is not always complete).

Xic contains a gene, called *Xist*, that is expressed only on the *inactive* X chromosome. The behavior of this gene is effectively the opposite from all other loci on the chromosome, which are turned off. Deletion of *Xist* prevents an X chromosome from being inactivated. However, it does not interfere with the counting mechanism (because other X chromosomes can be inactivated). So we can distinguish two features of *Xic:* an unidentified element(s) required for counting; and the *Xist* gene required for inactivation.

Figure 23.29 illustrates the role of *Xist* RNA in X-inactivation. *Xist* codes for an RNA that lacks open reading frames. The *Xist* RNA "coats" the X chromosome from which it is synthesized, suggesting that it has a structural role. Prior to X-inactivation, it is synthesized by both female X chromosomes. Following inactivation, the RNA is found only on the inactive X chromosome. The transcription rate remains the same before and after inactivation, so the transition depends on post-transcriptional events.

Prior to X-inactivation, *Xist* RNA decays with a half life of ~2 hr. X-inactivation is mediated by stabilizing the *Xist* RNA on the inactive X chromosome. The *Xist* RNA shows a punctate distribution along the X chromosome, suggesting that association with proteins to form particulate structures may be the means of stabilization. We do not know yet what other factors may be involved in this reaction and how the *Xist* RNA is limited to spreading in *cis* along the chromosome. The characteristic features of the inactive X chromosome, which include a lack of acetylation of histone H4, and methylation of CpG sequences (see *21.19 CpG islands are regulatory targets),* presumably occur later as part of the mechanism of inactivation.

The n–1 rule suggests that stabilization of *Xist* RNA is the "default," and that some blocking mechanism prevents stabilization at one X chromosome (which will be the active X). This means that, although *Xic* is necessary and sufficient for a chromosome to be *inactivated,* the products of other loci may be necessary for the establishment of an *active* X chromosome.

Silencing of *Xist* expression is necessary for the active X. Deletion of the gene for DNA methyltransferase prevents silencing of *Xist,* probably because methylation at the *Xist* promoter is necessary for cessation of transcription.

23.18 Chromosome condensation is caused by condensins

Key Concepts

- SMC proteins are ATPases that include the condensins and the cohesins.
- A heterodimer of SMC proteins associates with other subunits.
- The condensins cause chromatin to be more tightly coiled by introducing positive supercoils into DNA.
- Condensins are responsible for condensing chromosomes at mitosis.
- Chromosome-specific condensins are responsible for condensing inactive X chromosomes in *C. elegans.*

By Book_Crazy [IND]

The structures of entire chromosomes are influenced by interactions with proteins of the SMC (structural maintenance of chromosome) family. They are ATPases that fall into two functional groups. Condensins are involved with the control of overall structure, and are responsible for the condensation into compact chromosomes at mitosis. Cohesins are concerned with connections between sister chromatids that must be released at mitosis (see *29.19 Cohesins hold sister chromatids together).* Both consist of dimers formed by SMC proteins. Condensins form complexes that have a core of the heterodimer SMC2-SMC4 associated with other (non SMC) proteins. Cohesins have a similar organization based on the heterodimeric core of SMC1-SMC3.

Figure 23.30 shows that an SMC protein has a coiled-coil structure in its center, interrupted by a flexible hinge region. Both the amino and carboxyl termini have ATP- and DNA-binding motifs. Different models have been proposed for the actions of these proteins depending on whether they dimerize by intra- or inter-molecular interactions.

Experiments with the bacterial homologues of the SMC proteins suggest that a dimer is formed by an antiparallel interaction between the coiled coils, so that the N-terminus of one subunit bonds to the C-terminus of the other subunit. The existence of a flexible hinge region could allow cohesins and condensins to depend on a different mode of action by the dimer. **Figure 23.31** shows that cohesins have a V-shaped structure, with the arms separated by an 86° angle, whereas condensins are more sharply bent back, with only 6° between the arms. This enables cohesins to hold sister chromatids together, while condensins instead condense an individual chromosome. **Figure 23.32** shows that a cohesin could take the form of an extended dimer that cross-links two DNA molecules. **Figure 23.33** shows that a condensin could take the form of a V-shaped dimer—essentially bent at the hinge—that pulls together distant sites on the same DNA molecule, causing it to condense.

An alternative model is suggested by experiments to suggest that the yeast proteins dimerize by intramolecular interactions, that is, a homodimer is formed solely by interaction between two identical subunits. Dimers of two different proteins (in this case, SMC1 and SMC3) may then interact at both their head and hinge regions to form a circular structure as illustrated in **Figure 23.34.** Instead of binding directly to DNA, a structure of this type could hold DNA molecules together by encircling them.

Visualization of mitotic chromosomes shows that condensins are located all along the length of the chromosome, as can be seen in **Figure 23.35.** (By contrast, cohesins are found at discrete locations; see Figure 29.34).

The condensin complex was named for its ability to cause chromatin to condense *in vitro.* It has an ability to introduce positive supercoils into DNA in an action that uses hydrolysis of ATP and depends on the presence of topoisomerase I. This ability is controlled by the phosphorylation of the non-SMC subunits, which occurs at mitosis. We do not know yet how this connects with other modifications of chromatin, for example, the phosphorylation of histones. The activation of the condensin complex specifically at mitosis makes it questionable whether it is also involved in the formation of interphase heterochromatin.

Global changes occur in other types of dosage compensation. In *Drosophila,* a complex of proteins is found in males, where it localizes on the X chromosome. In *C. elegans,* a protein complex associates with both X chromosomes in XX embryos, but the protein components remain diffusely distributed in the nuclei of XO embryos. The protein complex contains an SMC core, and is similar to the condensin complexes that are associated with mitotic chromosomes in other species. This suggests that it has a structural role in causing the chromosome to take up a more condensed, inactive state. Multiple

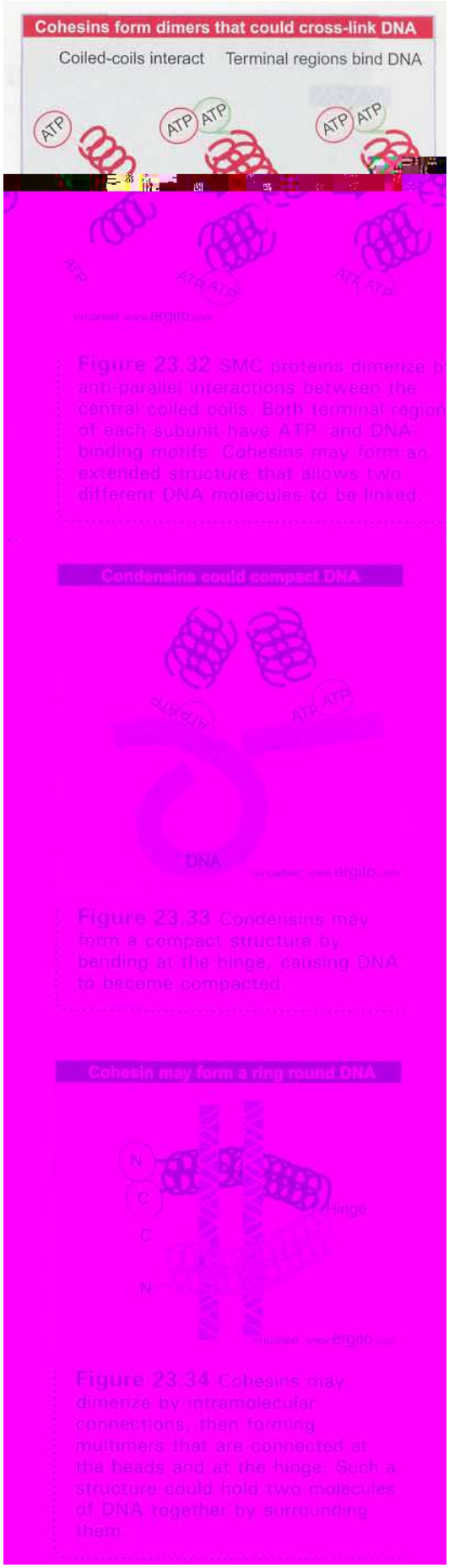


Figure 23.32 SMC proteins dimerize by anti-parallel interactions between the central coiled coils. Both terminal regions of each subunit have ATP- and DNA-binding motifs. Cohesins may form an extended structure that allows two different DNA molecules to be linked.

Figure 23.33 Condensins may form a compact structure by bending at the hinge, causing DNA to become compacted.

Figure 23.34 Cohesins may dimerize by intramolecular connections, then forming multimers that are connected at the heads and at the hinge. Such a structure could hold two molecules of DNA together by surrounding them.

Figure 23.35 Condensins are located along the entire length of a mitotic chromosome. DNA is red; condensins are yellow. Photograph kindly provided by Ana Losada and Tatsuya Hirano.

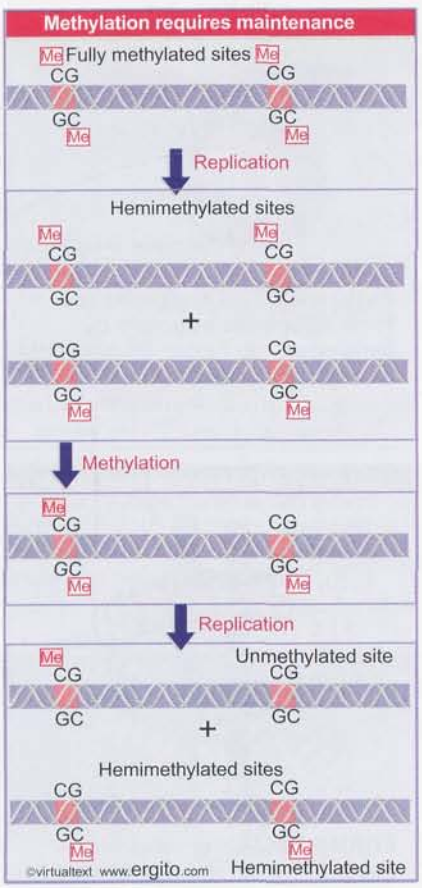


Figure 23.36 The state of methylated sites could be perpetuated by an enzyme that recognizes only hemimethylated sites as substrates.

sites on the X chromosome may be needed for the complex to be fully distributed along it.

Changes affecting all the genes on a chromosome, either negatively (mammals and *C. elegans)* or positively (*Drosophila*) are therefore a common feature of dosage compensation. However, the components of the dosage compensation apparatus may vary as well as the means by which it is localized to the chromosome, and of course its mechanism of action is different in each case.

23.19 DNA methylation is perpetuated by a maintenance methylase

Key Concepts

- Most methyl groups in DNA are found on cytosine on both strands of the CpG doublet.
- Replication converts a fully methylated site to a hemi-methylated site.
- Hemi-methylated sites are converted to fully methylated sites by a maintenance methylase.

Methylation of DNA occurs at specific sites. In bacteria, it is associated with identifying the particular bacterial strain, and also with distinguishing replicated and nonreplicated DNA (see *15.24 Controlling the direction of mismatch repair).* In eukaryotes, its principal known function is connected with the control of transcription; methylation is associated with gene inactivation (see *21.18 Gene expression is associated with demethylation).*

From 2-7% of the cytosines of animal cell DNA are methylated (the value varies with the species). Most of the methyl groups are found in CG "doublets," and, in fact, the majority of the CG sequences are methylated. Usually the C residues on both strands of this short palindromic sequence are methylated, giving the structure

5' ^{m}CpG 3'
3' GpC^{m} 5'

Such a site is described as fully methylated. But consider the consequences of replicating this site. **Figure 23.36** shows that each daughter duplex has one methylated strand and one unmethylated strand. Such a site is called hemi-methylated.

The perpetuation of the methylated site now depends on what happens to hemimethylated DNA. If methylation of the unmethylated strand occurs, the site is restored to the fully methylated condition. However, if replication occurs first, the hemimethylated condition will be perpetuated on one daughter duplex, but the site will become unmethylated on the other daughter duplex. **Figure 23.37** shows that the state of methylation of DNA is controlled by methylases, which add methyl groups to the 5 position of cytosine, and demethylases, which remove the methyl groups. (The more formal name for the enzymes uses methyltransferase as the description.)

There are two types of DNA methylase, whose actions are distinguished by the state of the methylated DNA. To modify DNA at a new position requires the action of the de novo methylase, which recognizes DNA by virtue of a specific sequence. It acts *only* on nonmethylated DNA, to add a methyl group to one strand. There are two *de novo* methylases (Dnmt3A and Dnmt3B) in mouse; they have different target sites, and both are essential for development.

By Book_Crazy [IND]

A maintenance methylase acts constitutively *only on hemimethylated sites* to convert them to fully methylated sites. Its existence means that any methylated site is perpetuated after replication. There is one maintenance methylase (Dnmt1) in mouse, and it is essential: mouse embryos in which its gene has been disrupted do not survive past early embryogenesis.

Maintenance methylation is virtually 100% efficient, ensuring that the situation shown on the left of Figure 23.36 usually prevails *in vivo*. The result is that, if a *de novo* methylation occurs on one allele but not on the other, this difference will be perpetuated through ensuing cell divisions, maintaining a difference between the alleles that does not depend on their sequences.

Methylation has various types of targets. Gene promoters are the most common target. The promoters are methylated when the gene is inactive, but unmethylated when it is active. The absence of Dnmtl in mouse causes widespread demethylation at promoters, and we assume this is lethal because of the uncontrolled gene expression. Satellite DNA is another target. Mutations in Dnmt3B prevent methylation of satellite DNA, which causes centromere instability at the cellular level. Mutations in the corresponding human gene cause a disease. The importance of methylation is emphasized by another human disease, which is caused by mutation of the gene for the protein McCp2 that binds methylated CpG sequences.

The methylases are conventional enzymes that act on a DNA target. However, there may also be a methylation system that uses a short RNA sequence to target a corresponding DNA sequence for methylation (see *11.18 Antisense RNA can be used to inactivate gene expression)* Nothing is known about the mechanism of operation of this system.

How are demethylated regions established and maintained? If a DNA site has not been methylated, a protein that recognizes the unmethylated sequence could protect it against methylation. Once a site has been methylated, there are two possible ways to generate demethylated sites. One is to block the maintenance methylase from acting on the site when it is replicated. After a second replication cycle, one of the daughter duplexes will be unmethylated (as shown on the right side of Figure 23.36). The other is actively to demethylate the site, as shown in **Figure 23.38,** either by removing the methyl group directly from cytosine, or by excising the methylated cytosine or cytidine from DNA for replacement by a repair system. We know that active demethylation can occur to the paternal genome soon after fertilization, but we do not know what mechanism is used.

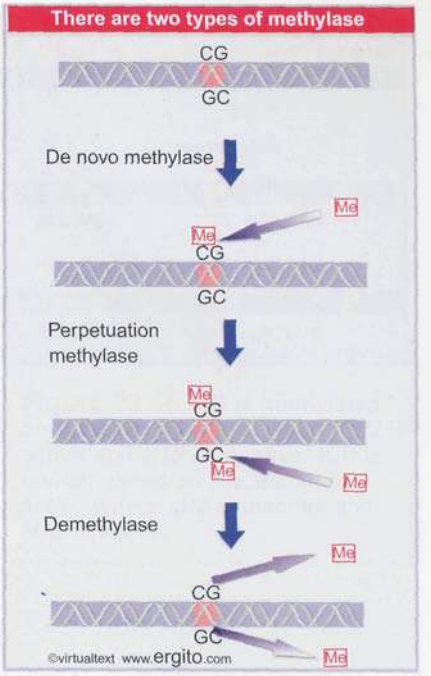


Figure 23.37 The state of methylation is controlled by three types of enzyme. *De novo* and perpetuation methylases are known, but demethylases have not been identified.

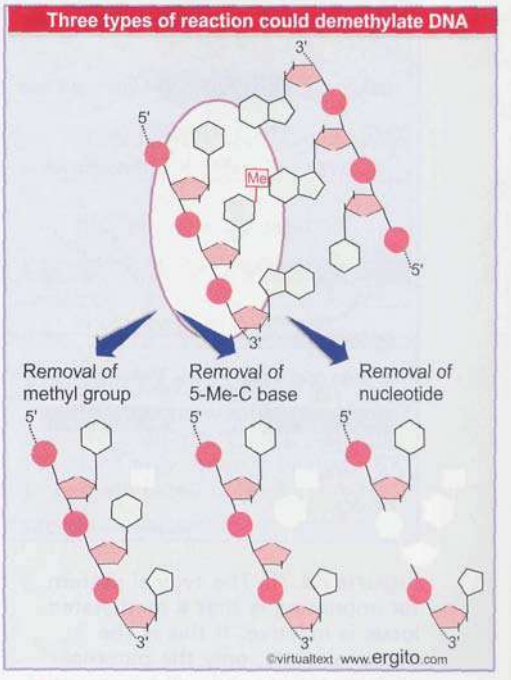


Figure 23.38 DNA could be demethylated by removing the methyl group, the base, or the nucleotide. Removal of the base or nucleotide would require its replacement by a repair system.

23.20 DNA methylation is responsible for imprinting

Key Concepts

- Paternal and maternal alleles may have different patterns of methylation at fertilization.
- Methylation is usually associated with inactivation of the gene.
- When genes are differentially imprinted, survival of the embryo may require that the functional allele is provided by the parent with the unmethylated allele.
- Survival of heterozygotes for imprinted genes is different depending on the direction of the cross.
- Imprinted genes occur in clusters and may depend on a local control site where *de novo* methylation occurs unless specifically prevented.

The pattern of methylation of germ cells is established in each sex during gametogenesis by a two stage process: first the existing pattern is erased by a genome-wide demethylation; then the pattern specific for each sex is imposed.

All allelic differences are lost when primordial germ cells develop in the embryo; irrespective of sex, the previous patterns of methylation are erased, and a typical gene is then unmethylated. In males, the pattern develops in two stages. The methylation pattern that is characteristic of mature sperm is established in the spermatocyte. But further changes are made in this pattern after fertilization. In females, the maternal pattern is imposed during oogenesis, when oocytes mature through meiosis after birth.

As may be expected from the inactivity of genes in gametes, the typical state is to be methylated. However, there are cases of differences between the two sexes, where a locus is unmethylated in one sex. A major question is how the specificity of methylation is determined in the male and female gametes.

Systematic changes occur in early embryogenesis. Some sites will continue to be methylated, but others will be specifically unmethylated in cells in which a gene is expressed. From the pattern of changes, we may infer that individual sequence-specific demethylation events occur during somatic development of the organism as particular genes are activated.

The specific pattern of methyl groups in germ cells is responsible for the phenomenon of imprinting, which describes a difference in behavior between the alleles inherited from each parent. The expression of certain genes in mouse embryos depends upon the sex of the parent from which they were inherited. For example, the allele coding for IGF-II (insulin-like growth factor II) that is inherited from the father is expressed, but the allele that is inherited from the mother is not expressed. The IGF-II gene of oocytes is methylated, but the IGF-II gene of sperm is not methylated, so that the two alleles behave differently in the zygote. This is the most common pattern, but the dependence on sex is reversed for some genes. In fact, the opposite pattern (expression of maternal copy) is shown for IGF-IIR, the receptor for IGF-II.

This sex-specific mode of inheritance requires that the pattern of methylation is established specifically during each gametogenesis. The fate of a hypothetical locus in a mouse is illustrated in **Figure 23.39.** In the early embryo, the paternal allele is nonmethylated and expressed, and the maternal allele is methylated and silent. What happens when this mouse itself forms gametes? If it is a male, the allele contributed to the sperm must be nonmethylated, irrespective of whether it was originally methylated or not. So when the maternal allele finds itself in a sperm, it must be demethylated. If the mouse is a female, the allele contributed to the egg must be methylated; so if it was originally the paternal allele, methyl groups must be added.

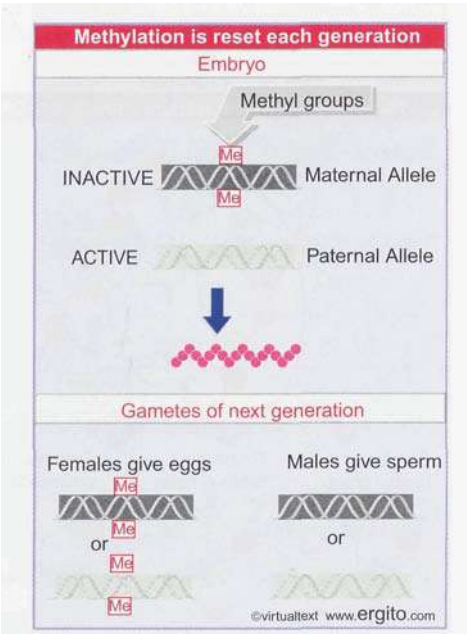


Figure 23.39 The typical pattern for imprinting is that a methylated locus is inactive. If this is the maternal allele, only the paternal allele is active, and will be essential for viability. The methylation pattern is reset when gametes are formed, so that all sperm have the paternal type, and all oocytes have the maternal type.

The consequence of imprinting is that an embryo requires a paternal allele for this gene. So in the case of a heterozygous cross where the allele of one parent has an inactivating mutation, the embryo will survive if the wild-type allele comes from the father, but will die if the wild-type allele is from the mother. This type of dependence on the directionality of the cross (in contrast with Mendelian genetics) is an example of epigenetic inheritance, where some factor other than the sequences of the genes themselves influences their effects (see *23.22 Epigenetic effects can be inherited).* Although the paternal and maternal alleles have identical sequences, they display different properties, depending on which parent provided them. These properties are inherited through meiosis and the subsequent somatic mitoses.

Imprinted genes are sometimes clustered. More than half of the 17 known imprinted genes in mouse are contained in two particular regions, each containing both maternally and paternally expressed genes. This

suggests the possibility that imprinting mechanisms may function over long distances. Some insights into this possibility come from deletions in the human population that cause the Prader-Willi and Angelman diseases. Most cases are caused by the same 4 Mb deletion, but the syndromes are different, depending on which parent contributed the deletion. The reason is that the deleted region contains at least one gene that is paternally imprinted and at least one that is maternally imprinted. There are some rare cases, however, with much smaller deletions. Prader-Willi syndrome can be caused by a 20 kb deletion that silences genes that are distant on either side of it. The basic effect of the deletion is to prevent a father from resetting the paternal mode to a chromosome inherited from his mother. The result is that these genes remain in maternal mode, so that the paternal as well as maternal alleles are silent in the offspring. The inverse effect is found in some small deletions that cause Angelman's syndrome. The implication is that this region comprises some sort of "imprint center" that acts at a distance to switch one parental type to the other.

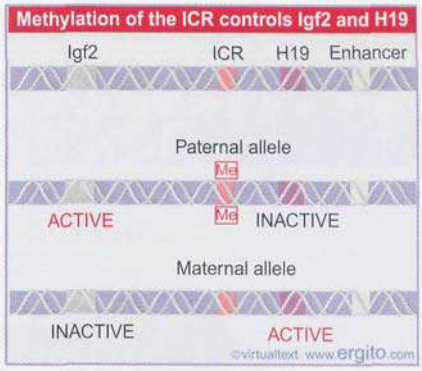


Figure 23.40 ICR is methylated on the paternal allele, where *Igf2* is active and *H19* is inactive. ICR is unmethylated on the maternal allele, where *Igf2* is inactive and *H19* is active.

23.21 Oppositely imprinted genes can be controlled by a single center

Key Concepts

- Imprinted genes are controlled by methylation of *cis*-acting sites.
- Methylation may be responsible for either inactivating or activating a gene.

Imprinting is determined by the state of methylation of a *cis*-acting site near a target gene or genes. These regulatory sites are known as DMDs (differentially methylated domains) or ICRs (imprinting control regions). Deletion of these sites removes imprinting, and the target loci then behave the same in both maternal and paternal genomes.

The behavior of a region containing two genes, *Igf2* and *H19*, illustrates the ways in which methylation can control gene activity. Figure 23.40 shows that these two genes react oppositely to the state of methylation at a site located between them, called the ICR. The ICR is methylated on the paternal allele. *H19* shows the typical response of inactivation. However, *Igf2* is expressed. The reverse situation is found on a maternal allele, where the ICR is not methylated. *H19* now becomes expressed, but *Igf2* is inactivated.

The control of *Igf2* is exercised by an insulator function of the ICR. Figure 23.41 shows that when the ICR is unmethylated, it binds the protein CTCF. This creates an insulator function that blocks an enhancer from activating the *Igf2* promoter. This is an unusual effect in which methylation indirectly activates a gene by blocking an insulator.

The regulation of *H19* shows the more usual direction of control in which methylation creates an inactive imprinted state. This could reflect a direct effect of methylation on promoter activity.

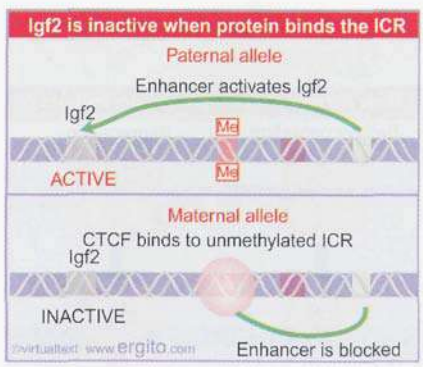


Figure 23.41 The ICR is an insulator that prevents an enhancer from activating Igf2. The insulator functions only when it binds CTCF to unmethylated DNA.

23.22 Epigenetic effects can be inherited

Key Concepts

- Epigenetic effects can result from modification of a nucleic acid after it has been synthesized or by the perpetuation of protein structures.

Epigenetic inheritance describes the ability of different states, which may have different phenotypic consequences, to be inherited without any change in the sequence of DNA. How can this occur?

We can divide epigenetic mechanisms into two general classes:

- DNA may be modified by the covalent attachment of a moiety that is then perpetuated. Two alleles with the same sequence may have different states of methylation that confer different properties.
- Or a self perpetuating protein state may be established. This might involve assembly of a protein complex, modification of specific protein(s), or establishment of an alternative protein conformation.

Methylation establishes epigenetic inheritance so long as the maintenance methylase acts constitutively to restore the methylated state after each cycle of replication, as shown in Figure 23.36. A state of methylation can be perpetuated through an indefinite series of somatic mitoses. This is probably the "default" situation. Methylation can also be perpetuated through meiosis: for example, in the fungus *Ascobolus* there are epigenetic effects that can be transmitted through both mitosis and meiosis by maintaining the state of methylation. In mammalian cells, epigenetic effects are created by resetting the state of methylation differently in male and female meioses.

Situations in which epigenetic effects appear to be maintained by means of protein states are less well understood in molecular terms. Position effect variegation shows that constitutive heterochromatin may extend for a variable distance, and the structure is then perpetuated through somatic divisions. Since there is no methylation of DNA in *Saccharomyces* and a vanishingly small amount in *Drosophila*, the inheritance of epigenetic states of position effect variegation or telomeric silencing in these organisms is likely to be due to the perpetuation of protein structures.

Figure 23.42 considers two extreme possibilities for the fate of a protein complex at replication:

- A complex could perpetuate itself if it splits symmetrically, so that half complexes associate with each daughter duplex. If the half complexes have the capacity to nucleate formation of full complexes, the original state will be restored. This is basically analogous to the maintenance of methylation. The problem with this model is that there is no evident reason why protein complexes should behave in this way.
- A complex could be maintained as a unit and segregate to one of the two daughter duplexes. The problem with this model is that it requires a new complex to be assembled *de novo* on the other daughter duplex, and it is not evident why this should happen.

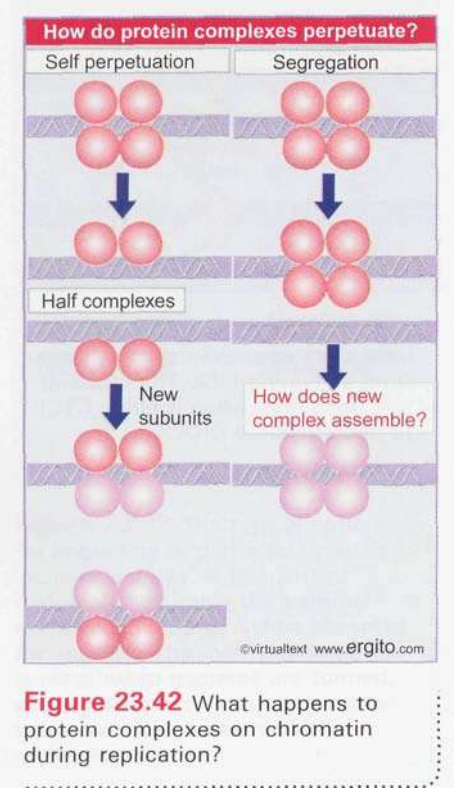


Figure 23.42 What happens to protein complexes on chromatin during replication?

Consider now the need to perpetuate a heterochromatic structure consisting of protein complexes. Suppose that a protein is distributed more or less continuously along a stretch of heterochromatin, as implied in Figure 23.20. If individual subunits are distributed at random to each daughter duplex at replication, the two daughters will continue to be marked by the protein, although its density will be reduced to half of the level before replication. If the protein has a self-assembling property that causes new subunits to associate with it, the original situation may be restored. *Basically, the existence of epigenetic effects forces us to the view that a protein responsible for such a situation must have some sort of self-templating or self-assembling capacity.*

In some cases, it may be the state of protein modification, rather than the presence of the protein *per se*, that is responsible for an epigenetic effect. There is a general correlation between the activity of chromatin and the state of acetylation of the histones, in particular the

acetylation of histones H3 and H4, which occurs on their N-terminal tails. Activation of transcription is associated with acetylation in the vicinity of the promoter; and repression of transcription is associated with deacetylation (see *23.7 Acetylases are associated with activators).* The most dramatic correlation is that the inactive X chromosome in mammalian female cells is underacetylated on histone H4.

The inactivity of constitutive heterochromatin may require that the histones are not acetylated. If a histone acetyltransferase is tethered to a region of telomeric heterochromatin in yeast, silenced genes become active. When yeast is exposed to trichostatin (an inhibitor of deacetylation), centromeric heterochromatin becomes acetylated, and silenced genes in centromeric regions may become active. *The effect may persist even after trichostatin has been removed.* In fact, it may be perpetuated through mitosis and meiosis. This suggests that an epigenetic effect has been created by changing the state of histone acetylation.

How might the state of acetylation be perpetuated? Suppose that the $H3_2 \cdot H4_2$ tetramer is distributed at random to the two daughter duplexes. This creates the situation shown in **Figure 23.43,** in which each daughter duplex contains some histone octamers that are fully acetylated on the H3 and H4 tails, while others are completely unacetylated. To account for the epigenetic effect, we could suppose that the presence of some fully acetylated histone octamers provides a signal that causes the unacetylated octamers to be acetylated.

(The actual situation is probably more complicated than shown in the figure, because transient acetylations occur during replication. If they are simply reversed following deposition of histones into nucleosomes, they may be irrelevant. An alternative possibility is that the usual deacetylation is prevented, instead of, or as well as, inducing acetylation.)

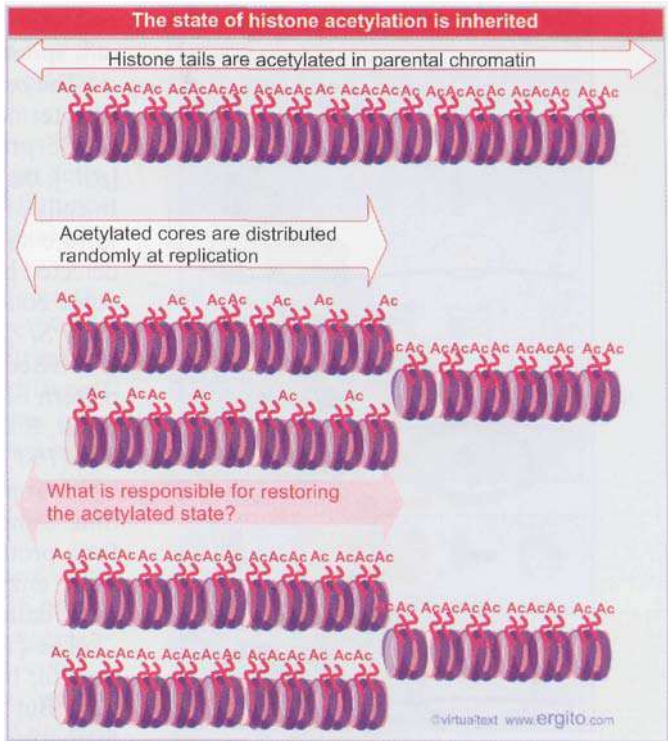


Figure 23.43 Acetylated cores are conserved and distributed at random to the daughter chromatin fibers at replication. Each daughter fiber has a mixture of old (acetylated) cores and new (unacetylated) cores.

23.23 Yeast prions show unusual inheritance

Key Concepts

- The Sup35 protein in its wild-type soluble form is a termination factor for translation.
- It can also exist in an alternative form of oligomeric aggregates, in which it is not active in protein synthesis.
- The presence of the oligomeric form causes newly synthesized protein to acquire the inactive structure.
- Conversion between the two forms is influenced by chaperones.
- The wild-type form has the recessive genetic state *psi*$^-$ and the mutant form has the dominant genetic state *PSI*$^+$.

One of the clearest cases of the dependence of epigenetic inheritance on the condition of a protein is provided by the behavior of prions—proteinaceous infectious agents. They have been characterized in two circumstances: by genetic effects in yeast; and as the causative agents of neurological diseases in mammals, including man. A striking epigenetic effect is found in yeast, where two different states can be inherited that map to a single genetic locus, *although the sequence of the gene is the same in both states.* The two different states are [*psi*$^-$]

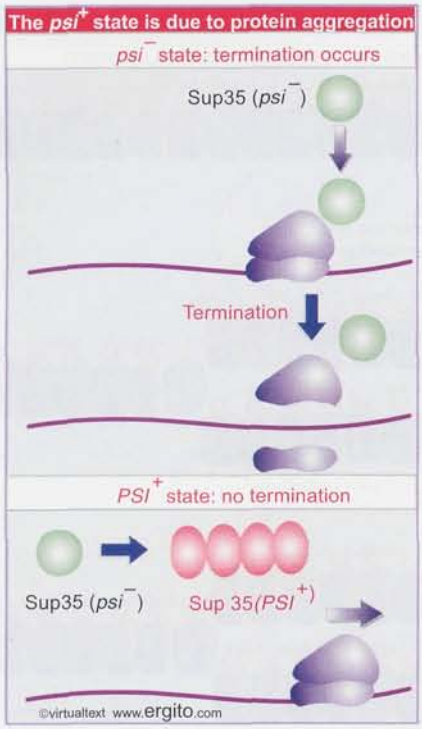


Figure 23.44 The state of the Sup35 protein determines whether termination of translation occurs.

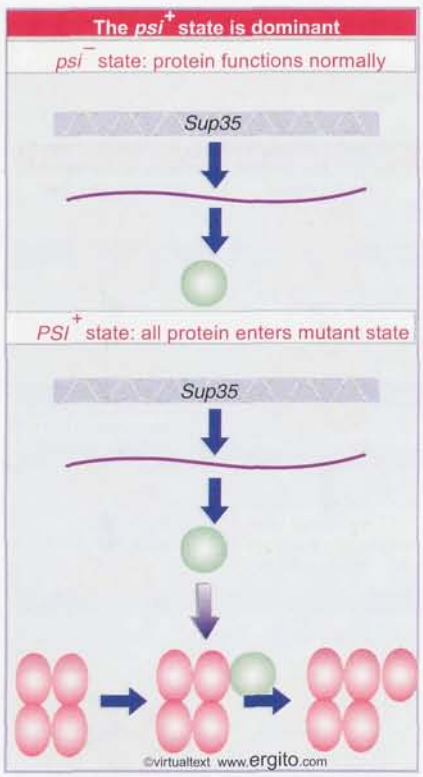


Figure 23.45 Newly synthesized Sup35 protein is converted into the [PSI+] state by the presence of pre-existing [PSI+] protein.

and [*PSI*+]. A switch in condition occurs at a low frequency as the result of a spontaneous transition between the states.

The *psi* genotype maps to the locus *sup35*, which codes for a translation termination factor. **Figure 23.44** summarizes the effects of the Sup35 protein in yeast. In wild-type cells, which are characterized as *[psi]*, the gene is active, and Sup35 protein terminates protein synthesis. In cells of the mutant [*PSI*+] type, the factor does not function, causing a failure to terminate protein synthesis properly. (This was originally detected by the lethal effects of the enhanced efficiency of suppressors of ochre codons in [*PSI*+] strains.)

[*PSI*+] strains have unusual genetic properties. When a [*psi*−] strain is crossed with a [*PSI*+] strain, *all of the progeny are [PSI+].* This is a pattern of inheritance that would be expected of an extrachromosomal agent, but the [*PSI*+] trait cannot be mapped to any such nucleic acid. The [*PSI*+] trait is metastable, which means that, although it is inherited by most progeny, it is lost at a higher rate than is consistent with mutation. Similar behavior is shown also by the locus *URE2*, which codes for a protein required for nitrogen-mediated repression of certain catabolic enzymes. When a yeast strain is converted into an alternative state, called *[URE3]*, the Ure2 protein is no longer functional.

The [*PSI*+] state is determined by the conformation of the Sup35 protein. In a wild-type [*psi*−]cell, the protein displays its normal function. But in a [*PSI*+] cell, the protein is present in an alternative conformation in which its normal function has been lost. To explain the unilateral dominance of [*PSI*+] over [*psi*−] in genetic crosses, we must suppose that *the presence of protein in the [PSI+] state causes all the protein in the cell to enter this state.* This requires an interaction between the [*PSI*+] protein and newly synthesized protein, probably reflecting the generation of an oligomeric state in which the [*PSI*+] protein has a nucleating role, as illustrated in **Figure 23.45.**

A feature common to both the Sup35 and Ure2 proteins is that each consists of two domains that function independently. The C-terminal domain is sufficient for the activity of the protein. The N-terminal domain is sufficient for formation of the structures that make the protein inactive. So yeast in which the N-terminal domain of Sup35 has been deleted cannot acquire the [*PSI*+] state; and the presence of an [*PSI*+] N-terminal domain is sufficient to maintain Sup35 protein in the [*PSI*+] condition. The critical feature of the N-terminal domain is that it is rich in glutamine and asparagine residues.

Loss of function in the [*PSI*+] state is due to the sequestration of the protein in an oligomeric complex. Sup35 protein in [*PSI*+] cells is clustered in discrete foci, whereas the protein in [*psi*−] cells is diffused in the cytosol. Sup35 protein from [*PSI*+] cells forms amyloid fibers *in vitro*—these have a characteristic high content of β sheet structures.

The involvement of protein conformation (rather than covalent modification) is suggested by the effects of conditions that affect protein structure. Denaturing treatments cause loss of the [*PSI*+] state. And in particular, the chaperone Hsp104 is involved in inheritance of [*PSI*+]. Its effects are paradoxical. Deletion of *HSP104* prevents maintenance of the [*PSI*+] state. And overexpression of Hsp104 also causes loss of the [*PSI*+] state. This suggests that Hsp104 is required for some change in the structure of Sup35 that is necessary for acquisition of the [*PSI*+] state, but that must be transitory.

Using the ability of Sup35 to form the inactive structure *in vitro*, it is possible to provide biochemical proof for the role of the protein. **Figure 23.46** illustrates a striking experiment in which the protein was converted to the inactive form *in vitro*, put into liposomes (when in effect the protein is surrounded by an artificial membrane), and then introduced directly into cells by fusing the liposomes with *[psi]* yeast. The

yeast cells were converted to [*PSI*$^+$]! This experiment refutes all of the objections that were raised to the conclusion that the protein has the ability to confer the epigenetic state. Experiments in which cells are mated, or in which extracts are taken from one cell to treat another cell, always are susceptible to the possibility that a nucleic acid has been transferred. But when the protein by itself does not convert target cells, but protein converted to the inactive state can do so, the only difference is the treatment of the protein—which must therefore be responsible for the conversion.

The ability of yeast to form the [*PSI*$^+$] prion state depends on the genetic background. The yeast must be [*PIN*$^+$] in order for the [*PSI*$^+$] state to form. The [*PIN*$^+$] condition itself is an epigenetic state. It can be created by the formation of prions from any one of several different proteins. These proteins share the characteristic of Sup35 that they have Gln/Asn-rich domains. Overexpression of these domains in yeast stimulates formation of the [*PSI*$^+$] state. This suggests that there is a common model for the formation of the prion state that involves aggregation of the Gln/Asn domains.

How does the presence of one Gln/Asn protein influence the formation of prions by another? We know that the formation of Sup35 prions is specific to Sup35 protein, that is, it does not occur by cross-aggregation with other proteins. This suggests that the yeast cell may contain soluble proteins that antagonize prion formation. These proteins are not specific for any one prion. As a result, the introduction of any Gln/Asn domain protein that interacts with these proteins will reduce the concentration. This will allow other Gln/Asn proteins to aggregate more easily.

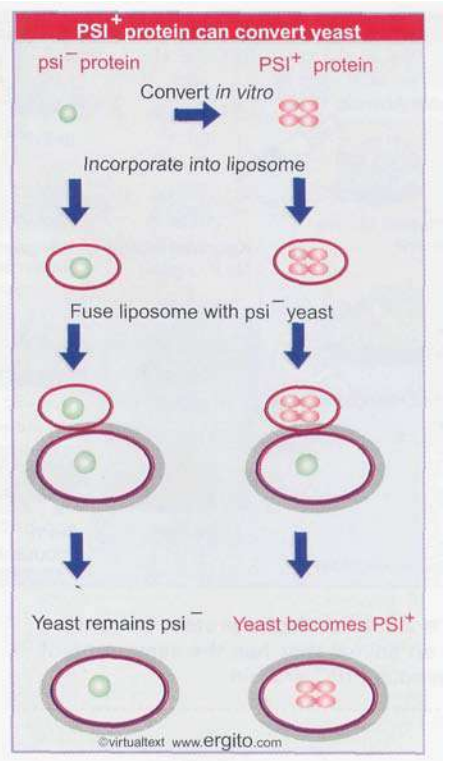


Figure 23.46 Purified protein can convert the [*psi*$^-$] state of yeast to [*PSI*$^+$].

23.24 Prions cause diseases in mammals

Key Concepts

- The protein responsible for scrapie exists in two forms, the wild-type noninfectious form PrP^C which is susceptible to proteases, and the disease-causing form PrP^{Sc} which is resistant to proteases.
- The neurological disease can be transmitted to mice by injecting the purified PrP^{Sc} protein into mice.
- The recipient mouse must have a copy of the *PrP* gene coding for the mouse protein.
- The PrP^{Sc} protein can perpetuate itself by causing the newly synthesized PrP protein to take up the PrP^{Sc} form instead of the PrP^C form.
- Multiple strains of PrP^{Sc} may have different conformations of the protein.

Prion diseases have been found in sheep and Man, and, more recently, in cows. The basic phenotype is an ataxia—a neurodegenerative disorder that is manifested by an inability to remain upright. The name of the disease in sheep, scrapie, reflects the phenotype: the sheep rub against walls in order to stay upright. Scrapie can be perpetuated by inoculating sheep with tissue extracts from infected animals. The disease kuru was found in New Guinea, where it appeared to be perpetuated by cannibalism, in particular the eating of brains. Related diseases in Western populations with a pattern of genetic transmission include Gerstmann- Straussler syndrome; and the related Creutzfeldt-Jakob disease (CJD) occurs sporadically. Most recently, a disease resembling CJD appears to have been transmitted by consumption of meat from cows suffering from "mad cow" disease.

By Book_Crazy [IND]

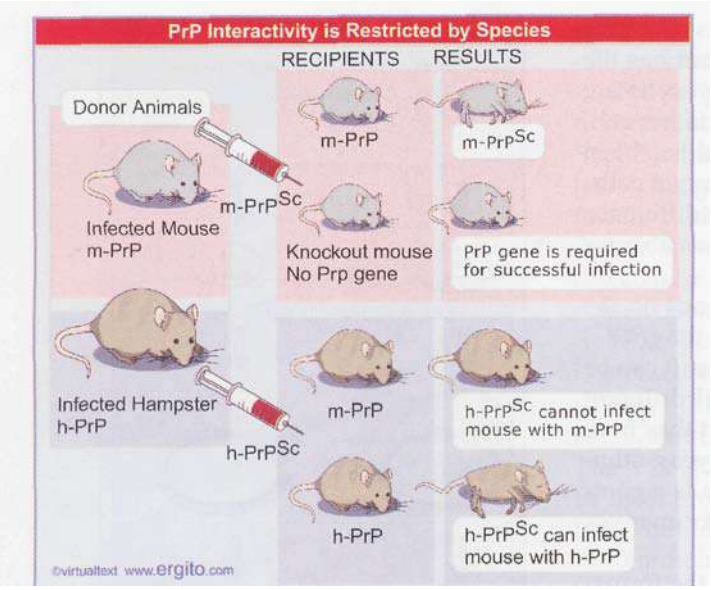


Figure 23.47 A PrpSc protein can only infect an animal that has the same type of endogenous PrPC protein.

When tissue from scrapie-infected sheep is inoculated into mice, the disease occurs in a period ranging from 75-150 days. The active component is a protease-resistant protein. The protein is coded by a gene that is normally expressed in brain. The form of the protein in normal brain, called PrP^C, is sensitive to proteases. Its conversion to the resistant form, called Prp^{Sc}, is associated with occurrence of the disease. The infectious preparation has no detectable nucleic acid, is sensitive to UV irradiation at wave lengths that damage protein, and has a low infectivity (1 infectious unit/10^5 PrP^{Sc} proteins). This corresponds to an epigenetic inheritance in which there is no change in genetic information, because normal and diseased cells have the same *PrP* gene sequence, but the PrP^{Sc} form of the protein is the infectious agent, whereas PrP^C is harmless.

The basis for the difference between the PrP^{Sc} and Prp^C forms appears to lie with a change in conformation rather than with any covalent alteration. Both proteins are glycosylated and linked to the membrane by a GPI-linkage. No changes in these modifications have been found. The PrP^{Sc} form has a high content of β sheets, which is absent from the PrP^C form.

The assay for infectivity in mice allows the dependence on protein sequence to be tested. **Figure 23.47** illustrates the results of some critical experiments. In the normal situation, PrP^{Sc} protein extracted from an infected mouse will induce disease (and ultimately kill) when it is injected into a recipient mouse. If the *PrP* gene is "knocked out", a mouse becomes resistant to infection. This experiment demonstrates two things. First, the endogenous protein is necessary for an infection, presumably because it provides the raw material that is converted into the infectious agent. Second, the cause of disease is not the removal of the PrP^C form of the protein, because a mouse with no PrP^C survives normally: the disease is caused by a gain-of-function in PrP^{Sc}.

The existence of species barriers allows hybrid proteins to be constructed to delineate the features required for infectivity. The original preparations of scrapie were perpetuated in several types of animal, but these cannot always be transferred readily. For example, mice are resistant to infection from prions of hamsters. This means that hamster-PrP^{Sc} cannot convert mouse-PrP^C to PrP^{Sc}. However, the situation changes if the mouse *PrP* gene is replaced by a hamster *PrP* gene. (This can be done by introducing the hamster *PrP* gene into the *PrP* knockout mouse.) A mouse with a hamster *PrP* gene is sensitive to infection by hamster PrP^{Sc}. This suggests that the conversion of cellular PrP^C protein into the Sc state requires that the PrP^{Sc} and PrP^C proteins have matched sequences.

There are different "strains" of PrP^{Sc}, which are distinguished by characteristic incubation periods upon inoculation into mice. This implies that the protein is not restricted solely to alternative states of PrP^C and PrP^{Sc}, but that there may be multiple Sc states. These differences must depend on some self-propagating property of the protein other than its sequence. If conformation is the feature that distinguishes PrP^{Sc} from PrP^C, then there must be multiple conformations, each of which has a self-templating property when it converts PrP^C.

The probability of conversion from PrP^C to PrP^{Sc} is affected by the sequence of PrP. Gerstmann-Straussler syndrome in man is caused by a single amino acid change in PrP. This is inherited as a dominant trait. If the same change is made in the mouse PrP gene, mice develop the disease. This suggests that the mutant protein has an increased probability of spontaneous conversion into the Sc state. Similarly, the sequence of the PrP gene determines the susceptibility of sheep to develop the dis-

ease spontaneously; the combination of amino acids at three positions (codons 136, 154, and 171) determines susceptibility.

The prion offers an extreme case of epigenetic inheritance, in which the infectious agent is a protein that can adopt multiple conformations, each of which has a self-templating property. This property is likely to involve the state of aggregation of the protein.

23.25 Summary

The existence of a preinitiation complex signals that the gene is in an "active" state, ready to be transcribed. The complex is stable, and may remain in existence through many cycles of replication. The ability to form a preinitiation complex could be a general regulatory mechanism. By binding to a promoter to make it possible for RNA polymerase in turn to bind, the factor in effect switches the gene on.

The variety of situations in which hypersensitive sites occur suggests that their existence reflects a general principle. *Sites at which the double helix initiates an activity are kept free of nucleosomes*. A transcription factor, or some other nonhistone protein concerned with the particular function of the site, modifies the properties of a short region of DNA so that nucleosomes are excluded. The structures formed in each situation need not necessarily be similar (except that each, by definition, creates a site hypersensitive to DNAase I).

Genes whose control regions are organized in nucleosomes usually are not expressed. In the absence of specific regulatory proteins, promoters and other regulatory regions are organized by histone octamers into a state in which they cannot be activated. This may explain the need for nucleosomes to be precisely positioned in the vicinity of a promoter, so that essential regulatory sites are appropriately exposed. Some transcription factors have the capacity to recognize DNA on the nucleosomal surface, and a particular positioning of DNA may be required for initiation of transcription.

Active chromatin and inactive chromatin are not in equilibrium. Sudden, disruptive events are needed to convert one to the other. Chromatin remodeling complexes have the ability to displace histone octamers by a mechanism that involves hydrolysis of ATP. Remodeling complexes are large and are classified according to the type of the ATPase subunit. Two common types are SWI/SNF and ISW. A typical form of this chromatin remodeling is to displace one or more histone octamers from specific sequences of DNA, creating a boundary that results in the precise or preferential positioning of adjacent nucleosomes. Chromatin remodeling may also involve changes in the positions of nucleosomes, sometimes involving sliding of histone octamers along DNA.

Acetylation of histones occurs at both replication and transcription and could be necessary to form a less compact chromatin structure. Some coactivators, which connect transcription factors to the basal apparatus, have histone acetylase activity. Conversely, repressors may be associated with deacetylases. The modifying enzymes are usually specific for particular amino acids in particular histones. The most common sites for modification are located in the N-terminal tails of histones H3 and H4, which extrude from nucleosomes between the turns of DNA. The activating (or repressing) complexes are usually large and often contain several activities that undertake different modifications of chromatin. Some common motifs found in proteins that modify chromatin are The chromo domain (concerned with protein-protein interactions), the bromo domain (which targets acetylated lysine), and the SET domain (part of the active sites of histone methyltransferases).

The formation of heterochromatin occurs by proteins that bind to specific chromosomal regions (such as telomeres) and that interact with histones. The formation of an inactive structure may propagate along the chromatin thread from an initiation center. Similar events occur in silencing of the inactive yeast mating type loci. Repressive

structures that are required to maintain the inactive states of particular genes are formed by the Pc-G protein complex in *Drosophila.* They share with heterochromatin the property of propagating from an initiation center.

Formation of heterochromatin may be initiated at certain sites and then propagated for a distance that is not precisely determined. When a heterochromatic state has been established, it is inherited through subsequent cell divisions. This gives rise to a pattern of epigenetic inheritance, in which two identical sequences of DNA may be associated with different protein structures, and therefore have different abilities to be expressed. This explains the occurrence of position effect variegation in *Drosophila.*

Modification of histone tails is a trigger for chromatin reorganization. Acetylation is generally associated with gene activation. Histories acetylases are found in activating complexes, and histone deacetylases are found in inactivating complexes. Histone methylation is associated with gene inactivation. Some histone modifications may be exclusive or synergistic with others.

Inactive chromatin at yeast telomeres and silent mating type loci appears to have a common cause, and involves the interaction of certain proteins with the N-terminal tails of histones H3 and H4. Formation of the inactive complex may be initiated by binding of one protein to a specific sequence of DNA; the other components may then polymerize in a cooperative manner along the chromosome.

Inactivation of one X chromosome in female (eutherian) mammals occurs at random. The *Xic* locus is necessary and sufficient to count the number of X chromosomes. The n-1 rule ensures that all but one X chromosome are inactivated. *Xic* contains the gene *Xist,* which codes for an RNA that is expressed only on the inactive X chromosome. Stabilization of *Xist* RNA is the mechanism by which the inactive X chromosome is distinguished.

Methylation of DNA is inherited epigenetically. Replication of DNA creates hemimethylated products, and a maintenance methylase restores the fully methylated state. Some methylation events depend on parental origin. Sperm and eggs contain specific and different patterns of methylation, with the result that paternal and maternal alleles are differently expressed in the embryo. This is responsible for imprinting, in which the nonmethylated allele inherited from one parent is essential because it is the only active allele; the allele inherited from the other parent is silent. Patterns of methylation are reset during gamete formation in every generation.

Prions are proteinaceous infectious agents that are responsible for the disease of scrapie in sheep and for related diseases in man. The infectious agent is a variant of a normal cellular protein. The PrP^{Sc} form has an altered conformation that is self-templating: the normal PrP^{C} form does not usually take up this conformation, but does so in the presence of PrP^{Sc}. A similar effect is responsible for inheritance of the *PSI* element in yeast.

References

23.2 Chromatin can have alternative states

rev Brown, D. D. (1984). The role of stable complexes that repress and activate eukaryotic genes. Cell 37, 359-365.

Weintraub, H. (1985). Assembly and propagation of repressed and derepressed chromosomal states. Cell 42, 705-711.

ref Bogenhagen, D. F., Wormington, W. M., and Brown, D. D. (1982). Stable transcription complexes of *Xenopus* 5S RNA genes: a means to maintain the differentiated state. Cell 28, 413-421.

Workman, J. L. and Roeder, R. G. (1987). Binding of transcription factor TFIID to the major late promoter during *in vitro* nucleosome assembly potentiates subsequent initiation by RNA polymerase II. Cell 51, 613-622.

23.3 Chromatin remodeling is an active process

rev Becker, P. B. and Horz, W. (2002). ATP-dependent nucleosome remodeling. Ann. Rev. Biochem. 71, 247-273.

Felsenfeld, G. (1992). Chromatin as an essential part of the transcriptional mechanism. Nature 355, 219-224.

Grunstein, M. (1990). Histone function in transcription. Ann. Rev. Cell Biol. 6, 643-678.
Narlikar, G. J., Fan, H. Y., and Kingston, R. E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell 108, 475-487.
Tsukiyama, T. (2002). The *in vivo* functions of ATP-dependent chromatin-remodelling factors. Nat. Rev. Mol. Cell Biol. 3, 422-429.
Vignali, M., Hassan, A. H., Neely, K. E., and Workman, J. L. (2000). ATP-dependent chromatin-remodeling complexes. Mol. Cell Biol. 20, 1899-1910.
ref Cairns, B. R., Kim, Y.- J., Sayre, M. H., Laurent, B. C, and Kornberg, R. (1994). A multisubunit complex containing the SWI/ADR6, SWI2/1, SWI3, SNF5, and SNF6 gene products isolated from yeast. Proc. Nat. Acad. Sci. USA 91, 1950-622.
Cote, J., Quinn, J., Workman, J. L., and Peterson, C. L. (1994). Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. Science 265, 53-60.
Gavin, I., Horn, P. J., and Peterson, C. L. (2001). SWI/SNF chromatin remodeling requires changes in DNA topology. Mol. Cell 7, 97-104.
Hamiche, A., Kang, J. G., Dennis, C, Xiao, H., and Wu, C. (2001). Histone tails modulate nucleosome mobility and regulate ATP-dependent nucleosome sliding by NURF. Proc. Nat. Acad. Sci. USA 98, 14316-14321.
Kingston, R. E. and Narlikar, G. J. (1999). ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. Genes Dev. 13, 2339-2352.
Kwon, H., Imbaizano, A. N., Khavari, P. A., Kingston, R. E., and Green, M. R. (1994). Nucleosome disruption and enhancement of activator binding of human SWI/SNF complex. Nature 370, 477-481.
Logie, C. and Peterson, C. L. (1997). Catalytic activity of the yeast SWI/SNF complex on reconstituted nucleosome arrays. EMBO J. 16, 6772-6782.
Lorch, Y., Cairns, B. R., Zhang, M., and Kornberg, R. D. (1998). Activated RSC-nucleosome complex and persistently altered form of the nucleosome. Cell 94, 29-34.
Lorch, Y., Zhang, M., and Kornberg, R. D. (1999). Histone octamer transfer by a chromatin-remodeling complex. Cell 96, 389-392.
Peterson, C. L. and Herskowitz, I. (1992). Characterization of the yeast SWI1, SWI2, and SWI3 genes, which encode a global activator of transcription. Cell 68, 573-583.
Robert, F., Young, R. A., and Struhl, K. (2002). Genome-wide location and regulated recruitment of the RSC nucleosome remodeling complex. Genes Dev. 16, 806-819.
Schnitzler, G., Sif, S., and Kingston, R. E. (1998). Human SWI/SNF interconverts a nucleosome between its base state and a stable remodeled state. Cell 94, 17-27.
Tamkun, J. W., Deuring, R., Scott, M. P., Kissinger, M., Pattatucci, A. M., Kaufman, T. C, and Kennison, J. A. (1992). brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. Cell 68, 561-572.
Tsukiyama, T., Daniel, C, Tamkun, J., and Wu, C. (1995). ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. Cell 83, 1021-1026.
Tsukiyama, T., Palmer, J., Landel, C. C, Shiloach, J., and Wu, C. (1999). Characterization of the imitation switch subfamily of ATP-dependent chromatin-remodeling factors in S. *cerevisiae.* Genes Dev. 13, 686-697.
Whitehouse, I., Flaus, A., Cairns, B. R., White, M. F., Workman, J. L., and Owen-Hughes, T. (1999). Nucleosome mobilization catalysed by the yeast SWI/SNF complex. Nature 400, 784-787.

23.4 Nucleosome organization may be changed at the promoter

ref Cosma, M. P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. Cell 97, 299-311.
Kadam, S., McAlpine, G. S., Phelan, M. L., Kingston, R. E., Jones, K. A., and Emerson, B. M. (2000). Functional selectivity of recombinant mammalian SWI/SNF subunits. Genes Dev. 14, 2441-2451.
Lohr, D. (1997). Nucleosome transactions on the promoters of the yeast GAL and PHO genes. J. Biol. Chem. 272,-26795-26798.
McPherson, C. E., Shim, E.-Y., Friedman, D. S., and Zaret, K. S. (1993). An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. Cell 75, 387-398.
Schmid, V. M., Fascher, K.-D., and Horz, W. (1992). Nucleosome disruption at the yeast PHO5 promoter upon PHO5 induction occurs in the absence of DNA replication. Cell 71, 853-864.
Truss, M., Barstch, J., Schelbert, A., Hache, R. J. G., and Beato, M. (1994). Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter in vivo. EMBO J. 14, 1737-1751.
Tsukiyama, T., Becker, P. B., and Wu, C. (1994). ATP-dependent nucleosome disruption at a heat shock promoter mediated by binding of GAGA transcription factor. Nature 367, 525-532.
Yudkovsky, N., Logie, C, Hahn, S., and Peterson, C. L. (1999). Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. Genes Dev. 13, 2369-2374.

23.5 Histone modification is a key event

rev Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. Science 293, 1074-1080.
ref Osada, S., Sutton, A., Muster, N., Brown, C. E., Yates, J. R., Sternglanz, R., and Workman, J. L. (2001). The yeast SAS (something about silencing) protein complex contains a MYST-type putative acetyltransferase and functions with chromatin assembly factor ASF1. Genes Dev. 15, 3155-3168.

23.6 Histone acetylation occurs in two circumstances

rev Hirose, Y. and Manley, J. L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev. 14, 1415-1429.
Verreault, A. (2000). De novo nucleosome assembly: new pieces in an old puzzle. Genes Dev. 14, 1430-1438.
ref Akhtar, A. and Becker, P. B. (2000). Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in *Drosophila.* Mol. Cell 5, 367-375.
Alwine, J. C, Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Nat. Acad. Sci. USA 74, 5350-5354.

Jackson, V., Shires, A., Tanphaichitr, N., and Chalkley, R. (1976). Modifications to histones immediately after synthesis. J. Mol. Biol. 104, 471-483.

Ling, X., Harkness, T. A., Schultz, M. C, Fisher-Adams, G., and Grunstein, M. (1996). Yeast histone H3 and H4 amino termini are important for nucleosome assembly *in vivo* and *in vitro:* redundant and position-independent functions in assembly but not in gene regulation. Genes Dev. 10, 686-699.

Shibahara, K., Verreault, A., and Stillman, B. (2000). The N-terminal domains of histones H3 and H4 are not necessary for chromatin assembly factor-1-mediated nucleosome assembly onto replicated DNA *in vitro.* Proc. Nat. Acad. Sci. USA 97, 7766-7771.

Turner, B. M., Birley, A. J., and Lavender, J. (1992). Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. Cell 69, 375-384.

23.7 Acetylases are associated with activators

rev Kingston, R. E. and Narlikar, G. J. (1999). ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. Genes Dev. 13, 2339-2352.

ref Brownell, J. E. et al. (1996). Tetrahymena histone acetyltransferase A: a homologue to yeast Gcn5p linking histone acetylation to gene activation. Cell 84, 843-851.

Chen, H. et al. (1997). Nuclear receptor coactivator ACTR is a novel histoneacetyltransferase and forms a multimeric activation complex with P/CAF and CP/p300. Cell 90, 569-580.

Grant, P. A. et al. (1998). A subset of $TAF_{II}s$ are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. Cell 94, 45-53.

Lee, T. I., Causton, H. C, Holstege, F. C, Shen, W. C, Hannett, N., Jennings, E. G., Winston, F., Green, M. R., and Young, R. A. (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. Nature 405, 701-704.

23.8 Deacetylases are associated with repressors

rev Richards, E. J., Elgin, S. C, and Richards, S. C. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. Cell 108, 489-500.

ref Ayer, D. E., Lawrence, Q. A., and Eisenman, R. N. (1995). Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. Cell 80, 767-776.

Kadosh, D. and Struhl, K. (1997). Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. Cell 89, 365-371.

Schreiber-Agus, N., Chin, L., Chen, K., Torres, R., Rao, G., Guida, P., Skoultchi, A. I., and DePinho, R. A. (1995). An amino-terminal domain of Mxi1 mediates anti-Myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor SIN3. Cell 80, 777-786.

23.9 Methylation of histones and DNA is connected

rev Richards, E. J., Elgin, S. C, and Richards, S. C. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. Cell 108, 489-500.

ref Ng, H. H., Feng, Q., Wang, H., Erdjument-Bromage, H., Tempst, P., Zhang, Y., and Struhl, K. (2002). Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. Genes Dev. 16, 1518-1527.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B. D., Sun, Z. W., Sun, M., Opravil, S., Mechtler, K., Ponting, C. P., Allis, C. D., and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature 406, 593-599.

Tamaru, H. and Selker, E. U. (2001). A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa.* Nature 414, 277-283.

Zhang, Y. and Reinberg, D. (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. Genes Dev. 15, 2343-2360.

23.11 Promoter activation involves an ordered series of events

rev Orphanides, G. and Reinberg, D. (2000). RNA polymerase II elongation through chromatin. Nature 407, 471-475.

ref Bortvin, A. and Winston, F. (1996). Evidence that Spt6p controls chromatin structure by a direct interaction with histones. Science 272, 1473-1476.

Cosma, M. P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. Cell 97, 299-311.

Hassan, A. H., Neely, K. E., and Workman, J. L. (2001). Histone acetyltransferase complexes stabilize swi/snf binding to promoter nucleosomes. Cell 104, 817-827.

Orphanides, G., LeRoy, G., Chang, C. H., Luse, D. S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell 92, 105-116.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., and Handa, H. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev. 12, 343-356.

23.12 Histone phosphorylation affects chromatin structure

ref Wang, Y., Zhang, W., Jin, Y., Johansen, J., and Johansen, K. M. (2001). The JIL-1 tandem kinase mediates histone H3 phosphorylation and is required for maintenance of chromatin structure in *Drosophila.* Cell 105, 433-443.

23.13 Heterochromatin propagates from a nucleation event

ref Ahmad, K. and Henikoff, S. (2001). Modulation of a transcription factor counteracts heterochromatic gene silencing in *Drosophila.* Cell 104, 839-847.

23.14 Some common motifs are found in proteins that modify chromatin

ref Dhalluin, C, Carlson, J. E., Zeng, L., He, C, Aggarwal, A. K., and Zhou, M. M. (1999). Structure and ligand of a histone acetyltransferase Nature 399, 491-496.

Eissenberg, J. C, Morris, G. D., Reuter, G., and Hartnett, T. (1992). The heterochromatin-associated protein HP-1 is an essential protein in *Drosophila* with dosage-dependent effects on position-effect variegation. Genetics 131, 345-352.

James, T. C. and Elgin, S. C. (1986). Identification of a nonhistone chromosomal protein associated with heterochromatin in *D. melanogaster* and its gene. Mol. Cell Biol. 6, 3862-3872.

Koonin, E. V., Zhou, S., and Lucchesi, J. C. (1995). The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nuc. Acids Res. 23, 4229-4233.

Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D., and Felsenfeld, G. (2001). Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. Science 293, 2453-2455.

Owen, D. J., Ornaghi, P., Yang, J. C, Lowe, N., Evans, P. R., Ballario, P., Neuhaus, D., Filetici, P., and Travers, A. A. (2000). The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p. EMBO J. 19, 6141-6149.

Platero, J. S., Hartnett, T., and Eissenberg, J. C. (1995). Functional analysis of the chromo domain of HP1. EMBO J. 14, 3977-3986.

Turner, B. M., Birley, A. J., and Lavender, J. (1992). Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. Cell 69, 375-384.

23.15 Heterochromatin depends on interactions with histones

rev Loo, S. and Rine, J. (1995). Silencing and heritable domains of gene expression. Ann. Rev. Cell Dev. Biol. 11, 519-548.

Moazed, D. (2001). Common themes in mechanisms of gene silencing. Mol. Cell 8, 489-498.

Thompson, J. S., Hecht, A., and Grunstein, M. (1993). Histones and the regulation of heterochromatin in yeast. Cold Spring Harbor Symp. Quant. Biol. 58, 247-256.

ref Nakayama , J., Rice, J. C, Strahl, B. D., Allis, C. D., and Grewal, S. I. (2001). Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science 292, 110-113.

Ahmad, K. and Henikoff, S. (2001). Modulation of a transcription factor counteracts heterochromatic gene silencing in *Drosophila.* Cell 104, 839-847.

Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. 0., Allshire, R. C, and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410, 120-124.

Bloom, K. S. and Carbon, J. (1982). Yeast centromere DNA is in a unique and highly ordered structure in chromosomes and small circular minichromosomes. Cell 29, 305-317.

Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S. M., and Grunstein, M. (1995). Histone H3 and H4 N-termini interact with the silent information regulators SIR3 and SIR4: a molecular model for the formation of heterochromatin in yeast. Cell 80, 583-592.

Imai, S., Armstrong, C. M., Kaeberlein, M., and Guarente, L. (2000). Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. Nature 403, 795-800.

Kayne, P. S., Kim, U. J., Han. M., Mullen, R. J., Yoshizaki, F., and Grunstein, M. (1988). Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell 55, 27-39.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410, 116-120.

Landry, J., Sutton, A., Tafrov, S. T., Heller, R. C, Stebbins, J., Pillus, L., and Sternglanz, R. (2000). The silencing protein SIR2 and its homologs are NAD-dependent protein deacetylases. Proc. Nat. Acad. Sci. USA 97, 5807-5811.

Manis, J. P., Gu, Y., Lansford, R., Sonoda, E., Ferrini, R., Davidson, L., Rajewsky, K., and Alt, F. W. (1998). Ku70 is required for late B cell development and immunoglobulin heavy chain class switching. J. Exp. Med. 187, 2081-2089.

Meluh, P. B. et al. (1998). Cse4p is a component of the core centromere of *S. cerevisiae.* Cell 94, 607-613.

Moretti, P., Freeman, K., Coodly, L., and Shore, D. (1994). Evidence that a complex of SIR proteins interacts with the silencer and telomere-binding protein RAP1. Genes Dev. 8, 2257-2269.

Nakagawa, H., Lee, J. K., Hurwitz, J., Allshire, R. C, Nakayama, J., Grewal, S. I., Tanaka, K., and Murakami, Y. (2002). Fission yeast CENP-B homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications. Genes Dev. 16, 1766-1778.

Palladino, F., Laroche, T., Gilson, E., Axelrod, A., Pillus, L., and Gasser, S. M. (1993). SIR3 and SIR4 proteins are required for the positioning and integrity of yeast telomeres. Cell 75, 543-555.

Sekinger, E. A. and Gross, D. S. (2001). Silenced chromatin is permissive to activator binding and PIC recruitment. Cell 105, 403-414.

Shore, D. and Nasmyth, K. (1987). Purification and cloning of a DNA-binding protein from yeast that binds to both silencer and activator elements. Cell 51, 721-732.

Smith, J. S., Brachmann, C. B., Celic, I., Kenna, M. A., Muhammad, S., Starai, V. J., Avalos, J. L., Escalante-Semerena, J. C, Grubmeyer, C, Wolberger, C, and Boeke, J. D. (2000). A phylogenetically conserved NAD +-dependent protein deacetylase activity in the Sir2 protein family. Proc. Nat. Acad. Sci. USA 97, 6658-6663.

Zhang, Y. and Reinberg, D. (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. Genes Dev. 15, 2343-2360.

23.16 Polycomb and trithorax are antagonistic repressors and activators

ref Chan, C.-S., Rastelli, L., and Pirrotta, V. (1994). A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. EMBO J. 13, 2553-2564.

Eissenberg, J. C, James, T. C, Fister-Hartnett, D. M., Hartnett, T., Ngan, V., and Elgin, S. C. R. (1990). Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in *D. melanogaster.* Proc. Nat. Acad. Sci. USA 87, 9923-9927.

Franke, A., DeCamillis, M., Zink, D., Cheng, N., Brock, H. W., and Paro, R. (1992). Polycomb and polyhomeotic are constituents of a multimeric protein complex in chromatin ofD. *melanogaster.* EMBO J. 11, 2941-29.

Geyer, P. K. and Corces, V. G. (1992). DNA position-specific repression of transcription by a *Drosophila* zinc finger protein. Genes Dev. 6, 1865-1873.

Orlando, V. and Paro, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using *in vivo* formaldehyde cross-linked chromatin. Cell 75, 1187-1198.

Strutt, H., Cavalli, G., and Paro, R. (1997). Colocalization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. EMBO J. 16, 3621-3632.

Zink, B. and Paro, R. (1989). *In vivo* binding patterns of a of the homeotic genes in *D. melanogaster.* Nature 337, 468-471.

23.17 X chromosomes undergo global changes

exp Lyon, M. (2002). The Discovery of X-Chromosome Inactivation (www.ergito.com/lookup.jsp?expt = lyon)

rev Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A., and Panning, B. (2002). Xist RNA and the mechanism of X chromosome inactivation. Ann. Rev. Genet. 36, 233-278.

ref Jeppesen, P. and Turner, B. M. (1993). The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. Cell 74, 281-289.

Lee, J. T. et al. (1996). A 450 kb transgene displays properties of the mammalian X-inactivation center. Cell 86, 83-94.

Lyon, M. F. (1961). Gene action in the X chromosome of the mouse. Nature 190, 372-373.

Panning, B., Dausman, J., and Jaenisch, R. (1997). X chromosome inactivation is mediated by Xist RNA stabilization. Cell 90, 907-916.

Penny, G. D. et al. (1996). Requirement for Xist in X chromosome inactivation. Nature 379, 131-137.

23.18 Chromosome condensation is caused by condensins

rev Hirano, T. (1999). SMC-mediated chromosome mechanics: a conserved scheme from bacteria to vertebrates? Genes Dev. 13, 11-19.

Hirano, T. (2000). Chromosome cohesion, condensation, and separation. Ann. Rev. Biochem. 69, 115-144.

Jessberger, R. (2002). The many functions of SMC proteins in chromosome dynamics. Nat. Rev. Mol. Cell Biol. 3, 767-778.

Nasmyth, K. (2002). Segregating sister genomes: the molecular biology of chromosome separation. Science 297, 559-565.

ref Haering, C. H., Lowe, J.,Hochwage, A., and Nasmyth, K. (2002). Molecular architecture of SMC proteins and the yeast cohesin complex. Mol. Cell 9, 773-788.

Hirano, T. (2002). The ABCs of SMC proteins: two-armed ATPases for chromosome condensation, cohesion, and repair. Genes Dev. 16, 399-414.

Kimura, K., Rybenkov, V. V., Crisona, N. J., Hirano, T., and Cozzarelli, N. R. (1999). 13S condensin actively reconfigures DNA by introducing global positive writhe: implications for chromosome condensation. Cell 98, 239-248.

23.19 DNA methylation is perpetuated by a maintenance methylase

rev Bird, A. P. (1986). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Nature 321, 209-213.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. 16, 6-21.

Matzke, M., Matzke, A. J., and Kooter, J. M. (2001). RNA: guiding gene silencing. Science 293, 1080-1083.

Sharp, P. A. (2001). RNA interference–2001. Genes Dev. 15, 485-490.

ref Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat. Genet. 23, 185-188.

Li, E., Bestor, T. H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69, 915-926.

Okano, M., Bell, D. W., Haber, D. A., and Li. E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247-257.

Xu, G. L., Bestor, T. H., Bourc'his, D., Hsieh, C. L, Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J. J., and Viegas-Paquignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature 402, 187-191.

23.20 DNA methylation is responsible for imprinting

rev Bartolomei, M. S. and Tilghman, S. (1997). Genomic imprinting in mammals. Ann. Rev. Genet. 31, 493-525.

ref Chaillet, J. R., Vogt, T. F., Beier, D. R., and Leder, P. (1991). Parental-specific methylation of an imprinted transgene is established during gametogenesis and progressively changes during embryogenesis. Cell 66, 77-83.

23.21 Oppositely imprinted genes can be controlled by a single center

ref Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405, 482-485.

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature 405, 486-489.

23.23 Yeast prions show unusual inheritance

rev Horwich, A. L. and Weissman, J. S. (1997). Deadly conformations: protein misfolding in prion disease. Cell 89, 499-510.

Lindquist, S. (1997). Mad cows meet psi-chotic yeast: the expansion of the prion hypothesis. Cell 89, 495-498.

Serio, T. R. and Lindquist, S. L. (1999). [PSI$^+$]: an epigenetic modulator of translation termination efficiency. Ann. Rev. Cell Dev. Biol. 15, 661-703.

Wickner, R. B. (1996). Prions and RNA viruses of S. *cerevisiae*. Ann. Rev. Genet. 30, 109-139.

ref Chernoff, Y. O. et al. (1995). Role of the chaperone protein Hsp104 in propagation of the yeast prion-like factor [PSI$^+$]. Science 268, 880-884.

Derkatch, I. L., Bradley, M.E., Masse, S. V., Zadorsky, S.P., Polozkov, G. V., Inge-Vechtomov, S. G., Liebman S. W. (2000). Dependence and independence of [PSK +)] and [PIN(+)]: a two-prion system in yeast? EMBO J. 19, 1942-1952.

Derkatch, I. L., Bradley, M. E., Hong, J. Y., and Liebman, S. W. (2001). Prions affect the appearance of other prions: the story of [PIN(+)]. Cell 106, 171-182.

Glover, J. R. et al. (1997). Self-seeded fibers formed by Sup35, the protein determinant of [PSI$^+$], a heritable prion-like factor of *S. cerevisiae.* Cell 89, 811-819.

Masison, D. C. and Wickner, R. B. (1995). Prion-inducing domain of yeast Ure2p and protease resistance of Ure2p in prion-containing cells. Science 270, 93-95.

Osherovich, L. Z. and Weissman, J. S. (2001). Multiple gln/asn-rich prion domains confer susceptibility to induction of the yeast. Cell 106, 183-194.

Sparrer, H. E., Santoso, A., Szoka F. C, Jr., and Weissman, J. S. (2000). Evidence for the prion hypothesis: induction of the yeast [PSI$^+$] factor by in vitro-converted sup 35 protein. Science 289, 595-599.

Wickner, R. B. (1994). [URE3] as an altered URE2 protein: evidence for a prion analog in *S. cerevisiae.* Science 264, 566-569.

23.24 Prions cause diseases in mammals

rev Prusiner, S. (1982). Novel proteinaceous infectious particles cause scrapie. Science 216, 136-144.

Prusiner, S. B. and Scott, M. R. (1997). Genetics of prions. Ann. Rev. Genet. 31, 139-175.

ref Basler, K., Oesch, B., Scott, M., Westaway, D., Walchli, M., Groth, D. F., McKinley, M. P., Prusiner, S. B., and Weissmann, C. (1986). Scrapie and cellular PrP isoforms are encoded by the same chromosomal gene. Cell 46, 417-428.

Bueler, H. et al. (1993). Mice devoid of PrP are resistant to scrapie. Cell 73, 1339-1347.

Hsiao, K. et al. (1989). Linkage of a prion protein missense variant to Gerstmann-Straussler syndrome. Nature 338, 342-345.

McKinley, M. P., Bolton, D. C, and Prusiner, S. B. (1983). A protease-resistant protein is a structural component of the scrapie prion. Cell 35, 57-62.

Oesch, B. et al. (1985). A cellular gene encodes scrapie PrP27-30 protein. Cell 40, 735-746.

Scott, M. et al. (1993). Propagation of prions with artificial properties in transgenic mice expressing chimeric PrP genes. Cell 73, 979-988.

Chapter 24

RNA splicing and processing

24.1 Introduction

Interrupted genes are found in all classes of organisms. They represent a minor proportion of the genes of the very lowest eukaryotes, but the vast majority of genes in higher eukaryotic genomes. Genes vary widely according to the numbers and lengths of introns, but a typical mammalian gene has 7-8 exons spread out over ~16 kb. The exons are relatively short (~ 100-200 bp), and the introns are relatively long (>1 kb) (see *2.7 Genes show a wide distribution of sizes).*

The discrepancy between the interrupted organization of the gene and the uninterrupted organization of its mRNA requires processing of the primary transcription product. The primary transcript has the same organization as the gene, and is sometimes called the pre-mRNA. Removal of the introns from pre-mRNA leaves a typical messenger of ~2.2 kb. The process by which the introns are removed is called RNA splicing.

Removal of introns is a major part of the production of RNA in all eukaryotes. (Although interrupted genes are relatively rare in lower eukaryotes such as yeast, the overall proportion underestimates the importance of introns, because most of the genes that are interrupted code for relatively abundant proteins. Splicing is therefore involved in the production of a greater proportion of total mRNA than would be apparent from analysis of the genome, perhaps as much as 50%.)

One of the first clues about the nature of the discrepancy in size between nuclear genes and their products in higher eukaryotes was provided by the properties of nuclear RNA. Its average size is much larger than mRNA, it is very unstable, and it has a much greater sequence complexity. Taking its name from its broad size distribution, it was called heterogeneous nuclear RNA (hnRNA). It includes pre-mRNA, but could also include other transcripts.

The physical form of hnRNA is a ribonucleoprotein particle (hnRNP), in which the hnRNA is bound by proteins. As characterized *in vitro*, an hnRNP particle takes the form of beads connected by a fiber. The structure is summarized in **Figure 24.1.** The most abundant proteins in the particle are the core proteins, but other proteins are present at lower stoichiometry, making a total of ~20 proteins. The proteins typically are present at ~10^8 copies per nucleus, compared with ~10^6 molecules of hnRNA. Some of the proteins may have a structural role in packaging the hnRNA; several are known to shuttle between the nucleus and cytoplasm, and play roles in exporting the RNA or otherwise controlling its activity.

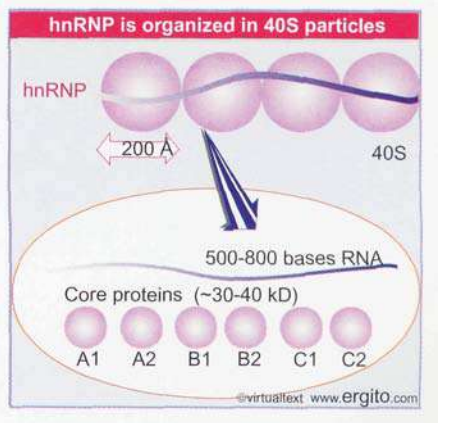


Figure 24.1 hnRNA exists as a ribonucleoprotein particle organized as a series of beads.

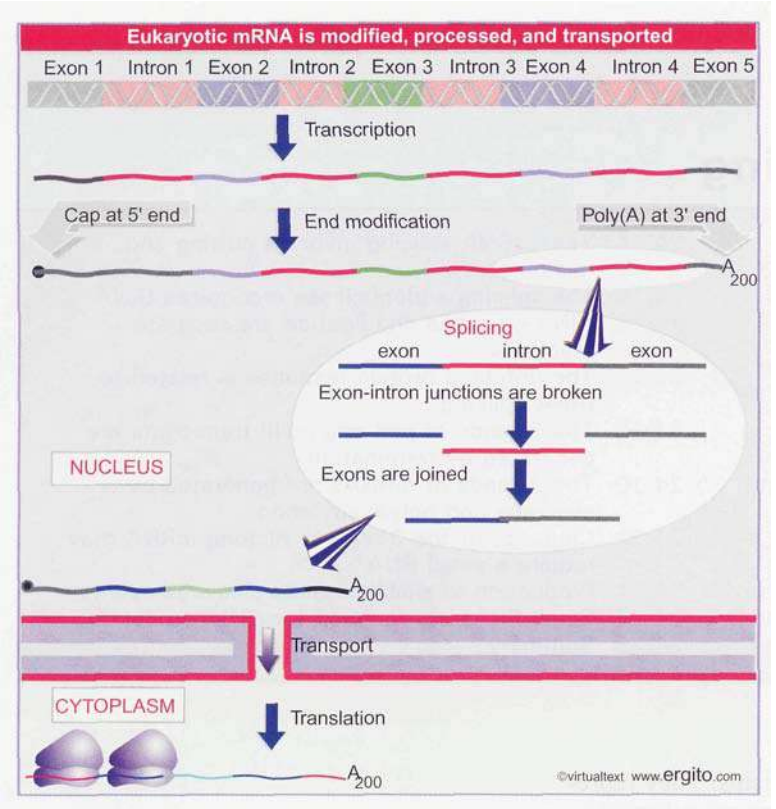


Figure 24.2 RNA is modified in the nucleus by additions to the 5' and 3' ends and by splicing to remove the introns. The splicing event requires breakage of the exon-intron junctions and joining of the ends of the exons. Mature mRNA is transported through nuclear pores to the cytoplasm, where it is translated.

Splicing occurs in the nucleus, together with the other modifications that are made to newly synthesized RNAs. The process of expressing an interrupted gene is reviewed in **Figure 24.2.** The transcript is capped at the 5' end (see *5.9 The 5' end of eukaryotic mRNA is capped),* has the introns removed, and is polyadenylated at the 3' end (see *5.10 The 3' terminus is polyadenylated).* The RNA is then transported through nuclear pores to the cytoplasm, where it is available to be translated.

With regard to the various processing reactions that occur in the nucleus, we should like to know at what point splicing occurs *vis-à-vis* the other modifications of RNA. Does splicing occur at a particular location in the nucleus; and is it connected with other events, for example, nucleocytoplasmic transport? Does the lack of splicing make an important difference in the expression of uninterrupted genes?

With regard to the splicing reaction itself, one of the main questions is how its specificity is controlled. What ensures that the ends of each intron are recognized in pairs so that the correct sequence is removed from the RNA? Are introns excised from a precursor in a particular order? Is the maturation of RNA used to *regulate* gene expression by discriminating among the available precursors or by changing the pattern of splicing?

We can identify several types of splicing systems:

- Introns are removed from the nuclear pre-mRNAs of higher eukaryotes by a system that recognizes only short consensus sequences conserved at exon-intron boundaries and within the intron. This reaction requires a large splicing apparatus, which takes the form of an array of proteins and ribonucleoproteins that functions as a large particulate complex (the spliceosome). The mechanism of splicing involves transesterifications, and the catalytic center includes RNA as well as proteins.
- Certain RNAs have the ability to excise their introns autonomously. Introns of this type fall into two groups, as distinguished by secondary/tertiary structure. Both groups use transesterification reactions in which the RNA is the catalytic agent (see *25 Catalytic RNA).*
- The removal of introns from yeast nuclear tRNA precursors involves enzymatic activities that handle the substrate in a way resembling the tRNA processing enzymes, in which a critical feature is the conformation of the tRNA precursor. These splicing reactions are accomplished by enzymes that use cleavage and ligation.

24.2 Nuclear splice junctions are short sequences

Key Concepts

- Splice sites are the sequences immediately surrounding the exon-intron boundaries. They are named for their positions relative to the intron.
- The 5' splice site at the 5' (left) end of the intron includes the consensus sequence GU.
- The 3' splice site at the 3' (right) end of the intron includes the consensus sequence AG.
- The GU-AG rule (originally called the GT-AG rule in terms of DNA sequence) describes the requirement for these constant dinucleotides at the first two and last two positions of introns in pre-mRNAs.

To focus on the molecular events involved in nuclear intron splicing, we must consider the nature of the splice sites, the two exon-intron boundaries that include the sites of breakage and reunion.

By comparing the nucleotide sequence of mRNA with that of the structural gene, the junctions between exons and introns can be assigned. There is no extensive homology or complementarity between the two ends of an intron. However, the junctions have well conserved, though rather short, consensus sequences.

It is possible to assign a specific end to every intron by relying on the conservation of exon-intron junctions. They can all be aligned to conform to the consensus sequence given in **Figure 24.3.**

The subscripts indicate the percent occurrence of the specified base at each consensus position. High conservation is found only *immediately within the intron* at the presumed junctions. This identifies the sequence of a generic intron as

GU. . . . . AG

Because the intron defined in this way starts with the dinucleotide GU and ends with the dinucleotide AG, the junctions are often described as conforming to the GT-AG rule. (This reflects the fact that the sequences were originally analyzed in terms of DNA, but of course the GT in the coding strand sequence of DNA becomes a GU in the RNA.)

Note that the two sites have different sequences and so they define the ends of the intron *directionally*. They are named proceeding from left to right along the intron as the 5' splice site (sometimes called the left or donor site) and the 3' splice site (also called the right or acceptor site). The consensus sequences are implicated as the sites recognized in splicing by point mutations that prevent splicing *in vivo* and *in vitro*.

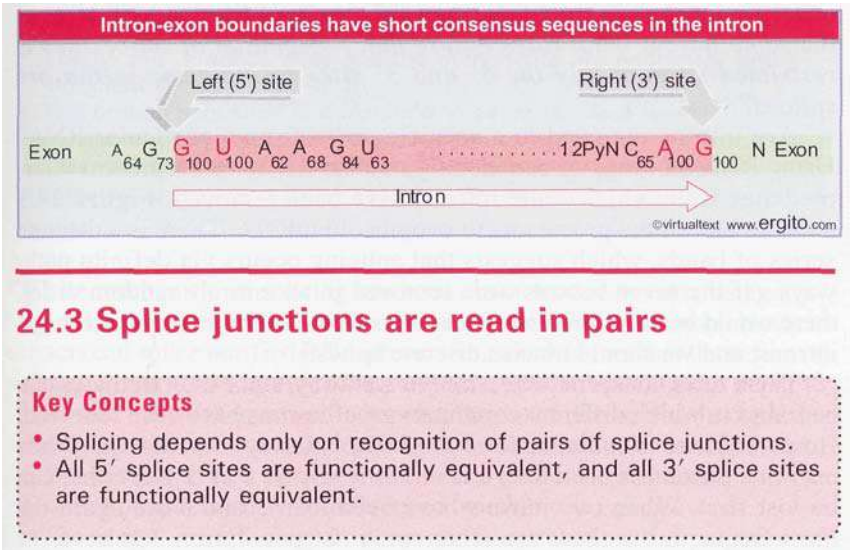


Figure 24.3 The ends of nuclear introns are defined by the GU-AG rule.

24.3 Splice junctions are read in pairs

Key Concepts

- Splicing depends only on recognition of pairs of splice junctions.
- All 5′ splice sites are functionally equivalent, and all 3′ splice sites are functionally equivalent.

A typical mammalian mRNA has many introns. The basic problem of pre-mRNA splicing results from the simplicity of the splice sites. This is illustrated in **Figure 24.4:** What ensures that the correct pairs of sites are spliced together? The corresponding GU-AG pairs must be connected across great distances (some introns are >10 kb long). We can imagine two types of principle that might be responsible for pairing the appropriate 5' and 3' sites:

- It could be an *intrinsic property* of the RNA to connect the sites at the ends of a particular intron. This would require matching of specific sequences or structures.
- Or all 5′ sites may be functionally equivalent and all 3′ sites may be similarly indistinguishable, but splicing could follow *rules* that ensure a 5′ site is always connected to the 3′ site that comes next in the RNA.

By Book_Crazy [IND]

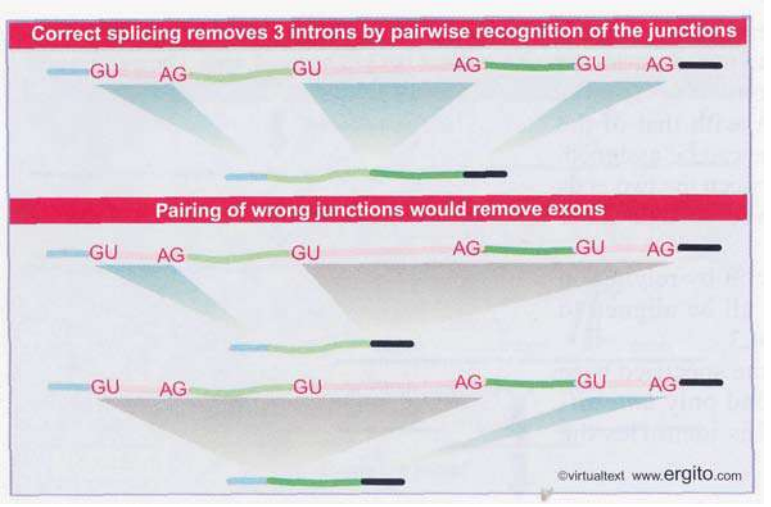


Figure 24.4 Splicing junctions are recognized only in the correct pairwise combinations.

Neither the splice sites nor the surrounding regions have any sequence complementarity, which excludes models for complementary base pairing between intron ends. And experiments using hybrid RNA precursors show that any 5' splice site can in principle be connected to any 3' splice site. For example, when the first exon of the early SV40 transcription unit is linked to the third exon of mouse β globin, the hybrid intron can be excised to generate a perfect connection between the SV40 exon and the β-globin exon. Indeed, this interchangeability is the basis for the exon-trapping technique described previously in Figure 2.12. Such experiments make two general points:

- *Splice sites are generic:* they do not have specificity for individual RNA precursors, and individual precursors do not convey specific information (such as secondary structure) that is needed for splicing.
- *The apparatus for splicing is not tissue specific;* an RNA can usually be properly spliced by any cell, whether or not it is usually synthesized in that cell. (We discuss exceptions in which there are tissue-specific alternative splicing patterns in *24.12 Alternative splicing involves differential use of splice junctions.*)

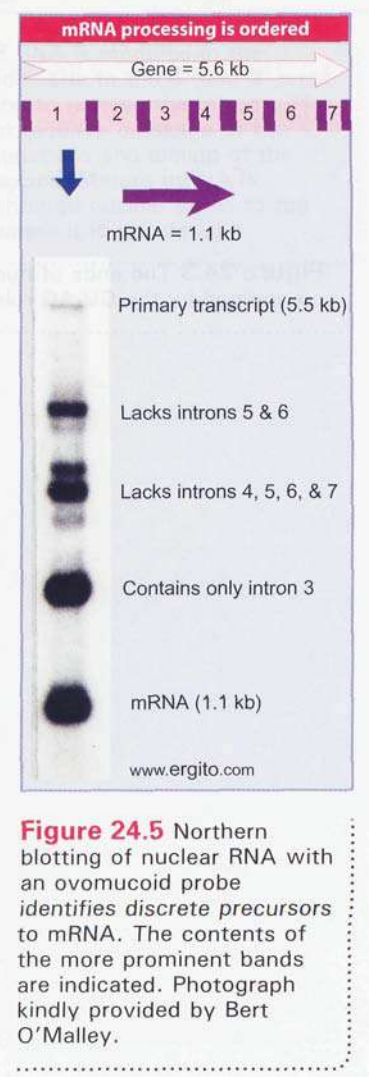


Figure 24.5 Northern blotting of nuclear RNA with an ovomucoid probe identifies discrete precursors to mRNA. The contents of the more prominent bands are indicated. Photograph kindly provided by Bert O'Malley.

Here is a paradox. Probably all 5' splice sites look similar to the splicing apparatus, and all 3' splice sites look similar to it. *In principle any 5' splice site may be able to react with any 3' splice site.* But in the usual circumstances splicing occurs only between the 5' and 3' sites of the *same* intron. *What rules ensure that recognition of splice sites is restricted so that only the 5' and 3' sites of the same intron are spliced?*

Are introns removed in a specific *order* from a particular RNA? Using RNA blotting, we can identify nuclear RNAs that represent intermediates from which some introns have been removed. Figure 24.5 shows a blot of the precursors to ovomucoid mRNA. There is a discrete series of bands, which suggests that splicing occurs via definite pathways. (If the seven introns were removed in an entirely random order, there would be more than 300 precursors with different combinations of introns, and we should not see discrete bands.)

There does not seem to be a *unique* pathway, since intermediates can be found in which different combinations of introns have been removed. However, there is evidence for a *preferred* pathway or pathways. When only one intron has been lost, it is virtually always 5 or 6. But either can be lost first. When two introns have been lost, 5 and 6 are again the most frequent, but there are other combinations. Intron 3 is never or very rarely lost at one of the first three splicing steps. From this pattern, we see that there is a preferred pathway in which introns are removed in the order 5/6, 7/4, 2/1,3. But there are other pathways, since (for example), there are some molecules in which 4 or 7 is lost last. A caveat in interpreting these results is that we do not have proof that all these intermediates actually lead to mature mRNA.

The general conclusion suggested by this analysis is that the conformation of the RNA influences the accessibility of the splice sites. As particular introns are removed, *the* conformation changes, and new pairs of splice sites become available. But the ability of the precursor to remove its introns in more than one order suggests that alternative conformations are available at each stage. Of course, the longer the molecule, the more structural options become available; and when we consider larger genes, it becomes difficult to see how specific secondary structures could

control the reaction. One important conclusion of this analysis is that *the reaction does not proceed sequentially along the precursor.*

A simple model to control recognition of splice sites would be for the splicing apparatus to act in a processive manner. Having recognized a 5′ site, the apparatus might scan the RNA in the appropriate direction until it meets the next 3' site. This would restrict splicing to adjacent sites. But this model is excluded by experiments that show that splicing can occur in *trans* as an intermolecular reaction under special circumstances (see *24.13 trans-splicing reactions use small RNAs)* or in RNA molecules in which part of the nucleotide chain is replaced by a chemical linker. This means that there cannot be a requirement for strict scanning along the RNA from the 5' splice site to the 3' splice site. Another problem with the scanning model is that it cannot explain the existence of alternative splicing patterns, where (for example) a common 5' site is spliced to more than one 3' site. The basis for proper recognition of correct splice site pairs remains incompletely defined.

24.4 pre-mRNA splicing proceeds through a lariat

Key Concepts

- A lariat is formed when the intron is cleaved at the 5' splice site, and the 5' end is joined to a 2' position at an A at the branch site in the intron.
- The intron is released as a lariat when it is cleaved at the 3' splice site, and the left and right exons are then ligated together.
- The 5' and 3' splice sites and the branch site are necessary and sufficient for splicing.
- The branch sequence is conserved in yeast but less well conserved in higher eukaryotes.
- The reactions occur by transesterifications in which a bond is transferred from one location to another.

The mechanism of splicing has been characterized *in vitro,* using systems in which introns can be removed from RNA precursors. Nuclear extracts can splice purified RNA precursors, which shows that the action of splicing is not linked to the process of transcription. Splicing can occur to RNAs that are neither capped nor polyadenylated. However, although the splicing reaction as such is independent of transcription or modification to the RNA, these events normally occur in a coordinated manner, and the efficiency of splicing may be influenced by other processing events.

The stages of splicing *in vitro* are illustrated in the pathway of Figure 24.6. We discuss the reaction in terms of the individual RNA species that can be identified, but remember that *in vivo* the species containing exons are not released as free molecules, but remain held together by the splicing apparatus.

The first step is to make a cut at the 5′ splice site, separating the left exon and the right intron-exon molecule. The left exon takes the form of a linear molecule. The right intron-exon molecule forms a lariat, in which the 5' terminus generated at the end of the intron becomes linked by a 5′–2′ bond to a base within the intron. The target base is an A in a sequence that is called the branch site.

Cutting at the 3' splice site releases the free intron in lariat form, while the right exon is ligated (spliced) to the left exon. The cleavage and ligation reactions are shown separately in the figure for illustrative purposes, but actually occur as one coordinated transfer.

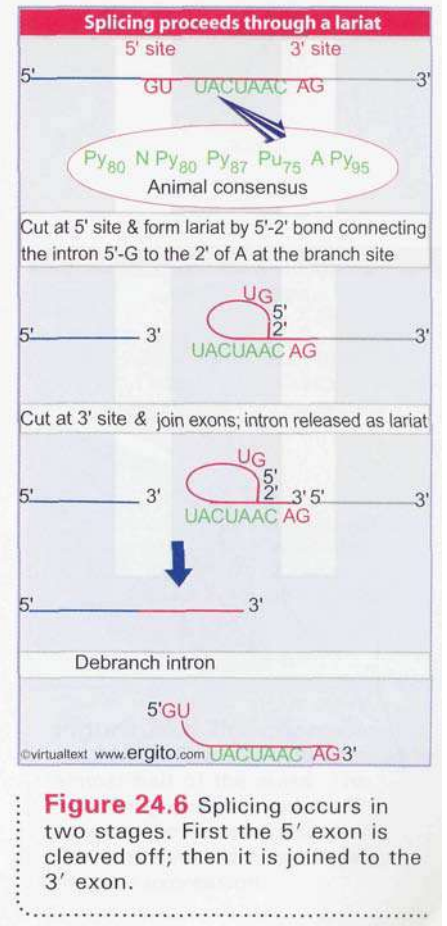


Figure 24.6 Splicing occurs in two stages. First the 5′ exon is cleaved off; then it is joined to the 3′ exon.

By Book_Crazy [IND]

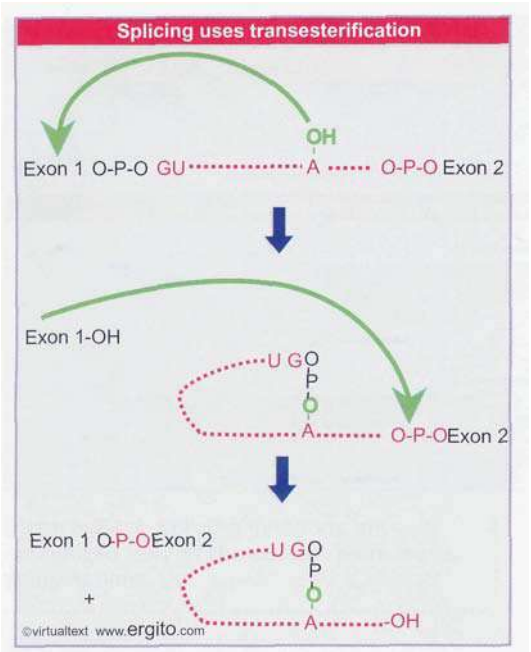


Figure 24.7 Nuclear splicing occurs by two transesterification reactions in which an OH group attacks a phosphodiester bond.

The lariat is then "debranched" to give a linear excised intron, which is rapidly degraded.

The sequences needed for splicing are the short consensus sequences at the 5′ and 3′ splice sites and at the branch site. Together with the knowledge that most of the sequence of an intron can be deleted without impeding splicing, this indicates that there is no demand for specific conformation in the intron (or exon).

The branch site plays an important role in identifying the 3′ splice site. The branch site in yeast is highly conserved and has the consensus sequence UACUAAC. The branch site in higher eukaryotes is not well conserved, but has a preference for purines or pyrimidines at each position and retains the target A nucleotide (see Figure 24.6).

The branch site lies 18-40 nucleotides upstream of the 3′ splice site. Mutations or deletions of the branch site in yeast prevent splicing. In higher eukaryotes, the relaxed constraints in its sequence result in the ability to use related sequences (called cryptic sites) when the authentic branch is deleted. Proximity to the 3′ splice site appears to be important, since the cryptic site is always close to the authentic site. A cryptic site is used only when the branch site has been inactivated. When a cryptic branch sequence is used in this manner, splicing otherwise appears to be normal; and the exons give the same products as wild type. *The role of the branch site therefore is to identify the nearest 3′ splice site as the target for connection to the 5′ splice site.* This can be explained by the fact that an interaction occurs between protein complexes that bind to these two sites.

The bond that forms the lariat goes from the 5′ position of the invariant G that was at the 5′ end of the intron to the 2′ position of the invariant A in the branch site. This corresponds to the third A residue in the yeast UACUAAC box.

The chemical reactions proceed by transesterification: a bond is in effect *transferred* from one location to another. **Figure 24.7** shows that the first step is a nucleophilic attack by the 2′–OH of the invariant A of the UACUAAC sequence on the 5′ splice site. In the second step, the free 3′–OH of the exon that was released by the first reaction now attacks the bond at the 3′ splice site. Note that the number of phosphodiester bonds is conserved. There were originally two 5′–3′ bonds at the exon-intron splice sites; one has been replaced by the 5′–3′ bond between the exons, and the other has been replaced by the 5′–2′ bond that forms the lariat.

24.5 snRNAs are required for splicing

Key Concepts

- The five snRNPs involved in splicing are U1, U2, U5, U4, and U6.
- Together with some additional proteins, the snRNPs form the spliceosome.
- All the snRNPs except U6 contain a conserved sequence that binds the Sm proteins that are recognized by antibodies generated in autoimmune disease.

The 5′ and 3′ splice sites and the branch sequence are recognized by components of the splicing apparatus that assemble to form a large complex. This complex brings together the 5′ and 3′ splice sites before any reaction occurs, explaining why a deficiency in any one of the sites may prevent the reaction from initiating. The complex assembles sequentially on the pre-mRNA, and several intermediates can be recognized by fractionating complexes of different sizes. Splicing occurs only after all the components have assembled.

By Book_Crazy [IND]

The splicing apparatus contains both proteins and RNAs (in addition to the pre-mRNA). The RNAs take the form of small molecules that exist as ribonucleoprotein particles. Both the nucleus and cytoplasm of eukaryotic cells contain many discrete small RNA species. They range in size from 100-300 bases in higher eukaryotes, and extend in length to ~1000 bases in yeast. They vary considerably in abundance, from 10^5-10^6 molecules per cell to concentrations too low to be detected directly.

Those restricted to the nucleus are called small nuclear RNAs (snRNA); those found in the cytoplasm are called small cytoplasmic RNAs (scRNA). In their natural state, they exist as ribonucleoprotein particles (snRNP and scRNP). Colloquially, they are sometimes known as snurps and scyrps. There is also a class of small RNAs found in the nucleolus, called snoRNAs, which are involved in processing ribosomal RNA (see *24.22 Small RNAs are required for rRNA processing).*

The snRNPs involved in splicing, together with many additional proteins, form a large particulate complex called the spliceosome. Isolated from the *in vitro* splicing systems, it comprises a 50-60S ribonucleoprotein particle. The spliceosome may be formed in stages as the snRNPs join, proceeding through several "presplicing complexes." The spliceosome is a large body, greater in mass than the ribosome.

Figure 24.8 summarizes the components of the spliceosome. The 5 snRNAs account for more than a quarter of the mass; together with their 45 associated proteins, they account for almost half of the mass. Some 70 other proteins found in the spliceosome are described as splicing factors. They include proteins required for assembly of the spliceosome, proteins required for it to bind to the RNA substrate, and proteins involved in the catalytic process. In addition to these proteins, another ~30 proteins associated with the spliceosome have been implicated in acting at other stages of gene expression, suggesting that the spliceosome may serve as a coordinating apparatus.

The spliceosome forms on the intact precursor RNA and passes through an intermediate state in which it contains the individual 5' exon linear molecule and the right lariat-intron-exon. Little spliced product is found in the complex, which suggests that it is usually released immediately following the cleavage of the 3' site and ligation of the exons.

We may think of the snRNP particles as being involved in building the structure of the spliceosome. Like the ribosome, the spliceosome depends on RNA-RNA interactions as well as protein-RNA and protein-protein interactions. Some of the reactions involving the snRNPs require their RNAs to base pair directly with sequences in the RNA being spliced; other reactions require recognition between snRNPs or between their proteins and other components of the spliceosome.

The importance of snRNA molecules can be tested directly in yeast by making mutations in their genes. Mutations in 5 snRNA genes are lethal and prevent splicing. All of the snRNAs involved in splicing can be recognized in conserved forms in animal, bird, and insect cells. The corresponding RNAs in yeast are often rather larger, but conserved regions include features that are similar to the snRNAs of higher eukaryotes.

The snRNPs involved in splicing are U1, U2, U5, U4, and U6. They are named according to the snRNAs that are present. Each snRNP contains a single snRNA and several (<20) proteins. The U4 and U6 snRNPs are usually found as a single (U4/U6) particle. A common structural core for each snRNP consists of a group of 8 proteins, all of which are recognized by an autoimmune antiserum called anti-Sm; conserved sequences in the proteins form the target for the antibodies. The other proteins in each snRNP are unique to it. The Sm proteins bind to the conserved sequence $PuAU_{3-6}Gpu$, which is present in all snRNAs

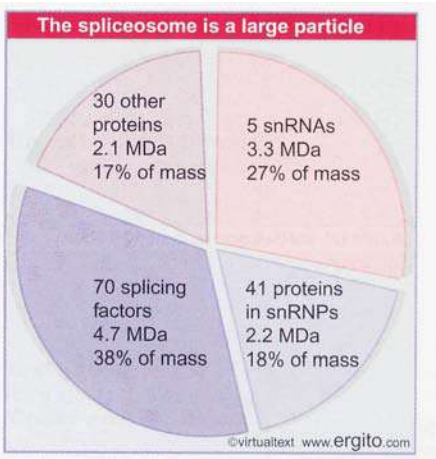


Figure 24.8 The spliceosome is ~12 MDa. Five snRNAPs account for almost half of the mass. The remaining proteins include known splicing factors and also proteins that are involved in other stages of gene expression.

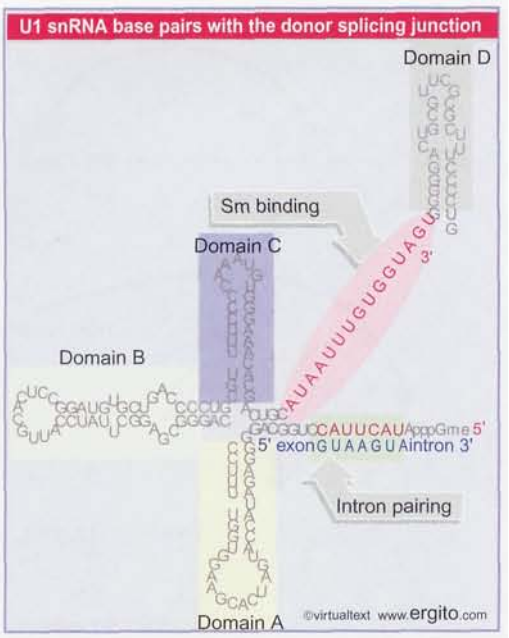


Figure 24.9 U1 snRNA has a base-paired structure that creates several domains. The 5′ end remains single stranded and can base pair with the 5′ splicing site.

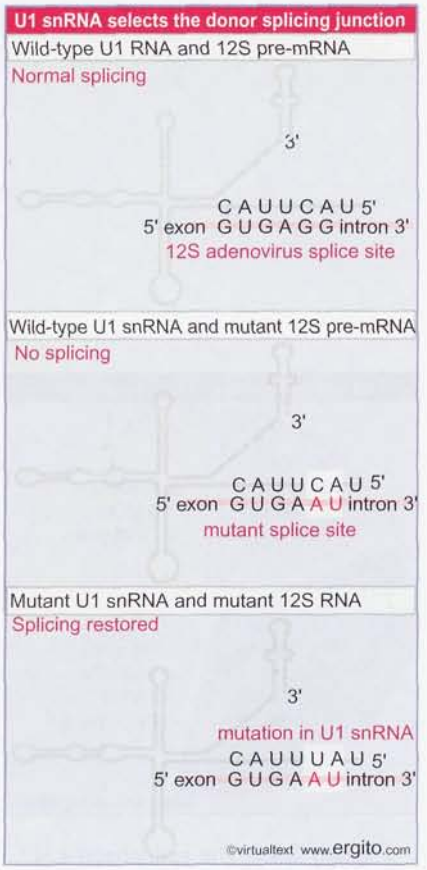


Figure 24.10 Mutations that abolish function of the 5′ splicing site can be suppressed by compensating mutations in U1 snRNA that restore base pairing.

except U6. The U6 snRNP contains instead a set of Sm-like (Lsm) proteins. The Sm proteins must be involved in the autoimmune reaction, although their relationship to the phenotype of the autoimmune disease is not clear.

Some of the proteins in the snRNPs may be involved directly in splicing; others may be required in structural roles or just for assembly or interactions between the snRNP particles. About one third of the proteins involved in splicing are components of the snRNPs. Increasing evidence for a direct role of RNA in the splicing reaction suggests that relatively few of the splicing factors play a direct role in catalysis; most are involved in structural or assembly roles.

24.6 U1 snRNP initiates splicing

Key Concepts

- U1 snRNP initiates splicing by binding to the 5' splice site by means of an RNA-RNA pairing reaction.
- The E complex contains U1 snRNP bound at the 5' splice site, the protein U2AF bound to a pyrimidine tract between the branch site and the 3' splice site, and SR proteins connecting U1 snRNP to U2AF.

plicing can be broadly divided into two stages:

- First the consensus sequences at the 5' splice site, branch sequence, and adjacent pyrimidine tract are recognized. A complex assembles that contains all of the splicing components.
- Then the cleavage and ligation reactions change the structure of the substrate RNA. Components of the complex are released or reorganized as it proceeds through the splicing reactions.

The important point is that all of the splicing components are assembled and have assured that the splice sites are available before any irreversible change is made to the RNA.

Recognition of the consensus sequences involves both RNAs and proteins. Certain snRNAs have sequences that are complementary to the consensus sequences or to one another, and base pairing between snRNA and pre-mRNA, or between snRNAs, plays an important role in splicing.

The human U1 snRNP contains 8 proteins as well as the RNA. The secondary structure of the Ul snRNA is drawn in **Figure 24.9.** It contains several domains. The Sm-binding site is required for interaction with the common snRNP proteins. Domains identified by the individual stem-loop structures provide binding sites for proteins that are unique to U1 snRNP.

Binding of Ul snRNP to the 5' splice site is the first step in splicing. The recruitment of Ul snRNP involves an interaction between one of its proteins (Ul-70k) and the protein ASF/SF2 (a general splicing factor in the SR class: see below). Ul snRNA base pairs with the 5' site by means of a single-stranded region at its 5′-terminus which usually includes a stretch of 4-6 bases that is complementary with the splice site.

Mutations in the 5' splice site and Ul snRNA can be used to test directly whether pairing between them is necessary. The results of such an experiment are illustrated in **Figure 24.10.** The wild-type sequence of the splice site of the 12S adenovirus pre-mRNA pairs at 5 out of 6

positions with U1 snRNA. A mutant in the 12S RNA that cannot be spliced has two sequence changes; the GG residues at positions 5-6 in the intron are changed to AU. The mutation changes the pattern of base pairing between Ul snRNA and the 5' splice site, although it does not alter the *overall* extent of pairing (because complementarity is lost at one position and gained at the other). The effect on splicing suggests that the base-pairing interaction is important.

When a mutation is introduced into U1 snRNA that restores pairing at position 5, normal splicing is regained. Other cases in which corresponding mutations are made in Ul snRNA to see whether they can suppress the mutation in the splice site suggests the general rule: complementarity between Ul snRNA and the 5' splice site is necessary for splicing, but the efficiency of splicing is not determined solely by the number of base pairs that can form. The pairing reaction is stabilized by the proteins of the Ul snRNP.

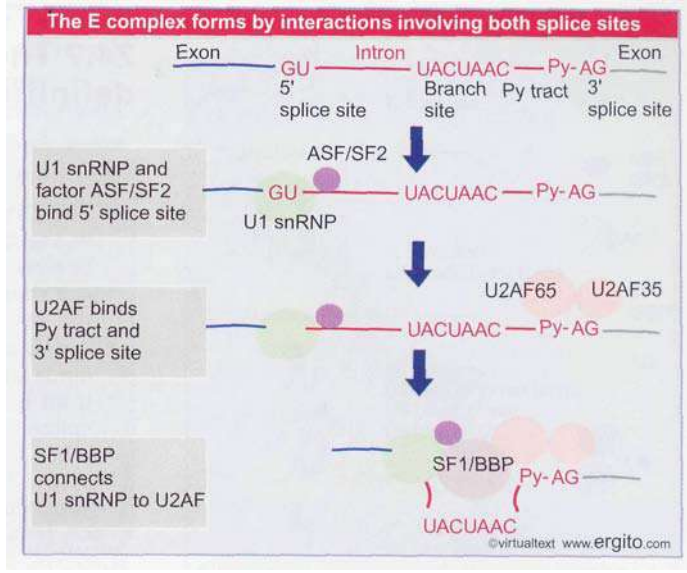


Figure 24.11 The commitment (E) complex forms by the successive addition of U1 snRNP to the 5' splice site, U2AF to the pyrimidine tract/3' splice site, and the bridging protein SF1/BBP.

Figure 24.11 shows the early stages of splicing. The first complex formed during splicing is the E (early presplicing) complex, which contains Ul snRNP, the splicing factor U2AF, and members of a family called SR proteins, which comprise an important group of splicing factors and regulators. They take their name from the presence of an Arg-Ser-rich region that is variable in length. SR proteins interact with one another via their Arg-Ser-rich regions. They also bind to RNA. They are an essential component of the spliceosome, forming a framework on the RNA substrate. They connect U2AF to Ul (see Figure 24.12). The E complex is sometimes called the commitment complex, because its formation identifies a pre-mRNA as a substrate for formation of the splicing complex.

In the E complex, U2AF is bound to the region between the branch site and the 3' splice site. The name of U2AF reflects its original isolation as the U2 auxiliary factor. In most organisms, it has a large subunit (U2AF65) that contacts a pyrimidine tract downstream of the branch site, while a small subunit (U2AF35) directly contacts the dinucleotide AG at the 3' splice site. In *S. cerevisiae*, this function is filled by the protein Mud2, which is a counterpart of U2AF65, and binds only to the pyrimidine tract. This marks a difference in the mechanism of splicing between *S. cerevisiae* and other organisms. In the yeast, the 3' splice site is not involved in the early stages of forming the splicing complex, but in all other known cases, it is required.

Another splicing factor, called SF1 in mammals and BBP in yeast, connects U2AF/Mud2 to the Ul snRNP bound at the 5' splice site. Complex formation is enhanced by the cooperative reactions of the two proteins; SF1 and U2AF (or BBP and Mud2) bind together to the RNA substrate ~10× more effectively than either alone. This interaction is probably responsible for making the first connection between the two splice sites across the intron.

The E complex is converted to the A complex when U2 snRNP binds to the branch site. Both Ul snRNP and U2AF/Mud2 are needed for U2 binding. The U2 snRNA includes sequences complementary to the branch site. A sequence near the 5' end of the snRNA base pairs with the branch sequence in the intron. In yeast this typically involves formation of a duplex with the UACUAAC box (see Figure 24.14). Several proteins of the U2 snRNP are bound to the substrate RNA just upstream of the branch site. The addition of U2 snRNP to the E complex generates the A presplicing complex. The binding of U2 snRNP requires ATP hydrolysis and commits a pre-mRNA to the splicing pathway.

24.7 The E complex can be formed by intron definition or exon definition

Key Concepts

- The direct way of forming an E complex is for U1 snRNP to bind at the 5' splice site and U2AF to bind at a pyrimidine tract between the branch site and the 3' splice site.
- Another possibility is for the complex to form between U2AF at the pyrimidine tract and U1 snRNP at a downstream 5' splice site.
- The E complex is converted to the A complex when U2 snRNP binds at the branch site.
- If an E complex forms using a downstream 5' splice site, this splice site is replaced by the appropriate upstream 5' splice site when the E complex is converted to the A complex.
- Weak 3' splice sites may require a splicing enhancer located in the exon downstream to bind SR proteins directly.

There is more than one way to form the E complex. **Figure 24.12** illustrates some possibilities. The most direct reaction is for both splice sites to be recognized across the intron. The presence of U1 snRNP at the 5′ splice site is necessary for U2AF to bind at the pyrimidine tract downstream of the branch site, making it possible that the 5' and 3' ends of the intron are brought together in this complex. The E complex is converted to the A complex when U2 snRNP binds at the branch site. *The basic feature of this route for splicing is that the two splice sites are recognized without requiring any sequences outside of the intron.* This process is called intron definition.

In an extreme case, the SR proteins may enable U2AF/U2 snRNP to bind *in vitro* in the absence of U1, raising the possibility that there could be a U1-independent pathway for splicing.

An alternative route to form the spliceosome may be followed when the introns are long and the splice sites are weak. As shown on the right of the figure, the 5' splice site is recognized by Ul snRNA in the usual way. However, the 3′ splice site is recognized as part of a complex that forms across the *next exon,* in which the next 5' splice site is also bound by Ul snRNA. This Ul snRNA is connected by SR proteins to the U2AF at the pyrimidine tract. When U2 snRNP joins to generate the A complex, there is a rearrangement in which the correct (leftmost) 5′splice site displaces the downstream 5′ splice site in the complex. The important feature of this route for splicing is that sequences downstream of the intron itself are

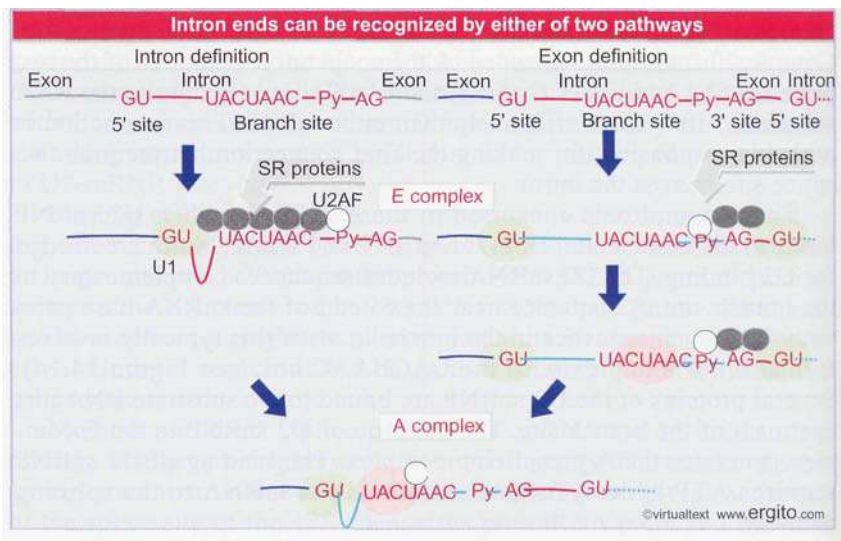


Figure 24.12 There may be multiple routes for initial recognition of 5' and 3' splice sites.

By Book_Crazy [IND]

required. Usually these sequences include the next 5' splice site. This process is called exon definition. This mechanism is not universal; neither SR proteins nor exon definition are found in *S. cerevisiae*.

"Weak" 3' splice sites do not bind U2AF and U2 snRNP effectively. Additional sequences are needed to bind the SR proteins, which assist U2AF in binding to the pyrimidine tract. Such sequences are called "splicing enhancers," and they are most commonly found in the exon downstream of the 3' splice site.

24.8 5 snRNPs form the spliceosome

Key Concepts

- Binding of U5 and U4/U6 snRNPs converts the A complex to the B1 spliceosome, which contains all the components necessary for splicing.
- The spliceosome passes through a series of further complexes as splicing proceeds.
- Release of U1 snRNP allows U6 snRNA to interact with the 5' splice site and converts the B1 spliceosome to the B2 spliceosome.
- When U4 dissociates from U6 snRNP, U6 snRNA can pair with U2 snRNA to form the catalytic active site.

Following formation of the E complex, the other snRNPs and factors involved in splicing associate with the complex in a defined order. Figure 24.13 shows the components of the complexes that can be identified as the reaction proceeds.

The B1 complex is formed when a trimer containing the U5 and U4/U6 snRNPs binds to the A complex containing U1 and U2 snRNPs. This complex is regarded as a spliceosome, since it contains the components needed for the splicing reaction. It is converted to the B2 complex after U1 is released. The dissociation of Ul is necessary to allow other components to come into juxtaposition with the 5' splice site, most notably U6 snRNA. At this point U5 snRNA changes its position; initially it is close to exon sequences at the 5' splice site, but it shifts to the vicinity of the intron sequences.

The catalytic reaction is triggered by the release of U4; this requires hydrolysis of ATP. The role of U4 snRNA may be to sequester U6 snRNA until it is needed. Figure 24.14shows the changes that occur in the base-pairing interactions between snRNAs during splicing. In the U6/U4 snRNP, a continuous length of 26 bases of U6 is paired with two separated regions of U4. When U4 dissociates, the region in U6 that is released becomes free to take up another structure. The first part of it pairs with U2; the second part forms an intramolecular hairpin. The interaction between U4 and U6 is mutually incompatible with the interaction between U2 and U6, so the release of U4 controls the ability of the spliceosome to proceed.

Although for clarity the figure shows the RNA substrate in extended form, the 5' splice site is actually close to the U6 sequence immediately on the 5' side of the stretch bound to U2. This sequence in U6 snRNA pairs with sequences in the intron just downstream of the conserved GU at the 5' splice site (mutations that enhance such pairing improve the efficiency of splicing).

So several pairing reactions between snRNAs and the substrate RNA occur in the course of splicing. They are summarized in Figure 24.15. The snRNPs have sequences that pair with the substrate and with one

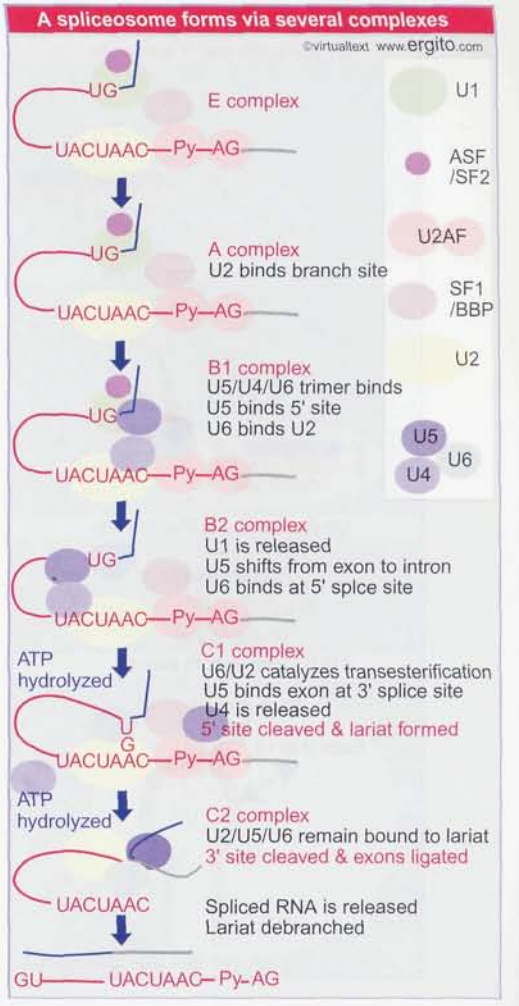


Figure 24.13 The splicing reaction proceeds through discrete stages in which spliceosome formation involves the interaction of components that recognize the consensus sequences.

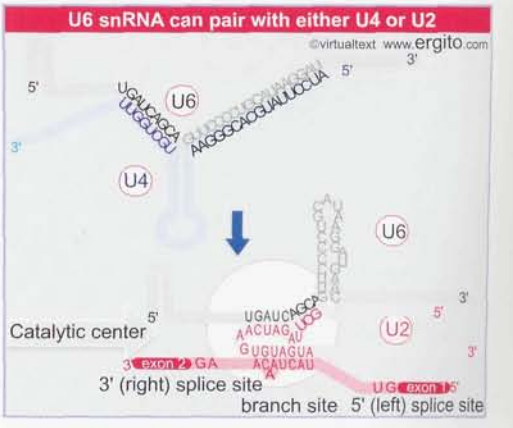


Figure 24.14 U6-U4 pairing is incompatible with U6-U2 pairing. When U6 joins the spliceosome it is paired with U4. Release of U4 allows a conformational change in U6; one part of the released sequence forms a hairpin (dark grey), and the other part (black) pairs with U2.

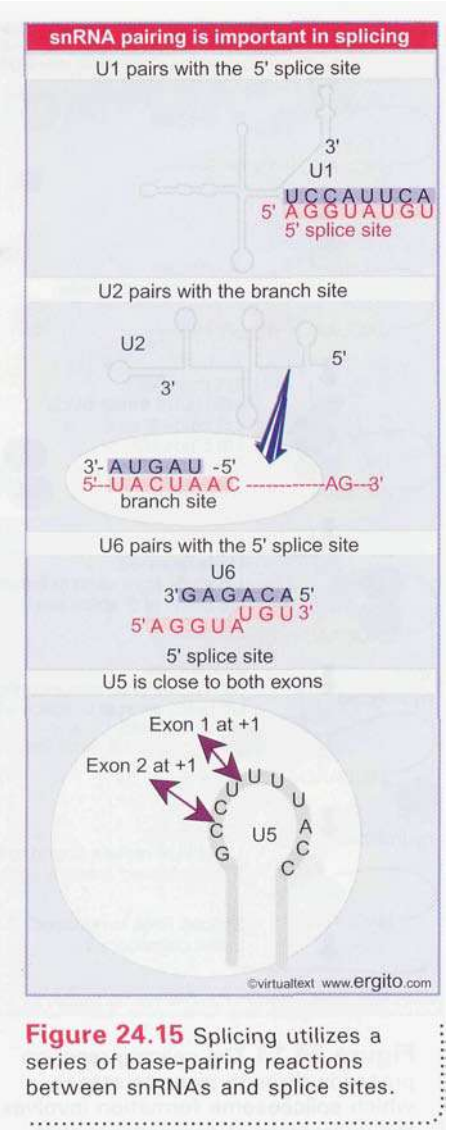


Figure 24.15 Splicing utilizes a series of base-pairing reactions between snRNAs and splice sites.

another. They also have single-stranded regions in loops that are in close proximity to sequences in the substrate, and which play an important role, as judged by the ability of mutations in the loops to block splicing.

The base pairing between U2 and the branch point, and between U2 and U6, creates a structure that resembles the active center of group II self-splicing introns (see Figure 24.20). This suggests the possibility that the catalytic component could comprise an RNA structure generated by the U2-U6 interaction. U6 is paired with the 5' splice site, and crosslinking experiments show that a loop in U5 snRNA is immediately adjacent to the first base positions in both exons. But although we can define the proximities of the substrate (5' splice site and branch site) and snurps (U2 and U6) at the catalytic center (as shown in Figure 24.14), the components that undertake the transesterifications have not been directly identified.

The formation of the lariat at the branch site is responsible for determining the use of the 3' splice site, since the 3' consensus sequence nearest to the 3′ side of the branch becomes the target for the second transesterification. The second splicing reaction follows rapidly. Binding of U5 snRNP to the 3′ splice site is needed for this reaction, but there is no evidence for a base pairing reaction.

The important conclusion suggested by these results is that *the snRNA components of the splicing apparatus interact both among themselves and with the substrate RNA by means of base pairing interactions, and these interactions allow for changes in structure that may bring reacting groups into apposition and may even create catalytic centers.* Furthermore, the conformational changes in the snRNAs are reversible; for example, U6 snRNA is not used up in a splicing reaction, and at completion must be released from U2, so that it can reform the duplex structure with U4 to undertake another cycle of splicing.

We have described individual reactions in which each snRNP participates, but as might be expected from a complex series of reactions, any particular snRNP may play more than one role in splicing. So the ability of U1 snRNP to promote binding of U2 snRNP to the branch site is independent of its ability to bind to the 5' splice site. Similarly, different regions of U2 snRNA can be defined that are needed to bind to the branch site and to interact with other splicing components.

An extensive mutational analysis has been undertaken in yeast to identify both the RNA and protein components of the spliceosome. Mutations in genes needed for splicing are identified by the accumulation of unspliced precursors. A series of loci that identify genes potentially coding for proteins involved in splicing were originally called *RNA*, but are now known as *PRP* mutants (for pre-RNA processing). Several of the products of these genes have motifs that identify them as RNA-binding proteins, and some appear to be related to a family of ATP-dependent RNA helicases. We suppose that, in addition to RNA-RNA interactions, protein-RNA interactions are important in creating or releasing structures in the pre-mRNA or snRNA components of the spliceosomes.

Some of the PRP proteins are components of snRNP particles, but others function as independent factors. One interesting example is PRP 16, a helicase that hydrolyzes ATP and associates transiently with the spliceosome to participate in the second catalytic step. Another example is PRP22, another ATP-dependent helicase, which is required to release the mature mRNA from the spliceosome. The conservation of bonds during the splicing reaction means that input of energy is not required to drive bond formation *per se*, which implies that the ATP hydrolysis is required for other purposes. The use of ATP by PRP 16 and PRP22 may be examples of a more general phenomenon: the use of ATP hydrolysis to drive conformational changes that are needed to proceed through splicing.

24.9 An alternative splicing apparatus uses different snRNPs

Key Concepts

- An alternative splicing pathway uses another set of snRNPs that comprise the U12 spliceosome.
- The target introns are defined by longer consensus sequences at the splice junctions, but usually include the same GU-AG junctions.
- Some introns have the splice junctions AU-AC, including some that are U1-dependent and some that are U12-dependent.

GU-AG introns comprise the vast majority (>98% of splicing junctions in the human genome). <1% use the related junctions GC-AG. And then there is a minor class of introns marked by the ends AU-AC (comprising 0.1% of introns). The first of these introns to be discovered required an alternative splicing apparatus, called the U12 spliceosome, consisting of U11 and U12 (related to U1 and U2, respectively), a U5 variant, and the $U4_{atac}$ and $U6_{atac}$ snRNAs. The splicing reaction is essentially similar to that at GU-AG introns, and the snRNAs play analogous roles. Whether there are differences in the protein components of this apparatus is not known.

It now turns out that the dependence on the type of spliceosome is also influenced by sequences in the intron, so that there are some AU-AC introns spliced by U2-type spliceosomes, and some GU-AG introns spliced by U12-type spliceosomes. A strong consensus sequence at the left end defines the U12-dependent type of intron: 5' $^{G}_{A}$UAUCCUUU... ...PyA$^{G}_{C}$ 3'. In fact, most U12-dependent introns have the GU....AG termini. In addition, they have a highly conserved branch point, UCCU-UPuAPy, which pairs with U12. For this reason, the term U12-dependent intron is used rather than AU-AC intron.

The two types of introns coexist in a variety of genomes, and in some cases are found in the same gene. U12-dependent introns tend to be flanked by U2-dependent introns. What is known about the phylogeny of these introns suggests that AU-AC U12-dependent introns may once have been more common, but tend to be converted to GU-AG termini, and to U2-dependence, in the course of evolution. The common evolution of the systems is emphasized by the fact that they use analogous sets of base pairing between the snRNAs and with the substrate pre-mRNA.

The involvement of snRNPs in splicing is only one example of their involvement in RNA processing reactions. snRNPs are required for several reactions in the processing of nuclear RNA to mature rRNAs. Especially in view of the demonstration that group I introns are self-splicing, and that the RNA of ribonuclease P has catalytic activity (as discussed in *25 Catalytic RNA)*, it is plausible to think that RNA-RNA reactions are important in many RNA processing events.

24.10 Splicing is connected to export of mRNA

Key Concepts

- The REF proteins bind to splicing junctions by associating with the spliceosome.
- After splicing, they remain attached to the RNA at the exon-exon junction.
- They interact with the transport protein TAP/Mex that exports the RNA through the nuclear pore.

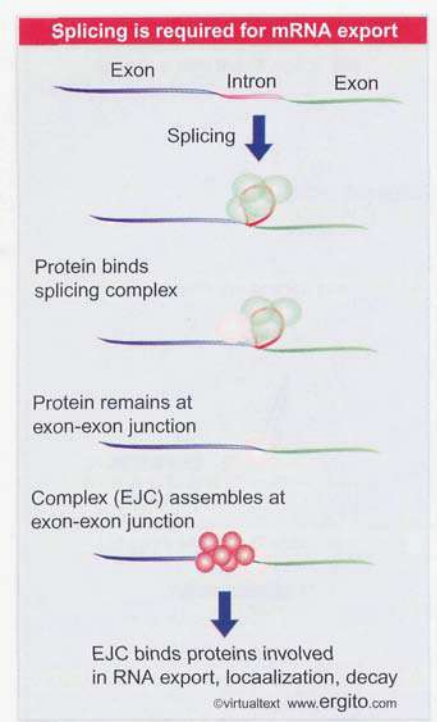


Figure 24.16 The EJC (exon junction complex) binds to RNA by recognizing the splicing complex.

After it has been synthesized and processed, mRNA is exported from the nucleus to the cytoplasm in the form of a ribonucleoprotein complex. The proteins that are responsible for transport "shuttle" between the nucleus and cytoplasm, remaining in the compartment only briefly (see *8.28 Transport receptors carry cargo proteins through the pore)*. One important question is how these proteins recognize their RNA substrates, and what ensures that only fully processed mRNAs are exported. Part of the answer may lie in the relative timing of events: spliceosomes may form to remove introns before transcription has been completed. However, there may also be a direct connection between splicing and export.

Introns may prevent export of mRNA because they are associated with the splicing apparatus. The spliceosome also may provide the initial point of contact for the export apparatus. **Figure 24.16** shows a model in which a protein complex binds to the RNA via the splicing apparatus. The complex consists of >9 proteins and is called the EJC (exon junction complex).

The EJC is involved in several functions of spliced mRNAs. Some of the proteins of the EJC are directly involved in these functions, and others recruit additional proteins for particular functions: The first contact in assembling the EJC is made with one of the splicing factors. Then after splicing, the EJC remains attached to the mRNA just upstream of the exon-exon junction. The EJC is not associated with RNAs transcribed from genes that lack introns, so its involvement in the process is unique for spliced products.

If introns are deleted from a gene, its RNA product is exported much more slowly to the cytoplasm. This suggests that the intron may provide a signal for attachment of the export apparatus. We can now account for this phenomenon in terms of a series of protein interactions, as shown in **Figure 24.17.** The EJC includes a group of proteins called the REF family (the best characterized member is called Aly). The REF proteins in turn interact with a transport protein (variously called TAP and Mex) which has direct responsibility for interaction with the nuclear pore.

A similar system may be used to identify a spliced RNA so that nonsense mutations prior to the last exon trigger its degradation in the cytoplasm (see *5.14 Nonsense mutations trigger a surveillance system).*

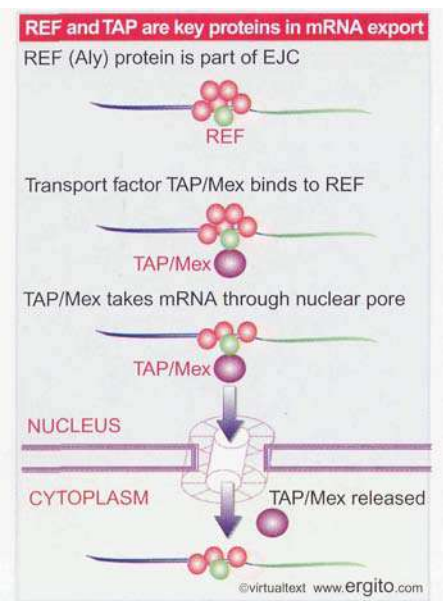


Figure 24.17 A REF protein binds to a splicing factor and remains with the spliced RNA product. REF binds to an export factor that binds to the nuclear pore.

24.11 Group II introns autosplice via lariat formation

Key Concepts

- Group II introns excise themselves from RNA by an autocatalytic splicing event.
- The splice junctions and mechanism of splicing of group II introns are similar to splicing of nuclear introns.
- A group II intron folds into a secondary structure that generates a catalytic site resembling the structure of U6-U2-nuclear intron.

Introns in protein-coding genes (in fact, in all genes except nuclear tRNA-coding genes) can be divided into three general classes. Nuclear pre-mRNA introns are identified only by the possession of the GU...AG dinucleotides at the 5' and 3' ends and the branch site/pyrimidine tract near the 3′ end. They do not show any common features of secondary structure. Group I and group II introns are found in organelles and in bacteria. (Group I introns are found also in the nucleus in lower eukaryotes.) Group I and group II introns are classified according to their

internal organization. Each can be folded into a typical type of secondary structure.

The group I and group II introns have the remarkable ability to excise themselves from an RNA. This is called autosplicing. Group I introns are more common than group II introns. There is little relationship between the two classes, but in each case the RNA can perform the splicing reaction *in vitro* by itself, without requiring enzymatic activities provided by proteins; however, proteins are almost certainly required *in vivo* to assist with folding (see *25 Catalytic RNA).*

Figure 24.18 shows that three classes of introns are excised by two successive transesterifications (shown previously for nuclear introns in Figure 24.6). In the first reaction, the 5' exon-intron junction is attacked by a free hydroxyl group (provided by an internal 2′–OH position in nuclear and group II introns, and by a free guanine nucleotide in group I introns). In the second reaction, the free 3′–OH at the end of the released exon in turn attacks the 3' intron-exon junction.

There are parallels between group II introns and pre-mRNA splicing. Group II mitochondrial introns are excised by the same mechanism as nuclear pre-mRNAs, via a lariat that is held together by a 5′–2′ bond. An example of a lariat produced by splicing a group II intron is shown in **Figure 24.19.** When an isolated group II RNA is incubated *in vitro* in the absence of additional components, it is able to perform the splicing reaction. This means that the two transesterification reactions shown in Figure 24.18 can be performed by the group II intron RNA sequence itself. Because the number of phosphodiester bonds is conserved in the reaction, an external supply of energy is not required; this could have been an important feature in the evolution of splicing.

A group II intron forms into a secondary structure that contains several domains formed by base-paired stems and single-stranded loops. Domain 5 is separated by 2 bases from domain 6, which contains an A residue that donates the 2′–OH group for the first transesterification. This constitutes a catalytic domain in the RNA. **Figure 24.20** compares this secondary structure with the structure formed by the combination of U6 with U2 and of U2 with the branch site. The similarity suggests that U6 may have a catalytic role.

The features of group II splicing suggest that splicing evolved from an autocatalytic reaction undertaken by an individual RNA molecule, in which it accomplished a controlled deletion of an internal sequence. Probably such a reaction requires the RNA to fold into a specific conformation, or series of conformations, and would occur exclusively in *cis* conformation.

The ability of group II introns to remove themselves by an autocatalytic splicing event stands in great contrast to the requirement of nuclear introns for a complex splicing apparatus. We may regard the snRNAs of the spliceosome as compensating for the lack of sequence information in the intron and providing the information required to form particular structures in RNA. The functions of the snRNAs may have evolved from the original autocatalytic system. These snRNAs act in *trans* upon the substrate pre-mRNA; we might imagine that the ability of U1 to pair with the 5' splice site, or of U2 to pair with the branch sequence, replaced a similar reaction that required the relevant sequence to be carried by the intron. So the snRNAs may undergo reactions with the pre-mRNA substrate and with one another that have substituted for the series of conformational changes that occur in RNAs that splice by group II mechanisms. In effect, these changes have relieved the substrate pre-mRNA of the obligation to carry the sequences needed to sponsor the reaction. As the splicing apparatus has become more complex (and as the number of potential substrates has increased), proteins have played a more important role.

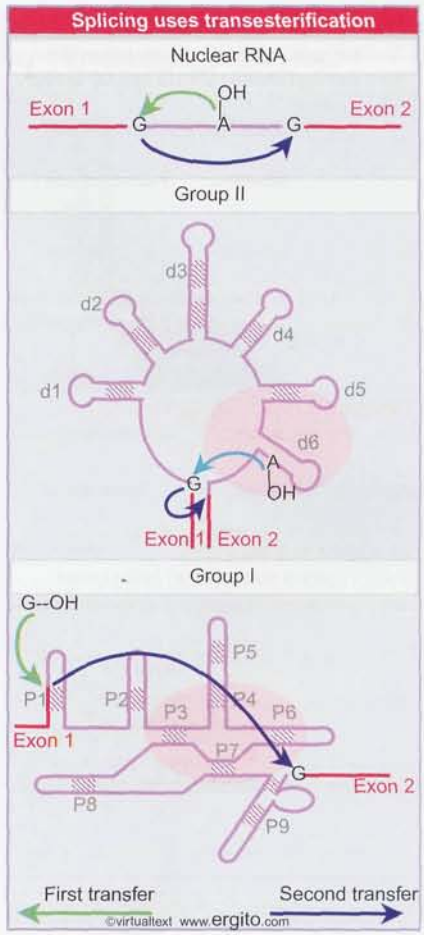


Figure 24.18 Three classes of splicing reactions proceed by two transesterifications. First, a free OH group attacks the exon 1–intron junction. Second, the OH created at the end of exon 1 attacks the intron–exon 2 junction.

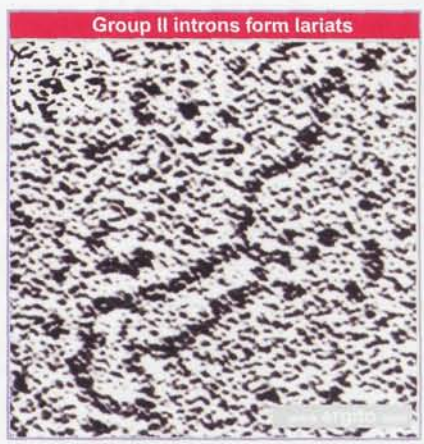


Figure 24.19 Splicing releases a mitochondrial group II intron in the form of a stable lariat. Photograph kindly provided by Leslie Grivell and Annika Arnberg.

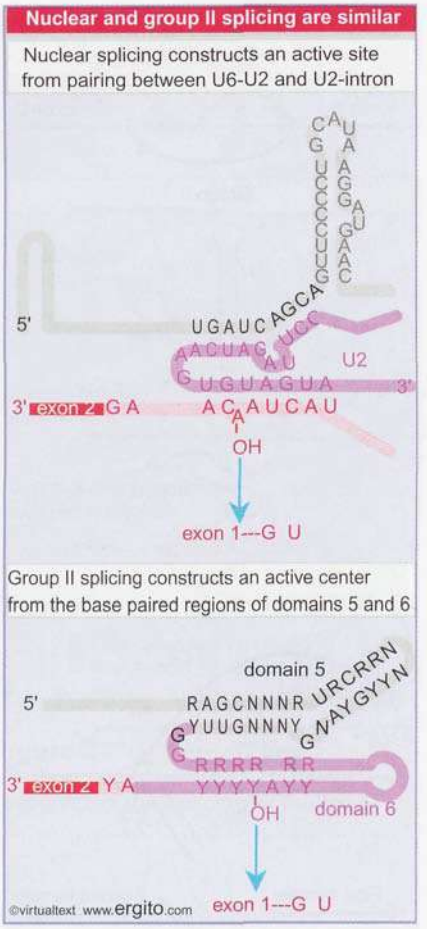


Figure 24.20 Nuclear splicing and group II splicing involve the formation of similar secondary structures. The sequences are more specific in nuclear splicing; group II splicing uses positions that may be occupied by either purine (R) or pyrimidine (Y).

24.12 Alternative splicing involves differential use of splice junctions

Key Concepts

- Specific exons may be excluded or included in the RNA product by using or failing to use a pair of splicing junctions.
- Exons may be extended by changing one of the splice junctions to use an alternative junction.
- Sex determination in *Drosophila* involves a series of alternative splicing events in genes coding for successive products of a pathway.
- P elements of *Drosophila* show germline-specific alternative splicing.

When an interrupted gene is transcribed into an RNA that gives rise to a single type of spliced mRNA, there is no ambiguity in assignment of exons and introns. But the RNAs of some genes follow patterns of alternative splicing, when a single gene gives rise to more than one mRNA sequence. In some cases, the ultimate pattern of expression is dictated by the primary transcript, because the use of different startpoints or the generation of alternative 3' ends alters the pattern of splicing. In other cases, a single primary transcript is spliced in more than one way, and internal exons are substituted, added, or deleted. In some cases, the multiple products all are made in the same cell, but in others the process is regulated so that particular splicing patterns occur only under particular conditions.

One of the most pressing questions in splicing is to determine what controls the use of such alternative pathways. Proteins that intervene to bias the use of alternative splice sites have been identified in two ways. In some mammalian systems, it has been possible to characterize alternative splicing *in vitro* and to identify proteins that are required for the process. In *D. melanogaster*, aberrations in alternative splicing may be caused either by mutations in the genes that are alternatively spliced or in the genes whose products are necessary for the reaction.

Figure 24.21 shows examples of alternative splicing in which one splice site remains constant, but the other varies. The large T/small t antigens of SV40 and the products of the adenovirus E1A region are generated by connecting a varying 5' site to a constant 3' site. In the case of the T/t antigens, the 5' site used for T antigen removes a termination codon that is present in the t antigen mRNA, so that T antigen is larger than t antigen. In the case of the E1A transcripts, one of the 5' sites connects to the last exon in a different reading frame, again making a significant change in the C-terminal part of the protein. In these examples, all the relevant splicing events take place in every cell in which the gene is expressed, so all the protein products are made.

There are differences in the ratios of T/t antigens in different cell types. A protein extracted from cells that produce relatively more small t antigen can cause preferential production of small t RNA in extracts from other cell types. This protein, which was called ASF (alternative splicing factor), turns out to be the same as the splicing factor SF2, which is required for early steps in spliceosome assembly and for the first cleavage-ligation reaction (see Figure 24.13). ASF/SF2 is an RNA-binding protein in the SR family. When a pre-mRNA has more than one 5' splice site preceding a single 3' splice site, increased concentrations of ASF/SF2 promote use of the 5' site nearest to the 3' site at the expense of the other site. This effect of ASF/SF2 can be counteracted by another splicing factor, SF5.

The exact molecular roles of the factors in controlling splice utilization are not yet known, but we see in general terms that alternative splicing involving different 5' sites may be influenced by proteins

involved in spliceosome assembly. In the case of T/t antigens, the effect probably rests on increased binding of the SR proteins to the site that is preferentially used. Alternative splicing also may be influenced by repression of one site. Exons 2 and 3 of the mouse troponin T gene are mutually exclusive; exon 2 is used in smooth muscle, but exon 3 is used in other tissues. Smooth muscle contains proteins that bind to repeated elements located on either side of exon 3, and which prevent use of the 3' and 5' sites that are needed to include it.

The pathway of sex determination in *D. melanogaster* involves interactions between a series of genes in which alternative splicing events distinguish male and female. The pathway takes the form illustrated in **Figure 24.22,** in which the ratio of X chromosomes to autosomes determines the expression of *sxl*, and changes in expression are passed sequentially hrough the other genes to *dsx*, the last in the pathway.

The pathway starts with sex-specific splicing of *sxl*. Exon 3 of the *sxl* gene contains a termination codon that prevents synthesis of functional protein. This exon is included in the mRNA produced in males, but is skipped in females. (Exon skipping illustrated for another example in Figure 24.23.) As a result, only females produce Sxl protein. The protein has a concentration of basic amino acids that resembles other RNA-binding proteins.

The presence of Sxl protein changes the splicing of the *transformer (tra)* gene. Figure 24.21 shows that this involves splicing a constant 5' site to alternative 3' sites. One splicing pattern occurs in both males and females and results in an RNA that has an early termination codon. The presence of Sxl protein inhibits usage of the normal 3' splice site by binding to the polypyrimidine tract at its branch site. When this site is skipped, the next 3' site is used. This generates a female-specific mRNA that codes for a protein.

So *tra* produces a protein only in females; this protein is a splicing regulator. *tra2* has a similar function in females (but is also expressed in the male germline). The Tra and Tra2 proteins are SR splicing factors that act directly upon the target transcripts. Tra and Tra2 cooperate (in females) to affect the splicing of *dsx*.

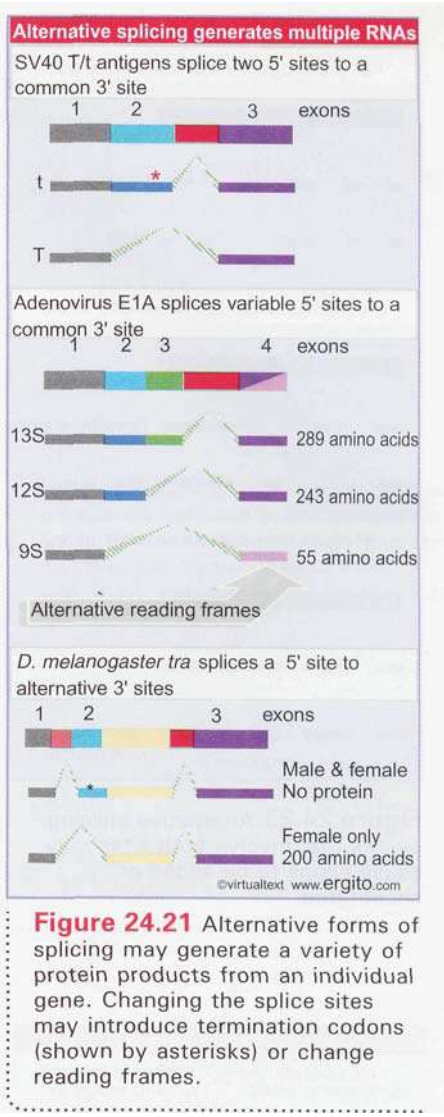


Figure 24.21 Alternative forms of splicing may generate a variety of protein products from an individual gene. Changing the splice sites may introduce termination codons (shown by asterisks) or change reading frames.

Figure 24.23 shows examples of cases in which splice sites are used to add or to substitute exons or introns, again with the consequence that different protein products are generated. In the *doublesex (dsx)* gene, females splice the 5′ site of intron 3 to the 3′ site of that intron; as a result translation terminates at the end of exon 4. Males splice the 5' site of intron 3 directly to the 3' site of intron 4, thus omitting exon 4 from the mRNA, and allowing translation to continue through exon 6. The result of the alternative splicing is that different proteins are produced in each sex: the male product blocks female sexual differentiation, while the female product represses expression of male-specific genes.

Alternative splicing of *dsx* RNA is controlled by competition between 3′ splice sites. *dsx* RNA has an element downstream of the leftmost 3′ splice site that is bound by Tra2; Tra and SR proteins associate with Tra2 at the site, which becomes an enhancer that assists binding of U2AF at the adjacent pyrimidine tract. This commits the formation of the spliceosome to use this 3' site in females rather than the alternative 3' site. The proteins recognize the enhancer cooperatively, possibly relying on formation of some secondary structure as well as sequence *per se*.

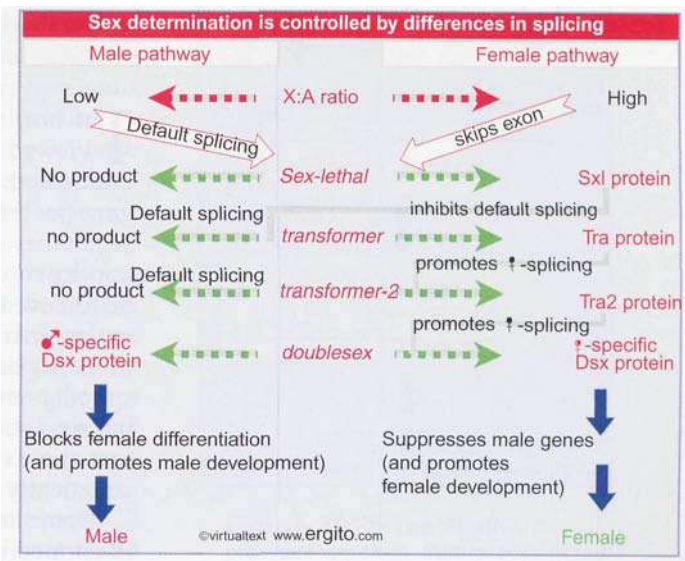


Figure 24.22 Sex determination in *D. melanogaster* involves a pathway in which different splicing events occur in females. Blocks at any stage of the pathway result in male development.

Sex determination therefore has a pleasing symmetry: the pathway starts with a female-specific splicing event that causes omission of an exon that has a termination codon, and ends with a female-specific splicing event that causes inclusion of an exon that has a termination codon. The events have different molecular bases. At the first control

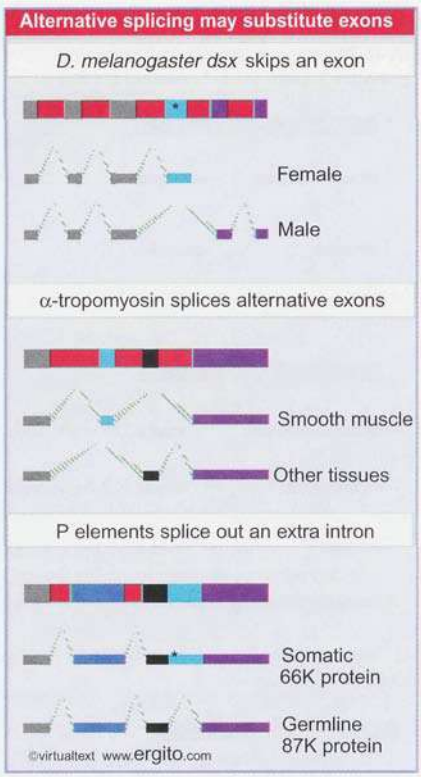


Figure 24.23 Alternative splicing events that involve both sites may cause exons to be added or substituted.

point, Sxl inhibits the default splicing pattern. At the last control point, Tra and Tra2 cooperate to promote the female-specific splice.

The Tra and Tra2 proteins are not needed for normal splicing, because in their absence flies develop normally (as males). As specific regulators, they need not necessarily participate in the mechanics of the splicing reaction; in this respect they differ from SF2, which is a factor required for general splicing, but can also influence choice of alternative splice sites.

P elements of *D. melanogaster* show a tissue-specific splicing pattern. In somatic cells, there are two splicing events, but in germline an additional splicing event removes another intron. Because a termination codon lies in the germline-specific intron, a longer protein (with different properties) is produced in germline. We discuss the consequences for control of transposition in *16.15 P elements are activated in the germline*, and note for now that the tissue specificity results from differences in the splicing apparatus.

The default splicing pathway of the P element pre-mRNA when the RNA is subjected to a heterologous (human) splicing extract is the germline pattern, in which intron 3 is excised. But extracts of somatic cells of *D. melanogaster* contain a protein that inhibits excision of this intron. The protein binds to sequences in exon 3; if these sequences are deleted, the intron is excised. The function of the protein is therefore probably to repress association of the spliceosome with the 5' site of intron 3.

24.13 trans-splicing reactions use small RNAs

Key Concepts

- Splicing reactions usually occur only in *cis* between splice junctions on the same molecule of RNA.
- *trans*-splicing occurs in trypanosomes and worms where a short sequence (SL RNA) is spliced to the 5' ends of many precursor mRNAs.
- SL RNA has a structure resembling the Sm-binding site of U snRNAs and may play an analogous role in the reaction.

In both mechanistic and evolutionary terms, splicing has been viewed as an *intramolecular* reaction, amounting essentially to a controlled deletion of the intron sequences at the level of RNA. In genetic terms, splicing occurs only in *cis*. This means that *only sequences on the same molecule of RNA can be spliced together*. The upper part of **Figure 24.24** shows the normal situation. The introns can be removed from each RNA molecule, allowing the exons of that RNA molecule to be spliced together, but there is no *intermolecular* splicing of exons between different RNA molecules. We cannot say that *trans* splicing never occurs between pre-mRNA transcripts of the same gene, but we know that it must be exceedingly rare, because if it were prevalent the exons of a gene would be able to complement one another genetically instead of belonging to a single complementation group.

Some manipulations can generate *trans*-splicing. In the example illustrated in the lower part of Figure 24.24, complementary sequences were introduced into the introns of two RNAs. Base pairing between the complements should create an H-shaped molecule. This molecule could be spliced in *cis*, to connect exons that are covalently connected by an intron, or it could be spliced in *trans*, to connect exons of the juxtaposed RNA molecules. Both reactions occur *in vitro*.

By Book_Crazy [IND]

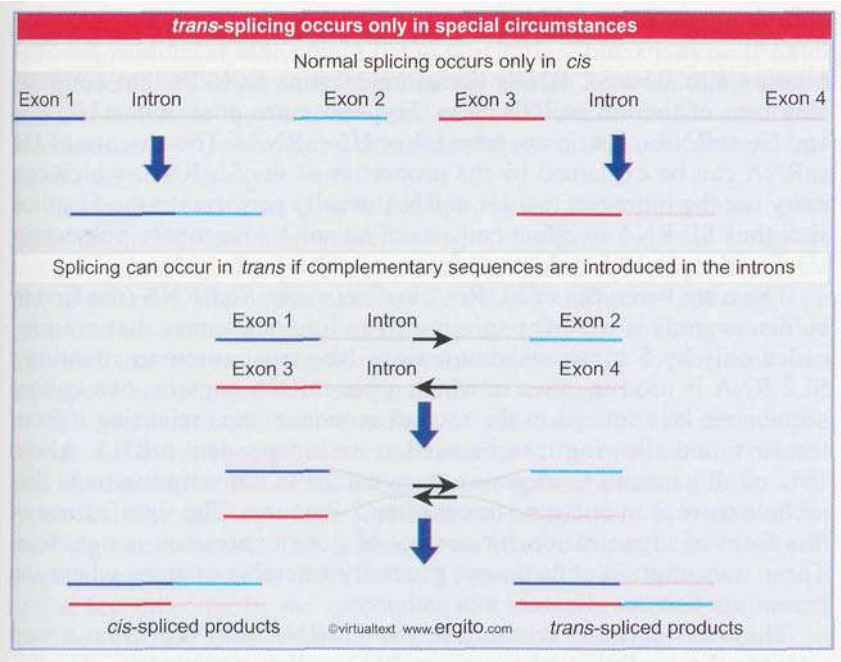


Figure 24.24 Splicing usually occurs only in *cis* between exons carried on the same physical RNA molecule, but trans splicing can occur when special constructs are made that support base pairing between introns.

Another situation in which *trans-splicing* is possible *in vitro* occurs when substrate RNAs are provided in the form of one containing a 5' splice site and the other containing a 3' splice site together with appropriate downstream sequences (which may be either the next 5′ splice site or a splicing enhancer). In effect, this mimics splicing by exon definition (see the right side of Figure 24.12), and shows that *in vitro* it is not necessary for the left and right splice sites to be on the same RNA molecule.

These results show that there is no *mechanistic* impediment to *trans*-splicing. They exclude models for splicing that require processive movement of a spliceosome along the RNA. It must be possible for a spliceosome to recognize the 5' and 3' splice sites of different RNAs when they are in close proximity.

Although *trans-splicing* is rare, it occurs *in vivo* in some special situations. One is revealed by the presence of a common 35 base leader sequence at the end of numerous mRNAs in the trypanosome. But the leader sequence is not coded upstream of the individual transcription units. Instead it is transcribed into an independent RNA, carrying additional sequences at its 3' end, from a repetitive unit located elsewhere in the genome. **Figure 24.25** shows that this RNA carries the 35 base leader sequence followed by a 5' splice site sequence. The sequences coding for the mRNAs carry a 3' splice site just preceding the sequence found in the mature mRNA.

When the leader and the mRNA are connected by a *trans-splicing* reaction, the 3' region of the leader RNA and the 5' region of the mRNA in effect comprise the 5' and 3' halves of an intron. When splicing occurs, a 5′–2′ link forms by the usual reaction between the GU of the 5' intron and the branch sequence near the AG of the 3' intron. Because the two parts of the intron are not covalently linked, this generates a Y-shaped molecule instead of a lariat.

A similar situation is presented by the expression of actin genes in *C. elegans*. Three actin mRNAs (and some other RNAs) share the same 22 base leader sequence at the 5' terminus. The leader sequence is not coded in the actin gene, but is transcribed independently as part of a 100 base RNA coded by a gene elsewhere. *trans*-splicing also occurs in chloroplasts.

The RNA that donates the 5' exon for *trans* splicing is called the SL RNA (spliced leader RNA). The SL RNAs found in several species of trypanosomes and also in the nematode *(C. elegans)* have some com-

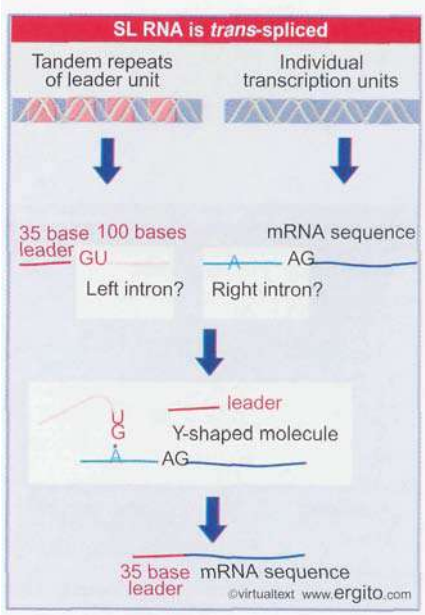


Figure 24.25 The SL RNA provides an exon that is connected to the first exon of an mRNA by *trans*-splicing. The reaction involves the same interactions as nuclear cis-splicing, but generates a Y-shaped RNA instead of a lariat.

mon features. They fold into a common secondary structure that has three stem loops and a single-stranded region that resembles the Sm-binding site. The SL RNAs therefore exist as snRNPs that count as members of the Sm snRNP class. Trypanosomes possess the U2, U4, and U6 snRNAs, but do not have U1 or U5 snRNAs. The absence of U1 snRNA can be explained by the properties of the SL RNA, which can carry out the functions that Ul snRNA usually performs at the 5' splice site; thus SL RNA in effect consists of an snRNA sequence possessing U1 function, linked to the exon-intron site that it recognizes.

There are two types of SL RNA in *C. elegans*. SL1 RNA (the first to be discovered) is used for splicing to coding sequences that are preceded only by 5' nontranslated regions (the most common situation). SL2 RNA is used in cases in which a pre-mRNA contains two coding sequences; it is spliced to the second sequence, thus releasing it from the first, and allowing it to be used as an independent mRNA. About 15% of all genes in *C. elegans* are organized in transcription units that include more than one gene (most often 2-3 genes). The significance of this form of organization for control of gene expression is not clear. These transcription units do not generally resemble operons where the genes function coordinately in a pathway.

The maws-splicing reaction of the SL RNA may represent a step towards the evolution of the pre-mRNA splicing apparatus. The SL RNA provides in *cis* the ability to recognize the 5' splice site, and this probably depends upon the specific conformation of the RNA. The remaining functions required for splicing are provided by independent snRNPs. The SL RNA can function without participation of proteins like those in Ul snRNP, which suggests that the recognition of the 5' splice site depends directly on RNA.

24.14 Yeast tRNA splicing involves cutting and rejoining

Key Concepts

- tRNA splicing occurs by successive cleavage and ligation reactions.

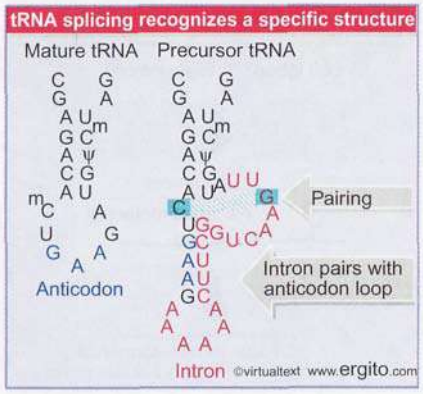


Figure 24.26 The intron in yeast $tRNA^{Phe}$ base pairs with the anticodon to change the structure of the anticodon arm. Pairing between an excluded base in the stem and the intron loop in the precursor may be required for splicing.

Most splicing reactions depend on short consensus sequences and occur by transesterification reactions in which breaking and making of bonds is coordinated. The splicing of tRNA genes is achieved by a different mechanism that relies upon separate cleavage and ligation reactions.

Some 59 of the 272 nuclear tRNA genes in the yeast *S. cerevisiae* are interrupted. Each has a single intron located just one nucleotide beyond the 3'side of the anticodon. The introns vary in length from 14 -60 bp. Those in related tRNA genes are related in sequence, but the introns in tRNA genes representing different amino acids are unrelated. *There is no consensus sequence that could be recognized by the splicing enzymes.* This is also true of interrupted nuclear tRNA genes of plants, amphibians, and mammals.

All the introns include a sequence that is complementary to the anticodon of the tRNA. This creates an alternative conformation for the anticodon arm in which the anticodon is base paired to form an extension of the usual arm. An example is drawn in **Figure 24.26.** Only the anticodon arm is affected—the rest of the molecule retains its usual structure.

The exact sequence and size of the intron is not important. Most mutations in the intron do not prevent splicing. *Splicing of tRNA depends*

principally on recognition of a common secondary structure in tRNA rather than a common sequence of the intron. Regions in various parts of the molecule are important, including the stretch between the acceptor arm and D arm, in the TψC arm, and especially the anticodon arm. This is reminiscent of the structural demands placed on tRNA for protein synthesis (see *6 Protein synthesis).*

The intron is not entirely irrelevant, however. Pairing between a base in the intron loop and an unpaired base in the stem is required for splicing. Mutations at other positions that influence this pairing (for example, to generate alternative patterns for pairing) influence splicing. The rules that govern availability of tRNA precursors for splicing resemble the rules that govern recognition by aminoacyl-tRNA synthetases (see *7.8 tRNAs are charged with amino acids by synthetases).*

In a temperature-sensitive mutant of yeast that fails to remove the introns, the interrupted precursors accumulate in the nucleus. The precursors can be used as substrates for a cell-free system extracted from wild-type cells. The splicing of the precursor can be followed by virtue of the resulting size reduction. This is seen by the change in position of the band on gel electrophoresis, as illustrated in **Figure 24.27.** The reduction in size can be accounted for by the appearance of a band representing the intron.

The cell-free extract can be fractionated by assaying the ability to splice the tRNA. The *in vitro* reaction requires ATR Characterizing the reactions that occur with and without ATP shows that the *two separate stages of the reaction are catalyzed by different enzymes.*

- The first step does not require ATP. It involves phosphodiester bond cleavage by an atypical nuclease reaction. It is catalyzed by an endonuclease.
- The second step requires ATP and involves bond formation; it is a ligation reaction, and the responsible enzyme activity is described as an RNA ligase.

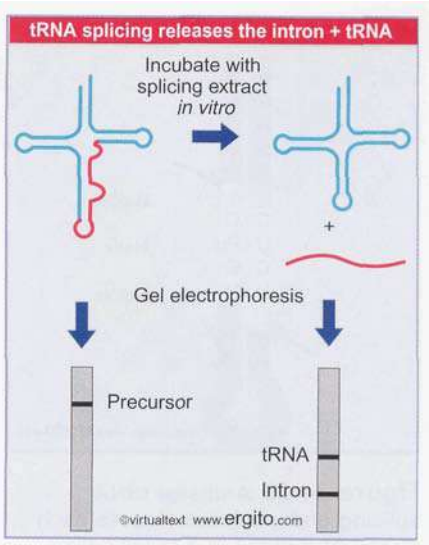


Figure 24.27 Splicing of yeast tRNA *in vitro* can be followed by assaying the RNA precursor and products by gel electrophoresis.

24.15 The splicing endonuclease recognizes tRNA

Key Concepts

- An endonuclease cleaves the tRNA precursors at both ends of the intron.
- The yeast endonuclease is a heterotetramer with two (related) catalytic subunits.
- It uses a measuring mechanism to determine the sites of cleavage by their positions relative to a point in the tRNA structure.
- The archaeal nuclease has a simpler structure and recognizes a bulge-helix-bulge structural motif in the substrate.

The endonuclease is responsible for the specificity of intron recognition. It cleaves the precursor at both ends of the intron. The yeast endonuclease is a heterotetrameric protein. Its activities are illustrated in **Figure 24.28.** The related subunits Sen34 and Sen2 cleave the 3' and 5' splice sites, respectively. Subunit Sen54 may determine the sites of cleavage by "measuring" distance from a point in the tRNA structure. This point is in the elbow of the (mature) L-shaped structure. The role of subunit Sen15 is not known, but its gene is essential in yeast. The base pair that forms between the first base in the anticodon loop and the base preceding the 3' splice site is required for 3' splice site cleavage.

An interesting insight into the evolution of tRNA splicing is provided by the endonucleases of archaea. These are homodimers or homotetramers, in which each subunit has an active site (although only two of

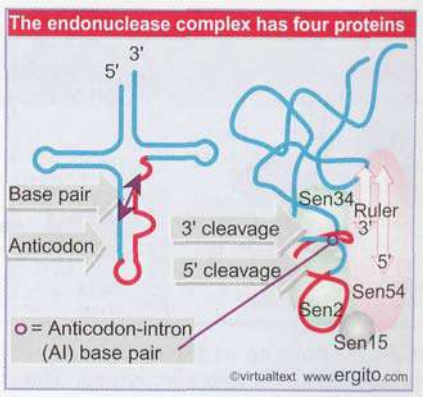


Figure 24.28 The 3′ and 5′ cleavages in *S. cerevisiae* pre-tRNA are catalyzed by different subunits of the endonuclease. Another subunit may determine location of the cleavage sites by measuring distance from the mature structure. The AI base pair is also important.

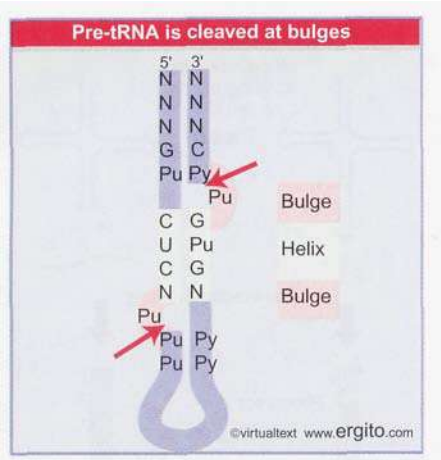


Figure 24.29 Archaeal tRNA splicing endonuclease cleaves each strand at a bulge in a bulge-helix-bulge motif.

the sites function in the tetramer) that cleaves one of the splice sites. The subunit has sequences related to the sequences of the active sites in the Sen34 and Sen2 subunits of the yeast enzyme. However, the archaeal enzymes recognize their substrates in a different way. Instead of measuring distance from particular sequences, they recognize a structural feature called the bulge-helix-bulge. **Figure 24.29** shows that cleavage occurs in the two bulges.

So the origin of splicing of tRNA precedes the separation of the archaea and the eukaryotes. If it originated by insertion of the intron into tRNAs, this must have been a very ancient event.

24.16 tRNA cleavage and ligation are separate reactions

Key Concepts

- Release of the intron generates two half-tRNAs that pair to form the mature structure.
- The halves have the unusual ends 5' hydroxyl and 2'-3'cyclic phosphate.
- The 5′-OH end is phosphorylated by a polynucleotide kinase, the cyclic phosphate group is opened by phosphodiesterase to generate a 2'-phosphate terminus and 3′-OH group, exon ends are joined by an RNA ligase, and the 2'-phosphate is removed by a phosphatase.

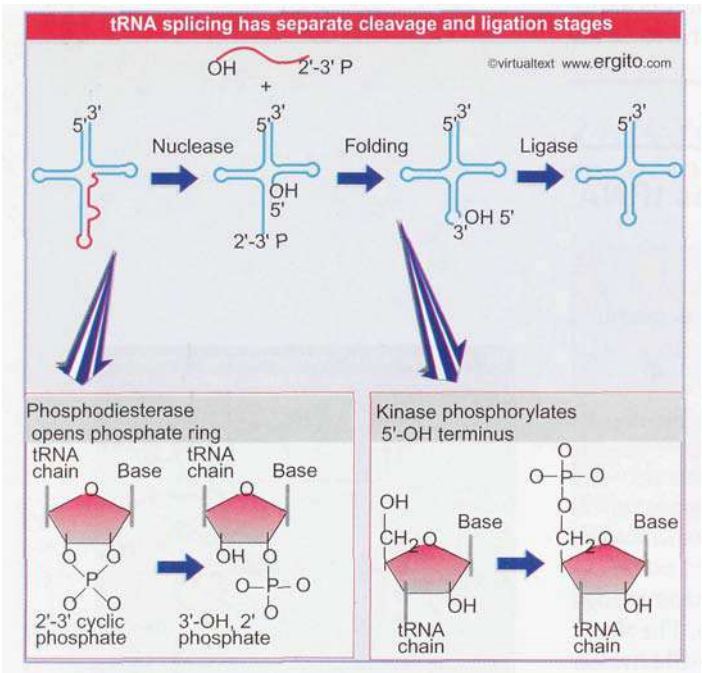


Figure 24.30 Splicing of tRNA requires separate nuclease and ligase activities. The exon-intron boundaries are cleaved by the nuclease to generate 2'-3' cyclic phosphate and 5′-OH termini. The cyclic phosphate is opened to generate 3′-OH and 2' phosphate groups. The 5′-OH is phosphorylated. After releasing the intron, the tRNA half molecules fold into a tRNA-like structure that now has a 3′-OH, 5'-P break. This is sealed by a ligase.

The overall tRNA splicing reaction is summarized in **Figure 24.30.** The products of cleavage are a linear intron and two half-tRNA molecules. These intermediates have unique ends. Each 5' terminus ends in a hydroxyl group; each 3' terminus ends in a 2',3'-cyclic phosphate group. (All other known RNA splicing enzymes cleave on the other side of the phosphate bond.)

The two half-tRNAs base pair to form a tRNA-like structure. When ATP is added, the second reaction occurs. Both of the unusual ends generated by the endonuclease must be altered.

The cyclic phosphate group is opened to generate a 2′–phosphate terminus. This reaction requires cyclic phosphodiesterase activity. The product has a 2′–phosphate group and a 3′–OH group.

The 5′–OH group generated by the nuclease must be phosphorylated to give a 5′–phosphate. This generates a site in which the 3′–OH is next to the 5′–phosphate. Covalent integrity of the polynucleotide chain is then restored by ligase activity.

All three activities—phosphodiesterase, polynucleotide kinase, and adenylate synthetase (which provides the ligase function)—are arranged in different functional domains on a single protein. They act sequentially to join the two tRNA halves.

The spliced molecule is now uninterrupted, with a 5′–3′phosphate linkage at the site of splicing, but it also has a 2′–phosphate group marking the event. The surplus group must be removed by a phosphatase.

Generation of a 2',3'-cyclic phosphate also occurs during the tRNA-splicing reaction in plants and mammals. The reaction in plants seems to be the same as in yeast, but the detailed chemical reactions are different in mammals.

By Book_Crazy [IND]

The yeast tRNA precursors also can be spliced in an extract obtained from the germinal vesicle (nucleus) of *Xenopus* oocytes. This shows that the reaction is not species-specific. *Xenopus* must have enzymes able to recognize the introns in the yeast tRNAs.

The ability to splice the products of tRNA genes is therefore well conserved, but is likely to have a different origin from the other splicing reactions (such as that of nuclear pre-mRNA). The tRNA-splicing reaction uses cleavage and synthesis of bonds and is determined by sequences that are external to the intron. Other splicing reactions use transesterification, in which bonds are transferred directly, and the sequences required for the reaction lie within the intron.

24.17 The unfolded protein response is related to tRNA splicing

Key Concepts

- Ire1p is an inner nuclear membrane protein with its N-terminal domain in the ER lumen and its C-terminal domain in the nucleus.
- Binding of an unfolded protein to the N-terminal domain activates the C-terminal nuclease by autophosphorylation.
- The activated nuclease cleaves Hac1 mRNA to release an intron and generate exons that are ligated by a tRNA ligase.
- The spliced Hac1 mRNA codes for a transcription factor that activates genes coding for chaperones that help to fold unfolded proteins.

An unusual splicing system that is related to tRNA splicing mediates the response to unfolded proteins in yeast. The accumulation of unfolded proteins in the lumen of the ER triggers a response pathway that leads to increased transcription of genes coding for chaperones that assist protein folding in the ER. A signal must therefore be transmitted from the lumen of the ER to the nucleus.

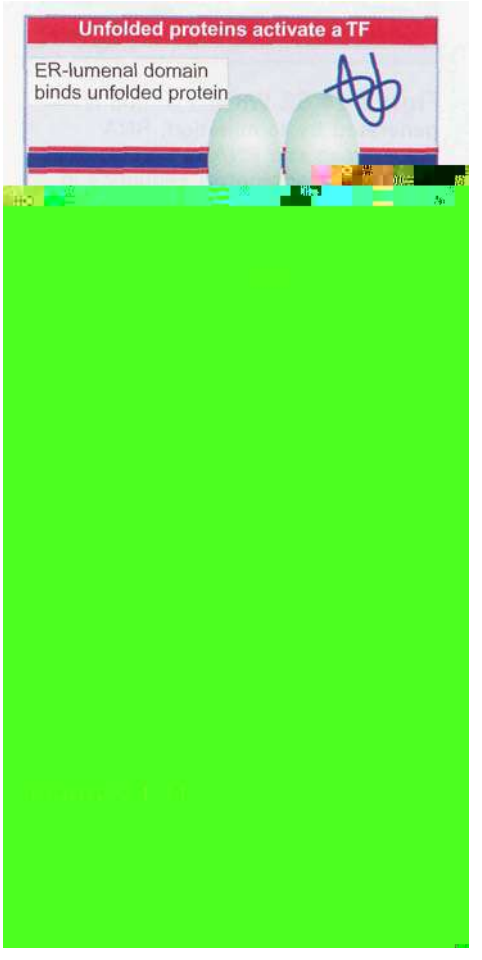
The sensor that activates the pathway is the protein Ire1p. It is an integral membrane protein (Ser/Thr) kinase that has domains on each side of the ER membrane. The N-terminal domain in the lumen of the ER detects the presence of unfolded proteins, presumably by binding to exposed motifs. This causes aggregation of monomers and activates the C-terminal domain on the other side of the membrane by autophosphorylation.

Genes that are activated by this pathway have a common promoter element, the UPRE (unfolded protein response element). The transcription factor Hac1p binds to the UPRE and is produced in response to accumulation of unfolded proteins. The trigger for production of Hac1p is the action of Ire1p on Hac1 mRNA.

The operation of the pathway is summarized in **Figure 24.31.** Under normal conditions, when the pathway is not activated, Hacl mRNA is translated into a protein that is rapidly degraded. The activation of Ire1p results in the splicing of the Hac1 mRNA to change the sequence of the protein to a more stable form. This form provides the functional transcription factor that activates genes with the UPRE.

Unusual splicing components are involved in this reaction. Ire1P has an endonuclease activity that acts directly on Hac 1 mRNA to cleave the two splicing junctions. The two junctions are ligated by the tRNA ligase that acts in the tRNA splicing pathway. The endonuclease reaction resembles the cleavage of tRNA during splicing.

Where does the modification of Hac1 mRNA occur? Ire1p is probably located in the inner nuclear membrane, with the N-terminal sensor



By Book_Crazy [IND]

domain in the ER lumen, and the C-terminal kinase/nuclease domain in the nucleus. This would enable it to act directly on Hac1 RNA before it is exported to the cytoplasm. It also would allow easy access by the tRNA ligase. There is no apparent relationship between the Ire1p nuclease activity and the tRNA splicing endonuclease, so it is not obvious how this specialized system would have evolved.

24.18 The 3' ends of polI and polIII transcripts are generated by termination

Key Concepts

- RNA polymerase I terminates transcription at an 18 base terminator sequence.
- RNA polymerase III terminates transcription in poly(U)$_4$ sequence embedded in a G·C-rich sequence.

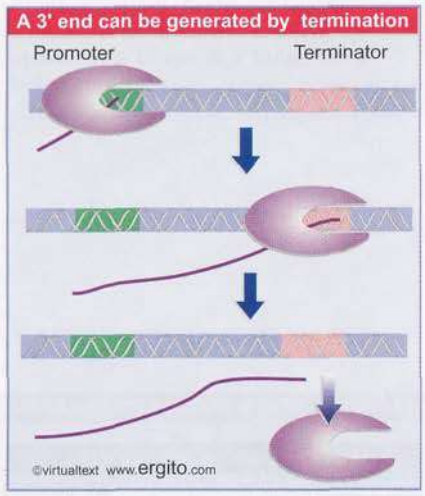


Figure 24.32 When a 3′ end is generated by termination, RNA polymerase and RNA are released at a discrete (terminator) sequence in DNA.

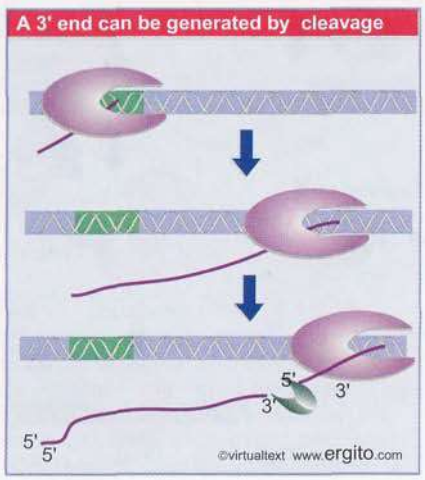


Figure 24.33 When a 3′ end is generated by cleavage, RNA polymerase continues transcription while an endonuclease cleaves at a defined sequence in the RNA.

3' ends of RNAs can be generated in two ways. Some RNA polymerases terminate transcription at a defined (terminator) sequence in DNA, as shown in **Figure 24.32.** Other RNA polymerases do not show discrete termination, but continue past the site corresponding to the 3′ end, which is generated by cleavage of the RNA by an endonuclease, as shown in Figure 24.33.

Information about the termination reaction for eukaryotic RNA polymerases is less detailed than our knowledge of initiation. RNA polymerases I and III have discrete termination events (like bacterial RNA polymerase), but it is not clear whether RNA polymerase II usually terminates in this way.

For RNA polymerase I, the sole product of transcription is a large precursor that contains the sequences of the major rRNA. The precursor is subjected to extensive processing. Termination occurs at a discrete site >1000 bp downstream of the mature 3' end, which is generated by cleavage. Termination involves recognition of an 18 base terminator sequence by an ancillary factor.

With RNA polymerase III, transcription *in vitro* generates molecules with the same 5' and 3' ends as those synthesized *in vivo.* The termination reaction resembles intrinsic termination by bacterial RNA polymerase (see *9.21 There are two types of terminators in E. coli*). Termination usually occurs at the second U within a run of 4 U bases, but there is heterogeneity, with some molecules ending in 3 or even 4 U bases. The same heterogeneity is seen in molecules synthesized *in vivo*, so it seems to be a *bona fide* feature of the termination reaction.

Just like the prokaryotic terminators, the U run is embedded in a G-C-rich region. Although sequences of dyad symmetry are present, they are not needed for termination, since mutations that abolish the symmetry do not prevent the normal completion of RNA synthesis. Nor are any sequences beyond the U run necessary, since all distal sequences can be replaced without any effect on termination.

The U run itself is not sufficient for termination, because regions of 4 successive U residues exist within transcription units read by RNA polymerase III. (However, there are no internal U_5 runs, which fits with the greater efficiency of termination when the terminator is a U_5 rather than U_4 sequence.) The critical feature in termination must therefore be the recognition of a U_4 sequence in a context that is rich in G·C base pairs.

How does the termination reaction occur? It cannot rely on the weakness of the rU-dA RNA-DNA hybrid region that lies at the end of the

transcript, because often only the first two U residues are transcribed. Perhaps the G·C-rich region plays a role in slowing down the enzyme, but there does not seem to be a counterpart to the hairpin involved in prokaryotic termination. We remain puzzled how the enzyme can respond so specifically to such a short signal. And in contrast with the initiation reaction, which RNA polymerase III cannot accomplish alone, termination seems to be a function of the enzyme itself.

24.19 The 3' ends of mRNAs are generated by cleavage and polyadenylation

Key Concepts

- The sequence AAUAAA is a signal for cleavage to generate a 3' end of mRNA that is polyadenylated.
- The reaction requires a protein complex that contains a specificity factor, an endonuclease, and poly(A) polymerase.
- The specificity factor and endonuclease cleave RNA downstream of AAUAAA.
- The specificity factor and poly(A) polymerase add ~200 A residues processively to the 3' end.
- AU-rich sequences in the 3' tail control cytoplasmic polyadenylation or deadenylation during *Xenopus* embryonic development.

It is not clear whether RNA polymerase II actually engages in a termination event at a specific site. It is possible that its termination is only loosely specified. In some transcription units, termination occurs >1000 bp downstream of the site corresponding to the mature 3' end of the mRNA (which is generated by cleavage at a specific sequence). Instead of using specific terminator sequences, the enzyme ceases RNA synthesis within multiple sites located in rather long "terminator regions." The nature of the individual termination sites is not known.

The 3′ ends of mRNAs are generated by cleavage followed by polyadenylation. Addition of poly(A) to nuclear RNA can be prevented by the analog 3'-deoxyadenosine, also known as cordycepin. Although cordycepin does not stop the transcription of nuclear RNA, its addition prevents the appearance of mRNA in the cytoplasm. This shows that polyadenylation is *necessary* for the maturation of mRNA from nuclear RNA.

Generation of the 3' end is illustrated in **Figure 24.34.** RNA polymerase transcribes past the site corresponding to the 3' end, and sequences in the RNA are recognized as targets for an endonucleolytic cut followed by polyadenylation. A single processing complex undertakes both the cutting and polyadenylation. The polyadenylation stabilizes the mRNA against degradation from the 3' end. Its 5' end is already stabilized by the cap. RNA polymerase continues transcription after the cleavage, but the 5' end that is generated by the cleavage is unprotected. As a result, the rest of the transcript is rapidly degraded. This makes it difficult to determine what is happening beyond the point of cleavage.

A common feature of mRNAs in higher eukaryotes (but not in yeast) is the presence of the highly conserved sequence AAUAAA in the region from 11-30 nucleotides upstream of the site of poly(A) addition. Deletion or mutation of the AAUAAA hexamer prevents generation of the polyadenylated 3′ end. The signal is needed for both cleavage and polyadenylation.

The development of a system in which polyadenylation occurs *in vitro* opened the route to analyzing the reactions. The formation and functions of the complex that undertakes 3' processing are illustrated in

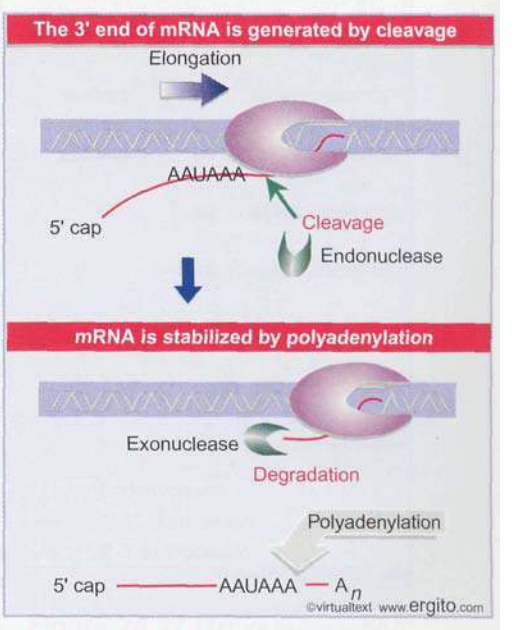


Figure 24.34 The sequence AAUAAA is necessary for cleavage to generate a 3′ end for polyadenylation.

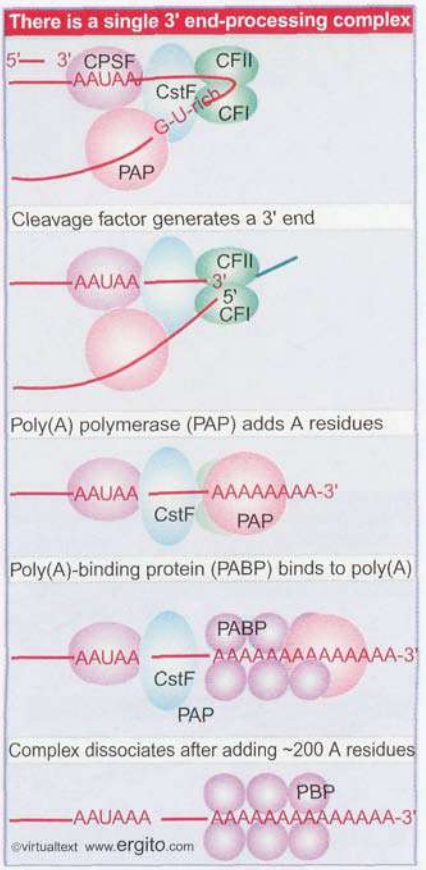


Figure 24.35 The 3′ processing complex consists of several activities. CPSF and CstF each consist of several subunits; the other components are monomeric. The total mass is >900 kD.

Figure 24.35. Generation of the proper 3' terminal structure requires an endonuclease (consisting of the components CFI and CFII) to cleave the RNA, a poly(A) polymerase (PAP) to synthesize the poly(A) tail, and a *specificity component* (CPSF) that recognizes the AAUAAA sequence and directs the other activities. A stimulatory factor, CstF, binds to a G-U-rich sequence that is downstream from the cleavage site itself.

The specificity factor contains 4 subunits, which together bind specifically to RNA containing the sequence AAUAAA. The individual subunits are proteins that have common RNA-binding motifs, but which by themselves bind nonspecifically to RNA. Protein-protein interactions between the subunits may be needed to generate the specific AAUAAA-binding site. CPSF binds strongly to AAUAAA only when CstF is also present to bind to the G-U-rich site.

The specificity factor is needed for both the cleavage and polyadenylation reactions. It exists in a complex with the endonuclease and poly(A) polymerase, and this complex usually undertakes cleavage followed by polyadenylation in a tightly coupled manner.

The two components CFI and CFII (cleavage factors I and II), together with specificity factor, are necessary and sufficient for the endonucleolytic cleavage.

The poly(A) polymerase has a nonspecific catalytic activity. When it is combined with the other components, the synthetic reaction becomes specific for RNA containing the sequence AAUAAA. The polyadenylation reaction passes through two stages. First, a rather short oligo(A) sequence (~10 residues) is added to the 3' end. This reaction is absolutely dependent on the AAUAAA sequence, and poly(A) polymerase performs it under the direction of the specificity factor. In the second phase, the oligo(A) tail is extended to the full ~200 residue length. This reaction requires another stimulatory factor that recognizes the oligo(A) tail and directs poly(A) polymerase specifically to extend the 3' end of a poly(A) sequence.

The poly(A) polymerase by itself adds A residues individually to the 3' position. Its intrinsic mode of action is distributive; it dissociates after each nucleotide has been added. However, in the presence of CPSF and PABP (poly(A)-binding protein), it functions processively to extend an individual poly(A) chain. The PABP is a 33 kD protein that binds stoichiometrically to the poly(A) stretch. The length of poly(A) is controlled by the PABP, which in some way limits the action of poly(A) polymerase to ~200 additions of A residues. The limit may represent the accumulation of a critical mass of PABP on the poly(A) chain. PABP binds to the translation initiation factor eIF4G, thus generating a closed loop in which a protein complex contains both the 5' and 3' ends of the mRNA (see Figure 6.20 in *6.9 Eukaryotes use a complex of many initiation factors).*

Polyadenylation is an important determinant of mRNA function. It may affect both stability and initiation of translation (see *5.10 The 3'* terminus is polyadenylated). In embryonic development in some organisms, the presence of poly(A) is used to control translation, and preexisting mRNAs may either be polyadenylated (to stimulate translation) or deadenylated (to terminate translation). During *Xenopus* embryonic development, polyadenylation of mRNA in the cytoplasm in *Xenopus* depends on a specific *cis-acting* element (the CPE) in the 3' tail. This is another AU-rich sequence, UUUUUAU.

In *Xenopus* embryos at least two types of *cis*-acting sequences found in the 3' tail can trigger deadenylation. EDEN (embryonic deadenylation element) is a 17 nucleotide sequence. ARE elements are AU-rich, usually containing tandem repeats of AUUUA. There is a poly(A)-specific RNAase (PARN) that could be involved in the degradation. Of course, deadenylation is not always triggered by specific elements; in some situations (including the normal degradation of mRNA as it ages), poly(A) is degraded unless it is specifically stabilized.

By Book_Crazy [IND]

24.20 Cleavage of the 3' end of histone mRNA may require a small RNA

Key Concepts

- Histone mRNAs are not polyadenylated; their 3' ends are generated by a cleavage reaction that depends on the structure of the mRNA.
- The cleavage reaction requires the SLBP to bind to a stem-loop structure and the U7 snRNA to pair with an adjacent single-stranded region.

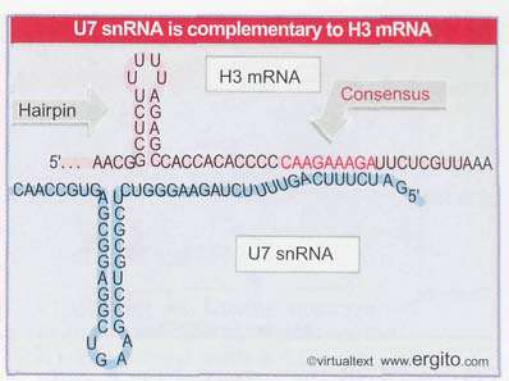


Figure 24.36 Generation of the 3′ end of histone H3 mRNA depends on a conserved hairpin and a sequence that base pairs with U7 snRNA.

Some mRNAs are not polyadenylated. The formation of their 3' ends is therefore different from the coordinated cleavage/polyadenylation reaction. The most prominent members of this mRNA class are the mRNAs coding for histones that are synthesized during DNA replication. Formation of their 3' ends depends upon secondary structure. The structure at the 3' terminus is a highly conserved stem-loop structure, with a stem of 6 bp and a loop of 4 nucleotides. Cleavage occurs 4-5 bases downstream of the stem loop. Two factors are required for the cleavage reaction: the stem-loop binding protein (SLBP) recognizes the structure; and the U7 snRNA pairs with a purine-rich sequence (the histone downstream element, or HDE) located ~10 nucleotides downstream of the cleavage site.

Mutations that prevent formation of the duplex stem of the stem loop prevent formation of the end of the RNA. Secondary mutations that restore duplex structure (though not necessarily the original sequence) behave as revertants. This suggests *that formation of the secondary structure is more important than the exact sequence.* The SLBP binds to the stem loop and then interacts with U7 snRNP to enhance its interaction with the downstream binding site for U7 snRNA. U7 snRNP is a minor snRNP consisting of the 63 nucleotide U7 snRNA and a set of several proteins (including Sm proteins; see *24.5 snRNAs are required for splicing).*

The reaction between histone H3 mRNA and U7 snRNA is drawn in **Figure 24.36.** The upstream hairpin and the HDE that pairs with U7 snRNA are conserved in histone H3 mRNAs of several species. The U7 snRNA has sequences towards its 5′ end that pair with the histone mRNA consensus sequences. 3' processing is inhibited by mutations in the HDE that reduce ability to pair with U7 snRNA. Compensatory mutations in U7 snRNA that restore complementarity also restore 3' processing. This suggests that U7 snRNA functions by base pairing with the histone mRNA. The sequence of the HDE varies among the various histone mRNAs, with the result that binding of snRNA is not by itself necessarily stable, but requires also the interaction with SLBP.

Cleavage to generate a 3' terminus occurs a fixed distance from the site recognized by U7 snRNA, which suggests that the snRNA is involved in defining the cleavage site. However, the factor(s) actually responsible for cleavage have not yet been identified.

24.21 Production of rRNA requires cleavage events

Key Concepts

- The large and small rRNAs are released by cleavage from a common precursor RNA.

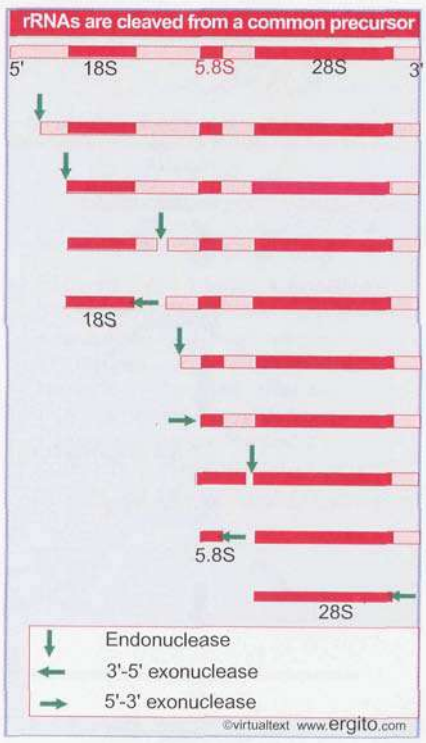


Figure 24.37 Mature eukaryotic rRNAs are generated by cleavage and trimming events from a primary transcript.

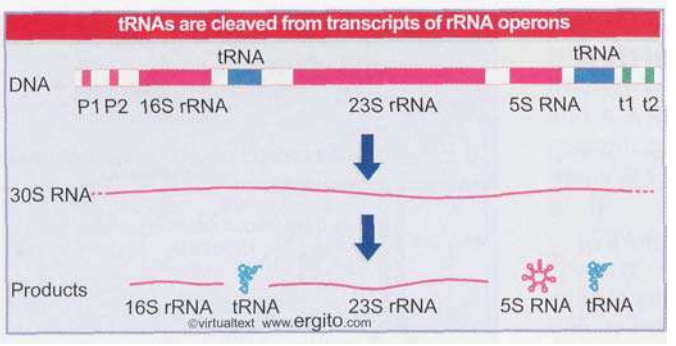


Figure 24.38 The *rrn* operons in *E. coli* contain genes for both rRNA and tRNA. The exact lengths of the transcripts depend on which promoters (P) and terminators (t) are used. Each RNA product must be released from the transcript by cuts on either side.

The major rRNAs are synthesized as part of a single primary transcript that is processed to generate the mature products. The precursor contains the sequences of the 18S, 5.8S, and 28S rRNAs. In higher eukaryotes, the precursor is named for its sedimentation rate as 45S RNA. In lower eukaryotes, it is smaller (35S in yeast).

The mature rRNAs are released from the precursor by a combination of cleavage events and trimming reactions. Figure 24.37 shows the general pathway in yeast. There can be variations in the order of events, but basically similar reactions are involved in all eukaryotes. Most of the 5' ends are generated directly by a cleavage event. Most of the 3' ends are generated by cleavage followed by a 3′-5′ trimming reaction.

Many ribonucleases have been implicated in processing rRNA, including the exosome, an assembly of several exonucleases that also participates in mRNA degradation (see *5.13 mRNA degradation involves multiple activities).* Mutations in individual enzymes usually do not prevent processing, suggesting that their activities are redundant and that different combinations of cleavages can be used to generate the mature molecules.

There are always multiple copies of the transcription unit for the rRNAs. The copies are organized as tandem repeats (see *4.9 The repeated genes for rRNA maintain constant sequence).*

5S RNA is transcribed from separate genes by RNA polymerase III. Usually the 5S genes are clustered, but are separate from the genes for the major rRNAs. (In the case of yeast, a 5S gene is associated with each major transcription unit, but is transcribed independently.)

There is a difference in the organization of the precursor in bacteria. The sequence corresponding to 5.8 S rRNA forms the 5' end of the large (23S) rRNA, that is, there is no processing between these sequences. Figure 24.38 shows that the precursor also contains the 5S rRNA and one or two tRNAs. In *E. coli,* the 7 *rrn* operons are dispersed around the genome; four *rrn* loci contain one tRNA gene between the 16S and 23S rRNA sequences, and the other *rrn* loci contain two tRNA genes in this region. Additional tRNA genes may or may not be present between the 5S sequence and the 3' end. So the processing reactions required to release the products depend on the content of the particular *rrn* locus.

In both prokaryotic and eukaryotic rRNA processing, ribosomal proteins (and possibly also other proteins) bind to the precursor, so that the substrate for processing is not the free RNA but is a ribonucleoprotein complex.

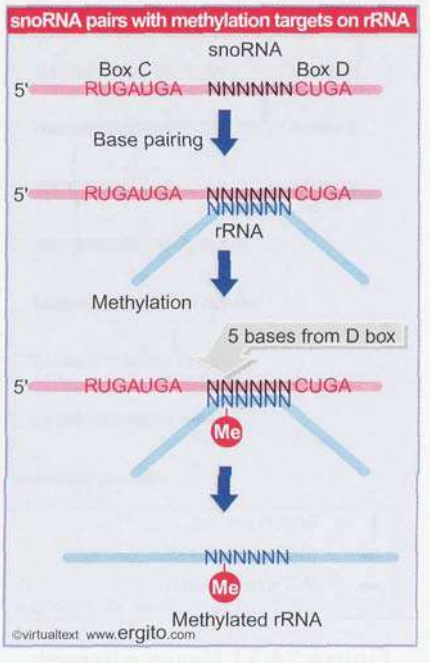


Figure 24.39 A snoRNA base pairs with a region of rRNA that is to be methylated.

24.22 Small RNAs are required for rRNA processing

Key Concepts

- The C/D group of snoRNAs is required for modifying the 2' position of ribose with a methyl group.
- The H/ACA group of snoRNAs is required for converting uridine to pseudouridine.
- In each case the snoRNA base pairs with a sequence of rRNA that contains the target base to generate a typical structure that is the substrate for modification.

Processing and modification of rRNA requires a class of small RNAs called snoRNAs (small nucleolar RNAs). There are 71 snoRNAs in the yeast *(S. cerevisiae)* genome. They are associated with the protein fibrillarin, which is an abundant component of the nucleolus (the region of the nucleus where the rRNA genes are transcribed). Some

snoRNAs are required for cleavage of the precursor to rRNA; one example is U3 snoRNA, which is required for the first cleavage event in both yeast and *Xenopus*. We do not know what role the snoRNA plays in cleavage. It could be required to pair with the rRNA sequence to form a secondary structure that is recognized by an endonuclease.

Two groups of snoRNAs are required for the modifications that are made to bases in the rRNA. The members of each group are identified by very short conserved sequences and common features of secondary structure.

The C/D group of snoRNAs is required for adding a methyl group to the 2' position of ribose. There are >100 2' -O-methyl groups at conserved locations in vertebrate rRNAs. This group takes its name from two short conserved sequences motifs called boxes C and D. Each snoRNA contains a sequence near the D box that is complementary to a region of the 18S or 28S rRNA that is methylated. Loss of a particular snoRNA prevents methylation in the rRNA region to which it is complementary.

Figure 24.39 suggests that the snoRNA base pairs with the rRNA to create the duplex region that is recognized as a substrate for methylation. Methylation occurs within the region of complementarity, at a position that is fixed 5 bases on the 5' side of the D box. Probably each methylation event is specified by a different snoRNA; ~40 snoRNAs have been characterized so far. The methylase(s) have not been characterized; one possibility is that the snoRNA itself provides part of the methylase activity.

Another group of snoRNAs is involved in the synthesis of pseudouridine. There are 43 ψ residues in yeast rRNAs and ~ 100 in vertebrate rRNAs. The synthesis of pseudouridine involves the reaction shown in **Figure 24.40** in which the N1 bond from uridylic acid to ribose is broken, the base is rotated, and C5 is rejoined to the sugar.

Pseudouridine formation in rRNA requires the H/ACA group of ~20 snoRNAs. They are named for the presence of an ACA triplet 3 nucleotides from the 3' end and a partially conserved sequence (the H box) that lies between two stem-loop hairpin structures. Each of these snoRNAs has a sequence complementary to rRNA within the stem of each hairpin. **Figure 24.41** shows the structure that would be produced by pairing with the rRNA. Within each pairing region, there are two unpaired bases, one of which is a uridine that is converted to pseudouridine.

The H/ACA snoRNAs are associated with a nucleolar protein called Gar1p, which is required for pseudouridine formation, but its function is unknown. The known pseudouridine synthases are proteins that function without an RNA cofactor. Synthases that could be involved in snoRNA-mediated pseudouridine synthesis have not been identified.

The involvement of the U7 snRNA in 3′end generation and the role of snoRNAs in rRNA processing and modification are consistent with the view we develop in *25 Catalytic RNA* that many—perhaps all—RNA processing events depend on RNA-RNA interactions. As with splicing reactions, the snRNA probably functions in the form of a ribonucleoprotein particle containing proteins as well as the RNA. It is common (although not the only mechanism of action) for the RNA of the particle to base pair with a short sequence in the substrate RNA.

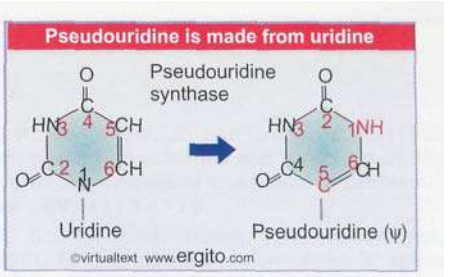


Figure 24.40 Uridine is converted to pseudouridine by replacing the N1-sugar bond with a C5-sugar bond and rotating the base relative to the sugar.

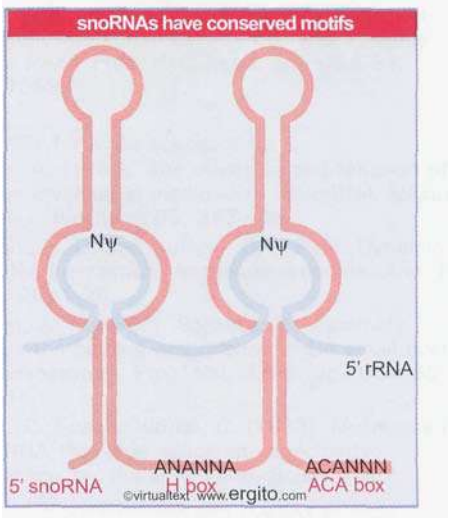


Figure 24.41 H/ACA snoRNAs have two short conserved sequences and two hairpin structures, each of which has regions in the stem that are complementary to rRNA. *Pseudouridine is formed by* converting an unpaired uridine within the complementary region of the rRNA.

24.23 Summary

Splicing accomplishes the removal of introns and the joining of exons into the mature sequence of RNA. There are at least four types of reaction, as distinguished by their requirements *in vitro* and the intermediates that they generate. The systems include eukaryotic nuclear introns, group I and group II introns, and tRNA introns. Each reaction involves a change of organization within an individual RNA molecule and is therefore a *cis*-acting event.

pre-mRNA splicing follows preferred but not obligatory pathways. Only very short consensus sequences are necessary; the rest of the intron appears irrelevant. All 5' splice sites are probably equivalent, as are all 3' splice sites. The required sequences are given by the GU-AG rule, which describes the ends of the intron. The UACUAAC branch site of yeast, or a less well conserved consensus in mammalian introns, is also required. The reaction with the 5' splice site involves formation of a lariat that joins the GU end of the intron via a 5'-2' linkage to the A at position 6 of the branch site. Then the 3′–OH end of the exon attacks the 3' splice site, so that the exons are ligated and the intron is released as a lariat. Both reactions are transesterifications in which bonds are conserved. Several stages of the reaction require hydrolysis of ATP, probably to drive conformational changes in the RNA and/or protein components. Lariat formation is responsible for choice of the 3' splice site. Alternative splicing patterns are caused by protein factors that either stimulate use of a new site or that block use of the default site.

pre-mRNA splicing requires formation of a spliceosome, a large particle that assembles the consensus sequences into a reactive conformation. The spliceosome most often forms by the process of intron definition, involving recognition of the 5' splice site, branch site, and 3' splice site. An alternative pathway involves exon definition, which involves initial recognition of the 5' splice sites of both the substrate intron and the next intron. Its formation passes through a series of stages from the E (commitment) complex that contains U1 snRNP and splicing factors, through the A and B complexes as additional components are added.

The spliceosome contains the U1, U2, U4/U6, and U5 snRNPs and some additional splicing factors. The U1, U2, and U5 snRNPs each contain a single snRNA and several proteins; the U4/U6 snRNP contains 2 snRNAs and several proteins. Some proteins are common to all snRNP particles. The snRNPs recognize consensus sequences. U1 snRNA base pairs with the 5' splice site, U2 snRNA base pairs with the branch sequence, U5 snRNP acts at the 5' splice site. When U4 releases U6, the U6 snRNA base pairs with U2, and this may create the catalytic center for splicing. An alternative set of snRNPs provides analogous functions for splicing the U12-dependent subclass of introns. The snRNA molecules may have catalytic-like roles in splicing and other processing reactions.

In the nucleolus, two groups of snoRNAs are responsible for pairing with rRNAs at sites that are modified; group C/D snoRNAs indicate target sites for methylation, and group ACA snoRNAs identify sites where uridine is converted to pseudouridine.

Splicing is usually intramolecular, but *trans*-(intermolecular) splicing occurs in trypanosomes and nematodes. It involves a reaction between a small SL RNA and the pre-mRNA. The SL RNA resembles U1 snRNA and may combine the role of providing the exon and the functions of U1. In worms there are two types of SL RNA, one used for splicing to the 5' end of an mRNA, the other for splicing to an internal site.

Group II introns share with nuclear introns the use of a lariat as intermediate, but are able to perform the reaction as a self-catalyzed property of the RNA. These introns follow the GT-AG rule, but form a characteristic secondary structure that holds the reacting splice sites in the appropriate apposition.

Yeast tRNA splicing involves separate endonuclease and ligase reactions. The endonuclease recognizes the secondary (or tertiary) structure of the precursor and cleaves both ends of the intron. The two half-tRNAs released by loss of the intron can be ligated in the presence of ATP.

The termination capacity of RNA polymerase II has not been characterized, and 3' ends of its transcripts are generated by cleavage. The sequence AAUAAA, located 11-30 bases upstream of the cleavage site, provides the signal for both cleavage and polyadenylation. An endonuclease and the poly(A) polymerase are associated in a complex with other factors that confer specificity for the AAUAAA signal.

References

24.1 Introduction

rev Dreyfuss, G. et al. (1993). hnRNP proteins and the biogenesis of mRNA. Ann. Rev. Biochem. 62, 289-321.
Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat. Rev. Mol. Cell Biol. 3, 195-205.
Lewin, B. (1975). Units of transcription and translation: sequence components of hnRNA and mRNA. Cell 4, 77-93.

24.2 Nuclear splice junctions are short sequences

rev Padgett, R. A. (1986). Splicing of messenger RNA precursors. Ann. Rev. Biochem. 55, 1119-1150.
Sharp, P. A. (1987). Splicing of mRNA precursors. Science 235, 766-771.

24.4 pre-mRNA splicing proceeds through a lariat

rev Sharp, P.A. (1994). Split genes and RNA splicing. Cell 77, 805-815.
Weiner, A. (1993). mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism. Cell 72, 161-164.

ref Reed, R. and Maniatis, T. (1985). Intron sequences involved in lariat formation during pre-mRNA splicing. Cell 41, 95-105.
Reed, R. and Maniatis, T. (1986). A role for exon sequences and splice-site proximity in splice-site selection. Cell 46, 681-690.
Zhuang, Y. A., Goldstein, A. M., and Weiner, A. M. (1989). UACUAAC is the preferred branch site for mammalian mRNA splicing. Proc. Nat. Acad. Sci. USA 86, 2752-2756.

24.5 snRNAs are required for splicing

rev Guthrie, C. (1991). Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science 253, 157-163.
Guthrie, C. and Patterson, B. (1988). Spliceosomal snRNAs. Ann. Rev. Genet. 22, 387-419.
Maniatis, T. and Reed, R. (1987). The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. Nature 325, 673-678.

ref Grabowski, P. J., Seiler, S. R., and Sharp, P. A. (1985). A multicomponent complex is involved in the splicing of messenger RNA precursors. Cell 42, 345-353.
Zhou, Z., Licklider, L. J., Gygi, S. P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. Nature 419, 182-185.

24.6 U1 snRNP initiates splicing

rev Brow, D. A. (2002). Allosteric cascade of spliceosome activation. Ann. Rev. Genet. 36, 333-360.

ref Abovich, N. and Rosbash, M. (1997). Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. Cell 89, 403-412.
Berglund, J. A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell 89, 781-787.
Burgess, S., Couto, J. R., and Guthrie, C. (1990). A putative ATP binding protein influences the fidelity of branchpoint recognition in yeast splicing. Cell 60, 705-717.
Parker, R., Siliciano, P. G., and Guthrie, C. (1987). Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. Cell 49, 229-239.
Singh, R., Valcaircel, J., and Green, M. R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science 268, 1173-1176.
Wu, S., Romfo, C. M., Nilsen, T. W., and Green, M. R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature 402, 832-835.
Zamore, P. D. and Green, M. R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc. Nat. Acad. Sci. USA 86, 9243-9247.
Zhang, D. and Rosbash, M. (1999). Identification of eight proteins that cross-link to pre-mRNA in the yeast commitment complex. Genes Dev. 13, 581-592.
Zhuang, Y. and Weiner, A. M. (1986). A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. Cell 46, 827-835.

24.7 The E complex can be formed by intron definition or exon definition

ref Bruzik, J. P. and Maniatis, T. (1995). Enhancer-dependent interaction between 5' and 3' splice sites in *trans*. Proc. Nat. Acad. Sci. USA 92, 7056-7059.

24.8 5 snRNPs form the spliceosome

rev Kramer, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. Ann. Rev. Biochem. 65, 367-409.
Madhani, H. D. and Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. Ann. Rev. Genet. 28, 1-26.

ref Lamond, A. I. (1988). Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein. Proc. Nat. Acad. Sci. USA 85, 411-415.
Lesser, C. F. and Guthrie, C. (1993). Mutations in U6 snRNA that alter splice site specificity: implications for the active site. Science 262, 1982-1988.
Madhani, H. D. and Guthrie, C. (1992). A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. Cell 71, 803-817.
Newman, A. and Norman, C. (1991). Mutations in yeast U5 snRNA alter the specificity of 5' splice site cleavage. Cell 65, 115-123.
Sontheimer, E. J. and Steitz, J. A. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science 262, 1989-1996.

24.9 An alternative splicing apparatus uses different snRNPs

ref Burge, C. B., Padgett, R. A., and Sharp, P. A. (1998). Evolutionary fates and origins of U12-type introns. Mol. Cell 2, 773-785.
Dietrich, R. C, Incorvaia, R., and Padgett, R. A. (1997). Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns Molecular. Mol. Cell 1, 151-160.
Tarn, W.-Y. and Steitz, J. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class AT-AC intron *in vitro*. Cell 84, 801-811.

24.10 Splicing is connected to export of mRNA

rev Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat. Rev. Mol. Cell Biol. 3, 195-205.

Reed, R. and Hurt, E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell 108, 523-531.

ref Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M. J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. 20, 4987-4997.

Le Hir, H., Izaurralde, E., Maquat, L. E., and Moore, M. J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. 19, 6860-6869.

Kataoka, N., Yong, J., Kim, V. N., Velazquez, F., Perkinson, R. A., Wang, F., and Dreyfuss, G. (2000). Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. Mol. Cell 6, 673-682.

Luo, M. J. and Reed, R. (1999). Splicing is required for rapid and efficient mRNA export in metazoans. Proc. Nat. Acad. Sci. USA 96, 14937-14942.

Luo, M. L., Zhou, Z., Magni, K., Christoforides, C, Rappsilber, J., Mann, M., and Reed, R. (2001). Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. Nature 413, 644-647.

Reichert, V. L., Le Hir, H., Jurica, M. S., and Moore, M. J. (2002). 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. Genes Dev. 16, 2778-2791.

Rodrigues, J. P., Rode, M., Gatfield, D., Blencowe, B., Blencowe, M., and Izaurralde, E. (2001). REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. Proc. Nat. Acad. Sci. USA 98, 1030-1035.

Strasser, K. and Hurt, E. (2001). Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p. Nature 413, 648-652.

Zhou, Z., Luo, M. J., Straesser, K., Katahira, J., Hurt, E., and Reed, R. (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature 407, 401-405.

24.1 1 Group II introns autosplice via lariat formation

rev Michel, F. and Ferat, J.-L. (1995). Structure and activities of group II introns. Ann. Rev. Biochem. 64, 435-461.

24.12 Alternative splicing involves differential use of splice junctions

rev Green, M. R. (1991). Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Ann. Rev. Cell Biol. 7, 559-599.

ref Handa, N., Nureki, 0., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. Nature 398, 579-585.

Lynch, K. W. and Maniatis, T. (1996). Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. Genes Dev. 10, 2089-2101.

Sun, Q., Mayeda, A., Hampson, R. K., Krainer, A. R., and Rottman, F. M. (1993). General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. Genes Dev. 7, 2598-2608.

Tian, M. and Maniatis, T. (1993). A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. Cell 74, 105-114.

Wu, J. Y. and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell 75, 1061-1070.

24.13 frans-splicing reactions use small RNAs

rev Nilsen, T. (1993). trans-splicing of nematode pre-mRNA. Ann. Rev. Immunol. 47, 413-440.

ref Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., and Kim, S. K. (2002). A global analysis of *C. elegans* operons. Nature 417, 851-854.

Hannon, G. J. et al. (1990). frans-splicing of nematode pre-mRNA *in vitro.* Cell 61, 1247-1255.

Huang, X. Y. and Hirsh, D. (1989). A second *trans*-spliced RNA leader sequence in the nematode *C. elegans.* Proc. Nat. Acad. Sci. USA 86, 8640-8644.

Krause, M. and Hirsh, D. (1987). A *trans*-spliced leader sequence on actin mRNA in *C. elegans.* Cell 49, 753-761.

Murphy, W. J., Watkins, K. P., and Agabian, N. (1986). Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans-splicing. Cell 47, 517-525.

Sutton, R. and Boothroyd, J. C. (1986). Evidence for frans-splicing in trypanosomes. Cell 47, 527-535.

24.15 The splicing endonuclease recognizes tRNA

ref Baldi, I. M. et al. (1992). Participation of the intron in the reaction catalyzed by the *Xenopus* tRNA splicing endonuclease. Science 255, 1404-1408.

Di Nicola Negri, E., Fabbri, S., Bufardeci, E., Baldi, M. I., Mattoccia, E., and Tocchini-Valentini, G. P. (1997). The eucaryal tRNA splicing endonuclease recognizes a tripartite set of RNA elements. Cell 89, 859-866.

Diener, J. L. and Moore, P. B. (1998). Solution structure of a substrate for the archaeal pre-tRNA splicing endonucleases: the bulge-helix-bulge motif. Mol. Cell 1, 883-894.

Kleman-Leyer, K., Armbruster, D. W., and Daniels, C. J. (2000). Properties of H. volcanii tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Cell 89, 839-847.

Lykke-Andersen, J. and Garrett, R. A. (1997). RNA-protein interactions of an archaeal homotetrameric splicing endoribonuclease with an exceptional evolutionary history. EMBO J. 16, 6290-6300.

Mattoccia, E. et al. (1988). Site selection by the tRNA splicing endonuclease of *X. laevis.* Cell 55, 731-738.

Reyes, V. M. and Abelson, J. (1988). Substrate recognition and splice site determination in yeast tRNA splicing. Cell 55, 719-730.

Trotta, C. R., Miao, F., Arn, E. A., Stevens, S. W., Ho, C. K., Rauhut, R., and Abelson, J. N. (1997). The yeast tRNA splicing endonuclease: a tetrameric enzyme with two active site subunits homologous to the archaeal tRNA endonucleases. Cell 89, 849-858.

24.17 The unfolded protein response is related to tRNA splicing

ref Gonzalez, T. N., Sidrauski, C, Dorfler, S., and Walter, P. (1999). Mechanism of non-spliceosomal mRNA splicing in the unfolded protein response pathway. EMBO J. 18, 31 19-3132.

Sidrauski, C. and Walter, P. (1997). The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. Cell 90, 1031-1039.
Sidrauski, C, Cox, J. S., and Walter, P. (1996). tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. Cell 87, 405-413.

24.19 The 3' ends of mRNAs are generated by cleavage and polyadenylation

rev Wahle, E. and Keller, W. (1992). The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors. Ann. Rev. Biochem. 61, 419-440.

ref Bouvet, P., Omilli, F., Arlot-Bonnemains, Y., Legagneux, V., Roghi, C, Bassez, T., and Osborne, H. B. (1994). The deadenylation conferred by the 3' untranslated region of a developmentally controlled mRNA in *Xenopus* embryos is switched to polyadenylation by deletion of a short sequence element. Mol. Cell Biol. 14, 1893-1900.
Conway, L. and Wickens, M. (1985). A sequence downstream of AAUAAA is required for formation of SV40 late mRNA 3' termini in frog oocytes. Proc. Nat. Acad. Sci. USA 82, 3949-3953.
Fox, C. A., Sheets, M. D., and Wickens, M. P. (1989). Poly(A) addition during maturation of frog oocytes: distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU. Genes Dev. 3, 2151-2162.

Gil, A. and Proudfoot, N. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA 3' end formation. Cell 49, 399-406.

Karner, C. G., Wormington, M., Muckenthaler, M., Schneider, S., Dehlin, E., and Wahle, E. (1998). The deadenylating nuclease (DAN) is involved in poly(A) tail removal during the meiotic maturation of *Xenopus* oocytes. EMBO J. 17, 5427-5437.
McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B., and Richter, J. D. (1989). Poly(A) elongation during *Xenopus* oocyte maturation is required for translational recruitment and is mediated by a short sequence element. Genes Dev. 3, 803-815.
Takagaki, Y., Ryner, L. C, and Manley, J. L. (1988). Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. Cell 52, 731-742.
Voeltz, G. K. and Steitz, J. A. (1998). AUUUA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. Mol. Cell Biol. 18, 7537-7545.

24.20 Cleavage of the 3' end of histone mRNA may require a small RNA

rev Birnstiel, M. L. (1985). Transcription termination and 3' processing: the end is in site. Cell 41, 349-359.

ref Bond, U. M., Yario, T. A., and Steitz, J. A. (1991). Multiple processing-defective mutations in a mammalian histone pre-mRNA are suppressed by compensatory changes in U7 RNA both *in vitro* and *in vitro.* Genes Dev. 5, 1709-1722.
Dominski, Z., Erkmann, J. A., Greenland, J. A., and Marzluff, W. F. (2001). Mutations in the RNA binding domain of stem-loop binding protein define separable requirements for RNA binding and for histone pre-mRNA processing. Mol. Cell. Biol. 21, 2008-2017.

Galli, G. et al. (1983). Biochemical complementation with RNA in the *Xenopus* oocyte: a small RNA is required for the generation of 3' histone mRNA termini. Cell 34, 823-828.

Mowry, K. L. and Steitz, J. A. (1987). Identification of the human U7 snRNP as one of several factors involved in the 3' end maturation of histone premessenger RNA's. Science 238, 1682-1687.
Wang, Z. F., Whitfield, M. L., Ingledue, T. C, Dominski, Z., and Marzluff, W. F. (1996). The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing. Genes Dev. 10, 3028-3040.

24.21 Production of rRNA requires cleavage events

rev Venema, J. and Tollervey, D. (1999). Ribosome synthesis in *S. cerevisiae.* Ann. Rev. Genet. 33, 261-311.

24.22 Small RNAs are required for rRNA processing

exp Kiss, T. (2002). Small nucleolar RNAs guide rRNA modification (www.ergito.com/lookup.jsp?expt= kiss)

ref Balakin, A. G., Smith, L., and Fournier, M. J. (1996). The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. Cell 86, 823-834.
Bousquet-Antonelli, C, Henry, Y., G'elugne, J. P., Caizergues-Ferrer, M., and Kiss, T. (1997). A small nucleolar RNP protein is required for pseudouridylation of eukaryotic ribosomal RNAs. EMBO J. 16, 4770-4776.
Ganot, P., Bortolin, M. L., and Kiss, T. (1997). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. Cell 89, 799-809.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev. 1 1, 941 -956.
Kass, S. et al. (1990). The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. Cell 60, 897-908.
Kiss-Laszlo, Z. et al. (1996). Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell 85, 1077-1068.
Kiss-Laszlo, Z., Henry, Y., and Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. EMBO J. 17, 797-807.

Ni, J., Tien, A. L., and Fournier, M. J. (1997). Small nucleolar RNAs direct site-specific synthesis of pseudouridine in rRNA. Cell 89, 565-573.

Chapter 25

Catalytic RNA

25.1 Introduction

The idea that only proteins have enzymatic activity was deeply rooted in biochemistry. (Yet devotees of protein function once thought that only proteins could have the versatility to be the genetic material!) A rationale for the identification of enzymes with proteins lies in the view that only proteins, with their varied three-dimensional structures and variety of side-groups, have the flexibility to create the active sites that catalyze biochemical reactions. But the characterization of systems involved in RNA processing has shown this view to be an over simplification.

Several types of catalytic reactions are now known to reside in RNA. Ribozyme has become a general term used to describe an RNA with catalytic activity, and it is possible to characterize the enzymatic activity in the same way as a more conventional enzyme. Some RNA catalytic activities are directed against separate substrates, while others are intramolecular (which limits the catalytic action to a single cycle).

Introns of the group I and group II classes possess the ability to splice themselves out of the pre-mRNA that contains them. Engineering of group I introns has generated RNA molecules that have several other catalytic activities related to the original activity.

The enzyme ribonuclease P is a ribonucleoprotein that contains a single RNA molecule bound to a protein. The RNA possesses the ability to catalyze cleavage in a tRNA substrate, while the protein component plays an indirect role, probably to maintain the structure of the catalytic RNA.

The common theme of these reactions is that the RNA can perform an intramolecular or intermolecular reaction that involves cleavage or joining of phosphodiester bonds *in vitro*. Although the specificity of the reaction and the basic catalytic activity is provided by RNA, proteins associated with the RNA may be needed for the reaction to occur efficiently *in vivo*.

RNA splicing is not the only means by which changes can be introduced in the informational content of RNA. In the process of RNA editing, changes are introduced at individual bases, or bases are added at particular positions within an mRNA. The insertion of bases (most commonly uridine residues) occurs for several genes in the mitochondria of certain lower eukaryotes; like splicing, it involves the breakage and reunion of bonds between nucleotides, but also requires a template for coding the information of the new sequence.

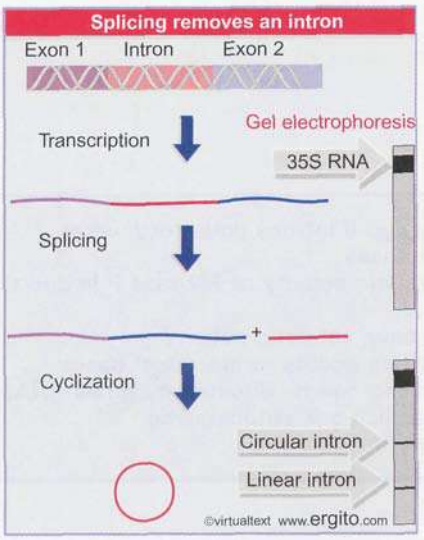


Figure 25.1 Splicing of the *Tetrahymena* 35S rRNA precursor can be followed by gel electrophoresis. The excised intron forms two bands, corresponding to circular and linear forms.

25.2 Group I introns undertake self-splicing by transesterification

Key Concepts

- The only factors required for autosplicing *in vitro* by group I introns are a monovalent cation, a divalent cation, and a guanine nucleotide.
- Splicing occurs by two transesterifications, without requiring input of energy.
- The 3′-OH end of the guanine cofactor attacks the 5′ end of the intron in the first transesterification.
- The 3′-OH end generated at the end of the first exon attacks the junction between the intron and second exon in the second transesterification.
- The intron is released as a linear molecule that circularizes when its 3′-OH terminus attacks a bond at one of two internal positions.
- The G^{414}–A^{16} internal bond of the intron can also be attacked by other nucleotides in a *trans*-splicing reaction.

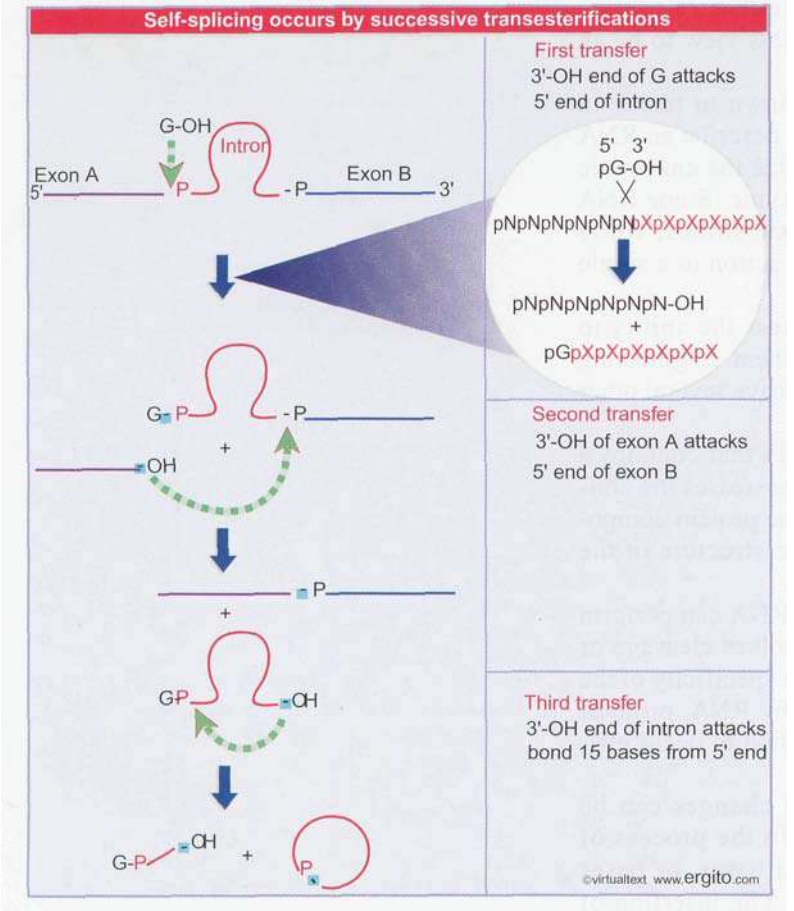


Figure 25.2 Self-splicing occurs by transesterification reactions in which bonds are exchanged directly. The bonds that have been generated at each stage are indicated by the shaded boxes.

Group I introns are found in diverse locations. They occur in the genes coding for rRNA in the nuclei of the lower eukaryotes *Tetrahymena thermophila* (a ciliate) and *Physarum polycephalum* (a slime mold). They are common in the genes of fungal mitochondria. They are present in three genes of phage T4 and also are found in bacteria. Group I introns have an intrinsic ability to splice themselves. This is called self-splicing or autosplicing.

Self-splicing was discovered as a property of the transcripts of the rRNA genes in *T. thermophila*. The genes for the two major rRNAs follow the usual organization, in which both are expressed as part of a common transcription unit. The product is a 35S precursor RNA with the sequence of the small rRNA in the 5′ part, and the sequence of the larger (26S) rRNA toward the 3' end.

In some strains of *T. thermophila*, the sequence coding for 26S rRNA is interrupted by a single, short intron. When the 35S precursor RNA is incubated *in vitro*, splicing occurs as an autonomous reaction. The intron is excised from the precursor and accumulates as a linear fragment of 400 bases, which is subsequently converted to a circular RNA. These events are summarized in **Figure 25.1.**

The reaction requires only a monovalent cation, a divalent cation, and a guanine nucleotide cofactor. No other base can be substituted for G; but a triphosphate is not needed; GTP, GDP, GMP, and guanosine itself all can be used, so there is no net energy requirement. The guanine nucleotide must have a 3′–OH group.

The fate of the guanine nucleotide can be followed by using a radioactive label. The radioactivity initially enters the excised linear intron fragment. The G residue becomes linked to the 5' end of the intron by a normal phosphodiester bond.

By Book_Crazy [IND]

Figure 25.2 shows that three transfer reactions occur. In the first transfer, the guanine nucleotide behaves as a cofactor that provides a free 3′–OH group that attacks the 5' end of the intron. This reaction creates the G-intron link and generates a 3′–OH group at the end of the exon. The second transfer involves a similar chemical reaction, in which this 3′–OH then attacks the second exon. The two transfers are connected; no free exons have been observed, so their ligation may occur as part of the same reaction that releases the intron. The intron is released as a linear molecule, but the third transfer reaction converts it to a circle.

Each stage of the self-splicing reaction occurs by a transesterification, in which one phosphate ester is converted directly into another, without any intermediary hydrolysis. Bonds are exchanged directly, and energy is conserved, so the reaction does not require input of energy from hydrolysis of ATP or GTP. (There is a parallel for the transfer of bonds without net input of energy in the DNA nicking-closing enzymes discussed in 75 *Recombination and repair.)*

If each of the consecutive transesterification reactions involves no net change of energy, why does the splicing reaction proceed to completion instead of coming to equilibrium between spliced product and nonspliced precursor? The concentration of GTP is high relative to that of RNA, and therefore drives the reaction forward; and a change in secondary structure in the RNA prevents the reverse reaction.

The in vitro system includes no protein so the ability to splice is intrinsic to the RNA. The RNA forms a specific secondary/tertiary structure in which the relevant groups are brought into juxtaposition so that a guanine nucleotide can be bound to a specific site and then the bond breakage and reunion reactions shown in Figure 25.2 can occur. Although a property of the RNA itself, the reaction is assisted *in vivo* by proteins, which stabilize the RNA structure.

The ability to engage in these transfer reactions resides with the sequence of the intron, which continues to be reactive after its excision as a linear molecule. **Figure 25.3** summarizes its activities.

The intron can circularize when the 3' terminal G attacks either of two positions near the 5' end. The internal bond is broken and the new 5' end is transferred to the 3′–OH end of the intron. The *primary cyclization* usually involves reaction between the terminal G and the A^{16}. This is the most common reaction (shown as the third transfer in Figure 25.2). Less frequently, the G^{414} reacts with U^{20}. Each reaction generates a circular intron and a linear fragment that represents the original 5' region (15 bases long for attack on A^{16}, 19 bases long for attack on U^{20}). The released 5' fragment contains the original added guanine nucleotide.

Either type of circle can regenerate a linear molecule *in vitro* by specifically hydrolyzing the bond (G^{414}–A^{16} or G^{414}–U^{20}) that had closed the circle. This is called a *reverse cyclization.* The linear molecule generated by reversing the primary cyclization at A^{16} remains reactive, and can perform a secondary cyclization by attacking U^{20}.

The final product of the spontaneous reactions following release of the intron is the L-19 RNA, a linear molecule generated by reversing the shorter circular form. This molecule has an enzymatic activity that allows it to catalyze the extension of short oligonucleotides (not shown in the figure, but see Figure 25.8).

The reactivity of the released intron extends beyond merely reversing the cyclization reaction. Addition of the oligonucleotide UUU reopens the primary circle by reacting with the G^{414}–A^{16} bond. The UUU (which resembles the 3' end of the 15-mer released by the primary cyclization) becomes the 5' end of the linear molecule that is formed. This is an *intermolecular* reaction, and thus demonstrates the ability to connect together two different RNA molecules.

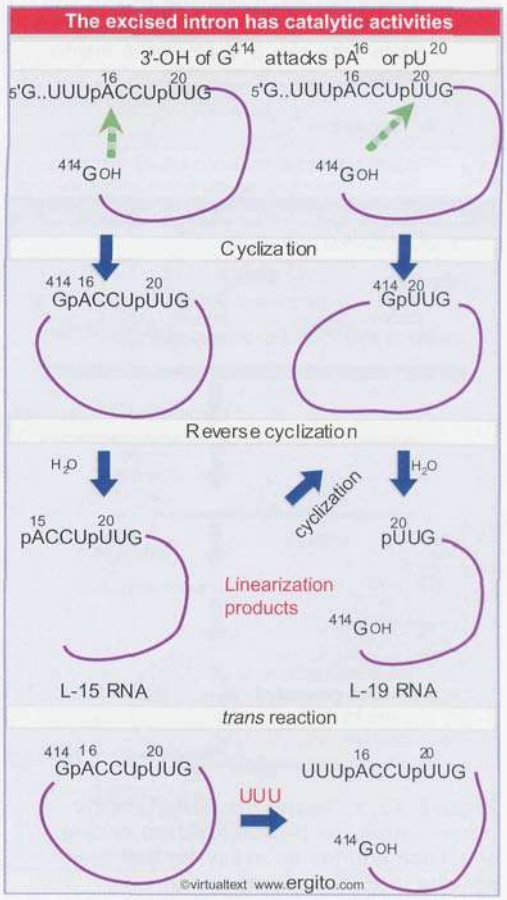


Figure 25.3 The excised intron can form circles by using either of two internal sites for reaction with the 5′ end and can reopen the circles by reaction with water or oligonucleotides.

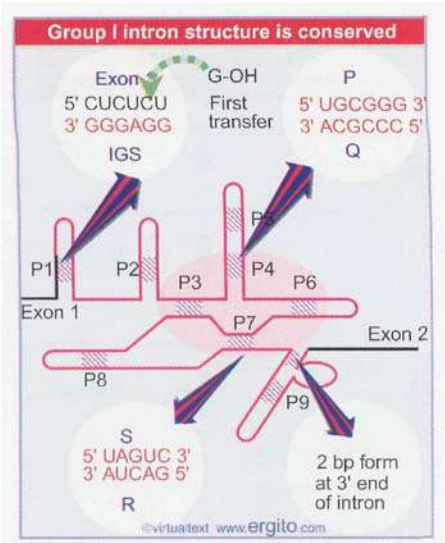


Figure 25.4 Group I introns have a common secondary structure that is formed by 9 base-paired regions. The sequences of regions P4 and P7 are conserved, and identify the individual sequence elements P, Q, R, and S. P1 is created by pairing between the end of the left exon and the IGS of the intron; a region between P7 and P9 pairs with the 3′ end of the intron.

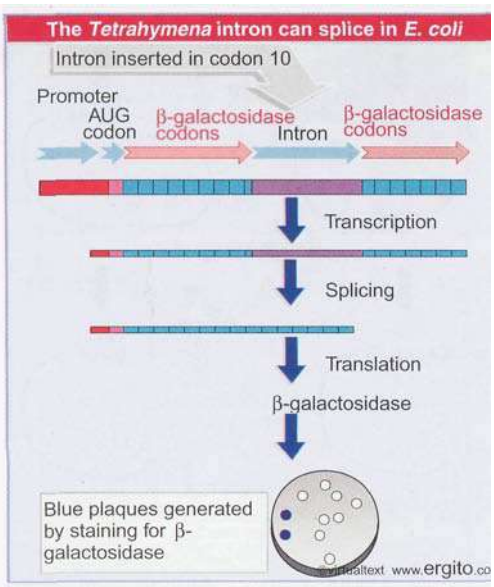


Figure 25.5 Placing the *Tetrahymena* intron within the β-galactosidase coding sequence creates an assay for self-splicing in *E. coli*. Synthesis of β-galactosidase can be tested by adding a compound that is turned blue by the enzyme. The sequence is carried by a bacteriophage, so the presence of blue plaques (containing infected bacteria) indicates successful splicing.

This series of reactions demonstrates vividly that the autocatalytic activity reflects a generalized ability of the RNA molecule to form an active center that can bind guanine cofactors, recognize oligonucleotides, and bring together the reacting groups in a conformation that allows bonds to be broken and rejoined. Other group I introns have not been investigated in as much detail as the *Tetrahymena* intron, but their properties are generally similar.

The autosplicing reaction is an intrinsic property of RNA *in vitro*, but to what degree are proteins involved *in vivo?* Some indications for the involvement of proteins are provided by mitochondrial systems, where splicing of group I introns requires the *trans-acting* products of other genes. One striking case is presented by the *cytl8* mutant *of N. crassa,* which is defective in splicing several mitochondrial group I introns. The product of this gene turns out to be the mitochondrial tyrosyl-tRNA synthetase! This is explained by the fact that the intron can take up a tRNA-like tertiary structure that is stabilized by the synthetase and which promotes the catalytic reaction.

This relationship between the synthetase and splicing is consistent with the idea that splicing originated as an RNA-mediated reaction, subsequently assisted by RNA-binding proteins that originally had other functions. The *in vitro* self-splicing ability may represent the basic biochemical interaction. The RNA structure creates the active site, but is able to function efficiently *in vivo* only when assisted by a protein complex.

25.3 Group I introns form a characteristic secondary structure

Key Concepts

- Group I introns form a secondary structure with 9 duplex regions.
- The core of regions P3, P4, P6, P7 has catalytic activity.
- Regions P4 and P7 are both formed by pairing between conserved consensus sequences.
- A sequence adjacent to P7 base pairs with the sequence that contains the reactive G.

All group I introns can be organized into a characteristic secondary structure with 9 helices (P1-P9). **Figure** 25.4shows a model for the secondary structure of the *Tetrahymena* intron. Two of the base-paired regions are generated by pairing between conserved sequence elements that are common to group I introns. P4 is constructed from the sequences *P* and *Q;* P7 is formed from sequences *R* and *S.* The other base-paired regions vary in sequence in individual introns. Mutational analysis identifies an intron "core," containing P3, P4, P6, and P7, which provides the minimal region that can undertake a catalytic reaction. The lengths of group I introns vary widely, and the consensus sequences are located a considerable distance from the actual splice junctions.

Some of the pairing reactions are directly involved in bringing the splice junctions into a conformation that supports the enzymatic reaction. P1 includes the 3' end of the left exon. The sequence within the intron that pairs with the exon is called the 1GS, or internal guide sequence. (Its name reflects the fact that originally the region immediately 3' to the IGS sequence shown in the figure was thought to pair with the 3' splice junction, thus bringing the two junctions together. This interaction may occur, but does not seem to be essential.) A very short sequence, sometimes as short as 2 bases, between P7 and P9, base

pairs with the sequence that immediately precedes the reactive G (position 414 in *Tetrahymena)* at the 3' end of the intron.

The importance of base pairing in creating the necessary core structure in the RNA is emphasized by the properties of *cis*-acting mutations that prevent splicing of group I introns. Such mutations have been isolated for the mitochondrial introns through mutants that cannot remove an intron *in vivo,* and they have been isolated for the *Tetrahymena* intron by transferring the splicing reaction into a bacterial environment. The construct shown in **Figure 25.5** allows the splicing reaction to be followed in *E. coli.* The self-splicing intron is placed at a location that interrupts the tenth codon of the β-galactosidase coding sequence. The protein can therefore be successfully translated from an RNA only after the intron has been removed.

The synthesis of β-galactosidase in this system indicates that splicing can occur in conditions quite distant from those prevailing in *Tetrahymena* or even *in vitro.* One interpretation of this result is that self-splicing can occur in the bacterial cell. Another possibility is that there are bacterial proteins that assist the reaction.

Using this assay, we can introduce mutations into the intron to see whether they prevent the reaction. Mutations in the group I consensus sequences that disrupt their base pairing stop splicing. The mutations can be reverted by making compensating changes that restore base pairing.

Mutations in the corresponding consensus sequences in mitochondrial group I introns have similar effects. A mutation in one consensus sequence may be reverted by a mutation in the complementary consensus sequence to restore pairing; for example, mutations in the R consensus can be compensated by mutations in the S consensus.

Together these results suggest that the group I splicing reaction depends on the formation of secondary structure between pairs of consensus sequences within the intron. The principle established by this work is that *sequences distant from the splice junctions themselves are required to form the active site that makes self-splicing possible.*

25.4 Ribozymes have various catalytic activities

Key Concepts

- By changing the substrate binding-site of a group I intron, it is possible to introduce alternative sequences that interact with the reactive G.
- The reactions follow classical enzyme kinetics with a low catalytic rate.
- Reactions using 2′-OH bonds could have been the basis for evolving the original catalytic activities in RNA.

The catalytic activity of group I introns was discovered by virtue of their ability to autosplice, but they are able to undertake other catalytic reactions *in vitro.* All of these reactions are based on transesterifications. We analyze these reactions in terms of their relationship to the splicing reaction itself.

The catalytic activity of a group 1 intron is conferred by its ability to generate a particular secondary and tertiary structure that creates active sites, equivalent to the active sites of a conventional (proteinaceous) enzyme. **Figure 25.6** illustrates the splicing reaction in terms of these sites (this is the same series of reactions shown previously in Figure 25.2).

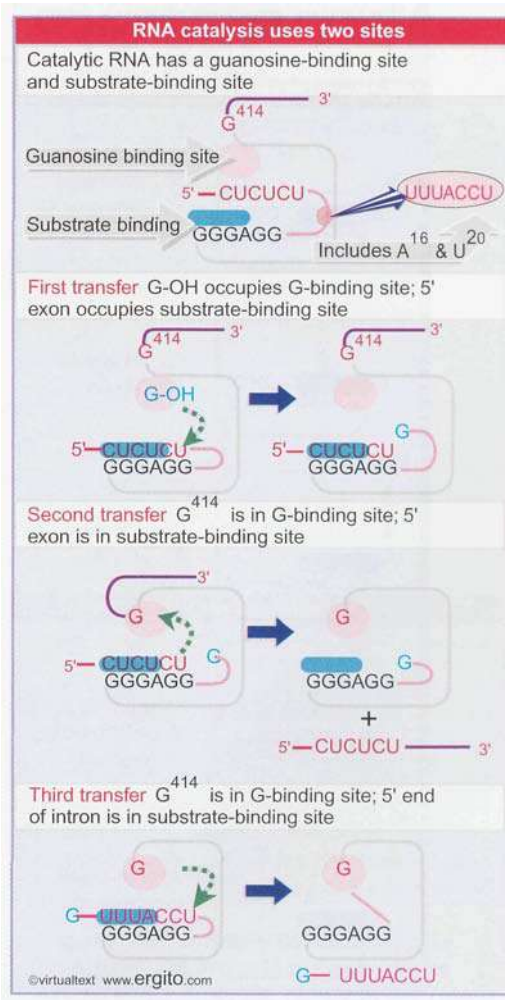


Figure 25.6 Excision of the group I intron in *Tetrahymena* rRNA occurs by successive reactions between the occupants of the guanosine-binding site and substrate-binding site. The left exon is red, and the right exon is purple.

By Book_Crazy [IND]

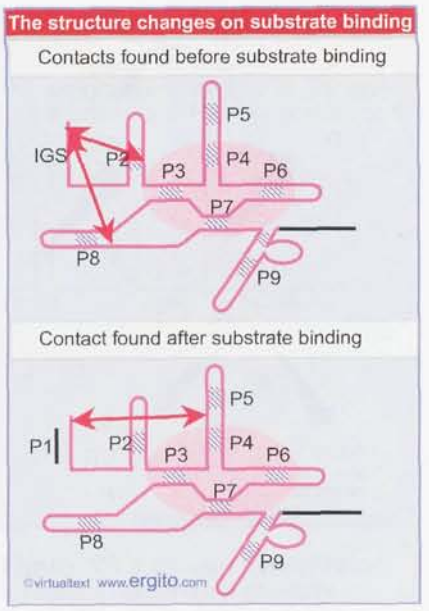


Figure 25.7 The position of the IGS in the tertiary structure changes when P1 is formed by substrate binding.

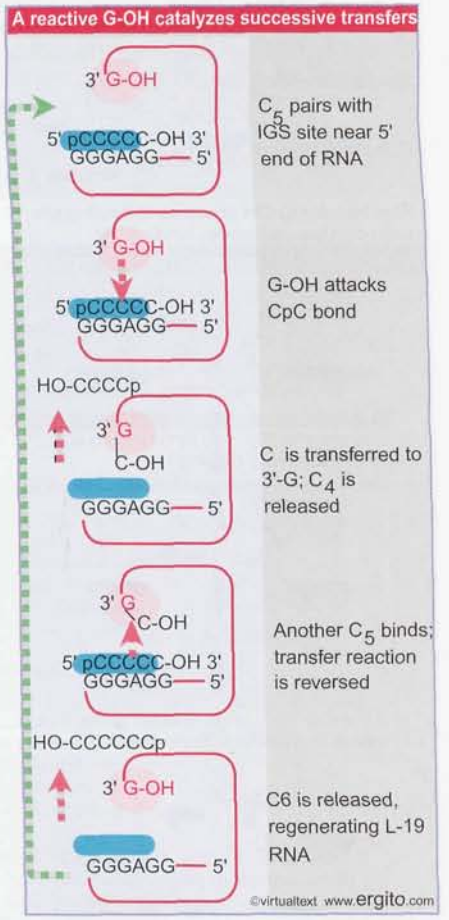


Figure 25.8 The L-19 linear RNA can bind C in the substrate-binding site; the reactive G-OH 3′ end is located in the G-binding site, and catalyzes transfer reactions that convert 2 C5 oligonucleotides into a C4 and a C6 oligonucleotide.

The substrate-binding site is formed from the P1 helix, in which the 3' end of the first intron base pairs with the IGS in an intermolecular reaction. A guanosine-binding site is formed by sequences in P7. This site may be occupied either by a free guanosine nucleotide or by the G residue in position 414. In the first transfer reaction, it is used by free guanosine nucleotide; but it is subsequently occupied by G^{414}. The second transfer releases the joined exons. The third transfer creates the circular intron.

Binding to the substrate involves a change of conformation; before substrate binding, the 5' end of the 1GS is close to P2 and P8, but after binding, when it forms the P1 helix, it is close to conserved bases that lie between P4 and P5. The reaction is visualized by contacts that are detected in the secondary structure in **Figure 25.7.** In the tertiary structure, the two sites alternatively contacted by P1 are 37 A apart, which implies a substantial movement in the position of P1.

The L-19 RNA is generated by opening the circular intron (shown as the last stage of the intramolecular rearrangements shown in Figure 25.3). It still retains enzymatic abilities. These resemble the activities involved in the original splicing reaction, and we may consider ribozyme function in terms of the ability to bind an intramolecular sequence complementary to the IGS in the substrate-binding site, while binding either the terminal $G^{4,4}$ or a free G-nucleotide in the G-binding site.

Figure 25.8 illustrates the mechanism by which the oligonucleotide C_5 is extended to generate a C_6 chain. The C_5 oligonucleotide binds in the substrate-binding site, while G^{414} occupies the G-binding site. By transesterification reactions, a C is transferred from C_5 to the 3′-terminal G, and then back to a new C_5 molecule. Further transfer reactions lead to the accumulation of longer cytosine oligonucleotides. The reaction is a true catalysis because the L-19 RNA remains unchanged and is available to catalyze multiple cycles. The ribozyme is behaving as a nucleotidyl transferase.

Some further enzymatic reactions are characterized in **Figure 25.9.** The ribozyme can function as a sequence-specific endoribonuclease by utilizing the ability of the IGS to bind complementary sequences. In this example, it binds an external substrate containing the sequence CUCU, instead of binding the analogous sequence that is usually contained at the end of the left exon. A guanine-containing nucleotide is present in the G-binding site and attacks the CUCU sequence in precisely the same way that the exon is usually attacked in the first transfer reaction. This cleaves the target sequence into a 5' molecule that resembles the left exon and a 3′molecule that bears a terminal G residue. By mutating the IGS element, it is possible to change the specificity of the ribozyme, so that it recognizes sequences complementary to the new sequence at the IGS region.

Altering the IGS, so that the specificity of the substrate-binding site is changed to enable other RNA targets to enter, can be used to generate a ligase activity. An RNA terminating in a 3′–OH is bound in the substrate site, and an RNA terminating in a 5'-G residue is bound in the G-binding site. An attack by the hydroxyl on the phosphate bond connects the two RNA molecules, with the loss of the G residue.

The phosphatase reaction is not directly related to the splicing transfer reactions. An oligonucleotide sequence that is complementary to the IGS and terminates in a 3′–phosphate can be attacked by the G^{414}. The phosphate is transferred to the G^{414}, and an oligonucleotide with a free 3′–OH end is then released. The phosphate can then be transferred either to an oligonucleotide terminating in 3′–OH (effectively reversing the reaction) or indeed to water (releasing inorganic phosphate and completing an authentic phosphatase reaction).

The reactions catalyzed by RNA can be characterized in the same way as classical enzymatic reactions in terms of Michaelis-Menten kinetics. **Figure 25.10** analyzes the reactions catalyzed by RNA. The K_M values for RNA-catalyzed reactions are low, and therefore imply that the RNA can bind its substrate with high specificity. The turnover numbers are low, which reflects a low catalytic rate. In effect, the RNA molecules behave in the same general manner as traditionally defined for enzymes, although they are relatively slow compared to protein catalysts (where a typical range of turnover numbers is 10^3–10^6).

How does RNA provide a catalytic center? Its ability seems reasonable if we think of an active center as a surface that exposes a series of active groups in a fixed relationship. In a protein, the active groups are provided by the side chains of the amino acids, which have appreciable variety, including positive and negative ionic groups and hydrophobic groups. In an RNA, the available moieties are more restricted, consisting primarily of the exposed groups of bases. Short regions are held in a particular structure by the secondary /tertiary conformation of the molecule, providing a surface of active groups able to maintain an environment in which bonds can be broken and made in another molecule. It seems inevitable that the interaction between the RNA catalyst and the RNA substrate will rely on base pairing to create the environment. Divalent cations (typically Mg^{2+}) play an important role in structure, typically being present at the active site where they coordinate the positions of the various groups. They play a direct role in the endonucleolytic activity of virusoid ribozymes (see *25.8 Viroids have catalytic activity*).

The evolutionary implications of these discoveries are intriguing. The split personality of the genetic apparatus, in which RNA is present in all components, but proteins undertake catalytic reactions, has always been puzzling. It seems unlikely that the very first replicating systems could have contained both nucleic acid and protein.

But suppose that the first systems contained only a self-replicating nucleic acid with primitive catalytic activities, just those needed to make and break phosphodiester bonds. If we suppose that the involvement of 2′– OH bonds in current splicing reactions is derived from these primitive catalytic activities, we may argue that the original nucleic acid was RNA, since DNA lacks the 2′–OH group and therefore could not undertake such reactions. Proteins could have been added for their ability to stabilize the RNA structure. Then the greater versatility of proteins could have allowed them to take over catalytic reactions, leading eventually to the complex and sophisticated apparatus of modern gene expression.

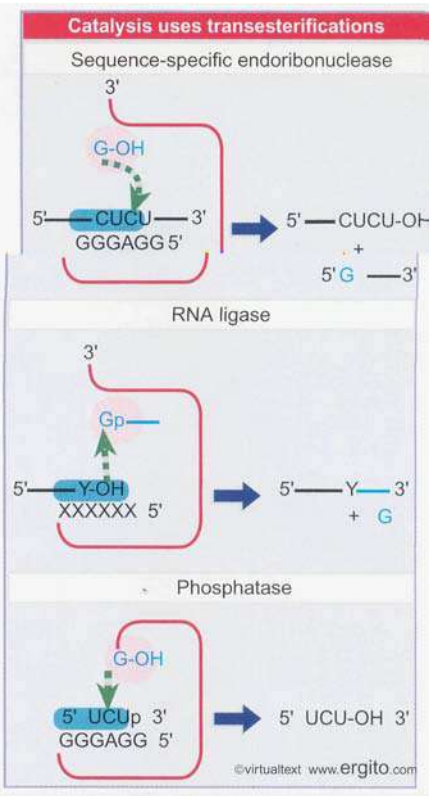


Figure 25.9 Catalytic reactions of the ribozyme involve transesterifications between a group in the substrate-binding site and a group in the G-binding site.

RNA catalysis is enzymatic

Enzyme	Substrate	K_M (mM)	Turnover (/min)
19 base virusoid	24 base RNA	0.0006	0.5
L-19 Intron	CCCCCC	0.04	1.7
RNAase P RNA	pre-tRNA	0.00003	0.4
RNAase P complete	pre-tRNA	0.00003	29
RNAase T1	GpA	0.05	5,700
β-galactosidase	lactose	4.0	12,500

Figure 25.10 Reactions catalyzed by RNA have the same features as those catalyzed by proteins, although the rate is slower. The K_M gives the concentration of substrate required for half-maximum velocity; this is an inverse measure of the affinity of the enzyme for substrate. The turnover number gives the number of substrate molecules transformed in unit time by a single catalytic site.

25.5 Some group I introns code for endonucleases that sponsor mobility

Key Concepts

- Mobile introns are able to insert themselves into new sites.
- Mobile group I introns code for an endonuclease that makes a double-strand break at a target site.
- The intron transposes into the site of the double-strand break by a DNA-mediated replicative mechanism.

Certain introns of both the group I and group II classes contain open reading frames that are translated into proteins. Expression of the proteins allows the intron (either in its original DNA form or as a DNA copy of the RNA) to be *mobile*: it is able to insert itself into

a new genomic site. Introns of both groups I and II are extremely widespread, being found in both prokaryotes and eukaryotes. Group I introns migrate by DNA-mediated mechanisms, whereas group II introns migrate by RNA-mediated mechanisms. Both types of intron may have maturase activities, which are needed for splicing out the particular intron from the pre-mRNA.

Intron mobility was first detected by crosses in which the alleles for the relevant gene differ with regard to their possession of the intron. Polymorphisms for the presence or absence of introns are common in fungal mitochondria. This is consistent with the view that these introns originated by insertion into the gene. Some light on the process that could be involved is cast by an analysis of recombination in crosses involving the large rRNA gene of the yeast mitochondrion.

This gene has a group I intron that contains a coding sequence. The intron is present in some strains of yeast (called ω^+) but absent in others (ω^-). Genetic crosses between ω^+ and ω^- are *polar:* the progeny are usually ω^+.

If we think of the ω^+ strain as a donor and the ω^- strain as a recipient, we form the view that in ω^+ X ω^- crosses a new copy of the intron is generated in the ω^- genome. As a result, the progeny are all ω^+.

Mutations can occur in either parent to abolish the polarity. Mutants show normal segregation, with equal numbers of ω^+ and ω^- progeny. The mutations indicate the nature of the process. Mutations in the ω^- strain occur close to the site where the intron would be inserted. Mutations in the ω^+ strain lie in the reading frame of the intron and prevent production of the protein. This suggests the model of **Figure 25.11,** in which the protein coded by the intron in an ω^+ strain recognizes the site where the intron should be inserted in an ω^- strain and causes it to be preferentially inherited.

What is the action of the protein? The product of the ω intron is an endonuclease *that recognizes the* ω^- *gene as a target for a double-strand break.* The endonuclease recognizes an 18 bp target sequence that contains the site where the intron is inserted. The target sequence is cleaved on each strand of DNA 2 bases to the 3′ side of the insertion site. So the cleavage sites are 4 bp apart and generate overhanging single strands.

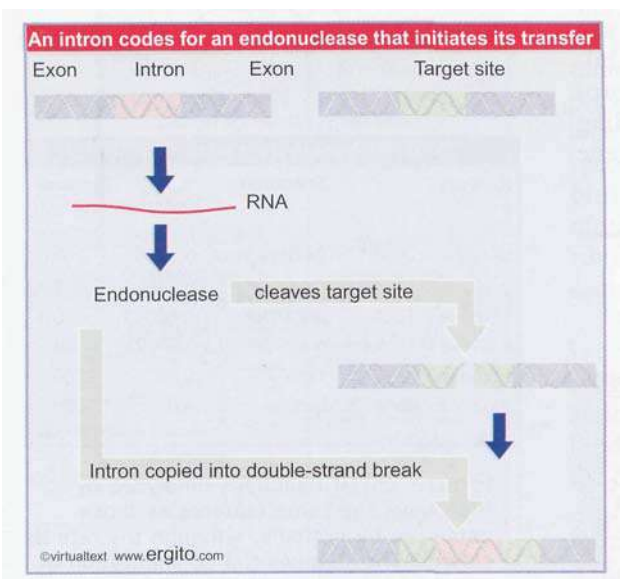


Figure 25.11 An intron codes for an endonuclease that makes a double-strand break in DNA. The sequence of the intron is duplicated and then inserted at the break.

This type of cleavage is related to the cleavage characteristic of transposons when they migrate to new sites (see *16 Transposons).* The double-strand break probably initiates a gene conversion process in which the sequence of the ω^+ gene is copied to replace the sequence of the ω^- gene. The reaction involves transposition by a duplicative mechanism, and occurs solely at the level of DNA. Insertion of the intron interrupts the sequence recognized by the endonuclease, thus ensuring stability.

Other group I introns that contain open reading frames also are mobile. The general mechanism of intron perpetuation appears to be the same: the intron codes for an endonuclease that cleaves a specific target site where the intron will be inserted. There are differences in the details of insertion; for example, the endonuclease coded by the phage T4 *td* intron cleaves a target site that is 24 bp upstream of the site at which the intron is itself inserted.

In spite of the common mechanism for intron mobility, there is no homology between the sequences of the target sites or the intron coding regions. We assume that the introns have a common evolutionary origin, but evidently they have diverged greatly. The target sites are among the longest and therefore the most specific known for any endonucleases. The specificity ensures that the intron perpetuates itself only by insertion into a single target site and not elsewhere in the genome. This is called intron homing.

Introns carrying sequences that code for endonucleases are found in a variety of bacteria and lower eukaryotes. These results strengthen the view that introns carrying coding sequences originated as independent elements that coded for a function involved in the ability to be spliced out of RNA or to migrate between DNA molecules. Consistent with this idea, the pattern of codon usage is somewhat different in the intron coding regions from that found in the exons.

25.6 Some group II introns code for reverse transcriptases

Key Concepts

- Some group II introns code for a reverse transcriptase that generates a DNA copy of the RNA sequence that transposes by a retroposon-like mechanism.

Most of the open reading frames contained in group II introns have regions that are related to reverse transcriptases. Introns of this type are found in organelles of lower eukaryotes and also in some bacteria. The reverse transcriptase activity is specific for the intron and is involved in homing. The reverse transcriptase generates a DNA copy of the intron from the pre-mRNA, and thus allows the intron to become mobile by a mechanism resembling that of retroviruses (see *17.2 The retrovirus life cycle involves transposition-like events).* The type of retrotransposition involved in this case resembles that of a group of retroposons that lack LTRs, and which generate the 3′–OH needed for priming by making a nick in the target (see Figure 17.20 in *17.12 LINES use an endonuclease to generate a priming end).*

The best characterized mobile group II introns code for a single protein in a region of the intron beyond its catalytic core. The typical protein contains an N-terminal reverse transcriptase activity, a central domain associated with maturase activity, and a C-terminal endonuclease domain. The endonuclease initiates the transposition reaction, and thus plays the same role in homing as its counterpart in a group I intron. The reverse transcriptase generates a DNA copy of the intron that is inserted at the homing site. The endonuclease also cleaves target sites that resemble, but are not identical to the homing site, at much lower frequency, leading to insertion of the intron at new locations.

Figure *25.12* *illustrates the transposition reaction for a typical* group II intron. The endonuclease makes a double-strand break at the target site. A 3' end is generated at the site of the break and provides a primer for the reverse transcriptase. The intron RNA provides the template for the synthesis of cDNA. Because the RNA includes exon sequences on either side of the intron, the cDNA product is longer than the region of the intron itself, so that it can span the double-strand break, allowing the cDNA to repair the break. The result is the insertion of the intron.

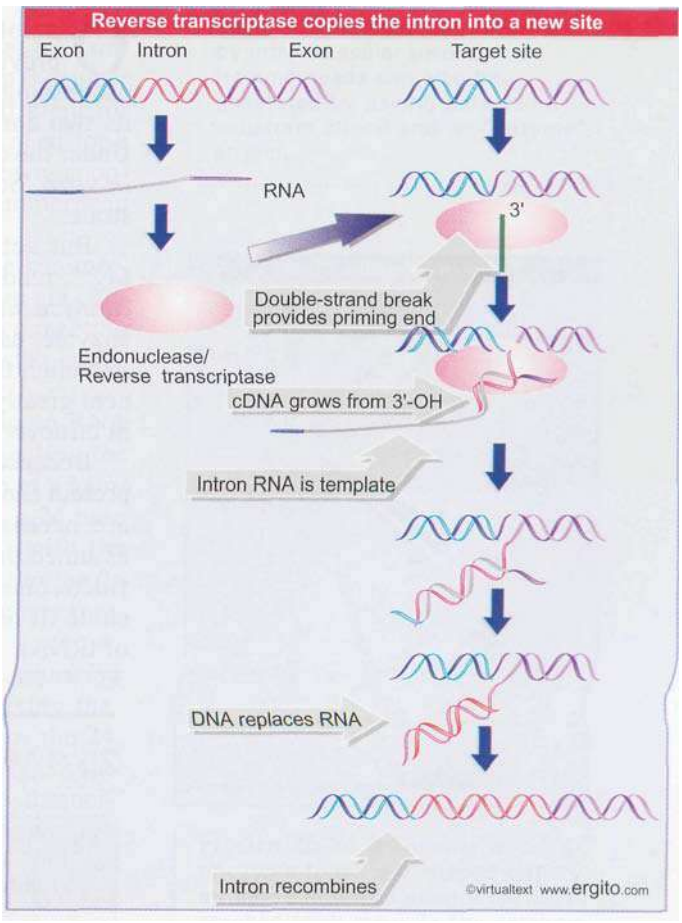


Figure 25.12 Reverse transcriptase coded by an intron allows a copy of the RNA to be inserted at a target site generated by a double-strand break.

An *in vitro* system for mobility can be generated by incubating a ribonucleoprotein preparation with a substrate DNA. The ribonucleoprotein includes the RNA containing a group II intron and its protein

product. It contains an endonuclease activity that makes a staggered double-strand break at the appropriate target site. Both the RNA and protein components of the ribonucleoprotein are required for cleavage, possibly both in catalytic capacities.

The maturase activity is required for splicing of the intron, rather than for mobility. Its basic role is to assist the folding of the catalytic core to form an active site. Some group II introns that do not code for maturase activities may use comparable proteins that are coded by sequences in the host genome. This suggests a possible route for the evolution of splicing factors. The factor may initially have been coded by a group II intron, the coding sequence became isolated from the intron in the host genome, and then it evolved to function with a wider range of substrates that the original intron sequence. The catalytic core of the intron could have evolved into an snRNA.

25.7 The catalytic activity of RNAase P is due to RNA

Key Concepts

- Ribonuclease P is a ribonucleoprotein in which the RNA has catalytic activity.

One of the first demonstrations of the capabilities of RNA was provided by the dissection of ribonuclease P, an *E. coli* tRNA-processing endonuclease. Ribonuclease P can be dissociated into its two components, the 375 base RNA and the 20 kD polypeptide. Under the conditions initially used to characterize the enzyme activity *in vitro*, both components were necessary to cleave the tRNA substrate.

But a change in ionic conditions, an increase in the concentration of Mg^{2+}, renders the protein component superfluous. *The RNA alone can catalyze the reaction!* Analyzing the results as though the RNA were an enzyme, each "enzyme" catalyzes the cleavage of multiple substrates. Although the catalytic activity resides in the RNA, the protein component greatly increases the speed of the reaction, as seen in the increase in turnover number (see Figure 25.10).

Because mutations in either the gene for the RNA or the gene for protein can inactivate RNAase P *in vivo*, we know that both components are necessary for natural enzyme activity. Originally it had been assumed that the protein provided the catalytic activity, while the RNA filled some subsidiary role, for example, assisting in the binding of substrate (it has some short sequences complementary to exposed regions of tRNA). But these roles are reversed!

25.8 Viroids have catalytic activity

Key Concepts

- Viroids and virusoids form a hammerhead structure that has a self-cleaving activity.
- Similar structures can be generated by pairing a substrate strand that is cleaved by an enzyme strand.
- When an enzyme strand is introduced into a cell, it can pair with a substrate strand target that is then cleaved.

Another example of the ability of RNA to function as an endonuclease is provided by some small plant RNAs (~350 bases) that undertake a self-cleavage reaction. As with the case of the *Tetrahymena* group I intron, however, it is possible to engineer constructs that can function on external substrates.

These small plant RNAs fall into two general groups: viroids and virusoids. The viroids are infectious RNA molecules that function independently, without enacapsidation by any protein coat. The virusoids are similar in organization, but are encapsidated by plant viruses, being packaged together with a viral genome. The virusoids cannot replicate independently, but require assistance from the virus. The virusoids are sometimes called satellite RNAs.

Viroids and virusoids both replicate via rolling circles (see Figure 13.16). The strand of RNA that is packaged into the virus is called the plus strand. The complementary strand, generated during replication of the RNA, is called the minus strand. Multimers of both plus and minus strands are found. Both types of monomer are generated by cleaving the tail of a rolling circle; circular plus strand monomers are generated by ligating the ends of the linear monomer.

Both plus and minus strands of viroids and virusoids undergo self-cleavage *in vitro*. The cleavage reaction is promoted by divalent metal cations; it generates 5′–OH and 2'-3'-cyclic phosphodiester termini. Some of the RNAs cleave *in vitro* under physiological conditions. Others do so only after a cycle of heating and cooling; this suggests that the isolated RNA has an inappropriate conformation, but can generate an active conformation when it is denatured and renatured.

The viroids and virusoids that undergo self-cleavage form a "hammerhead" secondary structure at the cleavage site, as drawn in the upper part of **Figure 25.13.** The sequence of this structure is sufficient for cleavage. When the surrounding sequences are deleted, the need for a heating-cooling cycle is obviated, and the small RNA self-cleaves spontaneously. This suggests that the sequences beyond the hammerhead usually interfere with its formation.

The active site is a sequence of only 58 nucleotides. The hammerhead contains three stem-loop regions whose position and size are constant, and 13 conserved nucleotides, mostly in the regions connecting the center of the structure. The conserved bases and duplex stems generate an RNA with the intrinsic ability to cleave.

An active hammerhead can also be generated by pairing an RNA representing one side of the structure with an RNA representing the other side. The lower part of Figure 25.13 shows an example of a hammerhead generated by hybridizing a 19 base molecule with a 24 base molecule. The hybrid mimics the hammerhead structure, with the omission of loops I and III. When the 19 base RNA is added to the 24 base RNA, cleavage occurs at the appropriate position in the hammerhead.

We may regard the top (24 base) strand of this hybrid as comprising the "substrate," and the bottom (19 base) strand as comprising the "enzyme." When the 19 base RNA is mixed with an excess of the 24 base RNA, multiple copies of the 24 base RNA are cleaved. This suggests that there is a cycle of 19 base-24 base pairing, cleavage, dissociation of the cleaved fragments from the 19 base RNA, and pairing of the 19 base RNA with a new 24 base substrate. The 19 base RNA is therefore a ribozyme with endonuclease activity. The parameters of the reaction are similar to those of other RNA-catalyzed reactions.

The crystal structure of a hammerhead shows that it forms a compact V-shape, in which the catalytic center lies in a turn, as indicated diagrammatically in **Figure 25.14.** An Mg^{2+} ion located in the catalytic site plays a crucial role in the reaction. It is positioned by the target cytidine and by

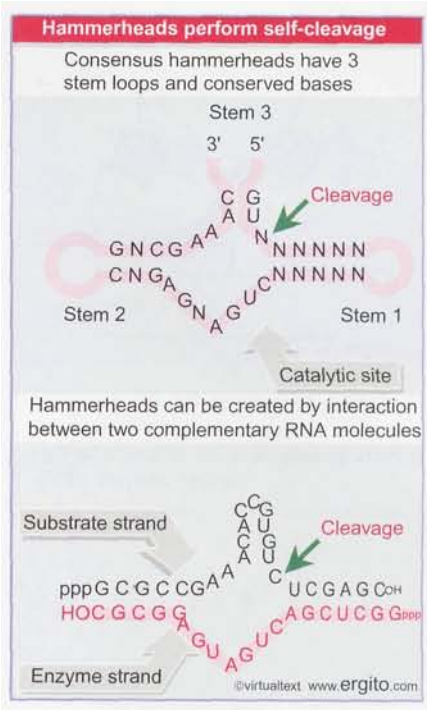


Figure 25.13 Self-cleavage sites of viroids and virusoids have a consensus sequence and form a hammerhead secondary structure by intramolecular pairing. Hammerheads can also be generated by pairing between a substrate strand and an "enzyme" strand.

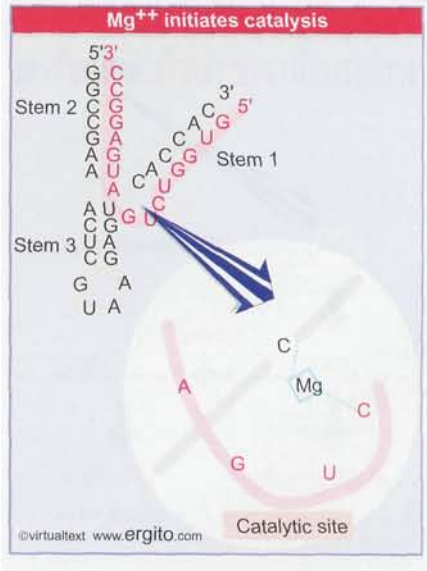


Figure 25.14 A hammerhead ribozyme forms a V-shaped tertiary structure in which stem 2 is stacked upon stem 3. The catalytic center lies between stem 2/3 and stem 1. It contains a magnesium ion that initiates the hydrolytic reaction.

the cytidine at the base of stem 1; it may also be connected to the adjacent uridine. It extracts a proton from the 2'-OH of the target cytidine, and then directly attacks the labile phosphodiester bond. Mutations in the hammerhead sequence that affect the transition state of the cleavage reaction occur in both the active site and other locations, suggesting that there may be a substantial rearrangement of structure prior to cleavage.

It is possible to design enzyme-substrate combinations that can form hammerhead structures, and these have been used to demonstrate that introduction of the appropriate RNA molecules into a cell can allow the enzymatic reaction to occur *in vivo*. A ribozyme designed in this way essentially provides a highly specific restriction-like activity directed against an RNA target. By placing the ribozyme under control of a regulated promoter, it can be used in the same way as (for example) antisense constructs specifically to turn off expression of a target gene under defined circumstances.

25.9 RNA editing occurs at individual bases

Key Concepts

- Apoplipoprotein-B and glutamate receptors have site specific deaminations catalyzed by cytidine and adenosine deaminases that change the coding sequence.

A prime axiom of molecular biology is that the sequence of an mRNA can only represent what is coded in the DNA. The central dogma envisaged a linear relationship in which a continuous sequence of DNA is transcribed into a sequence of mRNA that is in turn directly translated into protein. The occurrence of interrupted genes and the removal of introns by RNA splicing introduces an additional step into the process of gene expression: the coding sequences (exons) in DNA must be reconnected in RNA. But the process remains one of information transfer, in which the actual coding sequence in DNA remains unchanged.

Changes in the information coded by DNA occur in some exceptional circumstances, most notably in the generation of new sequences coding for immunoglobulins in mammals and birds. These changes occur specifically in the somatic cells (B lymphocytes) in which immunoglobulins are synthesized (see *26 Immune diversity*). New information is generated in the DNA of an individual during the process of reconstructing an immunoglobulin gene; and information coded in the DNA is changed by somatic mutation. The information in DNA continues to be faithfully transcribed into RNA.

RNA editing is a process in which *information changes at the level of mRNA*. It is revealed by situations in which the coding sequence in an RNA differs from the sequence of DNA from which it was transcribed. RNA editing occurs in two different situations, with different causes. In mammalian cells, there are cases in which a substitution occurs in an individual base in mRNA, causing a change in the sequence of the protein that is coded. In trypanosome mitochondria, more widespread changes occur in transcripts of several genes, when bases are systematically added or deleted.

Figure 25.15 summarizes the sequences of the apolipoprotein-B gene and mRNA in mammalian intestine and liver. The genome contains a single (interrupted) gene whose sequence is identical in all tissues, with a coding region of 4563 codons. This gene is transcribed into an mRNA that is translated into a protein of 512 kD representing the full coding sequence in the liver.

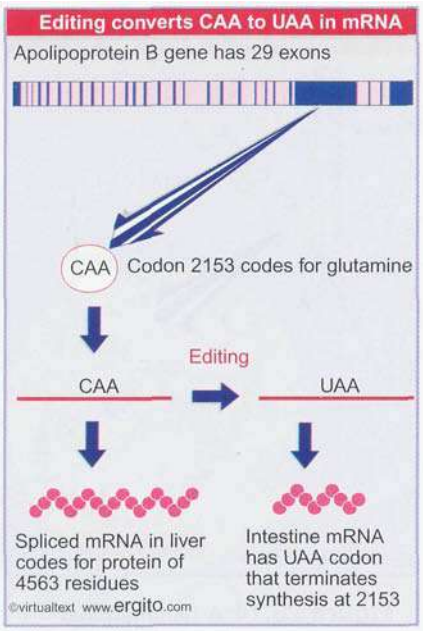


Figure 25.15 The sequence of the apo-B gene is the same in intestine and liver, but the sequence of the mRNA is modified by a base change that creates a termination codon in intestine.

By Book_Crazy [IND]

A shorter form of the protein, ~250 kD, is synthesized in intestine. This protein consists of the N-terminal half of the full-length protein. It is translated from an mRNA whose sequence is identical with that of liver except for a change from C to U at codon 2153. This substitution changes the codon CAA for glutamine into the ochre codon UAA for termination.

What is responsible for this substitution? No alternative gene or exon is available in the genome to code for the new sequence, and no change in the pattern of splicing can be discovered. We are forced to conclude that a change has been made directly in the sequence of the transcript.

Editing of this sort is rare, but apo-lipo-B is not unique. Another example is provided by glutamate receptors in rat brain. Editing at one position changes a glutamine codon in DNA into a codon for arginine in RNA; the change affects the conductivity of the channel and therefore has an important effect on controlling ion flow through the neurotransmitter. At another position in the receptor, an arginine codon is converted to a glycine codon.

The editing event in apo-B causes C_{2153} to be changed to U; both changes in the glutamate receptor are from A to I (inosine). These events are *deaminations* in which the amino group on the nucleotide ring is removed. Such events are catalyzed by enzymes called cytidine and adenosine deaminases, respectively.

What controls the specificity of an editing reaction? Enzymes that undertake deamination as such often have broad specificity—for example, the best characterized adenosine deaminase acts on any A residue in a duplex RNA region. Editing enzymes are related to the general deaminases, but have other regions or additional subunits that control their specificity. In the case of apoB editing, the catalytic subunit of an editing complex is related to bacterial cytidine deaminase, but has an additional RNA-binding region that helps to recognize the specific target site for editing. A special adenosine deaminase enzyme recognizes the target sites in the glutamate receptor RNA, and similar events occur in a serotonin receptor RNA.

The complex may recognize a particular region of secondary structure in a manner analogous to tRNA-modifying enzymes or could directly recognize a nucleotide sequence. The development of an *in vitro* system for the apo-B editing event suggests that a relatively small sequence (~26 bases) surrounding the editing site provides a sufficient target. **Figure 25.16** shows that in the case of the GluR-B RNA, a base-paired region that is necessary for recognition of the target site is formed between the edited region in the exon and a complementary sequence in the downstream intron. A pattern of mispairing within the duplex region is necessary for specific recognition. So different editing systems may have different types of requirement for sequence specificity in their substrates.

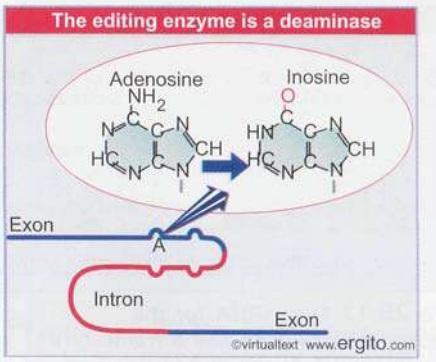


Figure 25.16 Editing of mRNA occurs when a deaminase acts on an adenine in an imperfectly paired RNA duplex region.

25.10 RNA editing can be directed by guide RNAs

Key Concepts

- Extensive RNA editing in trypanosome mitochondria occurs by insertions or deletions of uridine.
- The substrate RNA base pairs with a guide RNA on both sides of the region to be edited.
- The guide RNA provides the template for addition (or less often deletion) of uridines.
- Editing is catalyzed by a complex of endonuclease, terminal uridyltransferase activity, and RNA ligase.

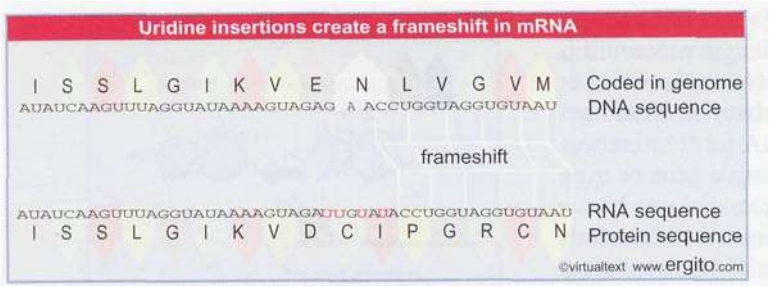


Figure 25.17 The mRNA for the trypanosome *coxII* gene has a frameshift relative to the DNA; the correct reading frame is created by the insertion of 4 uridines.

Another type of editing is revealed by dramatic changes in sequence in the products of several genes of trypanosome mitochondria. In the first case to be discovered, the sequence of the cytochrome oxidase subunit II protein has a frameshift relative to the sequence of the *coxII* gene. The sequences of the gene and protein given in **Figure 25.17** are conserved in several trypanosome species. How does this gene function?

The *coxII* mRNA has an insert of an additional four nucleotides (all uridines) around the site of frameshift. The insertion restores the proper reading frame; it inserts an extra amino acid and changes the amino acids on either side. No second gene with this sequence can be discovered, and we are forced to conclude that the extra bases are inserted during or after transcription. A similar discrepancy between mRNA and genomic sequences is found in genes of the SV5 and measles paramyxoviruses, in these cases involving the addition of G residues in the mRNA.

Similar editing of RNA sequences occurs for other genes and includes deletions as well as additions of uridine. The extraordinary case of the *coxIII* gene of *T. brucei* is summarized in **Figure 25.18.**

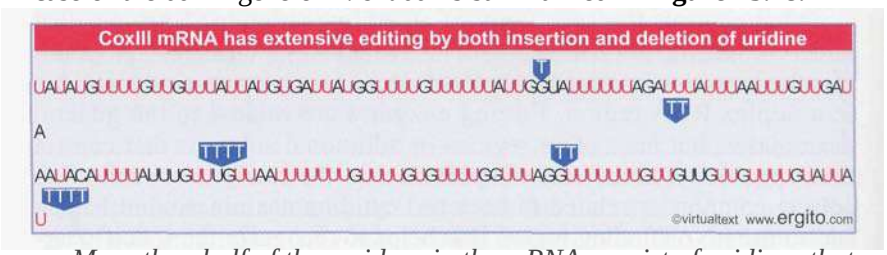


Figure 25.18 Part of the mRNA sequence of *T. brucei coxIII* shows many uridines that are not coded in the DNA (shown in red) or that are removed from the RNA (shown as T).

More than half of the residues in the mRNA consist of uridines that are not coded in the gene. Comparison between the genomic DNA and the mRNA shows that no stretch longer than 7 nucleotides is represented in the mRNA without alteration; and runs of uridine up to 7 bases long are inserted.

What provides the information for the specific insertion of uridines? A guide RNA contains a sequence that is complementary to the correctly edited mRNA. **Figure 25.19** shows a model for its action in the cytochrome *b* gene of *Leishmania.*

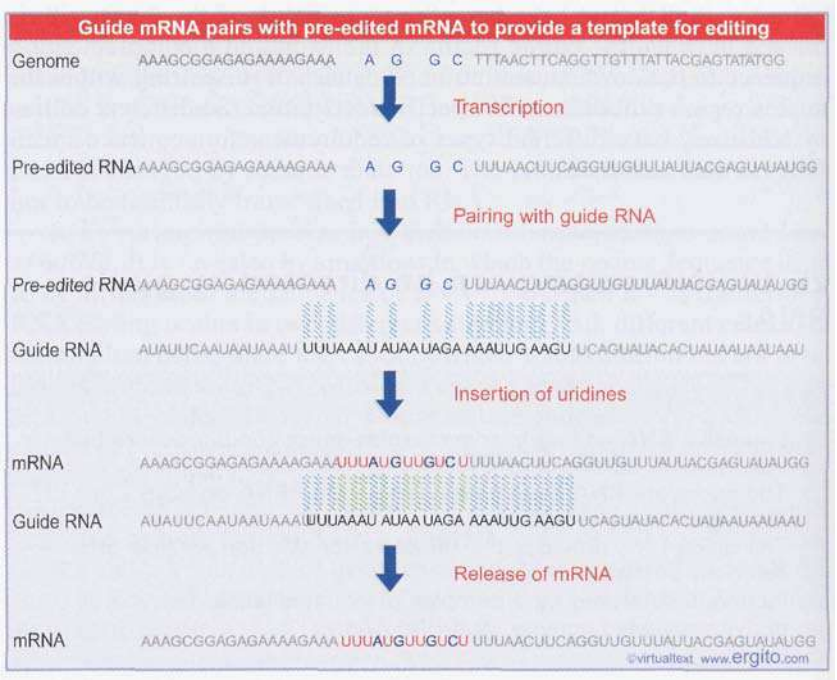


Figure 25.19 Pre-edited RNA base pairs with a guide RNA on both sides of the region to be edited. The guide RNA provides a template for the insertion of uridines. The mRNA produced by the insertions is complementary to the guide RNA.

By Book_Crazy [IND]

The sequence at the top shows the original transcript, or pre-edited RNA. Gaps show where bases will be inserted in the editing process. 8 uridines must be inserted into this region to create the valid mRNA sequence.

The guide RNA is complementary to the mRNA for a significant distance including and surrounding the edited region. Typically the complementarity is more extensive on the 3' side of the edited region and is rather short on the 5' side. Pairing between the guide RNA and the pre-edited RNA leaves gaps where unpaired A residues in the guide RNA do not find complements in the pre-edited RNA. The guide RNA provides a template that allows the missing U residues to be inserted at these positions. When the reaction is completed, the guide RNA separates from the mRNA, which becomes available for translation.

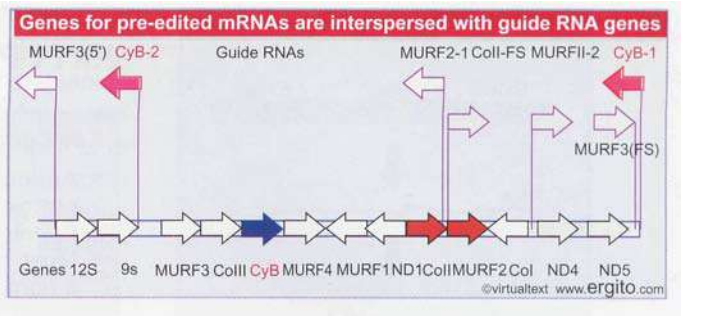


Figure 25.20 The *Leishmania* genome contains genes coding for pre-edited RNAs interspersed with units that code for the guide RNAs required to generate the correct mRNA sequences. Some genes have multiple guide RNAs.

Specification of the final edited sequence can be quite complex; in this example, a lengthy stretch of the transcript is edited by the insertion altogether of 39 U residues, and this appears to require two guide RNAs that act at adjacent sites. The first guide RNA pairs at the 3'-most site, and the edited sequence then becomes a substrate for further editing by the next guide RNA.

The guide RNAs are encoded as independent transcription units. **Figure 25.20** shows a map of the relevant region of the *Leishmania* mitochondrial DNA. It includes the "gene" for cytochrome *b*, which codes for the pre-edited sequence, and two regions that specify guide RNAs. Genes for the major coding regions and for their guide RNAs are interspersed.

In principle, a mutation in either the "gene" or one of its guide RNAs could change the primary sequence of the mRNA, and thus of the protein. By genetic criteria, each of these units could be considered to comprise part of the "gene." Since the units are independently expressed, they should of course complement in *trans*. If mutations were available, we should therefore find that 3 complementation groups were needed to code for the primary sequence of a single protein.

The characterization of intermediates that are partially edited suggests that the reaction proceeds along the pre-edited RNA in the 3′-5′ direction. The guide RNA determines the specificity of uridine insertions by its pairing with the pre-edited RNA.

Editing of uridines is catalyzed by a 20S enzyme complex that contains an endonuclease, a terminal uridyltransferase (TUTase), and an RNA ligase, as illustrated in **Figure 25.21.** It binds the guide RNA and uses it to pair with the pre-edited mRNA. The substrate RNA is cleaved at a site that is (presumably) identified by the absence of pairing with the guide RNA, a uridine is inserted or deleted to base pair with the guide RNA, and then the substrate RNA is ligated. UTP provides the source for the uridyl residue. It is added by the TUTase activity; it is not clear whether this activity, or a separate exonuclease, is responsible for deletion. (At one time it was thought that a stretch of U residues at the end of guide RNA might provide the source for added U residues or a sink for deleted residues, but transfer of U residues to guide RNAs appears to be an aberrant reaction that is not responsible for editing.)

The structures of partially edited molecules suggest that the U residues are added one at a time, and not in groups. It is possible that the reaction proceeds through successive cycles in which U residues are added, tested for complementarity with the guide RNA, retained if acceptable and removed if not, so that the construction of the correct edited sequence occurs gradually. We do not know whether the same types of reaction are involved in editing reactions that add C residues.

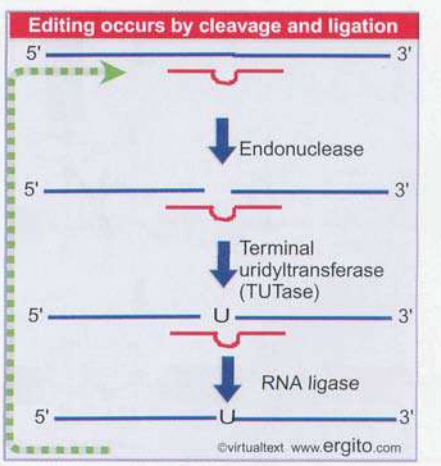


Figure 25.21 Addition or deletion of U residues occurs by cleavage of the RNA, removal or addition of the U, and ligation of the ends. The reactions are catalyzed by a complex of enzymes under the direction of guide RNA.

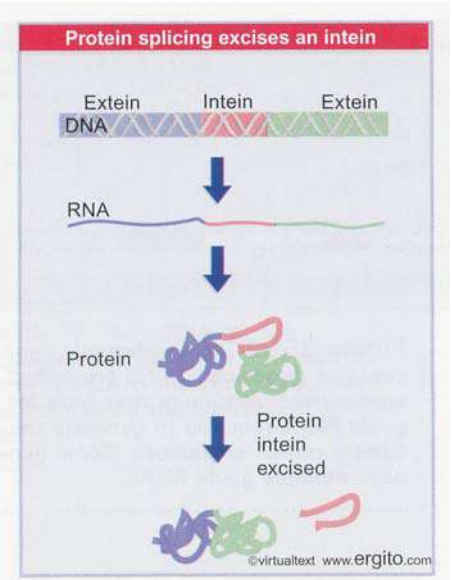


Figure 25.22 In protein splicing the exteins are connected by removing the intein from the protein.

25.11 Protein splicing is autocatalytic

Key Concepts

- An intein has the ability to catalyze its own removal from a protein in such a way that the flanking exteins are connected.
- Protein splicing is catalyzed by the intein.
- Most inteins have two independent activities: protein splicing and a homing endonuclease.

Protein splicing has the same effect as RNA splicing: a sequence that is represented within the gene fails to be represented in the protein. The parts of the protein are named by analogy with RNA splicing: exteins are the sequences that are represented in the mature protein, and inteins are the sequences that are removed. The mechanism of removing the intein is completely different from RNA splicing. **Figure 25.22** shows that the gene is translated into a protein precursor that contains the intein, and then theintein is excised from the protein. About 100 examples of protein splicing are known, spread through all classes of organisms. The typical gene whose product undergoes protein splicing has a single intein.

The first intein was discovered in an archaeal DNA polymerase gene in the form of an intervening sequence in the gene that does not conform to the rules for introns. Then it was demonstrated that the purified protein can splice this sequence out of itself in an autocatalytic reaction. The reaction does not require input of energy and occurs through the series of bond rearrangements shown in **Figure 25.23.** It is a function of the intein, although its efficiency can be influenced by the exteins.

The first reaction is an attack by an -OH or -SH side chain of the first amino acid in the intein on the peptide bond that connects it to the first extein. This transfers the extein from the amino-terminal group of the intein to an N-0 or N-S acyl connection. Then this bond is attacked by the -OH or -SH side chain of the first amino acid in the second extein. The result is to transfer extein 1 to the side chain of the amino-terminal acid of extein2. Finally, the C-terminal asparagine of the intein cyclizes, and the terminal NH of extein2 attacks the acyl bond to replace it with a conventional peptide bond. Each of these reactions can occur spontaneously at very low rates, but their occurrence in a coordinate manner rapidly enough to achieve protein splicing requires catalysis by the intein.

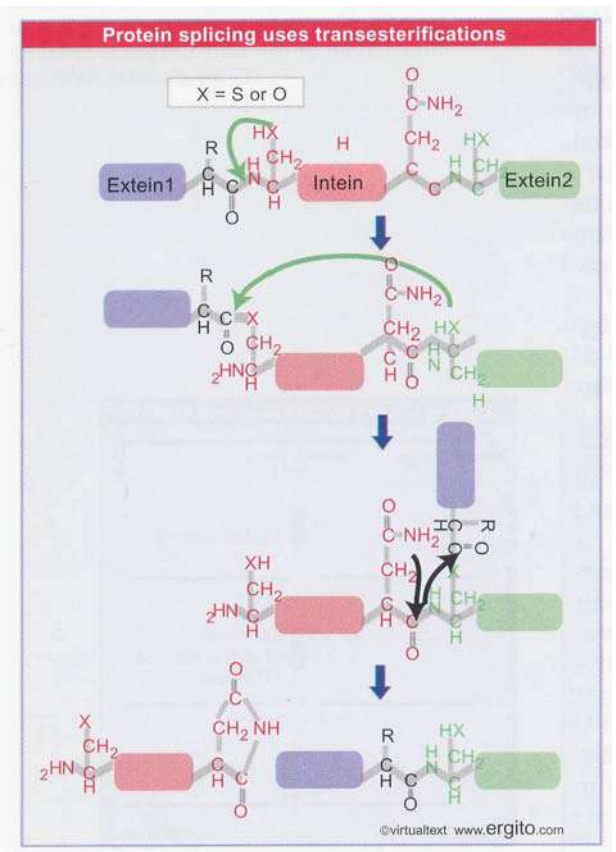


Figure 25.23 Bonds are rearranged through a series of transesterifications involving the -OH groups of serine or proline or the -SH group of cysteine until finally the exteins are connected by a peptide bond and the intein is released with a circularized C-terminus.

Inteins have characteristic features. They are found as in-frame insertions into coding sequences. They can be recognized as such because of the existence of homologous genes that lack the insertion. They have an N-terminal serine or cysteine (to provide the -XH side chain) and a C-terminal asparagine. A typical intein has a sequence of ~ 150 amino acids at the N-terminal end and ~50 amino acids at the C-terminal end that are involved in catalyzing the protein splicing reaction. The sequence in the center of the intein can have other functions.

An extraordinary feature of many inteins is that they have homing endonuclease activity. A homing endonuclease cleaves a target DNA to create a site into which the DNA sequence coding for the intein can be inserted (see Figure 25.11 in *25.5 Some group I introns code for endonucleases that sponsor mobility).* The protein splicing and homing endonuclease activities of an intein are independent.

We do not really understand the connection between the presence of both these activities in an intein, but two types of model have been suggested. One is to suppose that there was originally some sort of connection between the activities, but that they have since become independent and some inteins have lost the homing endonuclease. The other is to suppose that inteins may have originated as protein splicing units, most of which (for unknown reasons) were subsequently invaded by homing endonucleases. This is consistent with the fact that homing endonucleases appear to have invaded other types of units also, including most notably group I introns.

25.12 Summary

Self-splicing is a property of two groups of introns, which are widely dispersed in lower eukaryotes, prokaryotic systems, and mitochondria. The information necessary for the reaction resides in the intron sequence (although the reaction is actually assisted by proteins *in vivo).* For both group I and group II introns, the reaction requires formation of a specific secondary/tertiary structure involving short consensus sequences. Group I intron RNA creates a structure in which the substrate sequence is held by the IGS region of the intron, and other conserved sequences generate a guanine nucleotide binding site. It occurs by a transesterification involving a guanosine residue as cofactor. No input of energy is required. The guanosine breaks the bond at the 5' exon-intron junction and becomes linked to the intron; the hydroxyl at the free end of the exon then attacks the 3' exon-intron junction. The intron cyclizes and loses the guanosine and the terminal 15 bases. A series of related reactions can be catalyzed via attacks by the terminal G-OH residue of the intron on internal phosphodiester bonds. By providing appropriate substrates, it has been possible to engineer ribozymes that perform a variety of catalytic reactions, including nucleotidyl transferase activities.

Some group I and some group II mitochondrial introns have open reading frames. The proteins coded by group I introns are endonucleases that make double-stranded cleavages in target sites in DNA; the cleavage initiates a gene conversion process in which the sequence of the intron itself is copied into the target site. The proteins coded by group II introns include an endonuclease activity that initiates the transposition process, and a reverse transcriptase that enables an RNA copy of the intron to be copied into the target site. These types of introns probably originated by insertion events. The proteins coded by both groups of introns may include maturase activities that assist splicing of the intron by stabilizing the formation of the secondary/tertiary structure of the active site.

Catalytic reactions are undertaken by the RNA component of the RNAase P ribonucleoprotein. Virusoid RNAs can undertake self-cleavage at a "hammerhead" structure. Hammerhead structures can form between a substrate RNA and a ribozyme RNA, allowing cleavage to be directed at highly specific sequences. These reactions support the view that RNA can form specific active sites that have catalytic activity.

RNA editing changes the sequence of an RNA after or during its transcription. The changes are required to create a meaningful coding sequence. Substitutions of individual bases occur in mammalian systems; they take the form of deaminations in which C is converted to U, or A is converted to I. A catalytic subunit related to cytidine or adenosine deaminase functions as part of a larger complex that has specificity for a particular target sequence.

Additions and deletions (most usually of uridine) occur in trypanosome mitochondria and in paramyxoviruses. Extensive editing

reactions occur in trypanosomes in which as many as half of the bases in an mRNA are derived from editing. The editing reaction uses a template consisting of a guide RNA that is complementary to the mRNA sequence. The reaction is catalyzed by an enzyme complex that includes an endonuclease, terminal uridyltransferase, and RNA ligase, using free nucleotides as the source for additions, or releasing cleaved nucleotides following deletion.

Protein splicing is an autocatalytic reaction that occurs by bond transfer reactions and input of energy is not required. The intein catalyzes its own splicing out of the flanking exteins. Many inteins have a homing endonuclease activity that is independent of the protein splicing activity.

References

25.2 Group I introns undertake self-splicing by transesterification

exp Cech, T. (2002). RNA catalysis (www.ergito.com/lookup.jsp?expt=cech)

rev Cech, T. R. (1985). Self-splicing RNA: implications for evolution. Int. Rev. Cytol. 93, 3-22.
Cech, T. R. (1987). The chemistry of self-splicing RNA and RNA enzymes. Science 236, 1532-1539.

ref Been, M. D. and Cech, T. R. (1986). One binding site determines sequence specificity of Tetrahymena pre-rRNA self-splicing, *trans*-splicing, and RNA enzyme activity. Cell 47, 207-216.
Belfort, M., Pedersen-Lane, J., West, D., Ehrenman, K., Maley, G., Chu, F., and Maley, F. (1985). Processing of the intron-containing thymidylate synthase (td) gene of phage T4 is at the RNA level. Cell 41, 375-382.
Cech, T. R. et al. (1981). *In vitro* splicing of the rRNA precursor of *Tetrahymena:* involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell 27, 487-496.
Kruger, K. et al. (1982). Self-splicing RNA: autoexcision and autocyclization of the rRNA intervening sequence of Tetrahymena. Cell 31, 147-157.
Myers, C. A., Kuhla, B., Cusack, S., and Lambowitz, A. M. (2002). tRNA-like recognition of group I introns by a tyrosyl-tRNA synthetase. Proc. Nat. Acad. Sci. USA 99, 2630-2635.

25.3 Group I introns form a characteristic secondary structure

ref Burke, J. M. et al. (1986). Role of conserved sequence elements 9L and 2 in self-splicing of the Tetrahymena ribosomal RNA precursor. Cell 45, 167-176.
Michel, F. and Wetshof, E. (1990). Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Mol. Biol. 216, 585-610.

25.4 Ribozymes have various catalytic activities

rev Cech, T. R. (1990). Self-splicing of group I introns. Ann. Rev. Biochem. 59, 543-568.

25.5 Some group I introns code for endonucleases that sponsor mobility

ref Carignani, G. et al. (1983). An RNA maturase is encoded by the first intron of the mitochondrial gene for the subunit I of cytochrome oxidase in *S. cerevisiae.* Cell 35, 733-742.
Zimmerly, S. et al. (1995). A group II intron is a catalytic component of a DNA endonuclease involved in intron mobility. Cell 83, 529-538.

25.6 Some group II introns code for reverse transcriptases

rev Lambowitz, A. M. and Belfort, M. (1993). Introns as mobile genetic elements. Ann. Rev. Biochem. 62, 587-622.

ref Dickson, L., Huang, H. R., Liu, L., Matsuura, M., Lambowitz, A. M., and Perlman, P. S. (2001). Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. Proc. Nat. Acad. Sci. USA 98, 13207-13212.
Matsuura, M., Noah, J. W., and Lambowitz, A. M. (2001). Mechanism of maturase-promoted group II intron splicing. EMBO J. 20, 7259-7270.
Zimmerly, S. et al. (1995). Group II intron mobility occurs by target DNA-primed reverse transcription. Cell 82, 545-554.

25.8 Viroids have catalytic activity

rev Doherty, E. A. and Doudna, J. A. (2000). Ribozyme structures and mechanisms. Ann. Rev. Biochem. 69, 597-615.
Symons, R. H. (1992). Small catalytic RNAs. Ann. Rev. Biochem. 61, 641-71.

ref Forster, A. C. and Symons, R. H. (1987). Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. Cell 50, 9-16.
Guerrier-Takada, C. et al. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell 35, 849-857.
Scott, W. G., Finch, J. T., and Klug, A. (1995). The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. Cell 81, 991-1002.

25.9 RNA editing occurs at individual bases

ref Higuchi, M. et al. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell 75, 1361-1370.
Navaratnam, N. et al. (1995). Evolutionary origins of apoB mRNA editing: catalysis by a cytidine deaminase that has acquired a novel RNA-binding motif at its active site. Cell 81, 187-195.
Powell, L. M., Wallis, S. C, Pease, R. J., Edwards, Y. H., Knott, T. J., and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell 50, 831-840.
Sommer, B. et al. (1991). RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell 67, 11-19.

25.10 RNA editing can be directed by guide RNAs

exp Benne, R. (2002). RNA editing (www.ergito.com/lookup.jsp?expt=benne)

By Book_Crazy [IND]

ref Aphasizhev, R., Sbicego, S., Peris, M., Jang, S. H., Aphasizheva, I., Simpson, A. M., Rivlin, A., and Simpson, L. (2002). Trypanosome mitochondrial 3' terminal uridylyl transferase (TUTase): the key enzyme in U-insertion/deletion RNA editing. Cell 108, 637-648.

Benne, R., Van den Burg J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., and Tromp, M. C. (1986). Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell 46, 819-826.

Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. Cell 60, 189-198.

Feagin, J. E., Abraham, J. M., and Stuart, K. (1988). Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. Cell 53, 413-422.

Seiwert, S. D., Heidmann, S. and Stuart, K. (1996). Direct visualization of uridylate deletion *in vitro* suggests a mechanism for kinetoplastid editing. Cell 84, 831-841.

25.11 Protein splicing is autocatalytic

rev Paulus, H. (2000). Protein splicing and related forms of protein autoprocessing. Ann. Rev. Biochem. 69, 447-496.

ref Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, J. Z., and Belfort, M. (1997). Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution. Proc. Nat. Acad. Sci. USA 94, 11466-11471.

Perler, F. B. et al. (1992). Intervening sequences in an Archaea DNA polymerase gene. Proc. Nat. Acad. Sci. USA 89, 5577-5581.

Xu, M. Q., Southworth, M. W., Mersha, F. B., Hornstra, L. J., and Perler, F. B. (1993). In vitro protein splicing of purified precursor and the identification of a branched intermediate. Cell 75, 1371-1377.

Chapter 26

Immune diversity

26.1 Introduction

It is an axiom of genetics that the genetic constitution created in the zygote by the combination of sperm and egg is inherited by all somatic cells of the organism. We look to differential control of gene expression, rather than to changes in DNA content, to explain the different phenotypes of particular somatic cells.

Yet there are exceptional situations in which the reorganization of certain DNA sequences is used to regulate gene expression or to create new genes. The immune system provides a striking and extensive case in which the content of the genome changes, when recombination creates active genes in lymphocytes. Other cases are represented by the substitution of one sequence for another to change the mating type of yeast or to generate new surface antigens by trypanosomes (see *18 Rearrangement of DNA).*

The immune response of vertebrates provides a protective system that distinguishes foreign proteins from the proteins of the organism itself. Foreign material (or part of the foreign material) is recognized as comprising an antigen. Usually the antigen is a protein (or protein-attached moiety) that has entered the bloodstream of the animal—for example, the coat protein of an infecting virus. Exposure to an antigen initiates production of an immune response that *specifically recognizes the antigen and destroys it.*

Immune reactions are the responsibility of white blood cells—the B and T lymphocytes, and macrophages. The lymphocytes are named after the tissues that produce them. In mammals, B cells mature in the bone marrow, while T cells mature in the thymus. *Each class of lymphocyte uses the rearrangement of DNA as a mechanism for producing the proteins that enable it to participate in the immune response.*

The immune system has many ways to destroy an antigenic invader, but it is useful to consider them in two general classes. Which type of response the immune system mounts when it encounters a foreign structure depends partly on the nature of the antigen. The response is defined according to whether it is executed principally by B cells or T cells.

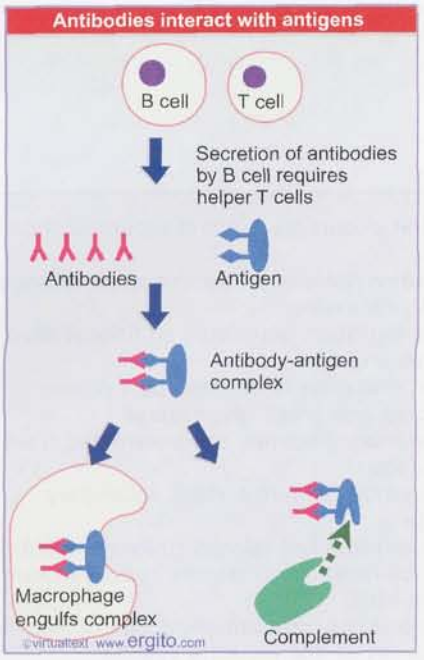


Figure 26.1 Humoral immunity is conferred by the binding of free antibodies to antigens to form antigen-antibody complexes that are removed from the bloodstream by macrophages or that are attacked directly by the complement proteins.

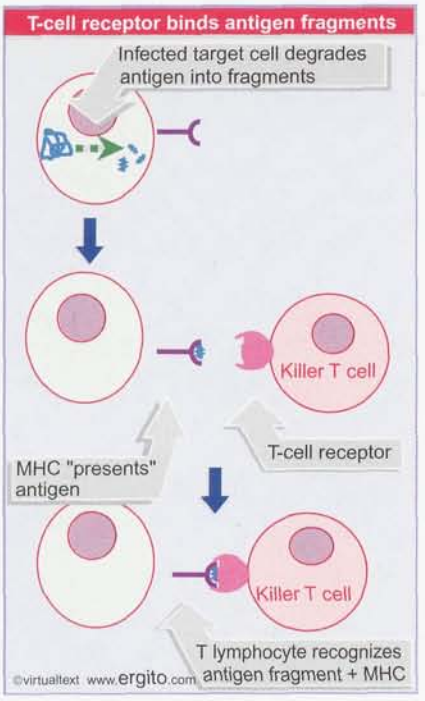


Figure 26.2 In cell-mediated immunity, killer T cells use the T-cell receptor to recognize a fragment of the foreign antigen which is presented on the surface of the target cell by the MHC protein.

The humoral response depends on B cells. It is mediated by the secretion of antibodies, which are immunoglobulin proteins. *Production of an antibody specific for a foreign molecule is the primary event responsible for recognition of an antigen.* Recognition requires the antibody to bind to a small region or structure on the antigen.

The function of antibodies is represented in **Figure 26.1.** Foreign material circulating in the bloodstream, for example, a toxin or pathogenic bacterium, has a surface that presents antigens. The antigen(s) are recognized by the antibodies, which form an antigen-antibody complex. This complex then attracts the attention of other components of the immune system.

The humoral *response depends on* these other *components in two* ways. First, B cells need signals provided by T cells to enable them to secrete antibodies. These T cells are called helper T cells, because they assist the B cells. Second, antigen-antibody formation is a trigger for the antigen to be destroyed. The major pathway is provided by the action of complement, a component whose name reflects its ability to "complement" the action of the antibody itself. Complement consists of a set of ~20 proteins that function through a cascade of proteolytic actions. If the target antigen is part of a cell, for example, an infecting bacterium, the action of complement culminates in lysing the target cell. The action of complement also provides a means of attracting macrophages, which scavenge the target cells or their products. Alternatively, the antigen-antibody complex may be taken up directly by macrophages (scavenger cells) and destroyed.

The cell-mediated response is executed by a class of T lymphocytes called cytotoxic T cells (also called killer T cells). The basic function of the T cell in recognizing a target antigen is indicated in **Figure 26.2.** A cell-mediated response typically is elicited by an intracellular parasite, such as a virus that infects the body's own cells. As a result of the viral infection, fragments of foreign (viral) antigens are displayed on the surface of the cell. These fragments are recognized by the T cell receptor (TCR), which is the T cells' equivalent of the antibody produced by a B cell.

A crucial feature of this recognition reaction is that *the antigen must be presented by a cellular protein that is a member of the* MHC (major histocompatibility complex). The MHC protein has a groove on its surface that binds a peptide fragment derived from the foreign antigen. The combination of peptide fragment and MHC protein is recognized by the T cell receptor. Every individual has a characteristic set of MHC proteins. They are important in graft reactions; a graft of tissue from one individual to another is rejected because of the difference in MHC proteins between the donor and recipient, an issue of major medical importance. The demand that the T lymphocytes recognize both foreign antigen and MHC protein ensures that the cell-mediated response acts only on host cells that have been infected with a foreign antigen. (We discuss the division of MHC proteins into the general types of class I and class II later in *26.20 The major histocompatibility locus codes for many genes of the immune system.)*

The purpose of each type of immune response is to attack a foreign target. Target recognition is the prerogative of B-cell immunoglobulins and T cell receptors. A crucial aspect of their function lies in the ability to distinguish "self" from "nonself." Proteins and cells of the body itself must *never* be attacked. Foreign targets must be *destroyed entirely.* The property of failing to attack "self" is called tolerance. Loss of this ability results in an autoimmune disease, in which the immune system attacks its own body, often with disastrous consequences.

What prevents the lymphocyte pool from responding to "self" proteins? Tolerance probably arises early in lymphocyte cell development when B and T cells that recognize "self" antigens are destroyed. This is

called clonal deletion. In addition to this negative selection, there is also positive selection for T cells carrying certain sets of T cell receptors.

A corollary of tolerance is that it can be difficult to obtain antibodies against proteins that are closely related to those of the organism itself. As a practical matter, therefore, it may be difficult to use (for example) mice or rabbits to obtain antibodies against human proteins that have been highly conserved in mammalian evolution. The tolerance of the mouse or rabbit for its own protein may extend to the human protein in such cases.

Each of the three groups of proteins required for the immune response—immunoglobulins, T cell receptors, MHC proteins—is diverse. Examining a large number of individuals, we find many variants of each protein. Each protein is coded by a large family of genes; and in the case of antibodies and the T cell receptors, the diversity of the population is increased by DNA rearrangements that occur in the relevant lymphocytes.

Immunoglobulins and T cell receptors are direct counterparts, each produced by its own type of lymphocyte. The proteins are related in structure, and their genes are related in organization. The sources of variability are similar. The MHC proteins also share some common features with the antibodies, as do other lymphocyte-specific proteins. In dealing with the genetic organization of the immune system, we are therefore concerned with a series of related gene families, indeed a superfamily that may have evolved from some common ancestor representing a primitive immune response.

26.2 Clonal selection amplifies lymphocytes that respond to individual antigens

Key Concepts

- Each B lymphocyte expresses a single immunoglobulin and each T lymphocyte expresses a single T cell receptor.
- There is a very large variety of immunoglobulins and T cell receptors.
- Antigen binding to an immunoglobulin or T cell receptor triggers clonal multiplication of the cell.

The name of the immune response describes one of its central features. After an organism has been exposed to an antigen, it becomes *immune* to the effects of a new infection. Before exposure to a particular antigen, the organism lacks adequate capacity to deal with any toxic effects. This ability is acquired during the immune response. After the infection has been defeated, the organism retains the ability to respond rapidly in the event of a re-infection.

These features are accommodated by the clonal selection theory illustrated in **Figure 26.3.** The pool of lymphocytes contains B cells and T cells carrying a large variety of immunoglobulins or T cell receptors. *But any individual B lymphocyte produces one immunoglobulin, which is capable of recognizing only a single antigen; similarly, any individual T lymphocyte produces only one particular T cell receptor.*

In the pool of immature lymphocytes, the unstimulated B cells and T cells are morphologically indistinguishable. But on exposure to antigen, a B cell whose antibody is able to bind the antigen, or a T cell whose receptor can recognize it, is stimulated to divide, probably by some feedback from the surface of the cell, where the antibody/receptor-antigen reaction occurs. The stimulated cells then develop into mature B or T lymphocytes, which includes morphological changes involving (for example) an increase in cell size (especially pronounced for B cells).

The initial expansion of a specific B- or T cell population upon first exposure to an antigen is called the primary immune response. Large

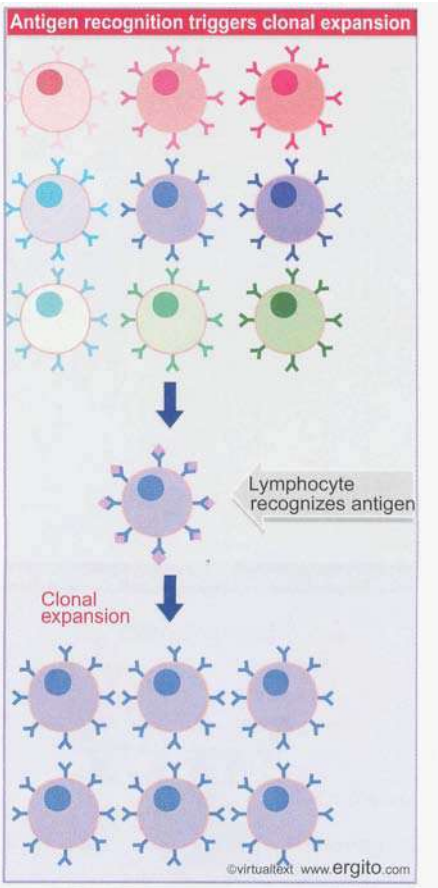


Figure 26.3 The pool of immature lymphocytes contains B cells and T cells making antibodies and receptors with a variety of specificities. Reaction with an antigen leads to clonal expansion of the lymphocyte with the antibody (B cell) or receptor (T cell) that can recognize the antigen.

numbers of B or T lymphocytes with specificity for the target antigen are produced. Each population represents a clone of the original responding cell. Antibody is secreted from the B cells in large quantities, and it may even come to dominate the antibody population.

After a successful primary immune response has been mounted, the organism retains B cells and T cells carrying the corresponding antibody or receptor. These memory cells represent an intermediate state between the immature cell and the mature cell. They have not acquired all of the features of the mature cell, but they are long-lived, and can rapidly be converted to mature cells. Their presence allows a secondary immune response to be mounted rapidly if the animal is exposed to the same antigen again.

The pool of immature lymphocytes in a mammal contains ~10^{12} cells. This pool contains some lymphocytes that have unique specificities (because a corresponding antigen has never been encountered), while others are represented by up to 10^6 cells (because clonal selection has expanded the pool to respond to an antigen).

What features are recognized in an antigen? Antigens are usually macromolecular. Although small molecules may have antigenic determinants and can be recognized by antibodies, usually they are not effective in provoking an immune response (because of their small size). But they do provoke a response when conjugated with a larger carrier molecule (usually a protein). A small molecule that is used to provoke a response by such means is called a hapten.

Only a small part of the surface of a macromolecular antigen is actually recognized by any one antibody. The binding site consists of only 5-6 amino acids. Of course, any particular protein may have more than one such binding site, in which case it provokes antibodies with specificities for different regions. The region provoking a response is called an antigenic determinant or epitope. When an antigen contains several epitopes, some may be more effective than others in provoking the immune response; in fact, they may be so effective that they entirely dominate the response.

How do lymphocytes find target antigens and where does their maturation take place? Lymphocytes are peripatetic cells. They develop from immature stem cells that are located in the adult bone marrow. They migrate to the peripheral lymphoid tissues (spleen, lymph nodes) either directly via the bloodstream (if they are B cells) or via the thymus (where they become T cells). The lymphocytes recirculate between blood and lymph; the process of dispersion ensures that an antigen will be exposed to lymphocytes of all possible specificities. When a lymphocyte encounters an antigen that binds its antibody or receptor, clonal expansion begins the immune response.

26.3 Immunoglobulin genes are assembled from their parts in lymphocytes

Key Concepts

- An immunoglobulin is a tetramer of two light chains and two heavy chains.
- Light chains fall into the lambda and kappa families; heavy chains form a single family.
- Each chain has an N-terminal variable region (V) and a C-terminal constant region (C).
- The V domain recognizes antigen and the C domain provides the effector response.
- V domains and C domains are separately coded by V gene segments and C gene segments.
- A gene coding for an intact immunoglobulin is generated by somatic recombination to join a V gene segment with a C gene segment.

A remarkable feature of the immune response is an animal's ability to produce an appropriate antibody whenever it is exposed to a new antigen. How can the organism be prepared to produce antibody proteins each designed specifically to recognize an antigen whose structure cannot be anticipated?

For practical purposes, we usually reckon that a mammal has the ability to produce 10^6-10^8 different antibodies. Each antibody is an immunoglobulin tetramer consisting of two identical light chains (L) and two identical heavy chains (H). If any light chain can associate with any heavy chain, to produce 10^6–10^8 potential antibodies requires 10 –10^4 different light chains and 10^3–10^4 different heavy chains.

There are 2 types of light chain and ~10 types of heavy chain. Different classes of immunoglobulins have different effector functions. The class is determined by the heavy chain constant region, which exercises the effector function (see Figure 26.16).

The structure of the immunoglobulin tetramer is illustrated in **Figure 26.4.** Light chains and heavy chains share the same general type of organization in which each protein chain consists of two principal regions: the N-terminal variable region (V region); and the C-terminal constant region (C region). They were defined originally by comparing the amino acid sequences of different immunoglobulin chains. As the names suggest, the variable regions show considerable changes in sequence from one protein to the next, while the constant regions show substantial homology.

Corresponding regions of the light and heavy chains associate to generate distinct domains in the immunoglobulin protein.

The variable (V) domain is generated by association between the variable regions of the light chain and heavy chain. *The V domain is responsible for recognizing the antigen.* An immunoglobulin has a Y-shaped structure in which the arms of the Y are identical, and each arm has a copy of the V domain. Production of V domains of different specificities creates the ability to respond to diverse antigens. The total number of variable regions for either light- or heavy-chain proteins is measured in hundreds. *So the protein displays the maximum versatility in the region responsible for binding the antigen.*

The number of constant regions is vastly smaller than the number of variable regions—typically there are only 1 -10 C regions for any particular type of chain. The constant regions in the subunits of the immunoglobulin tetramer associate to generate several individual C domains. The first domain results from association of the single constant region of the light chain (C_L) with the C_{H1} part of the heavy-chain constant region. The two copies of this domain complete the arms of the Y-shaped molecule. Association between the C regions of the heavy chains generates the remaining C domains, which vary in number depending on the type of heavy chain.

Comparing the characteristics of the variable and constant regions, we see the central dilemma in immunoglobulin gene structure. How does the genome code for a set of proteins in which any individual polypeptide chain must have one of <10 possible C regions, but can have any one of several hundred possible V regions? It turns out that the number of coding sequences for each type of region reflects its variability. There are many genes coding for V regions, but only a few genes coding for C regions.

In this context, *"gene" means a sequence of DNA coding for a discrete part of the final immunoglobulin polypeptide* (heavy or light chain). So V genes code for variable regions and C genes code for constant regions, although *neither type of gene is expressed as an independent unit.* To construct a unit that can be expressed in the form of an authentic light or heavy chain, a V gene must be joined physically to a C gene. In this system, two "genes" code for one polypeptide. To avoid confusion, we will refer to these units as "gene segments" rather than "genes."

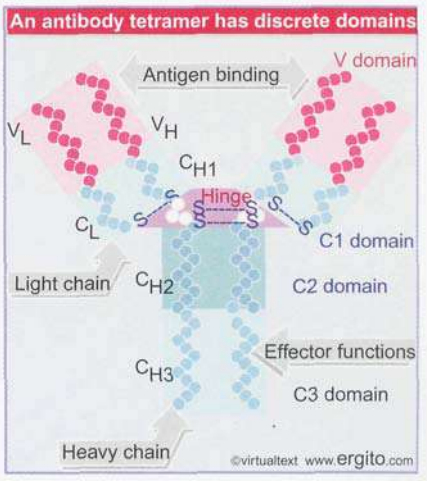


Figure 26.4 Heavy and light chains combine to generate an immunoglobulin with several discrete domains.

The sequences coding for light chains and heavy chains are assembled in the same way: *any one of many V gene segments may be joined to any one of a few C gene segments.* This somatic recombination occurs *in the B lymphocyte in which the antibody is expressed.* The large number of available V gene segments is responsible for a major part of the diversity of immunoglobulins. However, not all diversity is coded in the genome; some is generated by changes that occur during the process of constructing a functional gene.

Essentially the same description applies to the formation of functional genes coding for the protein chains of the T cell receptor. Two types of receptor are found on T cells, one consisting of two types of chain called a and β, the other consisting of 7 and 8 chains. Like the genes coding for immunoglobulins, the genes coding for the individual chains in T cell receptors consist of separate parts, including V and C regions, that are brought together in an active T cell.

The crucial fact about the synthesis of immunoglobulins, therefore, is that *the arrangement of V gene segments and C gene segments is different in the cells producing the immunoglobulins (or T cell receptors) from all other somatic cells or germ cells.*

The construction of a functional immunoglobulin or T cell receptor gene might seem to be a Lamarckian process, representing a change in the genome that responds to a particular feature of the phenotype (the antigen). At birth, the organism does not possess the functional gene for producing a particular antibody or T cell receptor. It possesses a large number of V gene segments and a smaller number of C gene segments. The subsequent construction of an active gene from these parts allows the antibody/receptor to be synthesized so that it is available to react with the antigen. The clonal selection theory requires that this rearrangement of DNA occurs *before the exposure to antigen,* which then results in *selection* for those cells carrying a protein able to bind the antigen. The entire process occurs in somatic cells and does not affect the germline; so the response to an antigen is not inherited by progeny of the organism.

There are two families of immunoglobulin light chains, K and *X,* and one family containing all the types of heavy chain (H). Each family resides on a different chromosome and consists of its own set of both V gene segments and C gene segments. This is called the *germline pattern,* and is found in the germline and in somatic cells of all lineages other than the immune system.

But in a cell expressing an antibody, each of its chains—one light type (either K or *X)* and one heavy type—is coded by a single intact gene. The recombination event that brings a V gene segment to partner a C gene segment creates an active gene consisting of exons that correspond precisely with the functional domains of the protein. The introns are removed in the usual way by RNA splicing.

Recombination between V and C gene segments to give functional loci occurs in a population of immature lymphocytes. A B lymphocyte usually has only one productive rearrangement of light-chain gene segments (either K or *X)* and one of heavy-chain gene segments. Similarly, a T lymphocyte productively rearranges an a gene and a β gene, or one 8 gene and one γ gene. The antibody or T cell receptor produced by any one cell is determined by the particular configuration of V gene segments and C gene segments that has been joined.

The principles by which functional genes are assembled are the same in each family, but there are differences in the details of the organization of the V and C gene segments, and correspondingly of the recombination reaction between them. In addition to the V and C gene segments, other short DNA sequences (including J segments and D segments) are included in the functional somatic loci.

26.4 Light chains are assembled by a single recombination

Key Concepts

- A lambda light chain is assembled by a single recombination between a V gene segment and a J-C gene segment.
- The V gene segment has a leader exon, intron, and variable-coding region.
- The J-C gene segment has a short J-coding exon, intron, and C-coding region.
- A kappa light chain is assembled by a single recombination between a V gene segment and one of five J segments preceding the C gene.

A λ light chain is assembled from two parts, as illustrated in **Figure 26.5.** The V gene segment consists of the leader exon (L) separated by a single intron from the variable (V) segment. The C gene segment consists of the J segment separated by a single intron from the constant (C) exon.

The name of the J segment is an abbreviation for joining, since it identifies the region to which the V segment becomes connected. So the joining reaction does not directly involve V and C gene segments, but occurs via the J segment; when we discuss the joining of "V and C gene segments" for light chains, we really mean V-JC joining.

The J segment is short and codes for the last few (13) amino acids of the variable region, as defined by amino acid sequences. In the intact gene generated by recombination, the V-J segment constitutes a single exon coding for the entire variable region.

The consequences of the K joining reaction are illustrated in **Figure 26.6.** A K light chain also is assembled from two parts, but there is a difference in the organization of the C gene segment. A group of five J segments is spread over a region of 500-700 bp, separated by an intron of 2-3 kb from the C_κ exon. In the mouse, the central J segment is nonfunctional (ψJ3). A V_κ segment may be joined to any one of the J segments.

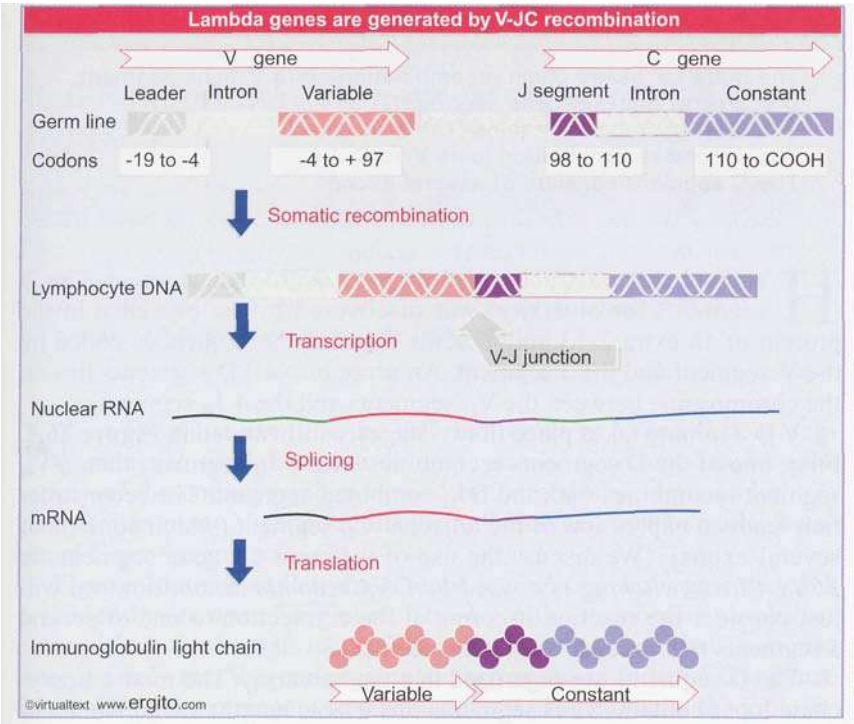


Figure 26.5 The lambda C gene segment is preceded by a J segment, so that V-J recombination generates a functional lambda light-chain gene.

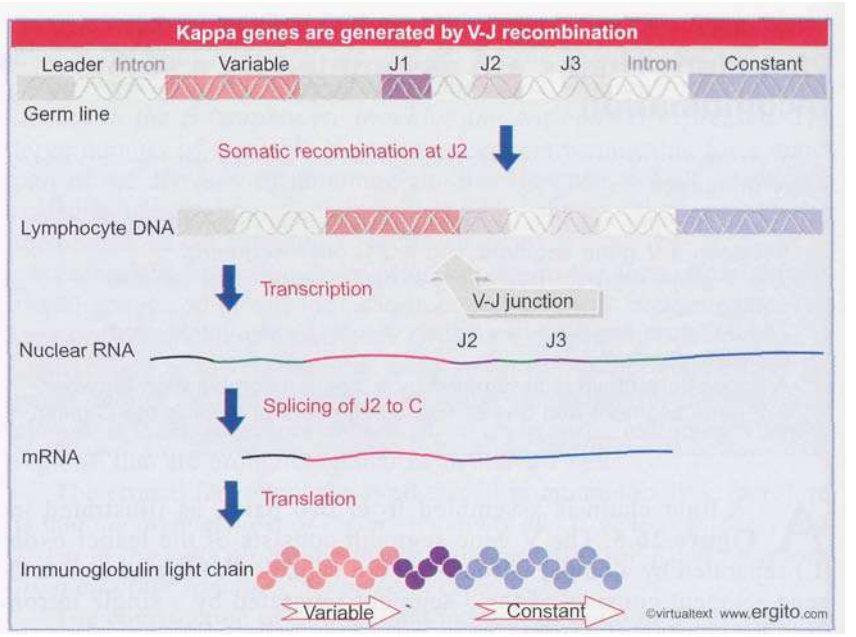


Figure 26.6 The kappa C gene segment is preceded by multiple J segments in the germ line. V-J joining may recognize any one of the J segments, which is then spliced to the C gene segment during RNA processing.

Whichever J segment is used becomes the terminal part of the intact variable exon. Any J segments on the left of the recombining J segment are lost (J1 has been lost in the figure). Any J segment on the right of the recombining J segment is treated as part of the intron between the variable and constant exons (J3 is included in the intron that is spliced out in the figure).

All functional J segments possess a signal at the left boundary that makes it possible to recombine with the V segment; and they possess a signal at the right boundary that can be used for splicing to the C exon. Whichever J segment is recognized in DNA joining uses its splicing signal in RNA processing.

26.5 Heavy chains are assembled by two recombinations

Key Concepts

- The units for heavy chain recombination are a V gene segment, D segment, and J-C gene segment.
- The first recombination joins D to J-C.
- The second recombination joins V to D-J-C.
- The C segment consists of several exons.

Heavy chain construction involves an additional segment. The D segment (for diversity) was discovered by the presence in the protein of an extra 2-13 amino acids between the sequences coded by the V segment and the J segment. An array of >10 D segments lies on the chromosome between the V_H segments and the 4 J_H segments.

V-D-J joining takes place in two stages, as illustrated in **Figure 26.7.** First, one of the D segments recombines with a J_H segment; then a V_H segment recombines with the DJ_H combined segment. The reconstruction leads to expression of the adjacent C_H segment (which consists of several exons). (We discuss the use of different C_H gene segments in *26.11 Class switching is caused by DNA recombination;* now we will just consider the reaction in terms of the connection to one of several J segments that precede a C_H gene segment.)

The D segments are organized in a tandem array. The mouse heavy-chain locus contains 12 D segments of variable length; the human locus

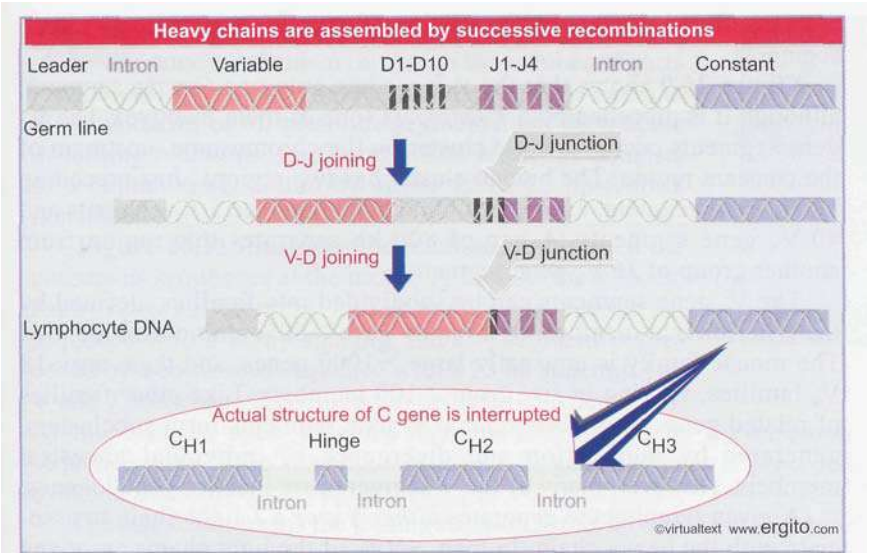


Figure 26.7 Heavy genes are assembled by sequential joining reactions. First a D segment is joined to a J segment; then a V gene segment is joined to the D segment.

has ~30 D segments (not all necessarily active). Some unknown mechanism must ensure that the *same* D segment is involved in the D-J joining and V-D joining reactions. (When we discuss joining of V and C gene segments for heavy chains, we assume the process has been completed by V-D and D-J joining reactions.)

The V gene segments of all three immunoglobulin families are similar in organization. The first exon codes for the signal sequence (involved in membrane attachment), and the second exon codes for the major part of the variable region itself (<100 codons long). The remainder of the variable region is provided by the D segment (in the H family only) and by a J segment (in all three families).

The structure of the constant region depends on the type of chain. For both K and λ light chains, the constant region is coded by a single exon (which becomes the third exon of the reconstructed, active gene). For H chains, the constant region is coded by several exons; corresponding with the protein chain shown in Figure 26.4, separate exons code for the regions C_{H1}, hinge, C_{H2}, and C_{H3}. Each C_H exon is ~100 codons long; the hinge is shorter. The introns usually are relatively small (~300 bp).

26.6 Recombination generates extensive diversity

Key Concepts

- A light chain locus can produce >1000 chains by combining 300 V genes with 4-5 C genes.
- An H locus can produce >4000 chains by combining 300 V genes, 20 D segments, and 4 J segments.

Now we must examine the different types of V and C gene segments to see how much diversity can be accommodated by the variety of the coding regions carried in the germline. In each light lg gene family, many V gene segments are linked to a much smaller number of C gene segments.

Figure 26.8 shows that the λ locus has ~6 C gene segments, each preceded by its own J segment. The λ locus in mouse is much less diverse than the human locus. The main difference is that in mouse there are only two V_λ gene segments; each is linked to two J-C regions. Of the 4 C_λ gene segments, one is inactive. At some time in the past, the

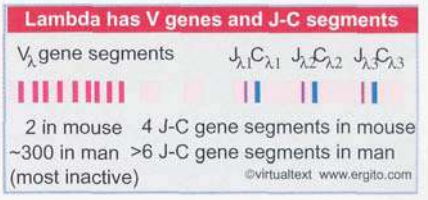


Figure 26.8 The lambda family consists of V gene segments linked to a small number of J-C gene segments.

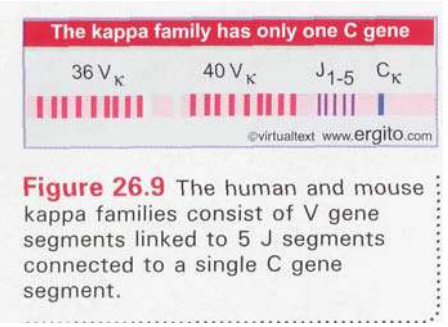


Figure 26.9 The human and mouse kappa families consist of V gene segments linked to 5 J segments connected to a single C gene segment.

mouse suffered a catastrophic deletion of most of its germline V_λ gene segments.

Figure 26.9 shows that the K locus has only one C gene segment, although it is preceded by 5 J segments (one of them inactive). The V_κ gene segments occupy a large cluster on the chromosome, upstream of the constant region. The human cluster has two regions. Just preceding the C_κ gene segment, a region of 600 kb contains the 5 J_κ segments and 40 V_κ gene segments. A gap of 800 kb separates this region from another group of 36 V_κ gene segments.

The V_κ gene segments can be subdivided into families, defined by the criterion that members of a family have >80% amino acid identity. The mouse family is unusually large, ~1000 genes, and there are ~18 V_κ families, varying in size from 2-100 members. Like other families of related genes, therefore, related V gene segments form subclusters, generated by duplication and divergence of individual ancestral members. However, many of the V segments are inactive pseudogenes.

A given lymphocyte generates *either* a K *or* a λ light chain to associate with the heavy chain. In man, ~60% of the light chains are K and ~40% are *X*. In mouse, 95% of B cells express the K type of light chain, presumably because of the reduced number of *X* gene segments.

The single locus for heavy chain production in Man consists of several discrete sections, as summarized in **Figure 26.10.** It is similar in the mouse, where there are more V_H gene segments, fewer D and J segments, and a slight difference in the number and organization of C gene segments. The 3' member of the V_H cluster is separated by only 20 kb from the first D segment. The D segments are spread over ~50 kb, and then comes the cluster of J segments. Over the next 220 kb lie all the C_H gene segments. There are 9 functional C_H gene segments and 2 pseudogenes. The organization suggests that a γ gene segment must have been duplicated to give the subcluster of γ-γ-ε-α, after which the entire group was then duplicated.

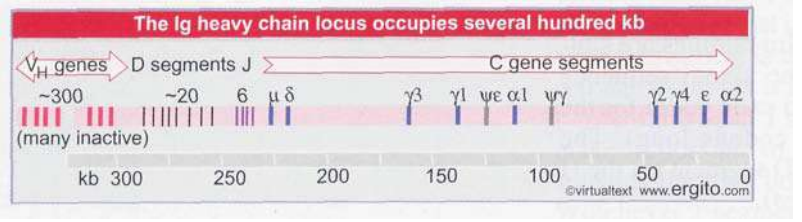


Figure 26.10 A single gene cluster in man contains all the information for heavy-chain gene assembly.

How far is the diversity of germline information responsible for V region diversity in immunoglobulin proteins? By combining any one of ~50 V gene segments with any one of 4-5 J segments, a typical light chain locus has the potential to produce some 250 chains. There is even greater diversity in the H chain locus; by combining any one of ~50 V_H gene segments, 20 D segments, and 4 J segments, the genome potentially can produce 4000 variable regions to accompany any C_H gene segment. In mammals, this is the starting point for diversity, but additional mechanisms introduce further changes. *When closely related variants of immunoglobulins are examined, there often are more proteins than can be accounted for by the number of corresponding V gene segments.* The new members are created by somatic changes in individual genes during or after the recombination process (see *26.14 Somatic mutation generates additional diversity in mouse and man).*

26.7 Immune recombination uses two types of consensus sequence

• ***Key Concepts***

The consensus sequence used for recombination is a heptamer separated by either 12 or 23 base pairs from a nonamer. Recombination occurs between two consensus sequences that have different spacings.

Assembly of light- and heavy-chain genes involves the same mechanism (although the number of parts is different). The same consensus sequences are found at the boundaries of all germline segments that participate in joining reactions. Each consensus sequence consists of a heptamer separated by either 12 or 23 bp from a nonamer.

Figure 26.11 illustrates the relationship between the consensus sequences at the mouse Ig loci. At the K locus, each V_κ gene segment is followed by a consensus sequence with a 12 bp spacing. Each J_κ segment is preceded by a consensus sequence with a 23 bp spacing. The V and J consensus sequences are inverted in orientation. At the λ locus, each V_λ gene segment is followed by a consensus sequence with 23 bp spacing, while each J_λ gene segment is preceded by a consensus of the 12 bp spacer type.

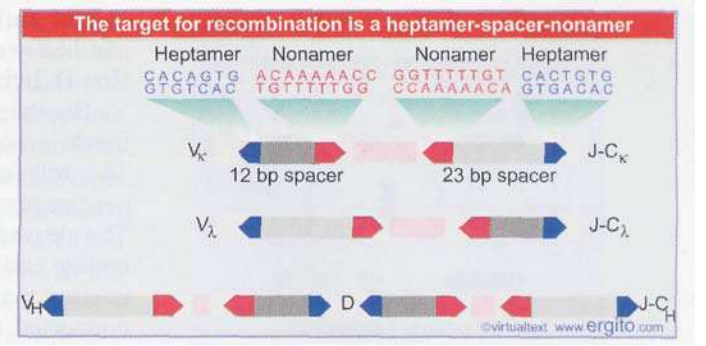


Figure 26.11 Consensus sequences are present in inverted orientation at each pair of recombining sites. One member of each pair has a spacing of 1 2 bp between its components; the other has 23 bp spacing.

The rule that governs the joining reaction is that *a consensus sequence with one type of spacing can be joined only to a consensus sequence with the other type of spacing.* Since the consensus sequences at V and J segments can lie in either order, the different spacings do not impart any directional information, but serve to prevent one V gene segment from recombining with another, or one J segment from recombining with another.

This concept is borne out by the structure of the components of the heavy gene segments. Each V_H gene segment is followed by a consensus sequence of the 23 bp spacer type. The D segments are flanked on either side by consensus sequences of the 12 bp spacer type. The J_H segments are preceded by consensus sequences of the 23 bp spacer type. So the V gene segment must be joined to a D segment; and the D segment must be joined to a J segment. A V gene segment cannot be joined directly to a J segment, because both possess the same type of consensus sequence.

The spacing between the components of the consensus sequences corresponds almost to one or two turns of the double helix. This may reflect a geometric relationship in the recombination reaction. For example, the recombination protein(s) may approach the DNA from one side, in the same way that RNA polymerase and repressors approach recognition elements such as promoters and operators.

26.8 Recombination generates deletions or inversions

Key Concepts

- Recombination occurs by double-strand breaks at the heptamers of two consensus sequences.
- The signal ends of the fragment between the breaks usually join to generate an excised circular fragment.
- The coding ends are covalently linked to join V to J-C (L chain) or D to J-C and V to D-J-C (H chain).
- If the recombining genes are in inverted instead of direct orientation, there is an inversion instead of deletion of an excised circle.

Recombination of the components of immunoglobulin genes is accomplished by a physical rearrangement of sequences, involving breakage and reunion, but the mechanism is different from homologous recombination. The general nature of the reaction is illustrated in

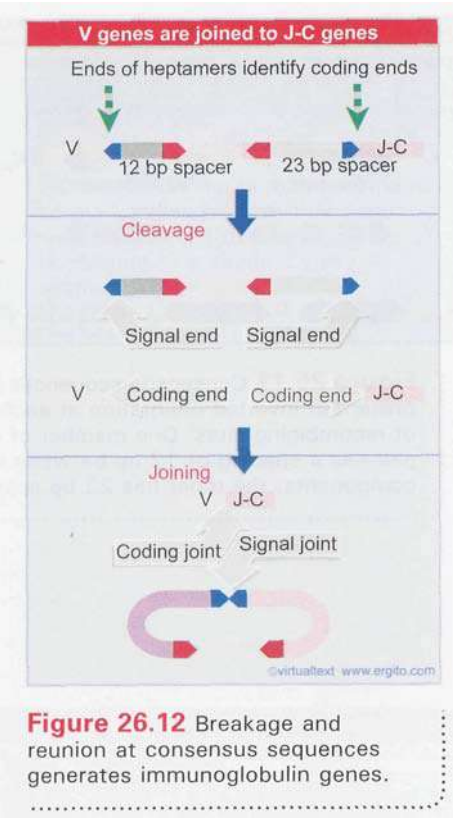


Figure 26.12 Breakage and reunion at consensus sequences generates immunoglobulin genes.

Figure 26.12 for the example of a K light chain. (The reaction is similar at a heavy chain locus, except that there are two recombination events: first D-J, then V-DJ.)

Breakage and reunion occur as separate reactions. A double-strand break is made at the heptamers that lie at the ends of the coding units. This releases the entire fragment between the V gene segment and J-C gene segment; the cleaved termini of this fragment are called signal ends. The cleaved termini of the V and J-C loci are called coding ends. The two coding ends are covalently linked to form a coding joint; this is the connection that links the V and J segments. If the two signal ends are also connected, the excised fragment would form a circular molecule.

We have shown the V and J-C loci as organized in the same orientation. As a result, the cleavage at each consensus sequence releases the region between them as a linear fragment. If the signal ends are joined, it is converted into a circular molecule, as indicated in Figure 26.12. Deletion to release an excised circle is the predominant mode of recombination at the immunoglobulin and TCR loci.

In some exceptional cases, the V gene segment is inverted in orientation on the chromosome relative to the J-C loci. In such a case, breakage and reunion inverts the intervening material instead of deleting it. The outcomes of deletion versus inversion are the same as shown previously for homologous recombination between direct or inverted repeats in Figure 16.9 and Figure 16.10. There is one further proviso, however; recombination with an inverted V gene segment makes it *necessary* for the signal ends to be joined, because otherwise there is a break in the locus. Inversion occurs in TCR recombination, and also sometimes in the K light chain locus.

26.9 The RAG proteins catalyze breakage and reunion

Key Concepts

- The RAG proteins are necessary and sufficient for the cleavage reaction.
- RAG1 recognizes the nonamer consensus sequences for recombination. RAG2 binds to RAG1 and cleaves at the heptamer.
- The reaction resembles the topoisomerase-like resolution reaction that occurs in transposition.
- It proceeds through a hairpin intermediate at the coding end; opening of the hairpin is responsible for insertion of extra bases (P nucleotides) in the recombined gene.
- Deoxynucleoside transferase inserts additional N nucleotides at the coding end.
- The codon at the site of the V-(D)J joining reaction has an extremely variable sequence and codes for amino acid 96 in the antigen-binding site.
- The double-strand breaks at the coding joints are repaired by the same system involved in nonhomologous end-joining of damaged DNA.
- An enhancer in the C gene activates the promoter of the V gene after recombination has generated the intact immunoglobulin gene.

The proteins RAG1 and RAG2 are necessary and sufficient to cleave DNA for V(D)J recombination. They are coded by two genes, separated by <10 kb on the chromosome, whose transfection into fibroblasts causes a suitable substrate DNA to undergo the V(D)J joining reaction. Mice that lack either RAG1 or RAG2 are unable to recombine their immunoglobulins or T cell receptors, and as a result

have immature B and T lymphocytes. The RAG proteins together undertake the catalytic reactions of cleaving and rejoining DNA, and also provide a structural framework within which the reactions occur.

RAG1 recognizes the heptamer/nonamer signals with the appropriate 12/23 spacing and recruits RAG2 to the complex. The nonamer provides the site for initial recognition, and the heptamer directs the site of cleavage.

The reactions involved in recombination are shown in **Figure 26.13.** The complex nicks one strand at each junction. The nick has 3′–OH and 5'-P ends. The free 3′–OH end then attacks the phosphate bond at the corresponding position *in the other strand of the duplex.* This creates a hairpin at the coding end, in which the 3′ end of one strand is covalently linked to the 5' end of the other strand; it leaves a blunt double-strand break at the signal end.

This second cleavage is a transesterification reaction in which bond energies are conserved. It resembles the topoisomerase-like reactions catalyzed by the resolvase proteins of bacterial transposons (see *16.9 TnA transposition requires transposase and resolvase).* The parallel with these reactions is supported further by a homology between RAG1 and bacterial invertase proteins (which invert specific segments of DNA by similar recombination reactions). In fact, the RAG proteins can insert a donor DNA whose free ends consist of the appropriate signal sequences (heptamer–12/23– spacer nonamer) into an unrelated target DNA in an *in vitro* transposition reaction. This suggests that somatic recombination of immune genes evolved from an ancestral transposon. It also suggests that the RAG proteins are responsible for chromosomal translocations in which Ig or TCR loci are connected to other loci (see *30.12 Proto-oncogenes can be activated by translocation).*

The hairpins at the coding ends provide the substrate for the next stage of reaction. If a single-strand break is introduced into one strand close to the

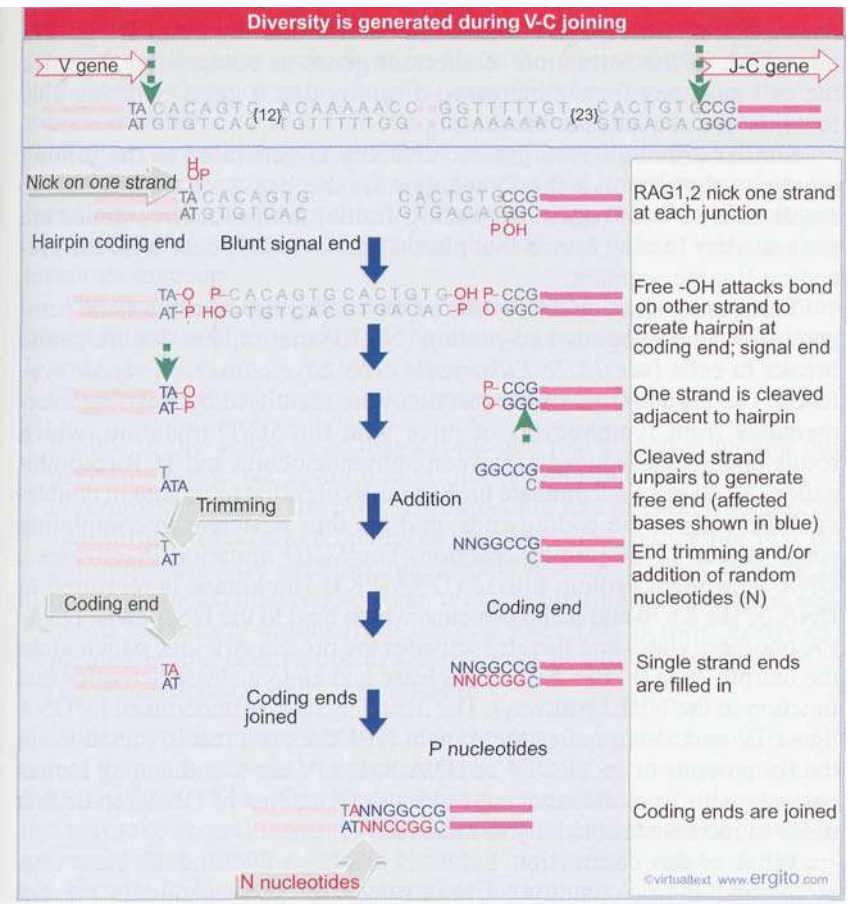


Figure 26.13 Processing of coding ends introduces variability at the junction.

hairpin, an impairing reaction at the end generates a single-stranded protrusion. Synthesis of a complement to the exposed single strand then converts the coding end to an extended duplex. This reaction explains the introduction of P nucleotides at coding ends; they consist of a few extra base pairs, related to, but reversed in orientation from, the original coding end.

Some extra bases also may be inserted, apparently with random sequences, between the coding ends. They are called N nucleotides. Their insertion occurs via the activity of the enzyme deoxynucleoside transferase (known to be an active component of lymphocytes) at a free 3' coding end generated during the joining process.

Changes in sequence during recombination are therefore a consequence of the enzymatic mechanisms involved in breaking and rejoining the DNA. In heavy chain recombination, base pairs are lost or inserted at the V_H-D or D-J or both junctions. Deletion also occurs in V_λ-J_λ joining, but insertion at these joints is unusual. The changes in sequence affect the amino acid coded at V-D and D-J junctions in heavy chains or at the V-J junction in light chains.

These various mechanisms together ensure that a coding joint may have a sequence that is different from what would be predicted by a direct joining of the coding ends of the V, D, and J regions.

Changes in the sequence at the junction make it possible for a great variety of amino acids to be coded at this site. It is interesting that the amino acid at position 96 is created by the V-J joining reaction. It forms part of the antigen-binding site and also is involved in making contacts between the light and heavy chains. So the maximum diversity is generated at the site that contacts the target antigen.

Changes in the number of base pairs at the coding joint affect the reading frame. The joining process appears to be random with regard to reading frame, so that probably only one third of the joined sequences retain the proper frame of reading through the junctions. If the V-J region is joined so that the J segment is out of phase, translation is terminated prematurely by a nonsense codon in the incorrect frame. We may think of the formation of aberrant genes as comprising the price the cell must pay for the increased diversity that it gains by being able to adjust the sequence at the joining site.

Similar although even greater diversity is generated in the joining reactions that involve the D segment of the heavy chain. The same result is seen with regard to reading frame; nonproductive genes are generated by joining events that place J and C out of phase with the preceding V gene segment.

The joining reaction that works on the coding end uses the same pathway of nonhomologous end-joining (NHEJ) that repairs double-strand breaks in cells (see *15.28 Eukaryotic cells have conserved repair systems)*. The initial stages of the reaction were identified by isolating intermediates from lymphocytes of mice with the *SCID* mutation, which results in a much reduced activity in immunoglobulin and TCR recombination. *SCID* mice accumulate broken molecules that terminate in double-strand breaks at the coding ends, and are thus deficient in completing some aspect of the joining reaction. The *SCID* mutation inactivates a DNA-dependent protein kinase (DNA-PK). The kinase is recruited to DNA by the Ku70 and Ku80 proteins, which bind to the DNA ends. DNA-PK phosphorylates and thereby activates the protein Artemis, which nicks the hairpin ends (it also has exonuclease and endonuclease activities that function in the NHEJ pathway). The actual ligation is undertaken by DNA ligase IV and also requires the protein XRCC4. As a result, mutations in the Ku proteins or in XRCC4 or DNA ligase IV are found among human patients who have diseases caused by deficiencies in DNA repair that result in increased sensitivity to radiation.

What is the connection between joining of V and C gene segments and their activation? Unrearranged V gene segments are not

actively represented in RNA. But when a V gene segment is joined productively to a C_κ gene segment, the resulting unit is transcribed. However, since the sequence upstream of a V gene segment is not altered by the joining reaction, *the promoter must be the same in unrearranged, nonproductively rearranged, and productively rearranged genes.*

A promoter lies upstream of every V gene segment, but is inactive. It is activated by its relocation to the C region. The effect must depend on sequences downstream. What role might they play? An enhancer located within or downstream of the C gene segment activates the promoter at the V gene segment. The enhancer is tissue specific; it is active only in B cells. Its existence suggests the model illustrated in **Figure 26.14,** in which the V gene segment promoter is activated when it is brought within the range of the enhancer.

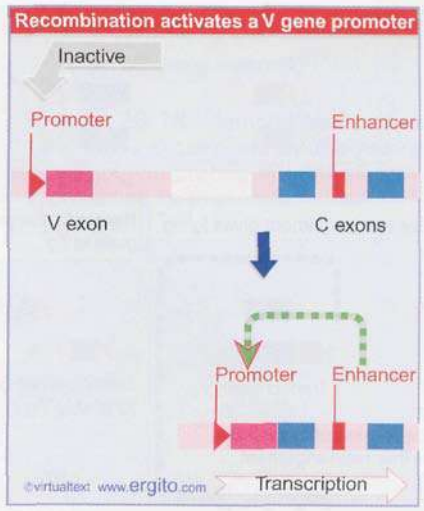


Figure 26.14 A V gene promoter is inactive until recombination brings it into the proximity of an enhancer in the C gene segment. The enhancer is active only in B lymphocytes.

26.10 Allelic exclusion is triggered by productive rearrangement

Key Concepts

- Recombination to generate an intact immunoglobulin gene is productive if it leads to expression of an active protein.
- A productive rearrangement prevents any further rearrangement from occurring, but a nonproductive rearrangement does not.
- Allelic exclusion applies separately to light chains (only one kappa *or* lambda may be productively rearranged) and to heavy chains (one heavy chain is productively rearranged).

Each B cell expresses a single type of light chain and a single type of heavy chain, because only a single productive rearrangement of each type occurs in a given lymphocyte, to produce one light and one heavy chain gene. Because each event involves the genes of only *one* of the homologous chromosomes, *the alleles on the other chromosome are not expressed in the same cell.* This phenomenon is called allelic exclusion.

The occurrence of allelic exclusion complicates the analysis of somatic recombination. A probe reacting with a region that has rearranged on one homologue will also detect the allelic sequences on the other homologue. We are therefore compelled to analyze the different fates of the two chromosomes together.

The usual pattern displayed by a rearranged active gene can be interpreted in terms of a deletion of the material between the recombining V and C loci.

Two types of gene organization are seen in active cells:

- Probes to the active gene may reveal one rearranged copy and one germline copy. We assume then that joining has occurred on one chromosome, while the other chromosome has remained unaltered.
- Two different rearranged patterns may be found, indicating that the chromosomes have suffered independent rearrangements. In some of these instances, material between the recombining V and C gene segments is entirely absent from the cell line. This is most easily explained by the occurrence of independent deletions (resulting from recombination) on each chromosome.

When two chromosomes both lack the germline pattern, usually only one of them has passed through a productive rearrangement to generate a functional gene. The other has suffered a nonproductive rearrangement; this may take several forms, but in each case the gene sequence cannot be expressed as an immunoglobulin chain. (It may be incomplete, for example, because D-J joining has occurred but V-D joining has not followed; or

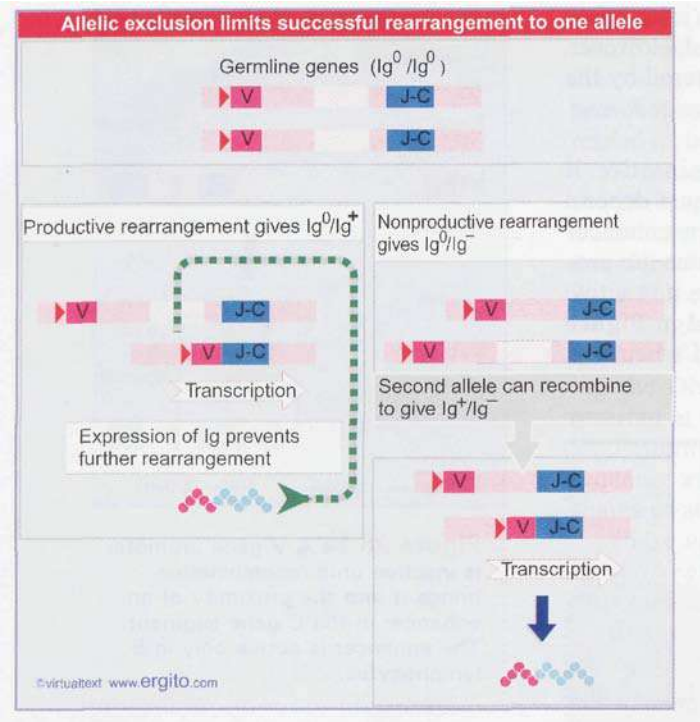


Figure 26.15 A successful rearrangement to produce an active light or heavy chain suppresses further rearrangements of the same type, and results in allelic exclusion.

it may be aberrant, with the process completed, but failing to generate a gene that codes for a functional protein.)

The coexistence of productive and nonproductive rearrangements suggests the existence of a feedback loop to control the recombination process. A model is outlined in **Figure 26.15.** Suppose that each cell starts with two loci in the unrearranged germline configuration Ig°. Either of these loci may be rearranged to generate a productive gene Ig^+ or a nonproductive gene Ig~.

If the rearrangement is productive, the synthesis of an active chain provides a trigger to prevent rearrangement of the other allele. The active cell has the configuration Ig^0/Ig^+.

If the rearrangement is nonproductive, it creates a cell with the configuration Ig^0/Ig^-. There is no impediment to rearrangement of the remaining germline allele. If this rearrangement is productive, the expressing cell has the configuration Ig^+/Ig^-. Again, the presence of an active chain suppresses the possibility of further rearrangements.

Two successive "nonproductive rearrangements produce the cell Ig-/Ig-. In some cases an Ig-/Ig-cell can try yet again. Sometimes the observed patterns of DNA can only have been generated by successive rearrangements.

The crux of the model is that the cell keeps trying to recombine V and C gene segments until a productive rearrangement is achieved. Allelic exclusion is caused by the suppression of further rearrangement as soon as an active chain is produced. The use of this mechanism *in vivo* is demonstrated by the creation of transgenic mice whose germline has a rearranged immunoglobulin gene. Expression of the transgene in B cells suppresses the rearrangement of endogenous genes.

Allelic exclusion is independent for the heavy- and light-chain loci. Heavy chain genes usually rearrange first. Allelic exclusion for light chains must apply equally to both families (cells may have *either* active κ or λ light chains). It is likely that the cell rearranges its K genes first, and tries to rearrange λ only if both K attempts are unsuccessful.

There is an interesting paradox in this series of events. The same consensus sequences and the same V(D)J recombinase are involved in the recombination reactions at H, K, and λ loci. Yet the three loci rearrange in a set order. What ensures that heavy rearrangement precedes light rearrangement and that K precedes λ? The loci may become accessible to the enzyme at different times, possibly as the result of transcription. Transcription occurs even before rearrangement, although of course the products have no coding function. The transcriptional event may change the structure of chromatin, making the consensus sequences for recombination available to the enzyme.

26.11 Class switching is caused by DNA recombination

Key Concepts

- Immunoglobulins are divided into five classes according to the type of constant region in the heavy chain.
- Class switching to change the C_H region occurs by a recombination between S regions that deletes the region between the old C_H region and the new C_H region.
- Multiple successive switch recombinations can occur.

By Book_Crazy [IND]

There are five types of heavy chain

Type	IgM	IgD	IgG	IgA	IgE
Heavy chain	μ	δ	γ	α	ε
Structure	$(\mu_2L_2)_5J$	δ_2L_2	γ_2L_2	$(\alpha_2L_2)_2$	ε_2L_2
Proportion	5%	1%	80%	14%	<1%
Effector function	Activates complement	Development of tolerance (?)	Activates complement	Found in secretions	Allergic response

Figure 26.16 Immunoglobulin type and function is determined by the heavy chain. J is a joining protein in IgM; all other Ig types exist as tetramers.

The *class* of immunoglobulin is defined by the type of C_H region. **Figure 26.16** summarizes the five Ig classes. IgM (the first immunoglobulin to be produced by any B cell) and IgG (the most common immunoglobulin) possess the central ability to activate complement, which leads to destruction of invading cells. IgA is found in secretions (such as saliva), and IgE is associated with the allergic response and defense against parasites.

All lymphocytes start productive life as immature cells engaged in synthesis of IgM. Cells expressing IgM have the germline arrangement of the C_H gene segment cluster shown in Figure 26.10. The V-D-J joining reaction triggers expression of the C_μ gene segment. A lymphocyte generally produces only a single class of immunoglobulin at any one time, but the class may change during the cell lineage. A change in expression is called class switching. It is accomplished by a substitution in the type of C_H region that is expressed. Switching can be stimulated by environmental effects; for example, the growth factor TGFβ causes switching from C_μ to C_α.

Switching involves only the C_H gene segment; the same V_H gene segment continues to be expressed. So a given V_H gene segment may be expressed successively in combination with more than one C_H gene segment. The same light chain continues to be expressed throughout the lineage of the cell. Class switching therefore allows the type of effector response (mediated by the C_H region) to change, while maintaining the same capacity to recognize antigen (mediated by the V regions).

Changes in the expression of C_H gene segments are made in two ways. The majority occur via further DNA recombination events, involving a system different from that concerned with V-D-J joining (and able to operate only later during B cell development). Another type of change occurs at the level of RNA processing, but generally this is involved with changing the C-terminal sequence of the C_H region rather than its class (see *26.13 Early heavy chain expression can be changed by RNA processing*).

Cells expressing downstream C_H gene segments have deletions of C_μ and the other gene segments preceding the expressed C_H gene segment. Class switching is accomplished by a recombination to bring a new C_H gene segment into juxtaposition with the expressed V-D-J unit. *The sequences of switched V-D-J-C_H units show that the sites of switching lie upstream of the C_H gene segments* themselves. The switching sites are called S regions. **Figure 26.17** depicts two successive switches.

In the first switch, expression of C_μ is succeeded by expression of $C_{\gamma 1}$. The $C_{\gamma 1}$ gene segment is brought into the expressed position by recombination between the sites S_μ and $S_{\gamma 1}$. The S_μ site lies between V-D-J and the C_μ gene segment. The $S_{\gamma 1}$ site lies upstream of the $C_{\gamma 1}$ gene segment. The DNA sequence between the two switch sites is excised as a circular molecule.

The linear deletion model imposes a restriction on the heavy-gene locus: *once a class switch has been made, it becomes impossible to express any C_H gene segment that used to reside between C_μ and the new C_H gene*

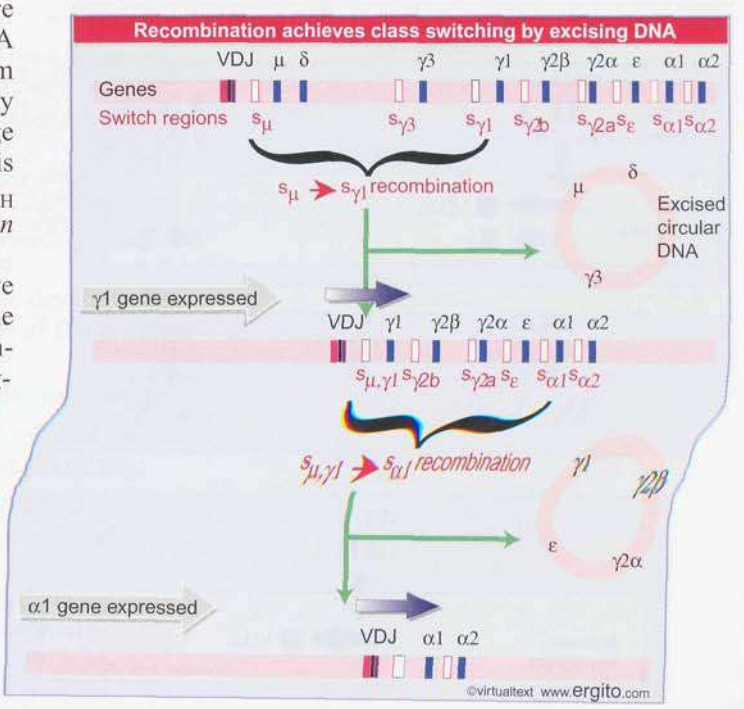


Figure 26.17 Class switching of heavy genes may occur by recombination between switch regions (S), deleting the material between the recombining S sites. Successive switches may occur.

segment. In the example of Figure 26.17, cells expressing $C_{\gamma 1}$ should be unable to give rise to cells expressing $C_{\gamma 3}$, which has been deleted.

However, it should be possible to undertake another switch to any C_H gene segment *downstream* of the expressed gene. Figure 26.17 shows a second switch to C_α expression, accomplished by recombination between $S_{\alpha 1}$ and the switch region $S_{\mu,\gamma 1}$ that was generated by the original switch.

We assume that all of the C_H gene segments have S regions upstream of the coding sequences. We do not know whether there are any restrictions on the use of S regions. Sequential switches do occur, but we do not know whether they are optional or an obligatory means to proceed to later C_H gene segments. We should like to know whether IgM can switch directly to *any* other class. Because the S regions lie within the introns that precede the C_H coding regions, switching does not alter the translational reading frame.

26.12 Switching occurs by a novel recombination reaction

Key Concepts

- Switching occurs by a double-strand break followed by the nonhomologous end joining reaction.
- The important feature of a switch region is the presence of inverted repeats.
- Switching requires activation of promoters that are upstream of the switch sites.

We know that switch sites are not uniquely defined because different cells expressing the same C_H gene segment prove to have recombined at different points. Switch regions vary in length (as defined by the limits of the sites involved in recombination) from 1-10 kb. They contain groups of short inverted repeats, with repeating units that vary from 20-80 nucleotides in length. The primary sequence of the switch region does not seem to be important; what matters is the presence of the inverted repeats.

An S region typically is located ~2 kb upstream of a C_H gene segment. The switching reaction releases the excised material between the switch sites as a circular DNA molecule. Two of the proteins required for the joining phase of VDJ recombination (and also for the general nonhomologous end-joining pathway, NHEJ), Ku and DNA-PKcs, are required, suggesting that the joining reaction may use the NHEJ pathway. Basically, this implies that the reaction occurs by a double-strand break followed by rejoining of the cleaved ends.

We can put together the features of the reaction to propose a model for the generation of the double-strand break. The critical points are

- transcription through the S region is required;
- the inverted repeats are crucial;
- and the break can occur at many different places within the S region.

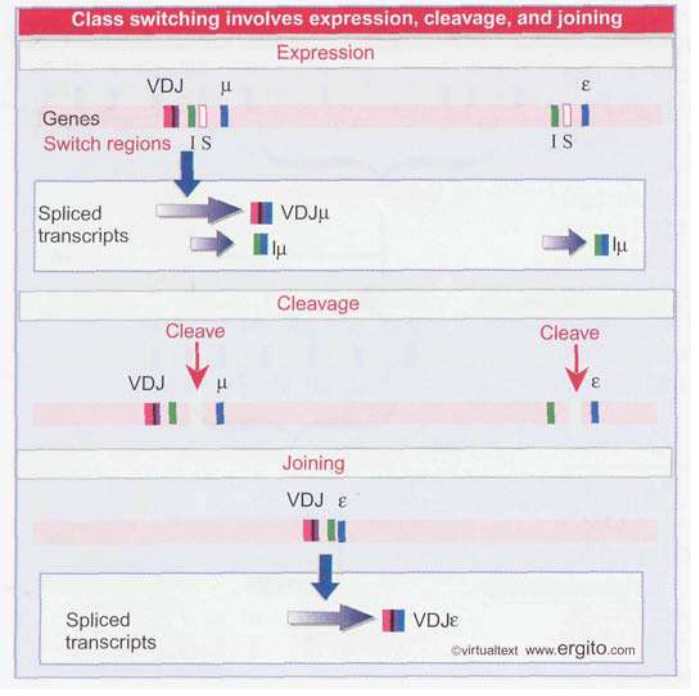


Figure 26.18 Class switching passes through discrete stages. The I promoters initiate transcription of sterile transcripts. The switch regions are cleaved. Joining occurs at the cleaved regions.

Figure 26.18 shows the stages of the class switching reaction. A promoter (I) lies immediately upstream of each switch region. Switching requires transcription from this promoter. The promoter may respond to activators that respond to environmental conditions, such as

stimulation by cytokines, thus creating a mechanism to regulate switching. The first stage in switching is therefore to activate the I promoters that are upstream of each of the switch regions that will be involved. When these promoters are activated, they generate sterile transcripts that are spliced to join the I region with the corresponding heavy constant region. The splicing reaction may be needed for switching.

Figure 26.19 shows that transcription can potentially affect the structure of sequences of DNA that are palindromic. This is a consequence of the separation of DNA strands during the process. When RNA polymerase passes through a palindromic sequence, the inverted repeats on the nontemplate strand might pair to form a hairpin. This hairpin could be a target for an endonuclease that specifically recognizes stem-loop structures. This would break one strand and possibly could trigger breakage of the other strand.

A cytidine deaminase is activated when class switching occurs. Known as AID (activation-induced deaminase) it is a member of a class of enzymes that usually act on RNA in connection with RNA editing to change a cytidine to a uridine (see *25.9 RNA editing occurs at individual bases)*. We do not know whether this particular enzyme acts on RNA or DNA, but class switching does not occur in its absence. The reaction is blocked before the nicking stage. Assuming the enzyme acts on RNA, there are two possibilities for its function. The sterile transcript may be involved in the double-strand break reaction in some direct structural capacity that requires RNA editing. Or the enzyme may be required for editing of an mRNA that codes for one of the essential enzymes.

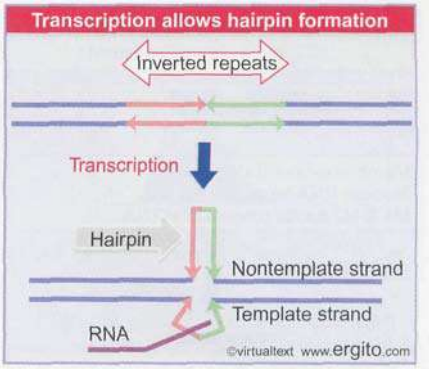


Figure 26.19 When transcription separates the strands of DNA, one strand may form an inverted repeat if the sequence is palindromic.

26.13 Early heavy chain expression can be changed by **RNA** processing

Key Concepts

- All lymphocytes start by synthesizing the membrane-bound form of IgM.
- A change in RNA splicing causes this to be replaced by the secreted form when the B cell differentiates.

The period of IgM synthesis that begins lymphocyte development *falls into two parts, during which different versions of the* μ *constant* region are synthesized.

As a stem cell differentiates to a pre-B lymphocyte, an accompanying light chain is synthesized, and the IgM molecule ($L_2\mu_2$) appears at the surface of the cell. This form of IgM contains the μ_m version of the constant region *(m* indicates that IgM is located in the membrane). The membrane location may be related to the need to initiate cell proliferation in response to the initial recognition of an antigen.

When the B lymphocyte differentiates further into a plasma cell, the μ_s version of the constant region is expressed. The IgM actually is secreted as a pentamer IgM_5J, in which J is a joining polypeptide (no connection with the J region) that forms disulfide linkages with μ chains. Secretion of the protein is followed by the humoral response depicted in Figure 26.1.

The μ_m and μ_s versions of the μ heavy chain differ only at the C-terminal end. The μ_m chain ends in a hydrophobic sequence that probably secures it in the membrane. This sequence is replaced by a shorter hydrophilic sequence in μ_s; the substitution allows the μ heavy chain to pass through the membrane. The change of C-terminus is accomplished

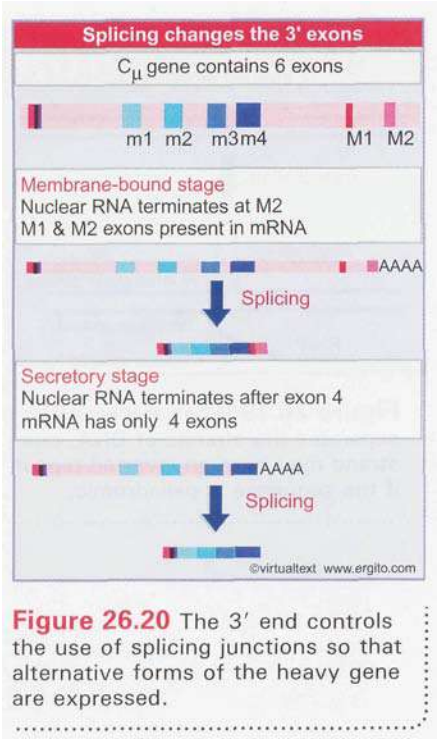


Figure 26.20 The 3′ end controls the use of splicing junctions so that alternative forms of the heavy gene are expressed.

by an alternative splicing event, which is controlled by the 3' end of the nuclear RNA, as illustrated in **Figure 26.20.**

At the membrane-bound stage, the RNA terminates after exon M2, and the constant region is produced by splicing together six exons. The first four exons code for the four domains of the constant region. The last two exons, M1 and M2, code for the 41-residue hydrophobic C-terminal region and its nontranslated trailer. The 5' splice junction within exon 4 is connected to the 3' splice junction at the beginning of M1.

At the secreted stage, the nuclear RNA terminates after exon 4. The 5' splice junction within this exon that had been linked to M1 in the membrane form is ignored. This allows the exon to extend for an additional 20 codons.

A similar transition from membrane to secreted forms is found with other constant regions. The conservation of exon structures suggests that the mechanism is the same.

26.14 Somatic mutation generates additional diversity in mouse and man

Key Concepts

- Active immunoglobulin genes have V regions with sequences that are changed from the germline because of somatic mutation.
- The mutations occur as substitutions of individual bases.
- The sites of mutation are concentrated in the antigen-binding site.
- The process depends on the enhancer that activates transcription at the Ig locus.

Comparisons between the sequences of expressed immunoglobulin genes and the corresponding V gene segments of the germline show that new sequences appear in the expressed population. Some of this additional diversity results from sequence changes at the V-J or V-D-J junctions that occur during the recombination process. However, other changes occur upstream at locations within the variable domain.

Two types of mechanism can generate changes in V gene sequences after rearrangement has generated a functional immunoglobulin gene. In mouse and Man, the mechanism is the induction of somatic mutations at individual locations within the gene specifically in the active lymphocyte. The process is sometimes called hypermutation. In chicken, rabbit, and pig, a different mechanism uses gene conversion to change a segment of the expressed V gene into the corresponding sequence from a different V gene (see *26.16 Avian immunoglobulins are assembled from pseudogenes).*

A probe representing an expressed V gene segment can be used to identify all the corresponding fragments in the germline. Their sequences should identify the complete repertoire available to the organism. Any expressed gene whose sequence is different must have been generated by somatic changes.

One difficulty is to ensure that every potential contributor in the germline V gene segments actually has been identified. This problem is overcome by the simplicity of the mouse λ chain system. A survey of several myelomas producing λ_1 chains showed that many have the sequence of the single germline gene segment. *But others have new sequences that must have been generated by mutation of the germline gene segment.*

To determine the frequency of somatic mutation in other cases, we need to examine a large number of cells in which the same V gene segment is expressed. A practical procedure for identifying such a group is

to characterize the immunoglobulins of a series of cells, all of which express an immune response to a particular antigen.

(Epitopes used for this purpose are small molecules—haptens—whose discrete structure is likely to provoke a consistent response, unlike a large protein, different parts of which provoke different antibodies. A hapten is conjugated with a nonreactive protein to form the antigen. The cells are obtained by immunizing mice with the antigen, obtaining the reactive lymphocytes, and sometimes fusing these lymphocytes with a myeloma [immortal tumor] cell to generate a hybridoma that continues to express the desired antibody indefinitely.)

In one example, 10 out of 19 different cell lines producing antibodies directed against the hapten phosphorylcholine had the same V_H sequence. This sequence was the germline V gene segment T15, one of four related V_H genes. The other 9 expressed gene segments differed from each other and from all 4 germline members of the family. They were more closely related to the T15 germline sequence than to any of the others, and their flanking sequences were the same as those around T15. This suggested that they arose from the T15 member by somatic mutation.

Figure 26.21 shows that sequence changes extend around the V gene segment. They take the form of substitutions of individual nucleotide pairs. The variation is different in each case. It represents ~3-15 substitutions, corresponding to <10 amino acid substitutions in the protein. They are concentrated in the antigen-binding site (thus generating the maximum diversity for recognizing new antigens). Only some of the mutations affect the amino acid sequence, since others lie in third-base coding positions as well as in nontranslated regions.

The large proportion of ineffectual mutations suggests that somatic mutation occurs more or less at random in a region including the V gene segment and extending beyond it. There is a tendency for some mutations to recur on multiple occasions. These may represent hotspots as a result of some intrinsic preference in the system.

Somatic mutation occurs during clonal proliferation, apparently at a rate $\sim 10^{-3}$ per bp per cell generation. Approximately half of the progeny cells gain a mutation; as a result, cells expressing mutated antibodies become a high fraction of the clone.

In many cases, a single family of V gene segments is used consistently to respond to a particular antigen. Upon exposure to an antigen, presumably the V region with highest intrinsic affinity provides a starting point. Then somatic mutation increases the repertoire. Random mutations have unpredictable effects on protein function; some inactivate the protein, others confer high specificity for a particular antigen. The proportion and effectiveness of the lymphocytes that respond is increased by selection among the lymphocyte population for those cells bearing antibodies in which mutation has increased the affinity for the antigen.

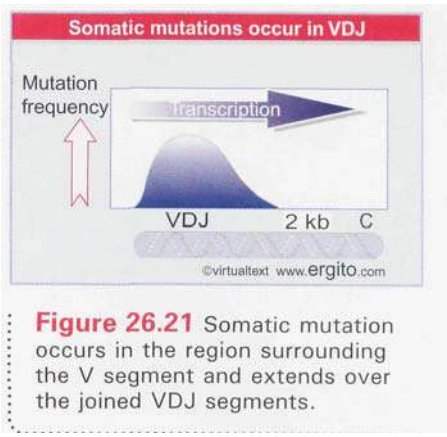


Figure 26.21 Somatic mutation occurs in the region surrounding the V segment and extends over the joined VDJ segments.

26.15 Somatic mutation is induced by cytidine deaminase and uracil glycosylase

Key Concepts

- A cytidine deaminase is required for somatic mutation as well as for class switching.
- Uracil-DNA glycosylase activity influences the pattern of somatic mutations.
- Hypermutation may be initiated by the sequential action of these enzymes.

Somatic mutation requires the enhancer that activates transcription at each Ig locus. There is a correlation between the occurrence of transcription and the induction of mutations, but we do not understand the role of transcription.

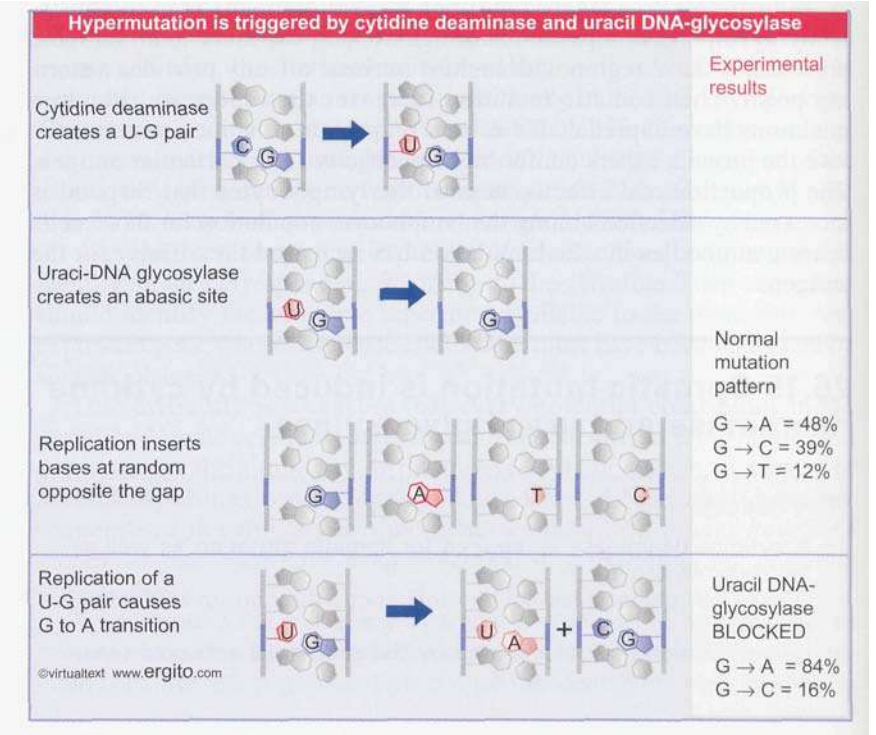
Somatic mutation requires the same cytidine deaminase (called AID for activation-induced deaminase) that is required during class switching. Parallels in the requirements for class switching and hypermutation suggest that similar reactions could be involved. One idea this suggested was that hypermutation could result from introducing a break into DNA and then repairing it with an error-prone system.

The main question has been whether cytidine deaminase is directly involved because it targets sites in the hypermutated region, or whether its action is indirect, for example because it is involved in some editing event elsewhere that is needed to generate products required for hypermutation. The answer that its action is direct is suggested by the involvement of uracil-DNA glycosylase.

Recall that deamination of cytidine in DNA generates uracil (see Figure 1.20). Uracil is removed from DNA by the enzyme uracil-DNA glycosylase (see Figure 15.41). **Figure 26.22** shows that what types of mutation result if we put these two events together.

Suppose that these enzymes are activated in B cells, and the generation and removal of uracil overwhelms the excision-repair systems that would usually respond. Error-prone replication past the missing base will probably insert bases at random into the daughter strand in which case 3 out of four daughter strands will have a mutation. We see from the results summarized in the figure that all three possible types of substitution are found (the exact proportions perhaps influenced by the extent to which repair systems have intervened, intrinsic preferences of the polymerase, etc.).

If the action of uracil-DNA glycosylase is blocked, however, we see a different result. If uracil is not removed from DNA, it should pair with adenine during replication. The ultimate result is to replace the original C-G pair with a T-A pair. Uracil-DNA glycosylase can be blocked by introducing into cells the gene coding for a protein that inhibits the enzyme.



By Book_Crazy [IND]

Figure 26.22 When the action of cytidine deaminase (top) is followed by that of uracil DNA-glycosylase, an abasic site is created. Replication past this site should insert all four bases at random into the daughter strand (center). If the uracil is not removed from the DNA, its replication causes a C-G to T-A transition.

(The gene is a component of the bacteriophage PSB-2, whose genome is unusual in containing uracil, so that the enzyme needs to be blocked during a phage infection.) When the gene is introduced into a lymphocyte cell line, there is a dramatic change in the pattern of mutations, with almost all comprising the predicted transition from C-G to A-T.

The fact that blocking uracil-DNA glycosylase changes the nature of the mutations found during immunoglobulin somatic mutation suggests that the enzyme is usually involved in the process. Together with the requirement for the AID cytidine deaminase, this suggests a pathway in which these two enzymes act on sites in the hypermutated region to initiate the mutagenic process. We don't know yet what restricts their action to the target region for hypermutation.

26.16 Avian immunoglobulins are assembled from pseudogenes

Key Concepts

- An immunoglobulin gene in chicken is generated by copying a sequence from one of 25 pseudogenes into the V gene at a single active locus.

The chick immune system is the paradigm for rabbits, cows, and pigs, which rely upon using the diversity that is coded in the genome. A similar mechanism is used by both the single light chain locus (of the *X* type) and the H chain locus. The organization of the *X* locus is drawn in Figure 26.23. It has only one functional V gene segment, J segment, and C gene segment. Upstream of the functional $V_{\lambda 1}$ gene segment lie 25 V_{λ} pseudogenes, organized in either orientation. They are classified as pseudogenes because either the coding segment is deleted at one or both ends, or proper signals for recombination are missing (or both). This assignment is confirmed by the fact that only the $V_{\lambda 1}$ gene segment recombines with the $J\text{-}C_{\lambda}$ gene segment.

But sequences of active rearranged $V_{\lambda}\text{-}J\text{-}C_{\lambda}$ gene segments show considerable diversity! A rearranged gene has one or more positions at which a cluster of changes has occurred in the sequence. A sequence identical to the new sequence can almost always be found in one of the pseudogenes (which themselves remain unchanged). The exceptional sequences that are not found in a pseudogene always represent changes at the junction between the original sequence and the altered sequence.

So a novel mechanism is employed to generate diversity. Sequences from the pseudogenes, between 10 and 120 bp in length, are substituted into the active $V_{\lambda 1}$ region by gene conversion. The unmodified $V_{\lambda 1}$ sequence is not expressed, even at early times during the immune response. A successful conversion event probably occurs every 10-20 cell divisions to every rearranged $V_{\lambda 1}$ sequence. At the end of the immune maturation period, a rearranged $V_{\lambda 1}$ sequence has 4-6 converted segments spanning its entire length, derived from different donor pseudogenes. If all pseudogenes participate, this allows 25 X 10^8 possible combinations!

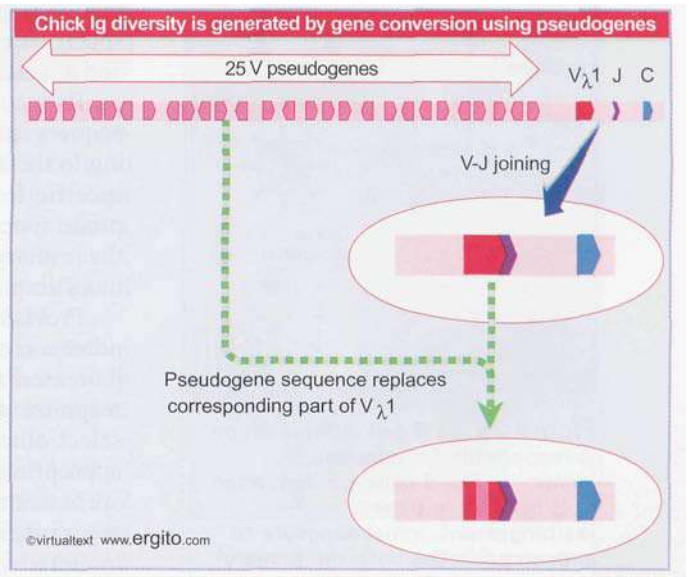


Figure 26.23 The chicken lambda light locus has 25 V pseudogenes upstream of the single functional V-J-C region.

The enzymatic basis for copying pseudogene sequences into the expressed locus depends on enzymes involved in recombination and is related to the mechanism for somatic hypermutation that introduces

diversity in mouse and man. Some of the genes involved in recombination are required for the gene conversion process; for example, it is prevented by deletion of *RAD54*. Deletion of other recombination genes (*XRCC2*, *XRCC3*, and *RAD51B*) has another very interesting effect: somatic mutation occurs at the V gene in the expressed locus. The frequency of the somatic mutation is ~10× greater than the usual rate of gene conversion.

These results show that the absence of somatic mutation in chick is not due to a deficiency in the enzymatic systems that are responsible in mouse and man. The most likely explanation for a connection between (lack of) recombination and somatic mutation is that unrepaired breaks at the locus trigger the induction of mutations. The reason why somatic mutation occurs in mouse and man but not in chick may therefore lie with the details of the operation of the repair system that operates on breaks at the locus. It is more efficient in chick, so that the gene is repaired by gene conversion before mutations can be induced.

26.17 B cell memory allows a rapid secondary response

Key Concepts

- The primary response to an antigen is mounted by B cells that do not survive beyond the response period.
- Memory B cells are produced that have specificity for the same antigen but that are inactive.
- A reexposure to antigen triggers the secondary response in which the memory cells are rapidly activated.

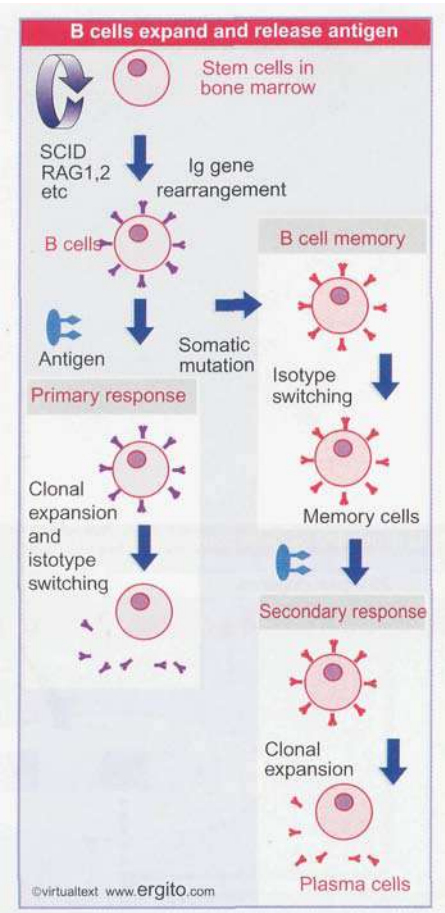


Figure 26.24 B cell differentiation is responsible for acquired immunity. Pre-B cells are converted to B cells by Ig gene rearrangement. Initial exposure to antigen provokes both the primary response and storage of memory cells. Subsequent exposure to antigen provokes the secondary response of the memory cells.

We are now in a position to summarize the relationship between the generation of high-affinity antibodies and the differentiation of the B cell. **Figure 26.24** shows that B cells are derived from a self-renewing population of stem cells in the bone marrow. Maturation to give B cells depends upon Ig gene rearrangement, which requires the functions of the *SCID* and *RAG 1,2* (and other) genes. If gene rearrangement is blocked, mature B cells are not produced. The antibodies carried by the B cells have specificities determined by the particular combinations of V(D)J regions, and any additional nucleotides incorporated during the joining process.

Exposure to antigen triggers two aspects of the immune response. The primary immune response occurs by clonal expansion of B cells responding to the antigen. This generates a large number of plasma cells that are specific for the antigen; isotype switching occurs to generate the appropriate type of effector response. The population of cells concerned with the primary response is a dead end; these cells do not live beyond the primary response itself.

Provision for a secondary immune response is made through the phenomenon of B cell memory. Somatic mutation generates B cells that have increased affinity for the antigen. These cells do not trigger an immune response at this time, although they may undergo isotype switching to select other forms of C_H region. They are stored as memory cells, with appropriate specificity and effector response type, but are inactive. They are activated if there is a new exposure to the same antigen. Because they are pre-selected for the antigen, they enable a secondary response to be mounted very rapidly, simply by clonal expansion; no further somatic mutation or isotype switching occurs during the secondary response.

The pathways summarized in Figure 26.24 show the development of acquired immunity, that is, the response to an antigen. In addition to these cells, there is a separate set of B cells, named the Ly-1 cells. These

cells have gone through the process of V gene rearrangement, and apparently are selected for expression of a particular repertoire of antibody specificities. They do not undergo somatic mutation or the memory response. They may be involved in natural immunity, that is, an intrinsic ability to respond to certain antigens.

A more detailed view of B cell development is shown in **Figure 26.25.** The first step is recombination between the D and J segments of the μ heavy chain. This is succeeded by V-D recombination, generating a μ heavy chain. Several recombination events, involving a succession of nonproductive and productive rearrangements, may occur, as shown previously in Figure 26.15. These cells express a protein resembling a λ chain, called the surrogate light (SL) chain, which is expressed on the surface and associates with the μ heavy chain to form the pre-B-receptor. It resembles an immunoglobulin complex, but does not function as one.

The production of μ chain represses synthesis of SL chain, and the cells divide to become small pre-B cells. Then light chain is expressed and functional immunoglobulin appears on the surface of the immature B cells. Further cell divisions occur, and the expression of δ heavy chain is added to that of μ chain, as the cells mature into B cells.

Immunoglobulins function both by secretion from B cells and by surface expression. **Figure 26.26** shows that the active complex on the cell surface is called the B cell receptor (BCR), and consists of an immunoglobulin associated with transmembrane proteins called Igα and Igβ. They provide the signaling components that trigger intracellular pathways in response to antigen-antibody binding. The activation of the BCR is also influenced by interactions with other receptors, for example, to mediate the interaction of antigen-activated B cells with helper T cells.

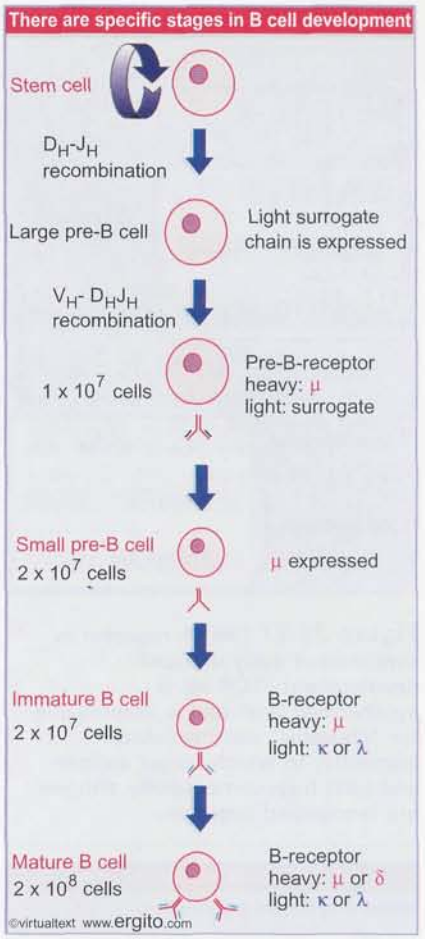


Figure 26.25 B cell development proceeds through sequential stages.

26.18 T cell receptors are related to immunoglobulins

Key Concepts

- T cells use a similar mechanism of V(D)J-C joining to B cells to produce either of two types of T cell receptor.
- TCR αβ is found on >95% of T lymphocytes, and TCR γδ is found on <5%.

The lymphocyte lineage presents an example of evolutionary opportunism: a similar procedure is used in both B cells and T cells to generate proteins that have a variable region able to provide significant diversity, while constant regions are more limited and account for a small range of effector functions. T cells produce either of two types of T cell receptor. The different T cell receptors are synthesized at different times during T cell development, as summarized in **Figure 26.27.**

The 78 receptor is found on <5% of T lymphocytes. It is synthesized only at an early stage of T cell development. In mice, it is the only receptor detectable at < 15 days of gestation, but has virtually been lost by birth at day 20.

TCR αβ is found on >95% of lymphocytes. It is synthesized later in T cell development than 78. In mice, it first becomes apparent at 15–17 days after gestation. By birth it is the predominant receptor. It is synthesized by a separate lineage of cells from those involved in TCR 78 synthesis, and involves independent rearrangement events.

Like immunoglobulins, a TCR must recognize a foreign antigen of unpredictable structure. The problem of antigen recognition by B cells and T cells is resolved in the same way, and the organization of the T

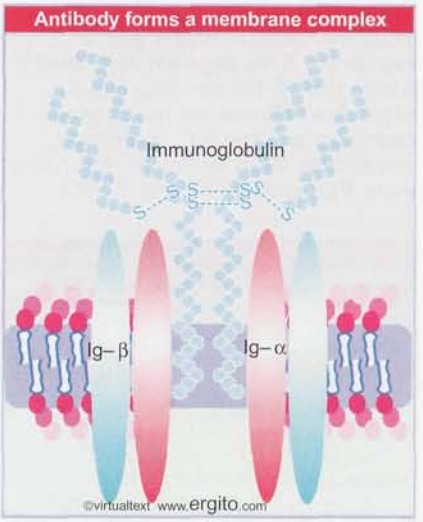


Figure 26.26 The B cell antigen receptor consists of an immunoglobulin tetramer (H_2L_2) linked to two copies of the signal-transducing heterodimer (Igαβ).

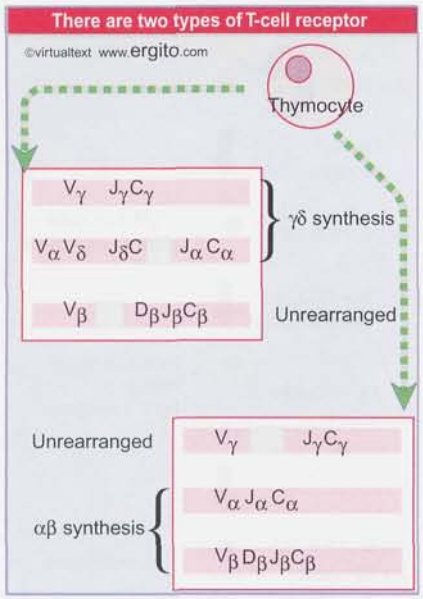


Figure 26.27 The γδ receptor is synthesized early in T-cell development. TCR αβ is synthesized later and is responsible for "classical" cell-mediated immunity, in which target antigen and host histocompatibility antigen are recognized together.

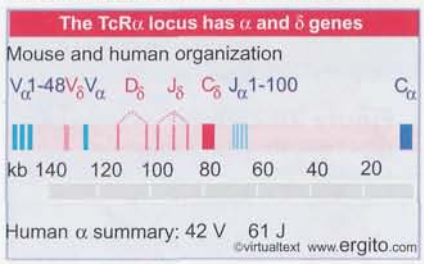


Figure 26.28 The human TCRα locus has interspersed α and δ segments. A V_δ segment is located within the V_α cluster. The D-J-C_δ segments lie between the V gene segments and the J-C_α segments. The mouse locus is similar, but has more V_δ segments.

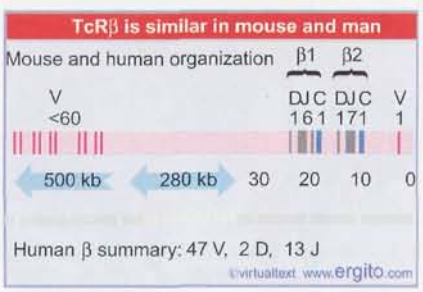


Figure 26.29 The TCRβ locus contains many V gene segments spread over ~500 kb, and lying ~280 kb upstream of the two D-J-C clusters.

cell receptor genes resembles the immunoglobulin genes in the use of variable and constant regions. *Each locus is organized in the same way as the immunoglobulin genes, with separate segments that are brought together by a recombination reaction specific to the lymphocyte.* The components are the same as those found in the three Ig families.

The organization of the TCR proteins resembles that of the immunoglobulins. The V sequences have the same general internal organization in both Ig and TCR proteins. The TCR C region is related to the constant Ig regions and has a single constant domain followed by transmembrane and cytoplasmic portions. Exon-intron structure is related to protein function.

The resemblance of the organization of TCR genes with the Ig genes is striking. As summarized in **Figure 26.28,** the organization of TCR α resembles that of Ig K, with V gene segments separated from a cluster of J segments that precedes a single C gene segment. The organization of the locus is similar in both Man and mouse, with some differences only in the number of V_α gene segments and J_α segments.

The components of TCR β resemble those of IgH. **Figure 26.29** shows that the organization is different, with V gene segments separated from two clusters each containing a D segment, several J segments, and a C gene segment. Again, the only differences between Man and mouse are in the numbers of the V_β and J_β units.

Diversity is generated by the same mechanisms as in immunoglobulins. Intrinsic diversity results from the combination of a variety of V, D, J, and C segments; some additional diversity results from the introduction of new sequences at the junctions between these components (in the form of P and N nucleotides; see Figure 26.13). Some TCR β chains incorporate two D segments, generated by D-D joins (directed by an appropriate organization of the nonamer and heptamer sequences). A difference between TCR and Ig is that somatic mutation does not occur at the TCR loci. Measurements of the extent of diversity show that the 10^{12} T cells in Man contain 2.5 X 10^7 different a chains associated with 10^6 different β chains.

The same mechanisms are likely to be involved in the reactions that recombine Ig genes in B cells and TCR genes in T cells. The recombining TCR segments are surrounded by nonamer and heptamer consensus sequences identical to those used by the Ig genes. This argues strongly that the same enzymes are involved. Most rearrangements probably occur by the deletion model (see Figure 26.12). We do not know how the process is controlled so that Ig loci are rearranged in B cells, while T cell receptors are rearranged in T cells.

The organization of the γ locus resembles that of Ig *X,* with V gene segments separated from a series of J-C segments. **Figure 26.30** shows that this locus has relatively little diversity, with ~8 functional V segments. The organization is different in Man and mouse. Mouse has 3 functional J-C loci, but some segments are inverted in orientation. Man has multiple J segments for each C gene segment.

The 8 subunit is coded by segments that lie at the TCR a locus, as illustrated previously in Figure 26.28. The segments D_δ-D_δ-J_δ-C_δ lie between the V gene segments and the J_α-C_α segments. Both of the D segments may be incorporated into the 8 chain to give the structure VDD J. The nature of the V gene segments used in the 8 rearrangement is an interesting question. Very few V sequences are found in active TCR 8 chains. In man, only one V gene segment is in general use for 8 rearrangement. In mouse, several V_δ segments are found; some are unique for 8 rearrangement, but some are also found in a rearrangements. The basis for specificity in choosing V segments in a and 8 rearrangement is not known. One possibility is that many of the V_α gene segments can be joined to the DDJ_δ segment, but that only some (therefore defined as V_δ) can give active proteins.

By Book_Crazy [IND]

While for the present we have labeled the V segments that are found in 8 chains as V§ gene segments, we must reserve judgment on whether they are really unique to 8 rearrangement. The interspersed arrangement of genes implies that synthesis of the TCR αβ receptor and the γδ receptor is mutually exclusive at any one allele, because the 8 locus is lost entirely when the V_α-J_α rearrangement occurs.

Rearrangements at the TCR loci, like those of immunoglobulin genes, may be productive or nonproductive. The β locus shows allelic exclusion in much the same way as immunoglobulin loci; rearrangement is suppressed once a productive allele has been generated. The a locus may be different; several cases of continued rearrangement suggest the possibility that substitution of V_α sequences may continue after a productive allele has been generated.

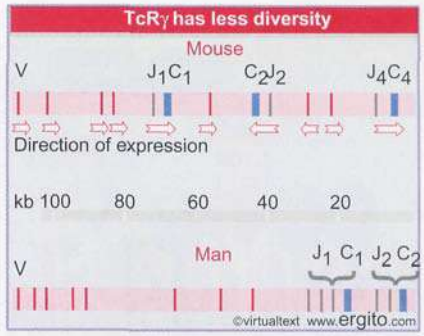


Figure 26.30 The TCRγ locus contains a small number of functional V gene segments (and also some pseudogenes; not shown), lying upstream of the J-C loci.

26.19 The T cell receptor functions in conjunction with the MHC

Key Concepts

- The TCR recognizes a short peptide that is bound in a groove of an MHC protein on the surface of the presenting cell.

T cells with αβ receptors are divided into several subtypes that have a variety of functions connected with interactions between cells involved in the immune response. Cytotoxic T cells (also known as killer T cells) possess the capacity to lyse an infected target cell. Helper T cells assist T cell-mediated target killing or B cell-mediated antibody-antigen interaction.

A major difference between the antibodies of B cells and the receptors of T cells is the way that they handle the antigen. An antibody recognizes a short region (an epitope) within the antigen. The T cell receptor binds to a small peptide (4-5 amino acids long) that has been processed from the antigen by cleavage reactions. The peptide fragment is "presented" to the T cell by an MHC protein. So the T cell simultaneously recognizes the foreign antigen and an MHC protein carried by the presenting cell, as illustrated previously in Figure 26.2. Both helper T cells and killer T cells work in this way, but they have different requirements for the presentation of antigen; different types of MHC protein are used in each case (see next section). Helper T cells require antigen to be presented by an MHC class II protein, while killer T cells require antigen to be presented by an MHC class I protein.

The TCR αβ receptor is responsible for helper T cell function in humoral *immunity, and* for killer T *cell* function in *cell-mediated immunity. This places upon it the* responsibility of recognizing both the foreign antigen and the host MHC protein. The MHC protein binds a *short peptide derived* from *the foreign* antigen, *and the TCR then* recognizes the peptide in a groove on the surface of the MHC. The MHC is said to *present* the peptide to the TCR. (The peptide is generated when the proteasome degrades the foreign protein to generate fragments of 8-10 residues long, as described in *8.32 The proteasome is a large machine that degrades ubiquitinated proteins.)* A given TCR has specificity for a particular MHC as well as for the foreign antigen. The basis for this dual capacity is one of the most interesting issues to be defined *about the αβ TCR.*

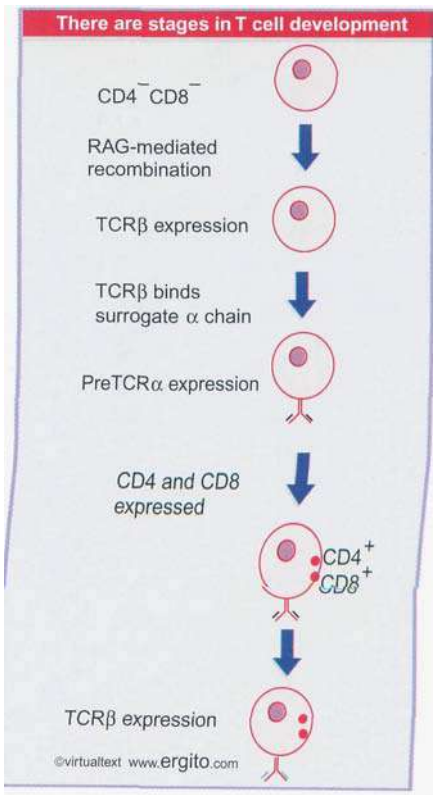


Figure 26.31 T cell development proceeds through sequential stages.

Recombination to generate functional TCR chains is linked to the development of the T lymphocyte, as summarized in **Figure 26.31.** The

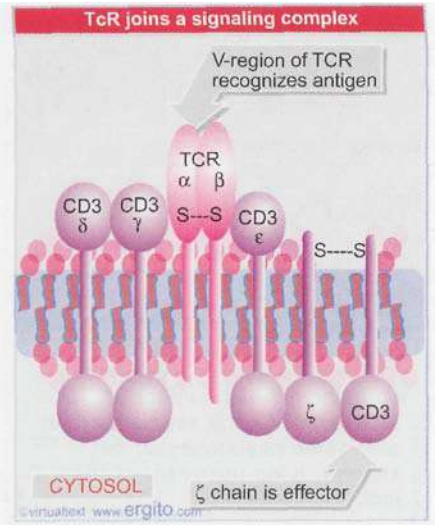


Figure 26.32 The two chains of the T-cell receptor associate with the polypeptides of the CD3 complex. The variable regions of the TCR are exposed on the cell surface. The cytoplasmic domains of the ζ chains of CD3 provide the effector function.

first stage is rearrangement to form an active TCR β chain. This binds a nonrearranging surrogate a chain, called preTCRα. At this stage, the lymphocyte has not expressed either of the surface proteins CD4 or CD8. The preTCR heterodimer then binds to the CD3 signaling complex (see below). Signaling from the complex triggers several rounds of cell division, during which a chains are rearranged, and the CD4 and CD8 genes are turned on, so that the lymphocyte is converted from $CD4^-CD8^-$ to $CD4^+CD8^+$. This point in development is called *ft* selection. It generates DP thymocytes.

α chain rearrangement continues in the DP thymocytes. The maturation process continues by both positive selection (for mature TCR complexes able to bind a ligand) and by negative selection (against complexes that interact with inappropriate—self—ligands). Both types of selection involve interaction with MHC proteins. The DP thymocytes either die after ~3 days or become mature lymphocytes as the result of the selective processes. The surface TCR *aft* heterodimer becomes crosslinked on the surface during positive selection (which rescues the thymocytes from cell death), and then, if they survive the subsequent negative selection, allows them to give rise to the separate T lymphocyte classes which are $CD4^+CD8^-$ and $CD4^-CD8^+$.

The T cell receptor is associated with a complex of proteins called CD3, which is involved in transmitting a signal from the surface of the cell to the interior when its associated receptor is activated by binding antigen. Our present picture of the components of the receptor complex on a T cell is illustrated in **Figure 26.32.** The important point is that the interaction of the TCR variable regions with antigen causes the ζ subunits of the CD3 complex to activate the T cell response. The activation of CD3 provides the means by which either α*ft* or 78 TCR signals that it has recognized an antigen. This is comparable to the constitution of the B cell receptor, in which immunoglobulin associates with the Igαβ signaling chains (see Figure 26.26).

A central dilemma about T cell function remains to be resolved. Cell-mediated immunity requires two recognition processes. Recognition of the foreign antigen requires the ability to respond to novel structures. Recognition of the MHC protein is of course restricted to one of those coded by the genome, but, even so, there are many different MHC proteins. So considerable diversity is required in both recognition reactions. Although helper and killer T cells rely upon different classes of MHC proteins, they use the same pool of a and *ft* gene segments to assemble their receptors. Even allowing for the introduction of additional variation during the TCR recombination process, it is not clear how enough different versions of the T cell receptor are made available to accommodate all these demands.

26.20 The major histocompatibility locus codes for many genes of the immune system

Key Concepts

- The MHC locus codes for the class I and class II proteins as well as for other proteins of the immune system.
- Class I proteins are the transplantation antigens that are responsible for distinguishing "self" from "nonself" tissue.
- An MHC class I protein is active as a heterodimer with β_2 microglobulin.
- Class II proteins are involved in interactions between T cells.
- An MHC class II protein is a heterodimer of α and β chains.

The major histocompatibility locus occupies a small segment of a single chromosome in the mouse (where it is called the H2 locus) and in man (called the HLA locus). Within this segment are many genes coding for functions concerned with the immune response. At those individual gene loci whose products have been identified, many alleles have been found in the population; the locus is described as highly *polymorphic,* meaning that individual genomes are likely to be different from one another. Genes coding for certain other functions also are located *in this region*.

Histocompatibility antigens are classified into three types by their immunological properties. In addition, other proteins found on lymphocytes and macrophages have a related structure and are important in the function of cells of the immune system:

MHC class I proteins are the transplantation antigens. They are present on every cell of the mammal. As their name suggests, these proteins are responsible for the rejection of foreign tissue, which is recognized as such by virtue of its particular array of transplantation antigens. In the immune system, their presence on target cells is required for the cell-mediated response. The types of class I proteins are defined serologically (by their antigenic properties). The murine class I genes code for the H2-K and H2-D/L proteins. Each mouse strain has one of several possible alleles for each of these functions. The human class I functions include the classical transplantation antigens, HLA-A, ***B*** ***n***

MHC class II proteins are found on the surfaces of both B and T lymphocytes as well as macrophages. These proteins are involved in communications between cells that are necessary to execute the immune response; in particular, they are required for helper T cell function. The murine class II functions are defined genetically as I-A and I-E. The human class II region (also called HLA-D) is arranged into four subregions: DR, DQ, DZ/DO, and DP.

The complement proteins are coded by a genetic locus that is also known as the S region; S stands for serum, indicating that the proteins are components of the serum. Their role is to interact with antibody-antigen complexes to cause the lysis of cells in the classical pathway of the humoral response.

The *Qα* and *Tla* loci proteins are found on murine hematopoietic cells. They are known as differentiation antigens, because each is found only on a particular subset of the blood cells, presumably related to their function. They are structurally related to the class I H2 proteins, and like them, are polymorphic.

We can now relate the types of proteins to the organization of the genes that code for them. The MHC region was originally defined by genetics in the mouse, where the classical H2 region occupies 0.3 map units. Together with the adjacent region where mutations affecting immune function are also found, this corresponds to a region of ~2000 kb of DNA. The MHC region has been completely sequenced in several mammals, and also some birds and fish. By comparing these sequences, we find that the organization has been generally conserved.

The gene organization in mouse and Man is summarized in **Figure** 26.33. The genomic regions where the class I and class II genes are located mark the original boundaries of the locus (going in the direction from telomere to centromere, right to left as shown in the figure). The genes in the region that separates the class I and class II genes code for a variety of functions; this is called the class III region. Defining the ends of the locus varies with the species, and the region beyond the class I genes on the telomeric side is called the extended class I region. Similarly, the region beyond the class II genes on the centromeric side is called the extended class II region. The major difference between

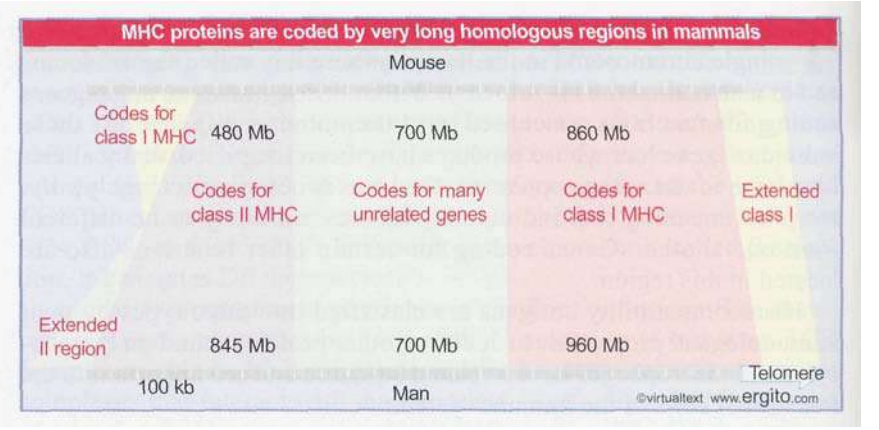


Figure 26.33 **The** MHC region extends for >2 Mb. MHC proteins of classes I and II are coded by two separate regions. The class III region is defined as **the** segment between them. The extended regions describe segments that are syntenic on either end of the cluster. The major difference between mouse and Man is **the** presence of H2 class I genes in the extended region on the left. The murine locus is located on chromosome 17, **and** the human locus on chromosome 6.

mouse and Man is that the extended class II region contains some class I (H2-K) genes in the mouse.

There are several hundred genes in the MHC regions of mammals, but it is possible for MHC functions to be provided by far fewer genes, as in the case of the chicken, where the MHC region is 92 Kb and has only 9 genes.

As in comparisons of other gene families, we find differences in the exact numbers of genes devoted to each function. Because the MHC locus shows extensive variation between individuals, the number of genes may be different in different individuals. As a general rule, however, a mouse genome has fewer active H2 genes than a human genome. The class II genes are unique to mammals (except for one subgroup), and as a rule, birds and fish have different genes in their place. There are ~8 functional class I genes in Man and ~30 in the mouse. The class I region also includes many other genes. The class III regions are closely similar in Man and mouse.

All MHC proteins are dimers located in the plasma membrane, with a major part of the protein protruding on the extracellular side. The structures of class I and class II MHC proteins are related, although they have different components, as summarized in **Figure 26.34.**

Class II antigens consist of two chains, a and β, whose combination generates an overall structure in which there are two extracellular domains.

All class I MHC proteins consist of a dimer between the class I chain itself and the β2-microglobulin protein. The class I chain is a 45 kD transmembrane component that has three external domains (each ~90 amino acids long, one of which interacts with β2 microglobulin), a transmembrane region of ~40 residues, and a short cytoplasmic domain of ~30 residues that resides within the cell.

The β2 microglobulin is a secreted protein of 12 kD. It is needed for the class I chain to be transported to the cell surface. Mice that lack the β2 microglobulin gene have no MHC class I antigen on the cell surface.

The organization of class I genes summarized in **Figure 26.35** coincides with the protein structure. The first exon codes for a signal sequence (cleaved from the protein during membrane passage). The next three exons code for each of the external domains. The fifth exon codes for the transmembrane domain. And the last three rather small exons together code for the cytoplasmic domain. The only difference in the genes for human transplantation antigens is that their cytoplasmic domain is coded by only two exons.

The exon coding for the third external domain of the class I genes is highly conserved relative to the other exons. The conserved domain probably represents the region that interacts with β2 microglobulin, which explains the need for constancy of structure. This domain also exhibits homologies with the constant region domains of immunoglobulins.

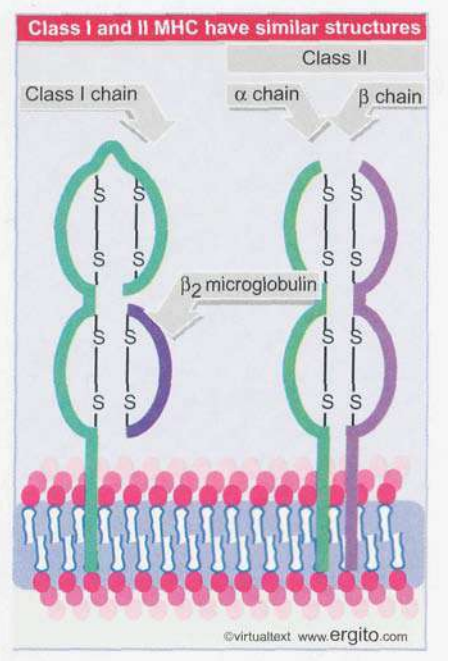


Figure 26.34 Class I and class II histocompatibility antigens have a related structure. Class I antigens consist of a single (α) polypeptide, with three external domains (α1, α2, α3), that interacts with β2 microglobulin (β2 m). Class II antigens consist of two (α and β) polypeptides, each with two domains (α1 & α2, β1 & β2) with a similar overall structure.

By Book_Crazy [IND]

What is responsible for generating the high degree of polymorphism in these genes? Most of the sequence variation between alleles occurs in the first and second external domains, sometimes taking the form of a cluster of base substitutions in a small region. One mechanism involved in their generation is gene conversion between class I genes. Pseudogenes are present as well as functional genes.

The gene for β2 microglobulin is located on a separate chromosome. It has four exons, the first coding for a signal sequence, the second for the bulk of the protein (from amino acids 3 to 95), the third for the last four amino acids and some of the nontranslated trailer, and the last for the rest of the trailer.

The length of β2 microglobulin is similar to that of an immunoglobulin V gene; there are certain similarities in amino acid constitution; and there are some (limited) homologies of nucleotide sequence between β2 microglobulin and Ig constant domains or type I gene third external domains. All the groups of genes that we have discussed in this chapter may have descended from a common ancestor that coded for a primitive domain.

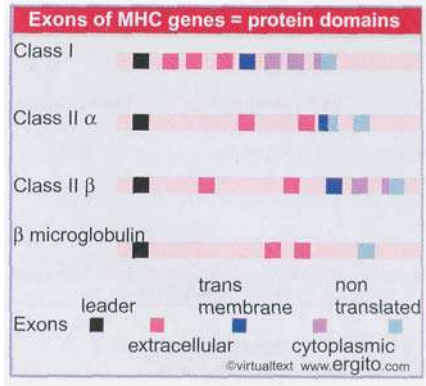


Figure 26.35 Each class of MHC genes has a characteristic organization, in which exons represent individual protein domains.

26.21 Innate immunity utilizes conserved signaling pathways

Key Concepts

- Innate immunity is triggered by receptors that recognize motifs (PAMPs) that are highly conserved in bacteria or other infective agents.
- Toll-like receptors are commonly used to activate the response pathway.
- The pathways are highly conserved from invertebrates to vertebrates, and an analogous pathway is found in plants.

The immune response described in this chapter comprises a set of reactions that respond to a pathogen by selecting lymphocytes (B cells or T cells) whose receptors (antibodies or TCR) have a high affinity for the pathogen. The basis for this selective process is the generation of a very large number of receptors so as to create a high possibility of recognizing a foreign molecule. Receptors that recognize the body's own proteins are screened out early in the process. Activation of the receptors on B cells triggers the pathways of the humoral response; activation of the receptors on T cells triggers the pathways of the cell-mediated response. The overall response to an antigen via selection of receptors on the lymphocytes is called adaptive immunity (or acquired immunity). The response typically is mounted over several days, following the initial activation of B cells or T cells that recognize the foreign pathogen. The organism retains a memory of the response, which enables it to respond more rapidly if it is exposed again to the same pathogen. The principles of adaptive immunity are similar through the vertebrate kingdoms, although details vary.

Another sort of immune response occurs more quickly and is found in a greater range of animals (including those that do not have adaptive immunity). Innate immunity depends on the recognition of certain, predefined patterns in foreign pathogens. These patterns are motifs that are conserved in the pathogens because they have an essential role in their junction, but they are not found in higher eukaryotes. The motif is typically recognized by a receptor dedicated to the purpose of triggering the innate response upon an infection. Receptors that trigger the innate

PAMPs are ubiquitous

Organism	Pathogen	Location
All bacteria	formyl-Methionine	Most proteins
Most bacteria	peptidoglycan	Cell wall
Gram-negative bacteria	lipopolysaccharide	Cell wall
Yeast	zymosan	Cell wall

Figure 26.36 Pathogen-associated molecular patterns (PAMPs) are compounds that are common to large ranges of bacteria or yeasts and are exposed when an infection occurs.

response are found on cells such as neutrophils and macrophages, and cause the pathogen to be phagocytosed and killed. The response is rapid, because the set of receptors is already present on the cells and does not have to be amplified by selection. It is widely conserved and is found in organisms ranging from flies to Man. When the innate response is able to deal effectively with an infection, the adaptive response will not be triggered. There is some overlap between the responses in that they activate some of the same pathways, so cells activated by the innate response may subsequently participate in the adaptive response.

The motifs that trigger innate immunity are sometimes called pathogen-associated molecular patterns (PAMPs). **Figure 26.36** shows that they are widely distributed across broad ranges of organisms. Formyl-methionine is used to initiate most bacterial proteins, but is not found in eukaryotes. The peptidoglycan of the cell wall is unique to bacteria. Lipopolysaccharide (LPS) is a component of the outer membrane of most gram-negative bacteria; also known as endotoxin, it is responsible for septic shock syndrome.

A key insight into the nature of innate immunity was the discovery of the involvement of Toll-like receptors (TLR). The receptor Toll, which is related to mammalian IL1 receptor, triggers the pathway in *Drosophila* that controls dorsal-ventral development (see *31.9 Dorsal-ventral development uses localized receptor-ligand interactions)*. This leads to activation of the transcription factor dorsal, a member of the Rel family, which is related to the mammalian factor NF-KB. The pathway of innate immunity is parallel to the Toll pathway, with similar components. In fact, one of the first indications of the nature of innate immunity in flies was the discovery of the transcription factor Dif (dorsal-related immunity factor), which is activated by one of the pathways.

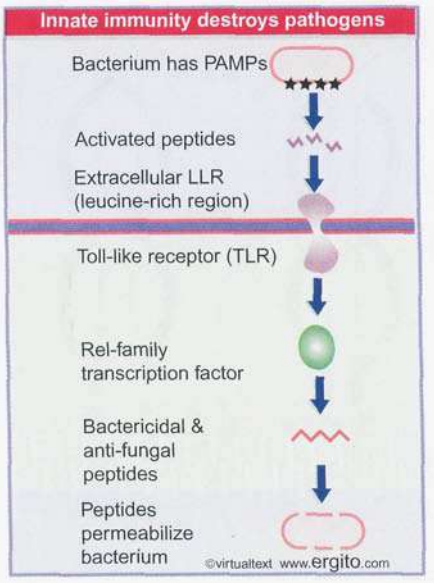


Figure 26.37 Innate immunity is triggered by PAMPs. In flies, they cause the production of peptides that activate Toll-like receptors. The receptors lead to a pathway that activates a transcription factor of the Rel family. Target genes for this factor include bactericidal and anti-fungal peptides. The peptides act by permeabilizing the membrane of the pathogenic organism.

Flies have no system of adaptive immunity, but are resistant to microbial infections. This is because their innate immune systems trigger synthesis of potent antimicrobial peptides. Seven distinct peptides have been identified in *Drosophila*, where they are synthesized in the fat body (the equivalent organ to the liver). Two of the peptides are antifungal, five act largely on bacteria. The general mode of action is to kill the target organism by permeabilizing its membrane. All of these peptides are coded by genes whose promoters respond to transcription factors of the Rel family. **Figure 26.37** summarizes the components of the innate pathways in *Drosophila*.

Two innate response pathways function in *Drosophila*, one responding principally to fungi, the other principally to gram-negative bacteria. Gram-positive bacteria may be able to trigger both pathways. **Figure 26.38** outlines the steps in each pathway. Fungi and gram-positive bacteria activate a proteolytic cascade that generates peptides that activate a Toll-like receptor. This is the NF-κB-like pathway. The dToll receptor activates the transcription factor Dif (a relative of NF-KB), leading ultimately to activation of the anti-fungal peptide drosomycin. Gram-negative bacteria trigger a pathway via a different receptor that activates the transcription factor Relish, leading to production of the bactericidal peptide attacin. This pathway is called the Imd pathway after one of its components, a protein that has a "death domain" related to those found in the pathways for apoptosis.

The key agents in responding to the bacteria are proteins called PGRPs because of their high affinities for bacterial peptidoglycans. There are two types of these proteins. PGRP-SAs are short extracellular proteins. They probably function by activating the proteases that trigger the Toll pathway. PGRP-LCs are transmembrane proteins with an extracellular PGRP domain. Their exact role has to be determined.

The innate immune response is highly conserved. Mice that are resistant to septic shock when they are treated with LPS have mutations in the Toll-like receptor TLR4. A human homologue of the Toll receptor can

activate some immune-response genes, suggesting that the pathway of innate immunity may also function in Man. The pathway downstream of the Toll-like receptors is generally similar in all cases, typically leading to activation of the transcription factor NF-KB. We do not yet know whether the upstream pathway is conserved, in particular whether the PAMPs function by generating ligands that in turn activate the Toll-like receptors or whether they might interact directly with them. The pathway upstream of the Toll-like receptors is different in mammals and flies, because the pathogens directly activate mammalian Toll-like receptors. In the case of LPS, the pathogen binds to the surface protein CD14; this enables CD14 to activate TLR4, triggering the innate response pathway. There are ~20 receptors in the TLR (Toll-like receptor) class in the human genome, which gives some indication of how many pathogens can trigger the innate response.

Plants have extensive defense mechanisms, among which are pathways analogous to the innate response in animals. The same principle applies that PAMPs are the motifs that identify the infecting agent as a pathogen. The proteins that respond to the pathogens are coded by a class of genes called the disease resistance genes. Many of these genes code for receptors that share a property with the TLR class of animal receptors: the extracellular domain has a motif called the leucine-rich region (LLR). The response mechanism is different from animal cells and is directed to activating a MAPK cascade. Many different pathogens activate the same cascade, which suggests that a variety of pathogen-receptor interactions converge at or before the activation of the first MAPK.

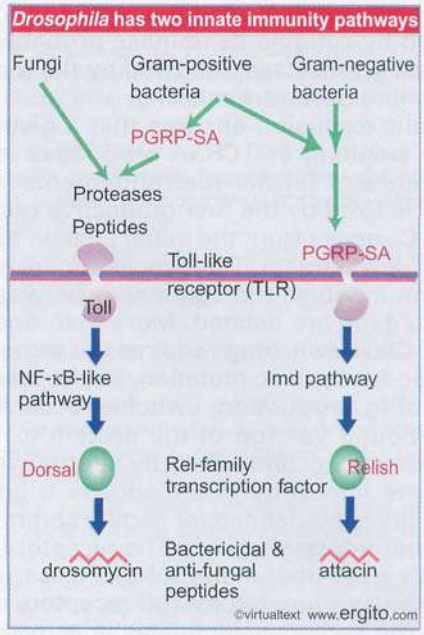


Figure 26.38 One of the *Drosophila* innate immunity pathways is closely related to the mammalian pathway for activating NF-kB; the other has components related to those of apoptosis pathways.

26.22 Summary

Immunoglobulins and T cell receptors are proteins that play analogous functions in the roles of B cells and T cells in the immune system. An Ig or TCR protein is generated by rearrangement of DNA in a single lymphocyte; exposure to an antigen recognized by the Ig or TCR leads to clonal expansion to generate many cells which have the same specificity as the original cell. Many different rearrangements occur early in the development of the immune system, creating a large repertoire of cells of different specificities.

Each immunoglobulin protein is a tetramer containing two identical light chains and two identical heavy chains. A TCR is a dimer containing two different chains. Each polypeptide chain is expressed from a gene created by linking one of many V segments via D and J segments to one of a few C segments. Ig L chains (either K or *X)* have the general structure V-J-C, Ig H chains have the structure V-D-J-C, TCR α and γ have components like Ig L chains, and TCR 8 and β are like Ig H chains.

Each type of chain is coded by a large cluster of V genes separated from the cluster of D, J, and C segments. The numbers of each type of segment, and their organization, are different for each type of chain, but the principle and mechanism of recombination appear to be the same. The same nonamer and heptamer consensus sequences are involved in each recombination; the reaction always involves joining of a consensus with 23 bp spacing to a consensus with 12 bp spacing. The cleavage reaction is catalyzed by the RAG1 and RAG2 proteins, and the joining reaction is catalyzed by the same NHEJ pathway that repairs double-strand breaks in cells. The mechanism of action of the RAG proteins is related to the action of site-specific recombination catalyzed by resolvases.

Although considerable diversity is generated by joining different V, D, J segments to a C segment, additional variations are introduced in the form of changes at the junctions between segments during the recombination process. Changes are also induced in immunoglobulin genes by somatic mutation, which requires the

actions of cytidine deaminase and uracil glycosylase. Mutations induced by cytidine deaminase probably lead to removal of uracil by uracil glycosylase, followed by the induction of mutations at the sites where bases are missing.

Allelic exclusion ensures that a given lymphocyte synthesizes only a single Ig or TCR. A productive rearrangement inhibits the occurrence of further rearrangements. Although the use of the V region is fixed by the first productive rearrangement, B cells switch use of C_H genes from the initial μ chain to one of the H chains coded farther downstream. This process involves a different type of recombination in which the sequences between the VDJ region and the new C_H gene are deleted. More than one switch occurs in C_H gene usage. Class switching requires the same cytidine deaminase that is required for somatic mutation, but its role is not known. At an earlier stage of Ig production, switches occur from synthesis of a membrane-bound version of the protein to a secreted version. These switches are accomplished by alternative splicing of the transcript.

Innate immunity is a response triggered by receptors whose specificity is predefined for certain common motifs found in bacteria and other infective agents. The receptor that triggers the pathway is typically a member of the Toll-like class, and the pathway resembles the pathway triggered by Toll receptors during embryonic development. The pathway culminates in activation of transcription factors that cause genes to be expressed whose products inactivate the infective agent, typically by permeabilizing its membrane.

References

26.3 Immunoglobulin genes are assembled from their parts in lymphocytes

rev Alt, F. W., Blackwell, T. K., and Yancopoulos, G. D. (1987). Development of the primary antibody repertoire. Science 238, 1079-1087.
Blackwell, T. K. and Alt, F. W. (1989). Mechanism and developmental program of immunoglobulin gene rearrangement in mammals. Ann. Rev. Genet. 23, 605-636.
Hood, L, Kronenberg, M., and Hunkapiller, T. (1985). T cell antigen receptors and the immunoglobulin supergene family. Cell 40, 225-229.
Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature 302, 575-581.
Yancopoulos, G. D. and Alt, F. W. (1986). Regulation of the assembly and expression of variable-region genes. Ann. Rev. Immunol. 4, 339-68.

ref Hozumi, N. and Tonegawa, S. (1976). Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc. Nat. Acad. Sci. USA 73, 3628-3632.

26.4 Light chains are assembled by a single recombination

ref Max, E. E., Seidman, J. G., and Leder, P. (1979). Sequences of five potential recombination sites encoded close to an immunoglobulin κ constant region gene. Proc. Nat. Acad. Sci. USA 76, 3450-3454.

26.7 Immune recombination uses two types of consensus sequence

ref Lewis, S., Gifford, A., and Baltimore, D. (1985). DNA elements are asymmetrically joined during the site-specific recombination of kappa immunoglobulin genes. Science 228, 677-685.

26.9 The RAG proteins catalyze breakage and reunion

exp Schatz, D. (2002). Identification of the V{D}J Recombination Activating Genes, (www.ergito.com/lookup.jsp?expt = schatz)

rev Gellert, M. (1992). Molecular analysis of VDJ recombination. Ann. Rev. Genet. 26, 425-446.
Jeggo, P. A. (1998). DNA breakage and repair. Adv. Genet. 38, 185-218.
Schatz, D. G., Oettinger, M. A., and Schlissel, M. S. (1992). VDJ recombination: molecular biology and regulation. Ann. Rev. Immunol. 10, 359-383.

ref Agrawal, A., Eastman, Q. M., and Schatz, D. G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394, 744-751.
Hiom, K., Melek, M., and Gellert, M. (1998). DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94, 463-470.
Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M. R. (2002). Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. Cell 108, 781-794.
Melek, M. and Gellert, M. (2000). RAG1/2-mediated resolution of transposition intermediates: two pathways and possible consequences. Cell 101, 625-633.
Qiu, J. X., Kale, S. B., Yarnell Schultz, H., and Roth, D. B. (2001). Separation-of-function mutants reveal critical roles for RAG2 in both the cleavage and joining steps of V(D)J recombination. Mol. Cell 7, 77-87.
Roth, D. B., Menetski, J. P., Nakajima, P. B., Bosma, M. J., and Gellert, M. (1992). V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in SCID mouse thymocytes. Cell 70, 983-991.
Schatz, D. G. and Baltimore, D. (1988). Stable expression of immunoglobulin gene V(D)J recombinase activity by gene transfer into 3T3 fibroblasts. Cell 53, 107-115.

By Book_Crazy [IND]

Schatz, D. G., Oettinger, M. A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. Cell 59, 1035-1048.

Tsai, C. L, Drejer, A. H., and Drejer, A. H. (2002). Evidence of a critical architectural function for the RAG proteins in end processing, protection, and joining in V(D)J recombination. Genes Dev. 16, 1934-1949.

Yarnell Schultz, H., Landree, M. A., Qiu, J. X., Kale, S. B., and Roth, D. B. (2001). Joining-deficient RAG1 mutants block V(D)J recombination *in vitro* and hairpin opening *in vitro*. Mol. Cell 7, 65-75.

26.10 Allelic exclusion is triggered by productive rearrangement

rev Storb, U. (1987). Transgenic mice with immunoglobulin genes. Ann. Rev. Immunol. 5, 151-174.

26.12 Switching occurs by a novel recombination reaction

rev Honjo, T., Kinoshita, K., and Muramatsu, M. (2002). Molecular mechanism of class switch recombination: linkage with somatic hypermutation. Ann. Rev. Immunol. 20, 165-196.

ref Gu, H., Zou, Y. R., and Rajewsky, K. (1993). Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting. Cell 73, 1155-1164.

Iwasato, T., Shimizu, A., Honjo, T., and Yamagishi, H. (1990). Circular DNA is excised by immunoglobulin class switch recombination. Cell 62, 143-149.

Kinoshita, K., Tashiro, J., Tomita, S., Lee, C. G., and Honjo, T. (1998). Target specificity of immunoglobulin class switch recombination is not determined by nucleotide sequences of S regions. Immunity 9, 849-858.

Manis, J. P., Gu, Y., Lansford, R., Sonoda, E., Ferrini, R., Davidson, L., Rajewsky, K., and Alt, F. W. (1998). Ku70 is required for late B cell development and immunoglobulin heavy chain class switching. J. Exp. Med. 187, 2081-2089.

Matsuoka, M., Yoshida, K., Maeda, T., Usuda, S., and Sakano, H. (1990). Switch circular DNA formed in cytokine-treated mouse splenocytes: evidence for intramolecular DNA deletion in immunoglobulin class switching. Cell 62, 135-142.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 102, 553-563.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, 0., Catalan, N., Forveille, M., Dufourcq-Labelouse, R., Gennery, A., Tezcan, I., Ersoy, F., Kayserili, H., Ugazio, A. G., Brousse, N., Muramatsu, M., Notarangelo, L. D., Kinoshita, K., and Honjo, T. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell 102, 565-575.

Rolink, A., Melchers, F., and Andersson, J. (1996). The SCID but not the RAG-2 gene product is required for S mu-S epsilon heavy chain class switching. Immunity 5, 319-330.

von Schwedler, U., Jack, H. M., and Wabl, M. (1990). Circular DNA is a product of the immunoglobulin class switch rearrangement. Nature 345, 452-456.

Xu, L., Gorham, B., Li, S. C, Bottaro, A., Alt, F. W., and Rothman, P. (1993). Replacement of germ-line epsilon promoter by gene targeting alters control of immunoglobulin heavy chain class switching. Proc. Nat. Acad. Sci. USA 90, 3705-3709.

26.14 Somatic mutation generates additional diversity in mouse and man

rev French, D. L., Laskov, R., and Scharff, M. D. (1989). The role of somatic hypermutation in the generation of antibody diversity. Science 244, 1152-1157.

Kocks, C. and Rajewsky, K. (1989). Stable expression and somatic hypermutation of antibody V regions in B-cell developmental pathways. Ann. Rev. Immunol. 7, 537-559.

ref Kim, S., Davis, M., Sinn, E., Patten, P., and Hood, L. (1981). Antibody diversity: somatic hypermutation of rearranged VH genes. Cell 27, 573-581.

26.15 Somatic mutation is induced by cytidine deaminase and uracil glycosylase

rev Honjo, T., Kinoshita, K., and Muramatsu, M. (2002). Molecular mechanism of class switch recombination: linkage with somatic hypermutation. Ann. Rev. Immunol. 20, 165-196.

Kinoshita, K. and Honjo, T. (2001). Linking class-switch recombination with somatic hypermutation. Nat. Rev. Mol. Cell Biol. 2, 493-503.

ref Di Noia, J. and Neuberger, J. (2002). Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature 419, 43-48.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 102, 553-563.

Peters, A. and Storb, U. (1996). Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. Immunity 4, 57-65.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, 0., Catalan, N., Forveille, M., Dufourcq-Labelouse, R., Gennery, A., Tezcan, I., Ersoy, F., Kayserili, H., Ugazio, A. G., Brousse, N., Muramatsu, M., Notarangelo, L. D., Kinoshita, K., and Honjo, T. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell 102, 565-575.

26.16 Avian immunoglobulins are assembled from pseudogenes

ref Reynaud, C. A., Anquez, V., Grimal, H., and Weill, J. C. (1987). A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell 48, 379-388.

Sale, J. E., Calandrini, D. M., Takata, M., Takeda, S., and Neuberger, M. S. (2001). Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation. Nature 412, 921-926.

26.17 B cell memory allows a rapid secondary response

rev Rajewsky, K. (1996). Clonal selection and learning in the antibody system. Nature 381, 751-758.

26.18 T cell receptors are related to immunoglobulins

rev Davis, M. M. (1990). T-cell receptor gene diversity and selection. Ann. Rev. Biochem. 59, 475-496.

Kronenberg, M., Siu, G., Hood, L. E., and Shastri, N. (1986). The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. Ann. Rev. Immunol. 4, 529-591.

Marrack, P. and Kappler, J. (1987). The T-cell receptor. Science 238, 1073-1079.

Raulet, D. H. (1989). The structure, function, and molecular genetics of the gamma/delta T-cell receptor. Ann. Rev. Immunol. 7, 175-207.

26.19 The T cell receptor functions in conjunction with the MHC

rev Goldrath, A. W. and Bevan, M. J. (1999). Selecting and maintaining a diverse T cell repertoire. Nature 402, 255-262.

26.20 The major histocompatibility locus codes for many genes of the immune system

rev Flavell, R. A., Allen, H., Burkly, L. C, Sherman, D. H., Waneck, G. L, and Widera, G. (1986). Molecular biology of the H-2 histocompatibility complex. Science 233, 437-443.

Kumnovics, A. , Takada, T. , and Lindahl, K. F. (2003). Genomic organization of the Mammalian MHC. Ann Rev. Immunol. 21, 629-657.

Steinmetz, M. and Hood, L. (1983). Genes of the MHC complex in mouse and man. Science 222, 727-732.

ref Kaufman, J. et al. (1999). The chicken B locus is a minimal essential major histocompatibility complex. Nature 401, 923-925.

26.21 Innate immunity utilizes conserved signaling pathways

rev Aderem, A. and Ulevitch, R. J. (2000). Toll-like receptors in the induction of the innate immune response. Nature 406, 782-787.

Dangl, J. L. and Jones, J. D. (2001). Plant pathogens and integrated defence responses to infection. Nature 41 1, 826-833.

Hoffmann, J. A., Kafatos, F. C, Janeway, C. A., and Ezekowitz, R. A. (1999). Phylogenetic perspectives in innate immunity. Science 284, 1313-1318.

Janeway, C. A. and Medzhitov, R. (2002). Innate immune recognition. Ann. Rev. Immunol. 20, 197-216.

ref Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W. L, Gomez-Gomez, L, Boiler, T., Ausubel, F. M., and Sheen, J. (2002). MAP kinase signalling cascade in Arabidopsis innate immunity. Nature 415, 977-983.

Ip, Y. T., Reach, M., Engstrom, Y., Kadalayil, L., Cai, H., GonzAilez-Crespo, S., Tatei, K., and Levine, M. (1993). Dif, a dorsal-related gene that mediates an immune response in *Drosophila.* Cell 75, 753-763.

Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M., and Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette spätzle/Toll/cactus controls the potent antifungal response in *Drosophila* adults. Cell 86, 973-983.

Medzhitov, R., Preston-Hurlburt, P., and Janeway, C. A. (1997). A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. Nature 388, 394-397.

Poltorak, A., He, X., Smirnova, I., Liu, M. Y., Huffel, C. V., Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C, Freudenberg, M., Ricciardi-Castagnoli, P., Layton, B., and Beutler, B. (1998). Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. Science 282, 2085-2088.

Rutschmann, S., Jung, A. C, Hetru, C, Reichhart, J. M., Hoffmann, J. A., and Ferrandon, D. (2000). The Rel protein DIF mediates the antifungal but not the antibacterial host defense in *Drosophila.* Immunity 12, 569-580.

Williams, M. J., Rodriguez, A., Kimbrell, D. A., and Eldon, E. D. (1997). The 18-wheeler mutation reveals complex antibacterial gene regulation in *Drosophila* host defense. EMBO J. 16, 6120-6130.

Chapter 27

Protein trafficking

27.1 Introduction

A great variety of molecules move out of and into the cell. At one extreme of the size range, proteins may be secreted from the cell into the extracellular fluid or may be internalized from the cell surface. At the other extreme, ions such as K^+, Na^+, and Ca^{2+} may be pumped out of or into the cell. In this chapter, we are concerned with the processes by which proteins are physically transported through membranous systems to the plasma membrane or other organelles, or from the cell surface to organelles within the cell. In *28 Signal transduction,* we discuss the pathways by which an interaction at the surface can trigger internal pathways.

Proteins enter the pathway that leads to secretion by co-translational transfer to the membranes of the endoplasmic reticulum. They are then transferred to the Golgi apparatus, where they are *sorted* according to their final intended destination. **Figure 27.1** summarizes the routes by which proteins are carried forward or diverted to other organelles. Their destinations are determined by specific sorting signals, which take the form of short sequences of amino acids or covalent modifications that are made to the protein.

The transport machinery consists of small membranous vesicles. A soluble protein is carried within the lumen of a vesicle, and an integral membrane protein is carried within its membrane. **Figure 27.2** illustrates the nature of the budding and fusion events by which the vesicles move between adjacent compartments. A vesicle buds off from a donor surface and then fuses with a target surface. Its proteins are released into the lumen or into the membrane of the target compartment, depending on their nature, and must be loaded into new vesicles for transport to the next compartment. The series of events is repeated at each transition between membrane surfaces, for example, during passage from the ER to the Golgi, or between cisternae of the Golgi stacks.

Once a protein enters a membranous environment, it remains in the membrane until it reaches its final destination. A membrane protein that

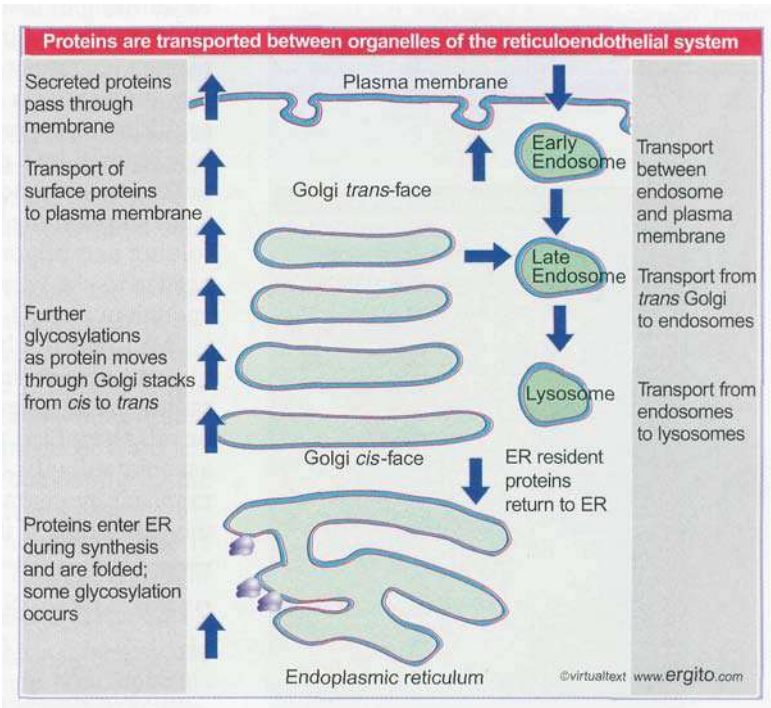


Figure 27.1 Proteins that enter the endoplasmic reticulum are transported to the Golgi and towards the plasma membrane. Specific signals cause proteins to be returned from the Golgi to the ER, to be retained in the Golgi, to be retained in the plasma membrane, or to be transported to endosomes and lysosomes. Proteins may be transported between the plasma membrane and endosomes.

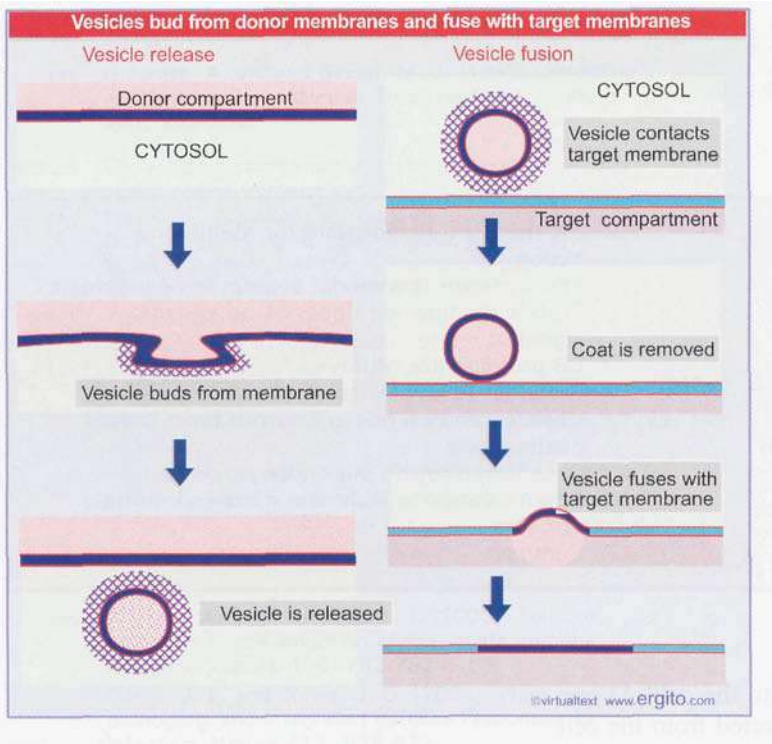


Figure 27.2 Vesicles are released when they bud from a donor compartment and are surrounded by coat proteins (left). During fusion, the coated vesicle binds to a target compartment, is uncoated, and fuses with the target membrane, releasing its contents (right).

enters the endoplasmic reticulum is inserted into the membrane with the appropriate orientation (N-terminal lumenal, C-terminal cytosolic for group I proteins, the reverse for group II proteins). The orientation is retained as it moves through the system. The process starts in the same way irrespective of whether the protein is destined to reside in the Golgi, lysosome or plasma membrane. In each case, it is transported in membrane vesicles along the secretory pathway to the appropriate destination, where some structural feature of the protein is recognized and it is permanently secured (or secreted from the cell).

Two important changes occur to a protein in the endoplasmic reticulum: it becomes folded into its proper conformation; and it is modified by glycosylation.

A protein is translocated into the ER in unfolded form. Folding occurs as the protein enters the lumen; probably a series of domains each folds independently as the protein passes through the membrane. Folding of a 50 kD protein is complete in <3-4 minutes, compared with the ~1 minute required to synthesize the chain.

Folding in the ER is associated with modification and is assisted by accessory proteins. Addition of carbohydrate may be required for correct folding; in fact, this may be an important function of the modification. Reshuffling of disulfide bonds by the enzyme PDI (protein disulfide isomerase) may be involved. And association with chaperones in the ER may be necessary to recognize the partially folded forms of proteins as they emerge from transport through the membrane, and to assist them in acquiring their proper conformation. Some or all of these activities could be exercised by a complex of enzymes as a protein enters the ER; that is, the necessary functions all could associate with the protein as it translocates through the membrane. Calculations of the spontaneous rates of folding and oligomerization suggest that these accessory activities are needed to catalyze the process in order for it to enable it to occur rapidly enough in the cell.

Multimeric glycoproteins usually oligomerize in the ER. In fact, oligomerization may be necessary for further transport. Oligomers are rapidly transported from the ER to the Golgi, but unassembled subunits or misassembled proteins are held back. Misfolded proteins are often associated with ER-specific chaperones. In due course, they are removed by degradation. So a protein is allowed to move forward into the Golgi only if it has been properly folded previously in the ER.

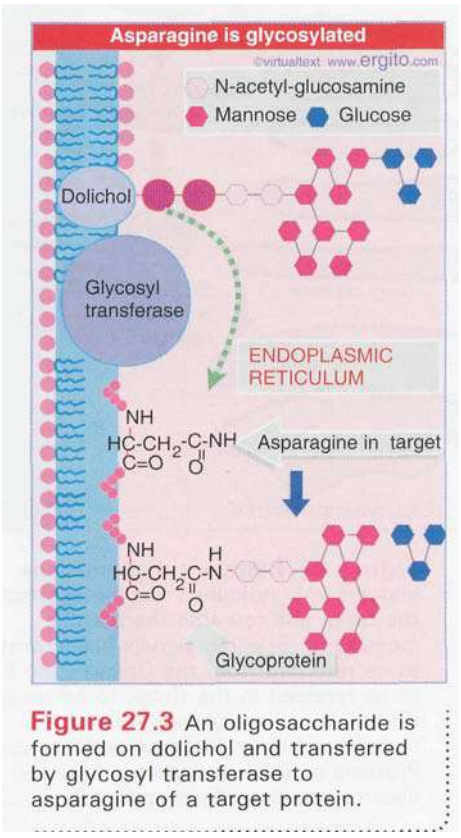


Figure 27.3 An oligosaccharide is formed on dolichol and transferred by glycosyl transferase to asparagine of a target protein.

27.2 Oligosaccharides are added to proteins in the ER and Golgi

Key Concepts

- A major function of the ER and Golgi is to glycosylate proteins as they pass through the system.
- N-linked oligosaccharides are initiated by transferring the saccharide from the lipid dolichol to an asparagine of the target protein in the ER.
- Sugars are trimmed in the ER to give a high mannose oligosaccharide.
- A complex oligosaccharide is generated in those cases in which further residues are added in the Golgi.

Virtually all proteins that pass through the secretory apparatus are glycosylated. Glycoproteins are generated by the addition of oligosaccharide groups either to the NH_2 group of asparagine (N-linked glycosylation) or to the OH group of serine, threonine, or hydroxylysine (O-linked glycosylation). N-linked glycosylation is initiated in the endoplasmic reticulum and completed in the Golgi; O-linked glycosylation occurs in the Golgi alone. The stages of N-glycosylation are illustrated in the next three figures.

The addition of all N-linked oligosaccharides starts in the ER by a common route, as illustrated in **Figure 27.3.** An oligosaccharide containing 2 N-acetyl glucosamine, 9 mannose, and 3 glucose residues is formed on a special lipid, dolichol. Dolichol is a highly hydrophobic lipid that resides within the ER membrane, with its active group facing the lumen. The oligosaccharide is constructed by adding sugar residues individually; it is linked to dolichol by a pyrophosphate group, and is transferred as a unit to a target protein by a membrane-bound glycosyl transferase enzyme whose active site is exposed in the lumen of the endoplasmic reticulum.

The acceptor group is an asparagine residue, located within the sequence Asn-X-Ser or Asn-X-Thr (where X is any amino acid except proline). It is recognized as soon as the target sequence is exposed in the lumen, when the nascent protein crosses the ER membrane.

Some trimming of the oligosaccharide occurs in the ER, after which a nascent glycoprotein is handed over to the Golgi. The oligosaccharide structures generated during transport through the ER and Golgi fall into two classes, determined by the fate of the mannose residues. Mannose residues are added only in the ER, although they can be trimmed subsequently:

- High mannose oligosaccharides are generated by trimming the sugar residues in the ER. **Figure 27.4** shows that almost immediately following addition of the oligosaccharide, the 3 glucose residues are removed by the enzymes glucosidases I and II. For proteins that reside in the ER, a mannosidase removes some of the mannose residues to generate the final structure of the oligosaccharide. The ER mannosidase attacks the first mannose quickly, and the next 3 more slowly; the total number of mannose residues that is removed varies with the individual substrate protein.
- Complex oligosaccharides result from additional trimming and further additions carried out in the Golgi. Golgi modifications occur in the fixed order illustrated in **Figure 27.5.** The first step is further trimming of mannose residues by Golgi mannosidase I. Then a single sugar residue is added by the enzyme N-acetyl-glucosamine transferase. Then Golgi mannosidase II removes further mannose residues. This generates a structure called the inner core, consisting of the sequence NAc-Glc·NAc-Glc·Man_3. At this point, the oligosaccharide becomes resistant to degradation by the enzyme endoglycosidase H (Endo H). *Susceptibility to Endo H is therefore used as an operational test to determine when a glycoprotein has left the ER.*

Additions to the inner core generate the terminal region. The residues that can be added to a complex oligosaccharide include N-acetyl-glucosamine, galactose, and sialic acid (N-acetyl-neuraminic acid). The pathway for processing and glycosylation is highly ordered, and the two types of reaction are interspersed in it. Addition of one sugar residue may be needed for removal of another, as in the example of the addition of N-acetyl-glucosamine before the final mannose residues are removed.

We do not know what determines how each protein undergoes its specific pattern of processing and glycosylation. We assume that the necessary information resides in the structure of the polypeptide chain; it cannot lie in the oligosaccharide, since all proteins subject to N-linked glycosylation start the pathway by addition of the same (preformed) oligosaccharide.

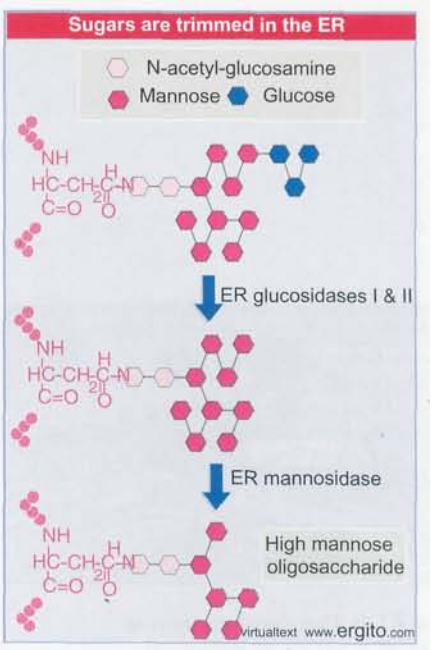


Figure 27.4 Sugars are removed in the ER in a fixed order, initially comprising 3 glucose and 1-4 mannose residues. This trimming generates a high mannose oligosaccharide.

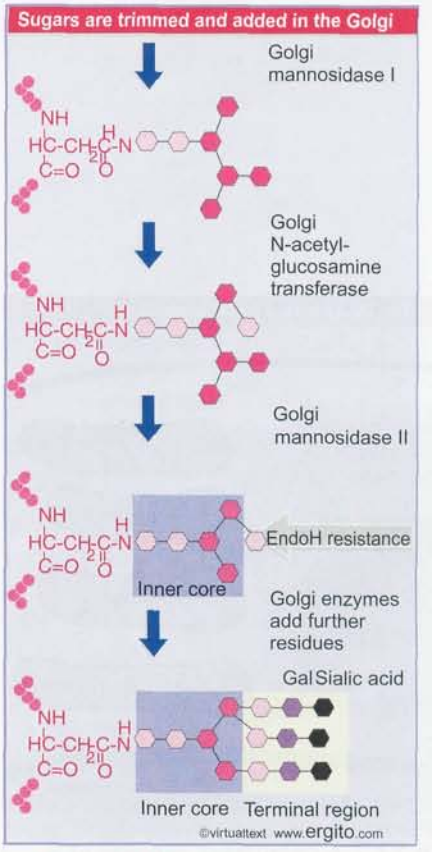


Figure 27.5 Processing for a complex oligosaccharide occurs in the Golgi and trims the original preformed unit to the inner core.

27.3 The Golgi stacks are polarized

Key Concepts

- The Golgi stacks change in lipid and protein constitution proceeding from the c/s-face near the ER to the *trans-face* near the plasma membrane.

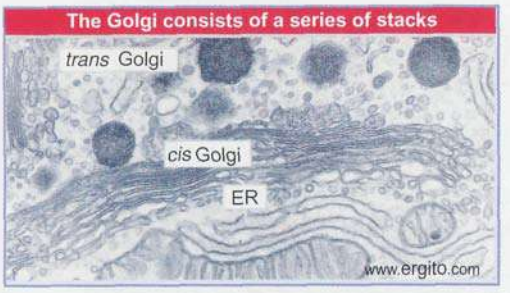


Figure 27.6 The Golgi apparatus consists of a series of individual membrane stacks. Photograph kindly provided by Alain Rambourg.

The individual cisternae of the Golgi are organized into a series of *stacks*, somewhat resembling a pile of plates. A typical stack consists of 4-8 flattened cisternae. Figure 27.6 shows an example. A major feature of the Golgi apparatus is its *polarity*. The *cis* side faces the endoplasmic reticulum; the *trans* side in a secretory cell faces the plasma membrane. The Golgi consists of compartments, which are named *cis*, *medial*, *trans*, and *TGN (trans-Golgi network)*, proceeding from the *cis* to the *trans* face. Proteins enter a Golgi stack at the *cis* face and are modified during their transport through the successive cisternae of the stack. When they reach the *trans* face, they are directed to their destination.

Membrane structure changes across the Golgi stack. The main difference is an increase in the content of cholesterol proceeding from *cis* to *trans*. As a result, fractionation of Golgi preparations generates a gradient in which the densest fractions represent the *cis* cisternae, and the lightest fractions represent the *trans* cisternae. The positions of enzymes on the gradient, and *in situ* immunochemistry with antibodies against individual enzymes, suggest that certain enzymes are differentially distributed proceeding from *cis* to *trans*. The difference between the *cis* and *trans* faces of the Golgi is clear, but it is not clear how the concept of compartments relates to individual cisternae; there may rather be a continuous series of changes proceeding through the cisternae.

Nascent proteins encounter the modifying enzymes as they are transported through the Golgi stack. Figure 27.7 illustrates the order in which the enzymes function. This may be partly determined by the fact that the modification introduced by one enzyme is needed to provide the substrate for the next, and partly by the availability of the enzymes proceeding through the cisternae.

The addition of a complex oligosaccharide can change the properties of a protein significantly. Glycoproteins often have a mass with a significant proportion of oligosaccharide. What is the significance of these extensive glycosylations? In some cases, the saccharide moieties play a structural role, for example, in the behavior of surface proteins that are involved in cell adhesion. Another possible role could be in promoting folding into a particular conformation. One modification—the addition of mannose-6-phosphate—confers a targeting signal.

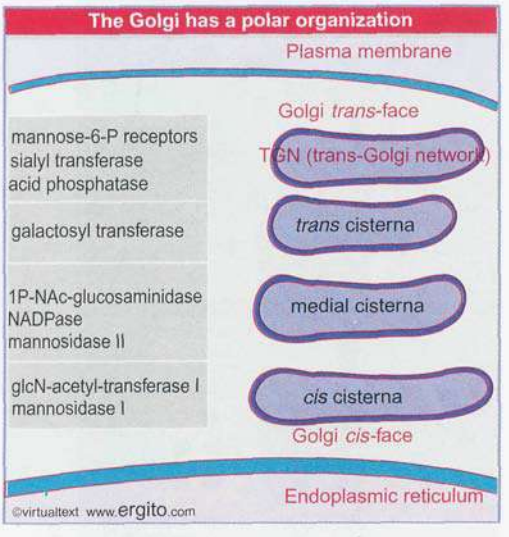


Figure 27.7 A Golgi stack consists of a series of cisternae, organized with *cis* to *trans* polarity. Protein modifications occur in order as a protein moves from the *cis* face to the *trans* face.

27.4 Coated vesicles transport both exported and imported proteins

Key Concepts

- Protein transport through the ER-Golgi system occurs in coated vesicles.
- Secreted proteins move in the forward, anterograde direction toward the plasma membrane.
- There is also movement within the system in the reverse, retrograde direction.
- Proteins that are imported through the plasma membrane also are incorporated in coated vesicles.
- Coated vesicles bud from a donor membrane surface and fuse with a target membrane surface.

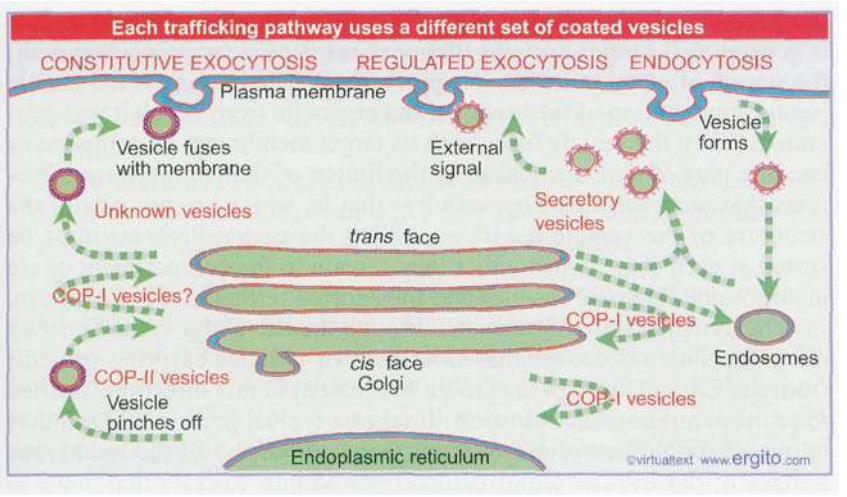


Figure 27.8 Proteins are transported in coated vesicles. Constitutive (bulk flow) transport from ER through the Golgi takes place by COP-coated vesicles. Clathrin-coated vesicles are used for both regulated exocytosis and endocytosis.

Secreted and transmembrane proteins start on the route to localization when they are translocated into the endoplasmic reticulum during synthesis. Transport from the ER, through the Golgi, to the plasma membrane occurs in vesicles. A protein is incorporated into a vesicle at one membrane surface, and is released from the vesicle at the next membrane surface. Progress through the system requires a series of such transport events. A protein changes its state of glycosylation as it passes through the Golgi from *cis* to *trans* compartments.

Vesicles are used to transport proteins both out of the cell and into the cell. The secretion of proteins is called exocytosis; internalization of proteins is called endocytosis. The pathways for vesicle movement are pictured in **Figure 27.8.** The cycle for each type of vesicle is similar, whether they are involved in export or import of proteins: *budding from the donor membrane is succeeded by fusion with the target membrane.*

Vesicles involved in transporting proteins have a protein layer surrounding their membranes, and for this reason are called coated vesicles. Examples are shown in the electron micrograph of **Figure 27.9.** Different types of vesicles are distinguished by the protein coats. The coat serves two purposes: it is involved with the processes of budding and fusion; and it enables the type of vesicle to be identified, so that it is directed to the appropriate target membrane. The coat may also play a role in the selection of proteins to be transported.

One of the most remarkable features of protein trafficking is the conservation of the vesicular apparatus, including structural components of the vesicles, and proteins required for budding or fusion. Many of these functions have been identified through mutations of the *sec* genotype in *S. cerevisiae,* which are unable to export proteins through the ER-Golgi pathway. Many of the genes identified by *sec* mutants in yeast have direct counterparts in animal cells. In particular, the proteins involved in budding, fusion, and targeting in mammalian brain (where release of proteins from the cell provides the means of propagating nerve impulses) have homologues in the yeast secretory pathway.

The process of generating a vesicle requires a membrane bilayer to protrude a vesicle that eventually pinches off as a bud (see Figure 27.2). Such events require deformation of the membrane, as illustrated in **Figure 27.10.** Proteins concerned with this process are required specifically for budding, and become part of the coat.

In the reverse reaction, fusion is a property of membrane surfaces. In order to fuse with a target membrane, a vesicle must be "uncoated" by removal of the protein layer. A coated vesicle recognizes its destination by a reaction between a protein in the vesicle membrane and a receptor in the target membrane.

A vesicle therefore follows a cycle in which it gains its coat, is released from a donor membrane, moves to the next membrane,

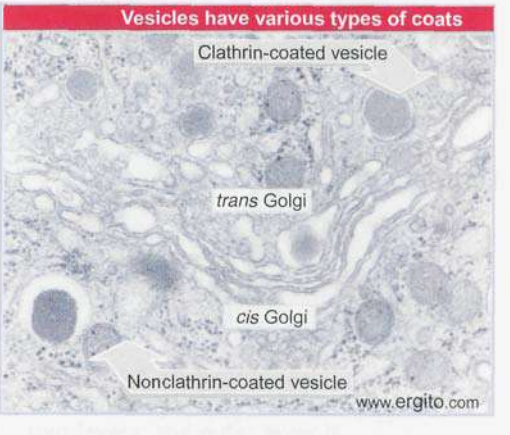


Figure 27.9 Coated vesicles are released from the *trans* face of the Golgi. The diameter of a vesicle is ~70 nm. Photograph kindly provided by Lelio Orci.

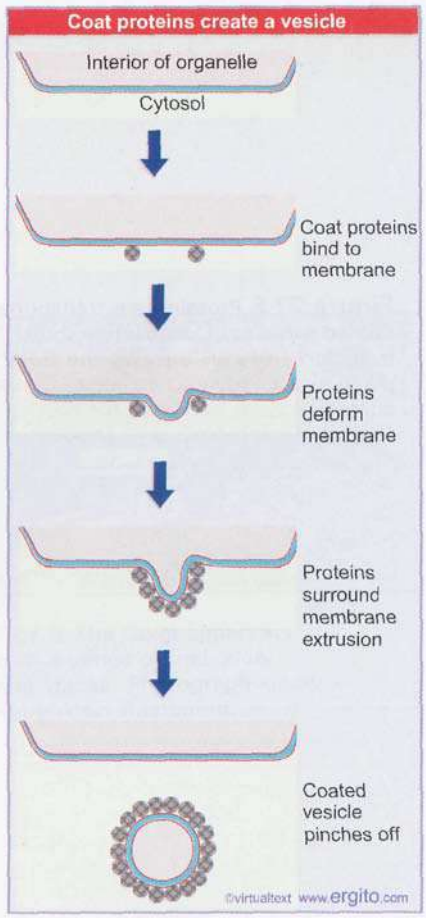


Figure 27.10 Vesicle formation results when coat proteins bind to a membrane, deform it, and ultimately surround a membrane vesicle that is pinched off.

becomes uncoated, and fuses with the target membrane. When a vesicle is generated, it carries proteins that were resident in (or associated with) the stretch of membrane that was pinched off. The interior of the vesicle has the constitution of the lumen of the organelle from which it was generated. When the vesicle fuses with its target membrane, its components become part of that membrane or the lumen of the compartment. Proteins that are transported by vesicles (that is, which are not part of the structure of the vesicle itself) are called the cargo. Proteins must be sorted at each stage, when either they remain in the compartment or are incorporated into new vesicles and transported farther along the system.

The dynamic state of transport through the ER-Golgi system poses a dilemma. There is continuous movement of vesicles carrying proteins from the ER and through the Golgi. Movement in this direction is called *forward* or anterograde transport. It takes a typical protein ~20 minutes to pass through the system. A significant proportion of the membrane surface of the ER and Golgi is incorporated into vesicles that move to the plasma membrane. Such a flow of membrane should rapidly denude the Golgi apparatus and enormously enlarge the plasma membrane, yet both are stable in size. The net amount (and types) of lipid in each membrane must remain unperturbed in spite of vesicle movement.

The need to maintain the structure of the reticuloendothelial system suggests that there is a pathway for returning membrane segments from the Golgi to ER, so that there is no net flow of membrane. Movement in this direction is called retrograde transport. We do not yet understand the balance of forward and retrograde flow. One possibility is that some vesicles engaged in retrograde movement do not carry cargo, except for returning components to earlier parts of the system. Alternatively, reverse flow might occur by structures that have a high surface to volume ratio, such as tubules, which could thus return large amounts of material.

27.5 Different types of coated vesicles exist in each pathway

Key Concepts

- Clathrin-coated vesicles are used for endocytosis.
- The clathrin coat forms when triskelions form a lattice at a coated pit on the plasma membrane.
- The clathrin is joined to the membrane by an AP adaptor, which is a heterotetramer of adaptin subunits.
- Different types of AP adaptors are found in vesicles used in different locations.
- COP-I coated vesicles have a coat consisting of a 7-component coatomer and are required for retrograde transport in the ER-Golgi system.
- COP-II coated vesicles have a coat consisting of a different protein complex.
- Cargo proteins may be bound by components of the vesicle coat.

Different groups of coated vesicles can be identified by the types of transport they undertake. In some cases, the vesicles can be distinguished by the biochemical components of their coats.

Newly synthesized proteins enter the ER and may be transported along the ER-Golgi system. This transport is undertaken by transition vesicles, and is common to all eukaryotic cells. Two different types of transition vesicles have been identified on the basis of their coats:

- The vesicles that were originally identified in transport between Golgi cisternae are now called COP-I-coated vesicles. (COP is an acronym

for coat protein.) They are involved in retrograde transport. It is an open question whether they also undertake anterograde transport.

- Transition vesicles that proceed from the ER to the Golgi have a different coat, called COP-II. Their major role appears to be forward transport.
- Some proteins are constitutively secreted, moving from the *trans*-Golgi to the plasma membrane. The vesicles that undertake this process have not been characterized biochemically.

The export of some proteins is regulated. These proteins are packaged into secretory vesicles. These vesicles provide a storage medium, and release their contents at the plasma membrane only following receipt of a particular signal (triggered, for example, by a hormone or Ca^{2+}). This occurs in cells that are specialized to produce the appropriate proteins. Vesicles that form at the *trans-face* of the Golgi for use in the regulated pathway may fuse to form secretory granules. They may also transport their cargoes to endosomes. Secretory vesicles may form at endosomes to transport proteins to the plasma membrane. The most common route for regulated transport is probably via the endosome.

Proteins enter the cell by packaging into endocytic vesicles, which are released from the plasma membrane, and transport their contents toward the interior of the cell. The cargo is released when the endocytic vesicle fuses with the membrane of a target compartment such as an endosome.

What controls the specificity of the cargo carried by a vesicle? It is necessary to distinguish those proteins that should be transported out of the compartment from resident proteins that should remain there. This may be a function of the coat. The COP-II coat can cause vesicles to bud when liposomes are mixed with the coat proteins. When the liposomes contain membrane proteins that are resident in the ER and other membrane proteins that are involved in targeting vesicles to the Golgi, only the latter class enters the vesicles. This suggests that specificity may be determined by a direct interaction with the coat proteins.

Endocytic and secretory vesicles have clathrin as the most prominent protein in their coats, and are therefore known as clathrin-coated vesicles. Their structure is known in some detail. The 180 kD chain of clathrin, together with a smaller chain of 35 kD, forms a polyhedral coat on the surface of the coated vesicle. The subunit of the coat consists of a triskelion, a three-pronged protein complex consisting of 3 light and 3 heavy chains. The triskelions form a lattice-like network on the surface of the coated vesicle, as revealed in the electron micrograph of **Figure 27.11.** Endocytic vesicles form and are coated at invaginations of the plasma membrane that are called coated pits. Similar structures can be observed at the *trans* face of the Golgi, where vesicles destined for endosomes and secretory vesicles originate.

The structure of clathrin-coated vesicles is shown schematically in **Figure 27.12.** The inner shell of the coat is made by proteins called adaptors, which bind both to clathrin and to integral membrane proteins of the vesicle. Different types of adaptors identify coated vesicles with different origins. There are several types of adaptors. Each is identified as AP (for adaptor complex) and a number. The most abundant adaptor is AP2, which is found at plasma membrane coated pits and identifies endocytic vesicles. These vesicles also contain an additional adaptor protein, AP180, which controls the size of the vesicle. AP4 is associated with the *trans*-Golgi network. AP 1 (founding member of the family) is found on vesicles that transport 6-mannose-phosphate receptors from the *trans*-Golgi network to the endosomes.

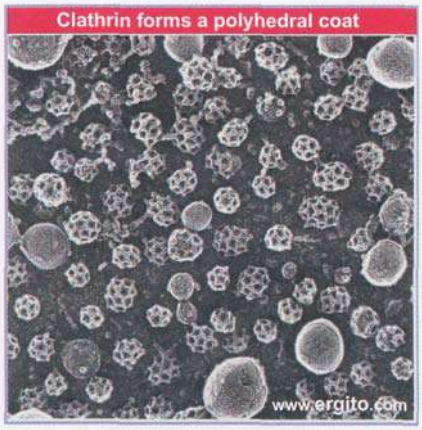


Figure 27.11 Coated vesicles have a polyhedral lattice on the surface, created by triskelions of clathrin. Photograph kindly provided by Tom Kirchhausen.

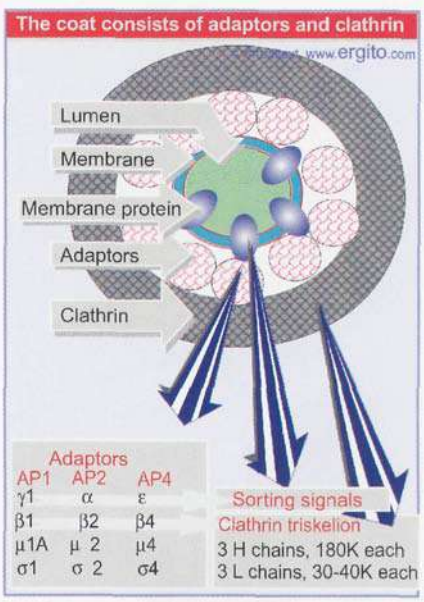


Figure 27.12 Clathrin-coated vesicles have a coat consisting of two layers: the outer layer is formed by clathrin, and the inner layer is formed by adaptors, which lie between clathrin and the integral membrane proteins.

Each AP is a heterotetramer. The individual subunits are called adaptins. The β adaptin usually binds to the clathrin skeleton. It may also interact with dileucine (KK) sorting signals in the membrane proteins. The μ adaptins recognize tyrosine-based sorting signals for internalization. The best characterized case is μ2, where phosphorylation of the subunit triggers a conformational change that enables it to recognize the target motif in a protein that is to be endocytosed. The a or γ adaptins (and presumably also the 8 and £ adaptins) are involved in interactions with the membrane where the vesicles are formed, that is, they are responsible for assembly of the full AP complex at the appropriate membrane, after which a clathrin coat assembles. The adaptor is therefore responsible not only for connecting the membrane of the vesicle to the clathrin skeleton, but also for incorporating the cargo proteins into the vesicle.

A variant of AP1, which contains the adaptin μ1B instead of the more common μ1A, is found in polarized epithelial cells, where it is involved in transporting proteins from the apical to the basolateral surface. We may expect that more variants like this will be discovered.

The variety of adaptor complexes, their localization to specific transport pathways, and their roles in recognizing cargo proteins, all argue that they are a key component of the targeting system that ensures that proteins are taken to the right location.

How many types of cargo protein can be carried by a single endocytic vesicle? It is not yet clear how many types of vesicle exist and what variety they display on the coats and in their cargo. We do know that some endocytic vesicles carry more than one type of cargo protein. Generally they are viewed as fairly specific carriers.

The coats of the COP-I-coated transition vesicles have 7 major protein components, called COPs. They exist as a high molecular weight complex (~700 kD), called coatomer, which is the precursor to the COP coat. The β-COP component has some homology to β-and β′-adaptins. This suggests a similar organization in which β-COP plays a similar role to the β- and β′-adaptins in connecting an outer coat protein (an analog of clathrin) to the membrane proteins in the vesicle.

COP-I-coated vesicles appear to be capable of performing both anterograde and retrograde transport. Such vesicles can be found carrying types of cargo that are transported in either direction; but any individual vesicles carries only anterograde or retrograde cargo, not both. Certain mutations in COP proteins block retrograde transport, which suggests that COP-I vesicles provide the sole (or at least major) capacity for retrograde transport. We do not know how COP-I vesicles moving in one direction are distinguished from those moving in the opposite direction: presumably there is some further component that has yet to be identified. Both directions of transport are probably supported through every level of the Golgi stack.

The coats of COP-II-coated vesicles consist of the protein complexes Sec23p/Sec24p (found as a 400 kD tetramer), Sec13p/Sec31p (which form a 700 kD complex), and Sar1p. There is no homology between the Sec protein components and the components of COP-I-coated vesicles. Sarlp is a small GTP-binding protein that regulates coat formation, and Sec23p is the GAP (GTPase-activating protein) that acts on Sarlp (see *27.7 Vesicles can bud and fuse with membranes)*. Sec24p is responsible for recruiting the cargo.

Another class of vesicle has the AP3 coat, whose subunits (8, β3, μ3, CT3) are related to those of the AP1 and AP2 adaptor complexes. This coat complex is found on some synaptic vesicles, which form at endosomes. This type of coat complex is also found on vesicles that transport cargo from the Golgi to lysosomes, and on storage vesicles.

27.6 Cisternal progression occurs more slowly than vesicle movement

Key Concepts

- An alternative model for ER-Golgi protein transport is that the stacks are mobile structures, and that *cis*-stacks mature into frans-stacks by a series of changes in lipid and protein composition.

Are coated vesicles responsible for all transport between membranous systems? There are conflicting models for the nature of forward transport from the ER, through Golgi cisternae, and then from the TGN to the plasma membrane.

The vesicular model for anterograde transport proposes that the Golgi cisternae are fixed structures that gain and lose proteins by the processes of vesicle fusion and budding. The process starts when COP-II coated vesicles bud at the ER and transport cargo to the Golgi. It is still not clear whether the vesicles that transport proteins along the Golgi are COP-II or COP-I coated. The natures of the coat(s) of vesicles that proceed from the Golgi to the plasma membrane remain unknown.

An alternative model for anterograde transport suggests that there is cisternal maturation. Instead of being fixed structures, cisternae move from the *cis* side of the Golgi to the *trans* side, maturing into more *trans-like* types of cisternae by changes in their protein constitution. Evidence for cisternal maturation has been provided by following the fate of a substrate protein that is too large to be incorporated into vesicles. Procollagen type I assembles into rod-like triple helices that are ~300 nm long in the lumen of the ER. These rods can be followed as they move into the *cis*-Golgi and through the Golgi to the TGN. Because they remain intact and are too large to be incorporated into vesicles (COP-coated vesicles are 60-90 nm in diameter), this means that the membrane-bound compartment containing the rods must itself have moved from the *cis* to the *trans* side of the Golgi. This shows at least the plausibility of cisternal maturation, although it does not demonstrate whether normal cargo proteins are carried by vesicles or also move by cisternal maturation.

To take the model for cisternal maturation to its extremes, the *cis*-Golgi could be formed by fusion between COP-II coated vesicles that bud from the ER; this process might also involve larger tubules. The *cis*-Golgi cisternae would move steadily forward until they mature into the *trans*-Golgi cisternae. At the TGN, secretory vesicles might form by fragmenting into tubular structures, without requiring any special type of coat. Of course, as cisternae mature, proteins that belong to more *cis*-like cisternae must be retrieved; this would occur by COP-I-mediated retrograde vesicular transport.

The outstanding question is the relative quantitative importance of cisternal maturation and vesicular transport for the anterograde direction. It is likely that cisternal progression is much slower than vesicle movement. This may mean that there is a two-track system, in which proteins can be transported rapidly in vesicles, but this will be accompanied by the much slower maturation of *cis*-stacks into *trans*-stacks.

Whichever model applies, the TGN provides the sorting center for directing proteins on the anterograde route to the plasma membrane, endosomes, or other membrane surfaces.

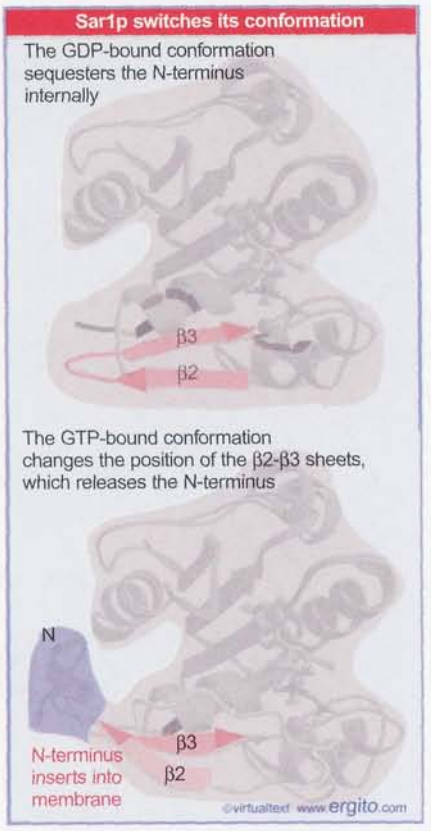


Figure 27.13 The conformation of Sar1p is controlled by the guanine nucleotide. Its N-terminal sequence is a hydrophobic stretch that can insert into membranes. This sequence is bound to the β2-β3 sheets and localized internally in the GDP-bound conformation. When Sar1p is bound to GTP, the β2-β3 sheets change their position and release the N-terminus, which projects out of the protein so that it can bind to membranes.

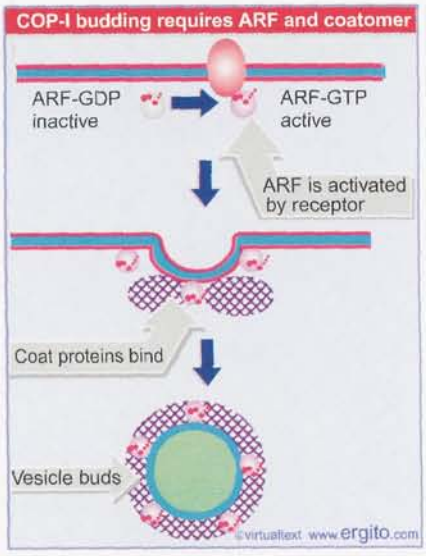


Figure 27.14 ARF and coatomer are required for the budding of COP-I-coated vesicles.

27.7 Vesicles can bud and fuse with membranes

Key Concepts

- Budding and fusion are controlled by a monomeric G protein, ARF/Sar1p.
- Budding requires ARF/Sar1p to be bound to GTP so that its N-terminus is available to insert into the membrane.
- Fusion occurs when the coat is destabilized by the hydrolysis of GTP.

Budding and fusion are essentially reversible reactions. Budding occurs when coat proteins assemble on a patch of membrane, ultimately causing its release as an independent vesicle. Fusion occurs when the coat proteins are removed, exposing the membrane surface, which can then fuse with a target membrane. Whether the coat proteins assemble or disassemble is controlled by the state of a monomeric G protein.

Budding of COP-I, clathrin-coated, and AP3 vesicles is initiated by ARF (ADP-ribosylation factor). Sar1p is a closely related protein that serves the same role for COP-II-coated vesicles. ARF is myristoylated at the N-terminus and can insert spontaneously into lipid bilayers. ARF-GTP is the active form; ARF-GDP is inactive. Sarlp behaves similarly, except that its N-terminus functions without requiring modification. ARF/Sar1p's activity (and ability to recycle) is controlled by GTP hydrolysis. The type of guanine nucleotide controls the conformation, so that the N-terminus is exposed only when GTP is bound. ARF/Sar1p is the key component that triggers both the budding and fusion processes in response to the condition of its guanine nucleotide.

Figure 27.13 shows how the conformation of ARF/Sarlp is controlled by its guanine nucleotide in such a way that the N-terminus is available to sponsor membrane insertion only when GTP is bound. The N-terminus has an amphipathic helix that can insert its hydrophobic side chains into a lipid bilayer. Contacts with the γ phosphate of GTP change the localization of a pair of β-sheets. This causes them to move within the protein, releasing the N-terminal region, which then protrudes and can insert into the membrane.

Figure 27.14 illustrates the initiation of budding when ARF/Sarlp is converted to the GTP-bound form. ARF/Sarlp is recruited to the appropriate membranes by interacting with a receptor (called ARF-GAP) that activates the factor. The receptor does not become a component of the vesicle, but acts catalytically to activate its target. The function of ARF/Sarlp is to provide the binding sites at which the other coat proteins can assemble; after ARF/Sarlp inserts into the membrane, the other coat proteins bind to it stoichiometrically. Coat proteins surround the membrane as a prerequisite for budding, but another function may be needed to complete the process of "pinching off."

Formation of a vesicle is energetically unfavorable—the membrane must be deformed and finally a small sphere is pinched off. Most is known about the process during endocytosis. When clathrin polymerizes into a basket-like structure at a coated pit, it pulls the membrane inward toward the cytoplasm. Then it is necessary to release the vesicle by separating it at the "neck" where it is joined to the plasma membrane. This requires input of energy in order to deform the membrane.

The first protein to be discovered that may act at the scission step was a GDP/GTP-binding protein called dynamin. The GDP-bound form of dynamin binds to the clathrin lattice. Replacement of the GDP by GTP causes the dynamin to form a ring around the neck of the forming vesicle. One model proposes that the dynamin uses hydrolysis of GTP to provide

the energy for the deformation and scission. (Dynamin is also involved in septation of mitochondria; see *13.24 How do mitochondria replicate and segregate?)* Other proteins have also since been discovered that can accomplish vesiculation, including amphiphysin and endophilin. The common feature in all these proteins is their ability to bind to the membrane, typically by interacting with a phosphoinositide head of a lipid.

The original evidence for the involvement of dynamin was the discovery that the *Drosophila* mutation *shibire*, which causes a temperature-sensitive paralysis as the result of a block in endocytosis, lies in a gene for dynamin. It is not clear whether dynamin is required for other vesicle trafficking, but one possibility is that different variants of dynamin (produced by alternative splicing) are required for different trafficking pathways.

Fusion is a reversal of budding. The coat of the vesicle is an impediment to fusion, and must be removed. If we suppose that uncoated vesicles would have an ability to fuse spontaneously with membranes in the cell, we can view the coat as a protective layer that preserves the vesicle until it reaches its destination. **Figure 27.15** shows that dissociation of the coat is triggered by hydrolysis of the GTP bound to ARE This causes ARF to withdraw from the membrane of the vesicle. The coat of COP-I-coated vesicles is unstable in the absence of ARF, so the coatomers then dissociate from the vesicle. In the case of clathrin-coated vesicles, the coat is stable, and further components, including a chaperone-like protein and an ATPase, are necessary to remove it. However, a significant proportion of the clathrin and the adaptors in the cell are found in a pool of free molecules, which suggests that both components are removed when vesicles become uncoated prior to fusion with their membrane targets.

Components involved in the fusion between the donor membrane of the vesicle and the target membrane were identified via a mammalian protein called NSF, identified by its sensitivity to the sulfhydryl agent NEM (N-ethyl-maleimide). NSF is the homologue for the product of yeast gene *sec18*, which is required for vesicle fusion during movement of transition vesicles. Fusion requires a 20S complex that consists of NSF (a soluble ATPase), a SNAP (Soluble NSF-*A*ttachment Protein), and SNAREs (*SNA*P-*re*ceptors) located in the membrane. The fusion particle is a basic part of the vesicular apparatus; it functions at all surfaces where vesicles fuse in the secretory and endocytic pathways. It functions in conjunction with the components necessary for vesicle-membrane recognition and uses ATP hydrolysis to allow them to recycle.

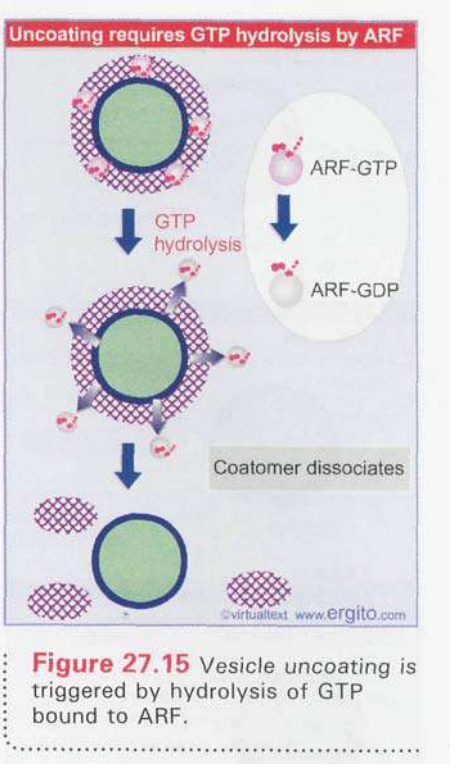


Figure 27.15 Vesicle uncoating is triggered by hydrolysis of GTP bound to ARF.

27.8 The exocyst tethers vesicles by interacting with a Rab

Key Concepts

- A tethering complex consists of a group of proteins that are localized to a target membrane and recognize a protein on a suitable vesicle.
- Rabs are prenylated monomeric G proteins that are specifically distributed to different membrane surfaces.
- One mode of action for a tethering complex is to recognize a specific Rab on the vesicle.

What controls the specificity of vesicle targeting? When a vesicle buds from a particular membrane, it has a specific target: vesicles leaving the ER have the *cis*-Golgi as their destination, vesicles leaving the *trans*-Golgi fuse with the plasma membrane, etc. The apparatus for budding and fusion is ubiquitous, so some additional component(s) must allow a vesicle to recognize the appropriate target membrane.

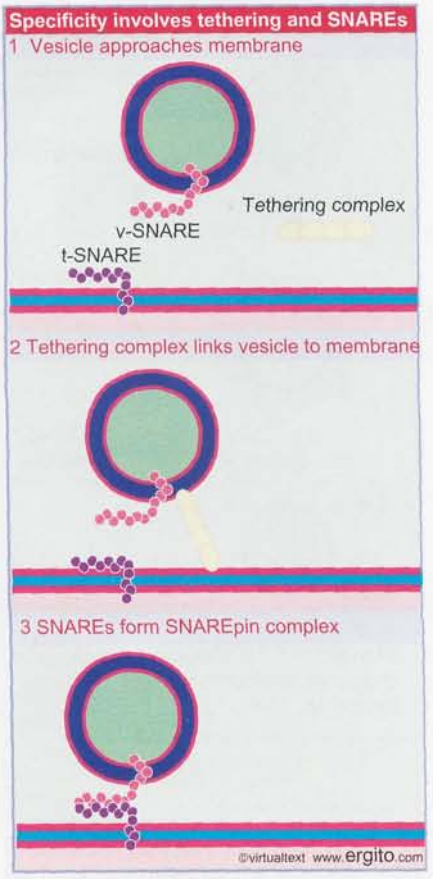


Figure 27.16 A vesicle makes its first contact with a target membrane via a tethering complex. Rab proteins are involved in the tethering reaction. When the v-SNARE and t-SNARE come into close proximity, they interact to bring the vesicle into contact with the membrane. Then the membranes fuse.

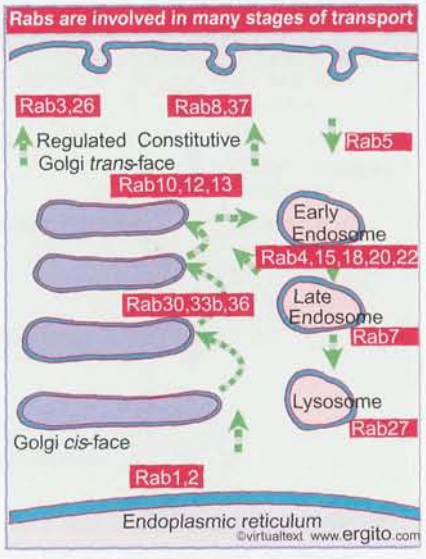


Figure 27.17 Rab proteins affect particular stages of vesicular transport.

Specific interactions occur between the vesicle and its target membrane at more than one stage, as summarized in **Figure 27.16.** A vesicle approaches a membrane either by diffusion or by some transport mechanism. Tethering complexes are found in the vicinity of the membrane and make the initial recognition of a suitable vesicle. Monomeric GTP-binding proteins called Rabs are important in the tethering reaction. They are either recognized by the tethering apparatus or involved in its activation. Tethering basically holds the vesicle in the vicinity of the membrane, but does not itself cause fusion. Tethering is followed by interactions between SNARE proteins that trigger fusion (see next section).

The best characterized tethering complex is the exocyst, an assembly of eight protein subunits that were originally identified as the products of yeast genes in which mutation causes the accumulation of vesicles that should have released proteins through the plasma membrane. The absence of any of the subunits makes the vesicles unable to fuse with the membrane, even though the fusion apparatus is not itself affected. The complex is localized at sites where exocytosis occurs. The homologues of this complex in mammalian cells are called the Sec6/8 complex.

The tethering complex recognizes an appropriate vesicle by interacting with a Rab protein on the vesicle. Rabs are attached to membranes via the addition of prenyl or palmityl groups at the C-terminus. There are ~30 Rabs, distributed to different membrane systems in the cell. **Figure 27.17** summarizes their distribution. Different Rabs are involved in ER to Golgi transport, in the constitutive and regulated pathways from the Golgi to the plasma membrane, and in stages of transport between endosomes. For example, mutations in the yeast genes *YPT1* or *SEC4* that code for two such (related) proteins block transport and cause the accumulation of vesicles in the Golgi stacks or between Golgi and plasma membrane, respectively.

The Rabs are GTP-binding proteins that are active in the form bound to GTP; but hydrolysis of the GTP converts the protein to an inactive form. As with other monomeric G proteins, their activities are affected by other proteins that influence the hydrolysis of GTP. There may be GAP (GTP-hydrolyzing) activities specific for certain Rabs, GEF proteins that stimulate dissociation of the guanine nucleotide, and GDI proteins that prevent dissociation of guanine nucleotide.

The exocyst binds to Sec4p, which is a Rab found on secretory vesicles. The reaction involves several of the exocyst subunits. Two of the exocyst subunits (Sec10p and Sec15p) bind to activated (GTP-bound) Sec4p. This subcomplex binds to Sec3p, which is a component of the exocyst responsible for localization at polarized secretion sites. The result is to tether the vesicle at the secretion site.

27.9 SNARES are responsible for membrane fusion

Key Concepts

- Vesicle-membrane fusion is specified by an interaction between a membrane-bound v-SNARE on the vesicle and t-SNARE on the target membrane.
- The SNAREs have a rod-like structure that lies parallel to the membrane surface.
- The v-SNARE and t-SNARE interact by a zipper-like reaction that creates a bundle of 4 helices parallel to the membrane, and which pulls the membranes into close contact.

The SNARE hypothesis proposes that the fusion reaction results from the interaction of a v-SNARE membrane protein carried by

the vesicle with a t-SNARE membrane protein that is present on the target membrane. Every membrane surface is marked by a particular constellation of SNAREs, as summarized in **Figure 27.18,** so that vesicle-membrane interactions are determined by pairwise combinations of v-SNAREs and t-SNAREs. This imposes further specificity at the step following the tethering reaction.

Figure 27.19 illustrates the interaction between SNAREs, for the example of a synaptic system (involving exocytosis by neurons: see next section). The v-SNARE is a transmembrane protein carried by the vesicle. The t-SNARE includes two proteins; syntaxin is a transmembrane protein, and SNAP-25 is connected to the membrane by a fatty acyl linkage. (The name of SNAP-25 has an independent origin, and it has no connection with the SNAPs of the fusion particle.) Homologues to these SNAREs are found in other systems, including other animal cell types and yeast cells.

The major part of each SNARE is exposed in the cytoplasm, and includes an extensive coiled-coil structure. Such structures commonly participate in protein-protein interactions. In fact, v-SNAREs can bind directly to the t-SNAREs *in vitro,* even without the other components of the fusion particle. The interaction between v-SNARE and t-SNARE is sufficient to sponsor membrane fusion. Liposomes containing v-SNAREs can fuse with liposomes containing t-SNAREs in the absence of any additional components. An energy source is not required, suggesting that activation energy is provided by changes in the conformation of the proteins. This *in vitro* reaction is slow, occurring over a time course of minutes. Comparison with the millisecond time course of fusion *in vivo* implies that other components will be needed to facilitate the reaction. But the basic apparatus involved in bringing the membranes together consists of the SNAREs. The idea that individual SNAREs influence the specificity of the reaction is supported by the observation that only certain pairs of SNAREs allow one liposome to interact with another.

A SNARE complex has a rod-like structure (~4 X 14 nm) in which the v-SNARE and t-SNARE are bound in parallel. Their membrane anchors are at the same end, implying that the rod must lie in a plane between the two membrane surfaces. This structure is called a SNARE-pin. **Figure 27.20** is based on the crystal structure, which shows that the complex consists of a 4-helix bundle. **Figure 27.21** shows a model for the SNAREpin superimposed at the appropriate scale on an electron micrograph of the complex.

The other components of the fusion particle bind to the far end of the complex. Hydrolysis of ATP and dissociation of the fusion complex is probably necessary not for fusion as such, but in order to release the SNAREs from the SNAREpin in order to allow them to be used again. (The 20S fusion complex was originally envisaged to play a role prior to fusion, possibly in providing the energy for fusion, but now we believe it is more likely to function post-fusion.)

It requires a lot of energy to fuse two membranes. When the v-SNARE and t-SNARE interact, they form the SNAREpin by a zipping reaction that moves along the rod. As this generates the 4-helical bundle, the facing leaflets of the two membranes are brought closer and closer, and ultimately they fuse together spontaneously to form the structure drawn in **Figure 27.22.** This stage is called hemifusion. The structure is energetically unstable because of the void space between the membranes. This leads to contacts between the two distal membrane layers. When they break down and reform, there is a small area of contact between the aqueous environment on either side. This is called the fusion pore. It rapidly expands and the membrane relaxes. The fusion pore is probably entirely lipid, with proteins acting to deform the exterior surface. The energy for driving the transformation of the membrane may come from the interactions between the SNARE proteins.

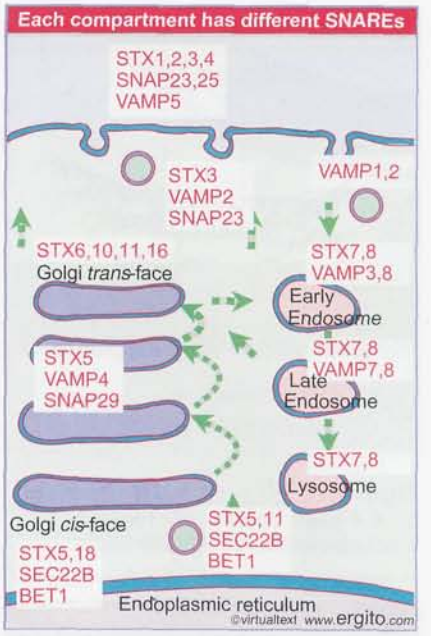


Figure 27.18 Each compartment or vesicle has a characteristic set of SNAREs. The constituent proteins fall into several families, including syntaxins (STX), VAMPs, and SNAP-25s.

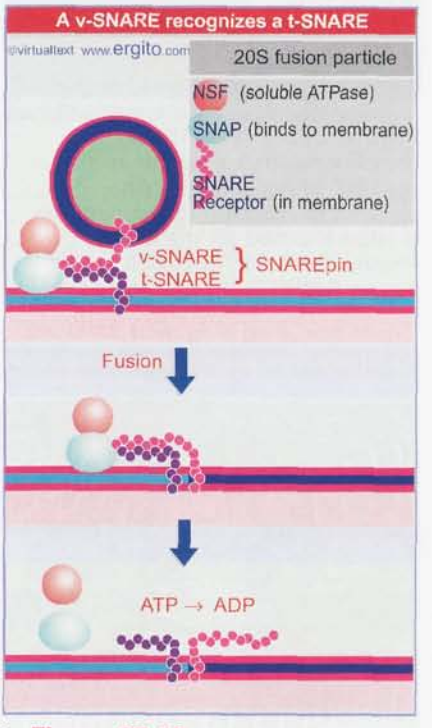


Figure 27.19 Specificity for docking is provided by SNAREs. The v-SNARE carried by the vesicle binds to the t-SNARE on the plasma membrane to form a SNAREpin. NSF and SNAP remain bound to the far end of the SNAREpin during fusion. After fusion, ATP is hydrolyzed and NSF and SNAP dissociate to release the SNAREs.

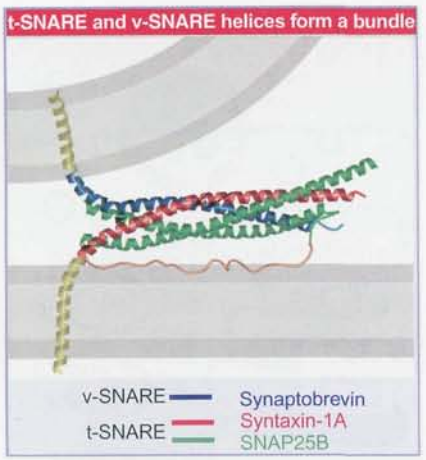


Figure 27.20 A SNAREpin forms by a 4-helix bundle. Photograph kindly provided by Axel Brunger.

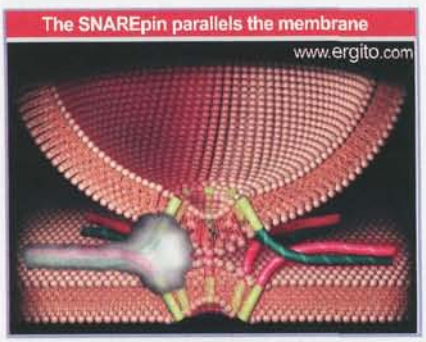


Figure 27.21 A SNAREpin complex protrudes parallel to the plane of the membrane. An electron micrograph of the complex is superimposed on the model. Photograph kindly provided by James Rothman.

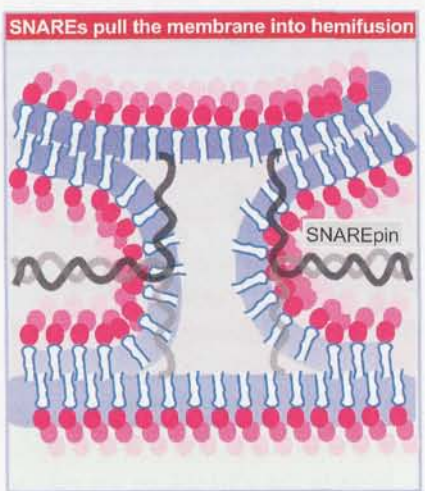


Figure 27.22 When a SNAREpin forms, it pulls the apposing membrane leaflets into juxtaposition in a state of hemifusion.

27.10 The synapse is a model system for exocytosis

Key Concepts

- The best characterized example of regulated exocytosis occurs at a synapse to release neurotransmitters from the donor neuron.

The synapse has been especially useful for investigating fusion because it offers the advantage of large numbers of vesicles of the same type which fuse with the plasma membrane when a specific trigger is applied.

Impulses in the nervous system are propagated by the passage of material from a donor (or presynaptic) cell to a recipient (or postsynaptic cell). **Figure 27.23** illustrates a nerve terminal. An impulse in the donor cell triggers the exocytic pathway. Stored coated vesicles (called synaptic vesicles) move to the plasma membrane and release their contents of neurotransmitters into the extracellular fluid. The neurotransmitters in turn act upon receptors at the plasma membrane of the recipient cell.

Exocytosis would lead to the accumulation of vesicle components in the plasma membrane if there were no means to retrieve them. There are two possibilities for the recycling of these vesicles, both of which may occur.

In the "kiss and run" model illustrated in **Figure 27.24,** a vesicle does not completely fuse with the plasma membrane, but contacts it transiently. The neurotransmitter is released through some sort of pore; then the vesicle reforms. Major questions about this pathway are how the vesicle maintains its integrity and what sort of structure forms the pore.

In the fusion model illustrated in **Figure 27.25,** the vesicle fuses with the plasma membrane in the conventional manner, releasing its contents into the extracellular space. Recycling occurs by the formation of clathrin-coated vesicles at coated pits, that is, by the endocytic pathway. This may occur at large invaginations of the plasma membrane. The importance of endocytosis in this pathway is emphasized by the fact that inhibition of the formation of the clathrin-coated vesicles affects neurotransmitter release from synaptic vesicles. A major question about the pathway is the relationship between the endocytic and exocytic vesicles. The synaptic vesicles are not clathrin-coated. It is probable that the clathrin-coated endocytic vesicles give rise to synaptic vesicles by losing their clathrin coats, but synaptic vesicles may also form by other pathways (as in the case of AP3-coated vesicles). It is probably true that removal of the clathrin coat takes place quite soon after budding for all classes of clathrin-coated vesicles; the process of removal is not well defined.

27.11 Protein localization depends on specific signals

Key Concepts

- A membrane protein is incorporated into a vesicle when a transport signal consisting of a short sequence of amino acids in its cytoplasmic tail binds an adaptor protein of the vesicle.

Various types of signals influence transport through the ER-Golgi system. A protein that has no special signals will presumably enter vesicles at a rate determined by its concentration in the compartment, and may move in the anterograde direction by bulk flow. However, most proteins appear to have specific signals that facilitate or retard transport.

A typical cargo protein has a transport signal that is responsible for its entry into budding vesicles. **Figure 27.26** shows that the transport signal in a transmembrane protein is usually a region in its cytoplasmic domain that binds to an adaptor protein of the vesicle coat. **Figure 27.27** shows that the transport signal in a soluble cargo protein (for example, a secreted protein that passes through the lumen) is a region that binds to the lumenal domain of a transmembrane cargo receptor, which in turn has a cytoplasmic domain that binds an adaptor protein. Interaction between the cargo and the coat thus directly or indirectly determines specificity of transport. Such mechanisms control anterograde transport from the ER to the cell surface and other destinations.

A protein may be prevented from leaving a compartment by a retention signal. Such signals are often found in transmembrane regions, perhaps because aggregation between them creates a structure that is too large to be incorporated into a budding vesicle.

We have detailed information about several types of signal: a conformation that is required for proteins to be internalized by endocytosis; an amino acid sequence that targets proteins to the ER; and a modification that targets proteins to lysosomes (small membranous bodies, where proteins are degraded).

Internalization of receptors via coated pits requires information in their cytoplasmic tails. The sequence for internalization is usually a short amino acid motif located near the C-terminus. Typically it makes a tight turn in the structure and exposes a tyrosine. Two signals of this sort are NPXY and YXRF. Although these are the basic signals for internalization, other sequences in the cytoplasmic tail influence the efficiency.

Enzymes that will be transported to lysosomes are recognized as targets for high mannose glycosylation, and are trimmed in the ER as described in Figure 27.5. Then *mannose-6-phosphate* residues are generated by a two-stage process in the Golgi. First the moiety is added to the 6 position of mannose by GlcNAc-phosphotransferase; then a glucosaminidase removes the N-acetyl-glucosamine (GlcNAc).

The action of the phosphotransferase provides the critical step in marking a protein for lysosomal transport. It occurs early in ER-Golgi transfer, possibly between the ER and the *cis* Golgi. The basis for the enzyme's specificity is its ability to recognize a structure that is common to lysosomal proteins. The structure consists of two short sequences, which are separated in the primary sequence, but form a common surface in the tertiary structure. Each of these sequences has a crucial lysine residue. The nature of this signal explains how proteins with little identity of sequence may share a common pathway for localization.

Lysosomal proteins continue to be transported along the Golgi stacks until they encounter receptors for mannose-6-phosphate. Recognition of mannose-6-phosphate targets a protein for transport in a coated vesicle to the lysosome. This final stage of sorting for the lysosome occurs in the *trans* Golgi, where the proteins are collected by specific transport vesicles that are coated with clathrin. The vesicles transport the lysosomal proteins to the late endosome, where they join the pathway for movement to the lysosome. A single pool of mannose-6-phosphate receptors is probably used for directing proteins to the lysosome whether they are newly synthesized or endocytosed. Most of the receptors in fact are located on endosomes, where they could recognize endocytosed proteins.

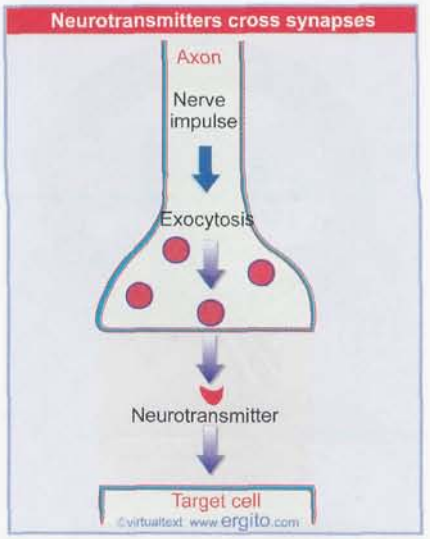


Figure 27.23 Neurotransmitters are released from a donor (presynaptic) cell when an impulse causes exocytosis. Synaptic (coated) vesicles fuse with the plasma membrane, and release their contents into the extracellular fluid.

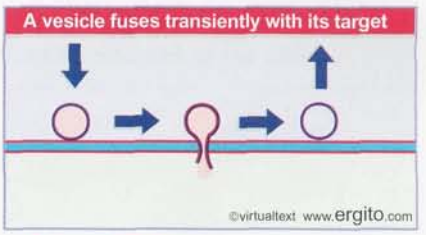


Figure 27.24 The kiss and run model proposes that a synaptic vesicle touches the plasma membrane transiently, releases its contents through a pore, and then reforms.

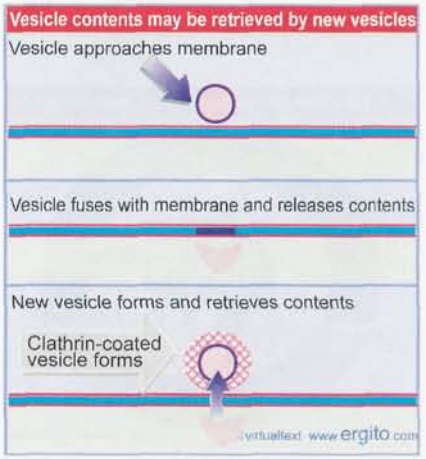


Figure 27.25 When synaptic vesicles fuse with the plasma membrane, their components are retrieved by endocytosis of clathrin-coated vesicles.

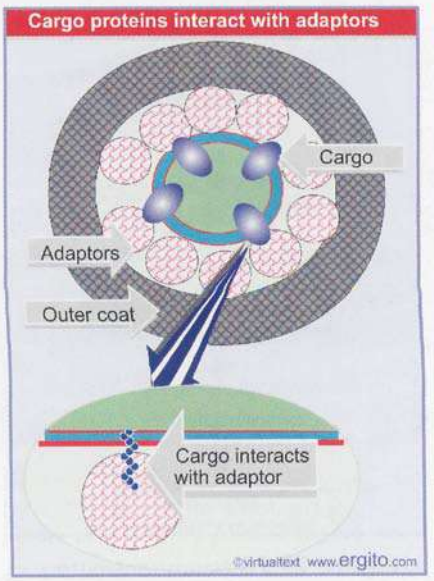


Figure 27.26 A transport signal in a *trans*-membrane cargo protein interacts with an adaptor protein.

27.12 ER proteins are retrieved from the Golgi

Key Concepts

- The KDEL or KKXX C-terminal sequences are required for a soluble protein to be retained in the ER.
- The KDEL signal binds to a receptor in the Golgi, which causes the receptor-KDEL protein to be transported back to the ER.
- The KKXX signal binds to components of the coatomer of COP-I vesicles, which return the protein from the Golgi to the ER.

Proteins that reside in the lumen of the endoplasmic reticulum possess a short sequence at the C-terminus, Lys-Asp-Glu-Leu (KDEL in single letter code). The alternative signals HDEL or DDEL are used in yeast. If this sequence is deleted, or if it is extended by the addition of other amino acids, the protein is secreted from the cell instead of remaining in the lumen. Conversely, if this tetrapeptide sequence is added to the C-terminus of lysozyme, the enzyme is held in the ER lumen instead of being secreted from the cell. This suggests that there is a mechanism to recognize the C-terminal tetrapeptide and cause it to be localized in the lumen.

An interesting question emerges from the behavior of proteins that have an ER-localization signal. Does this signal cause a protein to be held so that it cannot pass beyond the ER or is it the target for a more active localization process? The model shown in **Figure 27.28** suggests that *the KDEL sequence causes a protein to be returned to* the ER from an early Golgi stack.

Because the modification of proteins as they pass through the Golgi is ordered, we can use the types of sugar groups that are present on any particular species as a marker for its progress on the exocytic pathway. When a KDEL sequence is added to a protein that usually is targeted to the lysosome (because its oligosaccharide gains mannose-6-P residues), it causes the protein to be held in the ER. But the protein is modified by the addition of GlcNAc-P, which happens only in the Golgi. The GlcNAc is not removed, so the protein cannot have proceeded far enough through the Golgi stacks to encounter the second of the enzymes in the mannose-6-P pathway. This suggests that KDEL is recognized by a receptor which is located in Golgi before the stack containing the second enzyme.

Mutations in the *S. cerevisiae* genes *ERD1* and *ERD2* prevent retention of proteins with the HDEL signal in the ER; instead the proteins are secreted from the cell. The products of both these genes are integral membrane proteins. The *ERD1* mutation causes a general defect in the Golgi; this supports the idea that sorting of the ER proteins occurs by salvage from the Golgi. The *ERD2* mutation identifies the receptor for the HDEL sequence. One model for its role is that it cycles between the Golgi salvage compartment and the ER. This idea is supported by the localization of the corresponding receptor in mammalian cells: it is found largely in the Golgi, but overexpression of a protein with KDEL sequence causes it to concentrate in the ER. So binding of a KDEL-protein causes the receptor to move from Golgi to ER. It may have a high affinity for the HDEL sequence under the conditions prevailing in the Golgi, but a low affinity in the ER. This could enable ERD2 to seize proteins by their HDEL tails in the Golgi, and take them back to the ER, where they are then released.

Another signal is responsible for the localization of transmembrane proteins in the ER. This is KKXX, and thus consists of two Lys residues, located in the cytoplasmic tail just prior to the C-terminus.

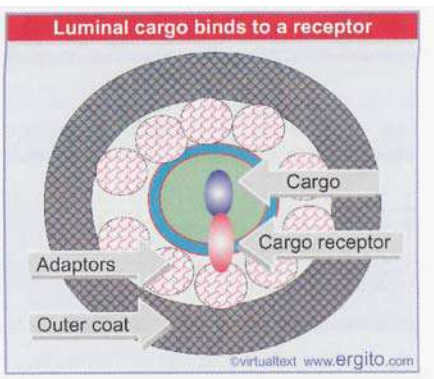


Figure 27.27 A transport signal in a luminal cargo protein interacts with a transmembrane receptor that interacts with an adaptor protein.

The dilysine motif of KKXX proteins binds to the β′- and α-COP components of coatomer. Yeast mutants that affect β′-, a-, or γ-COP are defective in retrieval of KKXX proteins from the Golgi. This suggests that vesicles with the COP-I coat are involved in retrieving proteins from the Golgi and returning them to the ER, that is, COP-I vesicles are responsible for retrograde transport.

Overall protein transport is a unidirectional process: proteins enter the ER and are transported through the Golgi, unless stopped *en route*. COP-II-coated vesicles are thought to provide the major capacity for anterograde transport from the ER to the Golgi. COP-I-coated vesicles provide transport capacity along the Golgi stacks. However, both COP-I- and COP-II coated vesicles can be observed to bud from the ER, so there is the possibility that they are involved at multiple stages (perhaps in transporting different types of cargoes).

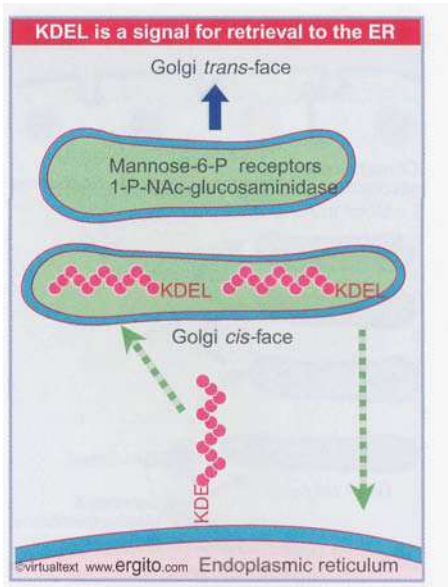


Figure 27.28 An (artificial) protein containing both lysosome and ER-targeting signals reveals a pathway for ER-localization. The protein becomes exposed to the first but not to the second of the enzymes that generates mannose-6-phosphate in the Golgi, after which the KDEL sequence causes it to be returned to the ER.

27.13 Brefeldin A reveals retrograde transport

Key Concepts

- Brefeldin A blocks the action of COP-II vesicles and thereby prevents anterograde (forward) transport.

Retrograde transport usually is obscured by anterograde transport, but is revealed when cells are treated with the drug brefeldin A (BFA), which specifically blocks the forward direction of transport. BFA blocks conversion of ARF from the GDP-bound to the GTP-bound form, and therefore prevents budding of coated vesicles. As a result of the block, a network of tubules forms between the cisternae of the Golgi (abolishing their usual independence) and joins them to the endoplasmic reticulum. There is resorption of most of the membranes of the *cis*-medial Golgi into the endoplasmic reticulum, which is accompanied by the redistribution of Golgi proteins into the ER, effecting a retrograde transport.

This happens because the COP-II vesicles involved in forward movement are more sensitive to the drug than the COP-I vesicles involved in retrograde movement. Retrograde transport may serve to retrieve membrane components to compensate for anterograde movement, and of course also provides for the retrieval of ER-proteins from the Golgi. It is possible there may also be other retrograde transport systems: certain toxins that are endocytosed at the plasma membrane can be found in the ER, but this retrograde transport does not appear to depend on the known systems.

The effects of brefeldin on ER-Golgi transport are universal, but in addition it inhibits other transport processes differently in different cells. In some cell types it inhibits transcytosis (transport from the basolateral surface to the apical surface in polarized cells); in other cells it inhibits transport from the *trans* Golgi network to endosomes. A related phenomenon is revealed by isolating cells that can grow in the presence of brefeldin. This identifies mutants in which transport is resistant in particular locations (such as endosomes or Golgi) but remains sensitive in others.

Brefeldin acts by binding to a common domain (the Sec7 domain) in the exchange factors (GEFs) that are responsible for regenerating ARF-GTP from ARF-GDP. BFA stabilizes the association of the GEF with the GDP-bound form of the ARF. This causes the ARF to remain in its inactive state. The differing effects of BFA on individual transport processes probably means that there is a variety of GEFs that act on ARFs

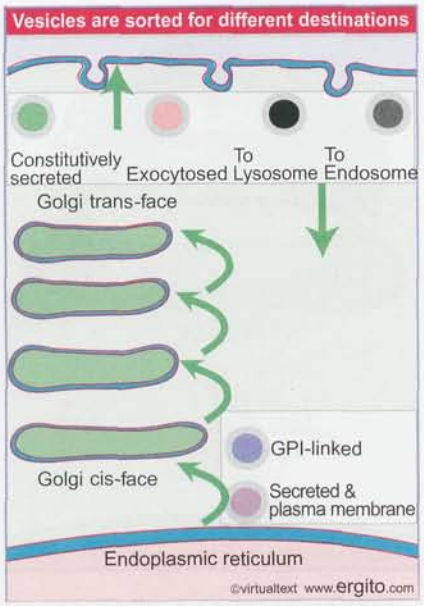


Figure 27.29 Sorting in the ER distinguishes GPI-linked proteins from other proteins, which are sorted into different vesicles at the Golgi *trans* face according to their ultimate destinations.

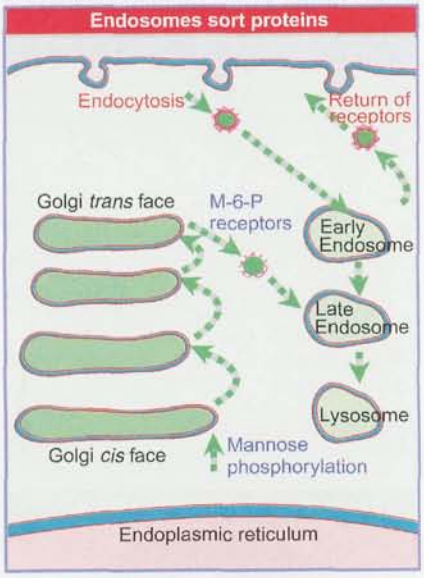


Figure 27.30 Endosomes sort proteins that have been endocytosed and provide one route to the lysosome. Proteins are transported via clathrin-coated vesicles from the plasma membrane to the early endosome, and may then either return to the plasma membrane or *proceed further to late endosomes* and lysosomes. Newly synthesized proteins may be directed to late endosomes (and then to lysosomes) from the Golgi stacks. The common signal in lysosomal targeting is the recognition of mannose-6-phosphate by a specific receptor.

on different membrane surfaces, so that the apparatus involved in assembling coated vesicles is specific for individual types of surface. The characteristic susceptibility of each GEF explains the effect of brefeldin on budding from its particular membrane.

27.14 Vesicles and cargos are sorted for different destinations

Key Concepts

- Proteins are loaded into vesicles in the ER and are sorted only subsequently according to their destinations.
- GPI-linked proteins are an exception and are sorted into separate vesicles in the ER.

A key question is how the loading of cargo into vesicles is coordinated with the targeting of the vesicles to their destinations. The first issue is whether and when different types of cargos are sorted into different vesicles. Most sorting occurs in the Golgi, but some proteins are sorted in the ER.

Secretory proteins and plasma membrane proteins can be colocalized in vesicles released from the ER. This suggests the general principle that proteins are packaged into vesicles in the ER irrespective of their final destination, and they are sorted only subsequently in the Golgi. However, there is at least one exception to this rule. Some proteins are linked to the extracellular side of the plasma membrane by glycosyl-phosphatidyl-inositol (GPI). GPI-linked proteins leave the ER in different vesicles from other secretory proteins, and so must be sorted from them at a very early stage.

Other secreted and membrane-bound proteins appear to be transported in the same sets of vesicles from the ER through the Golgi to the *trans* face. Figure 27.29 shows that at this point they are sorted into (at least) four groups of vesicles as defined by their destinations:

- Constitutively secreted proteins go in vesicles directly to the plasma membrane.
- Proteins whose exocytosis is regulated enter vesicles that will fuse with the plasma membrane when an appropriate signal is given.
- Vesicles containing proteins that have been modified by addition of mannose-6-phosphate are directed to endosomes (from where they continue to lysosomes).
- Other vesicles may also be directed to endosomes.

27.15 Receptors recycle via endocytosis

Key Concepts

- Surface receptors are internalized by incorporation into clathrin-coated vesicles at a coated pit.
- Some receptors are internalized continuously, others are internalized upon binding ligand.
- Different receptors have different fates in the cell; the receptor and/or ligand may recycle to the surface or may be degraded.

The systems involved in importing proteins into the cell are closely related to those used for exporting secreted proteins. Ingestion of

receptors starts by a common route, which leads to several pathways in which receptors have different fates. Some receptors are internalized continuously, but others remain exposed on the surface until a ligand is bound, which makes them susceptible to endocytosis. The signals that trigger internalization are different for ligand-independent and ligand-induced endocytosis.

In either case, the receptors slide laterally into coated pits, which are indented regions of the plasma membrane surrounded by clathrin. It is not clear whether simple lateral diffusion can adequately explain the movement into coated pits or whether some additional force is required. Coated pits invaginate into the cytoplasm and pinch off to form clathrin-coated vesicles. These vesicles move to early endosomes, are uncoated, fuse with the target membrane, and release their contents. The process is called receptor-mediated endocytosis.

The immediate destination for endocytic (clathrin-coated) vesicles is the endosome, a rather heterogeneous structure consisting of membrane-bounded tubules and vesicles. There are at least two types of endosome, as indicated in **Figure 27.30.** Early endosomes lie just beneath the plasma membrane and are reached by endocytosed proteins within ~1 minute. Late endosomes are closer to the nucleus, and are reached within 5-10 minutes.

The early endosome provides the main location for sorting proteins on the endocytic pathway. Its role is a counterpart to that played by the Golgi for newly synthesized proteins. The interior of the endosome is acidic, with a pH <6. Proteins that are transported to the endosome change their structure in response to the lowering of pH; this change is important in determining their fate.

Receptors that have been endocytosed to the early endosome behave in one of two ways. They may return to the plasma membrane (by vesicular transport). Or they may be transported further to the lysosome, where they are degraded. Transport to the lysosome is the default pathway, and applies to any material that does not possess a signal specifically directing it elsewhere.

The lysosome contains the cellular supply of hydrolytic enzymes, which are responsible for degradation of macromolecules. Like endosomes, the lysosome is an acidic compartment (pH = 5).

The relationship between the various types of endosomes and lysosomes is not yet clear. Vesicles may be used to transport proteins along the pathway from one pre-existing structure to the next; or early endosomes may "mature" into late endosomes, which in turn "mature" into lysosomes. At all events, the pathway is unidirectional, and a protein that has left the early endosome for the late endosome will end up in the lysosome.

There are two routes to the lysosome. Proteins endocytosed from the plasma membrane may be directed via the early endosome to the late endosome. Newly synthesized proteins may be directed from the *trans* Golgi via the late endosome, as described above.

The fate of a receptor-ligand complex depends upon its response to the acidic environment of the endosome. Exposure to low pH changes the conformation of the external domain of the receptor, causing its ligand to be released, and/or changes the structure of the ligand. But the receptor must avoid becoming irreversibly denatured by the acid environment; the presence of multiple disulfide bridges in the external domain may play an important role in maintaining this unusual stability.

Four possible fates for a receptor-ligand complex are described in the alternative pathways of the next four figures:

- *Receptor recycles to the surface in coated vesicles, while the ligand is degraded.* **Figure 27.31** shows that this pathway is used by receptors that transport ligands into cells at high rates. A receptor recycles

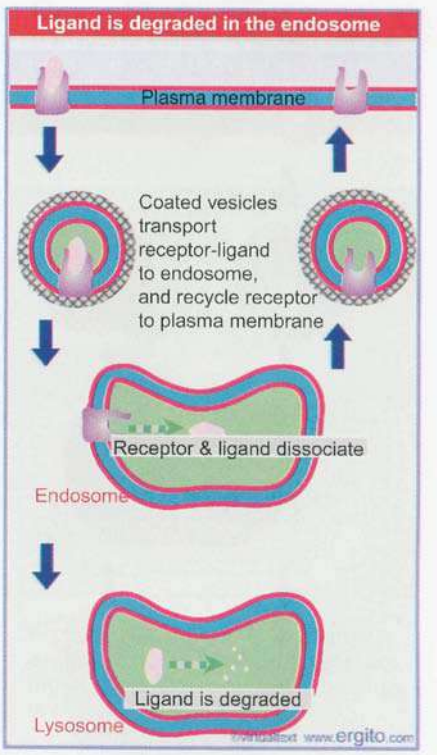


Figure 27.31 LDL receptor transports apo-B (and apo-E) into endosomes, where receptor and ligand separate. The receptor recycles to the surface, apo-B (or apo-E) continues to the lysosome and is degraded, and cholesterol is released.

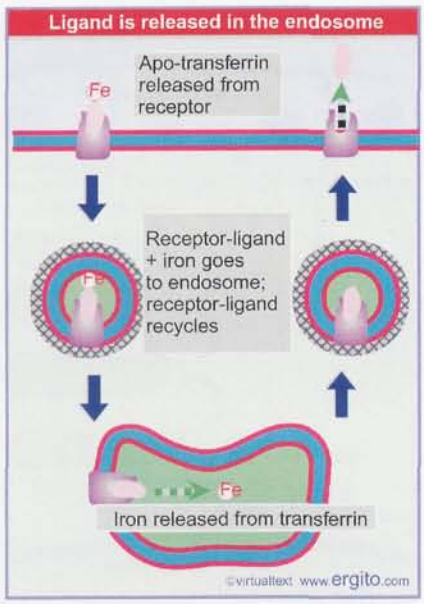


Figure 27.32 Transferrin receptor bound to transferrin carrying iron releases the iron in the endosome; the receptor now bound to apo-transferrin (lacking iron) recycles to the surface, where receptor and ligand dissociate.

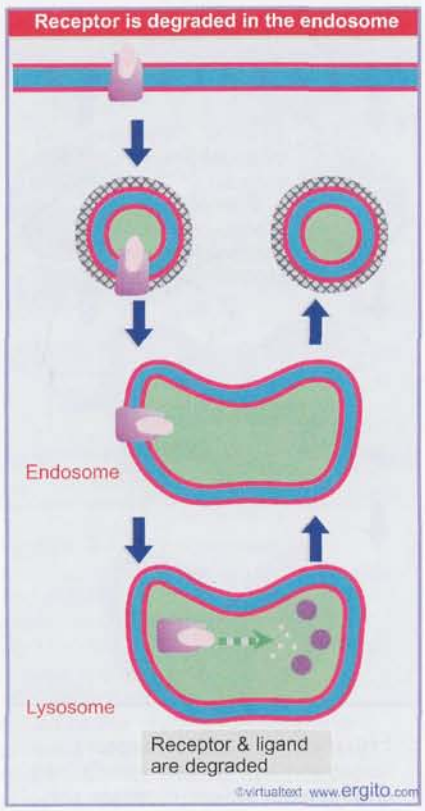


Figure 27.33 EGF receptor carries EGF to the lysosome where both the receptor and ligand are degraded.

every 1-20 minutes, and can undertake > 100 cycles during its lifetime of ~20 hours. The classic example of this pathway is the LDL receptor, whose ligands are the plasma low density lipoproteins apolipoprotein E and apolipoprotein B (collectively known as the LDLs). Apo-B is a very large (500 kD) protein that carries cholesterol and cholesterol esters. The LDL is released from its receptor in the endosome. The receptor recycles to the surface to be used again. The LDL and its cholesterol separate in the endosome; the LDL is sent on to the lysosome, where it is degraded, and the cholesterol is released for use by the cell. This constitutes the major route for removing cholesterol from the circulation. People with mutations in the LDL receptor accumulate large amounts of plasma cholesterol that cause the disease of familial hypercholesterolemia.

- *Receptor and ligand both recycle.* The transferrin receptor provides the classic example of this pathway, illustrated in **Figure 27.32.** The ligand for the receptor is the iron-carrying form of transferrin. When this reaches the endosome, the acid environment causes transferrin to release the iron. The iron-free ligand, called apo-transferrin, remains bound to the transferrin receptor, and recycles to the plasma membrane. In the neutral pH of the plasma membrane, apo-transferrin dissociates from the receptor. This leaves apo-transferrin free to bind another iron, while the transferrin receptor is available to internalize another iron-carrying transferrin. Again this cycle is quite intensively used; a transferrin receptor recycles every 15-20 minutes, and has a half-life of >30 hours. It provides the cell with the means of taking up iron.
- *Receptor and ligand both are degraded.* The EGF receptor binds its ligand as a requirement for internalization. Although EGF and its receptor appear to dissociate at low pH, they are both carried on to the lysosome, where they may be degraded, as indicated in **Figure 27.33.** We do not know how and whether these events are related to the ability of EGF to change the phenotype of a target cell via binding to the receptor.
- *Receptor and ligand are transported elsewhere.* The route illustrated in **Figure 27.34** is available in certain polarized cells. A receptor-ligand combination is taken up at one cell surface, transported to the endosome, and then released for transport to the far surface of the cell. This is called transcytosis. By this means, receptors can transport immunoglobulins across epithelial cells.

Rapid recycling in general occurs for receptors that bring ligands into the cell, not for those that trigger pathways of signal transduction. Receptors involved in signaling changes from the surface are usually degraded if they are endocytosed.

27.16 Internalization signals are short and contain tyrosine

Key Concepts

- Signals for internalization are short (~4 amino acids), are located in the cytoplasmic tail near the plasma membrane, and usually contain a tyrosine residue that is exposed by the protein conformation.

What features of protein structure are required for endocytosis? Mutations that prevent internalization can be used to identify the relevant sequences in the receptors. In fact, the characterization of

an internalization defect provided evidence that entry into coated pits is needed for receptor-mediated endocytosis of LDL. In cells from human patients with such defects in the LDL receptor, the receptor gathers in small clusters over the plasma membrane, and cannot enter coated pits in the manner observed for wild-type cells.

The mutations responsible for this type of defect all affect the *cytoplasmic domain* of the receptor, which functions independently in sponsoring endocytosis. A recombinant protein whose extracellular domain is derived from influenza virus hemagglutinin, which is not usually endocytosed, can be internalized if it is provided with a cytoplasmic domain from an endocytosed receptor.

We do not yet have a clear view of all of the sorting signals that mediate endocytosis, but two features are common. The relevant region of the protein comprises a relatively short part of the cytoplasmic tail close to the plasma membrane. And the presence of a tyrosine residue in this region often is necessary. The most common motif is YXXϕ (where ϕ is a hydrophobic amino acid). Removal of the tyrosine prevents internalization; conversely, substituting tyrosine in the relevant region of a protein that is not endocytosed (such as influenza virus hemagglutinin) allows it to be internalized.

Tyrosine is also found in another signal for internalization: NPXY (Asn-Pro-X-Tyr). The essential tyrosine can be replaced by other aromatic amino acids, but not by other types of amino acids. The NPXY motif is found in several group I proteins that are internalized, and is usually located close to the plasma membrane. (It was first identified in the LDL receptor.) It could be a general signal for endocytosis. In proteins that are internalized in response to ligand binding, the internalization signal may be generated by a change in conformation as a result of the binding.

Do internalized receptors interact directly with proteins on the vesicles that transport them? Clathrin forms an outer polyhedral layer on clathrin-coated vesicles, but other proteins form an inner layer. The adaptins recognize the appropriate sequences in the cytoplasmic domains of receptors that are to be internalized (see Figure 27.12). **Figure 27.35** illustrates a model in which, as a coated pit forms, the adaptins bind to the receptor cytoplasmic domain, immobilizing the receptor in the pit. As a result, the receptor is retained by the coated vesicle when it pinches off from the plasma membrane.

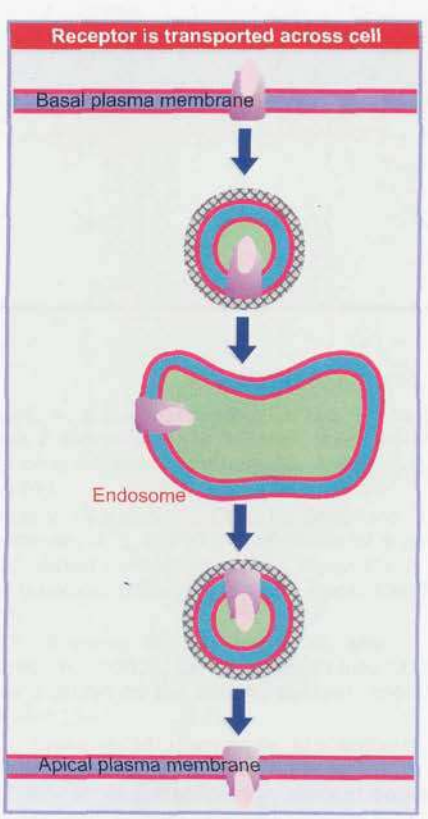


Figure 27.34 Ig receptor transports immunoglobulin across the cell from one surface to the other.

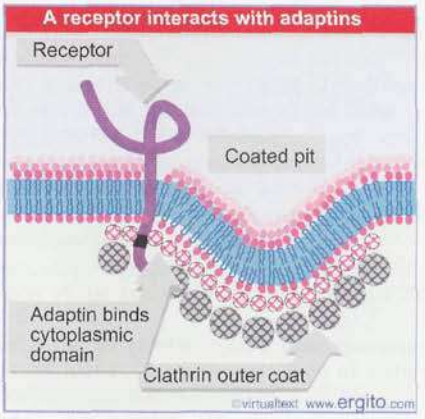


Figure 27.35 The cytoplasmic domain of an internalized receptor interacts with proteins of the inner layer of a coated pit.

27.17 Summary

Proteins that reside within the reticuloendothelial system or that are secreted from the plasma membrane enter the ER by cotranslational transfer directly from the ribosome. They are transported through the Golgi in the anterograde (forward) direction. Specific signals may cause them to be retained in the ER or a Golgi stack, or directed to other organelles such as endosomes. The default pathway is to be transported to the plasma membrane. Retrograde transport is less well characterized, but proteins that reside in the ER are retrieved from the Golgi by virtue of specific signals; an example is the C-terminal KDEL.

Proteins are transported between membranous surfaces as cargoes in membrane-bound coated vesicles. The vesicles form by budding from donor membranes; they unload their cargos by fusing with target membranes. The protein coats are added when the vesicles are formed and must be removed before they can fuse with target membranes. Anterograde transport does not result in any net flow of membrane from the ER to the Golgi and/or plasma membrane, so membrane moving with anterograde transport must be returned to the ER by a retrograde mechanism.

Modification of proteins by addition of a preformed oligosaccharide starts in the endoplasmic reticulum. High mannose oligosaccharides are trimmed. Complex oligosaccharides are generated by further modifications that are made during transport through the Golgi, determined by the order in which the protein encounters the enzymes localized in the various Golgi stacks. Proteins are sorted for different destinations in the *trans* Golgi. The signal for sorting to lysosomes is the presence of mannose-6-phosphate.

Different types of vesicles are responsible for transport to and from different membrane systems. The vesicles are distinguished by the nature of their protein coats.

COP-I-coated vesicles are responsible for retrograde transport from the Golgi to the ER. COP-I vesicles are coated with coatomer. One of the proteins of coatomer, β-COP, is related to the β-adaptin of clathrin-coated vesicles, suggesting the possibility of a common type of structure between COP-I-coated and clathrin-coated vesicles.

COP-II vesicles undertake forward movement from the ER to Golgi. Vesicles that transport proteins along the Golgi stacks have not yet been identified. Vesicles responsible for constitutive (bulk) movement from the Golgi to the plasma membrane also have not been identified. An alternative model for anterograde transport proposes that *cis*-Golgi cisternae actually become *trans*-Golgi cisternae, so that there is a continuous process of cisternal maturation from the *cis* to the *trans* face.

In the pathway for regulated secretion of proteins, proteins are sorted into clathrin-coated vesicles at the Golgi *trans* face. Some vesicles may fuse into (larger) secretory granules. Vesicles also move to endosomes, which control trafficking to the cell surface. Secretory vesicles are stimulated to unload their cargos at the plasma membrane by extracellular signals. Similar vesicles are used for endocytosis, the pathway by which proteins are internalized from the cell surface. The predominant protein in the outer coat of these vesicles is clathrin. The inner coat contains an adaptor complex, consisting of adaptin subunits, which bind to clathrin and to cargo proteins. There are (at least) three types of adaptor complex, with different specificities.

Budding and fusion of all types of vesicles is controlled by a small GTP-binding protein. This is ARF for clathrin and COP-I-coated vesicles, and Sar1P for COP-II-coated vesicles. When activated by GTP, ARF/Sar1p inserts into the membrane and causes coat proteins to assemble. This leads ultimately to budding. Further proteins, such as dynamin, may be required to "pinch off" the budding vesicle from the donor membrane. When ARF/Sar1p is inactivated because GTP is hydrolyzed to GDP, it withdraws from the membrane and the coat proteins either disassemble spontaneously (COP-coated vesicles) or are caused to do so by other proteins (clathrin-coated vesicles).

Vesicles initially recognize appropriate target membranes by a tethering reaction in which a tethering complex recognizes a Rab protein on the vesicle and brings the vesicle close to the membrane. Rabs are prenylated monomeric GTP-binding proteins. The fusion reaction is triggered when a v-SNARE on the vesicle pairs specifically with a t-SNARE on the target membrane. Pairing occurs by a coiled-coil interaction in which the SNARE complex lies parallel to the membrane surface. This causes the inner leaflets of the membranes to fuse to form a hemifusion complex; this is followed by fusion of the outer leaflets. The 20S fusion complex includes the soluble ATPase NSF and SNAP, and uses hydrolysis of ATP to release the SNAREs after pairing, which allows them to recycle.

Receptors may be internalized either continuously or as the result of binding to an extracellular ligand. Receptor-mediated endocytosis initiates when the receptor moves laterally into a coated pit. The cytoplasmic domain of the receptor has a signal that is recognized by proteins that are presumed to be associated with the coated pit. An exposed tyrosine located near the transmembrane domain is a common signal; it may be part of the sequence NPXY. When a recep-

tor has entered a pit, the clathrin coat pinches off a vesicle, which then migrates to the early endosome.

The acid environment of the endosome causes some receptors to release their ligands; the ligands are carried to lysosomes, where they are degraded, and the receptors are recycled back to the plasma membrane by means of coated vesicles. A ligand that does not dissociate may recycle with its receptor. In some cases, the receptor-ligand complex is carried to the lysosome and degraded.

References

27.1 Introduction

rev Hurtley, S. M. and Helenius, A. (1989). Protein oligomerization in the endoplasmic reticulum. Ann. Rev. Cell Biol. 5, 277-307.

Rothman, J. E. (1989). Polypeptide chain binding proteins: catalysts of protein folding and related processes in cells. Cell 59, 591-601.

27.4 Coated vesicles transport both exported and imported proteins

rev Farquhar, M. G. (1985). Progress in unraveling pathways of Golgi traffic. Ann. Rev. Cell Biol. 1, 447-488.

Pfeffer, S. R. and Rothman, J. E. (1987). Biosynthetic protein transport and sorting by the endoplasmic reticulum and Golgi. Ann. Rev. Biochem. 56, 829-852.

ref Novick, P., Field, C, and Schekman, R. (1980). Identification of 23 complementation groups required for posttranslational events in the yeast secretory pathway. Cell 21, 205-215.

Sollner, T., Whiteheart, S. W., Brunner, M., Erdjument-Bromage, H., Geromanos, S., Tempst, P., and Rothman, J. E. (1993). SNAP receptors implicated in vesicle targeting and fusion. Nature 362, 318-324.

27.5 Different types of coated vesicles exist in each pathway

rev Kirchhausen, T. (2000). Clathrin. Ann. Rev. Biochem. 69, 699-727.

Kirchhausen, T. (1999). Adaptors for clathrin-mediated traffic. Ann. Rev. Cell Dev. Biol. 15, 705-732.

Pearse, B. M. and Robinson, M. S. (1990). Clathrin, adaptors, and sorting. Ann. Rev. Cell Biol. 6, 151-171.

Schmid, S. L. (1997). Clathrin-coated vesicle formation and protein sorting: an integrated process. Ann. Rev. Biochem. 66, 511-548.

ref Barlowe, C, Orci, L., Yeung, T., Hosobuchi, M., Hamamoto, S., Salama, N., Rexach, M. F., Ravazzola, M., Amherdt, M., and Schekman, R. (1994). COP-II: a membrane coat formed by Sec proteins that drive vesicle budding from the ER. Cell 77, 895-907.

Boll, W. et al. (1999). Sequence requirements for the recognition of tyrosine-based endocytic signals by clathrin AP-2 complexes. EMBO J. 15, 5789-5795.

Collins, B. M., McCoy, A. J., Kent, H. M., Evans, P. R., and Owen, D. J. (2002). Molecular architecture and functional model of the endocytic AP2 complex. Cell 109, 523-535.

Faundez, V., Horng, J.-T., and Kelly, R. (1998). A function for the AP3 coat complex in synaptic vesicle formation from endosomes. Cell 93, 423-432.

Gallusser, A. and Kirchhausen, T. (1993). The beta 1 and beta 2 subunits of the AP complexes are the clathrin coat assembly components. EMBO J. 12, 5237-5244.

Malhotra, V., Serafini, T., Orci, L., Shepherd, J. C, and Rothman, J. E. (1989). Purification of a novel class of coated vesicles mediating biosynthetic protein transport through the Golgi stack. Cell 58, 329-36.

Miller, E., Antonny, B., Hamamoto, S., and Schekman, R. (2002). Cargo selection into COPII vesicles is driven by the Sec24p subunit. EMBO J. 21, 6105-6113.

Orci, L., Stamnes, M., Ravazzola, M., Amherdt, M., Perrelet, A., Sollner, T. H., and Rothman, J. E. (1997). Bidirectional transport by distinct populations of COP-I-coated vesicles. Cell 90, 335-349.

Rapoport, T. et al. (1999). Regulatory interactions in the recognition of endocytic sorting signals by AP-2 complexes. EMBO J. 17, 2148-2155.

Lindner R. and Ungewickell, E. (1999). Clathrin-associated proteins of bovine brain coated vesicles; an analysis of their number and assembly-promoting activity. J. Biol. Chem. 267, 16567-16573.

27.6 Cisternal progression occurs more slowly than vesicle movement

rev Griffiths, G. and Simons, K. (1986). The trans Golgi network: sorting at the exit site of the Golgi complex. Science 234, 438-443.

Rothman, J. E. (1994). Mechanisms of intracellular protein transport. Nature 372, 55-68.

Rothman, J. E. (1996). Protein sorting by transport vesicles. Science 272, 227-234.

Rothman, J. E. and Orci, L. (1992). Molecular dissection of the secretory pathway. Nature 355, 409-415.

ref Bonfanti, L. et al. (1998). Procollagen traverses the Golgi stack without leaving the lumen of cisternae: evidence for cisternal maturation. Cell 95, 993-1003.

Volchuk, A., Amherdt, M., Ravazzola, M., Brugger, B., Rivera, V. M., Clackson, T., Perrelet, A., Sollner, T. H., Rothman, J. E., and Orci, L. (2000). Megavesicles implicated in the rapid transport of intracisternal aggregates across the Golgi stack. Cell 102, 335-348.

27.7 Vesicles can bud and fuse with membranes

rev Hinshaw, J. E. (2000). Dynamin and its role in membrane fission. Ann. Rev. Cell Dev. Biol. 16, 483-519.

ref Bi, X., Corpina, R. A., and Goldberg, J. (2002). Structure of the Sec23/24-Sar1 pre-budding complex of the COPII vesicle coat. Nature 419, 271-277.

Clary, D. 0., Griff, I. C, and Rothman, J. E. (1990). SNAPs, a family of NSF attachment proteins involved in intracellular membrane fusion in animals and yeast. Cell 61, 709-21.

Ostermann, J. et al. (1993). Stepwise assembly of functionally active transport vesicles. Cell 75, 1015-1025.

van der Bliek, A. M. and Meyerowitz, E. M. (1991). Dynamin-like protein encoded by the *Drosophila shibire* gene associated with vesicular traffic. Nature 351, 41 1-414.

Wilson, D. W., Whiteheart, S. W., Wiedmann, M., Brunner, M., and Rothman, J. E. (1992). A multisubunit particle implicated in membrane fusion. J. Cell Biol. 117, 531-538.

27.8 The exocyst tethers vesicles by interacting with a Rab

rev Nuoffer, C. and Balch, W. E. (1994). GTPases: multifunctional molecular switches regulating vesicular traffic. Ann. Rev. Biochem. 63, 949-990.

Whyte, J. R. and Munro, S. (2002). Vesicle tethering complexes in membrane traffic. J. Cell Sci. 115, 2627-2637.

Zerial, M. and McBride, H. (2001). Rab proteins as membrane organizers. Nat. Rev. Mol. Cell Biol. 2, 107-117.

ref Guo, W., Roth, D., Walch-Solimena, C, and Novick, P. (1999). The exocyst is an effector for Sec4p, targeting secretory vesicles to sites of exocytosis. EMBO J. 18, 1071-1080.

Kee, Y., Yoo, J. S., Hazuka, C. D., Peterson, K. E., Hsu, S. C, and Scheller, R. H. (1997). Subunit structure of the mammalian exocyst complex. Proc. Nat. Acad. Sci. USA 94, 14438-14443.

TerBush, D. R., Maurice, T., Roth, D., and Novick, P. (1996). The Exocyst is a multiprotein complex required for exocytosis in *S. cerevisiae.* EMBO J. 15, 6483-6494.

27.9 SNARES are responsible for membrane fusion

exp Rothman, J. (2002). The SNARE complex and its role in the specificity of membrane fusion (www.ergito.com/lookup.jsp?expt=rothman2)

rev Chen, Y. A. and Scheller, R. H. (2001). SNARE-mediated membrane fusion. Nat. Rev. Mol. Cell Biol. 2, 98-106.

Jahn, R. and Sudhof, T. C. (1999). Membrane fusion and exocytosis. Ann. Rev. Biochem. 68, 863-911.

Lin, R. C. and Scheller, R. H. (2000). Mechanisms of synaptic vesicle exocytosis. Ann. Rev. Cell Dev. Biol. 16, 19-49.

ref McNew, J. A., Parlati, F., Fukuda, R., Johnston, R. J., Paz, K., Paumet, F., Sollner, T. H., and Rothman, J. E. (2000). Compartmental specificity of cellular membrane fusion encoded in SNARE proteins. Nature 407, 153-159.

Sollner, T., Whiteheart, S. W., Brunner, M., Erdjument-Bromage, H., Geromanos, S., Tempst, P., and Rothman, J. E. (1993). SNAP receptors implicated in vesicle targeting and fusion. Nature 362, 318-324.

Weber, T., Zemelman, B., McNew, J., Westermann, B., Gmachl, M., Parlati, F., Sollner, T. H., Rothman, J. E. (1998). SNAREpins: minimal machinery for membrane fusion. Cell 92, 759-772.

27.11 Protein localization depends on specific signals

rev Kornfeld, S. and Mellman, I. (1989). The biogenesis of lysosomes. Ann. Rev. Cell Biol. 5, 483-525.

von Figura, K. and Hasilik, A. (1986). Lysosomal enzymes and their receptors. Ann. Rev. Biochem. 55, 167-193.

ref Collawn, J. F. et al. (1990). Transferrin receptor internalization sequence YXRF implicates a tight turn as the structural recognition motif for endocytosis. Cell 63, 1061-1072.

Griffiths, G., Hoflack, B., Simons, K., Mellman, I., and Kornfeld, S. (1988). The mannose 6-phosphate receptor and the biogenesis of lysosomes. Cell 52, 329-341.

27.12 ER proteins are retrieved from the Golgi

exp Pelham, H. (2002). The Discovery of the KDEL Retrieval Signal (www.ergito.com/lookup.jsp?expt=pelham)

rev Pelham, H. R. (1989). Control of protein exit from the endoplasmic reticulum. Ann. Rev. Cell Biol. 5, 1-23.

Rose, J. K. and Doms, R. W. (1988). Regulation of protein export from the endoplasmic reticulum. Ann. Rev. Cell Biol. 4, 257-288.

ref Letourneur, F., Gaynor, E. C, Hennecke, S., Demolliere, C, Duden, R., Emr, S. D., Riezman, H., and Cosson, P. (1994). Coatomer is essential for retrieval of dilysine-tagged proteins to the endoplasmic reticulum. Cell 79, 1199-1207.

Lewis, M. J., Sweet, D. J., and Pelham, H. R. B. (1990). The ERD2 gene determines the specificity of the luminal ER protein retention system. Cell 61, 1359-1363.

Munro, S. and Pelham, H. R. (1987). A C-terminal signal prevents secretion of luminal ER proteins. Cell 48, 899-907.

Semenza, J. C, Hardwick, K. G., Dean, N., and Pelham, H. R. B. (1990). ERD2, a yeast gene required for the receptor-mediated retrieval of lumenal ER proteins from the secretory pathway. Cell 61, 1349-1357.

27.13 Brefeldin A reveals retrograde transport

ref Lippincott-Schwartz, J., Yuan, L. C, Bonifacino, J. S., and Klausner, R. D. (1989). Rapid redistribution of Golgi proteins into the ER in cells treated with brefeldin A: evidence for membrane cycling from Golgi to ER. Cell 56, 801-215.

27.14 Vesicles and cargos are sorted for different destinations

ref Kuehn, M. J., Schekman, R., and Ljungdahl, P. O. (1996). Amino acid permeases require COPII components and the ER resident membrane protein Shr3p for packaging into transport vesicles *in vitro.* J. Cell Biol. 135, 585-595.

Muniz, M., Morsomme, P., and Riezman, H. (2001). Protein sorting upon exit from the endoplasmic reticulum. Cell 104, 313-320.

27.15 Receptors recycle via endocytosis

rev Goldstein, J. L., Brown, M. S., Anderson, R. G., Russell, D. W., and Schneider, W. J. (1985). Receptor-mediated endocytosis: concepts emerging from the LDL receptor system. Ann. Rev. Cell Biol. 1, 1-39.

Chapter 28

Signal transduction

28.1 Introduction

The plasma membrane separates a cell from the surrounding environment. It is permeable only to small lipid-soluble molecules, such as the steroid hormones, which can diffuse through it into the cytoplasm. It is impermeable to water-soluble material, including ions, small inorganic molecules, and polypeptides or proteins.

The response of the cell to hydrophilic material in the environment depends on interactions that occur on the extracellular side of the plasma membrane. The hydrophilic material binds specifically to the extracellular domain of a protein embedded in the membrane. The extracellular molecule typically is called the ligand, and the plasma membrane protein that binds it is called the receptor.

Two fundamental types of response to an external stimulatory molecule that cannot penetrate the membrane are reviewed in **Figure 28.1:**

- *Material*—molecular or macromolecular—is physically transmitted from the outside of the membrane to the inside by transport through a proteinaceous channel in the lipid bilayer.
- *A signal* is transmitted by means of a change in the properties of a membrane protein that activates its cytosolic domain.

The physical transfer of material extends from ions to small molecules such as sugars, and to macromolecules such as proteins. Three major transport routes controlled by plasma membrane proteins are reviewed in **Figure 28.2:**

- Channels control the passage of ions: different channels exist for potassium, sodium, and calcium ions. By opening and closing in response to appropriate signals, the channels establish ionic levels within the cell (a feature of particular significance for cells of the neural network).
- One means to import small molecules is for a receptor to transport the molecule from one side of the membrane to the other. Transporters are responsible for the import of small molecules (such as sugars) across the membrane. The target molecule binds to the receptor on the extracellular side, but then is released on the cytoplasmic side.
- Ligand-binding may trigger the process of internalization, in which the receptor-ligand combination is brought into the cell by the

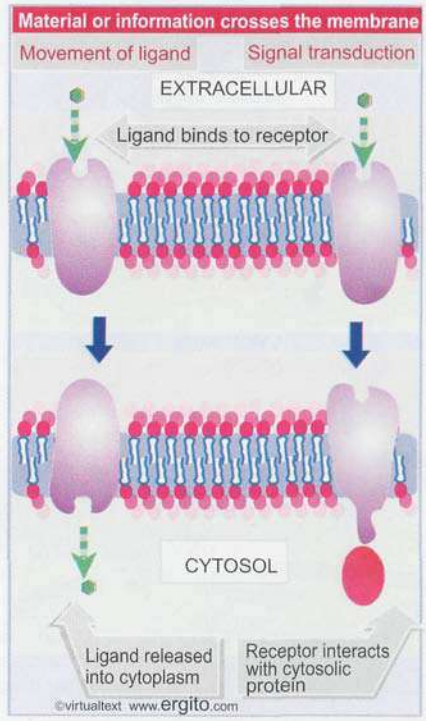


Figure 28.1 Overview: information may be transmitted from the exterior to the interior of the cell by movement of a ligand or by signal transduction.

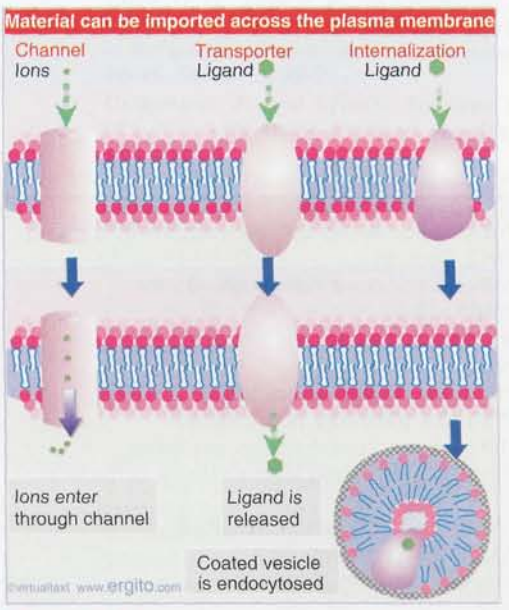


Figure 28.2 Ion channels, receptor-mediated ligand transport, and receptor internalization can transfer material into the cell.

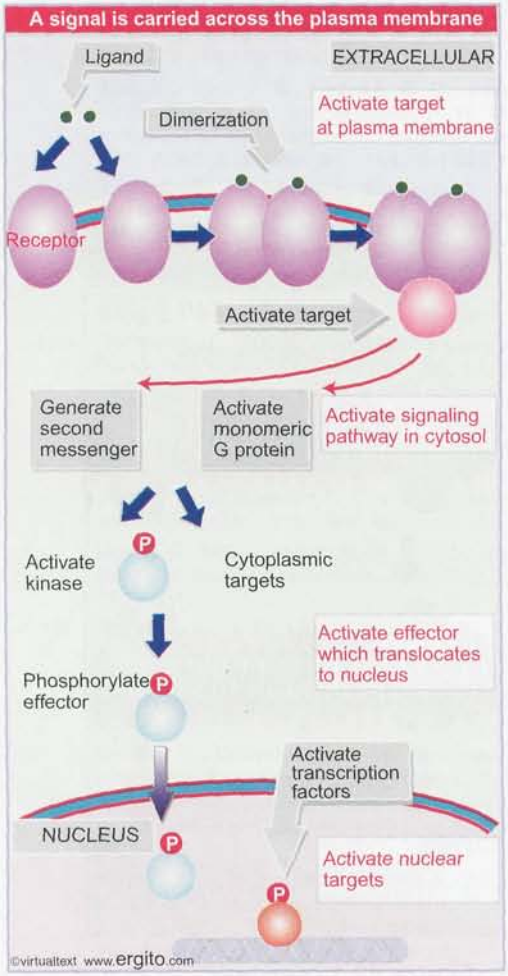


Figure 28.3 A signal transduction pathway carries a signal from the cell surface into the cytoplasm and (sometimes) into the nucleus.

process of endocytosis. In due course, the receptor and ligand are separated; the receptor may be returned to the surface for another cycle, or may be degraded. As described in *27.15 Receptors recycle via endocytosis*, endocytosis involves the passage of membrane proteins from one surface to another via coated vesicles.

The transmission of a signal involves the interaction of an extracellular ligand with a transmembrane protein that has domains on both sides of the membrane. Binding of ligand converts the receptor from an inactive to an active form. The basic principle of this interaction is that ligand binding on the extracellular side activates the receptor domain on the cytoplasmic side. The process is called signal transduction, because a signal has in effect been transduced across the membrane. The amplitude of the cytosolic signal is much greater than the original extracellular signal (the ligand), so signal transduction amplifies the original signal.

Figure 28.3 illustrates the principle of signal transduction. A receptor typically is activated by being caused to dimerize when a ligand binds to its extracellular domain. Its cytosolic domain then interacts with a protein at the plasma membrane. This interaction most often takes one of two forms. It increases the quantity of a small molecule (called a second messenger) inside the cell. Or it directly activates a protein whose role is to activate other proteins; one common type of protein that is activated early in such pathways is a monomeric GTP-binding protein.

However a signal transduction pathway is initiated, a common means of propagating the pathway through the cytosol is to activate a protein kinase, which activates a series of other protein kinases. Ultimately the signal leads to the activation of effectors, which trigger changes in the cell. Some of the effectors act in the cytosol (for example, to affect the cytoskeleton), but some carry the signal into the nucleus, where the ultimate target is the activation of transcription factors that cause new patterns of gene expression.

The major signal transduction pathways that we discuss in this chapter are extensively conserved throughout animal evolution, and can be found in most or all animal cells. They are essentially absent from plants, which have evolved different pathways for signal transduction.

Two major types of signal transduction are reviewed in **Figure 28.4:**

- The receptor has a protein kinase activity in its cytosolic domain. The activity of the kinase is activated when ligand binds to the extracellular domain. The kinase phosphorylates its own cytoplasmic domain; this autophosphorylation enables the receptor to associate with and activate a target protein, which in turn acts upon new substrates within the cell. The most common kinase receptors are tyrosine kinases, but there are also some serine/threonine kinase receptors. Such pathways most often lead to activation of a GTP-binding protein that activates a cascade of cytosolic kinases.
- The receptor may interact with a trimeric G protein that is associated with the cytosolic face of the membrane. G proteins are named for their ability to bind guanine nucleotides. The inactive form of the G protein is a trimer bound to GDP. Receptor activation causes the GDP to be replaced with GTP; as a result, the G protein dissociates into a single subunit carrying GTP and a dimer of the two other subunits. Either the monomer or the dimer then acts upon a target protein, often also associated with the membrane, which in turn reacts with a target(s) in the cytoplasm. This chain of events often stimulates the production of second messengers, the classic example being the production of cyclic AMP.

28.2 Carriers and channels form water soluble paths through the membrane

Key Concepts

- The electric gradient across the plasma membrane (inside is more negative) favors entry of cations and opposes entry of anions.
- The concentration gradient depends on the ion, typically with low intracellular levels of Na^+ and Cl^- and high levels of K^+.
- If the overall electrochemical gradient is favorable an ion can enter passively, otherwise it needs to be actively transported against the gradient.
- Carrier proteins that transport solutes across the membrane can be uniporters (one solute), symporters (two solutes) or antiporters (two solutes in opposite directions).
- Ion channels are water-soluble pores in the membrane that may be gated (controlled) by voltage or by ligands.

The impermeability of the plasma membrane to water-soluble compounds enables different aqueous conditions to be maintained on either side. The ionic environment of the cytosol is quite different from the extracellular ionic milieu. Within the cytoplasm, different organelles offer different ionic environments. A striking example is the maintenance of an acid pH in endosomes and lysosomes, with immediate implications for the functions of the proteins that enter them (see *27.15 Receptors recycle via endocytosis).* Another example is the maintenance of a store of Ca^{2+} in the endoplasmic reticulum. (The nucleus is an exception to the rule that conditions in membrane-bounded organelles usually differ from the cytosol. The nucleoplasm essentially is subjected to the same conditions as the cytosol. This happens because the nuclear pores form relatively large openings in the nuclear envelope, through which ions and other small molecules can diffuse freely.)

A notable feature of the cytosolic environment is that there are more free cations (positively charged) (~150 mM) than anions (~10 mM). The reason is that many cellular constituents are negatively charged—for example, nucleic acids have multiple negative charges for every phosphate group in the phosphodiester backbone. The superfluity of cations therefore establishes electrical neutrality by balancing these fixed charges.

The intracellular concentrations of Na^+ and Cl^- are low (~10 mM) while those outside the cell are high (>100 mM); and the situation for K^+ is reversed. This creates a concentration gradient across the membrane for each ion.

The plasma membrane is electrically charged (due to the different phospholipid compositions of the inner and outer leaflets). There is an electrical gradient in which the inside is negative compared to the outside. This voltage difference favors the entry of cations and opposes the entry of anions.

Together the concentration gradient and electrical gradient constitute the electrochemical gradient, which is characteristic for each solute. A solute whose gradient is favorable can enter the cell when a channel opens; the gradient is sufficient to drive passive transport of a solute such as Na^+ or Cl^- into the cell. But a solute that faces an unfavorable gradient requires active transport in which energy is used to pump it into the cell against the gradient.

The passage of ions (and other small solutes) through the plasma membrane is mediated by resident transmembrane proteins. A common feature of these proteins is their large size and the presence of multiple

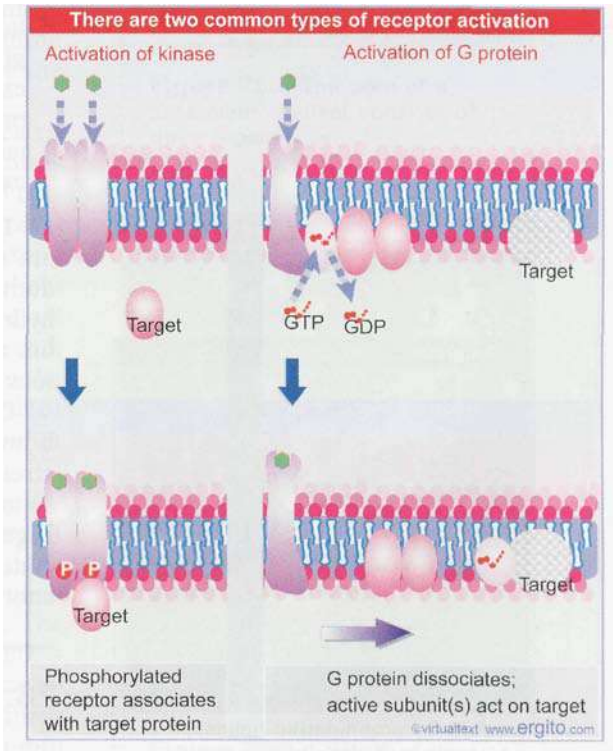


Figure 28.4 A signal may be transduced by activating the kinase activity of the cytoplasmic domain of a transmembrane receptor or by dissociating a G protein into subunits that act on target proteins in the membrane.

By Book_Crazy [IND]

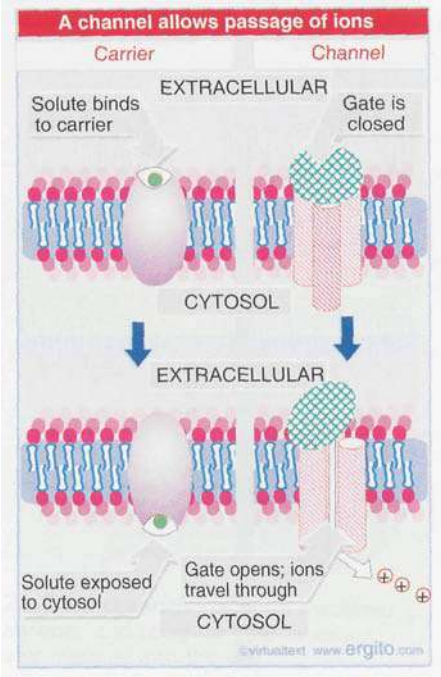


Figure 28.5 A carrier (porter) transports a solute into the cell by a conformational change that brings the solute-binding site from the exterior to the interior, while an ion channel is controlled by the opening of a gate.

membrane-spanning regions, features which together argue that they provide a relatively static feature of the membrane. **Figure 28.5** illustrates two general means of transport across the membrane:

- A carrier protein binds a solute on one side of the membrane and then experiences a conformational change that transports the solute to the other side of the membrane. By binding the solute on one side and releasing it on the other, the carrier in effect directly transports the solute across the membrane. Several types of carriers are distinguished by the number of solutes that they transport, and the directions in which they transport them. Carriers that transport a single solute across the membrane are called uniporters; carriers that simultaneously or sequentially transport two different solutes are called symporters; and carriers that transport one solute in one direction while transporting a different solute in the opposite direction are called antiporters. Carrier proteins may be used for passive transport or linked to an energy source to provide active transport. Energy for active transport is provided by hydrolysis of ATP, the classic example being the Na^+-K^+ pump that functions as an antiporter, pumping sodium out of the cell and potassium into it. Another source of energy is the electrochemical gradient itself; a symporter brings Na^+ into the cell together with some other solute, using the favorable gradient of sodium to overcome the unfavorable gradient of the other solute.
- An ion channel comprises a water-soluble pore in the membrane. Its activity is controlled by regulation of the opening and closing of the channel. When it is open, ions can diffuse passively, as driven by the electrochemical gradient. Ion channels allow *only* passive transport. The resting state of an ion channel is closed, and the gates that control channel activity usually open only briefly, in response to a specific signal. Ligand-gated channels are receptors that respond to binding of particular molecules, amongst which the neurotransmitters acetylcholine, glycine, GABA (γ-amino-butyric acid), and glutamate are prominent examples. Voltage-gated channels respond to electric changes, again a prominent feature of the neural system. Second messenger gated channels provide yet another means for signal transduction, one interesting example comprising channels that respond to activation of G proteins.

The structures of both carriers and channels present a paradox. They are transmembrane proteins that have multiple membrane-spanning domains, each consisting of a stretch of amino acids of sufficient hydrophobicity to reside in the lipid bilayer. Yet within these hydrophobic regions must be a highly selective, water-filled path that permits ions to travel through the membrane.

One solution to this problem lies in the structure of the transmembrane regions. Instead of comprising unremittingly hydrophobic stretches like those of single membrane-pass proteins, they contain some polar amino acids. They are likely to be organized as illustrated in **Figure 28.6** as amphipathic helices in which the hydrophobic face associates with the lipid bilayer, while the polar faces are aligned with one another to create the channel.

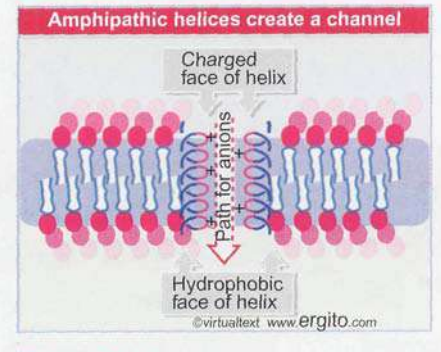


Figure 28.6 A channel may be created by amphipathic helices, which present their hydrophobic faces to the lipid bilayer, while juxtaposing their charged faces away from the bilayer. In this example, the channel is lined with positive charges, which would encourage the passage of anions.

28.3 Ion channels are selective

Key Concepts

- Channels typically consist of several protein subunits with the water-soluble pore at the axis of symmetry.
- Selectivity is determined by the properties of the pore.
- The gate acts by a mechanism resembling a ball and chain.

The importance of the interior of the channel is indicated by the ion selectivity. Different channels permit the passages of different ions or groups of ions. The channels are extremely narrow, so ions must be stripped of their associated water molecules in order to pass through. The channel possesses a "filter" at the entrance to the pore that has specificity for its particular ion, presumably based upon its geometry and electrostatic charge.

The structures of particular ion channels are beginning to reveal their general features. A common feature is that the constituent proteins are large and have several membrane-spanning regions. A channel probably consists of a "ring" of 4, 5, or 6 subunits, organized in a symmetrical or quasi-symmetrical manner. The water-filled pore is found at the central axis of symmetry. The size of the pore generally increases with the number of subunits in the ring. The subunits are always related in structure, and sometimes are identical. They may consist of separate proteins or of related domains in a single large protein.

Voltage-gated sodium channels have a single type of subunit, a protein of 1820 amino acids with a repetitive structure that consists of 4 related domains. Each domain has several membrane-spanning regions. The four domains are probably arranged in the membrane in a pseudo-symmetrical structure. Two smaller subunits are associated with the large protein.

Potassium channels have a smaller subunit, equivalent to one of the domains of the sodium channel; four identical subunits associate to create the channel. Six transmembrane domains are identified in the protein subunit by hydrophobicity analysis; they are numbered S1-S6. The S4 domain has an unusual structure for a transmembrane region: it is highly positively charged, with arginine or lysine residues present at every third or fourth position. The S4 motif is found in voltage-gated K^+, Na^+, and Ca^{2+} channels, so it seems likely that it is involved with a common property, thought to be channel opening. Some potassium channels have only the S5-S6 membrane-spanning domains, and they appear to be basically shorter versions of the protein.

Analysis of the *shaker* potassium channel of the fly has revealed some novel features, illustrated in **Figure 28.7.** The region that forms the pore has been identified by mutations that alter the response to toxins that inhibit channel function. It occupies the region between transmembrane domains S5 and S6, forming two membrane-spanning stretches that are not organized in the usual hydrophobic α-helical structure. The structure could be a rather extended β-hairpin. The state of the channel (open or closed) is controlled by the N-terminal end, which resembles a ball on a chain. The ball is in effect tethered to the channel by a chain, and plugs it on the cytoplasmic side. The length of the chain controls the rate with which the ball can plug the channel after it has been opened.

A major question about potassium channels is how their selectivity is maintained. K^+ and Na^+ ions are (positively charged) spheres of 1.33 A and 0.95 A, respectively. K^+ ions are selected over Na^+ ions by a margin of $10^4\times$, but at the same time, up to 10^8 ions per second can move through the pore, basically close to the diffusion limit. The salient features of the pore of a potassium channel, based on the crystal structure, are summarized in **Figure 28.8,** and shown as a cutaway model in **Figure 28.9.** The pore is ~45 A long and consists of three regions. It starts inside the cell with a long internal pore, opens out into a central cavity of ~10 A diameter, and then passes to the extracellular space with a narrow selectivity filter. The lining of the inner pore and central cavity is hydrophobic, providing a relatively inert surface to a diffusing potassium ion. The central cavity is aqueous, and may serve to lower the electrostatic barrier to crossing the membrane (which is at its maximum in the center). The selectivity filter has negative charges

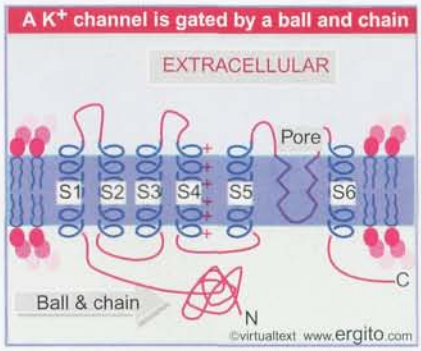


Figure 28.7 A potassium channel has a pore consisting of unusual transmembrane regions, with a gate whose mechanism of action resembles a ball and chain.

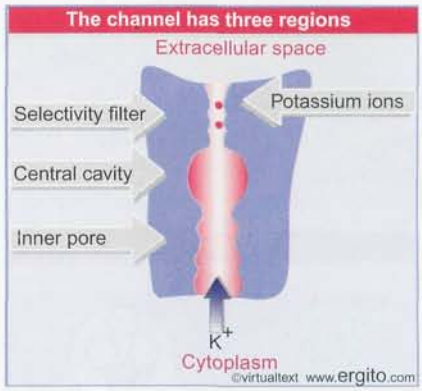


Figure 28.8 The pore of a potassium channel consists of three regions.

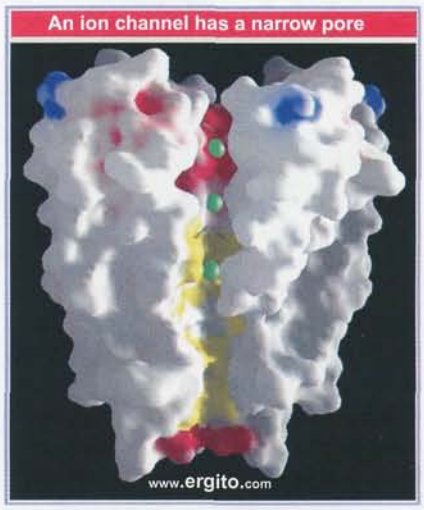


Figure 28.9 A model of the potassium channel pore shows electrostatic charge (blue = positive, white = neutral, red = negative) and hydrophobicity (= yellow). Photograph kindly provided by Rod MacKinnon.

and is lined with the polypeptide backbone. When a K^+ ion loses its hydrating water on entering the filter, the contacts that it made with the water will be replaced by contacts with the oxygens of the polypeptide carbonyl groups. The size of the pore may be set so that a smaller sodium ion would not be close enough to make these substitute contacts.

28.4 Neurotransmitters control channel activity

Key Concepts

- Neurotransmitter-gated receptors are ion channels that are controlled by neurotransmitters such as acetylcholine, glycine, or GABA.
- The nicotinic acetylcholine receptor is a 5-subunit ion channel that admits several cations but is largely used to control Na^+ uptake by the cell.

Neurotransmitter-gated receptors form a superfamily of related proteins in the 5-member channel class. The nicotinic acetylcholine receptor has been characterized in the most detail, and is a pentamer with the structure $\alpha_2\beta\delta\gamma$. As illustrated in **Figure 28.10,** the bulk of the 5 subunits projects above the plasma membrane into the extracellular space. The openings to the channel narrow from a diameter of ~25 A until reaching the pore itself. The entrance on the extracellular side is very deep, ~60 A; the distance on the cellular side is shorter, 20 A. The pore extends through the 30 A of the lipid bilayer and is only ~7 A in diameter.

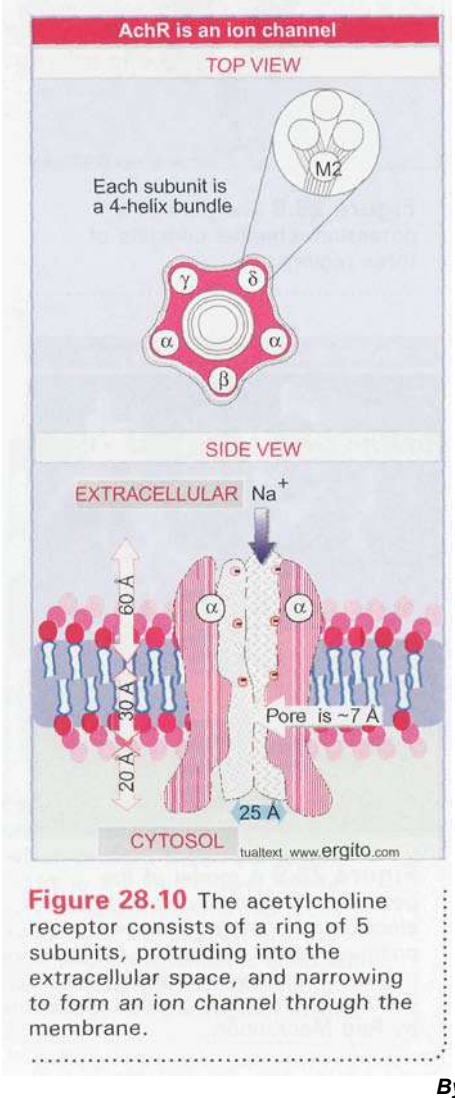


Figure 28.10 The acetylcholine receptor consists of a ring of 5 subunits, protruding into the extracellular space, and narrowing to form an ion channel through the membrane.

Ligand binding occurs on the a subunits. Both a subunits must bind an acetylcholine for the gate to open. Where is the gate? Since the channel is really narrow only in the region within the lipid bilayer, the gate seems likely to be located well within the receptor. Structural changes that occur upon opening seem greatest just by the cytoplasmic side of the lipid bilayer, so it is possible that the gate is located at the level of the phospholipid heads on the cytoplasmic boundary. So the acetylcholine receptor, like many other receptors, must transmit information about ligand binding internally, from the extracellular acetylcholine binding site to the near-cytoplasmic gate.

How does the gate function? It might consist of an electrostatic repulsion, in which positive groups are extruded into the channel to prevent passage of cations. Or it may take the form of a physical impediment to passage, in which a conformational change brings bulky groups to block the pore.

Ion selectivity may be determined by the walls of the wide entry passage. The walls lining the entrances to the pore have negatively charged groups; each subunit carries ~10 negative charges in its extracellular region. These charge clusters could modify the ionic environment at the entrance to the channel, concentrating the desired ions and diluting ions that are selected against. The structure of the acetylcholine receptor allows passage of Na^+, K^+, or Ca^{2+} ions, but because of the prevailing gradients, its main use in practice is to allow the entry of Na^+ into the cell.

The acetylcholine receptor is an example of a superfamily of receptors gated by neurotransmitters. All appear to have the same general organization, consisting of 5 subunits whose structures are related to one another. All the subunits are about the same size (~50 kD), and each is probably organized in the membrane as a bundle of 4 helices (each helix containing a transmembrane domain). In each case, one of the

four transmembrane domains (called M2) has an amphipathic structure and seems likely to be involved in lining the walls of the pore itself. The presence of serine and threonine residues, and some paired acid-basic residues, may assist ion passage. The sequences of subunits of the glycine and GABA receptors are related to the acetylcholine receptor subunits. Some changes in the sequences seem likely to reflect the ion selectivity. So the glycine and GABA receptors have positively charged groups in the entrance walls, consistent with their transport of anions such as Cl^-.

28.5 G proteins may activate or inhibit target proteins

Key Concepts

- Ligand binding to a serpentine membrane receptor causes it to activate a G protein.
- The G protein is a trimer bound to GDP in its inactive state.
- The mechanism of activation is that the receptor causes the GDP bound by the G-protein to be replaced with GTP.

G proteins transduce signals from a variety of receptors to a variety of targets. The components of the general pathway can be described as

- The receptor is a resident membrane protein that is activated by an extracellular signal.
- A G protein is converted into active form when an interaction with the activated receptor causes its bound GDP to be replaced with GTP.
- An effector is the target protein that is activated (or—less often—inhibited) by the G protein; sometimes it is another membrane-associated protein.
- Second messengers are small molecules that are released as the result of activation of (certain types) of effectors.

Another terminology that is sometimes used to describe the relationship of the components of the transduction pathway is to say that the receptor is *upstream* of the G protein, while the effector is *downstream.*

The effectors linked to different types of G proteins are summarized in **Figure 28.11.** The important point is that there is a large variety of G proteins, activated by a wide variety of receptors. The activation of an individual G protein may cause it to stimulate or to inhibit a particular effector; and some G proteins act upon multiple effectors (causing the activation in turn of multiple pathways). Two of the classic G proteins are G_s, which stimulates adenylate cyclase (increasing the level of cAMP), and G_t, which stimulates cGMP phosphodiesterase (decreasing

Figure 28.11 Classes of G proteins are distinguished by their effectors and are activated by a variety of transmembrane receptors.

By Book_Crazy [IND]

the level of cGMP). The cyclic nucleotides are a major class of second messengers; another important group consists of small lipid molecules, such as inositol phosphate or DAG.

Although the receptors that couple to G proteins respond to a wide variety of ligands, they have a common type of structure and mode of binding the ligand. They are serpentine receptors with 7 transmembrane regions, and they function as monomers. The greatest conservation of sequence is found in hydrophobic transmembrane regions, which in fact are used to classify the serpentine receptors into individual families.

The binding sites for small hydrophobic ligands lie in the transmembrane domains, so that the ligand becomes bound in the plane of the membrane. The smallest ligands, such as biogenic amines, may be bound by a single transmembrane segment. Larger ligands, such as extended peptides, may have more extensive binding sites in which extracellular domains provide additional points of contact. Large peptide hormones may be bound mainly by the extracellular domains.

When the ligand binds to its site, it triggers a conformational change in the receptor that causes it to interact with a G protein. A well-characterized (although not typical) case is that of rhodopsin, which contains a retinal chromophore covalently linked to an amino acid in a transmembrane domain. Exposure to light converts the retinal from the 11-*cis* to the all *trans* conformation, which triggers a conformational change in rhodopsin that causes its cytoplasmic domain to associate with the G_t protein (transducin).

28.6 G proteins function by dissociation of the trimer

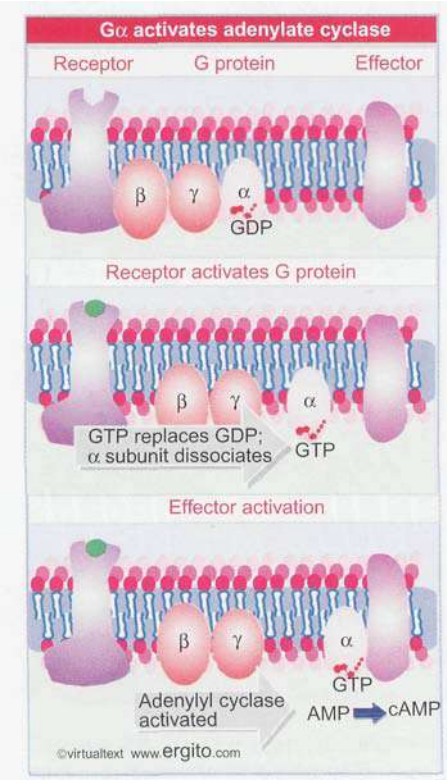


Figure 28.12 Activation of G_s causes the α subunit to activate adenylate cyclase.

Key Concepts

- When GDP is replaced by GTP, a trimeric G protein dissociates into an α-GTP subunit and a βγ dimer.
- It is most often the a subunit that activates the next component (the effector) in the pathway.
- Less often the βγ activates the effector.

G proteins are trimers whose function depends on the ability to dissociate into an a monomer and a βγ dimer. The dissociation is triggered by the activation of an associated receptor. In its inactive state, the a subunit of the G protein is bound to GDP. **Figure 28.12** shows that the activated receptor causes the GDP to be replaced by GTP. This causes the G protein to dissociate into a free a-GTP subunit and a free *fiy* dimer. G proteins are found in all classes of eukaryotes.

The interaction between receptor and G protein is catalytic. After a G protein has dissociated from an activated receptor, the receptor binds another (inactive) trimer, and the cycle starts again. So one ligand-receptor complex can activate many G protein molecules in a short period, amplifying the original signal.

The most common action for the next stage in the pathway calls for the activated a subunit to interact with the effector. In the case of G_s, the α_s subunit activates adenylate cyclase; in the case of G_t, the α_t subunit activates cGMP phosphodiesterase. In other cases, however, it is the βγ dimer that interacts with the effector protein. In some cases, *both* the a subunit and the βγ dimer interact with effectors.

Consistent with the idea that it is more often the a subunits that interact with effectors, there are more varieties of a subunits (16 known

in mammals) than of β (5) or γ subunits (11). However, irrespective of whether the a or βγ subunits carry the signal, *the common feature in all of these reactions is that a G protein usually acts upon an effector enzyme that in turn changes the concentration of some small molecule(s) in the cell.* (There are some other pathways in which G proteins behave by activating a kinase.)

In either the intact or dissociated state, G proteins are associated with the cytoplasmic face of the plasma membrane. But the individual subunits are quite hydrophilic, and none of them appears to have a transmembrane domain. The βγ dimer has an intrinsic affinity for the membrane because the 7 subunit is prenylated. The α_i and α_o types of subunit are myristoylated, which explains their ability to remain associated with the membrane after release from the βγ dimer. The α_s subunit is palmitoylated.

Because several receptors can activate the same G proteins, and since (at least in some cases) a given G protein has more than one effector, we must ask how specificity is controlled. The most common model is to suppose that receptors, G proteins, and effectors all are free to diffuse in the plane of the membrane. In this case, the concentrations of the components of the pathway, and their relative affinities for one another, are the important parameters that regulate its activity. We might imagine that an activated a-GTP subunit scurries along the cytoplasmic face of the membrane from receptor to effector. But it is also possible that the membrane constrains the locations of the proteins, possibly in a way that restricts interactions to local areas. Such compartmentation could allow localized responses to occur.

28.7 Protein kinases are important players in signal transduction

Key Concepts

- Protein kinases fall into groups that phosphorylate Ser/Thr or Tyr on target proteins.
- Receptor protein kinases are most often protein tyrosine kinases.
- Cytosolic protein kinases are most often protein Ser/Thr kinases.

There are many types of protein kinases involved in signal transduction. They all have the same basic catalytic activity: they add a phosphate group to an amino acid in a target protein. The phosphate is provided by hydrolyzing ATP to ADP. A protein kinase has an ATP-binding site and a catalytic center that can bind to the target amino acid. The phosphorylation of the target protein changes its properties so that it in turn acts to carry the signal transduction pathway to the next stage.

Protein kinases can be classified both by the types of amino acids that they phosphorylate in the protein target and by their location in the cell.

Three groups of protein kinases are distinguished by the types of amino acid targets:

- Protein serine/threonine kinases are responsible for the vast majority of phosphorylation events in the cell. As their name indicates, they phosphorylate either serine or threonine in the target protein.
- Protein tyrosine kinases phosphorylate tyrosine in the target protein.
- Dual specificity kinases are less common and can phosphorylate target proteins on either tyrosine or serine/threonine.

Protein kinases are found in two types of location:

Figure 28.13 The active site of a cytosolic kinase is identified by the catalytic loop (green). The target sequence for phosphorylation binds to the active site. Amino acid 0 is phosphorylated. The amino acid on the N-terminal side (indicated by N) is immediately adjacent to the phosphorylation lip, which is a sequence in the kinase that is often itself phosphorylated.

- *Cytosolic protein kinases* are most often Ser/Thr protein kinases. They are responsible for the vast majority of phosphorylation events in the cell. One particularly important class are the cdk (cyclin-dependent kinase) enzymes that control the cell cycle (see *29 Cell cycle and growth regulation).* Dual specificity kinases are found in the MAP kinase signal transduction pathway (see *28.16 A MAP kinase pathway is a cascade).* The products of some oncogenes, for which *src* is the paradigm, are protein tyrosine kinases (see *30.16 Src is the prototype for the proto-oncogenic cytoplasmic tyrosine kinases).*
- *Receptor protein kinases* are found in the plasma membrane. They have a domain on the exterior of the cell that binds a ligand, and a catalytic domain within the cell that can act on a target protein. Most receptors with protein kinase activity are protein tyrosine kinases (abbreviated as RTK for receptor tyrosine kinase), although there are also some receptors of the Ser/Thr kinase class.

All kinases have an active site that binds ATP and a short sequence of the target protein that includes the amino acid to be phosphorylated. The sequence bound at the active site usually conforms to a consensus, typically 3-4 amino acids long. Recognition of the target protein also depends on interactions involving other regions of both the kinase and the target.

Figure 28.13 illustrates the structure of a dual specificity kinase of the MAP kinase family, based on its crystal structure. The active site consists of a short loop of the enzyme that forms a deep cleft. The sequence of the catalytic loop is generally conserved. ATP binds at the bottom of the cleft. Adjacent to the catalytic loop, on the surface of the enzyme, is a sequence called the phosphorylation lip; many kinases in this group have amino acids in this sequence whose phosphorylation activates the enzyme activity. The phosphorylation lip contacts the amino acid on the N-terminal side of the amino acid that is phosphorylated.

Receptor tyrosine kinases have some common features. The extracellular domain often has characteristic repeating motifs. It contains a ligand-binding site. The transmembrane region is a single short membrane-spanning alpha helix. The catalytic domain is large (~250 amino acids), and often occupies the bulk of the cytoplasmic region. Certain conserved features are characteristic of all kinase catalytic domains. Sometimes the catalytic domain is broken into two parts by an interruption of some other sequence (which may have an important function in selecting the substrate).

Figure 28.14 illustrates the features of a receptor tyrosine kinase. Because the receptors are embedded in membranes, we do not yet have crystal structures of intact proteins. However, several extracellular and cytoplasmic domains have been independently crystallized. The extracellular and cytoplasmic domains of the RTK group both show large variations in size. The receptors are usually activated by binding a polypeptide ligand, which can be a significant size relative to the extracellular domain. The cytoplasmic domain of the RTK is large, and contains many sites involved in signaling, as well as the kinase catalytic domain. Crystal structures have identified the features of the active site. At the active site, the catalytic loop is adjacent to an activation loop, which contains 2-3 tyrosines. When these tyrosines are phosphorylated, the activation loop swings away from the catalytic loop, freeing it to bind the substrate.

Phosphorylation usually activates the target protein, but this is not a golden rule—there are some cases in which phosphorylation inhibits the activity of the target. One way to reverse the effects of a phosphorylation event is for a phosphatase (typically a cytosolic phosphatase) to remove the phosphate that was added by a protein kinase. There are phosphatases with specificity for the appropriate amino acids to match each type of kinase. Most phosphatases are cytosolic, although there are some receptor phosphatases.

28.8 Growth factor receptors are protein kinases

Key Concepts

- Binding of a ligand to the extracellular domain of a growth factor receptor activates the kinase activity of the cytoplasmic domain.
- The receptor may activate a second messenger or may activate a cascade of kinases.

Growth factor receptors take their names from the nature of their ligands, which usually are small polypeptides (casually called growth factors, more properly called cytokines) that stimulate the growth of particular classes of cells. The factors have a variety of effects, including changes in the uptake of small molecules, initiation or stimulation of the cell cycle, and ultimately cell division. The ligands most usually are secreted from one cell to act upon the receptor of another cell. Examples of secreted cytokines are EGF (epidermal growth factor), PDGF (platelet-derived growth factor), and insulin. In some cases, ligands instead take the form of components of the extracellular matrix, or membrane proteins on the surface of another cell (these are sometimes called counter-receptors).

The receptors share a general characteristic structure: they are group I integral membrane proteins, spanning the membrane once, with an N-terminal protein domain on the extracellular side of the membrane, and the C-terminal domain on the cytoplasmic side. Most receptors, such as those for EGF or PDGF, consist of single polypeptide chains. An exception is provided by receptors of the insulin family, which are disulfide-bonded dimers (each dimer being a group I protein).

The effector pathways that are activated by receptor tyrosine kinases *(RTKs)* fall into two groups:

- *An enzymatic activity is activated that leads to the production of a small molecule second messenger.* The second messenger may be the immediate product of an enzyme that is activated directly by the receptor, or may be produced later in the pathway. Lipids are common second messengers in these pathways. The enzymes include phospholipases (which cleave lipids from larger substrates) and kinases that phosphorylate lipid substrates. Some common pathways are summarized in **Figure 28.15.** The second messengers that are released in each pathway act in the usual way to activate or inactivate target proteins.

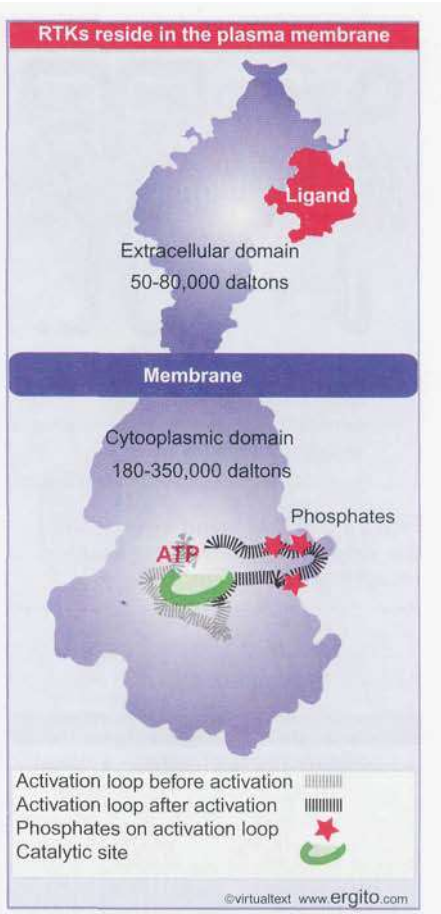


Figure 28.14 A receptor tyrosine protein kinase has an extracellular domain that binds a ligand, which is usually a polypeptide, and a cytoplasmic domain that includes a kinase region. ATP binds adjacent to the catalytic site. An important feature is the activation loop, which contains tyrosine residues whose phosphorylation activates the kinase activity. The activation loop containing these tyrosines changes its position when they are phosphoryated.

RTKs act on different types of target proteins that generate second messengers

Effector	Substrate	Products
PLC (phospolipase C) (3 families, PLCα,β,γ)	PIP2 (phosphatidylinositol 4,5-diphosphate) →	DAG (diacylglycerol) + IP3 (inositol 1,4,5-triphosphate) DAG activates protein kinase C IP3 mobilizes Ca^{2+}
PLA2 (phospholipase A2)	Phospholipids (phosphatidylcholine phosphatidylethanolamine phosphatidylinositol) →	Arachidonic acid Converted to prostaglandins & leukotriennes
PI3 kinase (phosphatidylinositol-3 kinase)	Phosphatidyl inositol →	PI3 (phosphatidyl inositol-3 phosphate)
PI4 kinase (phosphatidylinositol-4 kinase)	Phosphatidyl inositol →	PI4 (phosphatidyl inositol-4 phosphate) Converted to PIP2 (phosphatidyl diphosphate)

©virtualtext www.ergito.com

Figure 28.15 Effectors for receptor tyrosine kinases include phospholipases and kinases that act on lipids to generate second messengers.

By Book_Crazy [IND]

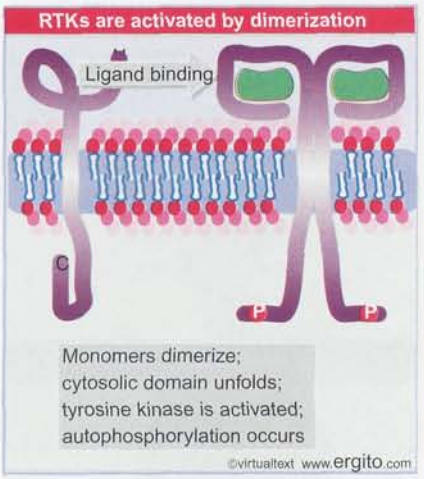


Figure 28.16 Ligand binding to the extracellular domain causes receptor dimerization.

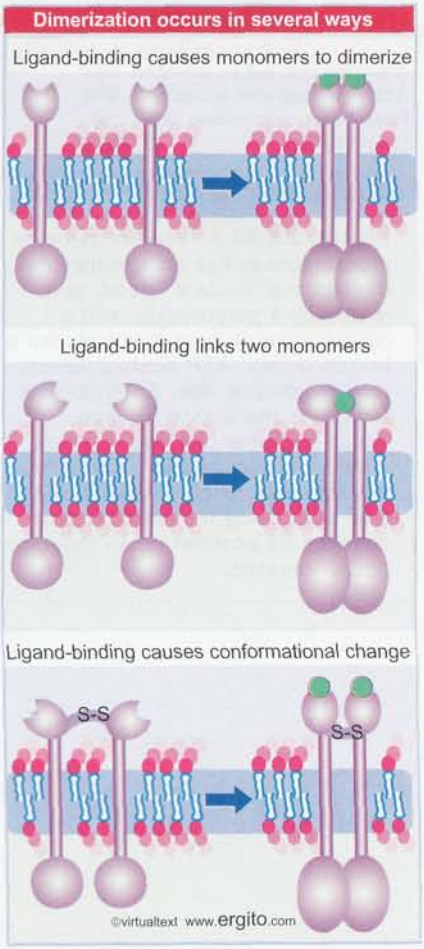


Figure 28.17 Binding of ligand to the extracellular domain can induce aggregation in several ways. The common feature is that this causes new contacts to form between the cytoplasmic domains.

- *The effector pathway is a cascade that involves a series of interactions between macromolecular components.* The most common components of such pathways are protein kinases; each kinase activates the next kinase in the pathway by phosphorylating it, and the ultimate kinases in the pathway typically act on proteins such as transcription factors that may have wide-ranging effects upon the cell phenotype.

The basic principle underlying the function of all types of effector pathway is that the signal is amplified as it passes from one component of the pathway to the next. When some components have multiple targets, the pathway branches, thus creating further diversity in the response to the original stimulus.

When a ligand binds to the extracellular domain of a growth factor receptor, the catalytic activity of the cytoplasmic domain is activated. *Phosphorylation of tyrosine is identified as the key event by which the growth factor receptors function because mutants in the tyrosine kinase domain are biologically inactive, although they continue to be able to bind ligand.*

28.9 Receptors are activated by dimerization

Key Concepts

- Ligand binding to receptor monomers causes them to dimerize by interactions between the extracellular domains.
- Dimerization is made possible by the ability of membrane proteins to move laterally within the membrane bilayer.
- Dimerization activates the cytoplasmic domains by an autophosphorylation in which the kinase activity of each monomer phosphorylates the other monomer.

A key question in the concept of how a signal is transduced across a membrane is how binding of the ligand to the extracellular domain activates the catalytic domain in the cytoplasm. *The general principle is that a conformational change is induced that affects the overall organization of the receptor.* An important factor in this interaction is that membrane proteins have a restricted ability to diffuse laterally (in contrast with the continuous motion of the lipids in the bilayer). This enables their state of aggregation to be controlled by external events.

Lateral movement plays a key role in transmitting information from one side of the membrane to the other. **Figure 28.16** shows that binding of ligand induces a conformation change in the N-terminal region of a group I receptor that causes the extracellular domains to dimerize. This causes the transmembrane domains to diffuse laterally, bringing the cytoplasmic domains into juxtaposition. The stabilization of contacts between the C-terminal cytosolic domains causes a change in conformation that activates the kinase activity. In some cases, phosphorylation also causes the receptor to interact with proteins present on the cytoplasmic surface of a coated pit, leading to endocytosis of the receptor. An extreme case of lateral diffusion is seen in certain cases of receptor internalization, when receptors of a given type aggregate into a "cap" in response to an extracellular stimulus.

Figure 28.17 shows that dimerization can take several forms. The most common is that a ligand binds to one or to both monomers to induce them to dimerize. A variation is that a dimeric ligand binds to two

monomers to bring them together. In the case of the insulin receptor family, the ligand binds to a dimeric receptor (which is stabilized by extracellular disulfide bridges) to cause an intramolecular change of conformation. The major consequence of dimerization is to allow transmission of a conformational change from the extracellular domain to the cytoplasmic domain without requiring a change in the structure of the transmembrane region.

Dimerization initiates the signaling pathway by triggering an autophosphorylation in the cytoplasmic domains of the receptor. When the two cytoplasmic domains are brought together in the dimer, each phosphorylates the other. It is necessary for *both* subunits to have kinase activity for the receptor to be activated; if one subunit is defective in kinase activity, the dimer cannot be activated.

Autophosphorylation has two consequences. Phosphorylation of tyrosines in the kinase domain causes the "activation **loop" to** swing away from the catalytic center, thus activating the ability of the kinase to bind its substrate (see Figure 28.14). Phosphorylation of tyrosines at other regions of the cytoplasmic domain provides the means by which substrate proteins are enabled to bind to the receptor. The existence of these phosphorylated tyrosine(s) in specific signaling motifs causes the cytoplasmic domain to associate with its target proteins.

28.10 Receptor kinases activate signal transduction pathways

Key Concepts

- Receptor activation causes phosphorylation of Tyr at several short sequence motifs in the cytoplasmic domain.
- Different substrate proteins bind to particular motifs.
- The substrate proteins may be docking proteins that bind other proteins, or signaling proteins that have an enzymatic activities that are activated by associating with the receptor.

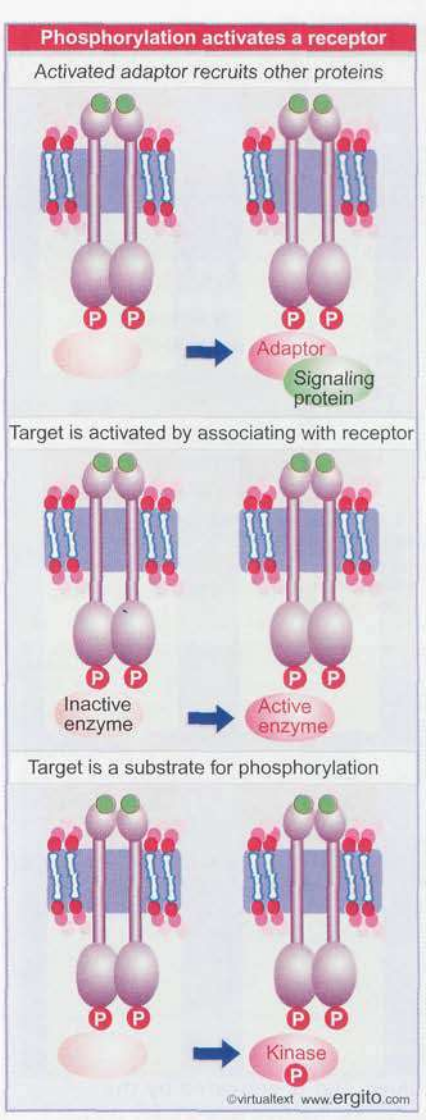


Figure 28.18 Phosphorylation of the intracellular domain of a receptor creates sites that bind cytosolic proteins.

Figure 28.18 shows that we can distinguish several types of proteins with which the activated receptor may interact:

- The protein may be an intermediary that has no catalytic activity of its own, but serves merely to bring other proteins to the receptor. "Docking proteins" or "adaptors" bind to an activated receptor, and then other proteins(s) bind to them, and may therefore become substrates for the receptor. By assembling complexes via such intermediaries, receptors can extend their range.
- The protein may be a *target* that is activated by its association with the receptor, but which is not itself phosphorylated. For example, some enzymes are activated by binding to a receptor, such as PI3 kinase (see Figure 28.15).
- If the protein is a *substrate* for the enzyme, it becomes phosphorylated. If the substrate is itself an enzyme, it may be activated by the phosphorylation (example: c-Src or PLC7; see Figure 28.15). Sometimes the substrate is a kinase, and the pathway is continued by a cascade of kinases that successively activate one another.
- Some substrates may be end-targets, such as cytoskeletal proteins, whose phosphorylation changes their properties, and causes assembly of a new structure.

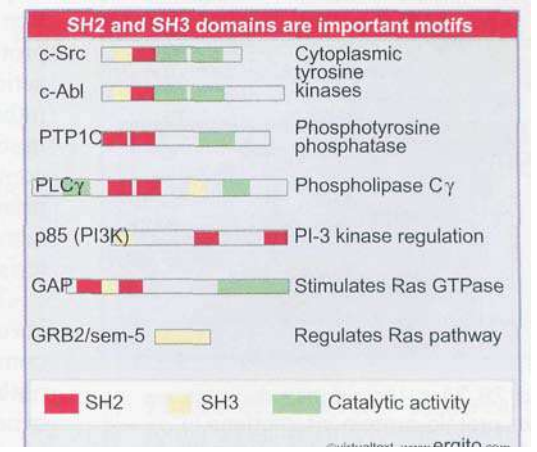


Figure 28.19 Several types of proteins involved in signaling have SH2 and SH3 domains.

A receptor tyrosine kinase can initiate a signaling cascade at the membrane. However, in many cases, the activation of the kinase is

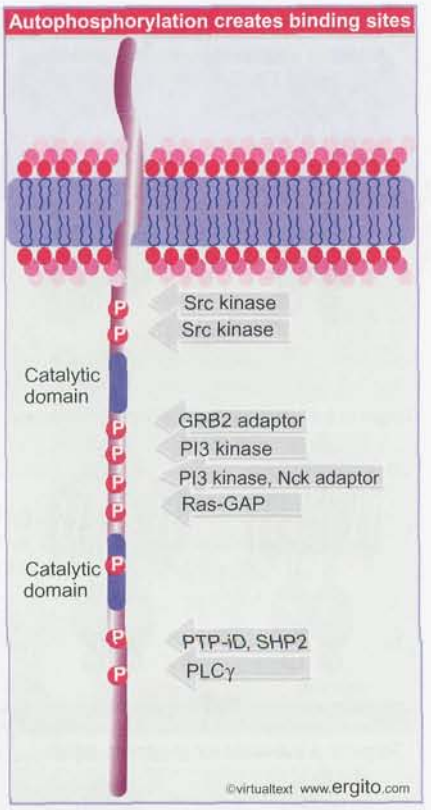


Figure 28.20 Autophosphorylation of the cytosolic domain of the PDGF receptor creates SH2-binding sites for several proteins. Some sites can bind more than one type of SH2 domain. Some SH2-containing proteins can bind to more than one site. The kinase domain consists of two separated regions (shown in blue), and is activated by the phosphorylation site in it.

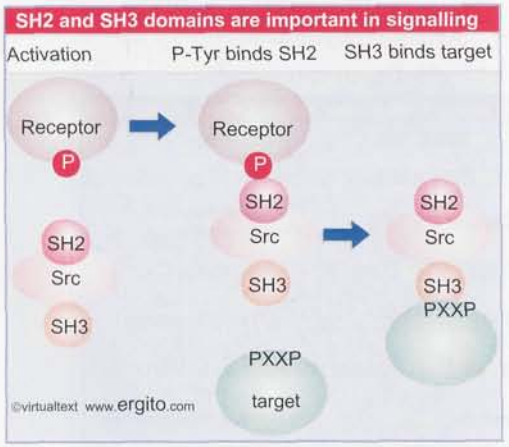


Figure 28.21 SH2 and SH3 domains are used for protein-protein interactions in signal transduction cascades. Typically a phosphorylated receptor recognizes an SH2 target in its substrate, and an SH3 domain in the substrate then recognizes the next protein in the cascade.

followed by its internalization, that is, it is removed from the membrane and transported to the interior of the cell by endocytosis of a vesicle carrying a patch of plasma membrane. The relationship between kinase activity and endocytosis is unclear. Phosphorylation at particular residues may be needed for endocytosis; whether the kinase activity as such is needed may differ for various receptors. It is possible that endocytosis of receptor kinases serves principally to clear receptor (and ligand) from the surface following the response to ligand binding (thus terminating the response). However, in some cases, movement of receptors to coated pits followed by internalization could be necessary for them to act on the target proteins.

Because growth factor receptors generate signals that lead to cell division, their activation in the wrong circumstances is potentially damaging to an organism, and can lead to uncontrolled growth of cells. Many of the growth factor receptor genes are represented in the oncogenes, a class of mutant genes active in cancers. The mutant genes are derived by changes in cellular genes; often the mutant protein is truncated in either or both of its N-terminal or C-terminal regions. The mutant protein usually displays two properties: the tyrosine kinase has been activated; and there is no longer any response to the usual ligand. As a result, the tyrosine kinase activity of the receptor is either increased or directed against new targets (see *30.15 Growth factor receptor kinases can be mutated to oncogenes).*

28.11 Signaling pathways often involve protein-protein interactions

Key Concepts

- An SH2-binding site has a phospho-Tyrosine residue that is recognized by an SH2 domain.
- A receptor may have several SH2-binding sites, which are recognized by the SH2 domains of different signaling proteins.
- The signaling protein may have an SH3 domain that recognizes the next protein in the pathway.

A common means for propagating a signal transduction pathway is for a protein specifically to recognize the next protein in the pathway by means of a physical interaction. (This contrasts with the generation of a small molecule [second messenger] that interacts with the next protein in the pathway.) The usual mechanism for a protein-protein interaction in a signal pathway is for a domain in one protein to recognize a rather short motif in a second protein. The salient feature of the target motif may be its sequence or its structure. Phospho-Tyr residues are often components of such motifs, allowing the motif to be made active by phosphorylation or made inactive by dephosphorylation. Another common feature of target motifs is the amino acid proline, which causes a characteristic turn in a polypeptide chain.

Two motifs found in a variety of cytoplasmic proteins that are involved in signal transduction are used to connect proteins to the components that are upstream and downstream of them in a signaling pathway. The domains are named SH2 and SH3, for *Sxc homology,* because they were originally described in the c-Src cytosolic tyrosine kinase.

The presence of SH2 and SH3 domains in various proteins is summarized in **Figure 28.19.** The cytoplasmic tyrosine kinases comprise one group of proteins that have these domains; other prominent members are phospholipase Cγ and the regulatory subunit (p85) of PI3

kinase (both targets for activation by receptor tyrosine kinases; see Figure 28.15). The extreme example of a protein with these domains is Grb2/sem5, which consists *solely* of an SH2 domain flanked by two SH3 domains (see later).

Some proteins contain multiple SH2 domains, which increases their affinity for binding to phosphoproteins or confers the ability to bind to different phosphoproteins. A receptor may contain different SH2-binding sites, enabling it to activate a variety of target proteins. **Figure 28.20** summarizes the organization of the cytoplasmic domain of the PDGF receptor, which has ~10 distinct SH2-binding sites, each created by a different phosphorylation event. Different pathways may be triggered by the proteins that bind to the various phosphorylated residues.

A protein that contains an SH2 domain is activated when it binds to an SH2-binding site. The activation may involve the SH2-containing protein directly (when it itself has an enzymatic activity) or may be indirect. The enzymatic activities that are regulated directly are most commonly kinases, phosphatases, or phospholipases. An example of a protein containing an SH2 domain that does not have a catalytic activity is provided by p85, the regulatory subunit of PI3 kinase; when p85 binds to a receptor, it is the associated PI3K catalytic subunit that is activated.

Figure 28.21 shows that the SH3 domain provides the effector function by which some SH2-containing proteins bind to a downstream component. The case of the "adaptor" Grb2 strengthens this idea; consisting only of SH2 and SH3 domains, it uses the SH2 domain to contact the component upstream in the pathway, and the SH3 domain to contact the component downstream. SH3 binds the motif PXXP in a sequence-specific manner (see *28.13 Prolines are important determinants in recognition sites).* When an activated receptor binds Grb2, the SH3 domain of Grb2 binds to a target protein that contains the PXXP motif.

SH3 domains often provide connections to small GTP-binding proteins (of which Ras is the paradigm). Another role that has been proposed for SH3 domains (and in particular for the SH3 domain of c-Src) is the ability to interact with proteins of the cytoskeleton, thus triggering changes in cell structure.

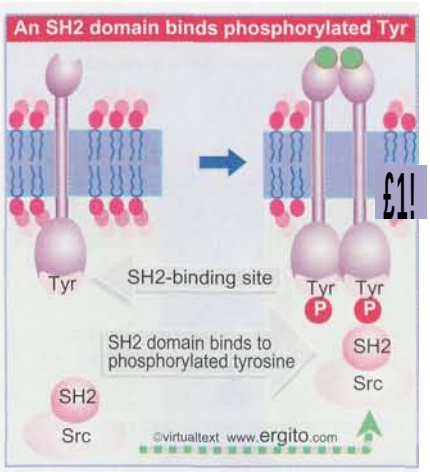


Figure 28.22 Phosphorylation of tyrosine in an SH2-binding domain creates a binding site for a protein that has an SH2 domain.

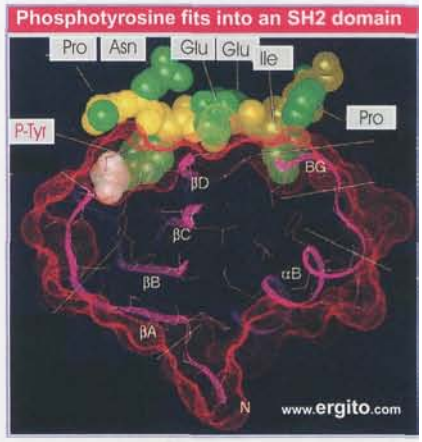


Figure 28.23 The crystal structure of an SH2 domain (purple strands) bound to a peptide containing phosphotyrosine shows that the P-Tyr (white) fits into the SH2 domain, and the 4 C-terminal amino acids in the peptide (backbone yellow, side chains green) also make contact. Photograph kindly provided by John Kuriyan.

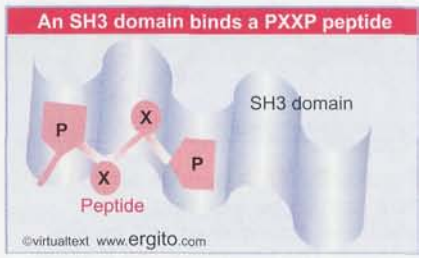


Figure 28.24 A PXXP sequence forms a helical structure in which the two flat prolines form a base and the intervening residues stick up out of the plane. The prolines bind to shallow grooves on the surface of the SH3 domain.

28.12 Phosphotyrosine is the critical feature in binding to an SH2 domain

Key Concepts

- An SH2-binding site consists of phospho-Tyrosine and <5 amino acids on its C-terminal side.
- An SH2 domain forms a globular structure with a pocket that binds the phospho-Tyrosine of the SH2-binding site of the target protein.

The SH2 domain is a region of ~ 100 amino acids that interacts with a target site in other proteins. The target site is called an SH2-binding site. **Figure 28.22** shows an example of a reaction in which SH2 domains are involved. Activation of a tyrosine kinase receptor causes autophosphorylation of a site in the cytosolic tail. Phosphorylation converts the site into an SH2-binding site. So a protein with a corresponding SH2 domain binds to the receptor only when the receptor is phosphorylated.

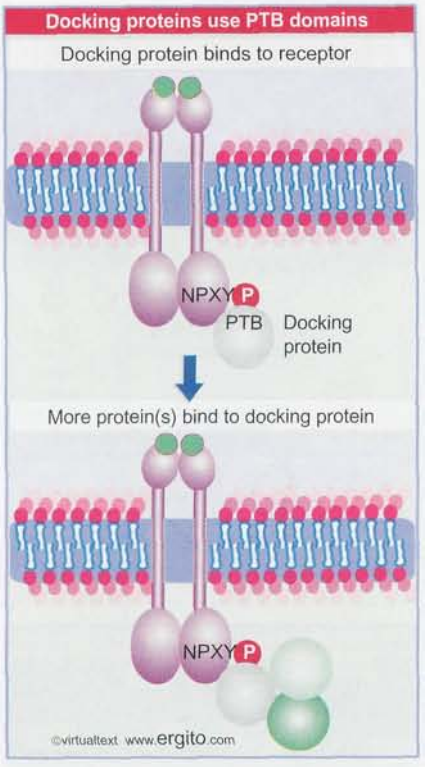


Figure 28.25 A docking protein is an adaptor that connects an activated receptor to a signaling protein(s). It may have a PTB domain that recognizes the motif NPXpY in the receptor.

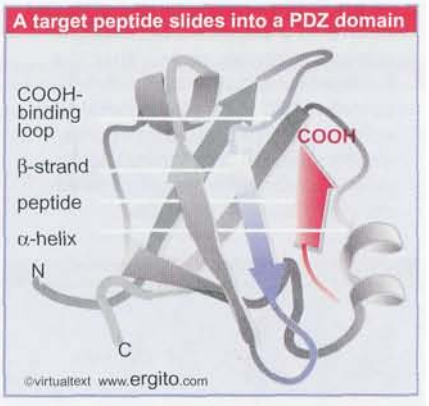


Figure 28.26 A peptide binds to a PDZ domain by inserting as an additional strand in an anti-parallel β-sheet. The strands of the β-sheet are shown as ribbons, with directionality indicated by the arrowheads. The peptide binds between one of the β-strands and an α-helix in the PDZ domain. Its C-terminal end makes contacts with the COOH-binding loop.

An SH2 domain specifically binds to a particular SH2-binding site. The specificity of each SH2 domain is different (except for a group of kinases related to Src, which seem to share the same specificity). The typical SH2-binding site is only 3-5 amino acids long, consisting of a phosphotyrosine and the amino acids on its C-terminal side. SH2 binding is a high-affinity interaction, as much as $10^3\times$ tighter than a typical kinase-substrate binding reaction.

The SH2 domain has a globular structure in which its N-terminal and C-terminal ends are close together, so that its structure is relatively independent of the rest of the protein. The phosphotyrosine in the SH2-binding site binds to a pocket in the SH2 domain, as illustrated in **Figure 28.23.**

28.13 Prolines are important determinants in recognition sites

Key Concepts

- An SH3 domain binds to the structure created by a PXXP amino acid sequence.
- Docking proteins often have a PTB domain that binds to the motif NPXpY in the receptor.
- A β-sheet in a PDZ domain binds a C-terminal β-strand in a target sequence.
- A WW domain recognizes a proline-rich target sequence.

Figure 28.24 shows the interaction between an SH3 domain and a PXXP target sequence. The surface of the SH3 domain is hydrophobic and has three shallow grooves where the target sequence binds. The turns in the polypeptide chain at the proline residues create a helix (called the PPII helix), which seen in cross section resembles a triangle with the two prolines on the base, and the other residues sticking up. The prolines fit into two of the grooves on the SH3 domain surface. The PPII helix is quite similar when viewed from either the N-terminus or the C-terminus, and the ligands for SH3 domains are classified as class I or class II depending on which orientation is bound. Binding is influenced by hydrophobic and other residues in and adjacent to the core PXXP sequence.

A domain that is often found in the targets of receptor kinases is the PTB (phosphotyrosine-binding motif). The PTB binds to the receptor at a motif that consists of a phosphotyrosine preceded by residues that form a β-turn, usually with the consensus Asn-Pro-X-phosphoY. It functions somewhat differently from SH2 or SH3 in being used principally to bind "docking proteins." A docking protein is an intermediary that recruits other proteins to the activated receptor. **Figure 28.25** shows that the docking protein uses its PTB to bind to a phosphorylated site on the receptor, and then other components of the signaling pathway in turn bind to other sites on it. The specificity with which the motif is recognized is determined by hydrophilic amino acids located a few amino acids on the N-terminal side of the phosphorylated Tyr in the receptor. The recognition reaction depends on peptide-peptide interactions, when the phosphopeptide forms an antiparallel β-strand juxtaposed to a β-sheet in the PTB. The fact that it does not depend exclusively on the phosphorylation event is emphasized by the fact that some PTB domains can bind to nonphosphorylated motifs.

The use of antiparallel β-strand interactions in protein recognition is a common theme. **Figure 28.26** shows the example of PDZ domain recognition. A PDZ domain is a 90-100 amino acid region, often represented in multiple repeats in a protein. It is particularly important in the

clustering of membrane proteins and in binding signaling proteins to membrane complexes. (It is named after three of the proteins in which these repeats are found.) It typically binds the last four amino acids at the C-terminus of a target protein. The consensus sequence for recognition is X-Thr/Ser-X-Val-COOH. In the example shown in the figure, the target sequence binds between a β-strand and an α-helix of the PDZ domain, and its terminal -COOH group is bound by the carboxy-binding loop. Basically, the β-sheet of the PDZ domain is extended by adding the target peptide as an additional β-strand.

SH3-binding sites and PDZ binding-sites have the opposite response to phosphorylation from SH2 binding-sites. Phosphorylation of a serine in the site prevents its recognition by the SH3 domain or PDZ domain, respectively.

The WW domain is another case in which a β-sheet interacts with a target peptide. The domain is ~38 amino acids, and has a high concentration of hydrophobic, aromatic, and proline residues. It binds target proline-rich target peptides. **Figure 28.27** shows an example in which the WW domain forms a β-sheet that interacts with the target consensus sequence PPXY (Pro-Pro-X-Tyr).

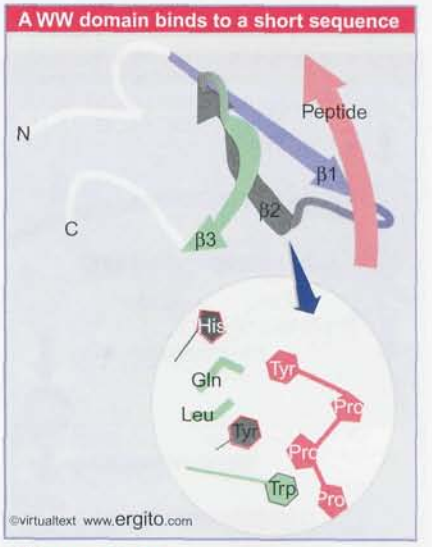


Figure 28.27 A WW domain has a β-sheet consisting of three β-strands that interacts with a target peptide. The insertion shows how amino acids from two of the β-strands specifically interact with the target peptide. The characteristic feature is the insertion of the Trp from the WW domain between the two Pro residues in the target.

28.14 The Ras/MAPK pathway is widely conserved

- **Key Concepts**

The Ras/MAPK pathway starts with activation of the monomeric G protein Ras and then continues by a cascade in which a series of kinases activate one another.

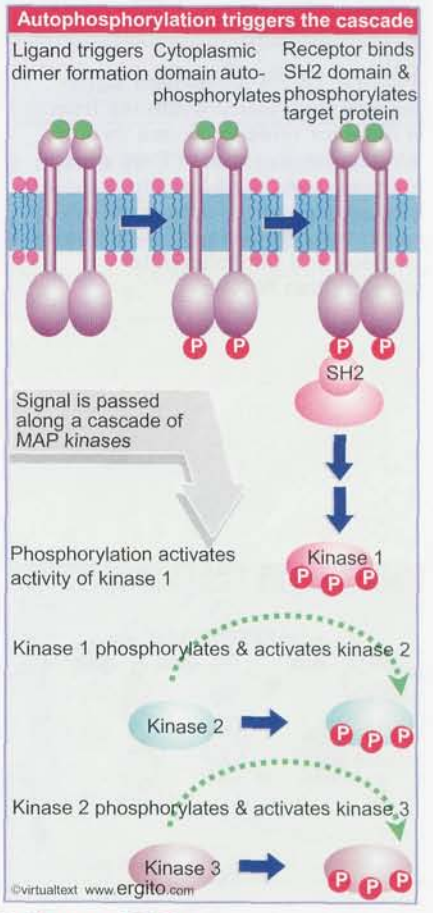


Figure 28.28 Autophosphorylation triggers the kinase activity of the cytoplasmic domain of a receptor. The target protein may be recognized by an SH2 domain. The signal may subsequently be passed along a cascade of kinases.

The best characterized pathway that is initiated by receptor tyrosine kinases passes through the activation of a monomeric G protein to activate a cascade of cytosolic kinases. Although there are still some gaps in the pathway to fill in, and branches that have not yet been identified, the broad outline is clear, as illustrated in **Figure 28.28.**

In mammalian cells, the cascade is often initiated by activation of a tyrosine kinase receptor, such as the EGF or PDGF receptors. The receptor activates the Ras pathway by means of an "adaptor" protein. The activation of Ras leads to the activation of the Raf Ser/Thr kinase, which in turn activates the kinase MEK (formerly known as MAP kinase kinase); its name reflects the fact that it is the kinase that phosphorylates, and thereby activates, a MAP kinase.

The name of the family of MAP kinases reflects their identification as *mitogen-activated protein kinases.* One MAPK family has also been called ERKs, for extracellular signal-regulated. Some major effects of Ras are conveyed via this pathway, but there is also a branch at Ras, which involves the activation of other monomeric G proteins.

The cascade from MEK to the end products is sometimes known as the MAP kinase pathway. Each kinase in this part of the cascade phosphorylates its target kinase, and the phosphorylation event activates the kinase activity of the target enzyme, as illustrated in **Figure 28.29.** The cascade of phosphorylation events leads ultimately to the phosphorylation of transcription factors that trigger changes in cell phenotype varying from growth to differentiation, depending on the cell type. Other targets for the kinases include cytoskeletal proteins that may directly influence cell structure.

The relationship between components of the pathway can be tested by investigating the effects of one component upon the action of another. For

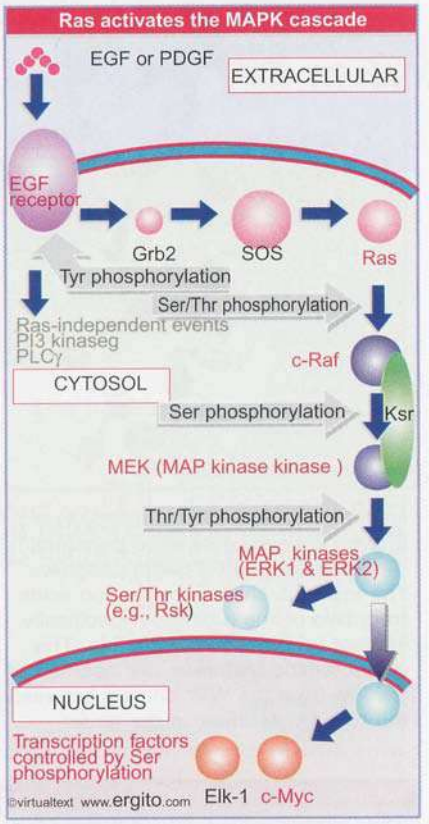


Figure 28.29 A common signal transduction cascade passes from a receptor tyrosine kinase through an adaptor to activate Ras, which triggers a series of Ser/Thr phosphorylation events. Finally, activated MAP kinases enter the nucleus and phosphorylate transcription factors.

example, a mutation that inactivates one component should make it impossible for the pathway to be activated by any components that act earlier. Using such tests allows components to be ordered in a pathway, and to determine whether one component is upstream or downstream of another.

The pathway has been characterized in several situations (see Figure 28.38): in terms of biochemical components responsible for growth of mammalian cultured cells, as the pathway involved in eye development in the fly *D. melanogaster*, as the pathway of vulval development in the worm *C. elegans*, and as the response to mating in the yeast *S. cerevisiae*.

The striking feature is that the pathway is activated by different means in each case (appropriate to the individual system), and it has different end effects in each system, but many of the intermediate components can be recognized as playing analogous roles. It is much as though Nature has developed a signal transduction cascade that can be employed wholesale by means of connecting the beginning to an appropriate stimulus and the end to an appropriate effector. The total pathway is sometimes known as the Ras pathway (named after one of the earlier components) or the Ras/MAPK pathway. Several of the components of this pathway in mammals are related to oncogenes, which suggests that the aberrant activation of this pathway at any one of various stages has a powerful potential to cause tumors.

Figure 28.30 shows how the events initiating the cascade occur at the plasma membrane. The activated receptor tyrosine kinase associates with the SH2 domain of the adaptor protein Grb2, which binds to the receptor but is not phosphorylated. The SH3 domain of Grb2 then binds to the protein SOS, which then activates Ras. The sole role of Grb2 in activating SOS appears to be fulfilled by binding to it. The binding reaction brings SOS to the membrane, and thus into the vicinity of Ras. SOS causes the GDP on Ras to be replaced by GTP, which is sufficient to activate Ras. (Grb2 is not the only adaptor that can activate Ras; an alternative pathway is provided by the adaptor SHC. Which adaptor is used depends on the cell type.)

When a tyrosine kinase receptor is activated, its intracellular domain may be phosphorylated at more than one site, and each site may trigger a different pathway (see Figure 28.20). The most common consequence of a phosphorylation is to activate a signal transduction pathway, but in some cases it may have a negative effect, providing a feedback loop to limit the action of the pathway. These effects may be direct or indirect. **Figure 28.31** illustrates an example of a system in which two phosphorylations counteract each other. Torso is a receptor tyrosine kinase that activates the Ras pathway during *Drosophila* embryogenesis. Two sites become phosphorylated when it is activated. Phosphorylation of Y630 is required to activate the downstream pathway. Phosphorylation of Y918 provides negative regulation.

The receptor binds the regulator RasGAP to the phosphorylated site Y918. This keeps RasGAP in an activated state in which it prevents Ras from functioning (see next section). Y918 is phosphorylated constitutively (or when Torso is activated at a low level). Under these circumstances, the pathway is turned off.

High activation of Torso results in phosphorylation ofY630. This creates a binding site for the cytosolic phosphatase corkscrew (CSW). Corkscrew then dephosphorylates Y918. The result is to release RasGAP, which becomes ineffective, allowing Ras to function. So corkscrew is required for Torso to activate Ras.

Corkscrew may have a second role in the pathway, which is to recruit the adaptor that in turn binds to SOS, which activates Ras.

We see from this example that phosphorylated sites may influence the signaling pathway positively or negatively. The state of one phos-

phorylated site may in fact control the state of another site. An activating site may act indirectly (to inactivate an inhibitory site) as well as directly to recruit components of the signaling pathway.

28.15 The activation of Ras is controlled by GTP

Key Concepts

- Ras is a monomeric G protein that is active when bound to GTP and inactive when bound to GDP.
- When GTP is hydrolyzed, the conformation of Ras changes.
- Constitutively active forms of Ras have mutations that affect GTP binding.
- SOS is the Ras-GEF that activates Ras by causing GDP to be replaced with GTP. SOS is activated by the adaptor Grb2, which is activated by a receptor.
- Ras-GAP is the protein that triggers the GTPase activity and deactivates Ras.

We turn now to the events involved in activating Ras. Ras is an example of a monomeric G protein. Other examples are found in protein trafficking (such as the Rabs) or in protein synthesis (such as EF-Tu). The activity of the G protein depends on whether it is bound to GTP (active state) or bound to GDP (inactive state). Like trimeric G proteins, a monomeric G protein possesses an intrinsic GTPase activity that converts it from the active state to the inactive state.

Figure 28.32 shows that two proteins control the conversion between the active and inactive states of Ras. Ras-GAP is the GAP (GTPase activating protein) that triggers the GTPase activity and thereby inactivates Ras in mammalian cultured cells. SOS is the Ras-GEF (guanine nucleotide exchange factor) that causes GDP to be replaced by GTP, and thereby activates Ras. (SOS is activated when phosphorylation of a receptor tyrosine kinase causes Grb2 to recruit it to the plasma membrane; see Figure 28.30.)

The general structure of mammalian Ras proteins is illustrated in **Figure 28.33.** Three groups of regions are responsible for the characteristic activities of Ras:

- The regions between residues 5-22 and 109-120 are implicated in guanine nucleotide binding by their homology with other G-binding proteins.
- Ras is attached to the cytoplasmic face of the membrane by farnesylation close to the C-terminus. Mutations that prevent the modification abolish oncogenicity, showing that membrane location is important for Ras function. After the farnesylation, the three C-terminal amino acids are cleaved from the protein, and the carboxyl group of the (now C-terminal) Cys^{186} is methylated; also, other Cys residues in the vicinity are reversibly palmitoylated. These changes further increase affinity for the membrane.
- The effector domain (residues 30-40) is the region that reacts with the target molecule when Ras has been activated. This region is required for the oncogenic activity of Ras proteins that have been activated by mutation at position 12. The same region is required for the interaction with Ras-GAP.

The crystal structure of Ras protein is illustrated schematically in **Figure 28.34.** The regions close to the guanine nucleotide include the domains that are conserved in other GTP-binding proteins. The potential effector loop is located near the phosphates; it consists of hydrophilic residues, and is potentially exposed in the cytoplasm.

When GTP is hydrolyzed, there is a switch in the conformation of Ras protein. The change involves L4, which includes position 61, at

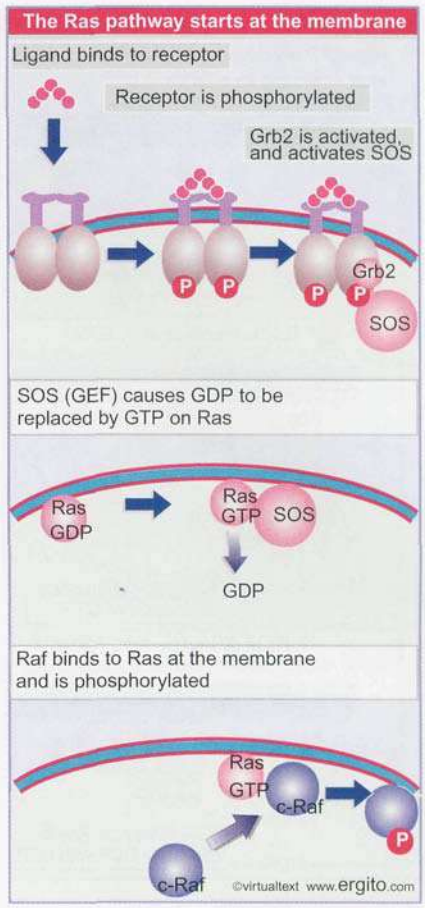


Figure 28.30 The Ras cascade is initiated by a series of activation events that occur on the cytoplasmic face of the plasma membrane.

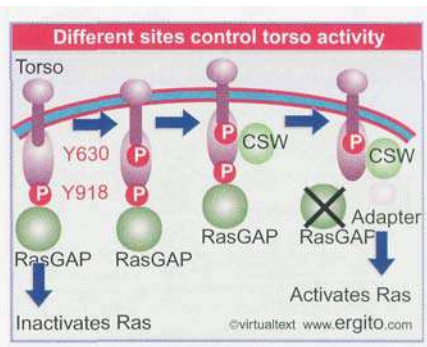


Figure 28.31 Phosphorylation at different sites on a receptor tyrosine kinase may either activate or inactivate the signal transduction pathway.

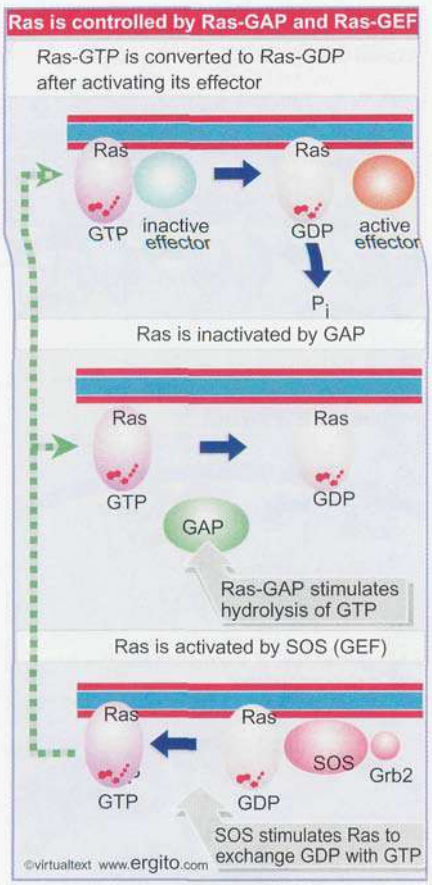


Figure 28.32 The relative amounts of Ras-GTP and Ras-GDP are controlled by two proteins. Ras-GAP inactivates Ras by stimulating hydrolysis of GTP. SOS (GEF) activates Ras by stimulating replacement of GDP by GTP, and is responsible for recycling of Ras after it has been inactivated.

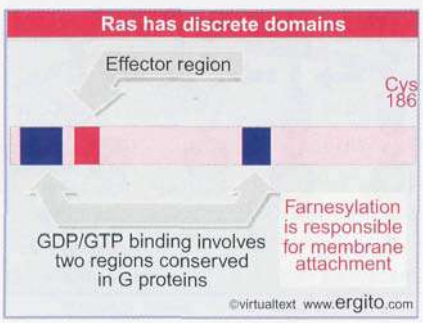


Figure 28.33 Discrete domains of Ras proteins are responsible for guanine nucleotide binding, effector function, and membrane attachment.

which some oncogenic mutations occur. Mutations that activate Ras constitutively (these are oncogenic as discussed in *30.10 Ras proto-oncogenes can be activated by mutation at specific positions)* occur at position 12 in loop 1, and directly affect binding to GTP. The changes between the wild-type and oncogenic forms are restricted to these regions, and impede the ability of the mutant Ras to make the conformational switch when GTP is hydrolyzed. The primary basis for the oncogenic property, therefore, lies in the reduced ability to hydrolyze GTP.

When mitogenesis is triggered by activation of a growth factor, or when a cell is transformed into the tumorigenic state (see 30 *Oncogenes and cancer),* there is a series of coordinated events, including changes in transcription and changes in cell structure. Activation of the Ras/MAPK pathway activates transcription factors that are responsible for one important set of changes. Other changes are triggered by the activation of a group of monomeric G proteins (Rac, Rho, and Cdc42). Each member of this group is responsible for particular types of structural change, as summarized in **Figure 28.35.**

The complete set of relationships that activates these factors is not known, but these structural changes generally can occur independently of the activation of Ras, suggesting that there are other pathways from growth factor receptors to the other monomeric G proteins.

Activation of Rho triggers the formation of actin stress fibers and their connection to the plasma membrane at sites called focal adhesions. Rho can be activated in response to addition of the lipid LPA (a component of serum), through activation of growth factors.

Rac can be activated by the activation of PI3 kinase (a kinase that phosphorylates a small lipid messenger) in response to (for example) activation of PDGF receptor. It stimulates membrane ruffling, formation of lamellipodia (transient structures that are driven by actin polymerization/depolymerization at the leading edge of the membrane), and progression into the G1 phase of the cell cycle. By an independent pathway it activates the stress kinases JNK and p38 (see *28.17 What determines specificity in signaling?).* (The two pathways can be distinguished by mutations in Rac that fail to activate one but not the other.)

Cdc42 activates the formation of filopodia (transient protrusions from the membrane that depend on actin polymerization), although we do not yet know in detail the pathway by which it is itself activated.

There is some crosstalk between these pathways, both laterally and vertically, as shown by the (grey) arrows in Figure 28.35. Rac can be activated by Ras. And Cdc42 can activate Rac, which in turn can activate Rho. This may help the coordination of the events they control; for example, lamellipodia often form along the membrane between two filopodia (making a web-like structure). All of these events are necessary for the full response to mitogenic stimulation, implying that the activation of multiple monomeric G proteins is required.

28.16 A MAP kinase pathway is a cascade

Key Concepts

- There are at least three families of MAP kinases.
- Ras activates the Ser/Thr kinase Raf, which activates the kinase MEK, which activates the ERK MAP kinase.
- An alternative pathway is for a G protein to activate MEKK, which activates MEK.
- Similar Ras/MAP kinase pathways are found in all eukaryotes.
- The end result of activating a MAP kinase pathway is to activate transcription by phosphorylating nuclear proteins.

One of the important features of signal transduction pathways is that they both diverge and converge, thus allowing different but overlapping responses to be triggered in different circumstances. Divergence may start with the initiating event. Activation of a receptor tyrosine kinase may itself trigger multiple pathways: for example, activation of EGF receptor activates the Ras pathway and also Ras-independent pathways involving second messengers (see Figure 28.29). There may also be "branches" later in a pathway.

Convergence of pathways is illustrated by the ability of different types of initiating signal to lead to the activation of MAP kinases. The original paradigm and best characterized example of a pathway leading to MAP kinase involves the activation of Ras, as summarized in Figure 28.28. Returning to the early events in this pathway, the next component after Ras is the Ser/Thr (cytosolic) kinase, Raf. The relationship between Ras and Raf has been puzzling. We know that Ras and Raf are on the same pathway, because both of them are required for the phosphorylation of the proteins later in the pathway (such as MAP kinase). Ras must be upstream of Raf because it is required for the activation of Raf in response to extracellular ligands. Similarly, Raf must be downstream of Ras because the pathway triggered by Ras can be suppressed by expression of a dominant-negative (kinase deficient) mutant of Raf. Ras is localized on the cytoplasmic side of the plasma membrane, and its activation results in binding of Raf, which as a result is itself brought to the vicinity of the plasma membrane. However, the events that then activate Raf, and in particular the kinase that phosphorylates it, are not yet known. The present model is that Ras activates Raf indirectly, perhaps because some kinase associated with the membrane is constitutively active (see Figure 28.30). The importance of localization of enzymatic activities is emphasized by the abilities of components both upstream and downstream of Ras (that is, SOS and Raf) to exercise their activating functions as a consequence of being brought from the cytosol to the plasma membrane.

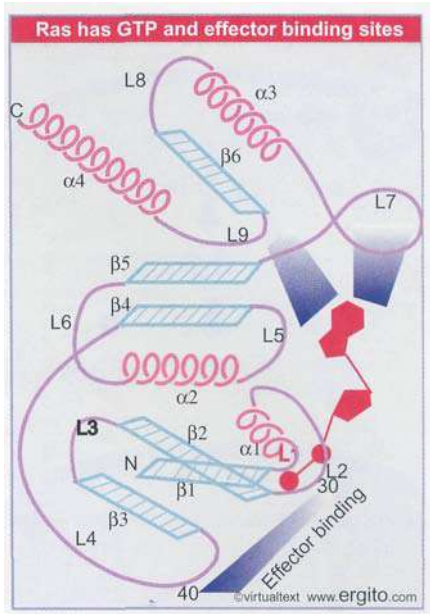


Figure 28.34 The crystal structure of Ras protein has 6 β strands, 4 α helices, and 9 connecting loops. The GTP is bound by a pocket generated by loops L9, L7, L2, and L1.

Raf activity leads to the activation of MEK. Raf directly phosphorylates MEK, which is activated by phosphorylation on two serine residues. MEK is an unusual enzyme with dual specificity, which can phosphorylate both threonine and tyrosine. Its target is the ERK MAP kinase.

Both types of phosphorylation are necessary to convert a MAP kinase into the active state. There are at least 3 MAP kinase families, and they provide important switching points in their pathways. They are activated in response to a wide variety of stimuli, including stimulation of cell growth, differentiation, etc., and appear to play central roles in controlling changes in cell phenotype. The MAP kinases are serine/threonine kinases. After this point in the pathway, all the activating events take the form of serine/threonine phosphorylations.

The ultimate effect of the MAP kinase pathway is a change in the pattern of transcription. So the initiating event occurs at the cell surface, but the final readout occurs in the nucleus, where transcription factors are activated (or inactivated). This type of response requires a nuclear localization step. General possibilities for this step are illustrated in Figure 28.36. In the classic MAP kinase pathway, it is accomplished by the movement of a MAP kinase itself to the nucleus, where it phosphorylates target transcription factors. An alternative pathway is to phosphorylate a cytoplasmic factor; this may be a transcription factor that then moves to the nucleus or a protein that regulates a transcription factor (for example, by releasing it to go to the nucleus).

The MAP kinases have several targets, including other kinases, such as Rsk, which extend the cascade along various branches. The ability of some MAP kinases to translocate into the nucleus after activation extends the range of substrates. In the classic pathway, ERK1 and ERK2 are the targets of MEK, and ERK2 translocates into the nucleus after phosphorylation. The direct end of one branch of the cascade is provided by the phosphorylation of transcription factors, including, c-Myc and Elk-1 (which cooperates with

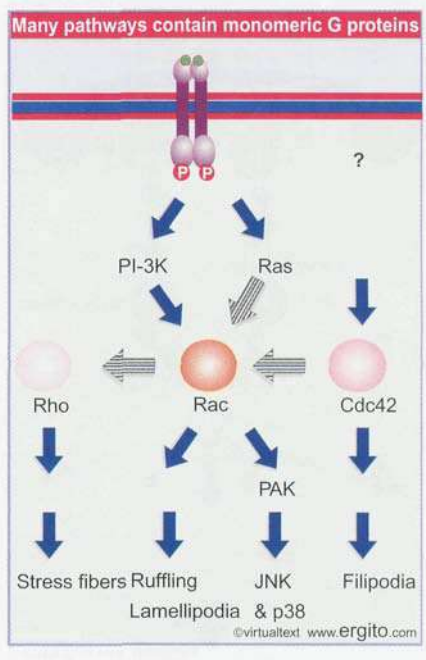


Figure 28.35 Changes in cell structure that occur during growth or transformation are mediated via monomeric G proteins.

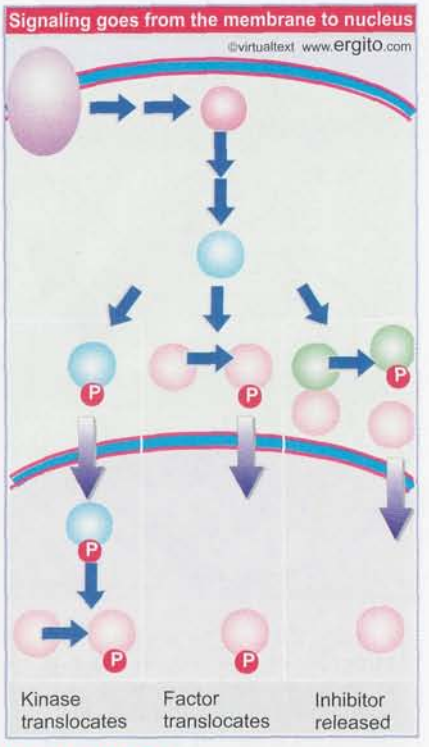


Figure 28.36 A signal transduction cascade passes to the nucleus by translocation of a component of the pathway or of a transcription factor. The factor may translocate directly as a result of phosphorylation or may be released when an inhibitor is phosphorylated.

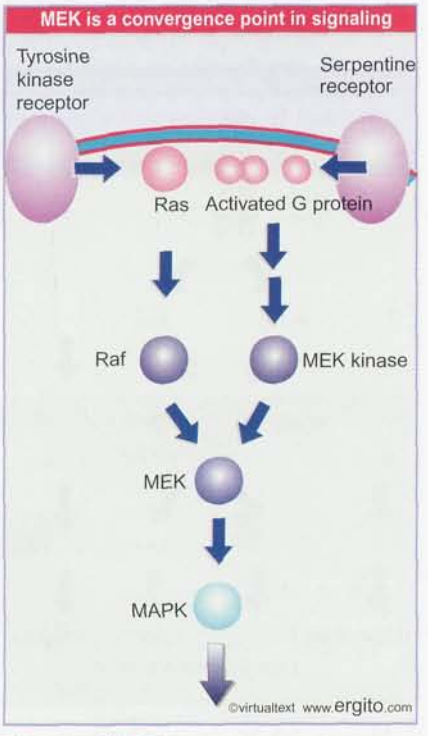


Figure 28.37 Pathways activated by receptor tyrosine kinases and by serpentine receptors converge upon MEK.

SRF [serum response factor]). This enables the cascade to regulate the activity of a wide variety of genes. (The important transcription factor c-Jun is phosphorylated by another MAPK, called JNK; see next section.)

In the MAP kinase pathway, MEK provides a convergence point. Ras activates Raf, which in turn activates MEK. Another kinase that can activate MEK is MEKK (MEK kinase), which is activated by G proteins, as illustrated in **Figure 28.37.** (We have not identified the component(s) that link the activated G protein to the MEKK.) So two principal types of stimulus at the cell surface—activation of receptor tyrosine kinases or of trimeric G proteins—both can activate the MAP kinase cascade. Formally, Raf and MEKK provide analogous functions in parallel pathways.

The counterparts for the components of the pathways in several organisms are summarized in **Figure 28.38.** And although signaling pathways are generally different in plants from animals, the MAPK cascade is triggered by the plant systems for defense against pathogenic infection (see *26.21 Innate immunity utilizes conserved signaling pathways)*.

In mammals, fly, and worm, it starts by the activation of a receptor tyrosine kinase; in mammals and worms the ligand is a polypeptide growth factor, and in *D. melanogaster* retina it is a surface transmembrane protein on an adjacent cell (a "counter-receptor"). The pathway continues through Grb2 in mammals, and through close homologues in the worm and fly. At the next stage, a homologue of SOS functions in the fly in the same way as in mammals. The pathway continues through Ras-like proteins (that is, monomeric guanine nucleotide-binding proteins) in all three higher eukaryotes. Mutations in a homologue of GAP also may influence the pathway in *D. melanogaster,* suggesting that there are alternative regulatory circuits, at least in flies. An interesting feature is that, although the Ras-dependent pathway is utilized in a variety of cells, the mutations in the SOS and GAP functions in *Drosophila* are specific for eye development; this implies that a common pathway may be regulated by components that are tissue-specific. There is a high degree of conservation of function; for example, Grb2 can substitute for Sem-5 in worms.

In yeast, the initiating event consists of the interaction of a polypeptide mating factor with a trimeric G protein, whose βγ dimer (STE2,3) activates the kinase STE20, which activates the MEKK, STE11. We do not know whether there are other components in addition to STE20 between Gβγ and STE11, but the yeast pathway at present provides the best characterization of the route from a G protein to the MAP kinase cascade. The pathway then continues through components all of which have direct counterparts in yeast and mammals. *STE7* is homologous to MEK, and *FUS3* and *KSS1* code for kinases that share with MAP kinase the requirement for activation by phosphorylation on both threonine and tyrosine. Their targets in turn directly execute the consequences of the cascade.

The MAP kinase cascade shown in Figure 28.38 is the best characterized, but there are also other, parallel cascades with related components. In yeast, in addition to the mating response pathway, cascades containing kinases homologous to MEKK, MEK, and MAPK respond to signaling initiated by changes in osmolarity, or activation of PKC (protein kinase C) in *S. cerevisiae*.

28.17 What determines specificity in signaling?

Key Concepts

- A MAP kinase has regions distinct from the active site that are involved in recognizing a substrate.
- Specificity in a MAP kinase pathway may be achieved by a scaffolding protein that binds several kinases that act successively.

The presence of multiple MAPK signaling pathways with analogous components is common. **Figure 28.39** summarizes the mammalian pathways. Each pathway functions in a linear manner, as indicated previously, but in addition there may be "crosstalk" between the pathways, when a component in one pathway can activate the subsequent component in other pathways as well as its own. Usually these "lateral" signals are weaker than those propagating down the pathway. At the very start of the pathway, there is also signaling from Ras to Rac. The strengths of these lateral signals, as well as the extent of activation of an individual pathway, may be important in determining the biological response.

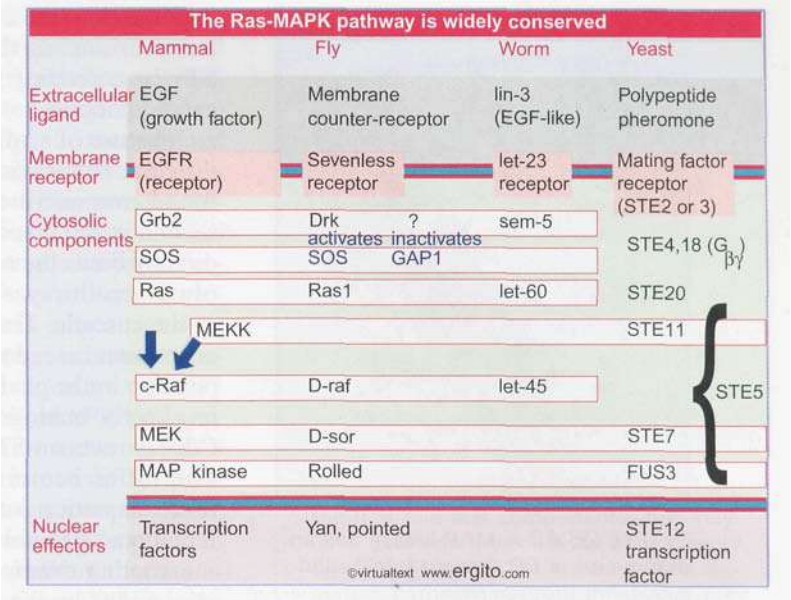


Figure 28.38 Homologous proteins are found in signal transduction cascades in a wide variety of organisms.

The kinases in the MAPK pathways all share a similar mode of action. Each kinase has an active site that binds and phosphorylates a short target sequence containing serine or threonine followed by a proline. The activity of the enzyme is controlled by its state of phosphorylation at a short sequence (Thr-X-Tyr), where it can be activated by a kinase or deactivated by a phosphatase. Given the very short sequences that are involved, these reactions cannot in themselves provide much specificity for target selection. However, each MAP kinase belongs in one (or sometimes more) specific pathways and has a narrow specificity that is appropriate for that pathway. For example, the substrate specificity of each MEK is quite narrow, and it is able to phosphorylate only one or a very few of the many MAP kinases.

Recognition of an appropriate substrate by a MAP kinase depends on two types of interactions: the catalytic site binds the substrate target sequence; and separate sites bind "docking" motifs in the target protein. **Figure 28.40** shows an example (characteristic of the ERK and p38 MAP kinases) in which two regions of the enzyme are involved in substrate recognition. One region is the C-terminal CD region (an abbreviation for common docking). The other is the docking groove, which is nearby but distinct from, the catalytic active site. The regions that are recognized in the target proteins are called the docking sites. The best characterized type of docking site is called the D domain, and is found in many of the target proteins for MAP kinases. It is characterized by a cluster of hydrophobic residues separated from two basic residues.

Another mechanism that is used to prevent a component of one MAP kinase cascade from activating a substrate in a parallel pathway is to localize the components. The first example of such a mechanism was provided by the yeast mating pathway. *STE5* in yeast is implicated in the cascade between *STE2,3* and *FUS3, KSS1*, but cannot be placed in a single position in the pathway. **Figure 28.41** shows that STE5 binds to three of the kinases, STE11 (MEKK), STE7 (MEK), and FUS3 (MAPK), suggesting that this complex has to form before each kinase can activate the next kinase in the pathway. Each of the kinases binds to a different region on STE5, which provides a scaffold. If the kinases can only function in the context of the STE5 scaffold, they may be prevented from acting upon kinases in other pathways. Similarly, Ksr is a scaffolding protein that holds Raf and MEK together, so as to direct Raf to phosphorylate MEK (see Figure 28.29).

Localization is important for holding the group of kinases together on the scaffold, and also for making them available to the upstream factors that activate the pathway. The monomeric G protein Cdc42p is localized on the inner side of the plasma membrane at the junction of the mother cell and the growing bud. It binds the kinase STE20 and

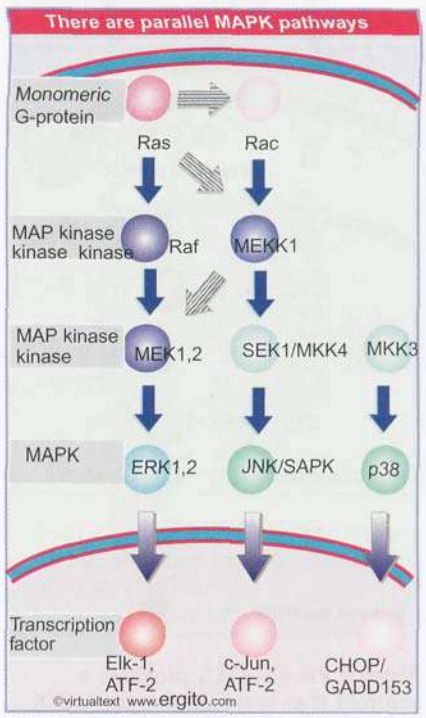


Figure 28.39 Three MAP kinase pathways have analogous components. Crosstalk between the pathways is shown by grey arrows.

By Book_Crazy [IND]

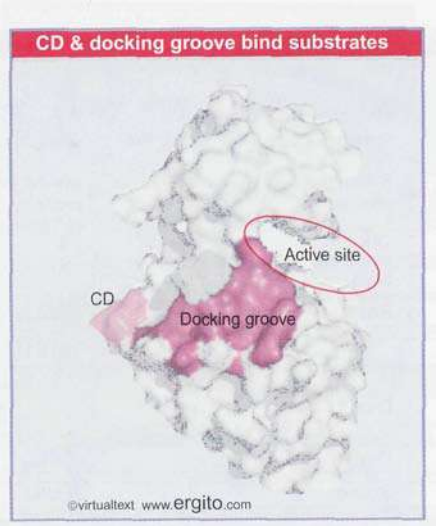


Figure 28.40 A MAP kinase has an active site, a CD domain (pink), and a docking groove (red).

localizes it to the cell cortex. Then the interaction of pheromone with its receptor causes the release of the Gβγ dimer, which also binds to STE20, activating the kinase. When STE20 phosphorylates STE11, the cascade begins.

The use of scaffolds enables the same components to be employed in different pathways. **Figure 28.42** compares two of the pathways. One is the pheromone-induced activation of the kinases bound by STE5. The second is the response to osmotic pressure, when an activated receptor directly binds the scaffold protein. A difference in the construction of the osmotic pathway is that the scaffolding protein is itself one of the kinases in the cascade. The striking feature is that the first kinase in the osmoadaptation cascade is STE11, the same enzyme that is used in the same position in the pheromone pathway. The function of the osmotic pathway is exactly analogous to the pheromone pathway: STE20, bound to Cdc42p, acts on STE11 that is brought to it as part of a kinase complex. The difference is that STE11 is linked to Pbs2p and Hog1p in the osmoadaptation pathway, so that the ultimate kinase in the cascade is different, and therefore a different set of responses is produced.

Another example of a pathway that proceeds through a MAP kinase is provided by the activation of the transcription factor Jun in response to stress signals. **Figure 28.43** shows that activation of the kinase JNK involves both convergence and divergence. JNK is regulated by two classes of extracellular signals: UV light (typical of a stress response); and also as a consequence of activation of Ras (by an unidentified branch of the Ras pathway). JNK is a (distant) relative of MAP kinases such as the ERKs, showing the classic features of being activated by phosphorylation of Thr and Tyr, and phosphorylating its targets on Ser. The proteins JIP1,2 provide a scaffold that may ensure the integrity of the pathway leading to JNK activation.

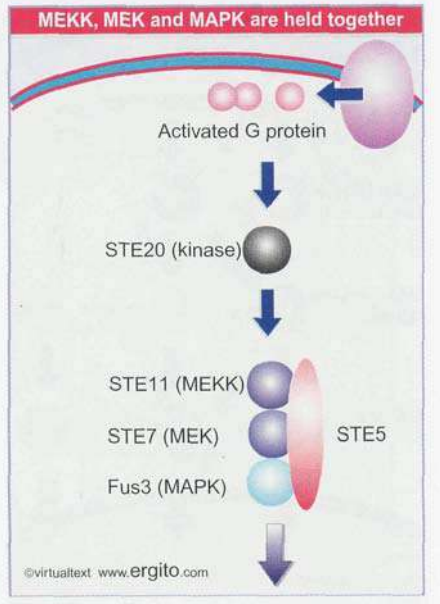


Figure 28.41 STE5 provides a scaffold that is necessary for MEKK, MEK, and MAPK to assemble into an active complex. The complex is activated by STE20, which is localized at the plasma membrane by binding to Cdc42p, and activated by the Gβγ dimer.

28.18 Activation of a pathway can produce different results

Key Concepts

- Transient activation of a MAP kinase may stimulate cell proliferation, whereas continued activation may trigger differentiation.

Signal transduction generates differential responses to stimuli that vary qualitatively (by activating different pathways) or quantitatively (by activating pathways with different intensities or for different durations). An individual stimulus may activate one or more pathways. The strength of activation of any particular pathway may influence the response, since there are cases in which more intense or long-term stimulation of a single pathway gives a different response from less intense or short-term stimulation. One of our major aims is to understand how differences in such stimuli are transduced into the typical cellular responses.

What degree of amplification is achieved through the Ras/MAPK pathway? Typically an ~10× amplification of signal can be achieved at each stage of a kinase cascade, allowing an overall amplification of $>10^4$ through the pathway. However, the combination of the last three kinases into one complex would presumably restrict amplification at these stages. In mammalian cells, the pathway can be fully activated by very weak signals; for example, the ERK1,2 MAP kinases are fully activated when $<5\%$ of the Raf protein molecules bind to Ras.

By Book_Crazy [IND]

A puzzling feature of the Ras/MAPK pathway is that activation of the same pathway under different circumstances can cause different outcomes. When PC 12 cells are treated with the growth factor NGF, they differentiate (by becoming neuronal-like) and stop dividing. When they are treated with EGF, however, they receive a signal for continued proliferation. In both cases, the principal signal transduction event is the activation of the ERK MAP kinase pathway. The differences in outcome might be explained, of course, by other (unidentified) pathways that are activated by the respective receptors. However, the major difference in the two situations is that NGF stimulation causes prolonged elevation of Ras-GTP, whereas EGF stimulation produces only a transient effect. (One reason for this difference is that EGF receptor is more susceptible to feedback mechanisms that reverse its activation.)

The idea that duration of the stimulus to the ERK MAPK pathway may be the critical parameter is supported by results showing that a variety of conditions that cause persistent activation of ERK MAP kinase all cause differentiation. By contrast, all conditions in which activation is transient lead instead to proliferation. More direct proof of the role of the ERK MAPK pathway is provided by showing that mutations constitutively activating MEK cause differentiation of PC 12 cells. So activation of the ERK MAPK pathway is sufficient to trigger the differentiation response. Another point is made by the fact that the same MEK mutation has different effects in a different host cell; in fibroblasts, it stimulates proliferation. This is another example of the ability of a cell to connect the same signal transduction pathway to different readouts.

How might the duration of the signal determine the type of outcome? The concentration of some active component in the pathway could increase with the duration of activation, and at some point would exceed a threshold at which it triggers a new response. One model for such an action is suggested by *Drosophila* development, in which increasing concentrations of a transcription factor activate different target genes, as the result of combinatorial associations with other factors that depend upon relative concentrations (see *31 Gradients, cascades, and signaling pathways).* Another possibility is suggested by the fact that prolonged activation is required before ERK2 translocates to the nucleus. The mechanism is unknown, but could mean that transient stimulation does not support the phosphorylation and activation of nuclear transcription factors, so the expression of new functions (such as those needed for differentiation) could depend upon the stimulus lasting long enough to cause translocation of ERK2.

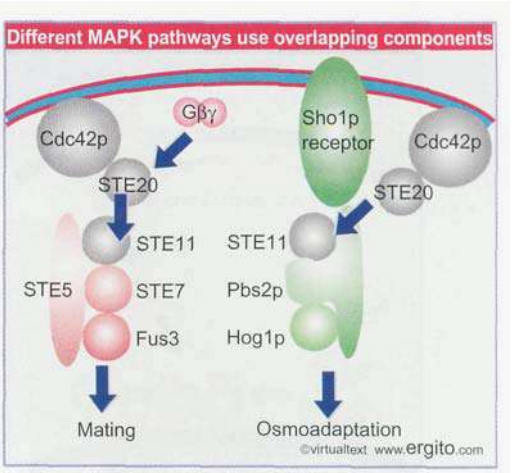


Figure 28.42 MAPK pathways differ in the first and last components, but may have common intermediate components. The components that are unique to the mating pathway are red, those that are unique to osmoadaptation are green, and those that are common to both pathways are gray.

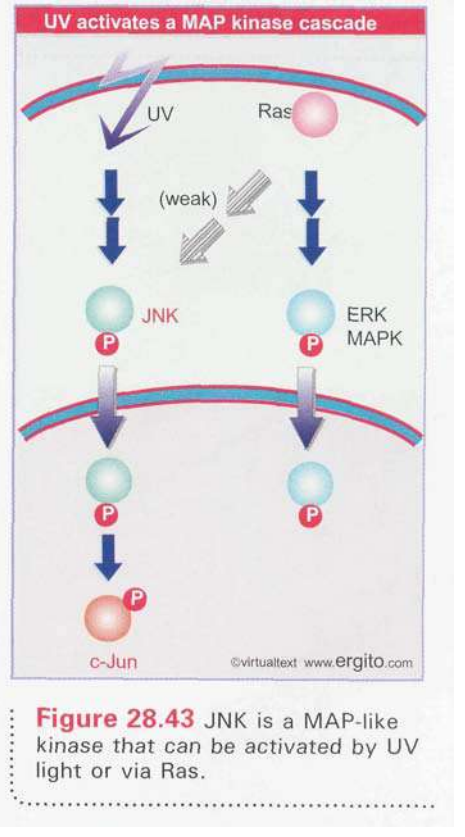


Figure 28.43 JNK is a MAP-like kinase that can be activated by UV light or via Ras.

28.19 Cyclic AMP and activation of CREB

Key Concepts

- Cyclic AMP is produced when a G protein activates adenylate cyclase at the plasma membrane.
- Cyclic AMP binds to the regulatory subunit of PKA (protein kinase A), releasing the catalytic subunit, which moves to the nucleus.
- One of the major nuclear targets for PKA is the transcription factor CREB, which is activated by phosphorylation.

Cyclic AMP is the classic second messenger, and its connection to transcription is by the activation of CREB (cAMP response element binding protein). **Figure 28.44** shows how the pathway proceeds through the Ser/Thr kinase, PKA.

By Book_Crazy [IND]

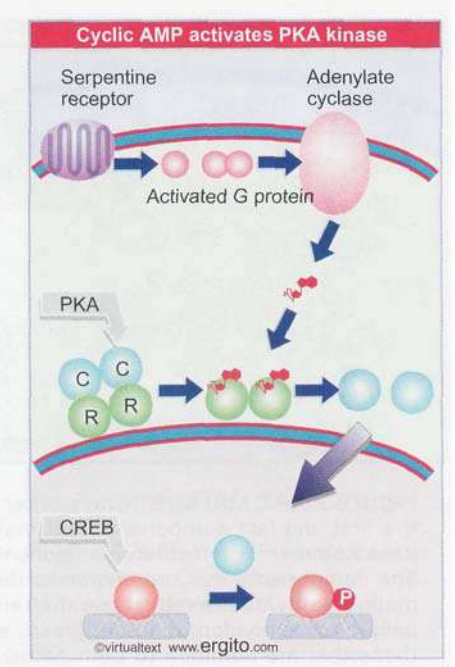


Figure 28.44 When cyclic AMP binds to the R subunit of PKA, the C subunit is released; some C subunits diffuse to the nucleus, where they phosphorylate CREB.

The initial step in the pathway is activation of adenylate cyclase at the plasma membrane by an activated G protein (see Figure 28.11). cAMP binds to the regulatory R subunit of PKA, which is anchored to membranes in the perinuclear region. This causes the R subunit to release the catalytic (C) subunit of PKA, which then becomes free to translocate to the nucleus. Translocation occurs by passive diffusion, and involves only a proportion of the released C subunits. In fact, the free C subunits phosphorylate targets in both the cytosol and nucleus.

The circuitry also has some feedback loops. The end-targets for PKA are also substrates for the phosphatase PPase I, which in effect reverses the action of PKA. However, PKA also has as a target a protein whose phosphorylation converts it into an inhibitor of PPase I, thus preventing the reversal of phosphorylation.

The transcription factor CREB is one of the major nuclear substrates for PKA. Phosphorylation at a single Ser residue greatly increases the activity of CREB bound to the response element CRE, which is found in genes whose transcription is induced by cAMP. The rate of transcription of these genes is directly proportional to the concentration of phosphorylated CREB in the nucleus. The kinetics of the response are limited by the relatively slow rate at which the free C subunit diffuses into the nucleus. Typically the phosphorylated C subunit reaches a maximum level in the nucleus after ~30 min, and then is slowly dephosphorylated (over several hours). Several circuits may be involved in the dephosphorylation, including direct control of phosphatases and indirect control by the entry into the nucleus of the protein PKI, which binds to the C subunit and causes it to be re-exported to the cytoplasm. The kinetics of activating PKA in the nucleus may be important in several situations, including learning, in which a weak stimulus of cAMP has only short-term effects, whereas a strong stimulus is required for long-term effects, including changes in transcription. This parallels the different consequences of short-term and long-term stimulation of the MAPK pathway (see previous section).

28.20 The JAK-STAT pathway

Key Concepts

- Some cytokine (growth factor) receptors activate JAK kinases.
- The JAK kinases phosphorylate STAT transcription factors.
- The activation of JAK and its activation of STAT occurs in a complex at the nuclear membrane.
- The phosphorylated STAT migrates to the nucleus where it activates transcription.

Some signal transduction pathways have large numbers of components (permitting a high degree of amplification) and many feedback circuits (permitting sensitive control of the duration and strength of the signal). The JAK-STAT pathway is much simpler, and consists of three components that function as illustrated in **Figure 28.45.**

JAK-STAT pathways are activated by several cytokine receptors. These receptors do not possess intrinsic kinase activities. However, binding of a cytokine causes its receptor to dimerize, which provides the signal to associate with and activate a JAK kinase. The JAK kinases take their name (originally Janus kinases) from the characteristic presence of two kinase domains in each molecule. Several members of the family are known (JAK1,2,3, etc.); each associates with a specific set of cytokine receptors. The interaction between the activated (dimeric) cytokine receptor and JAK kinase(s) in effect produces the same result

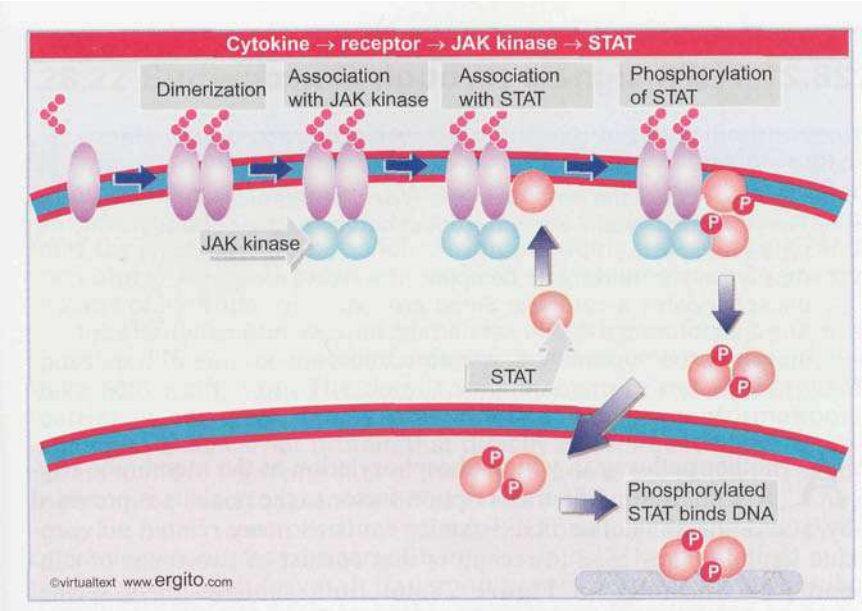


Figure 28.45 Cytokine receptors associate with and activate JAK kinases. STATs bind to the complex and are phosphorylated. They dimerize and translocate to the nucleus. The complex binds to DNA and activates transcription.

as the ligand-induced dimerization of a tyrosine kinase receptor: the difference is that the receptor and kinase activities are provided by different proteins instead of by the same protein.

The JAK kinases are tyrosine kinases whose major substrates are transcription factors called STATs. There are >7 STATs; each STAT is phosphorylated by a particular set of JAK kinases. The phosphorylation occurs while the JAK is associated with the receptor at the plasma membrane. A pair of JAK kinases associates with an activated receptor, and both may be necessary for the pathway to function. An example is that stimulation by the interferon IFNγ requires both JAK1 and JAK2.

STAT phosphorylation leads to the formation of both homodimers and heterodimers. The basis for dimerization is a reciprocal interaction between an SH2 domain in one subunit and a phosphorylated Tyr in the other subunit.

The STAT dimers translocate to the nucleus, and in some cases associate with other proteins. They bind to specific recognition elements in target genes, whose transcription is activated.

Given a multiplicity of related cytokine receptors, JAK kinases, and STAT transcription factors, how is specificity achieved? The question is sharpened by the fact that many receptors can activate the same JAKs, but activate different STATs. Control of specificity lies with formation of a multipartite complex containing the receptor, JAKs, and STATs. The STATs interact directly with the receptor as well as with the JAKs, and an SH2 domain in a particular STAT recognizes a binding site in a particular receptor. So the major control of specificity lies with the STAT.

Stimulation of a JAK-STAT pathway is only transient. Its activation may be terminated by the action of a phosphatase. An example is the pathway activated by binding of erythropoietin (red blood cell hormone) to its receptor. This activates JAK2 kinase. Recruitment of another component terminates the reaction; the phosphatase SH-PTP1 binds via its SH2 domain to a phosphotyrosine site in the erythropoietin receptor. This site in the receptor is probably phosphorylated by JAK2. The phosphatase then dephosphorylates JAK2 and terminates the activation of the corresponding STATs. This creates a simple feedback circuit: erythropoietin receptor activates JAK2, JAK2 acts on a site in the receptor, and this site is recognized by the phosphatase that in turn acts on JAK2. This again emphasizes the way in which formation of a multicomponent complex may be used to ensure specificity in controlling the pathway.

By Book_Crazy [IND]

28.21 TGFβ signals through Smads

Key Concepts

activates the heterodimeric type II receptor.
The activated type II receptor phosphorylates the heterodimeric type I receptor.
As part of the tetrameric complex, the type I receptor phosphorylates a cytosolic Smad protein.
The Smad forms a dimer with a related protein (Smad4) which moves to the nucleus and activates transcription.

Another pathway in which phosphorylation at the membrane triggers migration of a transcription factor to the nucleus is provided by TGFβ signaling. The TGFβ family contains many related polypeptide ligands. They bind to receptors that consist of two types of subunits, as illustrated in **Figure 28.46.** Both subunits have serine/threonine kinase activity. (Actually all serine/threonine receptor kinases are members of the TGFβ receptor family.)

The ligand binds to the type II receptor, creating a receptor-ligand combination that has high affinity for the type I receptor. A tetrameric complex is formed in which the type II receptor phosphorylates the type I receptor. (A variation occurs in a subset of these receptors that bind BMPs—bone morphogenetic proteins—which are members of the TGFβ family. In this case, both type I and type II subunits have low affinity for the ligand, but the combination of subunits has high affinity.)

Once the active complex has formed, the type 1 receptor phosphorylates a member of the cytosolic Smads family. Typically a Smad activator is phosphorylated at the motif SSXS at the C-terminus. This causes it to form a dimer with the common partner Smad4. The heterodimer is imported into the nucleus, where it binds to DNA and activates transcription.

The 9 Smad proteins fall into three functional categories. The pathway-specific activators are Smad2, 3 (which mediate TGFβ/activin signaling) and Smad1, 5 (which activate BMP signaling). Smad4 is a universal partner which can dimerize with all of the pathway-specific Smads. Inhibitory Smads act as competitive inhibitors of the activator Smads, providing another level of complexity to the pathway. Each ligand in the TGFβ superfamily activates a particular receptor that signals through a characteristic combination of Smads proteins. Various other proteins bind to the Smads dimers and influence their capacity to act on transcription.

Signaling systems of this type are important in early embryonic development, where they are part of the pathways that lead to development of specific tissues (typically bone formation and the development of mesoderm). Also, because TGFβ is a powerful growth inhibitor, this pathway is involved in tumor suppression. The TGFβ type II receptor is usually inactivated in hereditary nonpolyposis colorectal cancers, and mutations in Smad4 occur in 50% of human pancreatic cancers.

One striking feature of the JAK-STAT and TGFβ pathways is the simplicity of their organization, compared (for example) with the Ras-MAPK pathway. The specificity of these pathways depends on variation of the components that assemble at the membrane—different combinations of JAK-STATs in the first case, different Smad proteins in the second. Once the pathway has been triggered, it functions in a direct linear manner. The component that is phosphorylated at the plasma membrane (STAT in the JAK-STAT pathway, Smad in the TGFβ pathway) itself provides the unit that translocates to the nucleus to activate transcription—perhaps the ultimate demonstration of the role of localization.

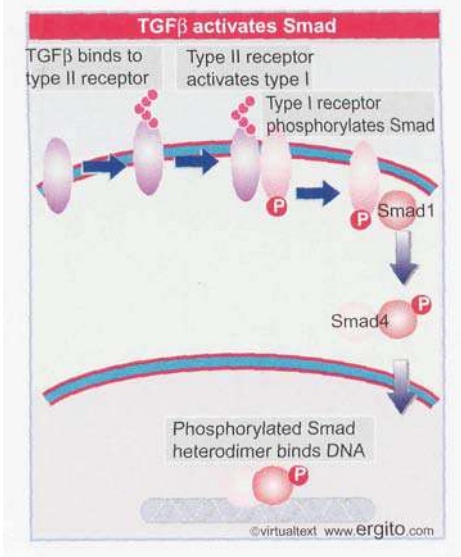


Figure 28.46 Activation of TGFβ receptors causes phosphorylation of a Smad, which is imported into the nucleus to activate transcription.

By Book_Crazy [IND]

28.22 Summary

Lipids may cross the plasma membrane, but specific transport mechanisms are required to promote the passage of hydrophilic molecules. Integral proteins of the plasma membrane offer several means for communication between the extracellular milieu and the cytoplasm. They include ion channels, transporters, and receptors. All such proteins reside in the plasma membrane by means of hydrophobic domains.

Ions may be transported by carrier proteins, which may utilize passive diffusion or may be connected to energy sources to undertake active diffusion. The detailed mechanism of movement via a carrier is not clear, but is presumed to involve conformational changes in the carrier protein that directly or indirectly allow a substrate to move from one side of the membrane to the other. Ion channels can be used for passive diffusion (driven by the gradient). They may be gated (controlled) by voltage, extracellular ligands, or cytoplasmic second messengers. Channels typically have multiple subunits, each with several transmembrane domains; hydrophilic residues within the transmembrane domains face inward so as to create a hydrophilic path through the membrane.

Receptors typically are group I proteins, with a single transmembrane domain, consisting exclusively of uncharged amino acids, connecting the extracellular and cytosolic domains. Many receptors for growth factors are protein tyrosine kinases. Such receptors have a binding site for their ligand in the extracellular domain, and a kinase activity in their cytoplasmic domain. When a ligand binds to the receptor, it causes the extracellular domain to dimerize; most often the product is a homodimer, but there are some cases where heterodimers are formed. The dimerization of the extracellular domains causes the transmembrane domains to diffuse laterally within the membrane, bringing the cytoplasmic domains into contact. This results in an autophosphorylation in which each monomeric subunit phosphorylates the other.

The phosphorylation creates a binding site for the SH2 motif of a target protein. Specificity in the SH2-binding site typically is determined by the phosphotyrosine in conjunction with the 4-5 neighboring amino acids on its C-terminal side. The next active component in the pathway may be activated indirectly or directly. Some target proteins are adaptors that are activated by binding to the phosphorylated receptor, and they in turn activate other proteins. An adaptor typically uses its SH2 domain to bind the receptor and uses an SH3 domain to bind the next component in the pathway. Other target proteins are substrates for phosphorylation, and are activated by the acquisition of the phosphate group.

One group of effectors consists of enzymes that generate second messengers, most typically phospholipases and kinases that generate or phosphorylate small lipids. Another type of pathway consists of the activation of a kinase cascade, in which a series of kinases successively activate one another, leading ultimately to the phosphorylation and activation of transcription factors in the nucleus. The MAP kinase pathway is the paradigm for this type of response.

The connection from receptor tyrosine kinases to the MAP kinase pathway passes through Ras. An adaptor (Grb2 in mammalian cells) is activated by binding to the phosphorylated receptor. Grb2 binds to SOS, and SOS causes GDP to be replaced by GTP on Ras. Ras is anchored to the cytoplasmic face of the membrane. The activated Ras binds the Ser/Thr kinase Raf, thus bringing Raf to the membrane, which causes Raf to be activated, probably because it is phosphorylated by a kinase associated with the membrane. Raf phosphorylates MEK, which is a dual-specificity kinase that phosphorylates ERK MAP kinases on both tyrosine and threonine. ERK MAP kinases activate other kinases; ERK2 MAP kinase also translocates to the nucleus, where it phosphorylates transcription factors

that trigger pathways required for cell growth (in mammalian cells) or differentiation (in fly retina, worm vulva, or yeast mating).

An alternative connection to the MAP kinase cascade exists from serpentine receptors. Activation of a trimeric G protein causes MEKK to be activated. One component in the pathway between Gβγ and MEKK in *S. cerevisiae* is the kinase STE20. The MEKK (STE11), MEK (STE7), and MAPK (Fus3) form a complex with the scaffold protein STE5 that may be necessary for the kinases to function. The use of scaffolding proteins allows the same kinases to participate in different pathways, but to signal to the downstream components only of the pathway that activates them.

The cyclic AMP pathway for activating transcription proceeds by releasing the catalytic subunit of PKA in the cytosol. It diffuses to the nucleus, where it phosphorylates the transcription factor CREB. The activity of this factor is responsible for activating cAMP-inducible genes. The response is down regulated by phosphatases that dephosphorylate CREB and by an inhibitor that exports the C subunit back to the cytosol.

JAK-STAT pathways are activated by cytokine receptors. The activated receptor associates with a JAK kinase and activates it. The target for the kinase is a STAT(s); STATs associate with a receptor-JAK kinase complex, are phosphorylated by the JAK kinase, dimerize, translocate to the nucleus, and form a DNA-binding complex that activates transcription at a set of target genes. In an analogous manner, TGFβ ligands activate type II/type I receptor systems that phosphorylate Smad proteins, which then are imported into the nucleus to activate transcription.

References

28.3 Ion channels are selective

rev Miller, C. (1989). Genetic manipulation of ion channels: a new approach to structure and mechanism. Neuron 2, 1195-1205.

Unwin, N. (1989). The structure of ion channels in membranes of excitable cells. Neuron 3, 665-676.

ref Doyle, D. A. et al. (1998). The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. Science 280, 69-77.

28.5 G proteins may activate or inhibit target proteins

rev Divecha, N. and Irvine, R. F. (1995). Phospholipid signaling. Cell 80, 269-278.

Pierce, K. L., Premont, R. T., and Lefkowitz, R. J. (2002). Seven-transmembrane receptors. Nat. Rev. Mol. Cell Biol. 3, 639-650.

Strader, D. (1994). Structure and function of G protein-coupled receptors. Ann. Rev. Biochem. 63, 101-132.

28.6 G proteins function by dissociation of the trimer

rev Clapham, D. E. and Neer, E. J. (1993). New roles of G protein βγ-dimers in transmembrane signaling. Nature 365, 403-406.

Neer, E. J. (1995). Heterotrimeric G proteins: organizers of transmembrane signals. Cell 80, 249-257.

Neer, E. J. and Clapham, D. E. (1988). Roles of G protein subunits in transmembrane signaling. Nature 333, 129-134.

Sprang, S. R. (1997). G protein mechanisms: insights from structural analysis. Ann. Rev. Biochem. 66, 639-678.

28.7 Protein kinases are important players in signal transduction

rev Hubbard, S. R. and Till, J. H. (2000). Protein tyrosine kinase structure and function. Ann. Rev. Biochem. 69, 373-398.

Hunter, T. (1987). A thousand and one protein kinases. Cell 50, 823-829.

Hunter, T. (1995). Protein kinases and phosphatases: the Yin and Yang of protein phosphorylation and signaling. Cell 80, 237-248.

Hunter, T. and Cooper, J. A. (1985). Protein-tyrosine kinases. Ann. Rev. Biochem. 54, 897-930.

Yarden, Y. and Ullrich, A. (1988). Growth factor receptor tyrosine kinases. Ann. Rev. Biochem. 57, 443-478.

ref Canagarajah, B. J., Khokhlatchev, A., Cobb, M. H., and Goldsmith, E. J. (1997). Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. Cell 90, 859-869.

Hubbard, S. R., Wei, L., Ellis, L., and Hendrickson, W. A. (1994). Crystal structure of the tyrosine kinase domain of the human insulin receptor. Nature 372, 746-754.

Mohammadi, M., Schlessinger, J., and Hubbard, S. R. (1996). Structure of the FGF receptor tyrosine kinase domain reveals a novel autoinhibitory mechanism. Cell 86, 577-587.

Plotnikov, A. N., Schlessinger, J., Hubbard, S. R., and Mohammadi, M. (1999). Structural basis for FGF receptor dimerization and activation. Cell 98, 641-650.

Zhang, F., Strand, A., Robbins, D., Cobb, M. H., and Goldsmith, E. J. (1994). Atomic structure of the MAP kinase ERK2 at 2.3 A resolution. Nature 367, 704-711.

28.9 Receptors are activated by dimerization

rev Heldin, C.-H. (1995). Dimerization of cell surface receptors in signal transduction. Cell 80, 213-223.
Hubbard, S. R. and **Till,** J. H. (2000). Protein tyrosine kinase structure and function. Ann. Rev. Biochem. 69, 373-398.
Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. Cell 103, 211-225.
Ullrich, A. and Schlessinger, J. (1990). Signal transduction by receptors with tyrosine kinase activity. Cell 61, 203-212.
van der Geer, P., Hunter, T., and Lindberg, R. A. (1994). Receptor protein-tyrosine kinases and their signal transduction pathways. Ann. Rev. Cell Biol. 10, 251-337.

ref Cunningham, B. C. et al. (1991). Dimerization of the extracellular domain of the human growth hormone receptor by a single hormone molecule. Science 254, 821-825.
Plotnikov, A. N., Schlessinger, J., Hubbard, S. R., and Mohammadi, M. (1999). Structural basis for FGF receptor dimerization and activation. Cell 98, 641-650.
Wiesmann, C, Fuh, G., Christinger, H. W., Eigenbrot, C, Wells, J. A., and de Vos, A. M. (1997). Crystal structure at 1.7 A resolution of VEGF in complex with domain 2 of the Flt-1 receptor. Cell 91, 695-704.

28.10 Receptor kinases activate signal transduction pathways

rev Pawson, T. and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. Science 278, 2075-2080.

28.11 Signaling pathways often involve protein-protein interactions

rev Cohen, G. B., Ren, R., and Baltimore, D. (1995). Molecular binding domains in signal transduction proteins. Cell 80, 237-248.
Kay, B. K., Williamson, M. P., and Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J. 14, 231-241.
Koch, C. A. (1991). SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. Science 252, 668-674.
Pawson, T. and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. Science 278, 2075-2080.
Yaffe, M. B. (2002). Phosphotyrosine-binding domains in signal transduction. Nat. Rev. Mol. Cell Biol. 3, 177-186.

ref Booker, G. W. et al. (1993). Solution structure and ligand-binding site of the SH3 domain of the p85αsubunit of phosphatidylinositol 3-kinase. Cell 73, 813-822.
Fantl, W. J. et al. (1992). Distinct phosphotyrosines on a growth factor receptor bind to specific molecules that mediate different signaling pathways. Cell 69, 413-423.

28.12 Phosphotyrosine is the critical feature in binding to an SH2 domain

ref Songyang, Z. et al. (1993). SH2 domains recognize specific phosphopeptide sequences. Cell 72, 767-778.

28.13 Prolines are important determinants in recognition sites

rev Harris, B. Z. and Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. J. Cell Sci. 114, 3219-3231.
Mayer, B. J. (2001). SH3 domains: complexity in moderation. J. Cell Sci. 114, 1253-1263.

ref Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85, 1067-1076.
Kavanaugh, W. M., Turck, C. W., and Williams, L. T. (1995). PTB domain binding to signaling proteins through a sequence motif containing phosphotyrosine. Science 268, 1177-1179.
Macias, M. J., Hyvonen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M., and Oschkinat, H. (1996). Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. Nature 382, 646-649.
Zhou M. M., Ravichandran K. S., Olejniczak E. F., Petros A. M., Meadows R. P., Sattler M., Harlan J. E., Wade W. S., Burakoff S. J., Fesik S. W. (1995). Structure and ligand recognition of the phosphotyrosine binding domain of Shc. Nature 378, 584-592.

28.14 The Ras/MAPK pathway is widely conserved

ref Aronheim, A. et al. (1994). Membrane targeting of the nucleotide exchange factor SOS is sufficient for activating the Ras signaling pathway. Cell 78, 949-961.
Buday, L. and Downward, J. (1993). EGF regulates $p21^{ras}$ through the formation of a complex of receptor, Grb2 adaptor protein, and SOS nucleotide exchange factor. Cell 73, 611-620.
Chardin, P. et al. (1 993). Human SOS1 : a guanine nucleotide exchange factor for Ras that binds to Grb2. Science 260, 1338-1343.
Lowenstein, E. J. et al. (1992). The SH2 and SH3-domain containing protein Grb2 links receptor tyrosine kinases to ras signaling. Cell 70, 431-442.

28.15 The activation of Ras is controlled by GTP

rev Boguski, M. S. and McCormick, F. (1993). Proteins regulating Ras and its relatives. Nature 366, 643-654.
Kaibuchi, K., Kuroda, S., and Amano, M. (1999). Regulation of the cytoskeleton and cell adhesion by the Rho family GTPases in mammalian cells. Ann. Rev. Biochem. 68, 459-486.

ref Lamarche, N. et al. (1996). Rac and Cdc42 induce actin polymerization and G cell cycle progression independently of p^{65}PAK and the JNK/SAPK MAP kinase cascade. Cell 87, 519-529.
Nobes, C. D. and Hall, A. (1995). Rho, Rac, and Cdc42 GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipodia, and filopodia. Cell 81, 53-62.
Ridley, A. J. et al. (1992). The small GTP-binding protein rac regulates growth factor-induced membrane ruffling. Cell 70, 401-410.
Ridley, A. J. and Hall, A. (1992). The small GTP-binding protein rho regulates the assembly of focal adhesions and actin stress fibers in response to growth factors. Cell 70, 389-399.
Simon, M. A. et al. (1991). Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the sevenless protein tyrosine kinase. Cell 67, 701-716.

28.16 A MAP kinase pathway is a cascade

rev Herskowitz, I. (1995). MAP kinase pathways in yeast: for mating and more. Cell 80, 187-198.
Hill, C. S. and Treisman, R. (1995). Transcriptional regulation by extracellular signals: mechanisms and specificity. Cell 80, 199-212.

ref Aroian, R. V. et al. (1990). The *let-23* gene necessary for *C. elegans* vulval induction encodes a tyrosine kinase of the EGF receptor subfamily. Nature 348, 693-699.

Hafen, E. et al. (1987). Sevenless, a cell-specific homeotic gene of *Drosophila,* encodes a putative transmembrane receptor with a tyrosine kinase domain. Science 236, 55-63.

Howe, L. R., Leevers, S. J., Gomez, N., Nakielny, S., Cohen, P., and Marshall, C. J. (1992). Activation of the MAP kinase pathway by the protein kinase raf. Cell 71, 335-342.

Lange-Carter, C. A. et al. (1993). A divergence in the MAP kinase regulatory network defined by MEK kinase and Raf. Science 260, 315-319.

Leevers, S. J., Paterson, H. F., and Marshall, C. J. (1994). Requirement for Ras in Raf activation is overcome by targeting Raf to the plasma membrane. Nature 369, 411-414.

Vojtek, A. B., Hollenberg, S. M., and Cooper, J. A. (1993). Mammalian Ras interacts directly with the serine/threonine kinase Raf. Cell 74, 205-214.

Wood, K. W. et al. (1992). Ras mediates nerve growth factor receptor modulation of three signal-transducing protein kinases: MAP kinase, Raf-1, and RSK. Cell 68, 1041-1050.

28.17 What determines specificity in signaling?

rev Elion, E. A. (2001). The Ste5p scaffold. J. Cell Sci. 114, 3967-3978.

Pearson, G., Robinson, F., Beers Gibson, T., Xu, B. E., Karandikar, M., Berman, K., Cobb, M. H. (2001). Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. Endocr. Rev. 22, 153-183.

ref Chang, C. I., Xu, B. E., Akella, R., Cobb, M. H., and Goldsmith, E. J. (2002). Crystal structures of MAP kinase p38 complexed to the docking sites on its nuclear substrate MEF2A and activator MKK3b. Mol. Cell 9, 1241-1249.

Choi, K.-Y. et al. (1994). Ste5 tethers multiple protein kinases in the MAP kinase cascade required for mating in *S. cerevisiae.* Cell 78, 499-512.

Derijard, B. et al. (1994). JNK1: a protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain. Cell 76, 1025-1037.

Roy, F., Laberge, G., Douziech, M., Ferland-McCollough, D., and Therrien, M. (2002). KSR is a scaffold required for activation of the ERK/MAPK module. Genes Dev. 16, 427-438.

Tanoue, T., Maeda, R., Adachi, M., and Nishida, E. (2001). Identification of a docking groove on ERK and p38 MAP kinases that regulates the specificity of docking interactions. EMBO J. 20, 466-479.

Yang, S. H., Yates, P. R., Whitmarsh, A. J., Davis, R. J., and Sharrocks, A. D. (1998). The Elk-1 ETS-domain transcription factor contains a mitogen-activated protein kinase targeting motif. Mol. Cell Biol. 18, 710-720.

Yasuda, J., Whitmarsh, A. J., Cavanagh, J., Sharma, M., and Davis, R. J. (1999). The JIP group of mitogen-activated protein kinase scaffold proteins. Mol. Cell Biol. 19, 7245-7254.

28.18 Activation of a pathway can produce different results

rev Marshall, C. J. (1995). Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. Cell 80, 179-186.

ref Cowley, S. et al. (1994). Activation of MAP kinase kinase is necessary and sufficient for PC 12 differentiation and for transformation of NIH-3T3 cells. Cell 77, 841-862.

28.19 Cyclic AMP and activation of CREB

ref Hagiwara, M. et al. (1992). Transcriptional attenuation following cAMP induction requires PPA-mediated dephosphorylation of CREB. Cell 70, 105-113.

Hagiwara, M. et al. (1993). Coupling of hormonal stimulation and transcription via the cAMP-responsive factor CREB is rate limited by nuclear entry of PKA Mol. Mol. Cell Biol. 13, 4852-4859.

28.20 The JAK-STAT pathway

rev Darnell, J. E., Kerr, I. M., and Stark, G. R. (1994). JAK-STAT pathways and transcriptional activation in response to IFN7 and other extracellular signaling proteins. Science 264, 1415-1421.

Schindler, C. and Darnell, J. E. (1995). Transcriptional responses to polypeptide ligands: the JAK-STAT pathway. Ann. Rev. Biochem. 64, 621-651.

ref Dale, T. C. et al. (1989). Rapid activation by interferon a of a latent DNA-binding protein present in the cytoplasm of untreated cells. Proc. Nat. Acad. Sci. USA 86, 1203-1207.

Klingmuller, U. et al. (1995). Specific recruitment of SH-PTP1 to the erythropoietin receptor causes inactivation of JAK2 and termination of proliferative signals. Cell 80, 729-738.

Shuai, K. et al. (1994). Interferon activation of the transcription factor STAT91 involves dimerization through SH2-phosphotyrosyl peptide interactions. Cell 76, 821-828.

Velazquez, L. et al. (1992). A protein tyrosine kinase in the interferon α/β signaling pathway. Cell 70, 313-322.

28.21 TGFβ signals through Smads

rev Attisano, L. and Wrana, J. L. (2002). Signal transduction by the TGF-beta superfamily. Science 296, 1646-1647.

Massague, J. (1996). TGFβ; signaling: receptors, transducers, and Mad proteins. Cell 85, 947-950.

Massague, J. (1998). TGFβ signal transduction. Ann. Rev. Biochem. 67, 753-791.

ref Macias-Silva, M. et al. (1996). Madr2 is a substrate of the TGFβ receptor and its phosphorylation is required for nuclear accumulation and signaling. Cell 87, 1215-1224.

28.22 Structural subunits can be messengers

ref Diebel, C. E., Proksch, R., Green, C. R., Neilson, P., Walker, M. M., Diebel, C. E., Proksch, R., Green, C. R., Neilson, P., and Walker, M. M. (2000). Magnetite defines a vertebrate magnetoreceptor. Nature 406, 299-302.

Chapter 29

Cell cycle and growth regulation

29.1 Introduction

The act of division is the culmination of a series of events that have occurred since the last time a cell divided. The period between two mitotic divisions defines the somatic cell cycle. The time from the end of one mitosis to the start of the next is called interphase. The period of actual division, corresponding to the visible mitosis, is called M phase.

In order to divide, a eukaryotic somatic cell must double its mass and then apportion its components equally between the two daughter cells. Doubling of size is a continuous process, resulting from transcription and translation of the genes that code for the proteins constituting the particular cell phenotype. By contrast, reproduction of the genome occurs only during a specific period of DNA synthesis.

Mitosis of a somatic cell generates two identical daughter cells, each bearing a diploid complement of chromosomes. Interphase is divided into periods that are defined by reference to the timing of DNA synthesis, as summarized in **Figure 29.1:**

- Cells are released from mitosis into G1 phase, when RNAs and proteins are synthesized, but there is no DNA replication.
- The initiation of DNA replication marks the transition from Gl phase to the period of S phase. S phase is defined as lasting until all of the DNA has been replicated. During S phase, the total content of DNA increases from the diploid value of $2n$ to the fully replicated value of An.
- The period from the end of S phase until mitosis is called G2 phase; during this period, the cell has two complete diploid sets of chromosomes.

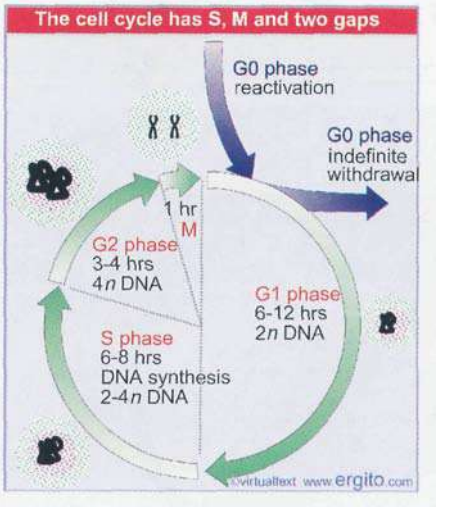


Figure 29.1 Overview: interphase is divided into the G1, S, and G2 periods. One cell cycle is separated from the next by mitosis (M). Cells may withdraw from the cycle into G0 or reenter from it.

By Book_Crazy [IND]

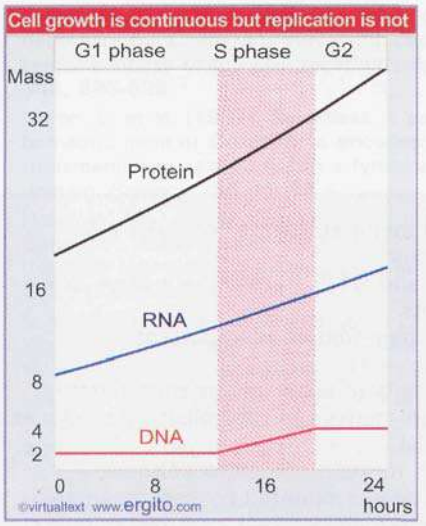


Figure 29.2 Synthesis of RNA and proteins occurs continuously, but DNA synthesis occurs only in the discrete period of S phase. The units of mass are arbitrary.

(S phase was so called as the synthetic period when DNA is replicated, G1 and G2 standing for the two "gaps" in the cell cycle when there is no DNA synthesis.)

The changes in cellular components are summarized in **Figure 29.2.** During interphase, there is little visible change in the appearance of the cell. The more or less continuous increase of RNA and protein contrasts with the discrete doubling of DNA. The nucleus increases in size predominantly during S phase, when proteins accumulate to match the production of DNA. Chromatin remains a compact mass in which no change of state is visible.

Mitosis segregates one diploid set of chromosomes to each daughter cell. Individual chromosomes become visible only during this period, when the nuclear envelope dissolves, and the cell is reorganized on a spindle. The mechanism for specific segregation of material applies only to chromosomes, and other components are apportioned essentially by the flow of cytoplasm into the two daughter cells. Virtually all synthetic activities come to a halt during mitosis.

In a cycling somatic animal cell, this sequence of events is repeated every 18-24 hours. Figure 29.1 shows that Gl phase usually occupies the bulk of this period, varying from ~6 h in a fairly rapidly growing animal cell to ~12 h in a more slowly growing cell. The duration of S phase is determined by the length of time required to replicate all the genome, and a period of 6-8 h is typical. G2 phase is usually the shortest part of interphase, essentially comprising the preparations for mitosis. M phase (mitosis) is a brief interlude in the cell cycle, usually <1 h in duration.

29.2 Cycle progression depends on discrete control points

Key Concepts

START in yeast cells and the restriction point in animal cells define the time in G1 when a cell makes a commitment to divide.

There are two points at which a decision may be taken on whether to proceed through another cell cycle. These are superimposed on the cycle in **Figure 29.3:**

- *Commitment to chromosome replication occurs in Gl phase.* If conditions to pass a "commitment point" are satisfied, there is a lag period and then a cell enters S phase. The commitment point has been defined most clearly in yeast cells, where it is called START. The comparable feature in animal cells is called the restriction point.
- *Commitment to mitotic division occurs at the end of G2.* If the cell does not divide at this point, it remains in the condition of having twice the normal complement of chromosomes.

How do cells use these two control points?

For animal cells growing in culture, Gl control is the major point of decision, and G2/M control is subsidiary. Cells spend the longest part of their cycle in G1, and it is the length of Gl that is adjusted in response to growth conditions. When a cell proceeds past G1, barring accidents it will complete S phase, proceed through G2, and divide. Cultured cells do not halt in G2. Control at Gl is probably typical of most diploid cells, in culture or *in vivo.*

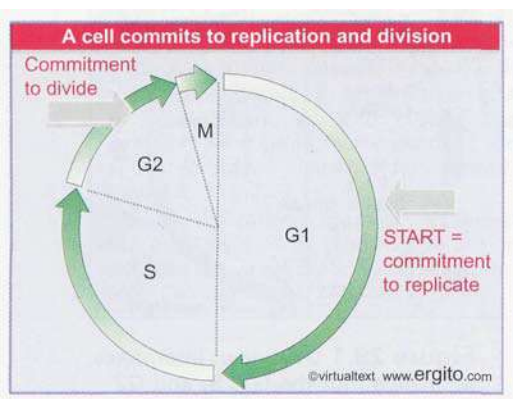


Figure 29.3 Commitment to replication occurs in G1, and commitment to division occurs in G2.

Some cell phenotypes do not divide at **all.** These cells are often considered to have withdrawn from the cell cycle into another state, resembling G1 but distinct from it because they are unable to proceed into S phase. This noncycling state is called GO. Certain types of cells can be stimulated to leave GO and reenter a cell cycle. Withdrawal from, or reentry into, the cell cycle can occur before the restriction point in G1 (see Figure 29.1).

Some cell types do halt in G2. In the diploid world, these are usually cells likely to be called upon to divide again; for example, nuclei at some stages of insect embryogenesis divide and rest in the tetraploid state. In the haploid world, it is more common for cells to rest in G2; this affords some protection against damage to DNA, since there are two copies of the genome instead of the single copy present in G1. Some yeasts can use either Gl or G2 as the primary control point, depending on the nutritional conditions. Some (haploid) mosses usually use G2 as the control point.

29.3 Checkpoints occur throughout the cell cycle

Key Concepts

- Orderly progress through the cell cycle is controlled by checkpoints, which prevent one stage from proceeding unless necessary earlier stages have been completed.

The upper part of **Figure 29.4** (in blue) identifies the critical points in the cell cycle:

For a cell with a cycle determined through G1 control, START marks the point at which the cell takes the basic decision: do I divide? Various parameters influence the ability of a cell to take this decision, including the response to external stimuli (such as supply of nutrients), and an assessment of whether cell mass is sufficient to support a division cycle. (Generally a cell is permitted to divide at a mass that is not absolute but is determined by a control that itself responds to growth rate.)

The beginning of S phase is marked by the point at which the replication apparatus begins to synthesize DNA.

The start of mitosis is identified by the moment at which the cell begins to reorganize **for** division.

Each of these events represents a discrete moment at which a molecular change occurs in a regulatory molecule. Once a cell has taken the decision to proceed at START, the other events will follow in order as the result of the cell cycle pathway.

Checkpoints that assess the readiness of the cell to proceed are superimposed on the pathway. The lower part of Figure 29.4 (in red) shows that *each checkpoint represents a control loop that makes the initiation of one event in the cell cycle dependent on the successful completion of an earlier event.* A checkpoint works by acting directly on the factors that control progression through the cell cycle.

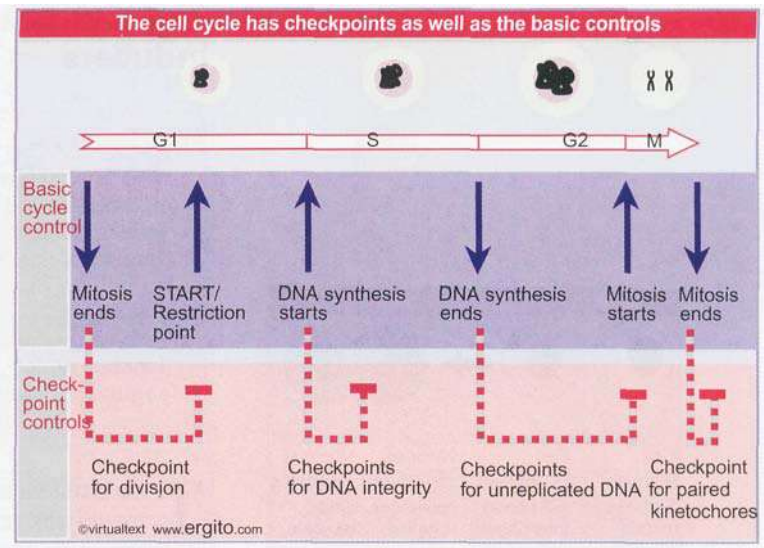


Figure 29.4 Checkpoints control the ability of the cell to progress through the cycle by determining whether earlier stages have been completed successfully. A horizontal red bar indicates the stage at which a checkpoint blocks the cycle.

By Book_Crazy [IND]

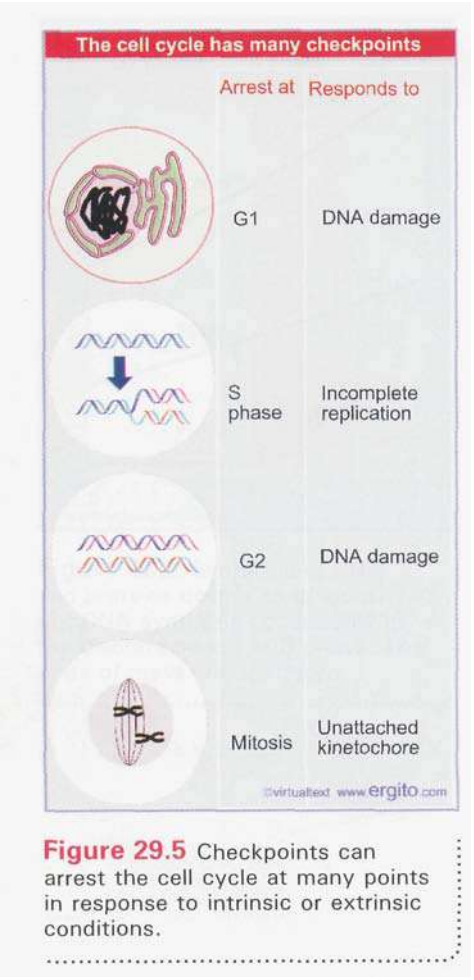


Figure 29.5 Checkpoints can arrest the cell cycle at many points in response to intrinsic or extrinsic conditions.

Figure 29.5 summarizes the events that trigger some important checkpoints. DNA damage is checked at every stage of the cycle. Checkpoints for specific events related to the individual stage occur at S phase and at mitosis.

Checkpoints operate within S phase to prevent replication from continuing if there are problems with the integrity of the DNA (for example, because of breaks or other damage). One important checkpoint establishes that all the DNA has been replicated. This explains why a common feature in the cycle of probably all somatic eukaryotic cells is that completion of DNA replication is a prerequisite for cell division.

Several checkpoints operate at mitosis, to ensure that the cell does not try to divide unless it has completed all of the necessary preceding events. There is a checkpoint that confirms that mitosis has been successfully completed before a cell can proceed through START to commit itself to another cycle.

Two types of cycle must be coordinated for a cell to divide:

- A cell must replicate every sequence of DNA once and only once. This is controlled by licensing factor (see *14.20 Licensing factor controls eukaryotic rereplication).* Having begun replication, it must complete it; *and it must not try to divide until replication has been completed.* This control is accomplished by the checkpoint at mitosis.
- The mass of the cell must double, so that there is sufficient material to apportion to the daughter cells. So *a cell must not try to start a replication cycle unless its mass will be sufficient to support division.* Cell mass influences ability to proceed through START and may also have a checkpoint at mitosis.

Some embryonic cycles bypass some of these controls and respond instead to a timer or oscillator. So the control of the cell cycle can be coupled as required to time, growth rate, mass, and the completion of replication.

29.4 Cell fusion experiments identify cell cycle inducers

Key Concepts

- The S phase activator is responsible for the ability of an S phase nucleus to induce DNA replication in a G1 nucleus in a heterokaryon.
- The M phase inducer is responsible for the ability of a mitotic cell to induce pseudo-mitosis in an interphase nucleus upon fusion of two cells.
- The M phase inducer is produced by activating the M phase kinase, which consists of a catalytic subunit and a cyclin regulatory subunit.

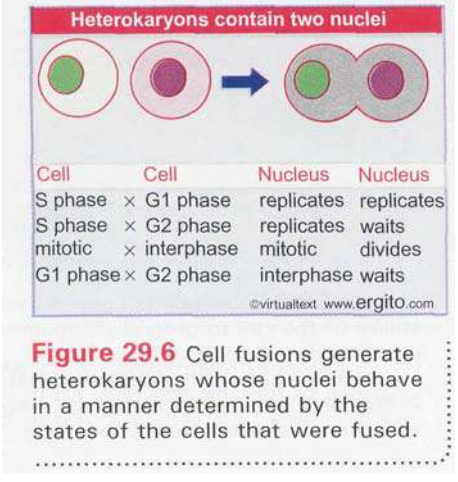


Figure 29.6 Cell fusions generate heterokaryons whose nuclei behave in a manner determined by the states of the cells that were fused.

The existence of different regulators at different stages of the cell cycle was revealed by early experiments that fused together cells in different stages of the cycle. As illustrated in **Figure 29.6,** cell fusion is performed by mixing the cells in the presence of either a chemical or viral agent that causes their plasma membranes to fuse, generating a hybrid cell (called a heterokaryon) that contains two (or more) nuclei in a common cytoplasm.

When a cell in S phase is fused with a cell in G1, both nuclei in the heterokaryon replicate DNA. This suggests that *the cytoplasm of the*

S phase cell contains an activator of DNA replication. The *quantity* of the activator may be important, because in fusions involving multiple cells, an increase in the ratio of S phase to G1 phase nuclei increases the rate at which Gl nuclei enter replication. The regulator identified by these fusions is called the S phase activator.

When a mitotic cell is fused with a cell at any stage of interphase, it causes the interphase nucleus to enter a pseudo-mitosis, characterized by a premature condensation of its chromosomes. This suggests that *an M phase inducer is present in dividing cells.*

Both the S phase and M phase inducers are present only transiently, because fusions between Gl and G2 cells do not induce replication or mitosis in either nucleus of the heterokaryon.

Figure 29.7 relates the cell cycle to the molecular basis for the regulatory cyclical events that control transitions between phases. Some of these events require the synthesis of new proteins or the degradation of existing proteins; other events occur by reversible activation or inactivation of pre-existing components. A minimum of three molecular activities must exist, and the striking thread that connects them is that all of these activities are controlled by phosphorylation:

- During G1, the cell passes START and becomes committed to a division cycle. The key event is a phosphorylation. The target protein is called RB, and its nonphosphorylated form represses transcription of genes that are needed for the cell cycle to advance. The repression is released when RB is phosphorylated.
- The period of S phase is marked by the presence of the S phase activator. This is a protein kinase. It is related to the kinase that activates mitosis (which was identified first and is better characterized).

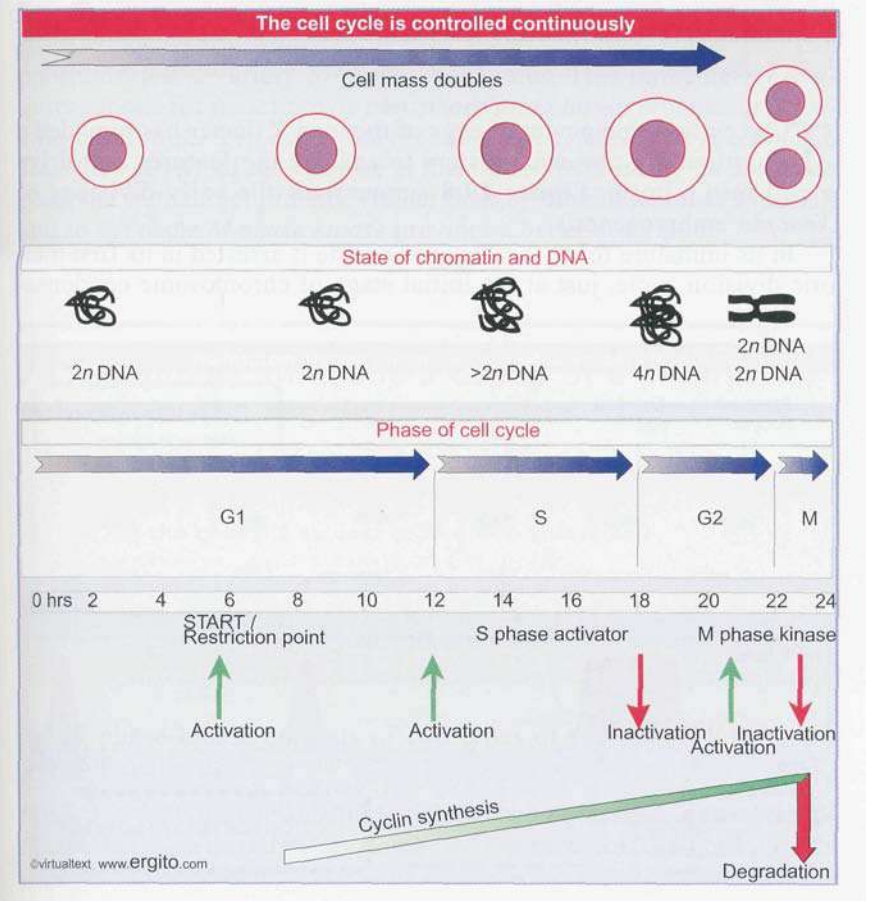


Figure 29.7 The phases of the cell cycle are controlled by discrete events that happen during G1, at S phase, and at mitosis.

By Book_Crazy [IND]

- Mitosis depends upon the activation of a pre-existing protein, the M phase kinase, which has two subunits. One is a kinase catalytic subunit that is activated by modification at the start of M phase. The other subunit is a cyclin, so named because it accumulates by continuous synthesis during interphase, but is destroyed during mitosis. Its destruction is responsible for inactivating M phase kinase and releasing the daughter cells to leave mitosis.

A striking feature of cell cycle regulation is that similar regulatory activities are employed in (probably) all eukaryotic systems. Some of these systems have cycles that superficially appear quite different from the normal somatic cycle. So very rapid divisions in which S phase alternates directly with mitosis characterize the development of the *Xenopus* egg, where entry into mitosis is controlled by M phase kinase, the very same factor that controls somatic mitosis. Yeast cells exist in a unicellular state, and certain species divide by an asymmetrical budding process; but control of entry into S phase and control of mitosis are determined by pathways that are related to those employed in the *Xenopus* egg. Homologous genes play related roles in organisms as distant as yeasts, insects, and mammals.

29.5 M phase kinase regulates entry into mitosis

Key Concepts

- MPF was originally identified as a factor that can induce G2 stage oocytes to enter meiosis.
- The same factor is responsible for inducing somatic cells to enter mitosis.
- Maturation promoting factor and M phase promoting factor are both manifestations of the M phase kinase.

The early development of eggs of the toad *X. laevis* has provided a particularly powerful system to analyze the features that drive a cell into mitosis. **Figure 29.8** summarizes the early divisions of *Xenopus* embryogenesis.

In its immature form, the *Xenopus* oocyte is arrested in its first meiotic division cycle, just at the initial stage of chromosome condensa-

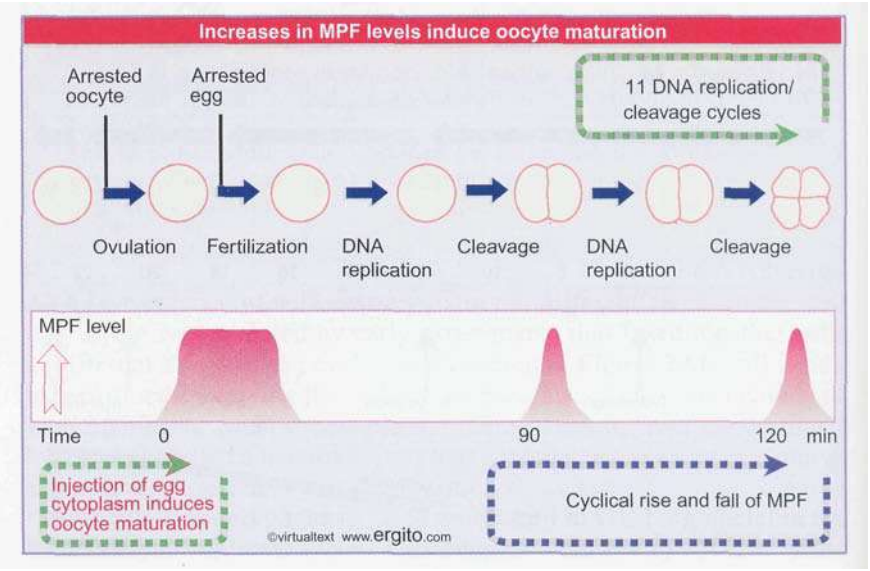


Figure 29.8 MPF induces M phase in early *Xenopus* development.

By Book_Crazy [IND]

tion. The closest correspondence to a somatic cycle is to the G2 stage. Ovulation occurs when the hormone progesterone releases the arrest, and the egg proceeds into meiosis. When the egg is laid, it is arrested towards the end of the second meiosis in a condition that corresponds to a somatic M phase.

The egg is a large structure (~1 mm. diameter in the case of *X. laevis)*, which contains a vast store of material needed for early divisions. Fertilization triggers a series of very rapid division cycles. The initial division takes ~90 minutes; then during the cleavage stage, another 11 divisions occur synchronously, each lasting ~30 minutes. These divisions in effect represent an alternation of S phase and mitosis; the major synthetic activity of the cleavage egg is the replication of DNA, all the required proteins having been previously synthesized and stored in the oocyte.

The size of the egg allows material to be purified from it. It is particularly useful that arrested oocytes (equivalent to G2 somatic cells) and arrested eggs (equivalent to M phase somatic cells) can be readily obtained. After eggs have been treated with progesterone, a factor can be obtained from their cytoplasm that, when injected into arrested oocytes, causes them to enter meiosis. The active component of the extract was called *maturation promoting factor* (MPF). The same factor can induce mitosis in somatic cells.

Because MPF turns out to have a general responsibility for causing somatic cells to enter M phase, MPF is now understood to stand for *M phase promoting factor*. The importance of MPF in inducing mitosis can be seen from its cyclical increase and decrease at the next stage of development, in the cleaving egg. As a synchronous wave of mitoses occurs in the cleavage egg, the level of MPF activity rises; as the mitoses are completed and S phase occurs, the MPF activity disappears.

MPF causes germinal vesicle (nuclear envelope) breakdown when injected into *Xenopus* oocytes, and induces several mitotic events in a cell-free system, including nuclear envelope disaggregation, chromosome condensation, and spindle formation. MPF is a kinase that can phosphorylate a variety of protein substrates. This immediately suggests a mode for its action: *by phosphorylating target proteins at a specific point in the cell cycle, MPF controls their ability to function.* In fact, the activity of the enzyme is most often assayed by its ability to phosphorylate target proteins (rather than by the induction of mitosis), and so the name *M phase kinase* provides a better description.

29.6 M phase kinase is a dimer of a catalytic subunit and a regulatory cyclin

Key Concepts

- Cdc2 is the catalytic subunit of M phase kinase and phosphorylates target proteins on Ser or Thr.
- The regulatory subunit is cyclin A or cyclin B.
- Activation results from covalent modification of the Cdc2 subunit.
- Inactivation results from degradation of the cyclin subunit.

M phase kinase consists of two types of subunit with different functions:

Cdc2 is the *catalytic subunit* which phosphorylates serine and threonine residues in target proteins. It is named for the homologous protein in *S. pombe*.

- Its partner is a cyclin; this is a *regulatory subunit* which is necessary for the kinase to function with appropriate substrates.

The crystal structure of a dimer shows that the cyclin induces a change in the conformation of its partner that is necessary to create the active kinase site. So a catalytic subunit by itself is inactive, and can be active only when joined by a cyclin partner.

An M phase kinase typically has a single type of catalytic subunit (Cdc2), but may have any one of several alternative cyclin partners. There are two general types of mitotic cyclins, A and B. They are characterized by the sequence of a stretch of ~150 amino acids (sometimes known as the cyclin box). In mammals and frogs, the B cyclins can be divided into the subtypes B1 and B2.

So there are (at least) two general forms of the M phase kinase: Cdc2-cyclin A and Cdc2-cyclin B. The common properties of the cyclins suggest that they have the same type of function: to influence the activity of the catalytic subunit of M phase kinase. Yet the two classes of cyclins have only weak similarity and follow a different temporal and spatial pattern of behavior. Although they are involved in the timing and localization of M phase kinase activity, we do not know how they function as regulatory subunits in determining the specificity of the catalytic subunit. We should like to know in particular whether Cdc2-cyclin A and Cdc2-cyclin B recognize different proteins as substrates.

The events that activate M phase kinase at G2/M and inactivate it during M phase identify crucial points in the cell cycle. Activation and inactivation are achieved by different types of action:

- *Activation requires modification of the catalytic subunit.* In most cells, the level of Cdc2 remains constant through the cycle, and is in excess compared to the cyclins. Cyclins are necessary to turn on the M phase specific kinase activity of Cdc2. However, they cannot provide the activating event because they accumulate to a maximum level before the kinase activity appears. The intact Cdc2-cyclin dimer accumulates in an inactive form; modification of Cdc2 is the critical event that triggers the G2/M transition.
- *Inactivation is achieved by the physical destruction of the cyclin.* Cyclins were originally named for their property of accumulating continuously through the cell cycle; then they are destroyed abruptly by proteolysis during mitosis (see Figure 29.7). The timing of cyclin destruction is characteristic; typically A precedes B (by a few minutes in embryonic divisions, by rather longer in a cultured cell cycle), and this difference appears to be common to all cells.

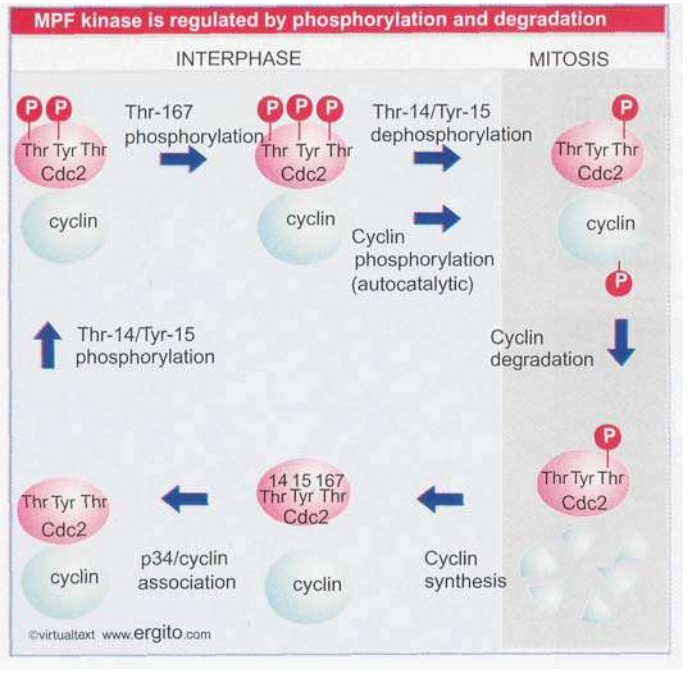


Figure 29.9 The activity of M phase kinase is regulated by phosphorylation, dephosphorylation, and protein proteolysis. The 3 phosphorylated amino acids are Thr-14, Tyr-15, and Thr-161. The first two are in the ATP-binding site.

What activates the catalytic activity of the M phase kinase? Cdc2 is itself a phosphoprotein, and its state of phosphorylation is a crucial determinant of activity. The events involved in the cycle of activation and inactivation of M phase kinase are summarized in **Figure 29.9.** For M phase kinase to be active, phosphate groups must be absent at some positions, but present at another position.

Two residues located within the ATP-binding site of Cdc2 must be *dephosphorylated* in order to activate the kinase (in mouse cells). The phosphates are located on Thr-14 and Tyr-15; both are removed by the same (dual specificity) phosphatase. M phase kinase is autocatalytic, that is, the activation of a small amount of the kinase is sufficient to trigger activation of the rest. This could be explained if M phase kinase itself activates the phosphatase.

Another phosphorylation occurs on Thr-161 of Cdc2. This phosphate group is added in G2 and removed at the end of mitosis. The phosphate is *required* for activity of Cdc2; mutations introducing an amino acid that cannot be phosphorylated at this site inactivate the kinase activity.

Association of the catalytic and cyclin subunits, and phosphorylation and dephosphorylation, occur in a specific order. Cyclin B associates specifically with the tyrosine-*dephosphorylated* form of Cdc2 *in vitro*. This generates a potentially active dimer. However, formation of the dimer causes Thr-14/Tyr-15 to be phosphorylated. So *association with cyclin induces the inactivating event.* The dimer is then maintained in its inactive form until the phosphates are removed.

Cyclins A and B both have a short motif near the N-terminus—the cyclin destruction box—which is required to make the cyclin a target for proteolysis. Cyclins are degraded by a common proteolytic system, the proteasome (a complex containing proteolytic activities that recognizes its targets when ubiquitin is added to them; see *8.32 The proteasome is a large machine that degrades ubiquitinated proteins*).

Destruction of the cyclin subunits is responsible for inactivating M phase kinase during mitosis, and is necessary for cells to exit mitosis. A truncated cyclin B that lacks the N-terminal region is resistant to proteolysis. When this protein is synthesized in *Xenopus* eggs, or in a cell-free extract that undertakes some of the typical cycling reactions, it causes anaphase arrest. This is the basis for concluding that *loss of kinase activity is a prerequisite for completing mitosis.*

An important question is how the roles of cyclin A and cyclin B differ in mitosis. A hint that their functions are different is provided by differences in the timing of their synthesis and destruction, but there is as yet no direct evidence to show whether both cyclins are required for passage through mitosis in an animal cell or whether they are redundant with one another.

29.7 Protein phosphorylation and dephosphorylation control the cell cycle

Key Concepts

- M phase kinase triggers mitosis by phosphorylating a wide variety of substrates.
- Mitosis is terminated by dephosphorylating the substrates.
- One of the best substrates for M phase kinase, which is often used to assay its activity, is histone H1.

Phosphorylation (catalyzed by kinases) and dephosphorylation (catalyzed by phosphatases) are the critical events that regulate the cell cycle. *They are used both to control the activities of the regulatory circuit itself and to control the activities of the substrates that execute the decisions of the regulatory circuit.*

The cell cycle regulatory circuit consists of a series of kinases and phosphatases that respond to external signals and checkpoints by phosphorylating or dephosphorylating the next member of the pathway. The ultimate readout of the circuit is to determine the activity of M phase kinase *(or* the S phase kinase) by controlling its state of phosphorylation.

Activation of M phase kinase is the event that triggers onset of M phase. Inactivation is necessary to exit M phase. This suggests that the

events regulated by M phase kinase are reversible: *phosphorylation of substrates is required for the reorganization of the cell into a mitotic spindle, and dephosphorylation of the same substrates is required to return to an interphase organization.*

What are the targets for M phase kinase? A major reorganization of the cell occurs at mitosis, and the ability of MPF to induce mitosis implies that the M phase kinase triggers these activities. Does M phase kinase act directly or indirectly upon the various potential substrates? Two general models could be proposed for its role:

- It may be a "master regulator" that phosphorylates target proteins that in turn act to regulate other necessary functions—a classic cascade.
- Or it may directly phosphorylate the proteins that are needed to execute the regulatory events or cell reorganization involved in the cycle.

The only common feature in substrates that are phosphorylated by M phase kinase is the presence of the duo Ser-Pro, flanked by basic residues (most often in the form Ser-Pro-X-Lys). Potential substrates, based upon the ability of M phase kinase preparations to phosphorylate targets *in vitro,* include H1 histone (perhaps required to condense chromosomes), lamins (possibly required for nuclear envelope breakdown), nucleolin (potentially involved in the arrest of ribosome synthesis), and other structural and enzymatic activities. The strength of the evidence varies as to which of these targets is phosphorylated *in vivo* in a cyclic manner and whether M phase kinase is in fact the active enzyme. From the variety of substrates, however, it seems likely that M phase kinase acts directly upon many of the proteins that are directly involved in the change in cell structure at mitosis.

How can we determine whether a potential substrate is an authentic target for Cdc2 in the cell cycle? The same sites should be phosphorylated by Cdc2 *in vitro* that are cyclically phosphorylated *in vivo* at time(s) when Cdc2 is known to be active. Ideally it should be possible to show that a mutation in Cdc2 kinase activity blocks the phosphorylation *in vivo,* but this is at present practical only in yeast. To conclude that the phosphorylation is a significant event in the cycle, some function of the protein must be altered by the presence of phosphate. This can be tested by making mutations at the amino acid that is phosphorylated to determine whether the absence of phosphorylation blocks a mitotic function.

The best characterized substrate for Cdc2 kinase is the H1 histone (one of the 5 histones that are the major protein constituents of chromatin; see *20 Nucleosomes).* It has been known for a long time that H1 is phosphorylated during the cell cycle, with 2 phosphate groups added during S phase, and 4 further phosphate groups added during mitosis. The major H1 kinase activity of the cell is provided by M phase kinase.

What purpose the phosphorylation of H1 serves in the cell cycle remains a matter for speculation, since no effects upon chromatin structure have been directly demonstrated. It is reasonable to suppose that it might be connected with chromosome condensation at M phase. Not enough is known about the timing of modification at S phase to wonder whether it is concerned with preparations for replication (which might require uncoiling) or with the consequences of replication (when preparations for mitosis could begin). But H1 histone is an exceedingly good substrate for kinases based on the Cdc2 engine, with the result that H1 kinase activity has become the usual means by which this enzyme is assayed *in vitro.* An illustration of the appeal of this assay is its application to *S. cerevisiae,* where H1 kinase activity is routinely measured to assess the cyclic activity of M phase kinase, although this yeast in fact is unusual in containing no H1 histone!

29.8 Many cell cycle mutants have been found by screens in yeast

Key Concepts

- Blocking the cell cycle is likely to kill a cell, so mutations in the cell cycle must be obtained as conditional lethals.
- Screens for conditional lethal cell cycle mutations identify ~80 genes in either *S. cerevisiae* or S. *pombe*.

To define the circuit for cell cycle control, we need to identify the genes that regulate progression through the cycle. We define a mutation in the cell cycle as one that *blocks the cell cycle at a specific stage,* a definition that excludes mutations in genes that control continuous processes of growth and metabolism. But the approach of characterizing mutants that are unable to proceed through the cycle has not been straightforward to develop. A mutation that prevents a cell from dividing will be lethal; and the reverse is also true, insofar as any lethal mutation will stop cells from dividing. It has therefore been difficult to devise procedures to isolate cell cycle mutants of animal cells, or, indeed, to demonstrate that potential mutants have specific blocks in the cell cycle.

Because of their nature, cell cycle mutants must be obtained as conditional lethals, so that although they are unable to grow under the conditions of isolation, they can be maintained by growth in other conditions. A series of such mutants has been isolated in two yeasts, in which the block to the cell cycle can be seen to affect the visible phenotype of the cell.

The mitotic cycles of fission yeast *(S. pombe)* and baker's yeast *(S. cerevisiae)* are summarized in **Figure 29.10.**

The cell cycle of *S. pombe* is divided into the usual phases (G1, S, G2, M). The cell grows longitudinally and then divides; progress through the cell cycle can be assessed (approximately) by the physical length of the cell (which doubles from 7-8 μm to 14-16 μm).

The cycle of *S. cerevisiae* is unusual. Cells proceed almost directly from S phase into division, so there is effectively very little or no G2 phase. (A short G2 phase is shown in this and subsequent figures for the purpose of localizing the relative timing of events concerned with the transition into M phase.) And instead of an equal division of the cell, the daughter cell grows as a bud off the mother cell, eventually obtaining its independence when it is released as a small separate cell. Again the cell cycle can be followed visually in terms of the growth of the bud. Mitosis itself shares some unusual features in both types of yeast; the nuclear membrane does not break down, and segregation of chromosomes therefore is compelled to occur within the nucleus.

Extensive screens for cell division cycle or cdc mutants have been performed in both yeasts. Initial isolation relies upon the criterion that cells accumulate at a particular stage of the cell cycle at an elevated temperature (36°C), but continue normally through the cycle at 23°C. The mutant phe-

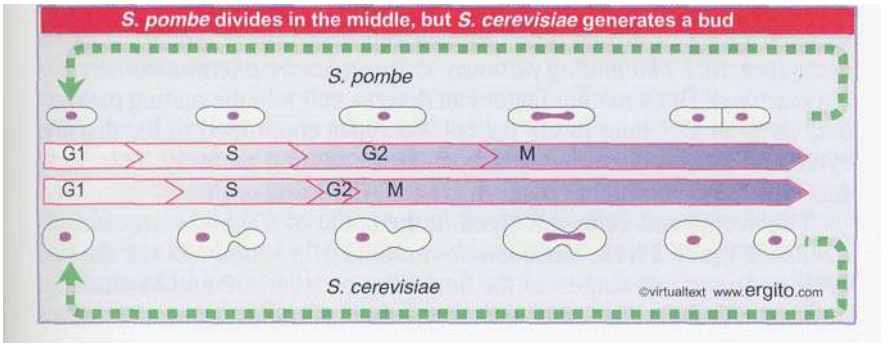


Figure 29.10 *S. pombe* lengthens and then divides during a conventional cell cycle, while *S. cerevisiae* buds during a cycle in which G2 is absent or very brief, and M occupies the greatest part.

By Book_Crazy [IND]

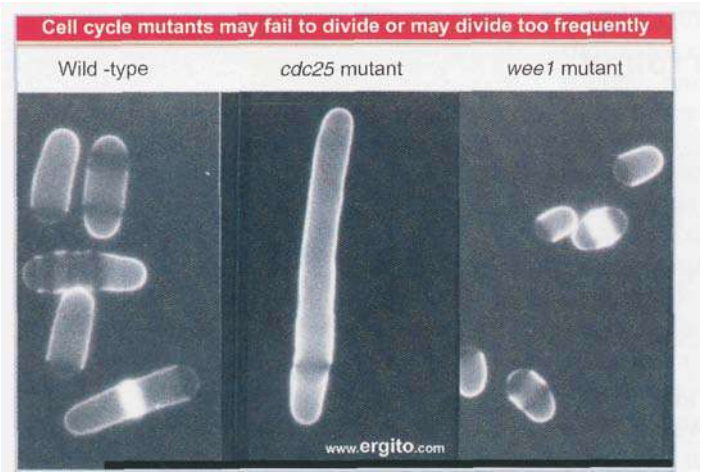


Figure 29.11 *S. pombe* cells are stained with calcofluor to identify the cell wall (surrounding the yeast cell) and the septum (which forms a central division when a cell is dividing). Wild-type cells double in length and then divide in half, but *cdc* mutants grow much longer, and *wee* mutants divide at a much smaller size. Photograph kindly provided by Paul Nurse.

notype allows cells to continue growing larger while the cycle is blocked, causing an obvious aberration; in *S. pombe* the cells become highly elongated, and in *S. cerevisiae* they fail to bud. **Figure 29.11** compares cell cycle mutants with wild-type *S. pombe*. The left panel shows a group of normal cells; the center panel shows a *cdc* mutant that is blocked in the ability to divide, and has therefore become elongated because it has continued to grow.

Because the mutants are temperature sensitive, the time at which a mutation takes effect can be determined by temperature shift protocols in which cells are shifted up in temperature at a specific point in the cycle. If this point is prior to the point at which the gene product acts, the cells halt at the execution point, but if the point is past the time when the protein function is needed, the cells continue their cycle.

Of the order of 80 *cdc* genes have been identified in each species, but not all of these loci are concerned with regulating the cell cycle. Many represent genes whose products are needed for metabolic purposes; for example, absence of enzymes that replicate DNA or synthesize the nucleotide precursors can block progression through S phase.

29.9 Cdc2 is the key regulator in yeasts

Key Concepts

- *cdc* mutants of S. *pombe* fall into two groups that block the cell cycle either at G2/M or at START in G1.
- Different *cdc2* mutant alleles may block the cycle at either of these stages.
- *CDC28* is the homologous gene in *S. cerevisiae.*
- Homologues of *cdc2* are found in all eukaryotic organisms.
- The active form of M phase kinase is phosphorylated on Thr-161; the inactive form is phosphorylated on Tyr-15 (and in animal cells also on Thr-14).

Yeast cells may exist in either haploid or diploid form. They have two forms of life cycle, as illustrated in **Figure 29.12.** Haploid cells double by mitosis. A haploid cell has a mating type of either *a* or a. Haploids of opposite types enter the sexual mating pathway, in which they conjugate (fuse) to form diploid cells. The diploid cells in turn sporulate to form haploid cells by meiosis (see *18.2 The mating pathway is triggered by pheromone-receptor interactions).*

A crucial point in the cell cycle is defined by the behavior of haploid cells. A haploid cell decides at a point early in G1 whether to proceed through a division cycle or to mate. The decision is influenced by environmental factors; for example, cells of opposite mating type must be present for conjugation to occur. In fact, a mating factor (a polypeptide hormone) secreted by a cell of one type causes a cell of the other type to arrest its cycle (see *18.2 The mating pathway is triggered by pheromone-receptor interactions*). But a mating factor can divert a cell into the mating pathway *only early in Gl;* later in Gl the cell becomes committed to the division cycle and cannot be stimulated to enter the mating pathway. This assay was how the commitment point in Gl (START) was originally identified.

The events and genes involved in the cycle of *S. pombe* are summarized in **Figure 29.13.** Various *cdc* mutants of *S. pombe* block the cell cycle at one of two stages: at the boundary between G2 and M phase; or in Gl at START.

cdc2 is identified as a crucial regulator by its involvement at *both* stages of cell cycle block: mutants of *cdc2* may be blocked prior to START or prior to M phase (depending on the point a cell had reached in the cycle when the mutation took effect).

The homologous gene in *S. cerevisiae* is called *CDC28*. A deficiency in the gene of either type of yeast can be corrected by the homologue from the other yeast. In both yeasts, proteins that resemble B-type cyclins associate with Cdc2 or CDC28 at G2/M.

The crucial breakthrough in understanding the nature and function of M phase kinase was the observation that *Xenopus Cdc2* is the homologue of *cdc2* of *S. pombe* and *CDC28* of *S. cerevisiae*. (When this discovery was made, the *Xenopus* protein was called p34, after its size, but then it was renamed Cdc2 after its name in *S. pombe.)*

The activity of the Cdc2 catalytic subunit in these dimers (and of the equivalent CDC28 in *S. cerevisiae)* is controlled by phosphorylation in the same way as Cdc2 in animal cells. A difference is that in yeast there is no Thr-14, so there are only two relevant sites: Tyr-15 where phosphorylation is inhibitory; and Thr-161 where phosphorylation is required.

The existence of a Cdc2 catalytic subunit, in organisms as diverse in evolution as yeasts, frogs, and mammals, identifies the key feature of cell cycle control. Conservation of function is indicated by the ability of the cloned human gene to complement the deficiency in *cdc2* mutants of *S. pombe*. This was the crucial experiment that identified Cdc2 as the universal regulator.

(Ability to compensate for the deficiency of a specific yeast mutant has been used with great effect to identify higher eukaryotic genes homologous to several cell cycle regulators of yeast. The assay introduces a cloned animal gene into a yeast mutant and then identifies cells that resume growth. It is quite remarkable that the control of the cell cycle has been so well conserved as to make this possible. However, it should be remembered that this assay does not impose very rigorous demands, and sometimes leads to the identification of a gene that is only loosely related to the mutant function.)

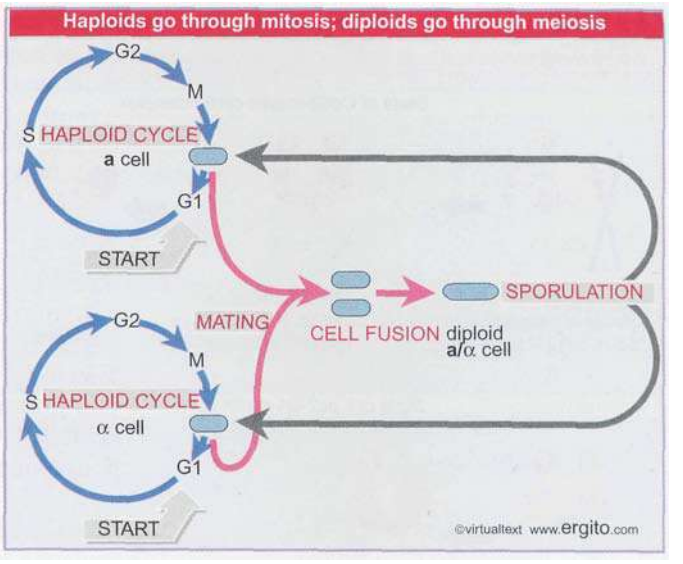


Figure 29.12 Haploid yeast cells of either a or a mating type may reproduce by a mitotic cycle. Cells of opposite type may mate to form an a/a diploid. The diploid may sporulate to generate haploid spores of both a and a types.

29.10 Cdc2 is the only catalytic subunit of the cell cycle activators in *S. pombe*

Key Concepts

- During mitosis Cdc2 exists in a dimer with one of the mitotic cyclins (A or B).
- During G1 Cdc2 forms a dimer with a G1 cyclin.
- G1 cyclins are distantly related to mitotic cyclins but are not regulated by degradation.
- The Cdc2-G1 cyclin dimer must be activated in order to enter S phase.

One of the key concepts of cell cycle regulation is that each stage is controlled by a dimer consisting of a Cdc2-like catalytic subunit and a cyclin-like regulatory subunit. Either or both of these subunits may be changed between the regulatory stages, and in some cases there may in fact be several alternative versions of the dimer at one stage.

In yeast, there is only one type of catalytic subunit, and the differences between regulatory stages are determined by its partner. Thus

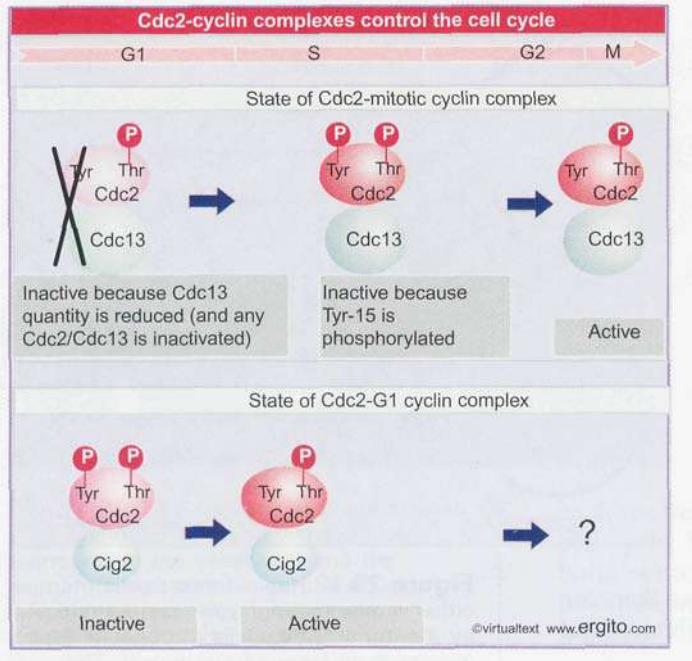


Figure 29.14 The state of the cell cycle in *S. pombe* is defined by the forms of the Cdc2-cyclin complexes.

Cdc2 in *S. pombe* has different partners at mitosis and during G1. At mitosis, its partner is the product of *cdcl3*, generating an M phase kinase that resembles the Cdc2-B cyclin dimer of animal cells. During G1, the active form of Cdc2 is associated with a B-like cyclin, cig2 (there is also a related cyclin, cig1).

Figure 29.14 summarizes changes in phosphorylation of the alternative dimers during the cell cycle.

The upper part of the figure summarizes the condition of the M phase kinase. At mitosis, the Cdc2 subunit of the Cdc2/Cdc13 dimer is in the active state that lacks the phosphate at Tyr-15 and has the phosphate at Thr-161. At the end of mitosis, kinase activity is lost when Cdc13 is degraded. The state of phosphorylation of Cdc2 does not change at this point. As new Cdc13 is synthesized, it associates with Cdc2, but the dimeric complex is maintained in an inactive state by another protein (see Figure 29.19). After START, however, activity of the Cdc2/Cdc13 dimer is inhibited by addition of the phosphate to Tyr-15. Removal of the inhibitory phosphate is the trigger that activates mitosis.

The lower part of the figure shows that Cdc2 forms a dimer with cig2 during G1. This forms a kinase with the same general structure as the M phase kinase. This kinase is in effect a counterpart to the M phase kinase, and it controls progression through the earlier part of the cell cycle. cig2 resembles a B-type cyclin. The dimer is converted from the inactive state to the active state by dephosphorylation of the Tyr-15 residue of Cdc2 at or after START. We do not know what happens to the Cdc2-cig2 dimer at the transition from S phase into G2.

The existence of alternative dimers during the cell cycle suggests the concept that *cyclins can be classified according to their period of active partnership with cdc2* as G1 cyclins or mitotic cyclins (also known as G2 cyclins). (Note that the cyclins defined in this way by their partnership with Cdc2 or a Cdc2-like protein do not necessarily have the property of cyclic degradation by which the original cyclins were defined. Cyclic degradation is a general property specifically for mitotic cyclins in animal cells.)

We may take the concept of cyclin classification further and suggest that *the phase of the cell cycle is defined by the nature of the active cdc2-cyclin dimer*. The period of Gl and S phase is defined by the presence of an active Cdc2-Gl cyclin dimer. Progression through G1 into S is controlled by activation of the Cdc2-Gl cyclin. The period of G2 and mitosis is characterized by an active Cdc2-mitotic cyclin dimer. The transition from G2 into mitosis is controlled by activation of the Cdc2-mitotic cyclin.

The original model for the employment of Cdc2 proposed that the available form of the Cdc2-cyclin dimer was regulated simply by replacement of one cyclin with another. However, it seems now that the mitotic and G1 forms may coexist in the yeast cell, but that their activities are differently regulated at each stage of the cycle.

29.11 *CDC28* acts at both START and mitosis in *S. cerevisiae*

Key Concepts

- CDC28 forms dimers with each of 4 B-type cyclins at mitosis.
- It forms dimers with each of 3 G1 cyclins, any one of which is sufficient to activate START.

The analysis of *cdc* mutants in *S. cerevisiae* shows that more than one type of cycle is required to proceed from one division to the next, although the cycles are connected at crucial points. (Remember that mitosis in *S. cerevisiae* is unusual morphologically, and chromosome segregation occurs within the intact nucleus; see Figure 29.10.) **Figure 29.15** shows how the three cycles of *S. cerevisiae* relate to the conventional phases of the overall cell cycle:

- The *chromosome cycle* comprises the events required to duplicate and separate the chromosomes, consisting of the initiation, continuation, and completion of S phase, and nuclear division. A mutation such as *cdc8* stops this cycle in S phase.
- Mutations in the chromosome cycle do not stop the *cytoplasmic cycle,* which consists of bud emergence and nuclear migration into the bud (visualized at the start of M phase in Figure 29.10). This cycle can be halted before bud emergence by *cdc24,* but the mutation does not prevent chromosome replication.
- The *centrosome cycle* consists of the events associated with the duplication and then separation of the spindle pole body (SPB), which in effect substitutes for the centrosome and organizes microtubules to allow chromosome segregation within the nucleus. Blocking this cycle, for example with *cdc31,* does not prevent S phase or bud emergence.

Completion of an entire cell cycle requires all three constituent cycles to be functional, since nuclear division requires both the chromosome and centrosome cycles, and cytokinesis requires these and the cytoplasmic cycle.

The decision on whether to initiate a division cycle is made before the point START. The crucial gene in passing START is *CDC28* (the homologue of *cdc2* of *S. pombe).* Mutations in *CDC28* prevent all three cycles from proceeding. The ability to pass START is determined by environmental conditions, since stationary phase populations that are limited by nutrients are arrested at START. Morphologically, the cells of such a population are arrested at point after cell separation and prior to SPB duplication, bud emergence, and DNA replication, which is to say that when *S. cerevisiae* exhaust their nutrients, they complete the current cycle and arrest all three cycles before START. In terms of phases of the cell cycle, this corresponds to a point early in G1.

Most mutants in *CDC28* are blocked at START, which has therefore been characterized as the major point for *CDC28* action. This contrasts with the better characterized function of M phase kinase at M phase. However, one mutant in *CDC28* passes START normally, but is inhibited in mitosis, suggesting that CDC28 is required to act at both points in the cycle. In fact, the function of *CDC28* is required at three stages: classically before START; then during S phase; and (of course) prior to M phase.

There is a difference in emphasis concerning control of the cell cycle in fission yeast and bakers' yeast, since in *S. pombe* we know most about control of mitosis, and in *S. cerevisiae* we know most about control of START. However, in both yeasts, the same principle applies that a single catalytic subunit (Cdc2/CDC28) is required for the G2/M and the G1/S transitions. It has different regulatory partners at each transition, using B-like (mitotic) cyclins at mitosis and G1 cyclins at the start of the cycle. A difference between these yeasts is found in the number of cyclin partners that are available at each stage.

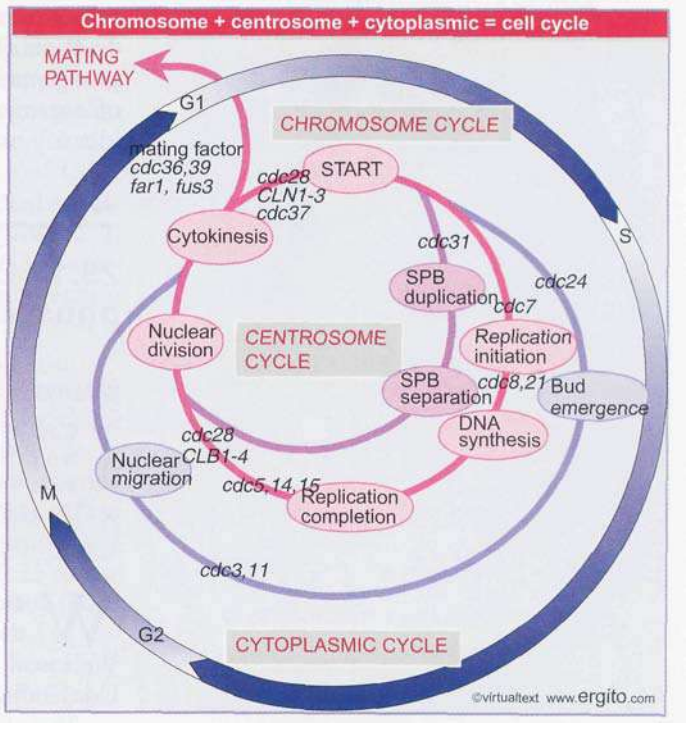


Figure 29.15 The cell cycle in *S. cerevisiae* consists of three cycles that separate after START and join before cytokinesis. Cells may be diverted into the mating pathway early in G1.

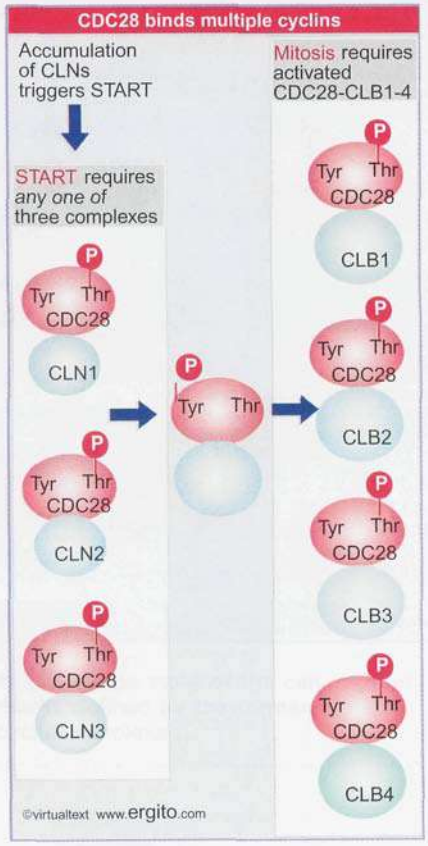


Figure 29.16 Cell cycle control in *S. cerevisiae* involves association of CDC28 with redundant cyclins at both START and G2/M.

A single regulatory partner for Cdc2 is used at mitosis in *S. pombe*, the product of *cdc13*. In *S. cerevisiae*, there are multiple alternative partners. These are coded by the *CLB1-4* genes, which code for products that resemble B cyclins and associate with CDC28 at mitosis. Sequence relationships place the genes into pairs, *CLB1-2* and *CLB3-4*. Mutation in any one of these genes fails to block division, but loss of the *CLB1-2* pair of genes is lethal. Constitutive expression of *CLB1* prevents cells from exiting mitosis.

The state of CDC28 in the *S. cerevisiae* cell cycle is summarized in **Figure 29.16.** The G1 cyclins were not immediately revealed by mutations that block the cell cycle in G1. The absence of such mutants has been explained by the discovery that three independent genes, *CLN1, CLN2, and CLN3 all must be inactivated* to block passage through START in *S. cerevisiae.* Mutations in any one or even any two of these genes fail to block the cell cycle; thus the *CLN* genes are functionally redundant. The *CLN* genes show a weak relationship to cyclins (resembling neither the A nor B class particularly well), although they are usually described as Gl cyclins.

Accumulation of the CLN proteins is the rate-limiting step for controlling the G1/S transition. Blockage of protein synthesis arrests the cycle by preventing the proteins from accumulating. The half-life of CLN2 protein is ~15 minutes; its accumulation to exceed a critical threshold level could be the event that triggers passage. This instability presents a different type of control from that shown by the abrupt destruction of the cyclin A and B types. Dominant mutations that truncate the protein by removing the C-terminal stretch (which contains sequences that target the protein for degradation) stabilize the protein, and as a result Gl phase is basically absent, with cells proceeding directly from M phase into S phase. Similar behavior is shown by the product of *CLN3*.

The redundancy of the *CLN* genes and the *CLB* genes is a feature found at several other stages of the cell cycle. In each case, the hallmark is that deletion of an individual gene produces a cell cycle mutant, but deletion of both or all members of the group is lethal. Members of a group may play overlapping rather than identical functions. This form of organization has the practical consequence of making it difficult to identify mutants in the corresponding function.

29.12 Cdc2 activity is controlled by kinases and phosphatases

Key Concepts

- Cdc25 is a phosphatase that removes the inhibitory phosphate from Tyr-15 of Cdc2 in order to activate the M phase kinase.
- Wee1 is a kinase that antagonizes Cdc25 by phosphorylating Tyr-15 when the cell size is small.

We can divide *cdc* genes into two classes, defined by the stage of the cell cycle at which the effects of a mutation are manifested. We know most about the circuit that controls the state of the Cdc2/Cdc 13 dimer at mitosis.

The states of Cdc2 complexes and the enzymes that act upon them during the *S. pombe* cell cycle are summarized in **Figure 29.17.** The Cdc2-cig2 dimer is phosphorylated during G1. By G2, the predominant form is the Cdc2/Cdc 13 dimer whose activity is controlled by antagonism between kinases and phosphatases that respond to environmental signals or to checkpoints.

cdc25 codes for a tyrosine phosphatase that is required to dephosphorylate Cdc2 in the Cdc2/Cdc13 dimer. It is responsible for the key dephosphorylating event in activating the M phase kinase. It is not a very powerful phosphatase (its sequence is atypical), and the quantities of Cdc25 and Cdc2 proteins are comparable, so the reaction appears almost to be stoichiometric rather than catalytic.

The level of Cdc25 increases at mitosis, and its accumulation over a threshold level could be important. Cdc25 executes the checkpoint that ensures S phase is completed before M phase can be activated. In mutants of *cdc2* that do not require *cdc25*, or in strains that over-express *cdc25*, blocking DNA replication does not impede mitosis. (Wild-type cells arrest in the cell cycle if DNA synthesis is prevented, for example, by treatment with hydroxyurea. But these mutant cells attempt to divide in spite of the deficiency in replication, with lethal consequences.)

Under normal conditions, the cell division cycle is related to the size of the cell. In poor growth conditions, when the cells increase in size more slowly, G1 becomes longer, because START does not occur until the cells attain a critical size. This is a protection against starting a cell division cycle and risking division before the amount of material is adequate to support two daughter cells. *cdc* mutants typically delay the onset of mitosis and lead either to cell cycle arrest or to division at increased size (as shown for *cdc25* in Figure 29.11).

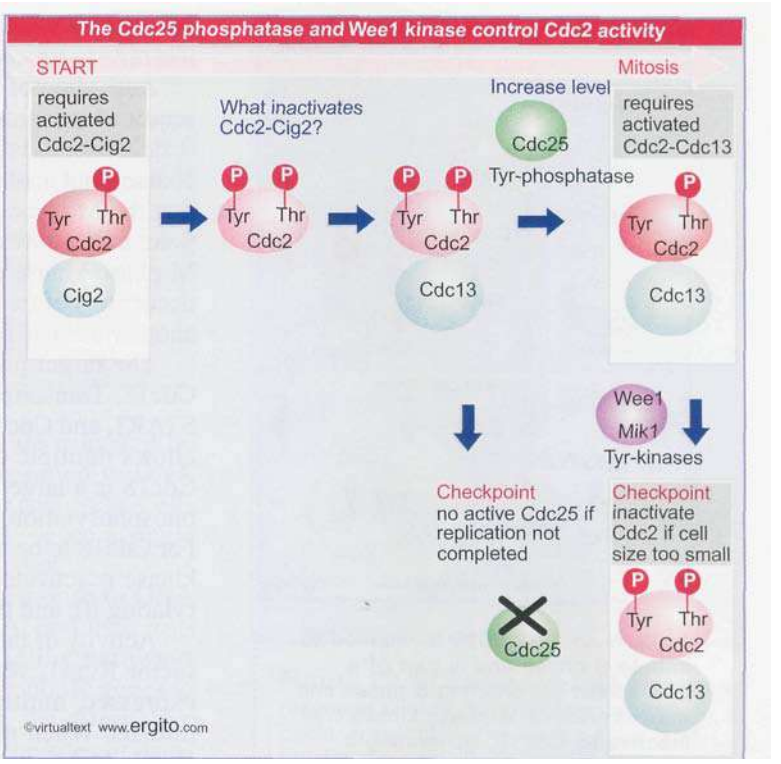


Figure 29.17 Cell cycle control in S. *pombe* involves successive phosphorylations and dephosphorylations of Cdc2. Checkpoints operate by influencing the state of the Tyr and Thr residues.

Genes involved in cell size control are identified by mutations with the opposite property: they advance cells into mitosis and therefore divide at reduced size. The *weel* gene takes its name from this phenotype (see the right panel of Figure 29.11). This behavior suggests that *weel* usually inhibits cells from initiating mitosis until their size is adequate. This identifies a checkpoint that prevents the activation of Cdc2 until an adequate mass has been attained. *weel* codes for a "dual specificity" kinase: it can phosphorylate serine/threonine and tyrosine. It inhibits Cdc2 by phosphorylating Tyr-15. Another gene, *mikl*, has similar effects. The deletion of both *weel* and *mikl* is lethal, suggesting that either gene can fulfill the same function. Redundancy of this sort is a common theme in the yeast cell cycle.

The products of *weel* and *cdc25* play antagonistic roles, as shown in Figure 29.17. The kinase activity of *weel* acts on Tyr-15 to inhibit Cdc2 function. The phosphatase activity of *cdc25* acts on the same site to activate Cdc2. Mutants that over-express *cdc25* have the same phenotype as mutants that lack *weel*. Regulation of Cdc2 activity is therefore important for determining *when* the yeast cell is ready to commit itself to a division cycle; inhibition by weel and activation by Cdc25 allow the cell to respond to environmental or other cues that control these regulators.

It is striking that all of the genes known to affect the G2/M boundary appear to have been widely conserved in evolution. Extending beyond the conservation of the components of the M phase kinase, *cdc25* has a counterpart in the *string* gene of *D. melanogaster*, and analogous proteins are found in amphibian and mammalian cells.

The basic principle established by this work is that the activity of the key regulator, Cdc2, is controlled by kinase and phosphatase activities that themselves respond to other signals. Cdc2 is the means by which

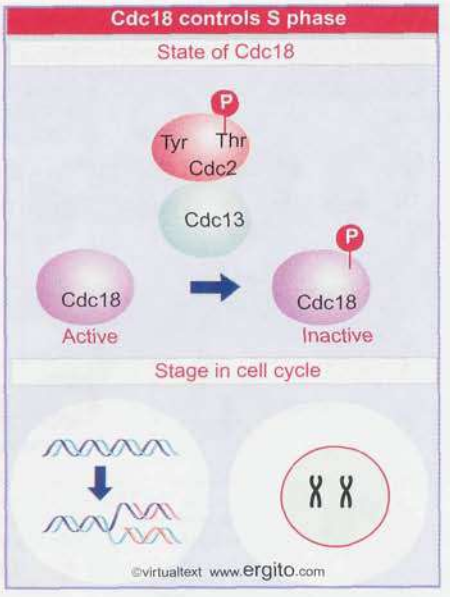


Figure 29.18 Cdc18 is required to initiate S phase and is part of a checkpoint for ordering S phase and mitosis. Active M phase kinase inactivates Cdc18, generating a reciprocal relationship between mitosis and S phase.

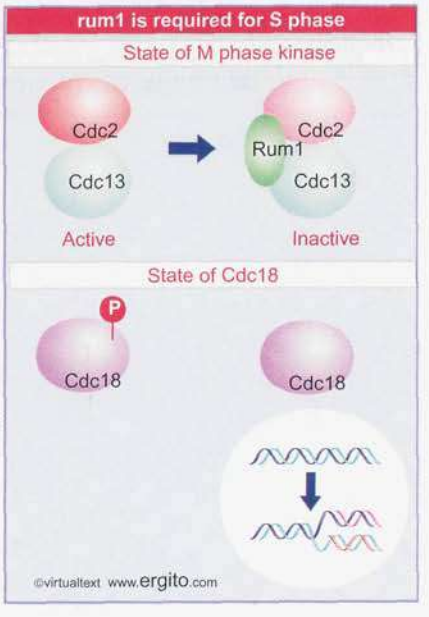


Figure 29.19 Rum1 inactivates the M phase kinase, preventing it from blocking the initiation of S phase.

all of these various signals are ultimately integrated into a decision on whether to proceed through the cycle.

Activation of the G1/S form of the kinase (Cdc2/Cig2 in *S. pombe*) is required to enter S phase, but inactivity of the M phase form (Cdc2/Cdc13) is also required. Mutants in *cdc13* fail to enter mitosis (of course), but also undergo multiple cycles of DNA replication, suggesting that the M phase kinase usually inhibits S phase. This provides a checkpoint that ensures the alternation of S phase and mitosis. Activation of the M phase kinase during G2 prevents further rounds of DNA replication occurring before mitosis; and inactivation of the M phase kinase prevents another mitosis from occurring before the next S phase has occurred.

The target protein through which this circuit functions is probably Cdc18. Transcription of *cdc18* is activated as a consequence of passing START, and Cdc 18 is required to enter S phase. Overexpression of *cdc 18* allows multiple cycles of DNA replication to occur without mitosis. If Cdc 18 is a target for Cdc2/Cdc13 M phase kinase, and is inactivated by phosphorylation, the circuit will take the form shown in **Figure 29.18.** For Cdc 18 to be active, Cdc2/Cdc13 must be inactive. When the M phase kinase is activated, it causes Cdc 18 to be inactive (possibly by phosphorylating it), and thereby prevents initiation of another S phase.

Activity of the Cdc2/Cdc13 M phase kinase is itself influenced by the factor Rum1, which controls entry into S phase. When *rum1* is overexpressed, multiple rounds of replication occur, and cells fail to enter mitosis. When *ruml* is deleted, cells enter mitosis prematurely. These properties suggest that ruml is an inhibitor of the M phase kinase. It is expressed between G1 and G2 and keeps any M phase kinase in an inactive state. (This is important during G1, before the inhibitory phosphate is added to Tyr-15; see Figure 29.14.) It also represses the level of Cdcl3 protein. The overall effect is to minimize M phase kinase activity, which is necessary to allow S phase to proceed. The consequences of the production of Rum1 on the state of M phase kinase, and consequently on the state of cdcl8, are illustrated in **Figure 29.19,** which suggests a model for the overall circuit to control S phase.

A general theme emerges from these results: the circuits that control the cell cycle have interlocking feedback loops to ensure orderly progression. And giving dual roles to a single component in which its activity is necessary to promote one event but to block another creates an intrinsic alternation of events. So an active M phase kinase simultaneously promotes mitosis as a legitimate event and inhibits S phase as an illegitimate event. This creates an intrinsic checkpoint: one event cannot be initiated until the state of the component responsible for the prior event has been reversed. By contrast, the pathway of Figure 29.23 is an example of an extrinsic checkpoint: in response to specific conditions, a pathway is activated whose end result is to alter the state of regulatory components in the cell cycle pathway.

A checkpoint is most often executed by controlling the state of phosphorylation of Cdc2. This may be achieved by activating or inhibiting the phosphatases and kinase that act on the activating or inhibitory tyrosines. Similar circuitry controls progress through meiosis as well as mitosis. For example, failure in recombination or in formation of synaptonemal complexes activates the pachytene checkpoint, which halts the cell during meiotic prophase. In *S. cerevisiae*, this is accomplished by activating the kinase Swe1 that phosphorylates the inhibitory tyrosine on Cdc28.

29.13 DNA damage triggers a checkpoint

Key Concepts

- Damage to DNA triggers a checkpoint that blocks the cell cycle until the damage has been repaired.
- The protein kinase ATM is a key component of the checkpoint pathway; it phosphorylates Chk2, which phosphorylates Cdc25, which is inactive in this form.
- The response pathway also activates genes that code for some repair proteins and directly activates other repair proteins.

Some of the most important cell cycle checkpoints are triggered by DNA damage. Cells contain many pathways to respond to DNA damage. A key aspect of the response is the control of the cell cycle. Cells are extremely sensitive to double-strand breaks in DNA—even a single break in the genome of *S. cerevisiae* can trigger a checkpoint.

Figure 29.20 illustrates the basic principle of the response to DNA damage. A complex of proteins binds to damaged DNA, and then triggers signaling pathways. The best characterized signaling pathways lead to a block in the cell cycle, which allows time for response genes to be activated and then for the damage to be repaired. An alternative response is to trigger death of the cell (thus preventing the sites of damage from inducing mutations) by apoptosis (see *29.25 Apoptosis is a property of many or all cells).*

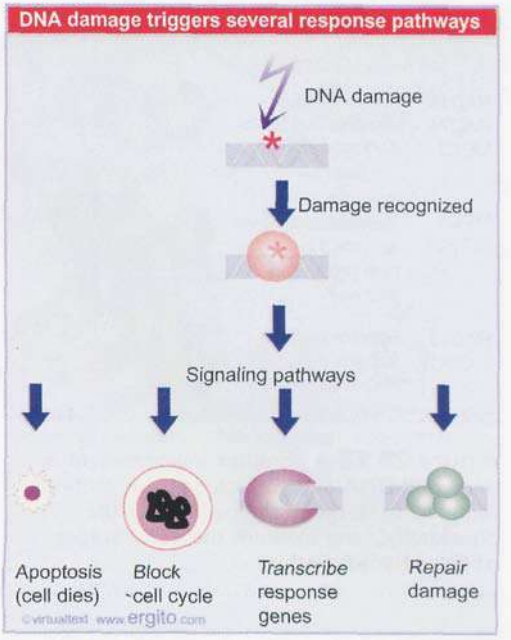


Figure 29.20 Damage to DNA triggers a response system that blocks the cell cycle, transcribes response genes, and repairs the damage, or that causes the cell to die.

Checkpoints respond to DNA damage at every stage of the cell cycle. There are also checkpoints to detect whether appropriate progress has been made through specific stages of the cycle. For example, replication must be completed before division is allowed to occur. All kinetochores must be paired before metaphase can give way to anaphase (see *29.19 Cohesins hold sister chromatids together).* **Figure 29.21** summarizes the checkpoints in *S. cerevisiae*. Note the distortion when the cycle is plotted in terms of checkpoints instead of real time. Although G1 is the longest part of the cycle, the checkpoints are concentrated in the later phases. This is because once a cell has left G1, it is committed to proceeding, but cannot be allowed to do so if the result will be to produce daughter cells with damaged DNA or other problems. So there are many checkpoints to ensure quality control.

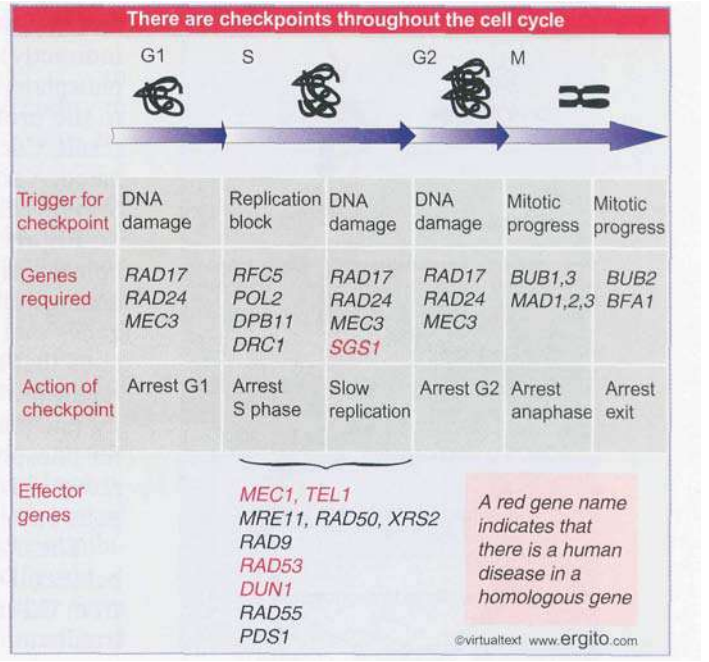


Figure 29.21 Checkpoints function at each stage of the cell cycle in *S. cerevisiae.* When damage to DNA is detected, the cycle is halted. The products of *RAD 17, RAD24, MEC3* are involved in detecting the damage. During S phase, there is a checkpoint for the completion of replication. During mitosis, there are checkpoints for progress, for example, for kinetochore pairing.

A checkpoint pathway typically involves three groups of proteins:

- *Sensor proteins* recognize the event that triggers the pathway. In the case of a checkpoint that responds to DNA damage, they bind to the damaged structure in DNA.
- *Transducer proteins* are activated by the sensor proteins. They are usually kinases that amplify the signal by phosphorylating the next group of proteins in the pathway.
- *Effector proteins* are activated by the transducer kinases. They execute the actions that are required by the particular pathway. They often include kinases whose targets are the final proteins in the pathway.

The three genes *RAD17, RAD24,* and *MEC3* code for a group of sensor proteins that are important in detecting DNA damage at each

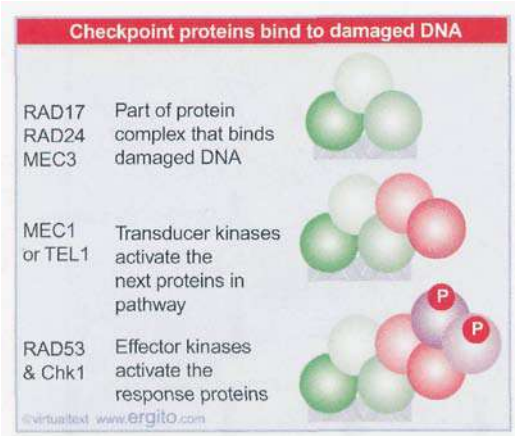


Figure 29.22 A complex assembles at a damaged DNA site. It includes the proteins required to detect damage, trigger the checkpoint, and execute the early stages of the effector pathway.

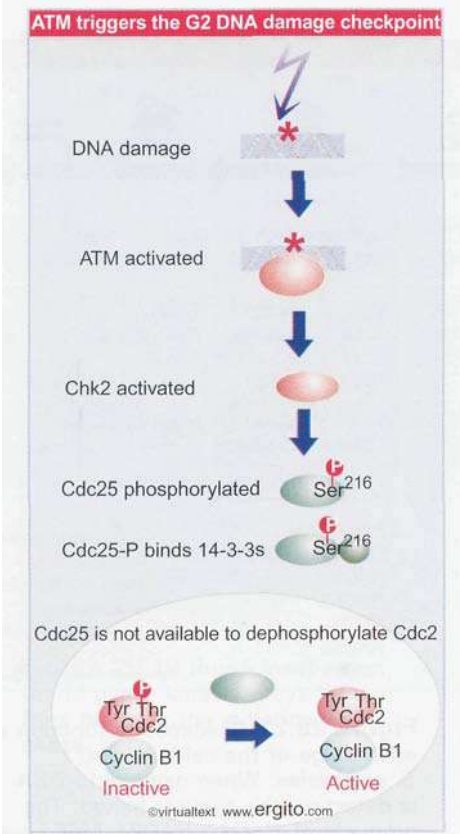


Figure 29.23 DNA damage triggers the G2 checkpoint.

stage of the cell cycle in *S. cerevisiae.* At replication, there is a second checkpoint, which also includes *SGS1,* that activates the same effector pathway. The least well characterized stage of the process is in fact the initial binding to DNA, but we believe that many of the early events in the pathway occur at a complex that assembles at the site of damage. **Figure 29.22** shows that this complex includes the three proteins RAD17, RAD24, and MEC3. They are joined by two transducer kinases, MEC1 and TEL1, that play partially redundant roles. These kinases phosphorylate and thereby activate the effector kinases RAD53 (called Chk2 in Man) and CHK1, which then phosphorylate proteins required to execute the pathway. The importance of this pathway is emphasized by the fact that many of its genes have human homologues that are implicated in genetic diseases.

One of the effectors of the response pathway in mammalian cells is the protein kinase ATM (the homologue of yeast *TEL1).* This is a central component of the DNA damage response. Loss of ATM function is responsible for the human disease ataxia telangiectasia. Patients with this disease are not only abnormally sensitive to agents that damage DNA, but also have many cellular defects that result from errors that occur during normal cell division, but which fail to be repaired because they did not induce the checkpoint. The protein kinase ATR (homologue of yeast *MEC1)* is related to ATM and also triggers this response.

ATM exists in unirradiated cells as a dimer in which its enzymatic activity is suppressed. DNA damage causes phosphorylation of a serine, which causes the dimer to dissociate. The monomer is an active kinase. The activation event may be the consequence of changes in chromatin structure that result from breaks in DNA caused by irradiation.

Figure 29.23 shows that DNA damage triggers a G2 checkpoint that involves both kinases and phosphatases. ATM activates (directly or indirectly) the kinase Chk2. Chk2 (and also the related kinase Chk1) phosphorylates Cdc25 on the residue Ser^{216}. This causes Cdc25 to bind to the protein 14-3-3σ, which maintains it in an inactive state. As a result, Cdc25 cannot dephosphorylate Cdc2, so M phase cannot be activated. A similar pathway is found in animal cells and in *S. pombe,* but in *S. cerevisiae* it is different.

Figure 29.24 shows that repair of the damaged DNA requires a combination of products of genes that are activated in response to damage and proteins that are activated by phosphorylation. Response genes that are transcribed as the result of DNA damage produce proteins that are involved directly in repairing DNA (such as the excision repair protein p48) and proteins that are needed in an ancillary capacity (such as the enzymes required for dNTP synthesis). Direct targets for phosphorylation by ATM include proteins involved in homologous recombination and the protein Nbs1 which is required for nonhomologous end-joining.

The *RAD9* mutant of *S. cerevisiae* reveals another connection between DNA and the cell cycle. Wild-type yeast cells cannot progress from G2 into M if they have damaged DNA. This can be caused by X-irradiation or by the result of replication in a mutant such as *cdc9* (DNA ligase). Mutation of *RAD9* allows these cells to divide in spite of the damage. *RAD9* therefore exercises a checkpoint that inhibits mitosis in response to the presence of damaged DNA. The reaction may be triggered by the existence of double-strand breaks. *RAD9*-dependent arrest and recovery from arrest can occur in the presence of cycloheximide, suggesting that the *RAD9* pathway functions at the post-translational level. At least 6 other genes are involved in this pathway; its ultimate regulatory target remains to be found.

29.14 The animal cell cycle is controlled by many cdk-cyclin complexes

Key Concepts

- Many dimeric kinases with a Cdc2-like catalytic subunit (cdk) are found in animal cells.
- cdk2 and cdk4 are the catalytic subunits that are usually bound to G1 cyclins.

Control of the cell cycle uses the same types of components in animal cells and in yeasts, although there is more diversity in each component in animal cells. The unifying theme is that progression to the next stage of the cell cycle is controlled by a kinase that consists of a catalytic subunit and a regulatory (cyclin) partner.

The components of the regulatory kinases for both G1/S and G2/M in yeasts and animal cells are summarized in **Figure 29.25.** The major difference is that animal cells have more variation in the subunits of the kinases. Instead of using the same catalytic subunit at both START and G2/M, animal cells use different catalytic subunits at each stage. They also have a larger number of cyclins.

At mitosis in animal cells, the Cdc2 catalytic subunit is provided by a single gene. The regulatory partner at mitosis usually is not unique, but is provided by a family of B-type cyclins, and sometimes also A-type cyclins.

During G1, animal cells have multiple kinases involved in cell cycle control, and they vary in *both* the catalytic subunit and the regulatory (cyclin) subunit. This contrasts with the retention of a common catalytic subunit in yeasts.

Just as families of cyclins can be defined by ability to interact with Cdc2, so may families of catalytic subunits be defined by the ability to interact with cyclins. Catalytic subunits that associate with cyclins are called cyclin-dependent kinases (cdks). Higher eukaryotes possess a large number of genes (~10) related to the true *cdc2* homologue. It is not entirely clear how many of these gene products are involved with the cell cycle and how many code for kinases with other functions. The cdk/cyclin dimers have the same general type of kinase activity as the Cdc2-cyclin dimers, and are often assayed in the same way, by H1 histone kinase activity. The involvement of the Cdc2/cdk kinase engines (and/or others related to them) at two regulatory points *in vivo* is consistent with an increase in H1 kinase activity at S phase as well as at M phase.

The pairwise associations between the catalytic and regulatory subunits are not exclusive, and a particular cyclin may associate with several potential catalytic subunits, while a catalytic subunit may associate with several potential cyclins. The trick is to determine which of these pairwise combinations form in the cell and are concerned with regulating the cell cycle.

Two of the *cdk* genes, *cdk2* and *cdk4,* code for proteins that form pairwise combinations with potential Gl cyclins. The first and best characterized is *cdk2;* it has 66% similarity to the *cdc2* homologue in the same organism.

G1 cyclins were originally identified as genes that could overcome the deficiency of *CLN* mutants in *S. cerevisiae.* Several new types of cyclins (including D and E) were identified by this means. They are distantly related to one another and to other cyclins. Cyclin E accumulates in a periodic manner through the cycle, but is not regulated by periodic destruction of protein. There are 3 D-type cyclins; they form dimers with cdk4 and cdk6.

Proteins described as "cyclins" are therefore now significantly more diverse than the A and B classes encompassed by the original definition.

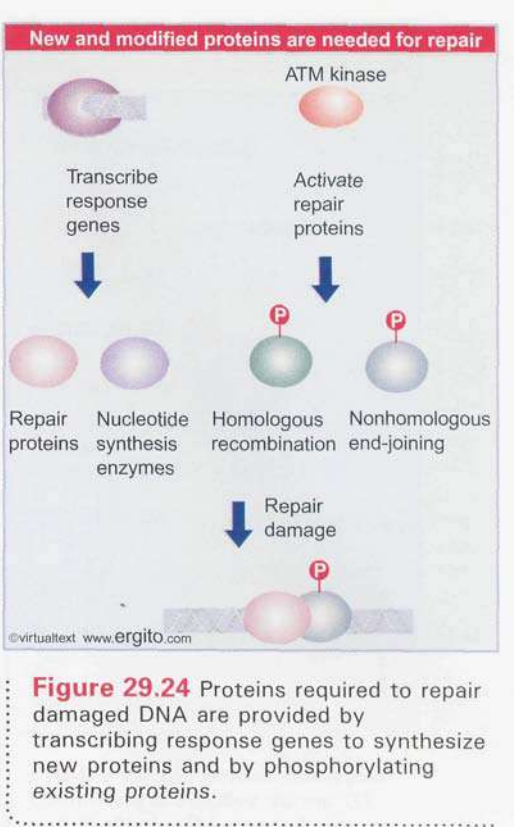


Figure 29.24 Proteins required to repair damaged DNA are provided by transcribing response genes to synthesize new proteins and by phosphorylating existing proteins.

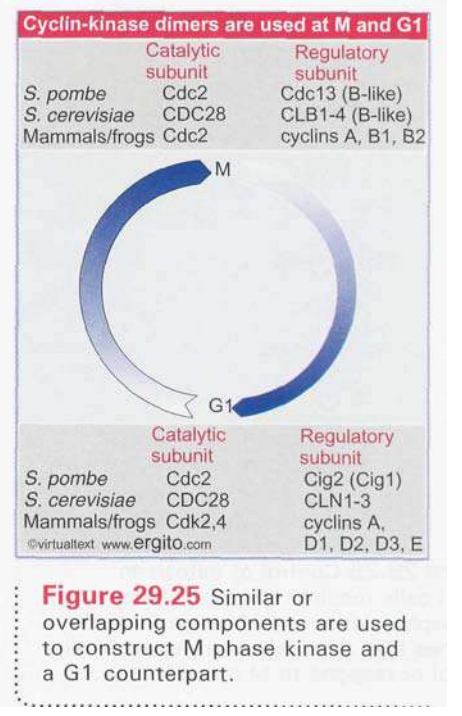


Figure 29.25 Similar or overlapping components are used to construct M phase kinase and a G1 counterpart.

For working purposes, we class as cyclins various proteins that have some sequence relationship to the original class, and which can participate in formation of a kinase by pairing with a Cdc2 or cdk2 or related catalytic subunit.

29.15 Dimers are controlled by phosphorylation of cdk subunits and by availability of cyclin subunits

The timing of activity of the various forms of cdk- and Cdc2-cyclins during the animal cell cycle suggests a model in which cdk2-Gl cyclin dimers function to regulate progression through G1 and S phase, while Cdc2-cyclinA,B dimers regulate passage through mitosis. We know most about the details of controlling the G2/M transition, but the principles are likely to be similar for G1/S, since the Cdc2 and cdk catalytic subunits conserve the residues that are involved in regulation.

Figure 29.26 shows that the regulatory events in an animal cell mitosis are similar to those in yeast cells (compare with Figure 29.17). The lower part of the figure shows the changes in M phase kinase; the upper part shows the enzymes that catalyze these changes. A cell leaves mitosis with Cdc2 monomers and no mitotic cyclins (because cyclins A and B were degraded during mitosis). Cyclins are then resynthesized. After a lag period, their level reaches a threshold at which they form dimers with Cdc2. But this does not activate the kinase activity; as we saw previously in Figure 29.9, the activity of the dimer is controlled by the state of certain Tyr and Thr residues:

- The phosphate that is necessary at Thr-161 is added by CAK (the Cdc2-activating kinase). CAK activity is probably constitutive.
- The Wee1 kinase is a counterpart to the enzyme of *S. pombe* and phosphorylates Tyr-15 to maintain the M phase kinase in inactive form.
- The Cdc25 phosphatase is a counterpart to the yeast enzyme and removes the phosphate from Tyr-15. Cdc25 is itself activated by phosphorylation; and M phase kinase can perform this phosphorylation, creating a positive feedback loop. Removal of the phosphate

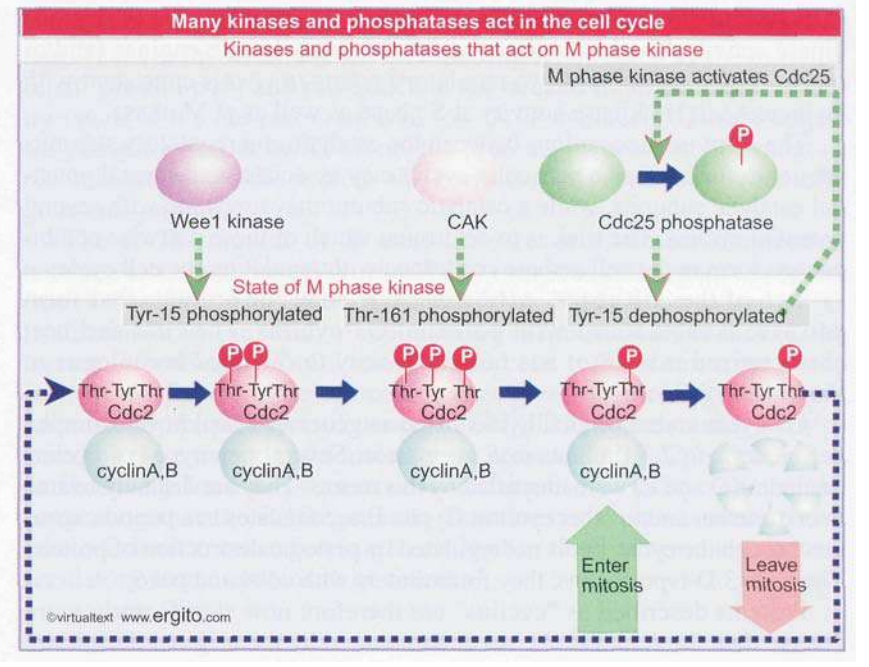


Figure 29.26 Control of mitosis in animal cells requires phosphorylations and dephosphorylations of M phase kinase by enzymes that themselves are under similar control or respond to M phase kinase.

By Book_Crazy [IND]

from Tyr-15 is the event that triggers the start of mitosis. Cdc25 is itself regulated by several pathways, including phosphatases that inactivate it, but these pathways are not yet well defined.

- Separate kinases and phosphatases that act on Thr-14 have not been identified; in some cases, the same enzyme may act on Thr-14 and Tyr-15.

This means of control is common, but not universal, since in some cells tyrosine phosphorylation does not seem to be critical, and in these cases other events must be used to control the activity of M phase kinase.

Various cdk-cyclin dimers regulate entry into S and progression through S in animal cells. Some may be concerned with entering the cycle from GO or exiting to it. The pairwise combinations of dimers that form during G1 and S are summarized in **Figure 29.27.** All of these dimers require phosphorylation on Thr-161 by CAK to generate the active form.

The synthesis of D cyclins is activated when growth factors stimulate cells to reenter the cycle from GO. The D cyclins have short half-lives, and their levels decline rapidly when the growth stimulus is removed. They may be involved with triggering reentry of quiescent cells into the cycle. Loss of D cyclins could be a trigger for a cell to leave the cell cycle for the GO state.

The activity of D cyclins is required during the latter part of G1, but not close to the G1/S boundary. Their functions may be partly redundant, but there are some differences between the D cyclins in their susceptibilities to inhibitors of the cell cycle. The significance of the ability of each D cyclin to associate with 3 different cdk subunits is not clear.

Activity of the cdk2-cyclin E complex is necessary to enter S phase. Cyclin E is synthesized during a period that spans the G1/S transition, but we do not yet know how and when it is inactivated or at what point it becomes dispensable. Cyclin E clearly has a unique role.

Progression through S phase requires the cdk2-cyclin A complex. Cyclin A is also required to associate with Cdc2 for entry into mitosis. The dual use of cyclin A in animal cells appears to be the only case in which a cyclin is used for both G1/S and G2/M transitions.

The states of cyclin-cdk complexes may influence the licensing system that prevents reinitiation of replication (see Figure 14.39). Cyclin B-cdk complexes prevent Cdc6 from loading on to the origin. This results in an orderly succession of events, because the degradation of cyclin B in mitosis releases the block and allows the procedure to start forming a prereplication complex at the origin.

Both the beginning and end of S phase are important points in the cell cycle. Just as a cell must know when it is ready to initiate replication, so it must have some means of recognizing the successful completion of replication. This may be accomplished by examining the state of DNA.

Inhibitors of DNA replication can block the cell cycle. The effect may depend on the presence of replication complexes on DNA. We do not know exactly how they trigger the checkpoint that blocks the activation of M phase kinase.

Controls that ensure an orderly progression through the cell cycle and that can be visualized in the context of checkpoints include monitoring the completion of DNA replication and checking cell size. It may be necessary to bypass some of these checkpoints in early embryogenesis; for example, early divisions in *Drosophila* in effect are nuclear only (because there are no cellular compartments: see *31 Gradients, cascades, and signaling pathways).* It may be significant that the product of *string* (the counterpart of *S. pombe cdc25* which exercises a size-dependent control) is present at high levels in these early cycles. After the 14th cycle, division becomes dependent on *string*

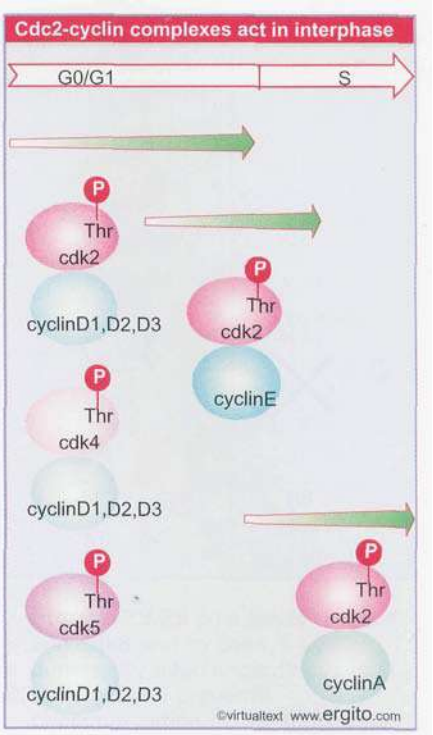


Figure 29.27 Several cdk-cyclin complexes are active during G1 and S phase. The shaded arrows show the duration of activity.

expression. It is at this point that cellular compartments develop, and it becomes appropriate for there to be a checkpoint for cell size. *string* may provide this checkpoint.

29.16 RB is a major substrate for cdk-cyclin complexes

Key Concepts

- RB is an important target for cdk-G1 cyclin complexes; its phosphorylation is required for initiation of S phase.

An important insight into control of the cell cycle at G1 has been provided by the identification of tumor suppressor genes that code for products that interact with cdk-cyclin complexes or with the downstream circuitry. Tumor suppressors are generally identified as genes in which loss of function causes tumor formation, either as seen by transformation of cells in culture, or by association of loss-of-function mutations with tumors in animals (see *30.3 Oncogenes and tumor suppressors have opposite effects).* The tumor suppressor RB is a key component in controlling the cell cycle (see also *30.19 RB is a tumor suppressor that controls the cell cycle).*

RB is a substrate for cdk-cyclin D complexes, and exerts its effects during the part of G1 that precedes the restriction point. **Figure 29.28** shows the basic circuit. In quiescent cells, or during the first part of G1, RB is bound to the transcription factor E2F. This has two effects. First, some genes whose products are essential for S phase depend upon the activity of E2F. By sequestering E2F, RB ensures that S phase cannot initiate. Second, the E2F-RB complex represses transcription of other genes. This may be the major effect in RB's ability to arrest cells in G1 phase.

RB may exert its repressive effects by interacting with chromatin. It binds histone deacetylases, which raises the possibility that it functions by causing them to remove acetyl groups from the histones at target promoters, thus inactivating the promoters (see *23.8 Deacetylases are associated with repressors).* It also interacts with components of a chromatin remodeling complex.

The nonphosphorylated form of RB forms a complex with cdk-cyclins. The complex with cdk4,6-cyclin D1, 2, 3 is the most prominent, but RB is also a substrate for cdk2-cyclin E. At or close to the restriction point, RB is phosphorylated by cdk4,6-cyclin D kinases. The phosphorylation causes RB to release E2F, which then activates transcription of the genes whose functions are required for S phase, and also releases repression of genes by the E2F-RB complex. The importance of E2F is seen by the result that expression of E2F in quiescent cells enables them to synthesize DNA.

There is an especially close relationship between RB and cyclin D1. Overexpression of D1 causes cells to enter S phase early. Inhibition of expression of D1 arrests cells before S phase. The sole role of cyclin D1 could be to inactivate RB and permit entry into S phase.

There are several related transcription factors in the E2F family, sharing the property that all recognize genes with the same consensus element. RB binds three of these factors. Two further proteins, p107 and p130, which are related to RB, behave in a similar way, and bind the other members of the E2F group. So together RB and p107 may control the activity of the E2F group of factors.

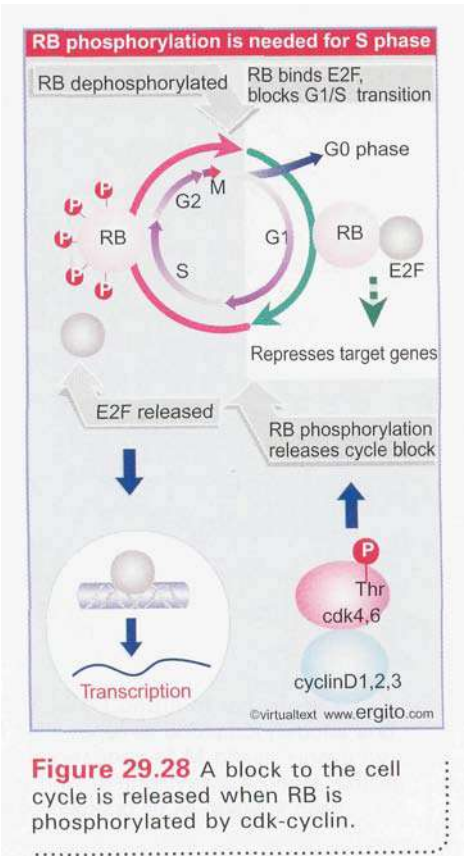


Figure 29.28 A block to the cell cycle is released when RB is phosphorylated by cdk-cyclin.

By Book_Crazy [IND]

29.17 G0/G1 and G1/S transitions involve cdk inhibitors

Key Concepts

- CKI proteins are inhibitors of cdk-cyclin dimers.
- Different CKI proteins have specificities for different cdk-cyclin complexes.

RB is a target for several pathways that inhibit growth, and may be the means by which growth inhibitory signals maintain cells in G1 (or G0). Several of these signals, including the growth inhibitory factor TGFβ , act through inhibitors of cdk-cyclin kinases. The inhibitors are called CKIs. They are found as proteins bound to cdk-cyclin dimers in inactive complexes, for example, in quiescent cells. By maintaining the cdk-cyclin complexes in inactive form, they prevent the phosphorylation of RB, making it impossible to release cells to enter S phase.

The CKI proteins fall into two classes. The INK4 family is specific for cdk4 and cdk6, and has four members: $p16^{INK4A}$, $p15^{INK4B}$, $p18^{INK4C}$, and $p19^{INK4D}$. The Kip family inhibit all G1 and S phase cdk enzymes, and have three members: $p21^{Cip1/WAF1}$, $p27^{Kip1}$, and $p57^{Kip2}$. (Each protein is identified by its size, with the casual name used as a superscript.)

INK4 protein binds specifically to cdk4 and cdk6. This suggests a connection with the G0/G1 transition. p16 cannot inhibit proliferation of cells that lack RB, which suggests that it functions by preventing cdk-cyclin kinase activity from using RB as a substrate, as illustrated in **Figure 29.29.** By binding to the cdk subunits, INK4 proteins inhibit both cdk4-cyclin D and cdk6-cyclin D activities. As exemplified by p16 and p19, they bind next to the ATP binding site of cdk6. This both inhibits catalytic activity and triggers a conformational change that prevents cyclin from binding (the conformational change is propagated to the cyclin-binding site).

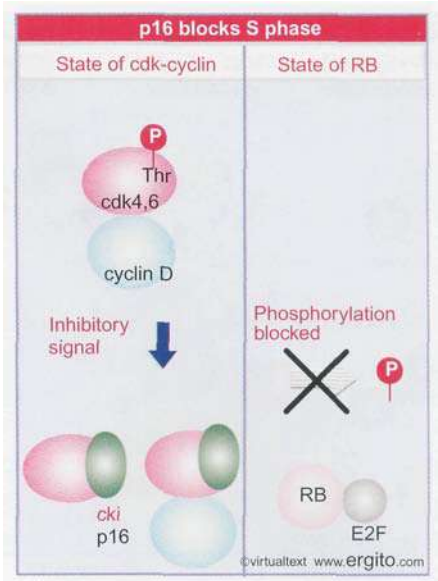


Figure 29.29 p16 binds to cdk4 and cdk6 and to cdk4,6-cyclin dimers. By inhibiting cdk-cyclin D activity, p16 prevents phosphorylation of RB and keeps E2F sequestered so that it is unable to initiate S phase.

p21 is a universal cdk inhibitor, binding to all complexes of cdk2, 4, 6. This suggests that it is likely to block progression through all stages of G1/S. In primary cultured cells (taken directly from the animal), cdk-cyclin dimers are usually found in the form of quaternary complexes that contain two further components. One is PCNA, a component of DNA polymerase δ, which may provide a connection with DNA replication. The other is the inhibitor p21. It may seem paradoxical that an inhibitor is consistently associated with the cdk-cyclin dimer, but it turns out that at a stoichiometry of 1 : 1 the p21 is not inhibitory. An increase in the number of p21 subunits associated with the cdk-cyclin dimer inhibits kinase function. In transformed cells (from lines that have been successfully perpetuated in culture), cdk-cyclin complexes lack p21 and PCNA. This suggests the possibility that p21 is involved in G1/S control, and that relaxation of this control is necessary for cells to be perpetuated in culture.

p27 has a sequence that is partly related to p21, and also binds promiscuously to cdk-cyclin complexes. Overexpression of p27 blocks progression through S phase, and levels of p27 are increased when cells are sent into a quiescent state by treatment with TGFβ. p21 and p27 block the catalytic subunit of cdk-cyclin dimers from being a substrate for activation by phosphorylation by CAK. They also prevent catalytic activity of the cdk-cyclin complex. The stages at which they function are illustrated in the summary of inhibitory pathways in **Figure 29.30.**

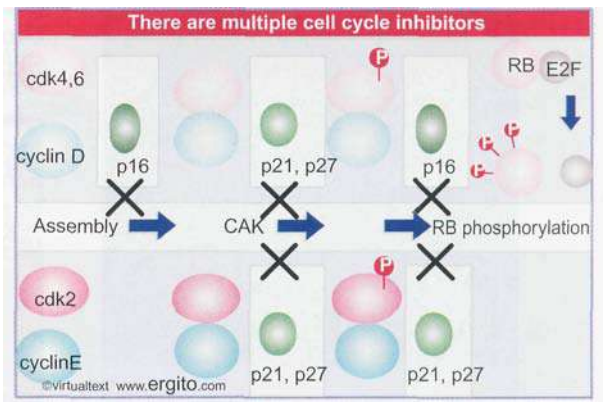


Figure 29.30 p21 and p27 inhibit assembly and activity of cdk4,6-cyclin D and cdk2-cyclin E by CAK. They also inhibit cycle progression independent of RB activity. p16 inhibits both assembly and activity of cdk4,6-cyclin D.

p21 and p27 are probably partially redundant in their functions. The pathway by which they inhibit the cell cycle is not entirely clear, but we know that it does not depend on controlling RB, because they can inhibit

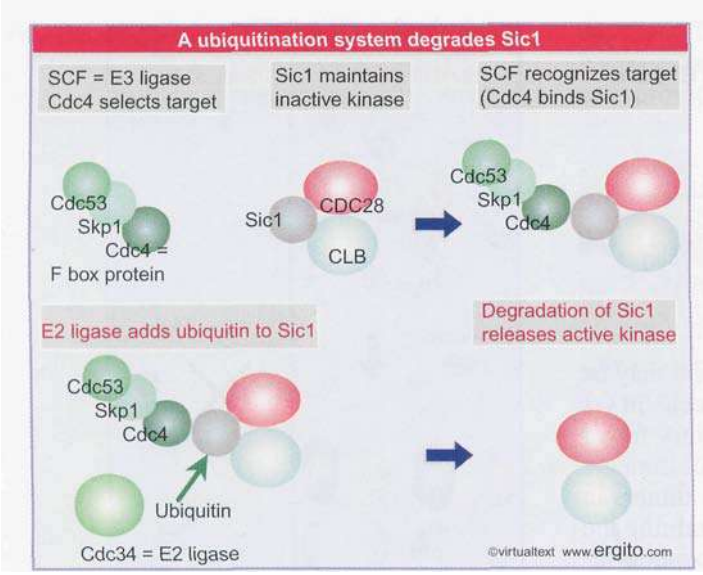


Figure 29.31 The SCF is an E3 ligase that targets the inhibitor Sid.

proliferation of cells that lack RB. This may mean that their inhibition of cdk2-cyclin E dimers is critical. Since both are present in proliferating cells, the normal progress of the cell cycle may require the levels of the cdk-cyclin dimers to increase to overcome an inhibitory threshold. p27 appears to be the major connection between extracellular mitogens and the cell cycle, with an inverse correlation between p27 activity and ability to proliferate.

The importance of the pathway from CKI proteins to RB is emphasized by the fact that tumor suppressors are found at every stage, including CKI proteins, cyclins D1, 2, and RB. The implication is that the CKI proteins are needed to suppress unrestrained growth of cells. In terms of controlling the cell cycle, this pathway is clearly central. It may be the key pathway by which cells are enabled to undertake a division cycle.

The CKI proteins are also involved in another level of control. In *S. cerevisiae*, the CKI Sic1 is bound to CDC28-CLB during G1, and this maintains the kinase in an inactive state. Entry into S phase requires degradation of Sicl to release the kinase. **Figure 29.31** shows how Sicl is targeted for degradation by a ubiquitinating system (see Figure 8.67 for a general description). The Sic1 target is recognized by a complex called the SCF, which functions as an E3 ligase (the component that selects the target). The SCF complex includes Cdc53, Skp1, and Cdc4. Cdc4 is the targeting component, which, together with Skpl, binds to Sicl. For this reason, the complex is described as SCF^{Cdc4}. Skpl is the connection to Cdc53, which interacts with the E2 ligase (Cdc34). The E2 ligase adds ubiquitin to Sic1, causing it to be degraded.

Cdc4 is a member of a class of proteins called F-box proteins. It uses the F-box motif to bind to Skp1. This is a general paradigm for the construction of SCF complexes. Other SCF complexes exist in which the targeting subunit is a different F-box protein, but the Cdc53 and Skpl components remain the same. An example relevant to the cell cycle is SCF in which the F-box protein Grr1 provides the targeting subunit, and causes the degradation of Gl cyclins.

There are further layers of control in this system. The substrates for the SCF must be phosphorylated to be recognized. The kinases that perform the phosphorylation are the cdk-cyclin complexes that are active at the appropriate stage of the cell cycle. The abundance of the SCF complexes is itself controlled by degrading the F-box subunits. SCF^{Cdc4} targets Cdc4, thus creating an autoregulatory limitation on its activity. The consequence of such feedbacks is to maintain a supply of the Cdc53-Skp1 cores that can be recruited as appropriate by the F-box subunits.

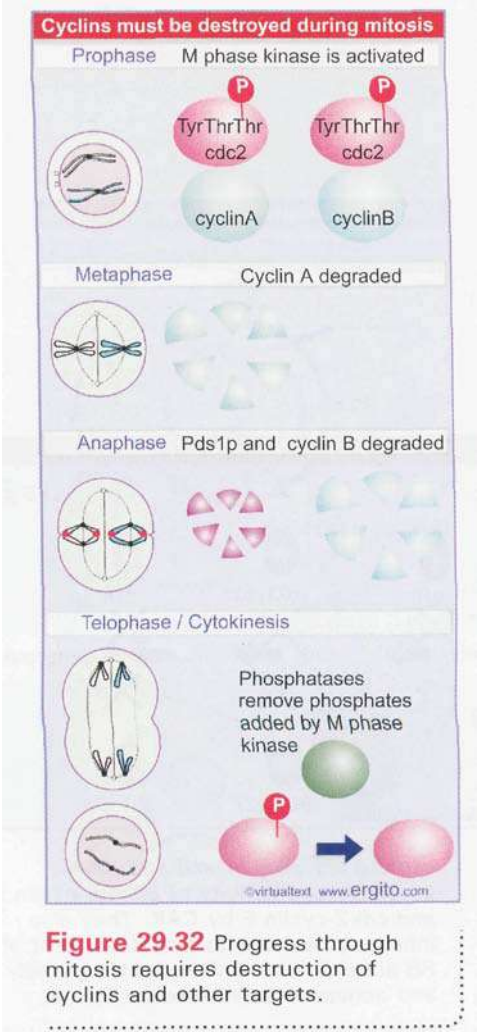


Figure 29.32 Progress through mitosis requires destruction of cyclins and other targets.

29.18 Protein degradation is important in mitosi

Key Concepts

- Cyclins are degraded at two points during mitosis, cyclin A during metaphase, and B cyclins to terminate mitosis.
- The APC (anaphase promoting complex) is an E3 ubiquitin ligase that targets a substrate for degradation.

The timing of the events that regulate mitosis is summarized in **Figure 29.32.** Mitosis is initiated by the activation of M phase kinase. Progress through mitosis requires degradation of cyclins, and also of other proteins. Several degradation events play different roles in mitotic progression:

- The first event to occur is the degradation of cyclin A at metaphase.
- The separation of chromosomes at anaphase requires the activity of the proteasomal system for protein degradation (but does not require cyclin degradation). The target for this event is the protein Pds1p, whose degradation triggers the pathway that enables sister chromatids to separate.
- Later during anaphase, B cyclins must be degraded in order to inactivate M phase kinase. This is necessary to allow the phosphorylations of substrate proteins that were catalyzed by M phase kinase to be reversed at the end of mitosis.

A large complex of 8 subunits is responsible for selecting the substrates that are degraded in anaphase. It is called the anaphase promoting complex (APC) (sometimes also known as the cyclosome). Similar complexes are found in both yeasts and vertebrates. The APC becomes active specifically during mitosis. It functions as an E3 ubiquitin ligase (see Figure 8.67), which is responsible for binding to the substrate protein so that ubiquitin is transferred to it. The ubiquitinated substrate is then degraded by the proteasome.

There are two separate routes to activating the APC, and each causes it to target a particular substrate. **Figure 29.33** shows that the regulatory factors CDC20 and Cdh1 bind to the APC and activate its ubiquitination activity. CDC20 is necessary for the APC to degrade Pdslp, which controls the transition from metaphase to anaphase. Cdhl is necessary for the APC to degrade Clb2 (yeast cyclin B), which is necessary to exit mitosis. The timing of activation and persistence of each type of complex is different.

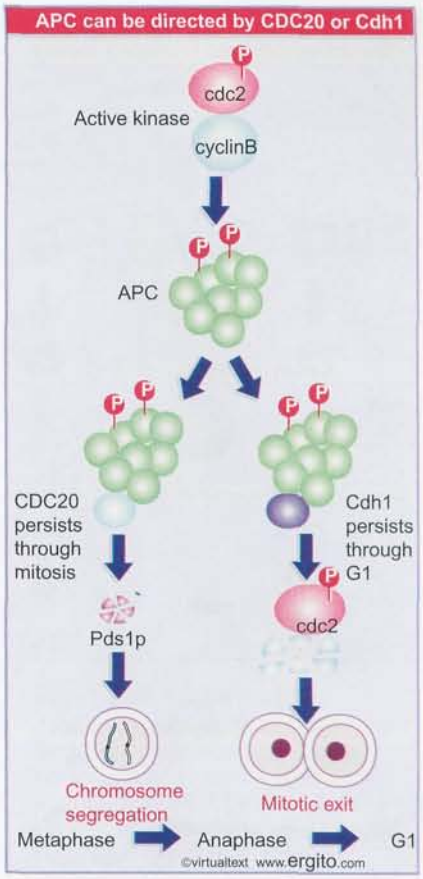


Figure 29.33 Two versions of the APC are required to pass through anaphase.

29.19 Cohesins hold sister chromatids together

Key Concepts

- Cohesins associate with chromatin at S phase and hold sister chromatids together.
- A securin is a protein that sequesters a separin and keeps it in an inactive state.
- Anaphase in yeast is triggered when CDC20 causes the APC to degrade the securin Pds1p.
- Degradation of the securin releases the separin (Esp1 in yeast) which is a protease that cleaves the cohesin Scc1p.
- Cleavage of Scc1p releases the sister chromatids.

After replication occurs at S phase, the products of each chromosome (sister chromatids) remain associated with one another, although this becomes evident only later at the beginning of mitosis. This is a crucial aspect of segregating the sister chromatids to different daughter cells, as shown in Figure 19.22. The sister chromatids are held together by a complex of cohesin proteins, which functions as a sort of "glue." The cohesins are the key players in assuring chromatid association at the start of mitosis, and dissociation during mitosis.

Figure 29.34 shows that cohesins are located at various sites along a pair of sister chromatids, with the appearance of being centrally localized between the chromatids. At metaphase, each sister chromatid pair is in equilibrium on the spindle, being connected to both poles by microtubules. When the connecting glue dissolves, the sister chromatids are released from one another and pulled toward opposite poles. This marks the transition from metaphase to anaphase.

At the metaphase-anaphase transition, the route to releasing sister chromatids so that they may segregate to opposite poles is indirect, as illustrated in **Figure 29.35.** There are three components in the system.

Figure 29.34 Cohesins are located sporadically along a pair of sister chromatids early in mitosis. DNA is in red; cohesins are in yellow. Photograph kindly provided by Ana Losada and Tatsuya Hirano.

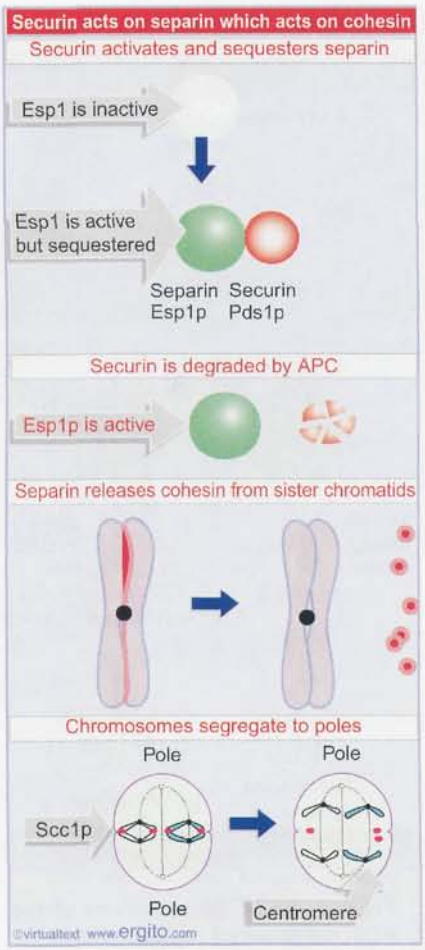


Figure 29.35 Anaphase progression requires the APC to degrade Pds1p to allow Esp1p to remove Scc1 from sister chromatids.

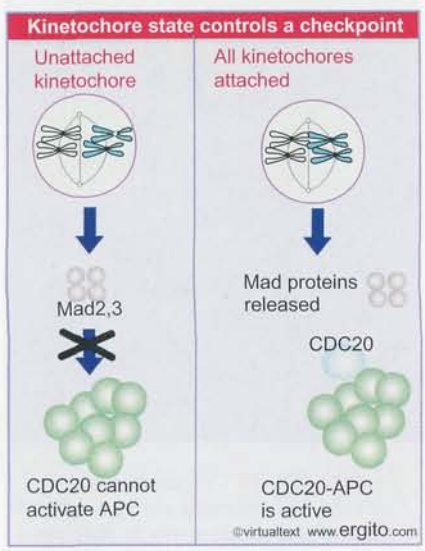


Figure 29.36 An unattached kinetochore causes the Mad pathway to inhibit CDC20 activity.

The cohesins connect the sister chromatids. A separin releases the cohesins. And the separin is controlled by a securin.

Prior to anaphase in yeast, the securin Pds1p binds to the separin Esp1p. When the APC degrades Pds1p (securin), Esp1p (separin) is released. The separin is a protease. It cleaves the protein Scc1p, which is a component of the cohesin complex. When Scc1p is released by the action of Esp1p, the cohesin complex can no longer hold the chromatids together, and they therefore become free to segregate on the spindle.

If securin simply sequesters separin, we would expect the loss of securin to cause premature chromosome separation. However, the opposite happens. If the securin gene is inactivated, the chromosomes have difficulty in separating, and the delay causes abnormalities in chromosome segregation. This suggests that securin plays two roles. Suppose that the separin is in an inactive state before it binds to securin. The securin activates it, but keeps it sequestered. When securin is degraded, the separin is released in the active state. If securin is absent altogether, the separin never becomes activated, so that the cohesins are not destroyed.

Scc1p is only one component of the cohesin complex. The core of the complex is a heterodimer of SMC proteins. (SMC stands for structural maintenance of chromosome; other SMC proteins form the condensins, which are involved in controlling chromosome condensation; see *23.18 Chromosome condensation is caused by condensins.)* The cohesin complex contains Scc1p, Scc3p, and a heterodimer of the two SMC proteins Smc1p and Smc3p. However, loss of Scc1p is sufficient to abolish the ability to hold sister chromatids together. The cohesins may function by cross-linking DNAs as shown in Figure 23.32.

Smc3p is involved in meiosis as well as mitosis. It is required for sister chromatid cohesion, together with a protein (Rec8p) that is related to Scc1p. This suggests that a cohesin complex, related to that of mitosis, may form at meiosis. Both of these components are required for synaptonemal complexes to form (see *15.5 Recombining chromosomes are connected by the synaptonemal complex).* The metaphase-anaphase transition at the first meiotic division in yeast is triggered in the same way as at mitosis, when the cohesin Rec8 is cleaved by separin to allow disjunction of homologous chromosomes.

The situation in higher eukaryotic cells is more complex. The cohesin Scc1p is mostly released from the chromosomes during prophase, although it is left in the centromeric regions. It is degraded at the start of anaphase, and its loss from the centromeric regions may be the trigger for chromosome separation.

The major target of the APC^{Cdh1} complex is the mitotic cyclin component of the cdk-cyclin kinase. Its destruction makes it possible to reverse the phosphorylations that triggered mitosis, which is necessary to exit mitosis. However, there is some overlap in the actions of the two APC complexes, and APC^{CDC20} may also act on mitotic targets. The relative timing of activation of the two APC complexes may be determined by a circuit in which a target of APC^{CDC20} is needed to activate APC^{Cdh1}.

Formation of the cohesin complex (or possibly its association with the chromosomes) occurs during S phase. Mutants in the locus *ctf7* have sister chromatids that never associate. The gene acts during the period of DNA replication, but does not code for a component of the cohesin complex. It is possible that the establishment of cohesion is triggered by replication in some way. So far as we can tell from examining individual sequences, the replicated copies remain tightly associated from replication until cell division.

The cohesion system is the target for a checkpoint. Progression through mitosis requires all kinetochores to be paired with their homologues. **Figure 29.36** shows that the checkpoint consists of a surveillance system that is triggered by the presence of an unattached kinetochore. Mutations in the *Mad* and *Bub* genes allow mitosis to continue (aberrantly) in the pres-

ence of unpaired kinetochores. The Mad proteins control the system for chromatid segregation. They bind to CDC20 and prevent it from activating the APC. When the kinetochores are all attached, some (unidentified) signal causes Mad proteins to be released from CDC20, which now activates the APC, allowing anaphase to continue. A similar role is played by Bub proteins in controlling the ability of Cdh1 to activate the APC.

29.20 Exit from mitosis is controlled by the location of Cdc14

Key Concepts

- During interphase the phosphatase Cdc14 is held in the nucleolus.
- When a spindle pole body migrates into the bud of *S. cerevisiae*, it carries the protein Tem1.
- Tem1 is a monomeric G protein that is activated by the local concentration of the exchange factor Lte1 in the bud.
- Activation of Tem1 triggers release of Cdc14.
- The action of Cdc14 triggers exit from mitosis.

The key event in leaving mitosis is the activation of the phosphatase Cdc14. **Figure 29.37** shows that during interphase, Cdc14 is sequestered in the nucleolus (because it binds to a nucleolar protein variously called RENT/Cfi1/Net1). When it is localized in the nucleolus, it cannot find any of its substrates, and therefore is inactive. When it is released from the nucleolus, it acts on (at least) two substrates. Dephosphorylation of Cdhl is necessary to activate the APC^{Cdh1} complex. And dephosphorylation of Sic1 enables it to inactivate mitotic cyclins. This is another example of belts and braces in cell cycle control, where parallel pathways lead to the same outcome.

What triggers the release of Cdc14 from the nucleolus? The pathway for leaving mitosis has many genes, and genetic relationships suggest that two key components are the GTP-binding protein Tem1 and the exchange factor Lte1. Like other monomeric G proteins, Teml is active when bound to GTP, and inactive when bound to GDP. The exchange factor activates it by causing bound GDP to be replaced with GTP. The ability of Ltel to activate Teml is controlled in an interesting way by the locations of the two components in the yeast cell.

Recall that *S. cerevisiae* has an asymmetrical division in which the daughter cell forms as a bud of the mother cell (see Figure 29.10). Ltel protein is present throughout the cell cycle, and when the bud forms at the beginning of S phase, the Lte1 is localized in it. By contrast, Tem1 is synthesized only at late S phase. At mitosis, the Teml is localized with one of the spindle pole bodies (the structures identifying the ends of the spindle where microtubules are nucleated). **Figure 29.38** shows that this is the spindle pole body that migrates into the bud! When Teml arrives in the concentration of Lte1 in the bud, it is activated. This triggers the release of Cdc14 from the nucleolus, which in turn triggers exit from mitosis.

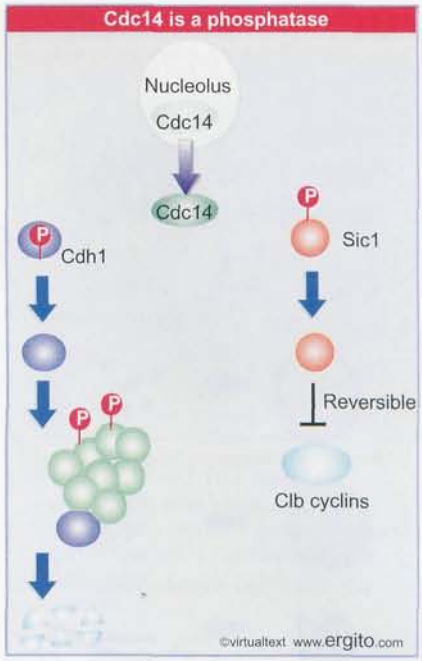


Figure 29.37 Cdc14 dephosphorylates both Cdh1 and Sic1. The first action leads to activation of the APC that degrades mitotic cyclins. The second action enables Sic1 to reversibly inactivate mitotic cyclins.

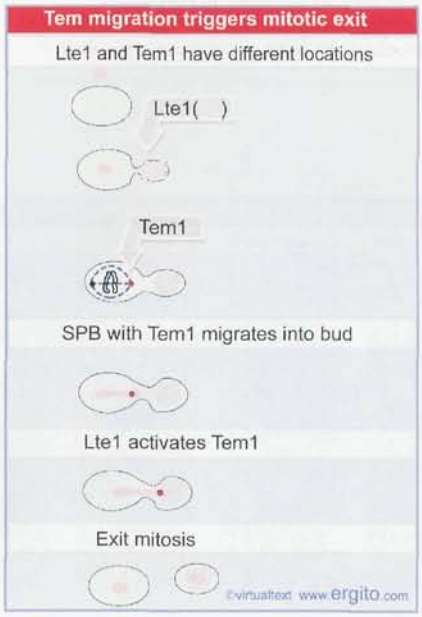


Figure 29.38 Exit from mitosis is triggered when the Tem1 that is localized on a spindle poly body migrates into the bud where Lte1 is localized.

29.21 The cell forms a spindle at mitosis

Key Concepts

- The activity of M phase kinase is responsible for condensation of chromatin into chromosomes, dissolution of the nuclear lamina and envelope breakdown, reorganization of actin filaments, and (by unknown means) reorganization of microtubules into the spindle.

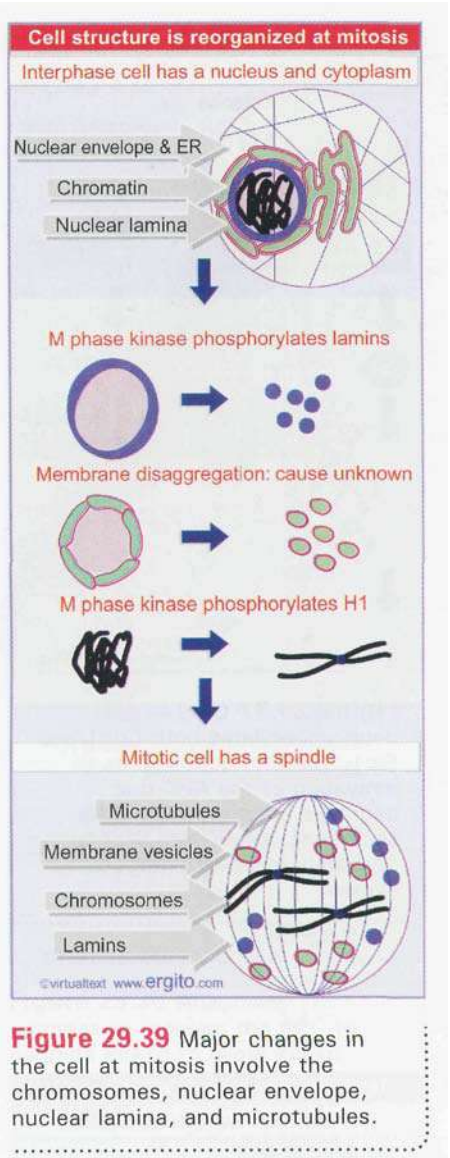


Figure 29.39 Major changes in the cell at mitosis involve the chromosomes, nuclear envelope, nuclear lamina, and microtubules.

The culmination of the cell cycle is the act of division, when the chromosomes segregate into two diploid sets and the other components of the cell are partitioned between the two daughter cells. The change in cell structure is dramatic, as summarized in Figure 29.39. The division between nucleus and cytoplasm is abolished, and the cytoskeleton is entirely reorganized. The relevant events include

- Condensation of chromatin to give recognizable chromosomes.
- Dissolution of the nuclear lamina and breakdown of the nuclear envelope. The lamina dissociates into individual lamin subunits, and the nuclear envelope, endoplasmic reticulum, and Golgi apparatus break down into small membrane vesicles. (Nuclear dissolution is typical of animal cells but does not occur in some lower eukaryotes, including yeasts, where mitosis involves nuclear division.)
- Dissociation and reconstruction of microtubules into a spindle. Microtubules dissociate into tubulin dimers, which reassemble into microtubules extended from the mitotic microtubule organizing centers.
- Reorganization of actin filaments to replace the usual network by the contractile ring that pinches the daughter cells apart at cytokinesis.

All of these changes are reversible; following the separation of daughter cells, the actin filaments resume their normal form, the microtubular spindle is dissolved, the nuclear envelope reforms, and chromosomes take a more dispersed structure in the form of interphase chromatin. Modification of appropriate substrate proteins (which could be either the structural subunits themselves or proteins associated with them) provides a plausible means to control the passage of mitosis. The question then becomes how the mitotic changes and their reversal depend upon the activation and inactivation of M phase kinase.

The best characterized substrate for M phase kinase is histone H1. As noted previously, we do not know what role the phosphorylation of H1 plays at either G1/S or M phase transitions. It is a reasonable assumption, however, that M phase kinase acts directly on chromatin by phosphorylating H1 and other target proteins, and that this is the cause of chromosomal condensation.

Two types of event have been implicated in nuclear envelope breakdown (sometimes abbreviated to NEBD). They affect the two components of the envelope: the membrane and the underlying lamina.

Mechanical effects are caused by the growth of microtubules into the envelope. This occurs when the centrosomes (the microtubule nucleating centers) move into the envelope at the start of mitosis. The motor dynein pulls the envelope along the microtubules. The stress causes the envelope to tear, and may also disrupt the structure of the lamina.

Nuclear integrity is abandoned when the lamina dissociates into its constituent lamins. The lamina is a dense network of fibers just underneath the inner nuclear membrane. It is responsible for maintaining the shape of the nucleus. The components of the lamina are three types of lamins, each of which has a domain structure similar to that of the protein subunits of the intermediate filaments that are found in the cytoplasm. The importance of lamins is emphasized by the existence of human diseases resulting from mutations in lamin genes, although we cannot yet relate the phenotypes of the diseases to the molecular functions of the lamins.

Lamins are phosphorylated at the start of mitosis, and the presence of phosphate groups on only two serine residues per lamin is sufficient to cause dissociation of the lamina. Mutations that change these serines into alanines prevent the lamina from dissociating at mitosis. So the reversible phosphorylation of these two serine residues induces a structural change in the individual lamin subunits that controls their ability to associate into the lamina.

The combined effects of tearing the membrane and dissociating the lamina generate a series of vesicles containing the membranes. The process is reversed at the end of mitosis, when lamins are closely involved in reassembly of the nuclear membrane.

Nuclear pore complexes dissociate from the envelope during mitosis. Nuclear pore components are phosphorylated, and as a result are released from the pores and dispersed in the cell.

These phosphorylation events are the direct responsibility of the M phase kinase, which can cause the nuclear lamina to dissociate *in vitro*. The reorganization of the endoplasmic reticulum and Golgi is not well defined.

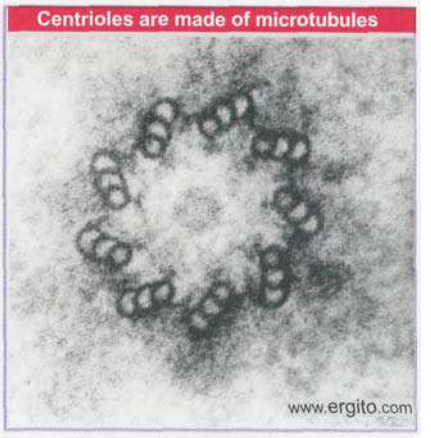


Figure 29.40 The centriole consists of nine microtubule triplets, apparent in cross section as the wall of a hollow cylinder. Photograph kindly provided by A. Ross.

29.22 The spindle is oriented by centrosomes

Key Concepts

- Microtubule assembly is nucleated by complexes located within the centrosomes that contain γ-tubulin.
- The centrioles duplicate by a mechanism involving orientation of the daughter relative to the parent.

The reorganization of microtubules into the spindle has been extensively described, but cannot yet be connected to the action of M phase kinase. Some microtubules extend from pole to pole, others connect the chromosomes to the poles. The ends of each microtubules are attached to MTOCs—microtubule organizing centers. The MTOCs at the poles lie in the regions of the centrosomes. The MTOC on a chromosome is located at the kinetochore.

The structure of centrosomes is not well defined, but in animal cells a centrosome contains a pair of centrioles, surrounded by a dense amorphous region. The centriole is a small hollow cylinder whose wall consists of a series of triplet fused microtubules. A centriole is shown in cross section in **Figure 29.40.**

The function of the centriole in mitosis is not clear. Originally it was thought that it might provide the structure to which microtubules are anchored at the pole, but the fibers seem instead to terminate in the amorphous region (the pericentriolar material) around the centrioles. It is possible that the centriole is concerned with orienting the spindle; it may also have a role in establishing directionality for cell movement. However, there are cell types in which centrosomes do not appear to contain centrioles.

Centrioles have their own cycle of duplication. When born at mitosis, a cell inherits two centrioles. During interphase they reproduce, so that at the start of mitosis there are four centrioles, two at each pole. Probably only the parental centriole is functional.

The centriole cycle is illustrated in **Figure 29.41.** Soon after mitosis, a procentriole is elaborated perpendicular to the parental centriole. It has the same structure as the mature parental centriole, but is only about half its length. Later during interphase, it is extended to full length. It plays no role in the next mitosis, but becomes a parental centriole when it is distributed to one of the daughter cells. The orientation of the parental centriole at the mitotic pole is responsible for establishing the direction of the spindle.

How are centrioles reproduced? The precise elaboration of the procentriole adjacent to the parental centriole suggests that some sort of template function is involved. The parental centriole cannot itself be seen to

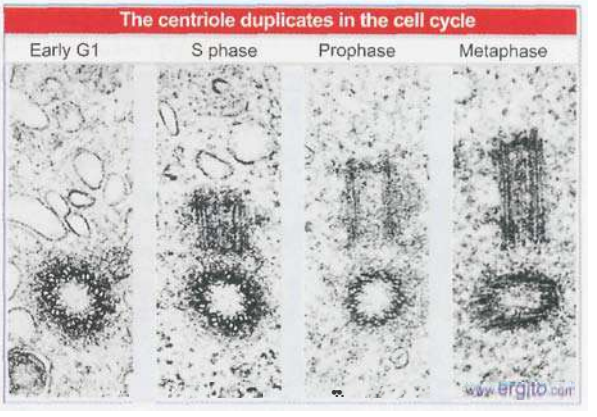


Figure 29.41 A centriole reproduces by forming a procentriole on a perpendicular axis; the procentriole is subsequently extended into a mature centriole. Photograph kindly provided by J. B. Rattner and S. G. Phillips.

reproduce or divide, but it could provide some nucleating structure onto which tubulin dimers assemble to extend the procentriole. Could a centriole be assembled in the absence of a pre-existing centriole?

Microtubules consist of hollow filaments made of 13 protofilaments that are constructed from dimers of α-tubulin and β-tubulin. Within the centrosome there is a related protein, γ-tubulin, which is part of a complex that provides the actual nucleating source for the microtubules. The complex is large, sedimenting at ~25S, and contains several other proteins in addition to γ-tubulin. The complex can nucleate the formation of microtubules from α-tubulin and β-tubulin *in vitro*. The complex takes the form of a ring, and probably the γ-tubulin-containing complex nucleates microtubules through some sort of end-binding mechanism.

The spindle is generally nucleated by the centrosomes, although cells somes. In addition to its mechanical role in cell reorganization, a centrosome is a regulatory target. Centrosome duplication is regulated during the cell cycle, and there is a checkpoint to stop the cycle proceeding until centrosome duplication has occurred. The components of the centrosome involved in nucleating microtubules are beginning to be defined, but the components involved in regulation mostly remain to be described.

29.23 A monomeric G protein controls spindle assembly

Key Concepts

- The active form of the G protein Ran (Ran-GTP) causes importin dimers to release proteins that trigger microtubule nucleation.
- The Ran-activating protein RCC is localized on chromosomes, generating a high local concentration of Ran-GTP.
- The proteins released by the importins have several different functions that assist microtubule nucleation.

The trigger for the reorganization of microtubules from the interphase network into the spindle may be the breakdown of the nuclear envelope, which exposes nuclear components to cytoplasmic components. Indirect evidence has been available for some time to indicate a connection, but only recently has a molecular mechanism been suggested. The important point here is that the ability of an MTOC to nucleate microtubules must be *controlled,* so that it happens only in the right time and place.

The critical component is a monomeric G protein called Ran, which controls the direction of protein transport through the nuclear envelope. Like all members of its class, Ran is active when bound to GTP, and inactive when bound to GDP. Conditions in the nucleus and cytosol differ so that typically there is Ran-GTP in the nucleus, but there is Ran-GDP in the cytosol. Protein export complexes are stable in the presence of Ran-GTP, whereas import complexes are stable in the presence of Ran-GDP. So export complexes are driven to form in the nucleus and dissociate in the cytosol, whereas the reverse is true of import complexes (see *8.28 Transport receptors carry cargo proteins through the pore)*.

Mutations in some proteins that bind to Ran cause the spindle to malfunction, and overexpression of the protein RanBPM (another Ran-binding protein) causes the formation of ectopic asters—arrays of microtubules emanating from centrosomes. The usual assay for these

experiments is to inject demembranated sperm into *Xenopus* eggs. The sperm centrioles assemble into centrosomes that nucleate microtubule asters. Using this assay identifies proteins that can stimulate nucleation. These include a mutant of Ran and the protein RCC that maintains Ran in the GTP-bound active state. The most likely explanation is that the breakdown of the nuclear envelope releases Ran-GTP, which then triggers microtubule nucleation by centrosomes.

Does Ran act directly or indirectly? One of the targets for Ran in the nuclear transport process is the import receptor importin-β, which (in combination with importin-α) transports cargo proteins from the cytoplasm to the nucleus. It turns out that the importin dimer binds to proteins that affect microtubules. One of these proteins is Xklp2, which connects a motor (a protein that moves other proteins) to microtubules at the poles; another is NuMA which cross links microtubules at the poles during mitosis. When the complex of importins with either of these proteins is exposed to Ran-GTP, it dissociates, releasing the cargo protein, which can then act to trigger microtubule nucleation.

How does the exposure of the importins to Ran-GTP change at mitosis? **Figure 29.42** shows that the situation in the cytoplasm of the interphase cell, and then correspondingly in the spindle, is that Ran is predominantly in the form of Ran-GDP, and therefore does not affect the importin complex. But there are localized areas where Ran-GTP is formed. The Ran-activating protein RCC is located on chromatin, so Ran-GTP forms in the vicinity of the chromosomes. This releases the proteins that are bound to the importin dimer, which activate the kinetochores to connect to microtubules.

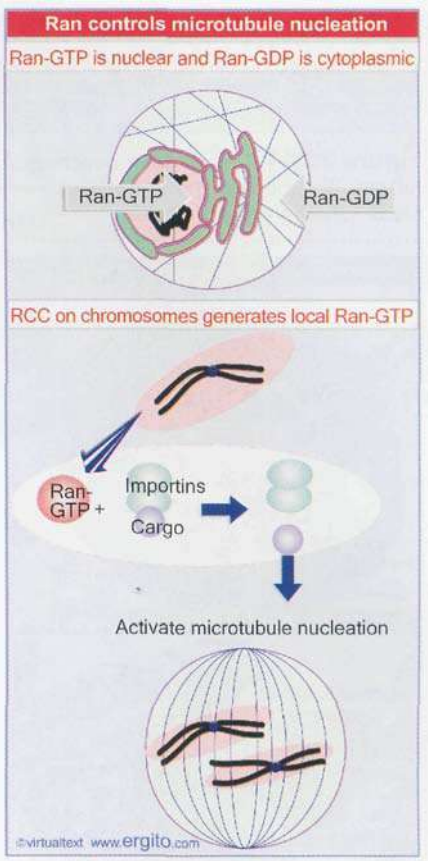


Figure 29.42 The Ran-activating protein RCC is localized on chromosomes. Ran-GTP is high in the nucleus in the interphase cell. When the nuclear envelope breaks down, RCC maintains a high level of Ran-GTP in the vicinity of the chromosomes. This causes importin dimers to release the proteins bound to them. These proteins cause microtubule nucleation.

29.24 Daughter cells are separated by cytokinesis

Once the two sets of daughter chromosomes have been separated at the poles, the cell must complete its division by physically separating into two parts. This process is called cytokinesis, and it passes through the stages illustrated in **Figure 29.43.**

The plane for division forms in the center of the spindle. We do not know exactly how its position is defined, but it seems to depend on the microtubules arrays that run to the poles. A local event that may be needed is the activation of the RhoA monomeric G protein. (This is one of the monomeric G proteins that controls actin filament behavior in the interphase cell; see Figure 28.35 in *28.15 The activation of Ras is controlled by GTP.)*

An invagination called the cleavage furrow appears in the plasma membrane soon after the start of anaphase. This is caused by the formation of the contractile ring, which forms from actomyosin fibers. It extends around the equator of the dividing cell and then pinches inward until it contacts a group of microtubules that run between the poles. This forms a structure connecting the future daughter cells that is called the midbody.

The final step in cytokinesis is to cut the cytoplasmic connection between the two cells by "resolving" the midbody. This requires changes in the organization of the plasma membrane, but we cannot yet account for these events at the molecular level.

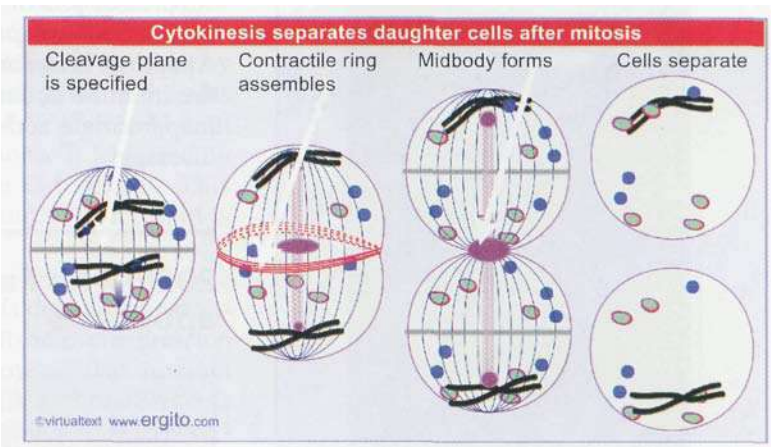


Figure 29.43 The spindle specifies the cleavage plane where the contractile ring assembles, the midbody forms in the center, and then the daughter cells separate.

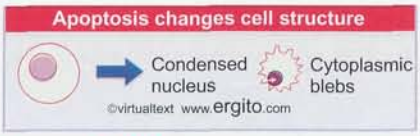


Figure 29.44 The nucleus becomes heteropycnotic and the cytoplasm blebs when a cell apoptoses.

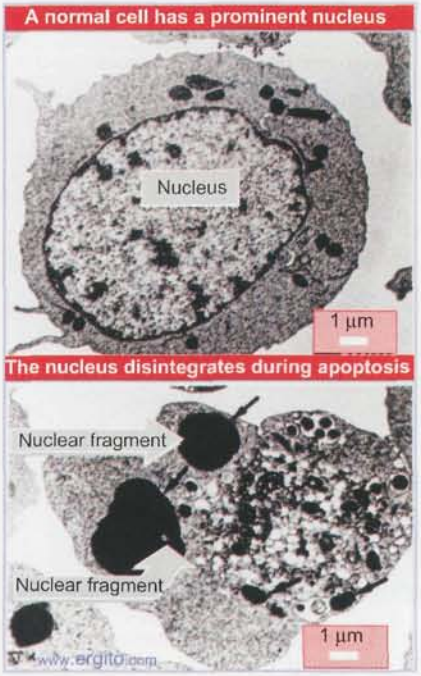


Figure 29.45 Cell structure changes during apoptosis. The top panel shows a normal cell. The lower panel shows an apoptosing cell. Photographs kindly provided by Shigekazu Nagata.

Figure 29.46 Fragmentation of DNA occurs ~2 hours after apoptosis is initiated in cells in culture. Photograph kindly provided by Shigekazu Nagata.

29.25 Apoptosis is a property of many or all cells

Key Concepts

- All cells possess the pathways that can cause death by apoptosis, which requires RNA and protein synthesis by the dying cell, but the pathway is activated only by appropriate stimuli.

During development of a multicellular eukaryotic organism, *some cells must die*. Unwanted cells are eliminated during embryogenesis, metamorphosis, and tissue turnover. This process is called programmed cell death or apoptosis. It provides a crucial control over the total cell number. In the worm *C. elegans* (in which somatic cell lineages have been completely defined), 131 of the 1090 adult somatic cells undergo programmed cell death—cells die predictably at a defined time and place in each animal. Similar, although less precisely defined, cell deaths occur during vertebrate development, most prominently in the immune system and nervous system. The proper control of apoptosis is crucial in probably all higher eukaryotes.

Apoptosis involves the activation of a pathway that leads to suicide of the cell by a characteristic process in which the cell becomes more compact, blebbing occurs at the membranes, chromatin becomes condensed, and DNA is fragmented (see **Figure 29.44).** *The pathway is an active process that depends on RNA and protein synthesis by the dying cell.* The typical features of a cell as it becomes heteropycnotic (condensed with a small, fragmented nucleus) are shown in **Figure** 29.45, and the course of fragmentation of DNA is shown in Figure 29.46. Ultimately the dead cells become fragmented into membrane-bound pieces, and may be engulfed by surrounding cells.

Apoptosis can be triggered by a variety of stimuli, including withdrawal of essential growth factors, treatment with glucocorticoids, 7-irradiation, and activation of certain receptors, as summarized in **Figure 29.47.** These all involve a molecular insult to the cell. Another means of initiating apoptosis is used in the immune system, where cytotoxic T lymphocytes attack target cells. Apoptosis is also an important mechanism for removing tumorigenic cells; the ability of the tumor suppressor p53 to trigger apoptosis is a key defense against cancer (see *30.20 Tumor suppressor p53 suppresses growth or triggers apoptosis).* Apoptosis is important, therefore, not only in tissue development, but in the immune defense and in the elimination of cancerous cells. Also, inappropriate activation of apoptosis is involved in neurodegenerative diseases.

29.26 The Fas receptor is a major trigger for apoptosis

Key Concepts

- The Fas receptor on a target cell is activated by interaction with the FasL protein on an activating cell plasma membrane.
- Fas is related to TNF receptor, and FasL is related to TNF.
- Fas is a trimer that aggregates upon interaction with FasL.
- Fas has a cytoplasmic domain called the "death domain" which is involved in protein-protein interactions.

The Fas receptor (called Fas or FasR) and Fas ligand (FasL) are a pair of plasma membrane proteins whose interaction triggers one of the major pathways for apoptosis. **Figure 29.48** shows that the cell bearing the Fas receptor apoptoses when it interacts with the cell carrying the Fas ligand.

Activation of Fas resembles other receptors in involving an aggregation step. However, **Figure 29.49** shows that there are some interesting differences from the growth receptor model. First, Fas forms a homomeric trimer. Second, the trimer assembles *before* the interaction with ligand. The effect of ligand may be to cause the trimers to cluster into large aggregates. At all events, when FasL interacts with Fas, there is an aggregation event that enables Fas to activate the next stage in the pathway.

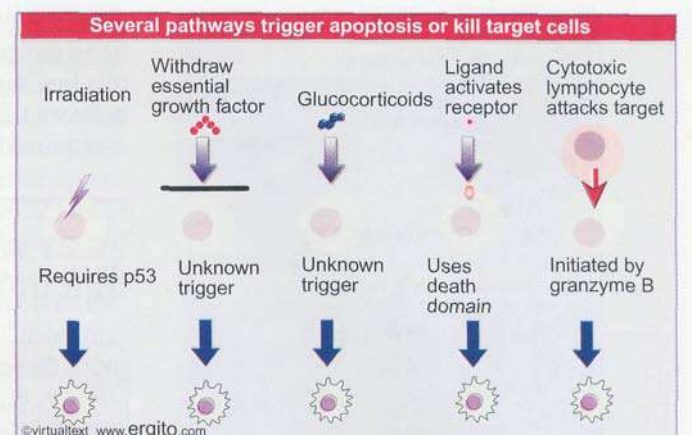


Figure 29.47 Apoptosis is triggered by a variety of pathways.

The names of the two proteins (Fas receptor and Fas ligand) reflect the way the system was discovered. An antibody directed against Fas protein kills cells that express Fas on their surface. The reason is that the antibody-Fas reaction activates Fas, which triggers a pathway for apoptosis. This defines Fas as a receptor that activates a cellular pathway.

Fas is a cell surface receptor related to the TNF (tumor necrosis factor) receptor. The FasL ligand is a transmembrane protein related to TNF. A family of related receptors includes two TNF receptors, Fas, and several receptors found on T lymphocytes. A corresponding family of ligands comprises a series of transmembrane proteins. This suggests that there are several pathways, each of which can be triggered by a cell-cell interaction, in which the "ligand" on one cell surface interacts with the receptor on the surface of the other cell. Both the Fas- and TNF-receptors can activate apoptosis.

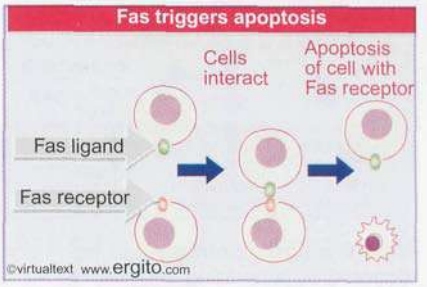


Figure 29.48 The Fas receptor and ligand are both membrane proteins. A target cell bearing Fas receptor apoptoses when it interacts with a cell bearing the Fas ligand.

Both of the Fas and TNF ligands are initially produced as membrane-bound forms, but can also be cleaved to generate soluble proteins, which function as diffusible factors. The soluble form of TNF is largely produced by macrophages, and is a pleiotropic factor that signals many cellular responses, including cytotoxicity. Most of its responses are triggered by interaction with one of the TNF receptors, TNF-R1. FasL is cleaved to generate a soluble form, but the soluble form is much less active than the membrane-bound form, so the reaction probably is used to reduce the activity of the cell bearing the ligand.

An assay for the capacity of the ligand-receptor interaction to trigger apoptosis is to introduce the receptor into cultured cells that do not usually express it. On treatment with the ligand, the transfected cells die by apoptosis, but the parental cells do not. Using this assay, similar results are obtained with FasL/Fas receptor and with TNF/TNF-R1. Mutant versions of the receptor show that the apoptotic response is triggered by an ~80 amino acid intracellular domain near the C-terminus. This region is loosely conserved (~28%) between Fas and TNF-R1, and is called the death domain.

An assay for components of the apoptotic pathway in the cell is to see whether their overexpression causes apoptosis. This is done by transfecting the gene for the protein into the cell (which results in overexpression of the protein). This assay identifies several proteins that interact specifically with the Fas and/or TNF receptors. All of these proteins themselves have death domains, and it is possible that a homomeric interaction between two death domains provides the means by which the signal is passed from the receptor to the next component of the pathway.

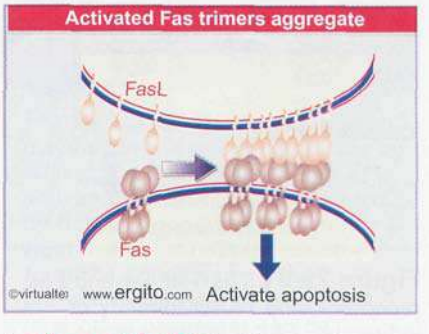


Figure 29.49 Fas forms trimers that are activated when binding to FasL causes aggregation.

The validity of this pathway *in vivo* was demonstrated by the discovery of the mouse mutation *lpr*. This is a recessive mutation in the gene for Fas. It causes proliferation of lymphocytes, resulting in a complex immune disorder affecting both B cells and T cells. Another mutation with similar

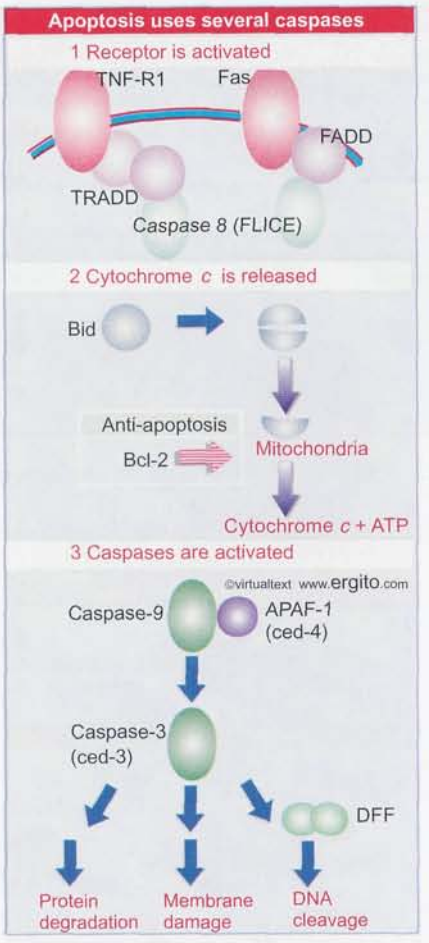


Figure 29.50 Apoptosis can be triggered by activating surface receptors. Caspase proteases are activated at two stages in the pathway. Caspase-8 is activated by the receptor. This leads to release of cytochome *c* from mitochondria. Apoptosis can be blocked at this stage by Bcl-2. Cytochrome *c* activates a pathway involving more caspases.

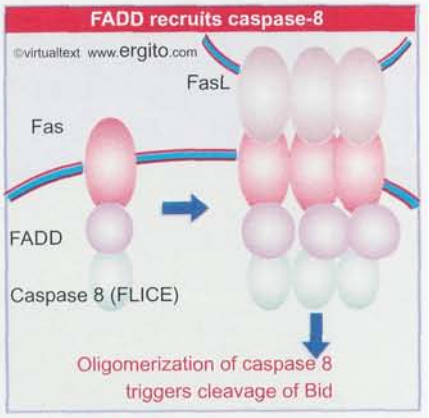


Figure 29.51 The TNF-R1 and Fas receptors bind FADD (directly or indirectly). FADD binds caspase-8. Activation of the receptor causes oligomerization of caspase-8, which activates the caspase.

effects is *gld* (generalized lymphoproliferative disease). This turns out to lie in the gene that codes the FasL ligand. The related properties of these two loci suggest that this apoptotic pathway is triggered by an interaction between the FasL ligand (*gld* product) and Fas (*lpr* product). The pathway is required for limiting the numbers of mature lymphocytes.

29.27 A common pathway for apoptosis functions via caspases

Key Concepts

- Caspases are proteases that are involved in multiple stages of the apoptotic pathway.
- Caspases are synthesized as inactive procaspases that are activated by autocleavage to form the active dimer.
- A complex forms at the Fas or TNF receptor that activates caspase-8 to initiate the intracellular pathway.

The "classical" pathway for apoptosis is summarized in **Figure 29.50.** A ligand-receptor interaction triggers the activation of a protease. This leads to the release of cytochrome *c* from mitochondria. This in turn activates a series of proteases, whose actions culminate in the destruction of cell structures.

A complex containing several components forms at the receptor. The exact components of the complex depends on the receptor. TNF receptor binds a protein called TRADD, which in turn binds a protein called FADD. Fas receptor binds FADD directly. **Figure 29.51** shows that, in either case, FADD binds the protein caspase-8 (also known as FLICE), which has a death domain as well as protease catalytic activity. The activation of caspase-8 activates a common pathway for apoptosis. The trigger for the activation event is the oligomerization of the receptor. In the case of the Fas system, the interaction of FasL with Fas causes the Fas trimers to interact, activating the pathway.

Members of the caspase family (cysteine aspartate proteases) are important downstream components of the pathway. Caspases have a catalytic cysteine, and cleave their targets at an aspartate. Individual enzymes have related, but not identical targets. For example, caspase-3 and ICE both cleave at tetrapeptide sequences in their substrates, but caspase-3 recognizes YVAD and ICE recognizes DEVD. There are ~14 mammalian members of the caspase family.

Caspases fall into two groups. The caspase-1 subfamily is involved in the response to inflammation. The caspase-3 subfamily (consisting of caspase 3 and caspases 6-10) is involved in apoptosis. All caspases are synthesized in the form of inactive procaspases, which have additional sequences at the N-terminus. **Figure 29.52** shows that the activation reaction involves cleavage of the prodomain followed by cleavage of the caspase sequence itself into a small subunit and large subunit. All procaspases except procaspase-9 probably exist as dimers.

Caspases with large prodomains are involved in initiating apoptosis. Dimerization causes an autocatalytic cleavage that activates the caspase. The prodomain of caspase-8 has two death domain motifs that are responsible for its association with the receptor complex. Cleavage to the active form occurs as soon as procaspase-8 is recruited to the receptor complex.

Caspases with small prodomains function later in the pathway. The first in the series is activated by an autocleavage when it forms an oligomer. Others later in the pathway typically are activated when another caspase cleaves them.

The first caspase to be discovered (ICE = caspase-1) was the IL-1β-converting enzyme, which cleaves the pro-IL-1β precursor into its active form. Although this caspase is usually involved with the inflammatory response, transfection of ICE into cultured cells causes apoptosis. The process is inhibited by CrmA (a product of cowpox virus). All caspases are inhibited by CrmA, although each caspase has a characteristic sensitivity. CrmA inhibits apoptosis triggered in several different ways, which demonstrates that the caspases play an essential role in the pathway, irrespective of how it is initiated. However, it turns out that ICE is not itself the protease commonly involved in apoptosis, because inactivation of the gene for ICE does not block general apoptosis in the mouse. (The ability of ICE to cause apoptosis demonstrates a danger of the transfection assay: overexpression may allow it to trigger apoptosis, although usually it does not do so. But ICE may be needed specifically for apoptosis of one pathway in lymphocytes.)

29.28 Apoptosis involves changes at the mitochondrial envelope

Key Concepts

- Caspase-8 cleaves Bid to release a C-terminal domain that translocates to the mitochondrion.
- Bid is a member of the Bcl2 family and acts together with other members of the family to cause mitochondria to release cytochrome *c*.
- Some members of the family, including Bcl2, inhibit the release of cytochrome *c*.

Changes in mitochondria occur during apoptosis (and also during other forms of cell death). These are typically detected by changes in permeability. The breakthrough in understanding the role of mitochondria was the discovery that cytochrome c is released into the cytosol. **Figure 29.53** summarizes the central role of the mitochondrion. In addition to releasing cytochrome c, it also releases other proteins from its intermembrane space that may either promote or inhibit apoptosis.

The pathway moves from the plasma membrane to the mitochondrion when caspase-8 cleaves a protein called Bid. The cleavage releases the C-terminal domain, which then translocates to the mitochondrial membrane. The action of Bid causes cytochrome *c* to be released.

Bid is a member of the important Bcl2 family. Some members of this family are required for apoptosis, while others counteract apoptosis. The eponymous Bcl2 inhibits apoptosis in many cells. It has a C-terminal membrane anchor, and is found on the outer mitochondrial, nuclear, and ER membranes. It prevents the release of cytochrome c, which suggests that in some way it counteracts the action of Bid.

bcl2 was originally discovered as a proto-oncogene that is activated in lymphomas by translocations resulting in its overexpression. (As discussed in more detail in *30 Oncogenes and cancer,* this means that Bcl2 is a member of a class of proteins that causes proliferation or tumorigenesis when inappropriately expressed.) Its role as an inhibitor of apoptosis was discovered when it was shown that its addition protects cultured lymphoid and myeloid cells from dying when the essential factor IL-3 is withdrawn.

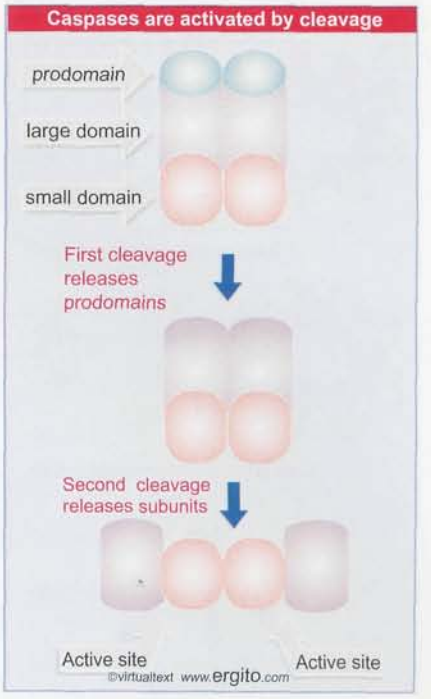


Figure 29.52 Caspase activation requires dimerization and two cleavages.

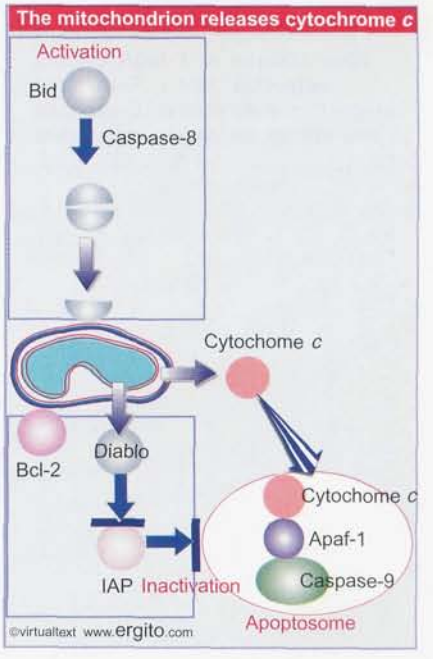


Figure 29.53 The mitochondrion plays a central role in apoptosis by releasing cytochrome *c*. This is activated by BID. It is inactivated by Bcl-2. Cytochrome *c* binds to *Apaf-1 and (pro)-Caspase-9* to form the apoptosome. The proteolytic activity of caspase-9 (and other caspases) can be inhibited by IAP proteins. Proteins that antagonize IAPs may be released from the mitochondrion.

Mammalian cells that are triggered into apoptosis by a wide variety of stimuli, including activation of the Fas/TNF-R1 pathways, can be rescued by expression of Bcl2. This suggests that these pathways converge on a single mechanism of cell killing, and that Bcl2 functions at a late, common stage of cell death. There are some systems in which Bcl2 cannot block apoptosis, so the pathway that it blocks may be common, but is not the only one.

Bcl2 belongs to a family whose members can homodimerize and heterodimerize. Two other members are bcl-x (characterized in chicken) and Bax (characterized in man). bcl-x is produced in alternatively spliced forms that have different properties. When transfected into recipient cells, bcl-x_L mimics Bcl2, and inhibits apoptosis. But bcl-x_S counteracts the ability of Bcl2 to protect against apoptosis. Bax behaves in the same way as bcl-x_S. This suggests that the formation of Bcl2 homodimers may be needed to provide the protective form, and that Bcl2/Bax or Bcl2/bcl-x_S heterodimers may fail to protect. Whether Bax or bcl-x_S homodimers actively assist apoptosis, or are merely permissive, remains to be seen. The general conclusion suggested by these results is that combinatorial associations between members of the family may produce dimers with different effects on apoptosis, and the relative proportions of the family members that are expressed may be important. The susceptibility of a cell to undergo apoptosis may be proportional to the ratio of Bax to Bcl2.

The mitochondrion is a crucial control point in the induction of apoptosis. The release of cytochrome *c* is preceded by changes in the permeability of the mitochondrial membrane. Bcl2 family members act at the mitochondrial membrane, and although their mode of action is not known, one possibility is that they form channels in the membrane. Apoptosis involves localization (or perhaps increased concentration) of Bcl2 family members at the mitochondrial membrane, including Bid (required to release cytochrome c) and Bax (perhaps involved in membrane permeability changes).

29.29 Cytochrome *c* activates the next stage of apoptosis

Key Concepts

* Cytochrome c causes Apaf-1 to aggregate with procaspase-9 to form the autosome, which then activates caspase-9 by autocleavage.
* Caspase-9 cleaves caspase-3 and other caspases to trigger the effector phase of apoptosis, when cellular structures are destroyed.

The release of cytochrome *c* is a crucial control point in the pathway. The basic role of cytochrome *c* is to trigger the activation of caspase-9. **Figure 29.54** shows the stages between cytochrome *c* release and caspase-9 activation. Cytochrome *c* triggers the interaction of the cytosolic protein Apaf-1 with caspase-9 in a complex called the apoptosome. The reaction takes place in several stages. Cytochrome *c* binds to Apaf-1. This enables Apaf-1 to bind ATP. This in turn enables it to oligomerize, which causes a change of conformation that exposes the caspase-binding domain; then Apaf-1 binds procapase-9. The incorporation of procaspase-9 into the apoptosome triggers the auto-activating cleavage.

The properties of mice lacking Apaf-1 or caspase-9 throw some light upon the generality of apoptotic pathways. Lack of caspase-9 is

lethal, because the mice have a malformed cerebrum as the result of the failure of apoptosis. Apoptosis is also reduced in thymocytes (immune precursors to lymphocytes). Apaf-1 deficient mice have less severe defects in brain development, implying that there are alternative means for activating caspase-9. Both types of deficient mice continue to show Fas-mediated apoptosis, implying that Fas has alternative means of triggering apoptosis.

Caspase-9 in turn cleaves procaspase-3 to generate caspase-3 (which is in fact the best characterized component of the downstream pathway. Caspase-3 is the homologue of the *C. elegans* protein ced-3; see below). Caspase-9 also activates caspases-6 and 7.

Caspase-3 acts at what might be called the effector stage of the pathway. We have not identified all of the targets of the protease activity that are essential for apoptosis. One known target is the enzyme PARP (poly[ADP-ribose] polymerase). Its degradation is not essential, but is a useful diagnostic for apoptosis.

One pathway that leads to DNA fragmentation has been identified. Caspase-3 cleaves one subunit of a dimer called DFF (DNA fragmentation factor). The other subunit then activates a nuclease that degrades DNA. However, the degradation of DNA by this pathway does not appear to be necessary for cell death, which continues in mice that lack the enzyme.

Another pathway for DNA degradation is triggered directly by release of an enzyme from the mitochondrion. The normal function of endonuclease G within the mitochondrion is concerned with DNA replication. However, in apoptosing cells it is released from the mitochondrion, and then degrades nuclear DNA. Interference with the function of the corresponding gene in *C. elegans* reduces DNA degradation and delays the appearance of cell corpses. This enzyme therefore appears to be important at least for the time course of apoptosis, even if it is not necessary for the eventual death of the cell.

The control of apoptosis involves components that inhibit the pathway as well as those that activate it. This first became clear from the genetic analysis of cell death in *C. elegans*, when mutants were found that either activate or inactivate cell death. Mutations in *ced-3* and *ced-4* cause the survival of cells that usually die, demonstrating that these genes are essential for cell death. *ced-3* codes for the protease activity (and was in fact the means by which caspases were first implicated in apoptosis). It is the only protease of this type in *C. elegans*. *ced-4* codes for the homologue to Apaf-1.

ced-9 inhibits apoptosis. It codes for the counterpart of Bcl2. A mutation that inactivates *ced-9* is lethal, because it causes the death of cells that should survive. This process requires *ced-3* and *ced-4*, and this was the original basis for the idea that *ced-9* blocks the apoptotic pathway(s) in which *ced-3* and *ced-4* participate. This relationship makes an important point: *ced-3* and *ced-4* are not expressed solely in cells that are destined to die, but are expressed also in other cells, where normally their action is prevented by *ced-9*. The proper control of apoptosis may therefore involve a balance between activation and inhibition of this pathway.

The apoptotic pathway can also be inhibited at the stages catalyzed by the later caspases. Proteins called IAP (inhibitor of apoptosis) can bind to procaspases and activated caspases to block their activities (see Figure 29.53). The blocking activities of the AIPs need to be antagonized in order for apoptosis to proceed. Vertebrate cells contain a protein called Diablo/Smac, which is released from mitochondria at the same time as cytochrome *c*, and acts by binding to IAPs.

The existence of mechanisms to inhibit as well as to activate apoptosis suggests that many (possibly even all) cells possess the intrinsic capacity to apoptose. If the components of the pathway are ubiquitous,

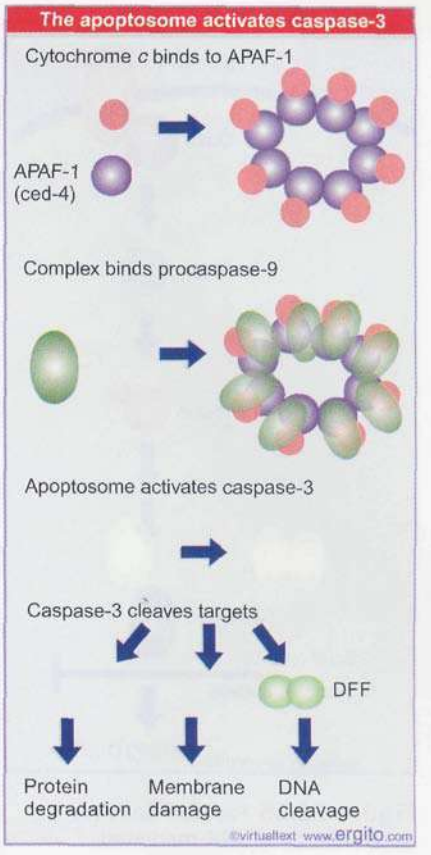


Figure 29.54 Cytochrome *c* causes Apaf-1 to interact with caspase-9, which activates caspase-3, which cleaves targets that cause apoptosis of the cell.

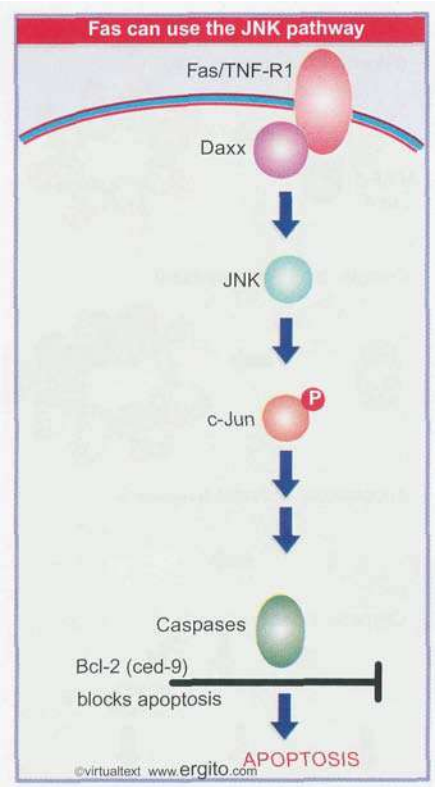


Figure 29.55 Fas can activate apoptosis by a JNK-mediated pathway.

the critical determinant of whether a cell lives or dies may depend on the regulatory mechanisms that determine whether the pathway is activated or repressed.

29.30 There are multiple apoptotic pathways

i Key Concepts

Fas activates apoptosis via the caspase pathway and also via the activation of the JNK kinase.

The pathway shown in Figure 29.50 is the prototypical pathway for activation of apoptosis via a protease cascade. However, Fas can also activate apoptosis by a pathway that involves the kinase JNK, whose most prominent substrate is the transcription factor c-Jun (see Figure 28.43). This leads by undefined means to the activation of proteases. **Figure 29.55** shows that this pathway is mediated by the protein Daxx (which does not have a death domain). Binding of FADD and Daxx to Fas is independent: each adaptor recognizes a different site on Fas. The two pathways function independently after Fas has engaged the adaptor. The TNF receptor also can activate JNK by means of distinct adaptor proteins.

In the normal course of events, activation of Fas probably activates both pathways. Overexpression experiments show that either pathway can cause apoptosis. The relative importance of the two pathways may vary with the individual cell type, in response to other signals that affect each pathway. For example, JNK is activated by several forms of stress independently of the Fas-activated pathway. This pathway is not inhibited by Bcl2, which may explain the variable ability of cells to resist apoptosis in response to Bcl2.

Another apoptotic pathway is triggered by cytotoxic T lymphocytes, which kill target cells by a process that involves the release of granules containing serine proteases and other lytic components. One such component is perforin, which can make holes in the target cell membrane, and under some conditions can kill target cells. The serine proteases in the granules are called granzymes. In the presence of perforin, granzyme B can induce many of the features of apoptosis, including fragmentation of DNA. It activates a caspase called Ich-3, which is necessary for apoptosis in this pathway.

29.31 Summary

The cell cycle consists of transitions from one regulatory state to another. The change in regulatory state is separated by a lag period from the subsequent changes in cell phenotype. The transitions take the form of activating or inactivating a kinase(s), which modifies substrates that determine the physical state of the cell. Checkpoints can retard a transition until some intrinsic or extrinsic condition has been satisfied.

The two key control points in the cell cycle are in G1 and at the end of G2. During G1, a commitment is made to enter a replication cycle; the decision is identified by the restriction point in animal cells, and by START in yeast cells. After this decision has been taken, cells are committed to beginning an S phase, although there is a lag period before DNA replication initiates. The end of G2 is marked by a decision that is executed immediately to enter mitosis.

By Book_Crazy [IND]

A unifying feature in the cell cycles of yeasts and animals is the existence of an M phase kinase, consisting of two subunits: Cdc2, with serine/threonine protein kinase catalytic activity; and a mitotic cyclin of either the A or B class. Homologous subunits exist in all eukaryotic cells. The genes that code for the catalytic subunit in yeasts are *cdc2* in *S. pombe* and *CDC28* in *S. cerevisiae.* Animal cells usually contain multiple mitotic cyclins (A, B1, B2); in *S. pombe,* there is only a single cyclin at M phase, a B class coded by *cdc13,* although *S. pombe* has several CLB proteins.

The activity of the M phase kinase is controlled by the state of phosphorylation of the catalytic subunit. The active form requires dephosphorylation on Tyr-15 (in yeasts) or Thr-14/Tyr-15 (in animal cells) and phosphorylation on Thr-161. The cyclins are also phosphorylated, but the significance of this modification is not known. In animal cells, the kinase is inactivated by degradation of the cyclin component, which occurs abruptly during mitosis. Cyclins of the A type are typically degraded before cyclins of the B type. Destruction of at least the B cyclins, and probably of both classes of cyclin, is required for cells to exit mitosis.

A comprehensive analysis of genes that affect the cell cycle has identified *cdc* mutants in both *S. pombe* and *S. cerevisiae.* The best characterized mutations are those that affect the components or activity of M phase kinase. Mutations *cdc25* and *wee1* in *S. pombe* have opposing effects in regulating M phase kinase in response to cell size (and other signals). Wee1 is a kinase that acts on Tyr-15 and maintains Cdc2 in an inactive state; Cdc25 is a phosphatase that acts on Tyr-15 and activates Cdc2. The existence of *wee1* and *cdc25* homologues in higher eukaryotes suggests that the apparatus for cell cycle control is widely conserved in evolution.

By phosphorylating appropriate substrates, the kinase provides MPF activity, which stimulates mitosis or meiosis (as originally defined in *Xenopus* oocytes). A prominent substrate is histone H1, and H1 kinase activity is now used as a routine assay for M phase kinase. Phosphorylation of H1 could be concerned with the need to condense chromatin at mitosis. Another class of substrates comprises the lamins, whose phosphorylation causes the dissolution of the nuclear lamina. A general principle governing these (and presumably other) events is that the state of the substrates is controlled reversibly in response to phosphorylation, so that the phosphorylated form of the protein is required for mitotic organization, while the dephosphorylated form is required for interphase organization. Phosphatases are required to reverse the modifications introduced by M phase kinase.

Transition from G1 into S phase requires a kinase related to the M phase kinase. In yeasts, the catalytic subunit is identical with that of the M phase kinase, but the cyclins are different (the combinations being CDC28-cig1,2 in *S. pombe,* Cdc2-CLN 1,2,3 in *S. cerevisiae).* Activity of the G1/S phase kinase and inactivity of the M phase kinase are both required to proceed through G1. Initiation of S phase in S. *pombe* requires rum1 to inactivate cdc2/cdc13 in order to allow the activation of Cdc18, which may be the S phase activator.

In mammalian cells, a family of catalytic subunits is provided by the *cdk* genes, named because they code for the catalytic subunits of cyclin-dependent kinases. There are ~10 *cdk* genes in an animal genome. Aside from the classic Cdc2, the best characterized product is cdk2 (which is well related to Cdc2). In a normal cell cycle, cdk2 is partnered by cyclin E during the G1/S transition and by cyclin A during the progression of S phase. cdk2, cdk4, and cdk5 all partner the D cyclins to form kinases that are involved with the transition from GO to G1. These cdk-cyclin complexes phosphorylate RB, causing it to release the transcription factor E2F, which then activates genes whose products are required for S phase. A group of CKI (inhibitor) proteins that are activated by treatments that inhibit growth can bind to cdk-cyclin complexes, and maintain them in an inactive form.

Checkpoints control progression of the cell cycle. One checkpoint responds to the presence of unreplicated or damaged DNA by

blocking mitosis. Others control progress through mitosis, for example, detecting unpaired kinetochores.

Apoptosis is achieved by an active pathway that executes a program for cell death. The components of the pathway may be present in many or all higher eukaryotic cells. Apoptosis may be triggered by various stimuli. A common pathway involves activation of caspase-8 by oligomerization at an activated surface receptor. Caspase-8 cleaves Bid, which triggers release of cytochrome *c* from mitochondria. The cytochrome *c* causes Apaf-1 to oligomerize with caspase-9. The activated caspase-9 cleaves procaspase-3, whose two subunits then form the active protease. This cleaves various targets that lead to cell death. The pathway is inhibited by Bcl2 at the stage of release of cytochrome *c.* An alternative pathway for triggering apoptosis that does not pass through Apaf-1 and caspase-9, and which is not inhibited by Bcl2, involves the activation of JNK. Different cells use these pathways to differing extents. Apoptosis was first shown to be necessary for normal development in *C. elegans,* and knockout mutations in mice show that this is also true of vertebrates. Every cell may contain the components of the apoptotic pathway and be subject to regulation of the balance between activation and repression of cell death.

References

29.1 Introduction

ref Howard, A. and Pelc, S. (1953). Synthesis of DNA in normal and irradiated cells and its relation to chromosome breakage. Heredity Suppl. 6, 261-273.

29.3 Checkpoints occur throughout the cell cycle

rev Hartwell, L. H., and Weinert, T. A. (1989). Checkpoints: Controls that ensure the order of cell cycle events. Science 246, 629-634.

29.5 M phase kinase regulates entry into mitosis

exp Masui (2002). The discovery of MPF (www.ergito.com/lookup.jsp?expt= masui)

29.6 M phase kinase is a dimer of a catalytic subunit and a regulatory cyclin

ref Draetta, G., Luca, F., Westendorf, J., Brizuela, L, Ruderman, J., and Beach, D. (1989). cdc2 protein kinase is complexed with both cyclin A and B: evidence for proteolytic inactivation of MPF. Cell 56, 829-38.

Evans, T. et al. (1983). Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. Cell 33, 389-396.

Gould, K. L. and Nurse, P. (1989). Tyrosine phosphorylation of the fission yeast $cdc2^{+}$ protein kinase regulates entry into mitosis. Nature 342, 39-45.

Murray, A. W., Solomon, M. J., and Kirschner, M. W. (1989). The role of cyclin synthesis and degradation in the control of maturation promoting factor activity. Nature 339, 280-286.

Riabowol, K., Draetta, G., Brizuela, L., Vandre, D., and Beach, D. (1989). The cdc2 kinase is a nuclear protein that is essential for mitosis in mammalian cells. Cell 57, 393-401.

Simanis, V. and Nurse, P. (1986). The cell cycle control gene $cdc2^{C}$ of fission yeast encodes a protein kinase potentially regulated by phosphorylation. Cell 45, 261-268.

29.7 Protein phosphorylation and dephosphorylation control the cell cycle

ref Arion, D., Meijer, L., Brizuela, L., and Beach, D. (1988). Cdc2 is a component of the M phase-specific histone H1 kinase: evidence for identity with MPF. Cell 55, 371-378.

Labbe, J. C, Picard, A., Peaucellier, G., Cavadore, J. C, Nurse, P., and Doree, M. (1989). Purification of MPF from starfish: identification as the H1 histone kinase $p34^{cdc2}$ and a possible mechanism for its periodic activation. Cell 57, 253-263.

29.8 Many cell cycle mutants have been found by screens in yeast

ref Hartwell, L., Culotti, J., Pringle, J. R., and Reid, B. J. (1974). Genetic control of the cell division cycle in yeast. Science 183, 46-51.

29.9 Cdc2 is the key regulator in yeasts

exp Nurse, P. (2002). The Discovery of cdc2 as the Key Regulator of the Cell Cycle (www.ergito.com/lookup.jsp?expt= nurse)

ref Dunphy, W. G., Brizuela, L., Beach, D., and Newport, J. (1988). The *Xenopus* cdc2 protein is a component of MPF, a cytoplasmic regulator of mitosis. Cell 54, 423-431.

Gautier, J., Norbury, C, Lohka, M., Nurse, P., and Mailer, J. (1988). Purified maturation-promoting factor contains the product of a *Xenopus* homologue of the fission yeast cell cycle control gene $cdc2^{+}$. Cell 54, 433-439.

29.1 1 *CDC28* acts at both START and mitosis in *S. cerevisiae*

rev Forsburg, S. L. and Nurse, P. (1991). Cell cycle regulation in the yeasts *S. cerevisiae* and *S. pombe.* Ann. Rev. Cell Biol. 7, 227-256.

29.12 Cdc2 activity is controlled by kinases and phosphatases

rev Murray, A. W. and Kirschner, M. W. (1989). Dominoes and clocks: the union of two views of the cell cycle. Science 246, 614-621.

Nurse, P. (1990). Universal control mechanism regulating onset of M phase. Nature 344, 503-508.

ref Gautier, J., Solomon, M. J., Booher, R. N., Bazan, J. F., and Kirschner, M. W. (1991). cdc25 is a specific tyrosine phosphatase that directly activates $p34^{cdc2}$. Cell 67, 197-211.

Hayles, J. et al. (1994). Temporal order of S phase and mitosis in fission yeast is determined by the state of the $p34^{cdc2}$ mitotic B cyclin complex. Cell 78, 813-822.

29.13 DNA damage triggers a checkpoint

rev Nyberg, K. A., Michelson, R. J., Putnam, C. W., and Weinert, T. A. (2002). Toward Maintaining the Genome: DNA Damage and Replication Checkpoints. Ann. Rev. Genet. 36, 617-656.

Zhou, B. B. and Elledge, S. J. (2000). The DNA damage response: putting checkpoints in perspective. Nature 408, 433-439.

ref Bakkenist, C. J. and Kastan, M. B. (2003). DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. Nature 421, 499-506.

Lee, S. E., Moore, J. K., Holmes, A., Umezu, K., Kolodner, R. D., and Haber, J. E. (1998). Saccharomyces Ku70, mre11/rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. Cell 94, 399-409.

Rouse, J. and Jackson, S. P. (2002). Interfaces between the detection, signaling, and repair of DNA damage. Science 297, 547-551.

Weinert, T. A., and Hartwell, L. H. (1988). The RAD9 gene controls the cell cycle response to DNA damage in *S. cerevisiae.* Science 241, 317-322.

29.14 The animal cell cycle is controlled by many cdk-cyclin complexes

rev Norbury, C. and Nurse, P. (1992). Animal cell cycles and their control. Ann. Rev. Biochem. 61, 441-470.

29.15 Dimers are controlled by phosphorylation of cdk subunits and by availability of cyclin subunits

rev Herr, C. J. (1993). Mammalian G1 cyclins. Cell 73, 1059-1065.

Nurse, P. (1994). Ordering S phase and M phase in the cell cycle. Cell 79, 547-550.

Reed, S. I. (1992). The role of p34 kinases in the G1 to S phase transition. Ann. Rev. Cell Biol. 8, 529-561.

Sherr, C. J. (1994). G2 phase progression: cycling on cue. Cell 79, 551-555.

ref Blow, J. J. and Nurse, P. (1990). A cdc2-like protein is involved in the initiation of DNA replication in *Xenopus* egg extracts. Cell 62, 855-862.

Fisher, R. P. and Morgan, D. O. (1994). A novel cyclin associates with MO15/cdk7 to form the cdk-activating kinase. Cell 78, 713-724.

Jeffrey, P. D. et al. (1995). Mechanism of cdk activation revealed by the structure of a cyclin A-cdk2 complex. Nature 376, 313-320.

29.16 RB is a major substrate for cdk-cyclin complexes

rev Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. Cell 81, 323-330.

ref Harbour, J. W. and Dean, D. C. (2000). The Rb/E2F pathway: expanding roles and emerging paradigms. Genes Dev. 14, 2393-2409.

29.17 G0/G1 and G1/S transitions involve cdk inhibitors

rev Deshaies, R. J. (1999). SCF and Cullin/Ring H2-based ubiquitin ligases. Ann. Rev. Cell Dev. Biol. 15, 435-467.

Hunter, T. and Pines, J. (1994). Cyclins and cancer II: cyclin D and CDK inhibitors come of age. Cell 79, 573-582.

Sherr, C. J. and Roberts, J. M. (1995). Inhibitors of mammalian G1 cyclin-dependent kinases. Genes Dev. 9, 1149-1163.

ref Skowyra, D., et al. (1997). F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. Cell 91, 209-219.

29.18 Protein degradation is important in mitosis

rev Harper, J. W., Burton, J. L., and Solomon, M. J. (2002). The anaphase-promoting complex: it's not just for mitosis any more. Genes Dev. 16, 2179-2206.

Page, A. M. and Hieter, P. (1999). The anaphase-promoting complex: new subunits and regulators. Ann. Rev. Biochem. 68, 583-609.

ref Fang, G., Yu, H. and Kirschner, M. W. (1998). Direct binding of CDC20 protein family members activates the anaphase-promoting complex in mitosis and G1. Mol. Cell 2, 163-171.

Glotzer, M., Murray, A. W., and Kirschner, M. W. (1991). Cyclin is degraded by the ubiquitin pathway. Nature 349, 132-138.

Holloway, S. L. et al. (1993). Anaphase is initiated by proteolysis rather than by the inactivation of MPF. Cell 73, 1393-1402.

King, R. W. et al. (1995). A 20S complex containing CDC27 and CDC16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin B. Cell 81, 279-288.

29.19 Cohesins hold sister chromatids together

rev Hirano, T. (2000). Chromosome cohesion, condensation, and separation. Ann. Rev. Biochem. 69, 115-144.

Hirano, T. (1999). SMC-mediated chromosome mechanics: a conserved scheme from bacteria to vertebrates? Genes Dev. 13, 11-19.

Nasmyth, K. (2001). Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. Ann. Rev. Genet. 35, 673-745.

Nasmyth, K. (2002). Segregating sister genomes: the molecular biology of chromosome separation. Science 297, 559-565.

ref Buonomo, S. B., Clyne, R. K., Fuchs, J., Loidl, J., Uhlmann, F., and Nasmyth, K. (2000). Disjunction of homologous chromosomes in meiosis I depends on proteolytic cleavage of the meiotic cohesin Rec8 by separin. Cell 103, 387-398.

Ciosk, R. et al. (1998). An ESP1/PDS1 complex regulates loss of sister chromatid cohesion at the metaphase to anaphase transition in yeast. Cell 93, 1067-1076.

Guacci, V., Hogan, E., and Koshland, D. (1994). Chromosome condensation and sister chromatid pairing in budding yeast. J. Cell Biol. 125, 517-530.

Jallepalli, P. V., Waizenegger, I. C, Bunz, F., Langer, S., Speicher, M. R., Peters, J. M., Kinzler, K. W., Vogelstein, B., and Lengauer, C. (2001). Securin is required for chromosomal stability in human cells. Cell 105, 445-457.

Losada, A., Hirano, M., and Hirano, T. (1998). Identification of *Xenopus* SMC protein complexes required for sister chromatid cohesion. Genes Dev. 12, 1986-1997.

Skibbens, R. V., Corson, L. B.,and Koshland, D. (1999). Ctf7p is essential for sister chromatid cohesion and links mitotic chromosome structure to the DNA replication machinery. Genes Dev. 13, 307-319.

Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V., and Nasmyth, K. (2000). Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. Cell 103, 375-386.

Waizenegger, I. C, Hauf, S., Meinke, A., and Peters, J. M. (2000). Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase. Cell 103, 399-410.

29.20 Exit from mitosis is controlled by the location of Cdc14

ref Bardin, A. J., Visintin, R., and Amon, A. (2000). A mechanism for coupling exit from mitosis to partitioning of the nucleus. Cell 102, 21-31.

Shirayama, M. et al. (1999). APC^{CDC20} promotes exit from mitosis by destroying the anaphase inhibitor Pds1 and cyclin Clb5. Nature 402, 203-207.

Shou, W., Seol, J. H., Shevchenko, A., Baskerville, C, Moazed, D., Chen, Z. W., Jang, J., Shevchenko, A., Charbonneau, H., and Deshaies, R. J. (1999). Exit from mitosis is triggered by Tem1 -dependent release of the protein phosphatase Cdc14 from nucleolar RENT complex. Cell 97, 233-244.

Straight, A. F., Shou, W., Dowd, G. J., Turck, C. W., Deshaies, R. J., Johnson, A. D., and Moazed, D. (1999). Net1, a Sir2-associated nucleolar protein required for rDNA silencing and nucleolar integrity. Cell 97, 245-256.

29.21 The cell forms a spindle at mitosis

rev Goldman, R. D., Gruenbaum, Y., Moir, R. D., Shumaker, D. K., and Spann, T. P. (2002). Nuclear lamins: building blocks of nuclear architecture. Genes Dev. 16, 533-547.

King, R. W., Jackson, P. K., and Kirschner, M. W. (1994). Mitosis in transition. Cell 79, 563-571.

McIntosh, J. R. and Koonce, M. P. (1989). Mitosis. Science 246, 622-628.

ref Beaudouin, J., Gerlich, D., Daigle, N., Eils, R., and Ellenberg, J. (2002). Nuclear envelope breakdown proceeds by microtubule-induced tearing of the lamina. Cell 108, 83-96.

Foisner, R. and Gerace, L. (1993). Integral membrane proteins of the nuclear envelope interact with lamins and chromosomes, and binding is modulated by mitotic phosphorylation. Cell 73, 1267-1279.

Peter, M. et al. (1990). *In vitro* disassembly of the nuclear lamina and M phase-specific phosphorylation of lamins by cdc2 kinase. Cell 61, 591-602.

Salina, D., Bodoor, K., Eckley, D. M., Schroer, T. A., Rattner, J. B., and Burke, B. (2002). Cytoplasmic dynein as a facilitator of nuclear envelope breakdown. Cell 108, 97-107.

Yang, L, Guan, T., and Gerace, L. (1997). Integral membrane proteins of the nuclear envelope are dispersed throughout the endoplasmic reticulum during mitosis. J. Cell Biol. 137, 1199-1210.

29.22 The spindle is oriented by centrosomes

rev Doxsey, S. (2001). Re-evaluating centrosome function. Nat. Rev. Mol. Cell Biol. 2, 688-698.

Mitchison, T. J. (1988). Microtubule dynamics and kinetochore function in mitosis. Ann. Rev. Cell Biol. 4, 527-549.

Murray, A. W. and Szostak, J. W. (1985). Chromosome segregation in mitosis and meiosis. Ann. Rev. Cell Biol. 1, 289-315.

ref Zheng, Y., Wong, M. L., Alberts, B., and Mitchison, T. (1995). Nucleation of microtubule assembly by a gamma-tubulin-containing ring complex. Nature 378, 578-583.

29.23 A monomeric G protein controls spindle assembly

ref Gruss, O. J., Carazo-Salas, R. E., Schatz, C. A., Guarguaglini, G., Kast, J., Wilm, M., Le Bot, N., Vernos, I., Karsenti, E., and Mattaj, I. W. (2001). Ran induces spindle assembly by reversing the inhibitory effect of importin alpha on TPX2 activity. Cell 104, 83-93.

Nachury, M. V., Maresca, T. J., Salmon, W. C, Waterman-Storer, C. M., Waterman-Storer, R., Heald, R., and Weis, K. (2001). Importin beta is a mitotic target of the small GTPase Ran in spindle assembly. Cell 104, 95-106.

Ohba, T., Nakamura, M., Nishitani, H., and Nishimoto, T. (1999). Self-organization of microtubule asters induced in *Xenopus* egg extracts by GTP-bound Ran. Science 284, 1356-1358.

Wilde, A. and Zheng, Y. (1999). Stimulation of microtubule aster formation and spindle assembly by the small GTPase Ran. Science 284, 1359-1362.

29.24 Daughter cells are separated by cytokinesis

rev Glotzer, M. (2001). Animal cell cytokinesis. Ann. Rev. Cell Dev. Biol. 17, 351-386.

29.25 Apoptosis is a property of many or all cells

rev Ellis, R. E., Yuan, J., and Horvitz, H. R. (1991). Mechanisms and functions of cell death. Ann. Rev. Cell Biol. 7, 663-698.

29.26 The Fas receptor is a major trigger for apoptosis

rev Nagata, S. (1999). Fas ligand-induced apoptosis. Ann. Rev. Genet. 33, 29-55.

ref Chan, F. K., Chun, H. J., Zheng, L., Siegel, R. M., Bui, K. L., Lenardo, M. J., Chan, F. K., Chun, H. J., Zheng, L., Siegel, R. M., Bui, K. L., and Lenardo, M. J. (2000). A domain in TNF receptors that mediates ligand-independent receptor assembly and signaling. Science 288, 2351-2354.

Ito, N. et al. (1991). The polypeptide encoded by the cDNA for human cell surface antigen Fas can mediate apoptosis. Cell 66, 233-243.

Siegel, R. M., Frederiksen, J. K., Zacharias, D. A., Chan, F. K., Johnson, M., Lynch, D., Tsien, R. Y., and Lenardo, M. J. (2000). Fas preassociation required for apoptosis signaling and dominant inhibition by pathogenic mutations. Science 288, 2354-2357.

Suda, T. et al. (1993). Molecular cloning and expression of the Fas ligand, a novel member of the TNF family. Cell 75, 1169-1178.

Tartaglia, L. A. et al. (1993). A novel domain within the 55 kD TNF receptor signals cell death. Cell 74, 845-853.

Watanabe-Fukunaga, R. et al. (1992). Lymphoproliferation disorder in mice explained by defects in Fas antigen that mediates apoptosis. Nature 356, 314-317.

29.27 A common pathway for apoptosis functions via caspases

rev Budihardjo, I. et al. (1999). Biochemical pathways of caspase activation during apoptosis. Ann. Rev. Cell Dev. Biol. 15, 269-290.

Earnshaw, W. C, Martins, L. M., and Kaufmann, S. H. (1999). Mammalian caspases: structure, activation, substrates, and functions during apoptosis. Ann. Rev. Biochem. 68, 383-424.

Strasser, A., O'Connor, L., and Dixit, V. M. (2000). Apoptosis signaling. Ann. Rev. Biochem. 69, 217-245.

ref Boldin, M. P, Goncharov, T.M., Goltsev, Y.V., Wallach, D. (1996). Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1- and TNF-receptor-induced cell death. Cell 85, 803-815.

Miura, M. et al. (1993). Induction of apoptosis in fibroblasts by IL-1 β-converting enzyme, a mammalian homologue of the *C. elegans* death gene *ced-3.* Cell 75, 653-660.

Muzio, M. et al. (1996). FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death—inducing signaling complex. Cell 85, 817-827.

29.28 Apoptosis involves changes at the mitochondrial envelope

exp Wang, X. (2002). The role of cytochrome c in apoptosis (www.ergito.com/lookup.jsp?expt = wang)

rev Chao, D. T. and Korsmeyer, S. J. (1998). Bcl2 family: regulators of cell death. Ann. Rev. Immunol. 16, 395-419.

Wang, X. (2001). The expanding role of mitochondria in apoptosis. Genes Dev. 1 5, 2922-2933.

ref Vander Heiden, M. G. et al. (1997). Bcl-xL regulates the membrane potential and volume homeostasis of mitochondria. Cell 91, 627-637.

Li, H., Zhu, H., Xu, C. J., and Yuan, J. (1998). Cleavage of BID by caspase 8 mediates the mitochondrial damage in the Fas pathway of apoptosis. Cell 94, 491-501.

Liu, X., Kim, C.N., Yang, J., Jemmerson, R., Wang, X. (1996). Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c. Cell 86, 147-157.

Luo, X., Budihardjo, I., Zou, H., Slaughter, C, Wang, X. (1998). Bid, a Bcl2 interacting protein, mediates cytochrome c release from mitochondria in response to activation of cell surface death receptors. Cell 94, 481-490.

29.29 Cytochrome c activates the next stage of apoptosis

rev Salvesen, G. S. and Duckett, C. S. (2002). IAP proteins: blocking the road to death's door. Nat. Rev. Mol. Cell Biol. 3, 401-410.

Wang, X. (2001). The expanding role of mitochondria in apoptosis. Genes Dev. 15, 2922-2933.

ref Du, C, Fang, M., Li, Y., Li, L., and Wang, X. (2000). Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition. Cell 102, 33-42.

Li, P., et al. (1997). Cytochrome c and ATP-dependent formation of Apaf-1/caspase-9 complex initiates an apoptotic protease cascade. Cell 91, 479-489.

Li, L. Y., Luo, X., and Wang, X. (2001). Endonuclease G is an apoptotic DNase when released from mitochondria. Nature 412, 95-99.

Liu, X. et al. (1997). DFF, a heterodimeric protein that functions downstream of caspase-3 to trigger DNA fragmentation during apoptosis. Cell 89, 175-184.

Parrish, J., Li, L., Klotz, K., Ledwich, D., Wang, X., and Xue, D. (2001). Mitochondrial endonuclease G is important for apoptosis in *C. elegans.* Nature 412, 90-94.

Verhagen, A. M., Ekert, P. G., Pakusch, M., Silke, J., Connolly, L. M., Reid, G. E., Moritz, R. L., Simpson, R. J., and Vaux, D. L. (2000). Identification of DIABLO, a mammalian protein that promotes apoptosis by binding to and antagonizing IAP proteins. Cell 102, 43-53.

Zhang, J., Liu, X., Scherer, D. C, van Kaer, L., Wang, X., and Xu, M. (1998). Resistance to DNA fragmentation and chromatin condensation in mice lacking the DNA fragmentation factor 45. Proc. Nat. Acad. Sci. USA 95, 12480-12485.

Zou, H., Li, Y., Liu, X., and Wang, X. (1999). An APAF-1 .cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9. J. Biol. Chem. 274, 11549-11556.

29.30 There are multiple apoptotic pathways

ref Yang, X. et al. (1997). Daxx, a novel Fas-binding protein that activates JNK and apoptosis. Cell 89, 1067-1076.

Chapter 30

Oncogenes and cancer

30.1 Introduction

A major feature of all higher eukaryotes is the defined life span of the organism, a property that extends to the individual somatic cells, whose growth and division are highly regulated. A notable exception is provided by cancer cells, which arise as variants that have lost their usual growth control. Their ability to grow in inappropriate locations or to propagate indefinitely may be lethal for the individual organism in which they occur.

Figure 30.1 shows that the incidence of cancer increases exponentially with age in a human population from the age of ~40 to ~80, when it plateaus. Immediately this suggests that cancer is the result of the occurrence of a series of independent events. From the trend of the curve we can estimate that a range of 4-10 stochastic events are required to generate a cancer.

The basic model for the occurrence of cancer is that cancer is a multistage process in which initiation of a tumor requires several steps, which may then be followed by further changes to strengthen the tumorigenic state. Tumor progression is then driven by selection among the tumor cells for those that can grow more aggressively. Many different types of events contribute to this process at the molecular level.

The two major types of change in the genome are the accumulation of somatic mutations and the development of genetic instability. There is still much debate about the relative importance of their contributions to the cancerous state.

Most cancer cells have an increased number of mutations compared to normal cells. As the cancer progresses, the number of mutations increases. However, the rate of somatic mutation is not sufficient to account for the accumulation of mutations that is observed in the cancer cells. **Figure 30.2** illustrates the view that some of the early mutations occur in mutator

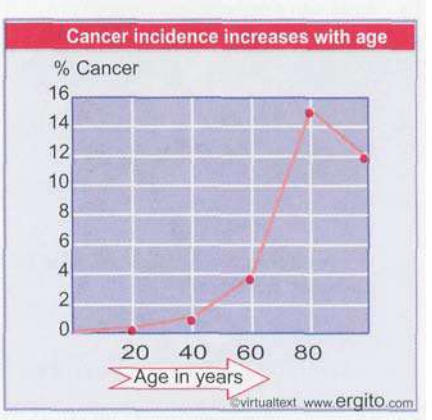


Figure 30.1 The frequency of newly diagnosed cancers increases sharply with age. Each point on the graph gives the incidence over a one year period for the preceding age group (the first point is for 0-20 years of age, etc.). The incidence rate plateaus above age 80.

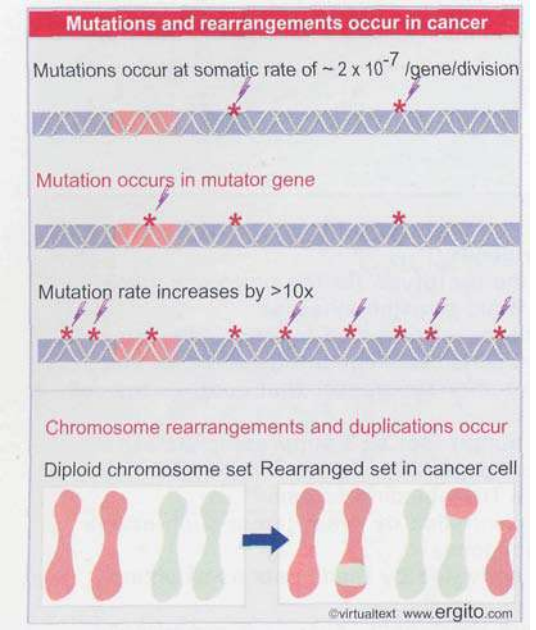


Figure 30.2 Cancer cells accumulate mutations at an increased rate as a result of the mutation of a mutator gene and also undergo changes in the chromosomal constitution.

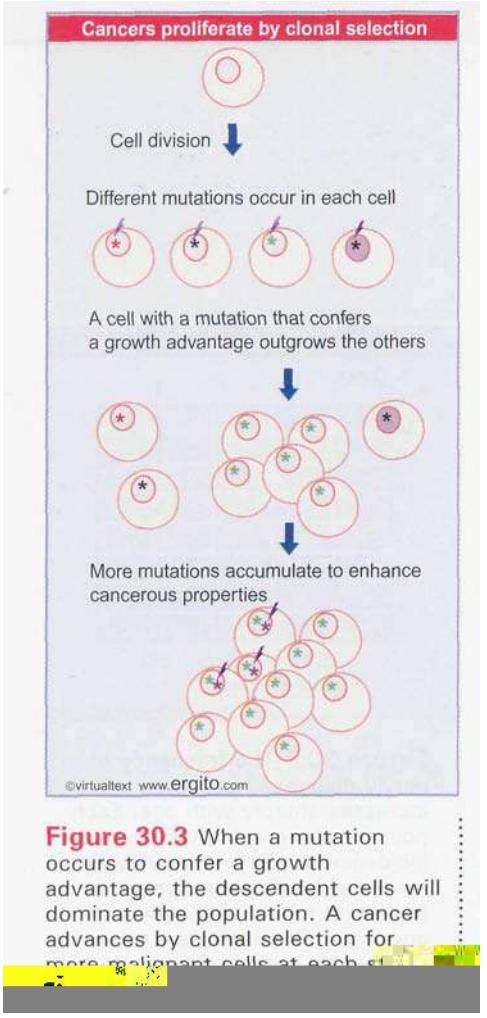


Figure 30.3 When a mutation occurs to confer a growth advantage, the descendent cells will dominate the population. A cancer advances by clonal selection for

genes. The inactivation of these genes decreases the repair of damaged DNA, and thereby increases the rate at which mutations occur.

Genetic instability is reflected in changes in the numbers of genes in cancer cells. This can be the result of small duplications or deletions, translocations of material from one chromosome to another, or even changes that affect entire chromosomes. Instability at the level of chromosome can be caused by systems that act on partitioning at mitosis.

The occurrence of different mutations creates an opportunity to select among the population for cells with particular properties. In the case of cancer, a mutation that increases the growth potential of a cell will give it a selective advantage. **Figure 30.3** illustrates the result that a cell that divides more often, perhaps because it does not respect the usual constraints on growth, will generate more descendants. At each stage during the progression of a cancer, the cell population is selected for those cells that can grow more aggressively (this meaning initially that they can grow more rapidly and later that they can migrate to start colonies in new locations).

Our current view of cancer is that it is driven by twin features: an increased rate of mutation is responsible for generating cells with altered growth properties; and the population of cells is then selected for those with an increased rate of proliferation. A cancer progresses by multiple cycles of mutation and selection.

By comparing cancer cells with normal cells, we can identify genes that have been changed by mutation. Those that have direct effects on the generation of a cancer can be divided into oncogenes (where a mutation has activated a gene whose function contributes to the tumorigenic state) and tumor suppressors (where a mutation has inactivated a gene whose function antagonizes the tumorigenic state). In most cases, a cancer arises because a series of mutations have accumulated in a somatic cell, activating oncogenes and/or inactivating tumor suppressors.

Genetic diversity in the population means that each individual may have a different set of alleles at these loci. Natural selection acts to eliminate alleles in the germline that contribute to cancer formation. However, there are some rare hereditary diseases that are caused by such alleles. In these cases, the affected individuals have a high probability of suffering from a cancer. In addition, susceptibility to cancer is influenced by many other loci, generally known as tumor modifiers. The products of these loci affect the functions of oncogenes or tumor suppressors, either directly or indirectly, but do not themselves have any direct effects.

30.2 Tumor cells are immortalized and transformed

Key Concepts

- Immortalization enables cells to overcome a limit on the number of cell divisions.
- Cultured cell lines have been immortalized.
- Transformation consists of a series of changes that release growth constraints on the immortalized cell.

Three types of changes that occur when a cell becomes tumorigenic are summarized in **Figure 30.4:**

Immortalization describes the property of indefinite cell growth (without any other changes in the phenotype necessarily occurring). Transformation describes the failure to observe the normal constraints of growth; for example, transformed cells become independent of factors usually needed for cell growth and survival.

- Metastasis describes the stage at which the cancer cell gains the ability to invade normal tissue, so that it can move away from the tissue of origin and establish a new colony elsewhere in the body.

To characterize the aberrant events that enable cells to bypass normal control and generate tumors, we need to compare the growth characteristics of normal and transformed cells *in vitro*. Transformed cells can be grown readily, but it is much more difficult to grow their normal counterparts.

When cells are taken from a vertebrate organism and placed in culture, they grow for several divisions, but then enter a senescent stage, in which growth ceases. This is followed by a crisis, in which most of the cells die. The number of divisions that occur before this happens is sometimes called the Hayflick limit, after the author who discovered the phenomenon.

The survivors that emerge from crisis are capable of dividing indefinitely, but their properties have changed in the act of emerging from crisis. This comprises the process of immortalization. (The features of crisis depend on both the species and tissue. Typically mouse cells pass through crisis at ~12 generations. Human cells enter crisis at ~40 generations, although it is rare for human cells to emerge from it, and only some types of human cells in fact can do so.)

The limitation of the life span of most cells by crisis restricts us to two options in studying nontransformed cells, neither entirely satisfactory:

- Primary cells are the immediate descendants of cells taken directly from the organism. They faithfully mimic the *in vivo* phenotype, but in most cases survive for only a relatively short period, because the culture dies out at crisis.
- Cells that have passed through crisis become established to form a (nontumorigenic) cell line. They can be perpetuated indefinitely, but their properties have changed in passing through crisis, and may indeed continue to change during adaptation to culture. These changes may partly resemble those involved in tumor formation, which reduces the usefulness of the cells.

An established cell line by definition has become immortalized, but usually is not tumorigenic. Nontumorigenic established cell lines display characteristic features similar to those of primary cultures, often including

- Anchorage dependence—a solid or firm surface is needed for the cells to attach to.
- Serum dependence (also known as growth factor dependence) — serum is needed to provide essential growth factors.
- Density-dependent inhibition—cells grow only to a limited density, because growth is inhibited, perhaps by processes involving cell-cell contacts.
- *Cytoskeletal organization*—cells are flat and extended on the surface on which they are growing, and have an elongated network of stress fibers (consisting of actin filaments).

The consequence of these properties is that the cells grow as a monolayer (that is, a layer one cell thick) on a substratum.

These properties provide parameters by which the normality of the cell may be judged. Of course, any established cell line provides only an approximation of *in vivo* control. The need for caution in analyzing the genetic basis for growth control in such lines is emphasized by the fact that almost always they suffer changes in the chromosome complement and are not true diploids. A cell whose chromosomal constitution has changed from the true diploid is said to be aneuploid.

Cells cultured from tumors instead of from normal tissues show changes in some or all of these properties. They are said to be transformed. A transformed cell grows in a much less restricted manner. It has reduced

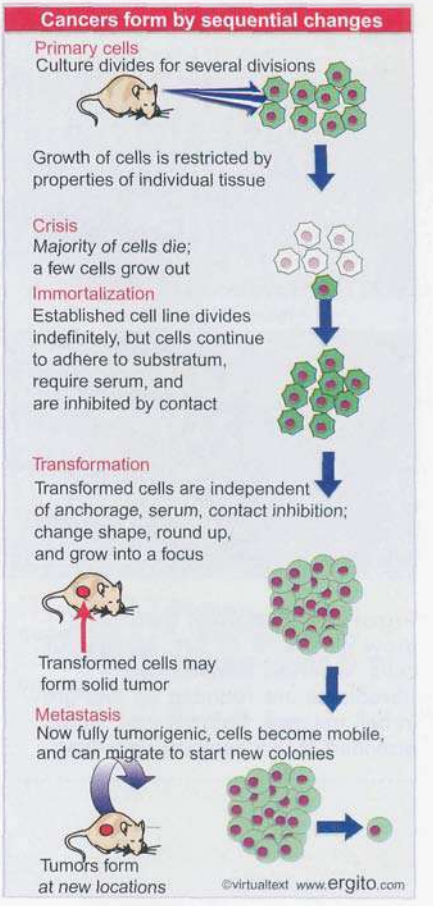


Figure 30.4 Three types of properties distinguish a cancer cell from a normal cell. Sequential changes in cultured cells can be correlated with changes in tumorigenicity.

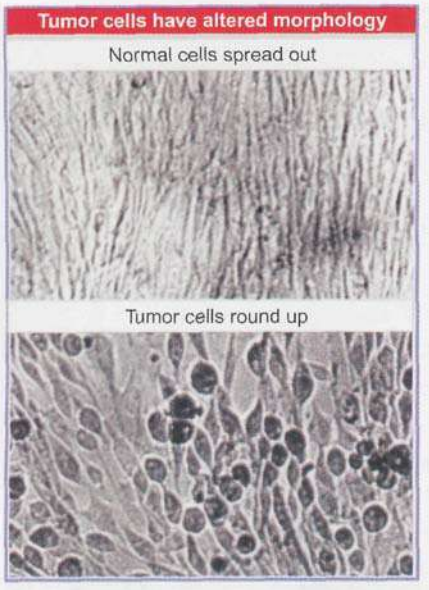


Figure 30.5 Normal fibroblasts grow as a layer of flat, spread-out cells, whereas transformed fibroblasts are rounded up and grow in cell masses. Photographs kindly provided by Hidesaburo Hanafusa.

serum-dependence, does not need to attach to a solid surface (so that individual cells "round-up" instead of spreading out) and the cells pile up into a thick mass of cells (called a focus) instead of growing as a surface monolayer. Furthermore, the cells may form tumors when injected into appropriate test animals. **Figure 30.5** compares views obtained by conventional microscopy of "normal" fibroblasts growing in culture with "transformed" variants. The difference can be seen more dramatically in the scanning electron microscope views of **Figure 30.6.**

The joint changes of immortalization and transformation of cells in culture provide a paradigm for the formation of animal tumors. By comparing transformed cell lines with normal cells, we hope to identify the genetic basis for tumor formation and also to understand the phenotypic processes that are involved in the conversion.

Certain events convert normal cells into transformed cells, and provide models for the processes involved in tumor formation. Usually multiple genetic changes are necessary to create a cancer; and sometimes tumors gain increased virulence as the result of a progressive series of changes.

A variety of agents increase the frequency with which cells (or animals) are converted to the transformed condition; they are said to be *carcinogenic*. Sometimes these carcinogens are divided into those that "initiate" and those that "promote" tumor formation, implying the existence of different stages in cancer development. Carcinogens may cause epigenetic changes or (more often) may act, directly or indirectly, to change the genotype of the cell.

30.3 Oncogenes and tumor suppressors have opposite effects

Key Concepts

- An oncogene results from a gain-of-function mutation of a proto-oncogene that generates a tumorigenic product.
- Mutation of a tumor suppressor causes a loss-of-function in the ability to restrain cell growth.

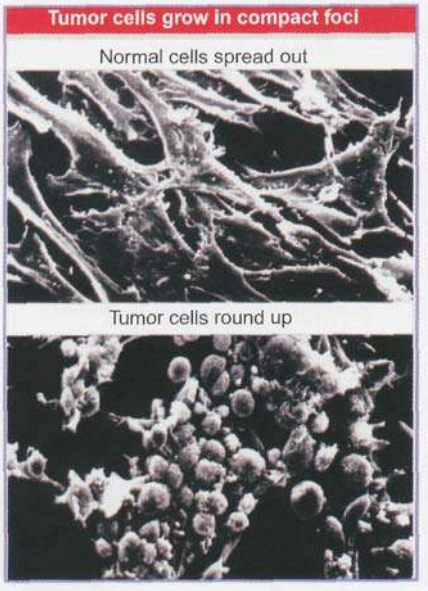


Figure 30.6 Scanning electron microscopy shows that normal cells spread out and form long processes, whereas tumor cells round up into tightly balled masses. Photographs kindly provided by J. Michael Bishop.

There are two classes of genes in which mutations cause transformation.

Oncogenes were initially identified as genes carried by viruses that cause transformation of their target cells. A major class of the viral oncogenes have cellular counterparts that are involved in normal cell functions. The cellular genes are called proto-oncogenes, and in certain cases their mutation or aberrant activation in the cell to form an oncogene is associated with tumor formation. About 100 oncogenes have been identified. The oncogenes fall into several groups, representing different types of activities ranging from transmembrane proteins to transcription factors, and the definition of these functions may therefore lead to an understanding of the types of changes that are involved in tumor formation.

The generation of an oncogene represents a gain-of-function in which a cellular proto-oncogene is inappropriately activated. This can involve a mutational change in the protein, or constitutive activation, overexpression, or failure to turn off expression at the appropriate time. The simple case of a somatic mutation is illustrated in **Figure 30.7.**

Tumor suppressors are detected by deletions (or other inactivating mutations) that are tumorigenic. The mutations represent loss-of-function in genes that usually impose some constraint on the cell cycle or cell growth; the release of the constraint is tumorigenic. It is necessary for both copies of the gene to be inactivated.

The most compelling evidence for the nature of tumor suppressors is provided by certain hereditary cancers, in which patients with the disease develop tumors that have lost both alleles, and therefore lack an active gene. There is also now evidence that changes in these genes may be associated with the progression of a wide range of cancers. About 10 tumor suppressors are known at present.

30.4 Transforming viruses carry oncogenes

Key Concepts

- A transforming virus usually carries a specific gene(s) that is responsible for transforming the target cell by changing its growth properties.

Transformation may occur spontaneously, may be caused by certain chemical agents, and, most notably, may result from infection with tumor viruses. There are many classes of tumor viruses, including both DNA and RNA viruses, and they occur widely in the avian and animal kingdoms.

The transforming activity of a tumor virus resides in a particular gene or genes carried in the viral genome. Oncogenes were given their name by virtue of their ability to convert cells to a tumorigenic (or oncogenic) state. An oncogene initiates a series of events that is executed by cellular proteins. In effect, the virus throws a regulatory switch that changes the growth properties of its target cell.

Figure 30.8 summarizes the general properties of the major classes of transforming viruses. The oncogenes carried by the DNA viruses specify proteins that inactivate tumor suppressors, so their action in part mimics loss-of-function of the tumor suppressors. The oncogenes carried by retroviruses are derived from cellular genes and therefore may mimic the behavior of gain-of-function mutations in animal proto-oncogenes.

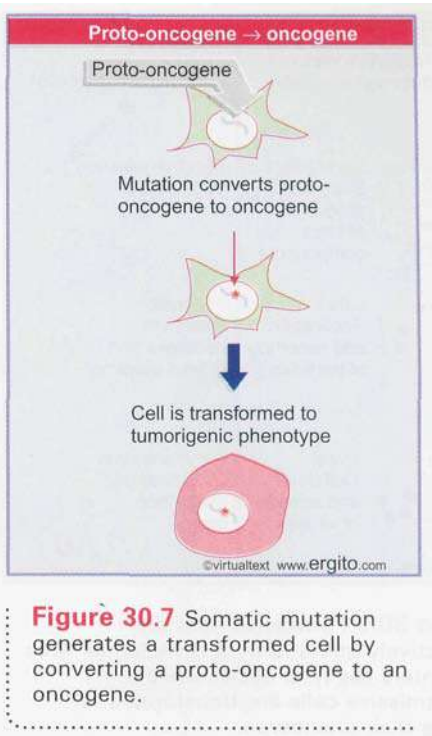


Figure 30.7 Somatic mutation generates a transformed cell by converting a proto-oncogene to an oncogene.

DNA and RNA oncogenic viruses carry different types of oncogenes

Viral Class	Genome	Size	Oncogenes	Origin of Oncogene	Action of Oncogene
Polyoma	dsDNA	5-6 kb	T antigens	Early viral gene	inactivates tumor suppressor
HPV	dsDNA	~8 kb	E6 & E7	Early viral gene	inactivates tumor suppressor
Adeno	dsDNA	~37 kb	E1A & E1B	Early viral gene	inactivates tumor suppressor
Retrovirus (acute)	ssRNA	6-9 kb	Individual	Cellular	activates oncogenic pathway

©virtualtext www.ergito.com

Figure 30.8 The oncogenes of DNA transforming viruses are early viral functions, whereas the oncogenes of retroviruses are modified from cellular genes.

30.5 Early genes of DNA transforming viruses have multifunctional oncogenes

Key Concepts

- The oncogenes of DNA transforming viruses are early viral functions.
- The oncogene becomes integrated into the host cell genome and is expressed constitutively.
- The oncogenes of polyomaviruses are T antigens, which are expressed by alternative splicing from a single locus.
- Adenoviruses express several E1A and E1B proteins from two genes.

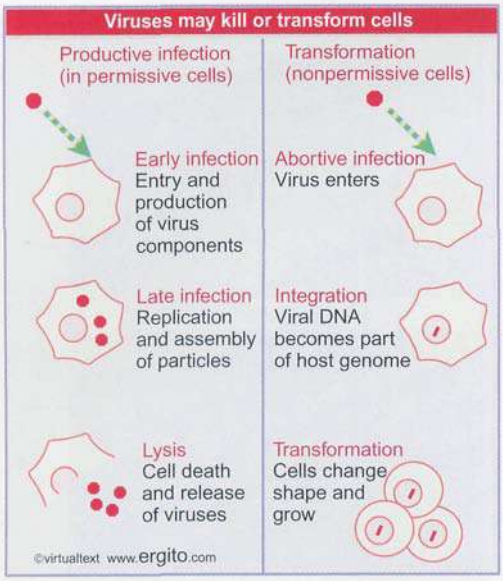


Figure 30.9 Permissive cells are productively infected by a DNA tumor virus that enters the lytic cycle, while nonpermissive cells are transformed to change their phenotype.

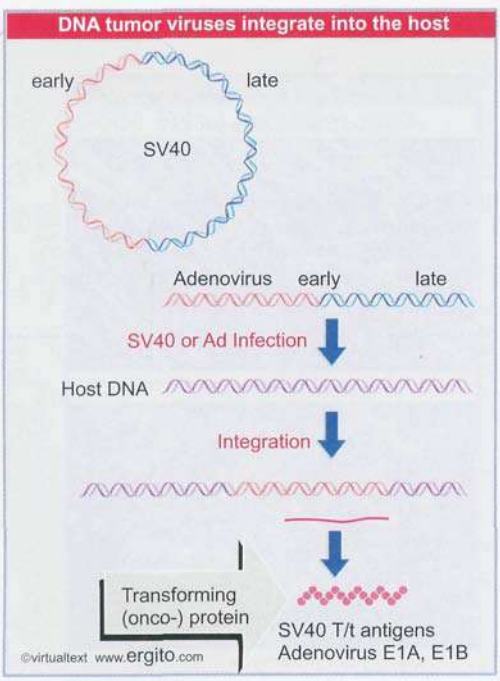


Figure 30.10 Cells transformed by polyomaviruses or adenoviruses have viral sequences that include the early region integrated into the cellular genome. Sites of integration are random.

Polyomaviruses and adenoviruses have been isolated from a variety of mammals. Although perpetuated in the wild in a single host species, a virus may be able to grow in culture on a variety of cells from different species. The response of a cell to infection depends on its species and phenotype and falls into one of two classes, as illustrated in **Figure 30.9:**

- Permissive cells are productively infected. The virus proceeds through a lytic cycle that is divided into the usual early and late stages. The cycle ends with release of progeny viruses and (ultimately) cell death.
- Nonpermissive cells cannot be productively infected, and viral replication is abortive. Some of the infected cells are transformed; in this case, the phenotype of the individual cell changes and the culture is perpetuated in an unrestrained manner.

A common mechanism underlies transformation by DNA tumor viruses. *Oncogenic potential resides in a single function or group of related functions that are active early in the viral lytic cycle. When transformation occurs, the relevant gene(s) are integrated into the genomes of transformed cells and expressed constitutively.* This suggests the general model for transformation by these viruses illustrated in **Figure 30.10,** in which the constitutive expression of the oncogene generates transforming protein(s) (oncoproteins).

Polyomaviruses are small. Polyomavirus itself is common in mice, the analogous virus SV40 (simian virus 40) was isolated from rhesus monkey cells, and more recently the human viruses BK and JC have been characterized. All of the polyomaviruses can cause tumors when injected into newborn rodents.

During a productive infection, the early region of each virus uses alternative splicing to synthesize overlapping proteins called T antigens. (The name reflects their isolation originally as the proteins found in tumor cells.) The various T antigens have a variety of functions in the lytic cycle. They are required for expression of the late region and for DNA replication of the virus.

Cells transformed by polyomaviruses contain integrated copies of part or all of the viral genome. The integrated sequences always include the early region. The T antigens have transforming activity, which rests upon their ability to interact with cellular proteins. This is independent of their ability to interact directly with the viral genome. SV40 requires "big T" and "little t" antigens, and polyoma requires "T" and "middle T" antigens for transformation.

Papillomaviruses are small DNA viruses that cause epithelial tumors; there are ~75 human papillomaviruses (HPVs); most are associated with benign growths (such as warts), but some are associated with cancers, in particular cervical cancers. Two virus-associated products are expressed in cervical cancers; these are the E6 and E7 proteins, which can immortalize target cells.

Adenoviruses were originally isolated from human adenoids; similar viruses have since been isolated from other mammals. They comprise a large group of related viruses, with >80 individual members. Human adenoviruses remain the best characterized, and are associated with respiratory diseases. They can infect a range of cells from different species.

Human cells are permissive and are therefore productively infected by adenoviruses, which replicate within the infected cell. But cells of some rodents are nonpermissive. All adenoviruses can transform nonpermissive cultured cells, but the oncogenic potential of the viruses varies; the most effective can cause tumors when they are injected into newborn rodents. The genomes of cells transformed by adenoviruses have gained a part of the early viral region that contains the E1A and E1B genes, which code for several nuclear proteins.

By Book_Crazy [IND]

Epstein-Barr is a human herpes virus associated with a variety of diseases, including infectious mononucleosis, nasopharyngeal carcinoma, African Burkitt lymphoma, and other lymphoproliferative disorders. EBV has a limited host range for both species and cell phenotype. Human B lymphocytes that are infected *in vitro* become immortalized, and some rodent cell lines can be transformed. Viral DNA is found in transformed cells, although it has been controversial whether it is integrated. It remains unclear exactly which viral genes are required for transformation.

30.6 Retroviruses activate or incorporate cellular genes

Key Concepts

- Acute transforming retroviruses have oncogenes that are derived from cellular genes.
- Nondefective transforming viruses do not have oncogenes, but activate an equivalent gene(s) in the host genome.

Retroviruses present a different situation from the DNA tumor viruses. They can transfer genetic information both horizontally and vertically, as illustrated in **Figure 30.11.** Horizontal transfer is accomplished by the normal process of viral infection, in which increasing numbers of cells become infected in the same host. Vertical transfer results whenever a virus becomes integrated in the germline of an organism as an endogenous provirus; like a lysogenic bacteriophage, it is inherited as a Mendelian locus by the progeny (see *12 Phage strategies).*

The retroviral life cycle propagates genetic information through both RNA and DNA templates. A retroviral infection proceeds through the stages illustrated previously in Figure 17.2, in which the RNA is reverse-transcribed into single-stranded DNA, then converted into double-stranded DNA, and finally integrated into the genome, where it may be transcribed again into infectious RNA. Integration into the genome leads to vertical transmission of the provirus. Expression of the provirus may generate retroviral particles that are horizontally transmitted. Integration is a normal part of the life cycle of every retrovirus, whether it is nontransforming or transforming.

The tumor retroviruses fall into two general groups with regard to the origin of their tumorigenicity:

- Nondefective viruses follow the usual retroviral life cycle. They provide infectious agents that have a long latent period, and often are associated with the induction of leukemias. Two classic models are FeLV (feline leukemia virus) and MMTV (mouse mammary tumor virus). *Tumorigenicity does not rely upon an individual viral oncogene, but upon the ability of the virus to activate a cellular proto-oncogene(s).*
- Acute transforming viruses *have gained new genetic information in the form of an oncogene.* This gene is not present in the ancestral (nontransforming virus); it originated as a cellular gene that was captured by the virus by means of a transduction event during an infective cycle. These viruses usually induce tumor formation *in vivo* rather rapidly, and they can transform cultured cells *in vitro.* Reflecting the fact that each acute transforming virus has specificity toward a particular type of target cell, these viruses are divided into classes

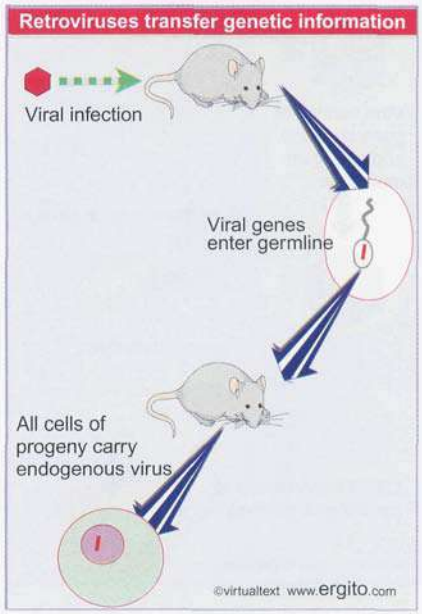


Figure 30.11 Retroviruses transfer genetic information horizontally by infecting new hosts; information is inherited vertically if a virus integrates in the genome of the germline.

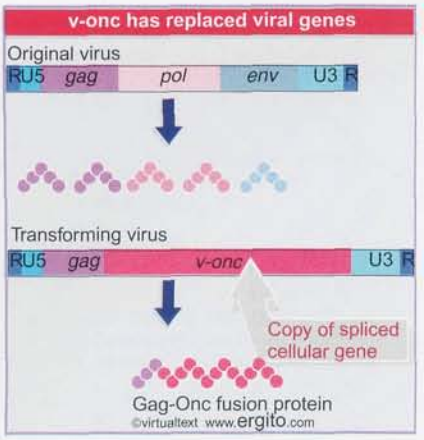


Figure 30.12 A transforming retrovirus carries a copy of a cellular sequence in place of some of its own gene(s).

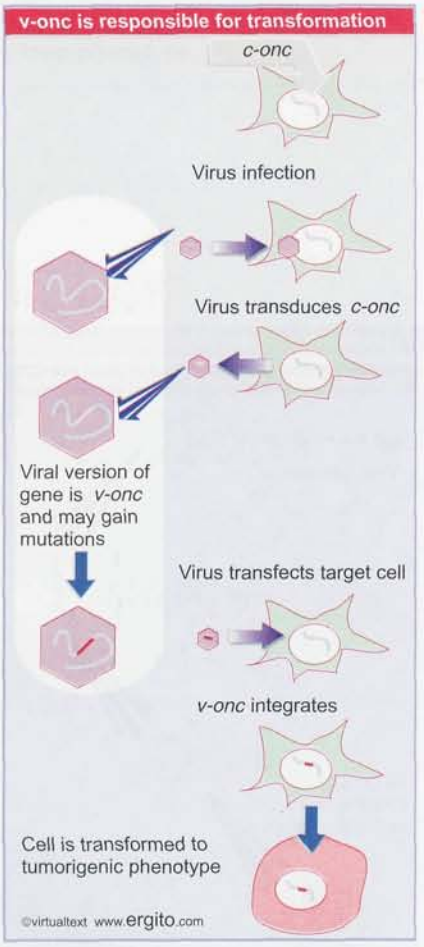


Figure 30.13 A retrovirus may incorporate a cellular proto-oncogene during an infection. Subsequent mutation or other changes in the proto-oncogene make it oncogenic when the virus transfers it to a new cell in another infectious cycle.

according to the type of tumor that is caused in the animal: leukemia, sarcoma, carcinoma, etc.

When a retrovirus captures a cellular gene by exchanging part of its own sequence for a cellular sequence (see *17.6 Retroviruses may transduce cellular sequences)*, it generates the structure summarized in **Figure 30.12.** Some of the original retroviral sequences (which are usually organized into the genes *gag-pol-env*, coding for coat proteins, reverse transcriptase, and other enzyme activities) are replaced by a sequence derived by reverse transcription of a cellular mRNA. This type of event is rare, but creates a transducing virus that has two important properties:

- Usually it cannot replicate by itself, because viral genes needed for reproduction have been lost by the exchange with cellular sequences. So almost all of these viruses are replication-defective. But they can propagate in a simultaneous infection with a wild-type "helper" virus that provides the functions that were lost in the recombination event. (RSV is an exceptional transducing virus that retains the ability to replicate.)
- During an infection, the transducing virus carries with it the cellular gene(s) that were obtained in the recombination event, and their expression may alter the phenotype of the infected cell. Any transducing virus whose cellular genetic information assists the growth of its target cells could have an advantage in future infective cycles. If a virus gains a gene whose product stimulates cell growth, the acquisition may enable the virus to spread by stimulating the growth of the particular cells that it infects. This is important also because a retrovirus can replicate only in a proliferating cell. After a virus has incorporated a cellular gene, the gene may gain mutations that enhance its ability to influence cell phenotype.

Of course, transformation is not the only mechanism by which retroviruses affect their hosts. A notable example is the HIV-1 retrovirus, which belongs to the retroviral group of lentiviruses. The virus infects and kills T lymphocytes carrying the CD4 receptor, devastating the immune system of the host, and inducing the disease of AIDS. The virus carries the usual *gag-pol-env* regions, and also has an additional series of reading frames, which overlap with one another, to which its lethal actions are attributed.

30.7 Retroviral oncogenes have cellular counterparts

Key Concepts

- A retroviral oncogene is derived by capturing a proto-oncogene from a host genome.

New sequences that are present in an acute transforming retrovirus can be delineated by comparing the sequence of the virus with that of the parental (nontumorigenic) virus. Usually the new region is closely related to a sequence in the cellular genome. The normal cellular sequence itself is not oncogenic—if it were, the organism could scarcely have survived—but it defines a proto-oncogene.

The general model for retroviral transformation is illustrated in **Figure 30.13.** The virus gains a copy of a proto-oncogene from a cellular genome. Sometimes the copy is different from the cellular sequence,

typically because it has been truncated. In some cases, the difference is sufficient to convert the proto-oncogene into an oncogene. In other cases, mutations occur in the viral sequence that convert the copy into an oncogene.

The viral oncogenes and their cellular counterparts are described by using prefixes v for viral and c for cellular. So the oncogene carried by Rous sarcoma virus is called *v-src*, and the proto-oncogene related to it in cellular genomes is called *c-src*. Comparisons between *v-onc* and *c-onc* genes can be used to identify the features that confer oncogenicity.

Oncogenes of some retroviruses are summarized in **Figure 30.14.** The type of tumor results from the combination of the particular oncogene with the time and place in which it is expressed. It is striking that usually the oncogenic activity resides in a single gene. AEV is one of a very few exceptions in which a retrovirus carries more than one oncogene.

More than 30 *c-onc* genes have been identified so far by their representation in retroviruses. Sometimes the same *c-onc* gene is represented in different transforming viruses; for example, the monkey virus SSV and the PI strain of the feline virus FeSV both carry a *v-onc* derived from *c-sis*. Some viruses carry related *v-onc* genes, such as in the Harvey and Kirsten strains of MuSV, which carry *v-ras* genes derived from two different members of the cellular *c-ras* gene family. In other cases the *v-onc* genes of related viruses represent unrelated cellular progenitors; for example, three different isolates of FeSV may have been derived from the same original (nontransforming) virus, but have transduced the *sis*, *fins*, and *fes* oncogenes. The events involved in formation of a transducing virus can be complex; some viruses include sequences derived from more than one cellular gene.

Given the rarity of the transducing event, it is significant that multiple independent isolates occur representing the same *c-onc* gene. For example, several viruses carry *v-myc* genes. They are all derived from a single *c-myc* gene, but the *v-myc* genes differ in their exact ends and in individual point mutations. The identification of such isolates probably means that we have identified most of the genes of the *c-onc* type that can be activated by viral transduction.

Direct evidence that expression of the *v-onc* sequence accomplishes transformation was first obtained with RSV. Temperature-sensitive mutations in *v-src* allow the transformed phenotype to be reverted by increase in temperature, and regained by decrease in temperature. This shows clearly that in this case the *v-src* gene is needed both to initiate and maintain the transformed state.

Transforming properties of retroviruses are determined by their v-onc genes

Virus	Name	Species	Tumor	Oncogene
Rous sarcoma	RSV	chicken	sarcoma	*src*
Harvey murine sarcoma	Ha-MuSV	rat	sarcoma & erythroleukemia	*H-ras*
Kirsten murine sarcoma	Ki-MuSV	rat	sarcoma & erythroleukemia	*K-ras*
Moloney murine sarcoma	Mo-MuSV	mouse	sarcoma	*mos*
FBJ murine osteosarcoma	FBJ-MuSV	mouse	chondrosarcoma	*fos*
Simian sarcoma	SSV	monkey	sarcoma	*sis*
Feline sarcoma	PI-FeSV	cat	sarcoma	*sis*
Feline sarcoma	SM-FeSV	cat	fibrosarcoma	*fms*
Feline sarcoma	ST-FeSV	cat	fibrosarcoma	*fes*
Avian sarcoma	ASV-17	chicken	fibrosarcoma	*jun*
Fujinami sarcoma	FuSV	chicken	sarcoma	*fps*
Avian myelocytomatosis	MC29	chicken	carcinoma, sarcoma, & myelocytoma	*myc*
Abelson leukemia	MuLV	mouse	B cell lymphoma	*abl*
Reticuloendotheliosis	REV-T	turkey	lymphatic leukemia	*rel*
Avian erythroblastosis	AEV	chicken	erythroleukemia & fibrosarcoma	*erbB,A*
Avian myeloblastosis	AMV	chicken	myeloblastic leukemia	*myb*

Figure 30.14 Each transforming retrovirus carries an oncogene derived from a cellular gene. Viruses have names and abbreviations reflecting the history of their isolation and the types of tumor they cause. This list shows some representative examples of the retroviral oncogenes.

30.8 Quantitative or qualitative changes can explain oncogenicity

Key Concepts

- An oncogene usually has qualitatively different (transforming) effects from the proto-oncogene as the result of mutational changes.
- Sometimes the oncogene is transforming because it is expressed at higher levels than the proto-oncogene.

Two general types of theory might explain the difference in properties between *v-onc* genes and *c-onc* genes:

- A quantitative model proposes that viral genes are functionally indistinguishable from the cellular genes, but are oncogenic because they are expressed in much greater amounts or in inappropriate cell types, or because their expression cannot be switched off.
- A qualitative model supposes that the *c-onc* genes intrinsically lack oncogenic properties, but may be converted by mutation into oncogenes whose devastating effects reflect the acquisition of new properties (or loss of old properties).

How closely related are *v-onc* genes to the corresponding *c-onc* genes? In some cases, the only changes are a very small number of point mutations. The *mos*, *sis*, and *myc* genes offer examples in which the entire *c-onc* gene has been gained by the virus; in this case, the small number of amino acid substitutions do not seem to affect function of the protein, and in fact are not required for transforming activity. So the *v-onc* product is likely to fulfill the same enzymatic or other functions as the *c-onc* product, but with some change in its regulation; in these cases, overexpression is responsible for oncogenicity. A good example is *c-myc*, where oncogenicity may be caused by overexpression either by a *v-myc* gene carried by a transforming retrovirus or by changes in the cellular genome that cause overexpression of *c-myc*.

Two cases in which point mutations play a critical role in creating an oncogenic protein are presented by *ras* and *src*.

In the case of *ras*, changes in the regulation of Ras activity that activate the protein can be directly attributed to the individual point mutations that have occurred in the *v-onc* gene. Overexpression of *c-ras* may have weak oncogenic effects, but full oncogenicity requires sequence changes in the protein.

In some cases, a *v-onc* gene is truncated by the loss of sequences from the N-terminus or C-terminus (or both) of the *c-onc* gene, probably as a result of the sites involved in the recombination event that generated it. Loss of these regions may remove some regulatory constraint that normally limits the activity of the *c-onc* product. Such sequence changes are required for oncogenicity of *src*. *v-src* is oncogenic at low levels of protein, but *c-src* is not oncogenic at high protein levels (>10X normal). The viral and cellular *src* genes are coextensive, but *v-src* has replaced the C-terminal 19 amino acids of *c-src* with a different sequence of 12 amino acids. This has an important regulatory consequence in activating the Src protein constitutively. In cases where *v-onc* genes are truncations of *c-onc* genes, point mutations may also contribute to the oncogenicity of the *v-onc* product. In the case of Src, changes in two tyrosine residues that are targets for phosphorylation have strong effects on oncogenicity (see *30.16 Src is the prototype for the proto-oncogenic cytoplasmic tyrosine kinases).*

By Book_Crazy [IND]

The characterization of transforming retroviruses played an important role in the definition of oncogenes. However, most events involved in human cancers do not involve viral intermediates, and other mechanisms are responsible for generating oncogenes. But the concept that oncogenes arise by activation of proto-oncogenes is an important paradigm for animal cancers.

30.9 Ras oncogenes can be detected in a transfection assay

Key Concepts

- Transformed cells can be distinguished from normal cells by the formation of foci in a culture dish.
- DNA extracted from tumor cells can transform 3T3 target cells.
- A transforming cellular *(c-onc)* gene often has a homologue *(v-onc)* in a transforming retrovirus.
- The *ras* genes are the most common transforming genes identified by this method.

Some oncogenes can be detected by using a direct assay for transformation in which "normal" recipient cells are transfected with DNA obtained from animal tumors. The procedure is illustrated in **Figure 30.15.** The established mouse NIH 3T3 fibroblast line usually is used as a recipient. Historically these experiments started by using DNA extracted *en masse*, but now they are usually performed with a purified oncogene. The ability of any individual gene to convert wild-type cells into the transformed state constitutes one form of proof that it is an oncogene. Another assay that can be used is to inject cells into "nude" mice (which lack the ability to reject such transplants immunologically). The ability to form tumors can then be measured directly in the animal.

When a cell is transformed in a 3T3 culture (or some other "normal" culture), its descendants pile up into a focus. The appearance of foci is used as a measure of the transforming ability of a DNA preparation. Starting with a preparation of DNA isolated from tumor cells, the efficiency of focus formation is low. However, once the transforming gene has been isolated and cloned, greater efficiencies can be obtained. In fact, the transforming "strength" of a gene can be characterized by the efficiency of focus formation by the cloned sequence.

DNA with transforming activity can be isolated only from tumorigenic cells; it is not present in normal DNA. The transforming genes isolated by this assay have two revealing properties:

- *They have closely related sequences in the DNA of normal cells.* This argues that transformation was caused by mutation of a normal cellular gene (a proto-oncogene) to generate an oncogene. The change may take the form of a point mutation or more extensive reorganization of DNA around the *c-onc* gene.
- *They may have counterparts in the oncogenes carried by known transforming viruses.* This suggests that the repertoire of proto-oncogenes is limited, and probably the same genes are targets for mutations to generate oncogenes in the cellular genome or to become viral oncogenes.

Oncogenes derived from the *c-ras* family are often detected in the transfection assay. The family consists of several active genes in both man and rat, dispersed in the genome. (There are also some pseudogenes.) The individual genes, *N-ras*, *H-ras*, and *K-ras*, are closely related, and code for protein products ~21 kD and known collectively as $p21^{ras}$.

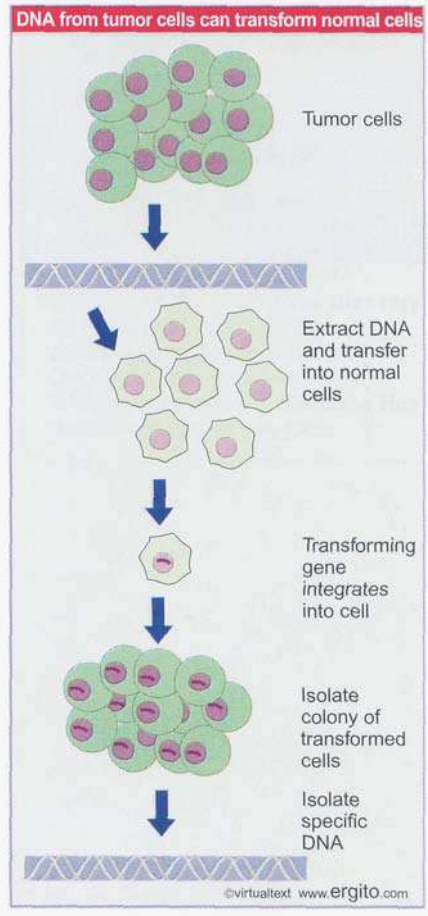


Figure 30.15 The transfection assay allows (some) oncogenes to be isolated directly by assaying DNA of tumor cells for the ability to transform normal cells into tumorigenic cells.

The *H-ras* and *K-ras* genes have *v-ras* counterparts, carried by the Harvey and Kirsten strains of murine sarcoma virus, respectively (see Figure 30.14). Each *v-ras* gene is closely related to the corresponding *c-ras* gene, with only a few individual amino acid substitutions. The Harvey and Kirsten virus strains must have originated in independent recombination events in which a progenitor virus gained the corresponding *c-ras* sequence.

30.10 Ras proto-oncogenes can be activated by mutation at specific positions

Key Concepts

- *v-ras* genes are derived by point mutations of *c-ras* genes.
- The same mutations occur in the *v-ras* genes of transforming viruses and the mutant *c-ras* genes of tumor cells.
- Almost any mutation at either position 12 or 61 converts a *c-ras* proto-oncogene into a transforming variety.
- The effect of the mutations is to increase Ras activity by inhibiting the hydrolysis of bound GTP to GDP.

Oncogenic variants of the *c-ras* genes are found in transforming DNA preparations obtained from various primary tumors and tumor cell lines. Each of the *c-ras* proto-oncogenes can give rise to a transforming oncogene by a single base mutation. *The mutations in several independent human tumors cause substitution of a single amino acid, most commonly at position 12 or 61, in one of the Ras proteins.*

Position 12 is one of the residues that is mutated in the *v-H-ras* and *v-K-ras* genes. *So mutations occur at the same positions in* v-ras *genes in retroviruses and in mutant* c-ras *genes in multiple rat and human tumors.* This suggests that the normal c-Ras protein can be converted into a tumorigenic form by a mutation in one of a few codons in rat or man (and perhaps any mammal).

The general principle established by this work is that *substitution in the coding sequence can convert a cellular proto-oncogene into an oncogene.* Such an oncogene can be associated with the appearance of a spontaneous tumor in the organism. It may also be carried by a retrovirus, in which case a tumor is induced by viral infection.

The *ras* genes appear to be finely balanced at the edge of oncogenesis. Almost any mutation at either position 12 or 61 can convert a *c-ras* proto-oncogene into an active oncogene:

- All three *c-ras* genes have glycine at position 12. If it is replaced *in vitro* by any other of the 19 amino acids except proline, the mutated *c-ras* gene can transform cultured cells. The particular substitution influences the strength of the transforming ability.
- Position 61 is occupied by glutamine in wild-type *c-ras* genes. Its change to another amino acid usually creates a gene with transforming potential. Some substitutions are less effective than others; proline and glutamic acid are the only substitutions that have no effect.

When the expression of a normal *c-ras* gene is increased, either by placing it under control of a more active promoter or by introducing multiple copies into transfected cells, recipient cells are transformed. Some mutant *c-ras* genes that have changes in the protein sequence also have a mutation in an intron that increases the level of expression (by increasing processing of mRNA ~10X). Also, some tumor lines have amplified *ras* genes. A 20-fold increase in the level of a nontransforming Ras protein is sufficient to allow the transformation of some cells.

By Book_Crazy [IND]

The effect has not been fully quantitated, but it suggests the general conclusion that oncogenesis depends on over-activity of Ras protein, and is caused either by increasing the amount of protein or (more efficiently) by mutations that increase the activity of the protein.

Transfection by DNA can be used to transform only certain cell types. Limitations of the assay explain why relatively few oncogenes have been detected by transfection. This system has been most effective with *ras* genes, where there is extensive correlation between mutations that activate *c-ras* genes in transfection and the occurrence of tumors.

Ras is a monomeric guanine nucleotide-binding protein that is active when bound to GTP and inactive when bound to GDP. It has an intrinsic GTPase activity. Figure 30.16 reviews the discussion of *28.15 The activation of Ras is controlled by GTP* in which we saw that the conversion between the two forms of Ras is catalyzed by other proteins. GAP proteins stimulate the ability of Ras to hydrolyze GTP, thus converting active Ras into inactive Ras. GEF proteins stimulate the replacement of GDP by GTP, thus reactivating the protein.

Constitutive activation of Ras could be caused by mutations that allow the GDP-bound form of Ras to be active or that prevent hydrolysis of GTP. What are the effects of the mutations that create oncogenic *ras* genes? Many mutations that confer transforming activity inhibit the GTPase activity. GAP cannot increase the GTPase activity of Ras proteins that have been activated by oncogenic mutations. In other words, Ras has become refractory to the interaction with GAP that turns off its activity. Inability to hydrolyze GTP causes Ras to remain in a permanently activated form; its continued action upon its target protein is responsible for its oncogenic activity.

This establishes an important principle: *constitutive activation of a cellular protein may be oncogenic*. In the case of Ras, its effects result from activating the ERK MAP kinase pathway and (possibly) other pathways. The level of expression is finely balanced, since overstimulation of Ras by either increase in expression or mutation of the protein has oncogenic consequences (although mutation is required for a full effect).

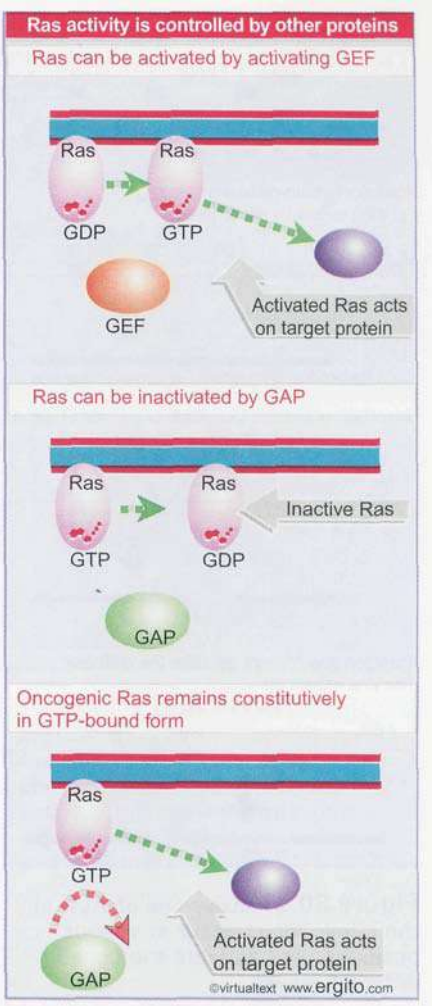


Figure 30.16 Pathways that rely on Ras could function by controlling either GEF or GAP. Oncogenic Ras mutants are refractory to control, because Ras remains in the active form.

30.11 Nondefective retroviruses activate proto-oncogenes

Key Concepts

- Increased expression of *c-myc* is transforming.
- *c-myc* can be activated by insertion of a nondefective retrovirus near the gene.

A variety of genomic changes can activate proto-oncogenes, sometimes involving a change in the target gene itself, sometimes activating it without changing the protein product. Insertion, translocation, and amplification can be causative events in tumorigenesis.

Many tumor cell lines have visible regions of chromosomal amplification, as shown by homogeneously staining regions (see Figure 18.32) or double minute chromosomes (see Figure 18.33). The amplified region may include an oncogene. Examples of oncogenes that are amplified in various tumors include *c-myc*, *c-abl*, *c-myb*, *c-erbB*, *c-K-ras*, and *Mdm2*.

Established cell lines are prone to amplify genes (along with other karyotypic changes to which they are susceptible). The presence of known oncogenes in the amplified regions, and the consistent amplification of particular oncogenes in many independent tumors of the same type, strengthens the correlation between increased expression and tumor growth.

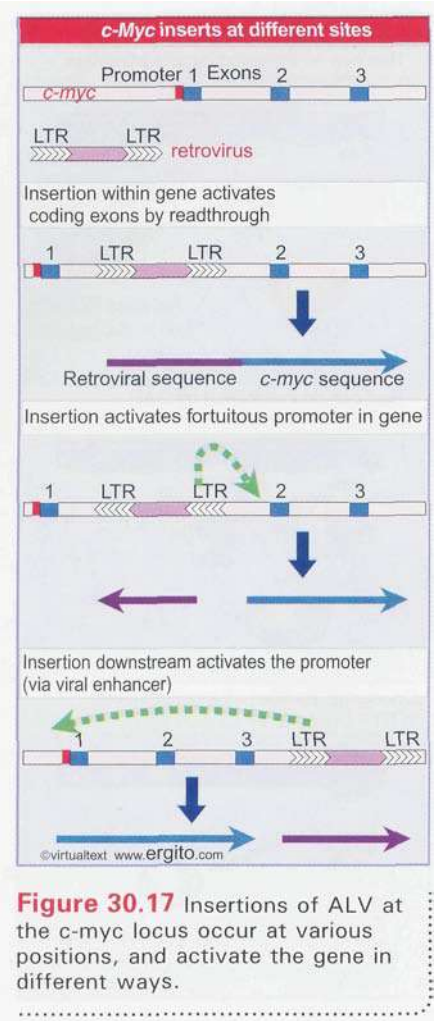


Figure 30.17 Insertions of ALV at the c-myc locus occur at various positions, and activate the gene in different ways.

Some proto-oncogenes are activated by events that change their expression, but which leave their coding sequence unaltered. The best characterized is *c-myc*, whose expression is elevated by several mechanisms. One common mechanism is the insertion of a nondefective retrovirus in the vicinity of the gene.

The ability of a retrovirus to transform without expressing a *v-onc* sequence was first noted during analysis of the bursal lymphomas caused by the transformation of B lymphocytes with avian leukemia virus. Similar events occur in the induction of T-cell lymphomas by murine leukemia virus. In each case, the transforming potential of the retrovirus is due to the ability of its LTR (the long terminal repeat of the integrated form) to cause expression of cellular gene(s).

In many independent tumors, the virus has integrated into the cellular genome within or close to the *c-myc* gene. **Figure 30.17** summarizes the types of insertions. The retrovirus may be inserted at a variety of locations relative to the *c-myc* gene.

The gene consists of three exons; the first represents a long nontranslated leader, and the second two code for the c-Myc protein. The simplest insertions to explain are those that occur within the first intron. The LTR provides a promoter, and transcription reads through the two coding exons. Transcription of *c-myc* under viral control differs from its usual control: the level of expression is increased (because the LTR provides an efficient promoter); expression cannot be switched off in B or T cells in response to the usual differentiation signals; and the transcript lacks its usual nontranslated leader (which may usually limit expression). All of these changes add up to increased constitutive expression.

Activation of *c-myc* in the other two classes of insertions reflects different mechanisms. The retroviral genome may be inserted within or upstream of the first intron, but in reverse orientation, so that its promoter points in the wrong direction. The retroviral genome also may be inserted downstream of the *c-myc* gene. In these cases, the enhancer in the viral LTR may be responsible for activating transcription of c-Myc, either from its normal promoter or from a fortuitous promoter.

In all of these cases, the coding sequence of c-myc is unchanged, so oncogenicity is attributed to the loss of normal control and increased expression of the gene.

Other oncogenes that are activated in tumors by the insertion of a retroviral genome include *c-erbB, c-myb, c-mos, c-H-ras,* and *c-raf* Up to 10 other cellular genes (not previously identified as oncogenes by their presence in transforming viruses) are implicated as potential oncogenes by this criterion. The best characterized among this latter class are *wnt1* and *int2*. The *wntl* gene codes for a protein involved in early embryogenesis that is related to the *wingless* gene of *Drosophila; int2* codes for an FGF (fibroblast growth factor).

30.12 Proto-oncogenes can be activated by translocation

Key Concepts

- *c-myc* can be activated in lymphocytes by translocations involving the Ig or TCR loci, giving B cell or T cell tumors.

Translocation to a new chromosomal location is another of the mechanisms by which oncogenes are activated. A reciprocal translocation occurs when an illegitimate recombination occurs between two chromosomes as illustrated in **Figure 30.18.** The involvement of such

By Book_Crazy [IND]

events in tumorigenesis was discovered via a connection between the loci coding immunoglobulins and the occurrence of certain tumors. Specific chromosomal translocations are often associated with tumors that arise from undifferentiated B lymphocytes. The common feature is that an oncogene on one chromosome is brought by translocation into the proximity of an Ig locus on another chromosome. Similar events occur in T lymphocytes to bring oncogenes into the proximity of a TCR locus.

In both man and mouse, the nonimmune partner is often the *c-myc* locus. In man, the translocations in B-cell tumors usually involve chromosome 8, which carries *c-myc,* and chromosome 14, which carries the IgH locus; ~10% involve chromosome 8 and either chromosome 2 (kappa locus) or chromosome 22 (lambda locus). The translocations in T-cell tumors often involve chromosome 8, and either chromosome 14 (which has the TCRα locus at the other end from the Ig locus) or chromosome 7 (which carries the TCRβ locus). Analogous translocations occur in the mouse.

Translocations in B cells fall into two classes, reflecting the two types of recombination that occur in immunoglobulin genes. One type is similar to those involved in the somatic recombination that generates the active genes. These events involve the consensus sequences used for V-D-J recombination. These can occur at all the Ig loci. In the other type, the translocation occurs at a switching site at the IgH locus, presumably reflecting the operation of the system for class switching.

When *c-myc* is translocated to the Ig locus, its level of expression is usually increased. The increase varies considerably among individual tumors, generally being in the range from 2-10X. Why does translocation activate the *c-myc* gene? The event has two consequences: *c-myc* is brought into a new region, one in which an Ig or TCR gene was actively expressed; and the structure of the *c-myc* gene may itself be changed (but usually not involving the coding regions). It seems likely that several different mechanisms can activate the *c-myc* gene in its new location (just as retroviral insertions activate *c-myc* in a variety of ways).

The correlation between the tumorigenic phenotype and the activation of *c-myc* by either insertion or translocation suggests that continued high expression of c-Myc protein is oncogenic. Expression of *c-myc* must be switched off to enable immature lymphocytes to differentiate into mature B and T cells; failure to turn off *c-myc* maintains the cells in the undifferentiated (dividing) state.

The oncogenic potential of *c-myc* has been demonstrated directly by the creation of transgenic mice. Mice carrying a *c-myc* gene linked to a B lymphocyte-specific enhancer (the IgH enhancer) develop lymphomas. The tumors represent both immature and mature B lymphocytes, suggesting that overexpression of *c-myc* is tumorigenic throughout the B cell lineage. Transgenic mice carrying a *c-myc* gene under the control of the LTR from a mouse mammary tumor virus, however, develop a variety of cancers, including mammary carcinomas. This suggests that increased or continued expression of *c-myc* transforms the type of cell in which it occurs into a corresponding tumor.

c-myc exhibits three means of oncogene activation: retroviral insertion, chromosomal translocation, and gene amplification. The common thread among them is deregulated expression of the oncogene rather than a qualitative change in its coding function, although in at least some cases the transcript has lost the usual (and possibly regulatory) nontranslated leader. *c-myc* provides the paradigm for oncogenes that may be effectively activated by increased (or possibly altered) expression.

Translocations are now known in many types of tumors. Often a specific chromosomal site is commonly involved, creating the supposition that a locus at that site is involved in tumorigenesis. However, every translocation generates reciprocal products; sometimes a known oncogene is activated in one of the products, but in other cases it is not

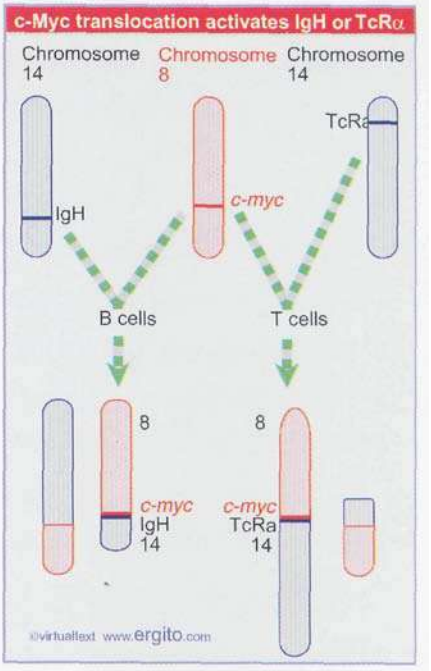


Figure 30.18 A chromosomal translocation is a reciprocal event that exchanges parts of two chromosomes. Translocations that activate the human *c-myc* proto-oncogene involve Ig loci in B cells and TCR loci in T cells.

evident which of the reciprocal products has responsibility for oncogenicity. Also, it is not axiomatic that the gene(s) at the breakpoint have responsibility; for example, the translocation could provide an enhancer that activates another gene nearby.

A variety of translocations found in B and T cells have identified new oncogenes. In some cases, the translocation generates a hybrid gene, in which an active transcription unit is broken by the translocation. This has the result that the exons of one gene may be connected to another. In such cases, there are two potential causes of oncogenicity. The proto-oncogene part of the protein may be activated in some way that is independent of the other part, for example, because it is overexpressed under its new management (a situation directly comparable to the example of *c-myc).* Or the other partner in the hybrid gene may have some positive effect that generates a gain-of-function in the part of the protein coded by the proto-oncogene.

30.13 The Philadelphia translocation generates a new oncogene

Key Concepts

* The Philadelphia translocations create new genes with N-terminal sequences from *bcr* and C-terminal sequences from *c-abl.*
* Both parts of the fusion protein contribute to oncogenicity, which results from activation of the Ras/MAPK pathway.

One of the best characterized cases in which a translocation creates a hybrid oncogene is provided by the *Philadelphia* (Ph^1) chromosome present in patients with chronic myelogenous leukemia (CML). This reciprocal translocation is too small to be visible in the karyotype, but links a 5000 kb region from the end of chromosome 9 carrying *c-abl* to the *bcr* gene of chromosome 22. The *bcr* (breakpoint cluster region) was originally named to describe a region of ~5.8 kb within which breakpoints occur on chromosome 22.

The consequences of this translocation are summarized in **Figure 30.19.** The *bcr* region lies within a large (>90 kb) gene, which is now known as the *bcr* gene. The breakpoints in CML usually occur within one of two introns in the middle of the gene. The same gene is also involved in translocations that generate another disease, ALL (acute lymphoblastic leukemia); in this case, the breakpoint in the *bcr* gene occurs in the first intron.

The *c-abl* gene is expressed by alternative splicing that uses either of the first two exons. The breakpoints in both CML and ALL occur in the intron that precedes the first common exon. Although the exact breakpoints on both chromosomes 9 and 22 vary in individual cases, the common outcome is the production of a transcript coding for a Bcr-Abl fusion protein, in which N-terminal sequences derived from *bcr* are linked to *c-abl* sequences. In ALL, the fusion protein has ~45 kD of the Bcr protein; in CML the fusion protein has ~70 kD of the Bcr protein.

In each case, the fusion protein contains ~140 kD of the usual ~145 kD c-Abl protein, that is, it has lost just a few N-terminal amino acids of the *c-abl* sequence. Changes at the N-terminus are involved in

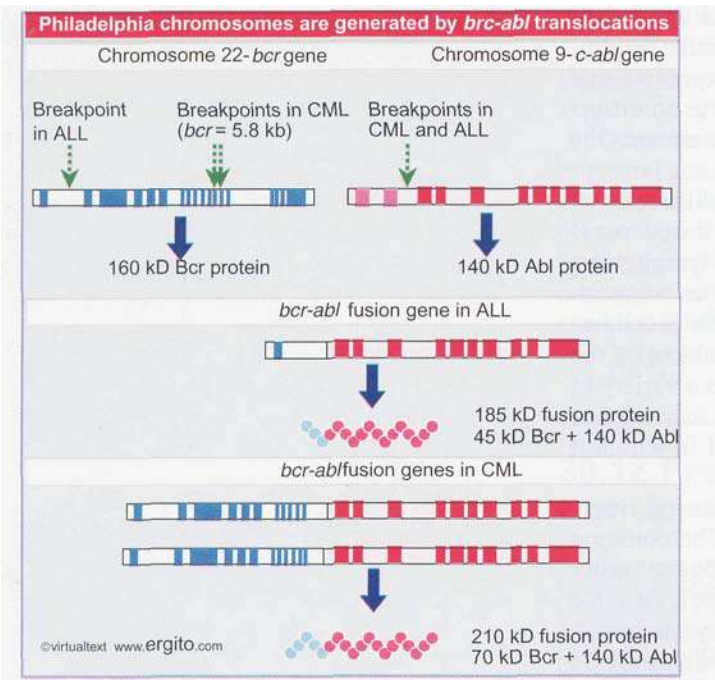


Figure 30.19 Translocations between chromosome 22 and chromosome 9 generate Philadelphia chromosomes that synthesize *bcr-abl* fusion transcripts that are responsible for two types of leukemia.

By Book_Crazy [IND]

activating the oncogenic activity of *v-abl*, a transforming version of the gene carried in a retrovirus. The *c-abl* gene codes for a tyrosine kinase activity; this activity is essential for transforming potential in oncogenic variants. Deletion (or replacement) of the N-terminal region activates the kinase activity and transforming capacity. So the N-terminus provides a domain that usually regulates kinase activity; its loss may cause inappropriate activation.

Why is the fusion protein oncogenic? The Bcr-Abl protein activates the Ras pathway for transformation. It may have multiple ways of doing so, including activation of the adaptors Grb2 and Shc (see *28.14 The Ras/MAPK pathway is widely conserved).* Both the Bcr and Abl regions of the joint protein may be important in transforming activity.

30.14 Oncogenes code for components of signal transduction cascades

Key Concepts

- Oncogenes can be derived from any part of a signal transduction cascade, from the initiating growth factor or receptor to the transcription factor that is the ultimate effector.
- Upstream and downstream components of the Ras pathway are often involved, although not the MAP kinases themselves.

Whether activated by quantitative or qualitative changes, oncogenes may be presumed to influence (directly or indirectly) functions connected with cell growth. Transformed cells lack restrictions imposed on normal cells, such as dependence on serum or inhibition by cell-cell contact. They may acquire new properties, such as the ability to metastasize. Many phenotypic properties are changed when we compare a normal cell with a tumorigenic counterpart, and it is striking indeed that individual genes can be identified that trigger many of the changes associated with this transformation.

We assume that oncogenes, individually or in concert, set in train a series of phenotypic changes that involve the products of many genes. In this description, we see at once a similarity with genes that regulate developmental pathways: they do not themselves necessarily code for the products that characterize the differentiated cells, but they may direct a cell and its progeny to enter a particular pathway. The same analogy suggests itself for oncogenes and developmental regulators: they provide switches responsible for causing transitions between one discrete phenotypic state and another.

Taking this relationship further, we may ask what activities the products of proto-oncogenes play in the normal cell, and how they are changed in the transformed cell. Could some proto-oncogenes be regulators of normal development whose malfunction results in aberrations of growth that are manifested as tumors? There are some examples of such relationships, but do not yet have any systematic understanding of the connection.

Signal transduction pathways are often involved in oncogenesis. The best characterized example is c-Ras, which plays a central role in transmitting the signal from receptor tyrosine kinases (see *28.14 The Ras/MAPK pathway is widely conserved).* Oncogenic mutations change the regulation of Ras activity.

Oncoproteins are organized according to their types of functions in **Figure 30.20.** The left part of the figure groups the oncogenes according

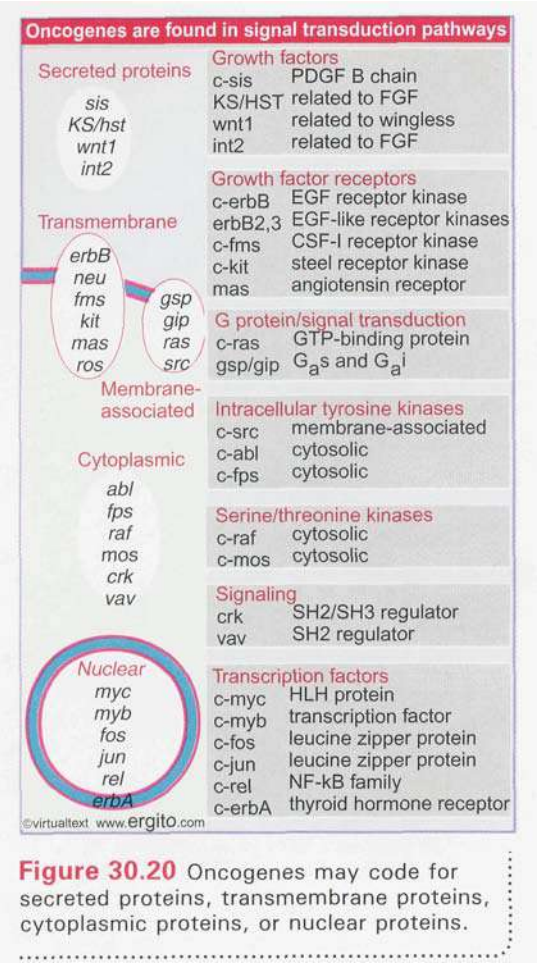


Figure 30.20 Oncogenes may code for secreted proteins, transmembrane proteins, cytoplasmic proteins, or nuclear proteins.

to the locations of their products. The boxes on the right give details of the corresponding proto-oncogenes. The functions of many oncogenes remain unknown, and further groups will no doubt be identified:

- Growth factors are proteins secreted by one cell that act on another. The oncoprotein counterparts can only transform cells bearing the appropriate receptor.
- The growth factor receptors are transmembrane proteins that are activated by binding an extracellular ligand (usually a polypeptide). Most often the receptor is a protein tyrosine kinase. Oncogenicity may result from constitutive (that is, ligand-independent) activation of the kinase activity. Other early stages in signal transduction are identified by Gsp and Gip, which are mutant forms of the a subunits of the G_s and G_i trimeric G proteins.
- An important group of intracellular protein kinases phosphorylate tyrosine residues in target proteins. c-Src, which associates with the cytoskeleton as well as with the membrane, is the prototype of a family of kinases with similar catalytic activities (including c-Yes, c-Fgr, Lck, c-Fps, and Fyn). We understand the effects of oncogenic mutations on the Src kinase activity in some detail, although we have yet to explain why the altered kinase activity is oncogenic. Other protein tyrosine kinases in the intracellular group are cytosolic; c-Abl is found in both cytosol and nucleus.
- A group of cytosolic enzymes are protein serine/threonine kinases, that is, they phosphorylate target proteins on serine or threonine. Little is known about the effects of oncogenic mutations beyond the fact they probably increase or constitutively activate the kinase activities. Mos is an example which can activate ERK MAPK.
- Nuclear proteins include transcription factors of several types. The functions of these proto-oncoproteins are rather well described (see *22 Activating Transcription).* Generally we understand what effects the oncogenic mutations have on the factors, but we cannot yet relate these changes to the activation or repression of a set of target genes that defines the oncogenic state.

The common feature is that each type of protein is in a position to trigger general changes in cell phenotypes, either by initiating or responding to changes associated with cell growth, or by changing gene expression directly. Before we consider in detail the potential of each group for initiating a series of events that has an oncogenic outcome, we need to consider how many independent pathways are identified by these factors.

Recall the example of the best characterized mitogenic pathway, the MAPK pathway which consists of the following stages:

growth factor
i
growth factor receptor (tyrosine kinase)
i
Ras
i
kinase cascade (serine/threonine kinases)
i
transcription factor(s)

When a growth factor interacts with its receptor, it activates the tyrosine kinase activity. The signal is passed (via an adaptor) to Ras. At this point, the pathway switches to a series of serine/threonine kinases. The targets at the end of the pathway may be controlled directly or indirectly by phosphorylation, and include transcription factors, which are in a position to make widespread changes in the pattern of gene expression.

By Book_Crazy [IND]

If a pathway functions in a linear manner, in which the signal passes directly from one component to the next, the same results should be achieved by constitutive activation of any component (so that it no longer needs to be activated by a signal from an earlier component).

A signal transduction pathway, of course, is likely to branch at several stages, so that an initial stimulus may trigger a variety of responses. The activation of components that are downstream will therefore activate a smaller number of end-functions than the activation of components at the start of the pathway. But we can analyze any individual part of the pathway by tracing it back to the beginning as though it were strictly linear.

In the example of the Ras pathway, we know that it is activated by many growth factors to generate a mitogenic response. Mutations in the early part of this pathway, including the *ras* and ra/genes, may be oncogenic. But oncogenic mutations are not usually found in the following components of the cascade, the MEK and MAP kinases. This suggests that there may be a branch in the pathway at the stage of *ras* or *raf*, and that activation of this branch is also necessary for oncogenicity. Ras activates a cytoskeletal GTPase called Rac, which may identify this branch. However, the ERK MAPK pathway terminates in the activation of several "immediate early" genes, including *fos* and *jun*, which themselves have oncogenic counterparts, suggesting that the targets of the MAPK pathway can be sufficient for oncogenicity.

The central role of this pathway is indicated by the number of its components that are coded by proto-oncogenes. One explanation of the discrepancies between the susceptibilities of MAP kinases and other components to oncogenic mutation may be that the *level* or *duration* of expression is important. It could be the case that mutations in MEK or MAP kinases do not activate the enzymes sufficiently to be oncogenic. Alternatively, the oncogenic mutations (which, after all, represent gain-of-function) may cause new targets to be activated in addition to the usual pathway. The general principle is clear: *that aberrant activation of mitogenic pathways can contribute to oncogenicity*, but we cannot yet explain exactly how the activation of these pathways changes the properties of the cell in terms of immortalization or transformation.

30.15 Growth factor receptor kinases can be mutated to oncogenes

Key Concepts

- Oncogenes are generated by mutations that constitutively activate growth factor receptor genes.
- The type of tumor reflects the phenotypes of the cells in which the receptor is expressed.

The protein tyrosine kinases constitute a major class of oncoproteins, and fall into two general groups: transmembrane receptors for growth factors; and cytoplasmic enzymes. We have more understanding about the biological functions of the receptors, because we know the general nature of the signal transduction cascades that they initiate, and we can see how their inappropriate activation may be oncogenic. The normal roles in the cell of the cytoplasmic tyrosine kinases are not so well defined, but in several cases it appears that they provide catalytic functions for receptors that themselves lack kinase activity; that is, the activation of the receptor leads to activation of the cytoplasmic tyrosine kinase. We have a great deal of information about their

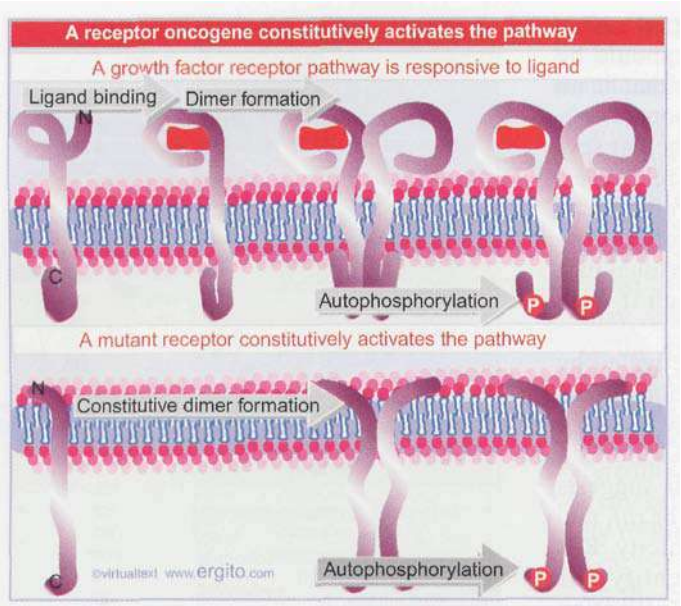


Figure 30.21 Activation of a growth factor receptor involves ligand binding, dimerization, and autophosphorylation. A truncated oncogenic receptor that lacks the ligand-binding region is constitutively active because it is not repressed by the N-terminal domain.

enzymatic activities and the molecular effects of oncogenic mutations, although it has been more difficult to identify their physiological targets.

Receptors for many growth factors have kinase activity. They tend to be large integral membrane proteins, with domains assembled in modular fashion from a variety of sources. We discussed the general nature of transmembrane receptors and the means by which they are activated to initiate signal transduction cascades in *28.8 Growth factor receptors are protein kinases.* The EGF receptor is the paradigm for tyrosine kinase receptors. The extracellular N-terminal region binds the ligand that activates the receptor. The intracellular C-terminal region includes a domain that has tyrosine kinase activity. Most of the receptors that are coded by cellular proto-oncogenes have a similar form of organization.

Dimerization of the extracellular domain of a receptor activates the tyrosine kinase activity of the intracellular domain. Various forms of this reaction were summarized previously in Figure 28.17. When the cytoplasmic domains of the monomers are brought into contact, they trigger an autophosphorylation reaction, in which each monomer phosphorylates the other (see *28.9 Receptors are activated by dimerization*).

A (generalized) relationship between a growth factor receptor and an oncogenic variant is illustrated in **Figure 30.21.** The wild-type receptor is regulated by ligand binding. In the absence of ligand, the monomers do not interact. Growth factor binding triggers an interaction, allowing the receptor to form dimers. This in turn activates the receptor, and triggers signal transduction. By contrast, the oncogenic variant spontaneously forms dimers that are constitutively active. Different types of events may be responsible for the constitutive dimerization and activation in different growth factor receptors.

The oncogene *v-erb* is a truncated version of *c-erbB,* the gene coding for the EGF receptor. The oncoprotein retains the tyrosine kinase and transmembrane domains, but lacks the N-terminal part of the protein that binds EGF, and does not have the C-terminus. The deletions at both ends may be needed for oncogenicity. The change in the extracellular N-terminal domain allows the protein to dimerize spontaneously; and the C-terminal deletion removes a cytosolic domain that inhibits transforming activity. There is also an activating mutation in the catalytic domain. So the basis for oncogenicity is the combination of mutations that activate the receptor constitutively.

The general principle that constitutive or altered activity may be responsible for oncogenicity applies to the group of growth factor receptors summarized previously in Figure 30.20. Another example of an activation event is provided by *erbB2,* which codes for a receptor closely related to the EGF receptor. An oncogenic form has a key mutation in its transmembrane region; this increases the propensity of the receptor monomers to form dimers.

Some proto-oncogenes code for receptors or factors involved in the development of particular cell types. Mutation of such a receptor (or growth factor) may promote unrestricted growth of cells of the appropriate type. The proto-oncogene *c-fms* codes for the CSF-I receptor, which mediates the action of colony stimulating factor I, a macrophage growth factor that stimulates the growth and maturation of myeloid precursor cells. *c-fms* can be rendered oncogenic by a mutation in the extracellular domain; this increases dimerization and makes the protein constitutively active in the absence of CSF-I. Oncogenicity is enhanced by C-terminal mutations, which could act by inactivating an inhibitory intracellular domain.

By Book_Crazy [IND]

30.16 Src is the prototype for the proto-oncogenic cytoplasmic tyrosine kinases

Key Concepts

- The cytoplasmic tyrosine kinases phosphorylate tyrosine residues in cytosolic proteins.
- Myristoylation of the N-terminus enables Src to associate with the plasma membrane.
- The crucial cellular targets for Src remain unidentified.

The cellular action and basis for oncogenicity of the cytoplasmic group of protein tyrosine kinases is more obscure. The cytoplasmic group is characterized by the viral oncogenes *src*, *yes*, *fgr*, *fps/fes*, *abl*, and *ros*. (c-Src is actually associated with membranes.) A major stretch of the sequences of all these genes is related, corresponding to residues 80-516 of *c-src*. This includes the SH2 and SH3 domains and the catalytic domain responsible for kinase activity. Presumably the regions outside this domain control the activities of the individual members of the family. In few cases, however, do we know the cellular function of a *c-onc* member of this group.

The paradigm for a cytoplasmic tyrosine kinase in search of a role is presented by the Src proteins. Since its isolation by Rous in 1911, RSV has been perpetuated under a variety of conditions, and there are now several "strains," carrying variants of *v-src*. The common feature in the sequence of *v-src* is that the C-terminal sequence of *c-src* has been replaced. The various strains contain different point mutations within the *src* sequence.

Proteins in the Src family were the first oncoproteins of the kinase type to be characterized. Src was also the first example of a kinase whose target is a tyrosine residue in protein. The level of phosphotyrosine is increased about 10 X in cells that have been transformed by RSV. In addition to acting on other proteins, Src is able to phosphorylate itself.

Src proteins have several interesting features. **Figure 30.22** summarizes their activities in terms of protein domains.

Both v-Src and c-Src are modified at the N-terminus. The N-terminal amino acid is cleaved, and myristic acid (a rare fatty acid of 14 carbon residues) is covalently added to the N-terminal glycine. Myristoylation enables Src proteins to attach to the cytosolic face of membranes in the cytoplasm. Most of the protein is associated with the cytoplasmic face of the endosomes, and it is enriched in regions where there are cell-to-cell contacts and adhesion plaques.

Myristoylation is essential for oncogenic activity of v-Src, since N-terminal mutants that cannot be myristoylated have reduced tumorigenicity. The simplest explanation for the dependence of transformation on the membrane location of v-Src is that important substrates for Src are located in the membrane.

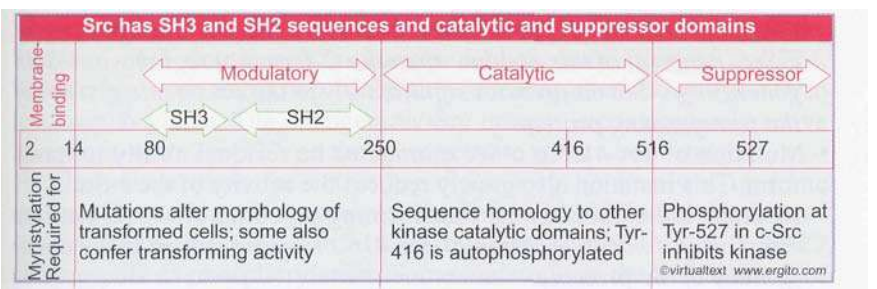


Figure 30.22 A Src protein has an N-terminal domain that associates with the membrane, a modulatory domain that includes SH2 and SH3 motifs, a kinase catalytic domain, and (c-Src only) a suppressor domain.

By Book_Crazy [IND]

The biological action of v-Src is qualitatively different from that of c-Src, since increased concentrations of c-Src cannot fully transform cells. The major biochemical difference between v-Src and c-Src lies in their kinase activities. The activity of v-Src is ~20× greater than that of c-Src. The transforming activity of *src* mutants is correlated with the level of kinase activity, and we believe that oncogenicity results from phosphorylation of target protein(s). We do not know whether the increased activity is itself responsible for oncogenicity or whether there is also a change in the specificity with which target proteins are recognized.

Kinase activity plays two roles in Src function. First, attempts to identify a function for the phosphorylation in cell transformation have concentrated on identifying cellular substrates that may be targets for v-Src (especially those that may not be recognized by c-Src). A variety of substrates has been identified, but none has yet been equated with the cause of transformation. Second, the state of phosphorylation of Src itself controls the transforming activity (see next section).

30.17 Src activity is controlled by phosphorylation

Key Concepts

- Src autophosphorylates and its activity is controlled by the state of phosphorylation at two Tyr residues.
- Oncogenic variants are derived from c-Src by mutations that cause decreased phosphorylation at Tyr-527 and increased phosphorylation at Tyr-416.
- v-Src lacks Tyr-527 and is constitutively active.
- Src was the protein in which the SH2 and SH3 motifs were originally identified.

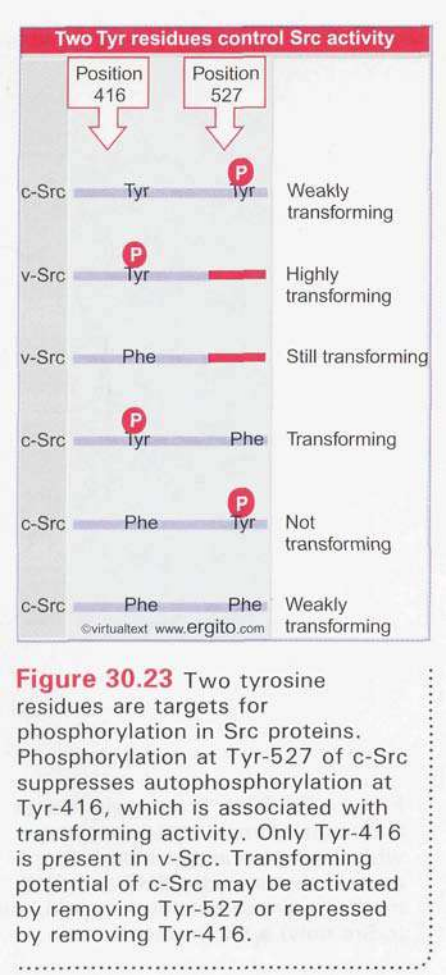


Figure 30.23 Two tyrosine residues are targets for phosphorylation in Src proteins. Phosphorylation at Tyr-527 of c-Src suppresses autophosphorylation at Tyr-416, which is associated with transforming activity. Only Tyr-416 is present in v-Src. Transforming potential of c-Src may be activated by removing Tyr-527 or repressed by removing Tyr-416.

Two sites in Src control its kinase activity. It is inactivated by phosphorylation at tyrosine residue 527, which is part of the C-terminal sequence of 19 amino acids that is missing from v-Src. The c-Src protein is phosphorylated *in vivo* at this position by the kinase Csk, which maintains it in an inactive state. Src is activated by phosphorylation at Tyr-416, which is located in the activation loop of the kinase domain.

The importance of these phosphorylations can be tested by mutating the tyrosine residues at 416 and 527 to prevent addition of phosphate groups. The mutations have opposite effects, as summarized in **Figure 30.23:**

- Mutation of Tyr-527 to the related amino acid phenylalanine activates the transforming potential of c-Src. The protein c-Src$^{Phe\text{-}527}$ becomes phosphorylated on Tyr-416, has its kinase activity increased ~10×, and it transforms target cells, although not as effectively as v-Src. *Phosphorylation of Tyr-527 therefore represses the oncogenicity of c-Src. Removal of this residue when the C-terminal region was lost in generating v-Src contributes significantly to the oncogenic activity of the transforming protein.*
- Mutation of Tyr-416 in c-Src eliminates its residual ability to transform. This mutation also greatly reduces the activity of the c-Src$^{Phe\text{-}527}$ mutant. It also reduces the transforming potential of v-Src, but less effectively. *Phosphorylation at Tyr-416 therefore activates the oncogenicity of Src proteins.*

By Book_Crazy [IND]

Point mutations at other positions in c-Src support a correlation in which oncogenicity is associated with decreased phosphorylation at Tyr-527 and increased phosphorylation at Tyr-416. The states of these tyrosines may therefore be a general indicator of the oncogenic potential of c-Src. The reduced phosphorylation at Tyr-527 is responsible for the increased phosphorylation at Tyr-416, which is the crucial event. However, v-Src is less dependent on the state of Tyr-416, and mutants at this position retain transforming activity; presumably *v-src* has accumulated other mutations that increase transforming potential.

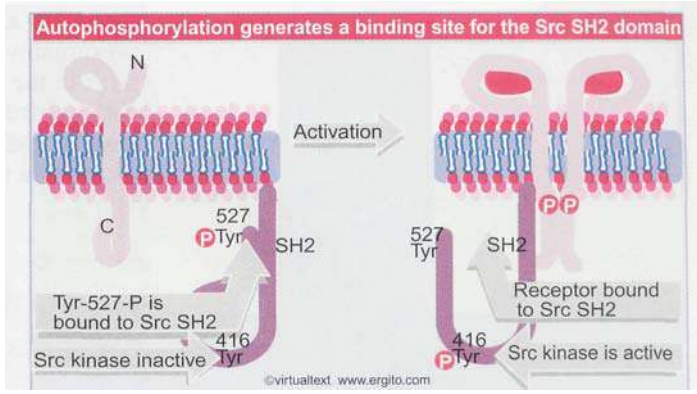


Figure 30.24 When a receptor tyrosine kinase is activated, autophosphorylation generates a binding site for the Src SH2 domain, Tyr-527 is released and dephosphorylated, Tyr-416 becomes phosphorylated, and Src kinase is activated.

What is the function of c-Src; and how is it related to the oncogenicity of v-Src? The c-Src and v-Src proteins are very similar: they share N-terminal modification, cellular location, and protein tyrosine kinase activity. c-Src is expressed at high levels in several types of terminally differentiated cells, which suggests that it is not involved in regulating cell proliferation. But we have so far been unable to determine the normal function of c-Src. A very large number of proteins have been identified as targets for the Src kinase, most of them identified with signaling pathways, and some with the interactions of the cell with the environment. c-Src is activated by growth factor receptors, such as the PDGF receptor, suggesting the general view that, like other oncogenes, it is involved in signaling pathways that regulate growth which can be tumorigenic when constitutively activated.

The modulatory region of c-Src contains two motifs that are found in a variety of other cytoplasmic proteins that are involved in signal transduction: these may connect a protein to the components that are upstream and downstream of it in a signaling pathway. The names of these two domains, SH2 and SH3, reflect their original identification as regions of *Sxc* homology (see *28.11 Signaling pathways often involve protein-protein interactions).*

How is c-Src usually activated? Most mutations in the SH2 region reduce transforming activity (suggesting that the SH2 function is required to activate c-Src), and most mutations in SH3 increase transforming activity (suggesting that SH3 has a negative regulatory role). **Figure 30.24** shows a more detailed model for the function of the SH2 domain. The state of phosphorylation at Tyr-527 is critical. In the inactive state, Tyr-527 is phosphorylated, and this enables the C-terminal region of c-Src itself to bind to the N-terminal SH2 domain. When an appropriate receptor tyrosine kinase (such as PDGF receptor) is activated, its autophosphorylation creates an SH2-binding site that displaces Tyr-527. This leads to its dephosphorylation, which triggers a change in conformation allowing Tyr-416 to be phosphorylated.

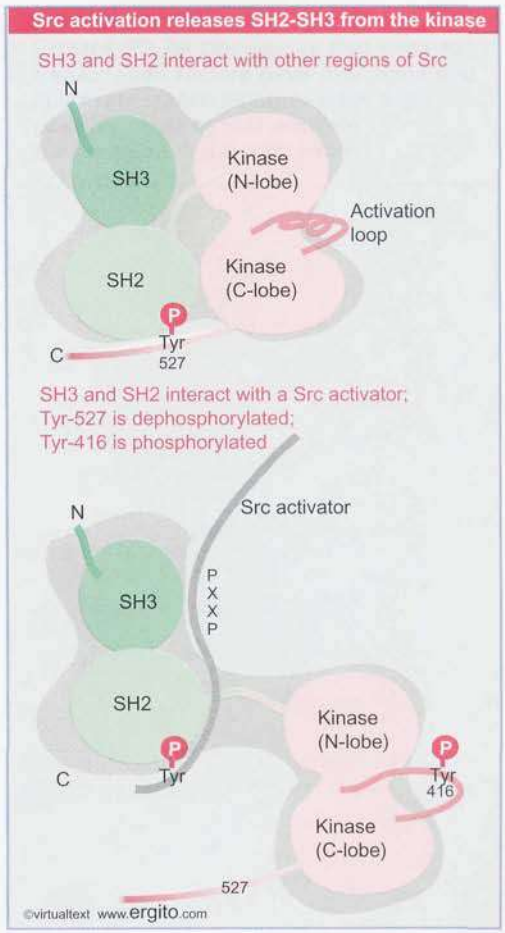


Figure 30.25 Src is inactive when the SH2 domain binds the Tyr-phosphate at position 527, and the SH3 domain binds the connector between kinase and SH2 domains. These interactions are disrupted when a Src activator binds to the SH2 and SH3 domains. This causes Tyr-527 to be dephosphorylated, and changes the conformation of the activation loop of the kinase domain so that Tyr-416 can be phosphorylated.

Figure 30.25 shows a schematic model based on the crystal structure of Src. The SH2 domain binds to a C-terminal projection of the kinase domain that contains Tyr-527. The SH3 domain binds to a short sequence that connects the SH2 and catalytic domains. The SH2 and SH3 domains are at the back of the kinase domain, so these interactions lock the enzyme in an inactive state. The activation loop in the kinase domain is in a state that does not allow Tyr-416 to be phosphorylated. An activator (such as an activated membrane receptor) binds to both the SH2 and SH3 domains. This causes dephosphorylation of Tyr-527, which triggers an unfolding of the activator loop that allows Tyr-416 to be phosphorylated. The oncogenic v-Src protein is constitutively active because it lacks Tyr-527, so the inactive state cannot be formed.

Alternative ways for activating c-Src may be involved in some oncogenic reactions. For example, the polyoma middle T antigen activates c-Src by binding to the C-terminal region including Tyr-527 and prevents its phosphorylation. Some mutations in the SH2 domain of

c-Src can activate the kinase activity (with oncogenic consequences), presumably because they prevent it from sequestering Tyr-527. Mutations in the SH2 and SH3 domains of c-Src can influence its specificity with regard to transforming different types of target cells, which suggests that these regions provide the connections to other (cell specific) proteins in the pathway.

30.18 Oncoproteins may regulate gene expression

Key Concepts

- The *rel* oncogene is a member of the NFKB family.
- The *jun* and fos oncogenes code for the subunits of the transcription factor AP1.
- Screening with high density DNA microarrays allows the pattern of gene expression to be compared between tumor cells and normal cells.

It goes almost without saying that it is necessary to make changes in gene expression in order to convert a cell to the transformed phenotype. Many oncogenes act at early stages in pathways that lead ultimately to changes in gene expression. Some act directly at the level of transcription. Retroviral oncogenes include examples derived from the major classes of cellular transcription factors. Several prominent gene families coding for transcription factors are identified by *v-onc* genes: *rel, jun, fos, erbA, myc,* and *myb.* In the cases of Rel, Jun, and ErbA proteins, there are differences in transcriptional activity between the c-Onc and v-Onc proteins that may be related to transforming capacity.

The actions of *v-onc* genes may in principle be quantitative or qualitative; and those that affect transcription might either increase or decrease expression of particular genes. By virtue of increased expression or activity they could turn up transcription of genes whose products can be tolerated only in small amounts. Failure to respond to normal regulation of activity by other cellular factors also might lead to increased gene expression. A less likely possibility is the acquisition of specificity for new target promoters. Alternatively, if the oncoproteins are defective in the ability to activate transcription, they might function as dominant negative suppressors of the cellular transcription factors. The first steps towards distinguishing these possibilities lie with determining which functions are altered in v-Onc compared with c-Onc proteins: is DNA-binding altered either quantitatively or quantitatively; is the ability to activate transcription altered? **Figure 30.26** summarizes the properties of some of these oncoproteins.

The oncogene *v-rel* was identified as the transforming function of the avian (turkey) reticuloendotheliosis virus. The retrovirus is highly oncogenic in chickens, where it causes B-cell lymphomas. *v-rel* is a truncated version of *c-rel,* lacking the ~100 C-terminal amino acids, and has a small number of point mutations in the remaining sequence.

The *rel* gene belongs to a family whose best characterized member is the transcription factor NF-KB. This is a dimer of two subunits, p65 and p50, which is held in the cytoplasm by a regulator, I-κB. (Binding of I-κB masks the nuclear localization sequence in NF-KB.) When I-KB is phosphorylated, it is degraded and therefore releases NF-KB, which enters the nucleus and activates transcription of target genes whose promoters or enhancers have the KB motif. This regulatory story is essentially recapitulated in *Drosophila* development, where *dorsal* codes for

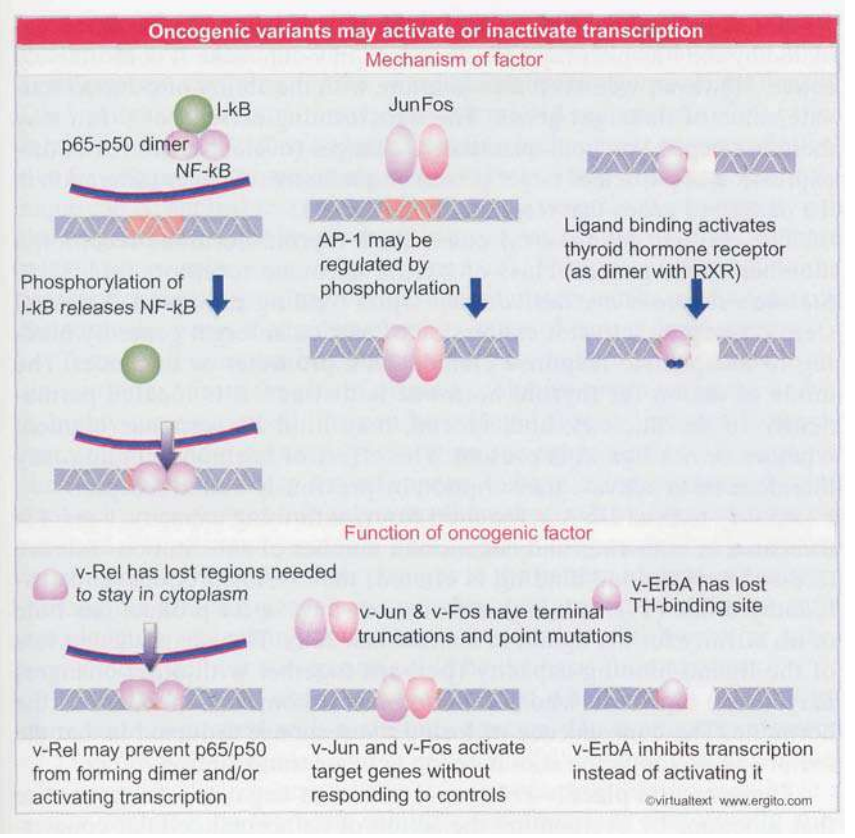


Figure 30.26 Oncogenes that code for transcription factors have mutations that inactivate transcription (v-erbA and possibly v-rel) or that activate transcription (v-jun and v-fos).

an NF-KB homologue that is held in the cytoplasm by the *cactus* product, an IκB homologue (see *31.11 Dorsal protein forms a gradient of nuclear localization).* The two subunits of NF-KB have related sequences, and *c-rel* has 60% similarity with p50.

NF-KB is one of the most pleiotropic transcription factors; indeed, it has been suggested that it may constitute a general second messenger. Many types of stimulus to the cell result in activation of NF-Kb and a broad range of genes is activated via the presence of KB binding sites. The members of the NF-KB family form various pairwise combinations that regulate transcription. When v-Rel forms dimers with cellular family members, it may influence their activities either negatively or positively, thus changing the pattern of gene expression. v-Rel is exclusively nuclear, because it has lost the sequences required for export to the cytoplasm.

The transcription factor AP1 is the nuclear factor required to mediate transcription induced by phorbol ester tumor promoters (such as TPA). An AP1 binding site confers TPA-inducibility upon a target gene. The canonical AP1 factor consists of a dimer of two subunits, coded by the genes *c-jun* and *c-fos,* which activates genes whose promoters or enhancers have an AP1 binding site.

Jun and Fos are transcription factors of the leucine zipper class. Each protein is a member of a family, and a series of pairwise interactions between Jun family members and Fos family members may generate a series of transcription factors related to AP1. Mutations of *v-jun* or *v-fos* that abolish the ability to bind DNA or that damage the transactivation function also render the product non-transforming, providing a direct proof that ability to activate transcription is required for transforming activity.

c-Jun is activated by phosphorylation on two serine residues by the action of the kinase JNK, which is activated by the Ras pathway, and this contributes to the transforming action of Ras. The transforming activity of v-Jun has a more complex basis. v-Jun has a deletion of amino acids

34-60 that includes both these sites of phosphorylation, and so is not regulated by the Ras pathway. Other changes in v-Jun make it constitutively active. However, v-Jun can also interfere with the ability of c-Jun to activate some of its target genes. The transforming activity of v-Jun may therefore depend on both quantitative changes (overexpression or underexpression of particular target genes) or qualitative changes (alteration in the pattern of genes that responds to the factor).

The cellular gene *c-erbA* codes for a thyroid hormone receptor, a member of the general class of steroid hormone receptors (see *22.10 Steroid receptors are activators).* Upon binding its ligand, a typical steroid receptor activates expression of particular target genes by binding to its specific response element in a promoter or enhancer. The mode of action for thyroid hormone is distinct: it is located permanently in the nucleus, and, indeed, may bind its response element whether or not ligand is present. The effect of hormone binding may therefore be to activate transcription by previously bound receptor.

Ability to bind DNA is required for transforming capacity. *v-erbA* is truncated at both ends and has a small number of substitutions relative to *c-erbA.* Hormone binding is altered; the *c-erbA* product binds triiodothyronine (T_3) with high affinity, but the *v-erbA* product has little or no affinity for the ligand in mammalian cells. This suggests that loss of the ligand-binding capacity (perhaps together with other changes) may create a protein whose function has become independent of the hormone. The consequence of losing the response to ligand is that the factor can no longer be stimulated to activate transcription.

These results place *v-erbA* as a dominant negative oncogene, one that functions by overcoming the action of its normal cellular counterpart. Its action is to prevent transcription of genes that usually are activated by c-ErbA. The implication is that genes activated by c-ErbA act to *suppress* transformation. In this particular case, it seems likely that these genes usually promote differentiation; blocking this action allows the cells to proliferate.

c-jun, c-fos, c-rel, and also *c-myc* are "immediate early" genes, members of a class of genes that are rapidly induced when resting cells are treated with mitogens, which suggests that they may be involved in a cascade that initiates cycling. So their targets are likely to be concerned with initiating or promoting growth. We should therefore expect an increase in their activities to be associated with oncogenesis, an expectation that may be fulfilled for *v-fos* and *v-myc,* but does not explain the behavior of *v-rel.*

The adenovirus oncogene E1A provides an example of a protein that regulates gene expression indirectly, that is, without itself binding to DNA. The E1A region is expressed as three transcripts, derived by alternative splicing, as indicated in **Figure 30.27.** The 13S and 12S mRNAs code for closely related proteins and are produced early in infection. They possess the ability to immortalize cells, and can cooperate with other oncoproteins (notably Ras) to transform primary cells (see later). No other viral function is needed for this activity.

The E1A proteins exercise a variety of effects on gene expression. They activate the transcription of some genes, but repress others. Mutation of the E1A proteins suggests that transcriptional activation requires only the short region of domain 3, found only in the 289 amino acid protein coded by 13S mRNA. Repression of transcription, induction of DNA synthesis, and morphological transformation all require domains 1 and 2, common to both the 289 and 243 amino acid proteins. This suggests that repression of target genes is required to cause transformation. E1A proteins act by binding to several cellular proteins that in turn repress or activate transcription of appropriate target genes. Among these targets are the CBP and p300 coactivators, the TBP basal transcription factor, and the cell cycle regulators RB and p27.

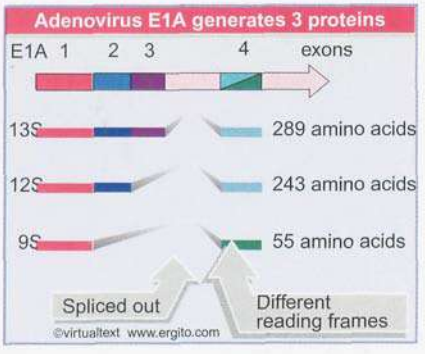


Figure 30.27 The adenovirus E1A region is spliced to form three transcripts that code for overlapping proteins. Domain 1 is present in all proteins, domain 2 in the 289 and 243 residue proteins, and domain 3 is unique to the 289 residue protein. The C-terminal domain of the 55 residue protein is translated in a different reading frame from the common C-terminal domains of the other two proteins.

A powerful new approach to analyzing the roles of individual genes in cancer has been made possible by the development of techniques to allow simultaneous screening for the expression of many genes. High density DNA microarrays contain probes to the mRNAs of up to 20,000 genes (typically immobilized on a glass slide). The technique is at its most effective for comparing the genes expressed in two related cell types. The technique can be applied to a tumor cell when it is possible to compare it with the original cell type from which it arose, or can be used to compare related tumor cells with different properties. This gives insights into the extent of change in gene expression, and ultimately can be used to identify the particular genes that are involved in stages of cancer development.

Tumor cell lines can be obtained that vary in their ability to metastasize (to spread from the site of origin to colonize new sites in the body). A highly metastatic cell line can be selected from a line that is poorly metastatic without apparently changing the properties of the tumor as such—only the ability to spread appears to be affected. A comparison in two such cases showed that only a small number (<20) of genes have a significant change in transcription. This suggests the possibility that metastasis involves relatively minor changes in the pattern of gene expression. Among the affected genes are several whose products act on the actin cytoskeleton, suggesting that changes in cellular motility and/or adhesion to other cells or substratum could be the basis for metastasis.

More widespread changes were detected when benign prostate tumors were compared with metastatic cancers. On average, expression of 55 genes was increased, and expression of 480 genes was decreased. This is interesting in suggesting that cancer progression may have a major contribution from the inhibition of functions that usually suppress cell growth or motility. One of the genes whose expression is increased is *EZH2,* a member of the Polycomb group that forms complex involved in repressing the activity of chromatin (see *23.16 Polycomb and trithorax are antagonistic repressors and activators)*. *EZH2* is responsible for repressing about one third of the genes whose expression is decreased in the metastatic prostate cancer.

30.19 RB is a tumor suppressor that controls the cell cycle

Key Concepts

- Loss of both alleles of *RB* causes retinoblastomas.
- Nonphosphorylated RB prevents cell proliferation.
- In the normal cell cycle, phosphorylation of RB by cdk-cyclin kinases is necessary to proceed into S phase.
- Certain tumor antigens suppress the inhibitory action by sequestering the nonphosphorylated form of RB.
- Several tumor suppressors act by blocking the cdk-kinase complexes that phosphorylate RB.

The common theme in the role of oncogenes in tumorigenesis is that increased or altered activity of the gene product is oncogenic. Whether the oncogene is introduced by a virus or results from a mutation in the genome, it is dominant over its allelic proto-oncogene(s). A mutation that activates a single allele is tumorigenic. Tumorigenesis then results from gain of a function.

Certain tumors are caused by a different mechanism: loss of both alleles at a locus is tumorigenic. Propensity to form such tumors may be inherited through the germline; it also occurs as the result of somatic

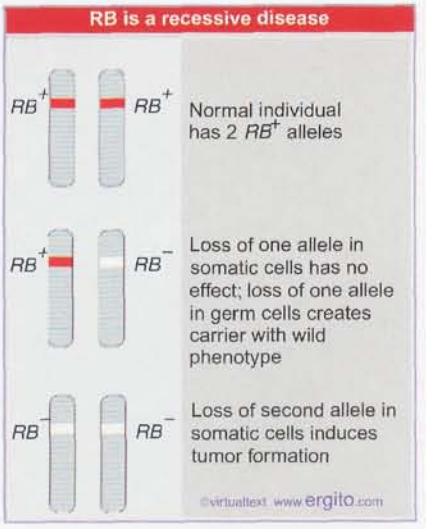


Figure 30.28 Retinoblastoma is caused by loss of both copies of the RB gene in chromosome band 13q14. In the inherited form, one chromosome has a deletion in this region, and the second copy is lost by somatic mutation in the individual. In the sporadic form, both copies are lost by individual somatic events.

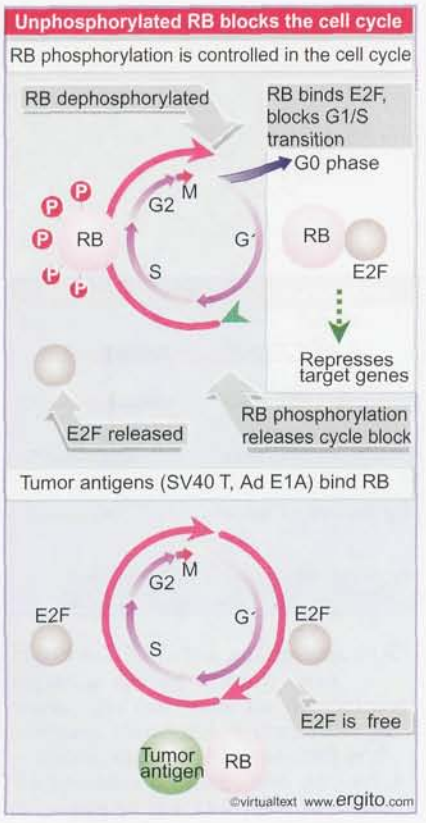


Figure 30.29 A block to the cell cycle is released when RB is phosphorylated or when it is sequestered by a tumor antigen.

change in the individual. Such cases identify tumor suppressors: genes whose products are needed for normal cell function, and whose loss of function causes tumors. The two best characterized genes of this class code for the proteins RB and p53.

Retinoblastoma is a human childhood disease, involving a tumor of the retina. It occurs both as a heritable trait and sporadically (by somatic mutation). It is often associated with deletions of band q14 of human chromosome 13. The *RB* gene has been localized to this region by molecular cloning.

Figure 30.28 summarizes the situation. Retinoblastoma arises when both copies of the *RB* gene are inactivated. In the inherited form of the disease, one parental chromosome carries an alteration in this region. A somatic event in retinal cells that causes loss of the other copy of the *RB* gene causes a tumor. In the sporadic form of the disease, the parental chromosomes are normal, and both *RB* alleles are lost by (individual) somatic events.

The cause of retinoblastoma is therefore loss of protein function, usually resulting from mutations that prevent gene expression (as opposed to point mutations that affect function of the protein product). Loss of *RB* is involved also in other forms of cancer, including osteosarcomas and small cell lung cancers.

RB is a nuclear phosphoprotein that influences the cell cycle (see *29.17 G0/G1 and G1/S transitions involve cdk inhibitors).* In resting (G0/G1) cells, RB is not phosphorylated. RB is phosphorylated during the cell cycle by cyclin/cdk complexes, most particularly at the end of G1; it is dephosphorylated during mitosis. The nonphosphorylated form of RB specifically binds several proteins, and these interactions therefore occur only during part of the cell cycle (prior to S phase). Phosphorylation releases these proteins.

The target proteins include the E2F group of transcription factors, which activate target genes whose products are essential for S phase. Binding to RB inhibits the ability of E2F to activate transcription, which suggests that RB blocks the expression of genes dependent on E2F. In this way, RB indirectly prevents cells from entering S phase. Also, the RB-E2F complex directly represses some target genes, so its dissociation allows them to be expressed.

Certain viral tumor antigens bind specifically to the nonphosphorylated form of RB. The best characterized are SV40 T antigen and adenovirus E1A. This suggests the model shown in **Figure 30.29.** Nonphosphorylated RB prevents cell proliferation; this activity must be suppressed in order to pass through the cell cycle, which is accomplished by the cyclic phosphorylation. And it may also be suppressed when a tumor antigen sequesters the nonphosphorylated RB. Because the RB-tumor antigen complex does not bind E2F, the E2F is permanently free to allow entry into S phase (and the RB-E2F complex is not available to repress its target genes).

Overexpression of RB impedes cell growth. An indication of the importance of RB for cell proliferation is given by the properties of an osteosarcoma cell line that lacks RB; when RB is introduced into this cell line, its growth is impeded. However, the inhibition can be overcome by expression of D cyclins, which form cdk-cyclin combinations that phosphorylate RB. RB is not the only protein of its type: proteins with related sequences, called p107 and p130, have similar properties.

The connection between the cell cycle and tumorigenesis is illustrated in **Figure 30.30.** Several regulators are identified as tumor suppressors by the occurrence of inactivating mutations in tumors. As well as occurring in RB itself, mutations are found in the small inhibitory proteins (most notably p16 and possibly p21), and D cyclin(s). Although these proteins (most notably RB) play a role in the cycle of a proliferating cell, the role that is relevant for tumorigenesis is more

probably their function in the quiescent (GO) state. In quiescent cells, RB is not phosphorylated, D cyclin levels are low or absent, and p16, p21, and p27 ensure inactivity of cdk-cyclin complexes.

Cyclin D activity is needed for proliferation. Loss of the circuit that suppresses it (p16 or p21) causes unrestrained growth. The control of cyclin D is particularly important in breast cancer. The cyclin D1 gene is amplified in 20% of human breast cancers, and the protein is overexpressed in >50% of human mammary carcinomas. The role of cyclin D1 has been confirmed in mouse models by showing that increased Dl expression causes increased breast cancer, while the deletion of the cyclin D1 gene prevents certain oncogenes from causing breast cancer. Mammary tumors can be caused in cyclin D1-deficient mice by the oncogenes *myc* or *Wnt-1*, but not by the *ras* or *neu* oncogenes. The *neu* oncogene codes for a growth factor receptor that activates Ras, and the Ras pathway leads to activation of the promoter of the cyclin Dl gene, which explains the result. The Myc and Wnt-1 oncogenes must cause breast cancer by a different pathway.

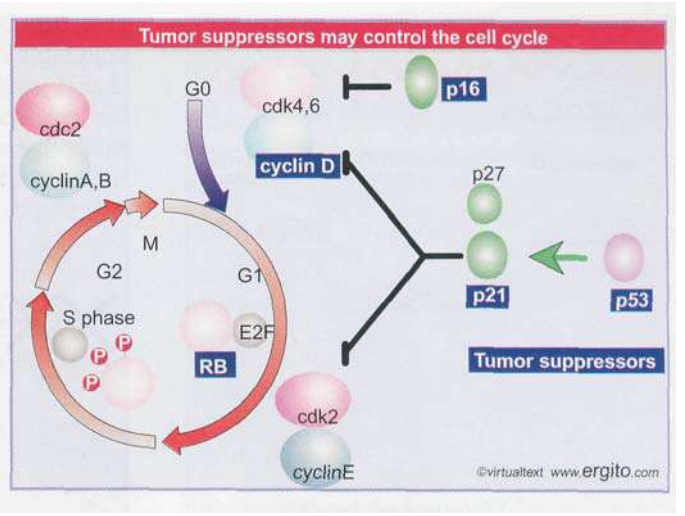


Figure 30.30 Several components concerned with G0/G1 or G1/S cycle control are found as tumor suppressors.

30.20 Tumor suppressor p53 suppresses growth or triggers apoptosis

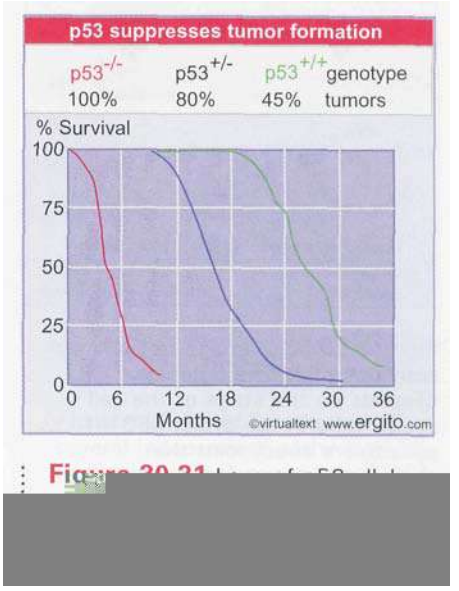
Key Concepts

- p53 is a tumor suppressor that is lost or inactivated in >50% of all human cancers.
- Wild-type p53 is activated by damage to DNA.
- The response may be to block cell cycle progression or to cause apoptosis depending on the circumstances.

The most important tumor suppressor is p53 (named for its molecular size). More than half of all human cancers either have lost p53 protein or have mutations in the gene, making loss of p53 by far and away the most common alteration in human cancer. Its effects have been demonstrated directly in mice, where loss of p53 alleles causes the occurrence of tumors. **Figure 30.31** shows the survival curves for wild type mice ($p53^{+/+}$), heterozygotes who have lost one allele ($p53^{+/-}$), and homozygotes who have lost both alleles ($p53^{-/-}$). The frequency of tumors is increased from 45% to 80% by loss of the first allele, causing the mice to die sooner; and loss of both alleles shortens the life span dramatically due to the occurrence of tumors in virtually 100% of the mice.

p53 is a nuclear phosphoprotein. It was originally discovered in SV40-transformed cells, where it is associated with the T antigen coded by the virus. T antigen is needed to transform cells, and it was thought it might be acting through its effect on p53. A large increase in the amount of p53 protein is found in many transformed cells or lines derived from tumors. In early experiments, the introduction of cloned p53 was found to immortalize cells. These experiments caused p53 to be classified as an oncogene, with the usual trait of dominant gain-of-function.

But all the transforming forms of p53 turned out to be mutant forms of the protein! They fall into the category of dominant negative mutants, which function by overwhelming the wild-type protein and preventing it from functioning. The most common form of a dominant negative



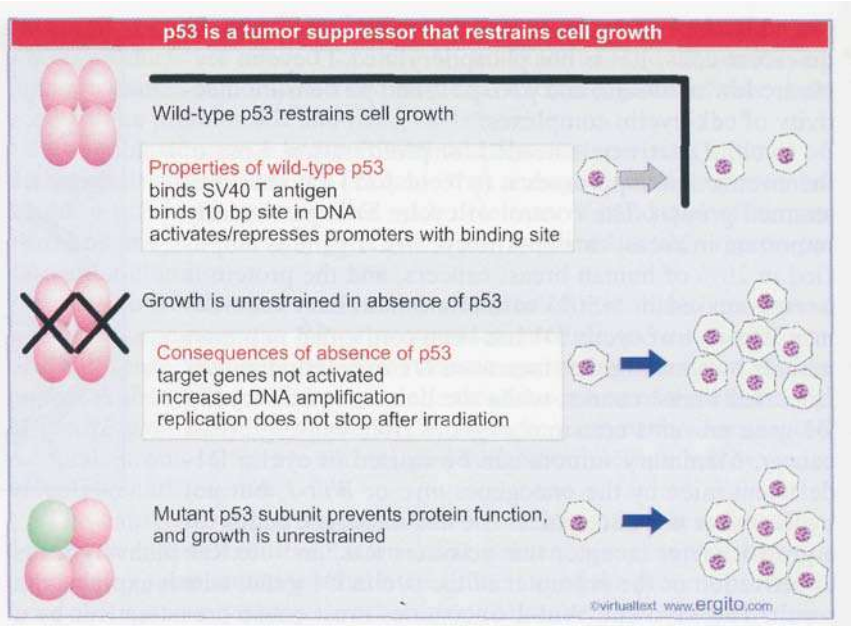


Figure 30.32 Wild-type p53 is required to restrain cell growth. Its activity may be lost by deletion of both wild-type alleles or by a dominant mutation in one allele.

mutant is one that forms a heteromeric protein containing both mutant and wild-type subunits, in which the wild-type subunits are unable to function. p53 exists as a tetramer. When mutant and wild-type subunits of p53 associate, the tetramer takes up the mutant conformation.

Figure 30.32 shows that the same phenotype is produced either by the deletion of both alleles or by a missense point mutation in one allele that produces a dominant negative subunit. Both situations are found in human cancers. Mutations in p53 accumulate in many types of human cancer, probably because loss of p53 provides a growth advantage to cells; that is, wild-type p53 restrains growth. The diversity of these cancers suggests that p53 is not involved in a tissue-specific event, but in some general and rather common control of cell proliferation; and the loss of this control may be a secondary event that occurs to assist the growth of many tumors. p53 is defined as a tumor suppressor also by the fact that wild-type p53 can suppress or inhibit the transformation of cells in culture by various oncogenes. Mutant p53 cells also have an increased propensity to amplify DNA, which is likely to reflect p53's role in the characteristic instability of the genome that is found in cancer cells.

Mutation in p53 is a cause of Li-Fraumeni syndrome, which is a rare form of inherited cancer. Affected individuals display cancers in a variety of tissues. They are heterozygotes that have missense mutations in one allele. These mutations behave as dominant negatives, overwhelming the function of the wild-type allele. This explains the occurrence of the disease as an autosomal dominant.

All normal cells have low levels of p53. A paradigm for p53 function is provided by systems in which it becomes activated, the most usual cause being irradiation or other treatments that damage DNA. This results in a large increase in the amount of p53. Two types of event can be triggered by the activation of p53: growth arrest and apoptosis (cell death). The outcome depends in part on which stage of the cell cycle has been reached. **Figure 30.33** shows that in cells early in G1, p53 triggers a checkpoint that blocks further progression through the cell cycle. This allows the damaged DNA to be repaired before the cell tries to enter S phase. But if a cell is committed to division, then p53 triggers a program of cell death. The typical results of this apoptosis are the collapse of the cell into a small heteropycnotic mass and the fragmentation of nuclear DNA (see *29.25 Apoptosis is a property of many or all cells).* The stage of the cell cycle is not the only determinant of the outcome; for example, some cell types are more prone to show an apoptotic response than others.

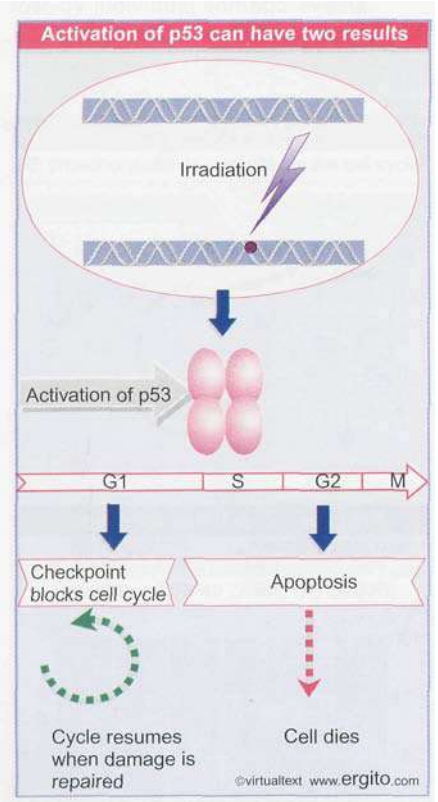


Figure 30.33 Damage to DNA activates p53. The outcome depends on the stage of the cell cycle. Early in the cycle, p53 activates a checkpoint that prevents further progress until the damage has been repaired. If it is too late to exercise the checkpoint, p53 triggers apoptosis.

By Book_Crazy [IND]

We may rationalize the existence of these two outcomes by supposing that damage to DNA can activate oncogenic pathways, and that the purpose of p53 is to protect the organism against the consequences. If it is possible, a checkpoint is triggered to allow the damage to be repaired, but if this is not possible, the cell is destroyed. We do not know in molecular terms how p53 triggers one pathway or the other, depending on the conditions, but we have an understanding of individual activities of p53 that may be relevant to these pathways.

30.21 p53 is a DNA-binding protein

Key Concepts

- p53 binds to promoters that contain a 10 bp recognition sequence, and may (usually) activate or (less often) inactive transcription.
- p53 also binds to single-stranded regions that are generated in damaged DNA, including those at the telomeres.
- p53 activates the CKI p21, which inhibits the cell cycle in the G1 phase.
- p53 activates the GADD45 repair protein that responds to radiation damage.
- The pathway leading to apoptosis has not been identified.
- p53 is usually present at low levels and has a short half-life.

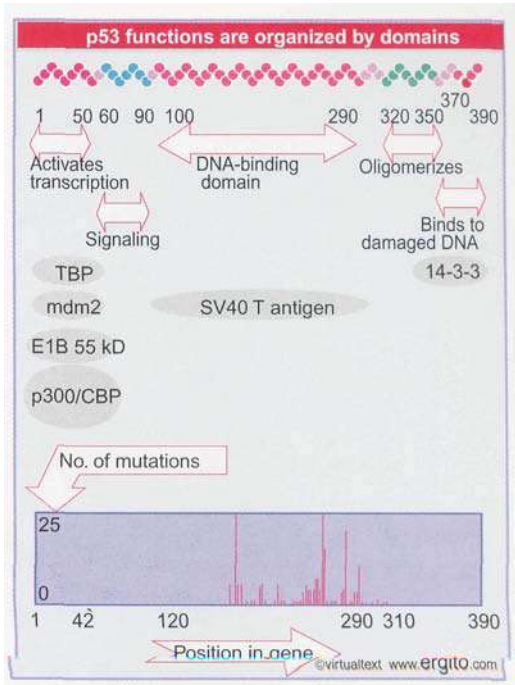


Figure 30.34 Different domains are responsible for each of the activities of p53.

P53 has a variety of molecular activities. **Figure 30.34** summarizes the responsibilities of individual domains of the protein for these activities:

- p53 is a DNA-binding protein that recognizes an interrupted palindromic 10 bp motif. The ability to bind to its specific target sequences is conferred by the central domain.
- p53 activates transcription at promoters that contain multiple copies of this motif. The immediate N-terminal region provides the transactivator domain. p53 may repress other genes; the mechanism is unknown.
- p53 also has the ability to bind to damaged DNA. The C-terminal domain recognizes single-stranded regions in DNA.
- p53 is a tetramer (oligomerization is a prerequisite for mutants to behave in a dominant negative manner). Oligomerization requires the C-terminal region.
- A (putative) signaling domain contains copies of the sequence PXXP, which forms a binding site for SH3 domains.

Mutations in p53 have various effects on its properties, including increasing its half-life from 20 minutes to several hours, causing a change in conformation that can be detected with an antibody, changing its location from the nucleus to the cytoplasm, preventing binding to SV40 T antigen, and preventing DNA-binding. As shown in Figure 30.34, the majority of these mutations map in the central DNA-binding domain, suggesting that this is an important activity.

p53 activates various pathways through its role as a transcription factor. The pathways can be divided into the three groups summarized in **Figure 30.35.** The major pathway leading to inhibition of the cell cycle at G1 is mediated via activation of p21, which is a CKI (cell cycle inhibitor) that is involved with preventing cells from proceeding through Gl (see Figure 29.30 and Figure 30.30). Activation of GADD45 identifies the pathway that is involved with maintaining genome stability. GADD45 is a repair protein that is activated also by other pathways that respond to irradiation damage.

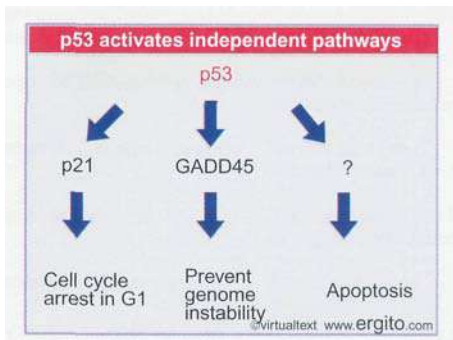


Figure 30.35 p53 activates several independent pathways. Activation of cell cycle arrest together with inhibition of genome instability is an alternative to apoptosis.

When it functions as a transcription factor, p53 uses the central domain to bind to its target sequence. The N-terminal transactivation domain interacts directly with TBP (the TATA box-binding protein). This region of p53 is also a target for several other proteins. An interaction with E1B 55 kD enables adenovirus to block p53 action, which is an essential part of its transforming capacity. Other regions of p53 can also be targets for inhibition; the SV40 T antigen binds to the specific DNA-binding region, thereby preventing the recognition of target genes.

The stability of p53 is an important parameter. It usually has a short half-life. The response to DNA damage stabilizes the protein and activates p53's transactivation activity.

To function as a transcription factor, p53 requires the coactivators p300/CBP (which are also used by many other transcription factors). The coactivator binds to the transactivation (N-terminal) domain of p53. The interaction between p53 and p300 is also necessary in order for p53 to bind the protein Mdm2, which inhibits its activity (see next section).

The C-terminal domain of p53 binds without sequence-specificity to short (<40 base) single-stranded regions of DNA and to mismatches generated by very short (1-3 base) deletions and insertions of bases. Such targets are generated by DNA damage. One important example of a single-stranded region that activates p53 is the overhanging end that is generated at the telomere of a senescent cell (see *30.25 Immortalization depends on loss of p53*).The consequence of this interaction is to activate the sequence-specific binding activity of the central domain, so that p53 stimulates transcription of its target genes. The nature of this connection is not clear, but may be a two-stage process. When p53 binds through its C-terminal domain to a damaged site on DNA, a change occurs in its properties; it then dissociates from the damaged site and binds to a target gene, which it activates.

The ability of p53 to trigger apoptosis is less well understood. It probably depends on the transactivation of a different set of target genes from those involved in activating the G1 checkpoint. The two activities can be separated by the response to adenovirus E1B 19 kD protein, which blocks the apoptotic activity of p53, but does not block its activity to activate target genes. The independence of the effects of p53 on growth arrest and apoptosis is emphasized by the fact that the E1B 55 kD protein blocks transactivation capacity but does not interfere with apoptosis.

p53 can activate apoptosis pathways in two ways. One is to cause the production of proteins that act on the mitochondrion to trigger its apoptotic functions (see *29.28 Apoptosis involves changes at the mitochondrial envelope).* The other is to produce or activate the cell surface receptors that trigger apoptosis (see *29.26 The Fas receptor is a major trigger for apoptosis).*

The importance of the connection between tumorigenesis and loss of apoptosis is also shown by the properties of the *bcl2* oncogene. *bcl2* was originally identified as a target that is activated by translocations in certain tumors. It turns out to have the property of inhibiting most pathways for apoptosis (see *29.28 Apoptosis involves changes at the mitochondrial envelope).* This suggests that apoptosis plays an important role in inhibiting tumorigenesis, probably because it eliminates potentially tumorigenic cells. When apoptosis is prevented because *bcl2* is activated, these cells survive instead of dying.

Cells with defective p53 function have a variety of phenotypes; this pleiotropy makes it difficult to determine which (if any) of these effects is directly connected to the tumor suppressor function. Most of our knowledge about p53 action comes from situations in which it has been activated. We assume that the pathways it triggers—growth arrest or apoptosis—are connected to its ability to suppress tumors. Certainly it is

clear that the failure of p53 to respond to DNA damage is likely to increase susceptibility to mutational changes that are oncogenic. However, we do not know whether this is the sole role played by p53. $p53^{-}$ mice develop normally, implying that p53's role is not essential for development.

The general definition of their properties shows that both RB and p53 are tumor suppressors that usually control cell proliferation; their absence removes this control, and contributes to tumor formation.

30.22 p53 is controlled by other tumor suppressors and oncogenes

Key Concepts

- Most human cancer cells either have mutations that inactivate p53 directly or have mutations in other loci that lead to loss of p53.
- p53 is destabilized by Mdm2, which targets it for degradation and also directly inhibits its transactivation activity.
- The INK4a-ARF locus codes for $p16^{INK4a}$, which controls RB, and for $p19^{ARF}$, which controls p53 via inactivating Mdm2.
- Loss of INK4a-ARF is a common cause of human cancer.

P53 does not function correctly in most human tumors, but the cause of the problem lies with the gene itself in only about half of the cases. **Figure 30.36** summarizes the causes of p53 deficiency:

- The mutations in p53 itself most often lie in the DNA-binding region, and prevent the protein from binding to promoters to activate the protective response. In some cases, the mutations lie in the C-terminal region that is responsible for forming tetramers, so that active proteins are not produced.
- The major pathway controlling p53 is mediated through the protein Mdm2, which inactivates p53, so the Mdm2 locus behaves as an oncogene. Amplification of the Mdm2 gene causes an increase in expression of the protein, which reduces p53 function.
- Mdm2 is itself inactivated by the protein $p19^{ARF}$, so deletions of the $p19^{ARF}$ gene lead to increase in Mdm2 and thus to decrease in p53.

Because p53 inhibits growth or triggers apoptosis when it is activated, it is obviously crucial for the cell to restrain the activity unless it is needed. The circuitry that controls p53's activity is illustrated in the upper part of **Figure 30.37.** Proteins that activate p53 behave as tumor suppressors; proteins that inactivate p53 behave as oncogenes.

A major feature in controlling p53 activity is its interaction with Mdm2 (which was originally identified as the product of an oncogene). Mdm2 inhibits p53 activity in two ways:

Mdm2 affects p53's stability by acting as an E3 ubiquitin ligase that causes p53 to be targeted by the degradation apparatus.

Mdm2 also acts directly at the N-terminus to inhibit the transactivation activity of p53.

In the reverse direction, p53 induces transcription of Mdm2. The consequence of this circuit is that Mdm2 limits p53 activity; and the activation of p53 increases the amount of Mdm2, so the interaction between p53 and Mdm2 forms a negative feedback loop in which the two components limit each other's activities.

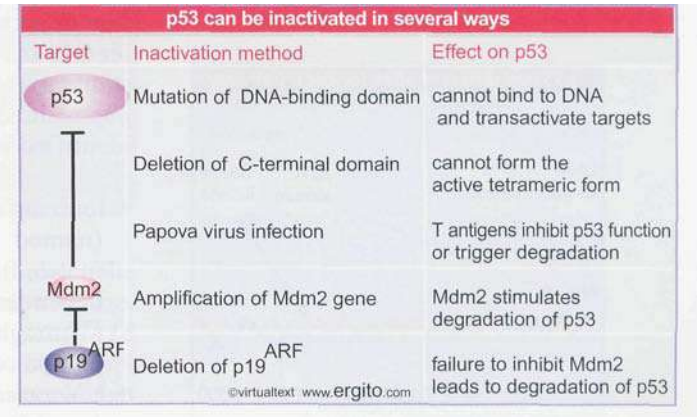


Figure 30.36 p53 is inactivated in human cancers as the result of mutations directly affecting its function or in other genes that affect the level of protein.

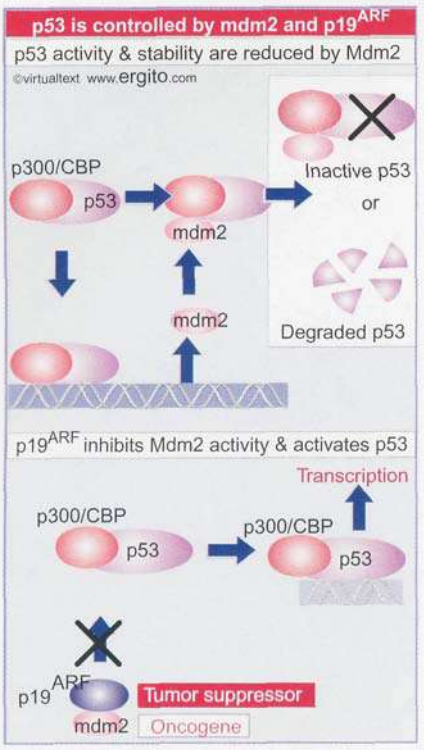


Figure 30.37 p53 activity is antagonized by Mdm2, which is neutralized by $p19^{ARF}$.

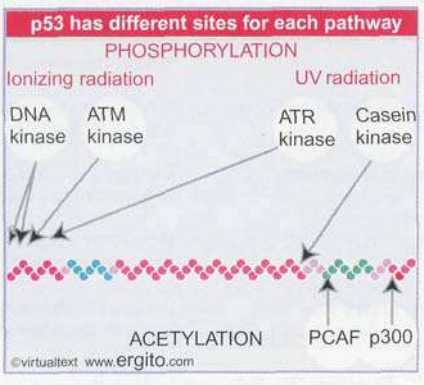


Figure 30.38 Each pathway that activates p53 causes modification of a particular set of residues.

INK4A-ARF is an important locus that controls both p53 and RB. The transcript of the INK4A-ARF gene is alternatively spliced to give two mRNAs that code for proteins with no sequence relationship. $p16^{INK4a}$ is upstream of RB. The second protein is called $p19^{ARF}$ in mouse and $p14^{ARF}$ in man. We will use $p19^{ARF}$ to describe it irrespective of source. As we have just seen, p19 ^{F} is upstream of p53. Deletions of the locus are common in human cancers (almost as common as mutations in p53), and have a highly significant effect, because they eliminate both $p16^{INK4a}$ and p19 and therefore lead to loss of both the RB and p53 tumor suppressor pathways.

$p16^{INK4a}$ inhibits the cdk4/6 kinase (see Figure 29.30). So it prevents the kinase from phosphorylating RB. In the absence of this phosphorylation, progress through the cell cycle (and therefore growth) is inhibited. $p16^{IN}$ is often inhibited by point mutations in human tumors.

$p19^{ARF}$ antagonizes Mdm2, as shown in the lower part of Figure 30.37. $p19^{ARF}$ binds to Mdm2 and directly prevents it from ubiquitinating p53. This stabilizes p53 and allows it to accumulate. In effect, therefore, $p19^{ARF}$ functions as a tumor suppressor by inhibiting the inhibitor of the p53 tumor suppressor. Loss of $p19^{ARF}$ or loss of p53 have similar effects on cell growth (and tumors usually lose one or the other but not both), suggesting that they function in the same pathway, that is, $p19^{ARF}$ in effect functions exclusively through p53. The cellular oncogene c-myc, and the adenoviral oncogene E1, both act via $p19^{ARF}$ to activate p53-dependent pathways.

30.23 p53 is activated by modifications of amino acids

Key Concepts

- p53 is modified by (mostly) phosphorylation or by acetylation in response to treatments that damage DNA.
- Different sensor pathways act through kinases that modify different target sites on p53.

P53 responds to environmental signals that affect cell growth, and many of these signals act by causing specific sites on p53 to be modified. The most common form of modification is the phosphorylation of serine, but acetylation of lysine also occurs. Different pathways lead to the modification of different amino acid residues in p53, as summarized in **Figure 30.38.** There is often overlap between the various residues activated by each pathway. Three principal pathways are identified by the agents that act on p53:

- Ionizing radiation induces DNA breaks that activate the kinase ATM (named for the disease ataxia telangiectasia in which its gene is mutated). ATM phosphorylates S^{15}. DNA breaks also activate a DNA-dependent kinase that acts on other sites in the N-terminal region. (Through unknown pathways, ionizing radiation also causes phosphorylation of S^{33}, dephosphorylation of S^{376}, and acetylation of L^{382}.)
  UV radiation and other types of stress activate the kinases ATR (ataxia telangiectasia related) and casein kinase II, which phosphorylate S^{15} and S^{33}, and also cause phosphorylation of S^{392}.
  Some aberrant growth signals, such as those produced by the oncogenes Ras or Myc, may activate $p19^{ARF}$. This inactivates Mdm2 and thus activates p53.

By Book_Crazy [IND]

The target sites for these various pathways are located in the terminal regulatory domains of the protein. The modifications may affect stability of the protein, oligomerization, DNA-binding, and binding to other proteins. So p53 acts as a sensor that integrates information from many pathways that affect the cell's ability to divide. *The important point is that each pathway leads to modification of specific residues in p53 that activate its response.*

Phosphorylation can change the properties of a protein by altering the structure in the immediate vicinity, creating or abolishing sites that interact with other proteins. At some of the sites in p53, the phosphorylation has a more widespread effect, and changes the conformation of the polypeptide backbone of the protein. The usual conformation of a protein chain is a series of amino acids arranged in *trans* conformation about the peptide bonds that connect them. This means that the adjacent amino acids lie on opposite sides of the peptide bond. However, as shown in **Figure 30.39,** the ring structure of proline allows a rearrangement in which it lies in *cis* configuration relative to the preceding amino acid. This has a major effect on protein conformation. The reaction is catalyzed by a class of proteins called peptidyl-prolyl-isomerases.

Some of the target sites for phosphorylation in p53 are serine or threonine residues followed by a proline. Phosphorylation changes the energetics to favor the *trans-cis* isomerization. The reaction is assisted by a particular peptidyl-prolyl-isomerase, Pin1, which binds to the dipeptide sequence only when the first amino acid is phosphorylated. The importance of the reaction is demonstrated by the fact that mutations in Pin1 impair the ability of p53 to respond to damage in DNA. Although Pin1 is essentially part of the mechanism by which p53 is activated, in formal terms it behaves like a tumor suppressor.

There are two aspects to the activation of p53. The amount of p53 protein in the cell is determined principally by its degradation, which is a reflection of the activity of Mdm2. Stabilization of the protein is a prerequisite for the response, but is not sufficient. Conformational changes must be triggered by one or more of the modification events that respond to the various sensor systems.

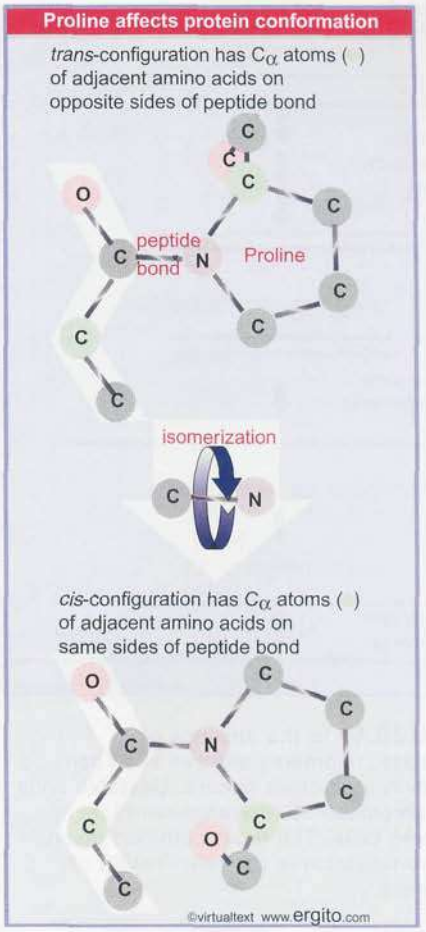


Figure 30.39 Amino acids usually lie in *trans* on either side of the peptide bond connecting them in a protein chain. The ring structure of proline allows isomerization around the bond connecting it to the preceding amino acid, generating the *cis* configuration in which both amino acids lie on the same side of the peptide bond.

30.24 Telomere shortening causes cell senescence

Key Concepts

- Somatic cells usually lack telomerase activity, which means that telomeres shorten with each cell division.
- Cultured cells may go into crisis as the result of reaching zero telomere length.
- Reactivation of telomerase enables cells to survive crisis and to become immortal.

A key event in the limited divisions of growth in cell culture is the shortening of telomeres. A special ribonucleoprotein enzyme, called telomerase, is responsible for extending telomeres (see *19.18 Telomeres are synthesized by a ribonucleoprotein enzyme).* The function of telomerase is to compensate for the shortening of telomeres that occurs at each replication cycle. Telomerase is turned off in many somatic cells, typically when differentiation occurs.

Continued division in cells that lack telomerase activity (for example, when primary cells are placed into culture) will cause the telomeres to shorten in each generation. The cells become unable to propagate

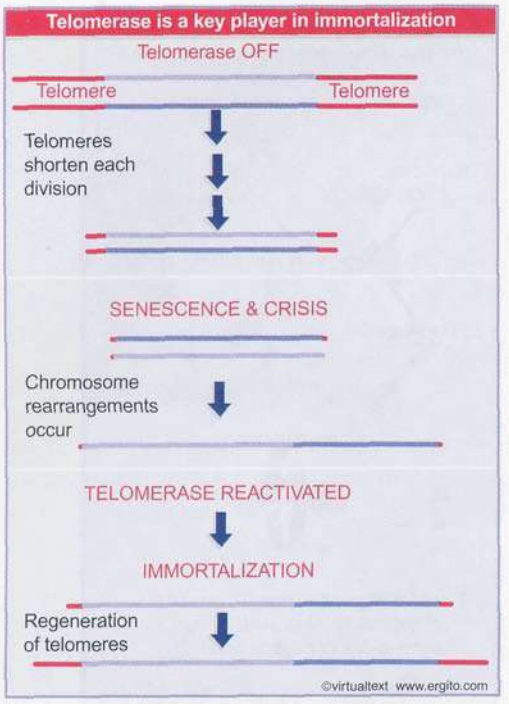


Figure 30.40 In the absence of telomerase, telomeres shorten each cell generation until crisis occurs. Unstable ends cause chromosome rearrangements in senescent cells. The most common means of immortalization is the reactivation of telomerase.

properly when the telomeres become too short to ensure stability at the ends of the chromosomes (see *19.19 Telomeres are essential for survival)*. The consequences are visible as changes in the appearance of the culture, and the cells are said to be senescent. Crisis occurs when the cells cannot divide any longer, and the culture dies out. Figure 30.40 summarizes the events involved in approaching crisis and passing through it.

The rare cells that survive crisis pass through a stage at which the ends of their chromosomes are unstable. These ends interact with one another or with other chromosomal regions, and this is the probable cause of the frequent occurrence of chromosomal abnormalities in cultured cells. The mutations that are caused by these abnormalities can contribute to the tumorigenic state.

To continue to divide after passing through crisis, a cell must regain the capacity to replicate its telomeres. This suggests that a critical parameter for immortalization might be the reactivation of telomerase. Telomerase activity can be restored by transfecting the gene for the catalytic subunit into target cells, and this allows them to be perpetuated in culture without passing through crisis. We might view the finite replicative capacity of primary human cells, in "general, or their inability to continue propagation once telomere lengths have become too short, in particular, as a tumor suppression mechanism that in effect prevents cells from undertaking the indefinite replication that is needed to make a tumor.

Immortalization is required for cells to be perpetuated indefinitely in tissue culture, and we have to ask how the relevant events relate to the formation of a tumor *in vivo*. The behavior of telomerase shows a parallel between the immortalization of a cell in culture and the generation of a tumor *in vivo*. The limiting step in production of telomerase is the transcription of the catalytic subunit, which is repressed in differentiated somatic cells and restored in tumor cells. This resembles the reactivation of telomerase activity in cultured cells that have emerged from crisis.

Can we test the role of telomerase in intact animals? Mice in which the gene coding for the telomerase RNA has been inactivated can survive for >6 generations. Mouse telomeres are exceptionally long, and range from 10-60 kb. In the absence of telomerase, telomeres shorten at 50-100 bp per cell division. There are ~60 divisions in sperm cell production, and ~25 divisions in oocyte production, which fits with the observed rate of shortening of ~4.8 kb per male mouse generation. This gives an expectation that after about 7 generations, a telomerase-negative mouse will have run down its telomeres to around zero length.

By the 6th generation, chromosomal abnormalities become more frequent, and the mice become infertile (due to the inability to produce sperm). The effects of lack of telomerase are first seen in tissues consisting of highly proliferative cells (as might be expected). All of these observations demonstrate the importance of telomerase for continued cell division. However, cells from the telomerase-negative mice can pass through crisis and can be transformed to give tumorigenic cells, so the presence of telomerase is not essential, or at least is not the only means of supporting an immortal state (although reactivation of telomerase is by far the most common mechanism) (see *19.19 Telomeres are essential for survival).*

Telomerase-negative mice can develop tumors, but do so at a rate lower than wild-type mice. The effect of telomere loss on formation of a cancer cell is therefore confined to its role in provoking a genetic instability that stimulates tumor initiation. After that, it is in fact inhibitory to cancer formation.

There is a curious inconsistency between the results obtained with cultured cells and the survival of telomerase-negative mice. Crisis of mouse cells occurs typically after 10-20 divisions in culture, but we

would not expect the telomeres to have reached a limiting length at this point. Mice of the first telomerase-negative generation have passed a greater number of cell divisions without telomerase, and without suffering any ill effects. Mice of the third telomerase-negative generation are to all intents and purposes normal, although their cells have gone through more divisions than would have triggered crisis in culture.

Lack of telomerase is clearly associated with inability to continue growth, and reactivation of telomerase is one means by which cells can behave as immortal. It is not clear whether telomerase is the only relevant factor in driving cells into crisis and to what extent other mechanisms might be able to compensate for lack of telomerase. We do not know what pathway is responsible for controlling telomerase production *in vivo*, and how it is connected to pathways that control cell growth.

30.25 Immortalization depends on loss of p53

Key Concepts

- Loss of p53 is the crucial step in immortalization.

When cells enter senescence as the result of telomere shortening, p53 is activated, leading to growth arrest or apoptosis. **Figure 30.41** shows that the trigger that activates p53 is the loss of the telomere-binding protein TRF2 from the chromosome ends. In effect, TRF2 protects the end of the DNA, but when it is lost, the free 3' overhanging end activates p53.

The loss of p53 is the crucial event that allows the cells to survive and divide. A variety of events can be associated with immortalization, but they converge upon causing either the loss or inactivation of p53 protein. Remember that p53 was discovered as the protein of host cells that binds the T antigen (the transforming protein) of polyomaviruses. A major part of the activity of T antigen is its ability to inactivate p53. The T antigens of different viruses work in different ways, but the consequences of the interaction are especially clear in the case of HPV E6 (the equivalent of T antigen), which targets p53 for degradation. In effect, HPV converts a target cell into a $p53^-$ state.

p53 provides an important function in immortalization, but may not be sufficient by itself. Established cell lines have usually lost p53 function, which suggests that the role of p53 is connected with the acquisition of ability to support prolonged growth. However, loss of the known functions of p53 is not enough by itself to explain immortalization, since, for example, a $p53^-$ mouse is viable, and therefore is able to undergo the usual pattern of cell cycle arrest and differentiation. Primary cells from a $p53^-$ mouse can pass into the established state more readily than cells that have p53 function, which suggests that loss of p53 activity facilitates or is required for immortalization.

An interesting convergence is seen in the properties of the tumor antigens from different DNA tumor viruses. The antigens always bind to both the cellular tumor suppressor products RB and p53. The two cellular proteins are recognized independently. Either different T antigens of the virus bind separately to RB and to p53, or different domains of the same antigen do so. So adenovirus E1A binds RB, while E1B binds p53; HPV E7 binds RB, while E6 binds p53. SV40 T antigen can bind both RB and p53. The loss of p53 (and/or RB) is a major step in the transforming action of DNA tumor viruses, and explains some significant part of the action of the T antigens. The critical events are inhibition of

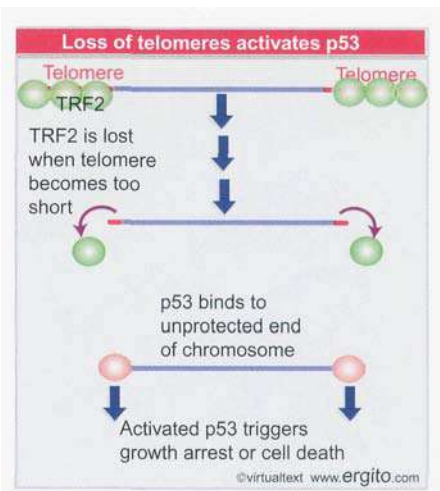


Figure 30.41 TRF2 protects telomeres, but when it is lost, the exposed ends can bind and activate p53.

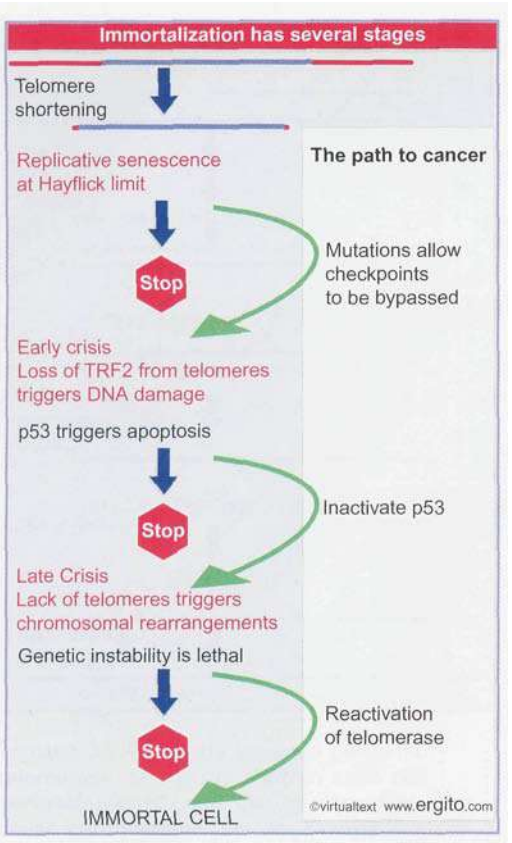


Figure 30.42 Several changes are required to allow a cell to pass the replicative limit and to become immortalized, including bypassing the checkpoints that respond to short telomeres, losing or preventing the ability of p53 to trigger apoptosis, and reactivating telomerase or finding other means to stabilize telomeres.

p53's ability to activate transcription, and loss of RB's ability to bind substrates such as E2F. Loss of the tumor suppressors (especially p53) is the major route in the immortalization pathway.

Inability to trigger either growth arrest or apoptosis could lead to continued growth. We do not know whether only one or both of these activities are required for immortalization *in vitro*. We know that more than the growth arrest pathway is needed for p53's contribution to tumorigenesis, because a p21~ mouse shows deficiencies in the G1 checkpoint (as would be expected) but does not develop tumors. The contrast with the increased susceptibility of a p53⁻ mouse to tumors shows that other functions of p53 are involved besides its control of p21.

We are now in a position to put together the various events involved in immortalization. **Figure 30.42** summarizes the order in which they occur. Checkpoints normally stop a cell from dividing when its telomeres become too short. Cells then enter replicative senescence and stop growing. If they manage to bypass the checkpoints to enter early crisis, the loss of TRF2 from the telomeres activates p53, which causes growth arrest and/or apoptosis. In the absence of p53 activity, they pass into late crisis, where the dysfunction of the telomeres causes genetic instability as seen in large scale chromosomal rearrangements. To survive this stage, they must activate telomerase or find an alternative means of maintaining the telomeres. A cell that survives through all of these stages will be immortal, but almost certainly will have an altered genetic constitution.

30.26 Different oncogenes are associated with immortalization and transformation

Key Concepts

- A tumor cell has independently acquired changes necessary to immortalize it and to transform it.
- Established cell lines grown in culture usually have been immortalized and need to acquire only transforming properties.
- Primary cells require the actions of different oncogenes to be immortalized and to be transformed.

Most tumors arise as the result of multiple events. Some of these events involve the activation of oncogenes, while others take the form of inactivation of tumor suppressors. The requirement for multiple events reflects the fact that normal cells have multiple mechanisms to regulate their growth and differentiation, and several separate changes may be required to bypass these controls. Indeed, the existence of single genes in which mutations were tumorigenic would no doubt be deleterious to the organism, and has been selected against. Nonetheless, oncogenes and tumor suppressors define genes in which mutations create a predisposition to tumors, that is, they represent one of the necessary events. It is an open question as to whether the oncogenes and tumor suppressor genes identified in available assays are together sufficient to account entirely for the occurrence of cancers, but it is clear that their properties explain at least many of the relevant events.

Figure 30.43 gives an overview of the stages of tumor formation. There are two discrete stages, which may loosely be viewed as being concerned with immortalization or with transformation.

The immortalization step is to bypass crisis (or its equivalent in the *in vivo* situation). Crisis is provoked when cells continue to divide in the absence of telomerase (see *30.24 Telomere shortening causes cell senescence).* When the telomeres become too short, damage to DNA is

caused by attempts at replication, and this triggers the activation of p53. The role of p53 is to cause cell death. If p53 is absent, a cell may survive, although at the expense of a genetic catastrophe in which telomere malfunction leads to chromosome fusions and other rearrangements.

Immortalized cells can pass through an unlimited number of cell divisions, but they do not have other tumorigenic properties, such as independence of factors required for growth. The second step, transformation, converts immortalized cells into tumorigenic cells. Whether further changes are involved in creating a cancerous state depends on the nature of the tumor cell. A leukemia cell can multiply freely in the blood. However, a cell type that forms a solid tumor needs to develop a blood supply for the tumor (requiring angiogenic development), and may later pass to the stage of metastasis, when cells are detached from the tumor and migrate to form new tumors at other locations.

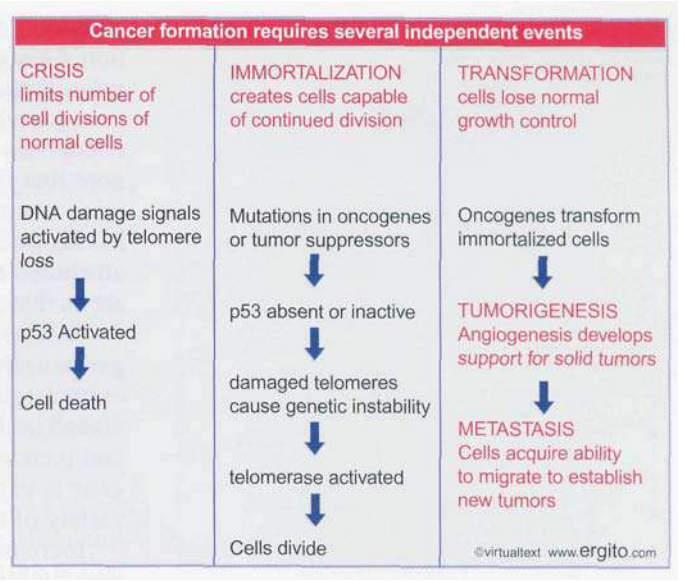


Figure 30.43 Crisis is induced by attempts to divide in the absence of telomerase. Immortalization occurs when p53 activity is lost, as the result of mutation in p53 or in a pathway that acts on it. Transformation requires oncogenes to induce further changes in the growth properties of the cells. For cells that develop solid tumors, angiogenic development is required for provision of nutrients. Metastasis is the result of further changes that allow a cell to migrate to form a colony in a new location.

The minimum requirement to enter the tumorigenic state is therefore the occurrence of successive, independent events that involve different tumor suppressors and/or oncogenes. The need for multiple functions of different types is sometimes described as the requirement for cooperativity.

The involvement of multiple functions fits with the pattern established by some DNA tumor viruses, in which (at least) two types of functions are needed to transform the usual target cells:

- Adenovirus carries the *E1A* region, which allows primary cells to grow indefinitely in culture, and the *E1B* region, which causes the morphological changes characteristic of the transformed state.
- Polyoma produces three T antigens; large T elicits indefinite growth, middle T is responsible for morphological transformation, and small T is without known function. Large T and middle T together can transform primary cells.
- Consistent with the classification of oncogenic functions, adenovirus E1A together with polyoma middle T can transform primary cells. This suggests that one function of each type is needed.

In the same way, expression of two or more oncogenes in the cellular transfection assay is usually needed to convert a primary cell (one taken directly from the organism) into a tumor cell.

Several cellular oncogenes have been identified by transforming ability in the 3T3 transfection assay; 10-20% of spontaneous human tumors have DNA with detectable transforming activity in this assay. Of course, 3T3 cells have been adapted to indefinite growth in culture over many years, and have passed through some of the changes characteristic of tumor cells. The exact nature of these changes is not clear, but generally they can be classified as involving functions concerned with immortalization. Oncogenic activity in this assay therefore depends on the ability to induce further changes in an established cell line.

The principal products of 3T3 transfection assays are mutated *c-ras* genes. They do not have the ability to transform primary cells *in vitro*, and this supports the implication that their functions are concerned with the act of transforming cells that have previously been immortalized. *ras* oncogenes clearly provide one major pathway for transforming immortalized cells; we do not know how many other transforming pathways may exist that are independent of *ras*.

Although *ras* oncogenes alone cannot transform primary fibroblasts, dual transfection with *ras* and another oncogene can do so. The ability to transform primary cells in conjunction with *ras* provides

a general assay for oncogenes that have an immortalization-like function. This group includes several retroviral oncogenes, *v-myc*, *v-jun*, and *v-fos*. It also includes adenovirus E1A and polyoma large T. Mutant p53 genes have the same effect. In fact, the action of the immortalizing oncogenes is most likely to cause inactivation or loss of p53. However, note that the distinction between immortalizing and transforming proteins is not crystal clear. For example, although E1A is classified as having an immortalizing function, it has (some) of the functions usually attributed to transforming proteins, and loss of p53 confers some properties that are usually considered transforming.

One way to investigate the oncogenic potential of individual oncogenes independently of the constraints that usually are involved in their expression is to create transgenic animals in which the oncogene is placed under control of a tissue-specific promoter. A general pattern is that increased proliferation often occurs in the tissue in which the oncogene is expressed. Oncogenes whose expression have this effect with a variety of tissues include SV40 T antigen, *v-ras*, and *c-myc*.

Increased proliferation (hyperplasia) is often damaging and sometimes fatal to the animal (usually because the proportion of one cell type is increased at the expense of another). However, the expression of a single oncogene does not usually cause malignant transformation (neoplasia), with the production of tumors that kill the animal. Tumors resulting from the introduction of an oncogene (for example, in transgenic mice) are probably due to the occurrence of a second event.

The need for two types of event in malignancy is indicated by the difference between transgenic mice that carry either the *v-ras* or activated *c-myc* oncogene, and mice that carry both oncogenes. Mice carrying either oncogene develop malignancies at rates of 10% for *c-myc* and 40% for *v-ras;* mice carrying both oncogenes develop 100% malignancies over the same period. These results with transgenic mice are even more striking than the comparable results on cooperation between oncogenes in cultured cells.

In some systems, immortalization may be connected with an inability of the cells to differentiate. Growth and differentiation are often mutually exclusive, because a cell must stop dividing in order to differentiate. An oncoprotein that blocks differentiation may allow a cell to continue proliferating (in a sense resembling the immortalization of cultured cells); continued proliferation in turn may provide an opportunity for other oncogenic mutations to occur. This may explain the occurrence among the oncoproteins of products that usually regulate differentiation.

A connection between differentiation and tumorigenesis is shown by avian erythroblastosis virus (AEV). The AEV-H strain carries only *v-erbB*, but the AEV-E54 strain carries two oncogenes, *v-erbB* and *v-erbA*. The major transforming activity of AEV is associated with *v-erbB*, a truncated form of the EGF receptor, which is equivalent to the single oncogene carried by other tumor retroviruses: it can transform erythroblasts and fibroblasts. The other gene, *v-erbA,* cannot transform target cells alone, but it increases the transforming efficiency of *v-erbB.* Expression of *v-erbA* itself has two phenotypic effects upon target cells: it prevents the spontaneous differentiation (into erythrocytes) of erythroblasts that have been transformed by *v-erbB;* and it expands the range of conditions under which transformed erythroblasts can propagate. *v-erbA* may therefore contribute to tumorigenicity by a combination of inhibiting differentiation and stimulating proliferation. In fact, *v-erbA* has a similar effect in extending the efficacy of transformation by other oncogenes that induce sarcomas, notably *v-src, v-fps,* and *v-ras*.

Correlations between the activation of oncogenes and the successful growth of tumors are strong in some cases, but by and large the nature of the initiating event remains open. It seems clear that oncogene activity assists tumor growth, but activation could occur (and be selected for)

after the initiation event and during early growth of the tumor. We hope that the functions of *c-onc* genes will provide insights into the regulation of cell growth in normal as well as aberrant cells, so that it will become possible to define the events needed to initiate and establish tumors.

30.27 p53 may affect aging

Key Concepts

- Shortening of telomeres below a critical length is associated with reduced longevity.
- Increase of p53 above wild-type levels can decrease tumor formation, but also decreases longevity.

We have very little idea what is responsible for aging of an animal. The general drift of evolutionary theories of aging is that natural selection operates only via reproduction, and therefore there is little advantage to the survival of the organism past the stage when it is reproductively active. In other words, there is no selection for longevity beyond the reproductive state.

Wedo not know how aging of the organism relates to changes in individual cells, but one possibility is that aging results from the accumulation of damage at the cellular level. Within this model, one contribution could be inappropriate expression of genes resulting from failure of regulation.

It is an open question whether aging of the organism is connected with the senescence of individual cells. Cessation of telomerase activity in adult lineages causes telomeres to shorten as cells divide. When telomere lengths reach zero, cells enter the senescent state. Shortened telomeres can reduce lifespan. **Figure 30.44** shows that mice from the fourth or fifth telomerase-negative generations have a slightly reduced lifespan, and mice from the sixth generation have a much reduced lifespan. Whereas normal mice have a 50% survival rate at ~25 months, the figure for the sixth generation mice is ~17 months. Increased cancer incidence accounts for only half of the accelerated deaths. The other mice die from unknown causes (this is typical of aging), and they prematurely show several of the characteristics of aging (such as reduced wound healing).

We know that loss of p53 is one of the mechanisms allowing cultured cells to pass through the crisis provoked by loss of telomeres (see *30.24 Telomere shortening causes cell senescence).* What role might p53 loss play in telomerase-negative mice? A major role is seen in the pattern of inheritance. An increased proportion of the progeny are $p53^{-}$. This is because loss of p53 reduces the apoptosis (death) of germ line cells that is caused by loss of telomerase.

Direct attempts to see whether $p53$ might have any effect on aging have been unsuccessful because $p53^{-}$ mice die early as the result of accumulating tumors, and mice that over-express p53 cannot be made, probably because of deleterious effects of excess p53 during embryonic development. However, striking results have been obtained from the serendipitous production of a mouse that has a mutant form of p53. The mutant gene, called the *m* allele, has lost its first 6 exons, and makes a truncated protein. Heterozygous $p53^{+/m}$ mice have a reduced frequency of tumor formation. Comparison with wild-type and hemizygous p53 mice shows the following:

p53 +/- (one active allele): >80% tumors
p53 +/+ (two active alleles): >45% tumors
p53 +/m: 6% tumors.

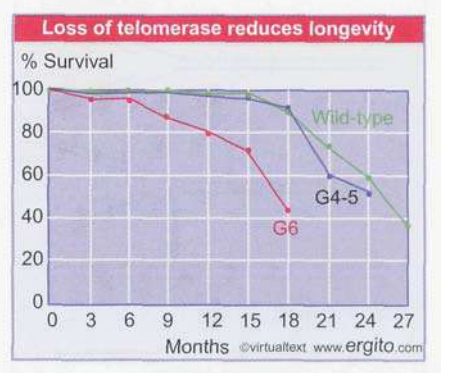


Figure 30.44 Telomerase-negative mice of the sixth generation (G6) have sharply reduced longevity, and mice of the fourth to fifth generations (G4-5) have slightly reduced longevity.

By Book_Crazy [IND]

Mice with two active alleles form tumors at about half the frequency of mice with only one active allele, which corresponds to the relative rate at which one allele is likely to be spontaneously inactivated compared to two alleles (see Figure 30.31). The much reduced rate of tumor formation in the $p53^{+/m}$ mice suggests that the *m* allele has the unexpected effect of increasing p53 activity. This is confirmed by directly measuring some of the known responses to p53 in cells from the $p53^{+/m}$ mice.

We might expect the reduction in tumor formation in $p53^{+/m}$ mice to be associated with an increase in longevity, but exactly the reverse is found when they are compared with wild-type mice. Although the $p53^{+/m}$ mice have only 6% tumors and the wild-type mice develop 45% tumors, half of the $p53^{+/m}$ mice have died by 22 months, whereas the wild-type mice survive on average to 27 months. (Mice that are p53 or $p53^{-/-}$ die more quickly because of the high accumulation of tumors.) The effect seems to result from the interaction of the *m* mutant protein with the wild-type protein in the heterozygote, because $p53^{-/m}$ mice have just as many tumors and die just as quickly as p53 mice. So the *m* allele does not have any protective effect on its own.

The $p53^{+/m}$ mice show no differences from wild-type mice for the first 12 months, but by 18 months show signs of premature aging. This suggests that increased activity of p53 has a direct effect in promoting aging. This raises the possibility that the very same activities of p53 that are needed for protection against cancer also have the effect of causing aging! This clearly makes the level of p53 activity something that must be very tightly controlled.

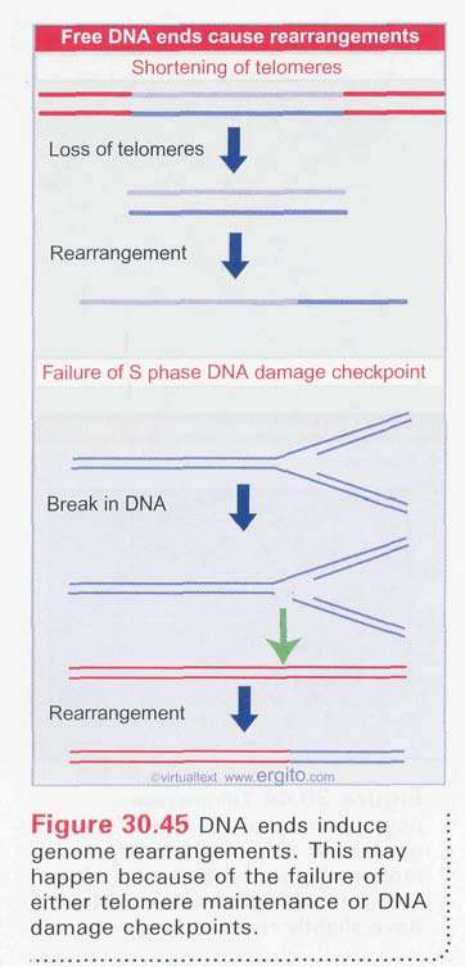


Figure 30.45 DNA ends induce genome rearrangements. This may happen because of the failure of either telomere maintenance or DNA damage checkpoints.

30.28 Genetic instability is a key event in cancer

Key Concepts

- Tumor cells have rates of genetic change that are increased above the usual rate of somatic mutation.
- Gross chromosomal changes are observed in most types of colorectal cancer.
- Chromosome rearrangements can be generated by mutations of checkpoint pathways and other pathways that act on the genome in a yeast model system.

The inactivation of tumor suppressors and the activation of oncogenes are key events in creating a tumor, but several such events (typically 4-10 in the case of human cancers) are required to generate a fully tumorigenic state. This number of events would not be predicted to occur during the life of a cell or organism if the individual changes occurred at the normal rate of spontaneous mutation. Many cancers are associated with genetic instability that significantly increases the number of events.

Genetic instability is revealed by increases in the frequency of genomic changes. These range from reorganizations at the level of the chromosome to individual point mutations. We can get a sense of their relative importance from their occurrence in colon cancers, where the genetic changes have been well characterized. The majority of colorectal tumors show high rates of gross alteration in chromosomes, often involving changes in the number of copies of a gene. This is the most common way to generate the basic changes that fuel tumorigenesis. In the minority of cases (~15%), there are no gross changes, but there are many individual mutations, resulting from a highly increased rate of

mutation (see next section). Either of these types of change in the cell can propagate a tumor; it is rare for both to occur together.

Gross chromosomal alterations involve deletion, duplication, or translocation. They may result in changes in the number of copies of a gene. **Figure 30.45** illustrates two major causes:

- loss of telomeres because cells continue to divide in the absence of telomerase;
- creation of free double-strand ends that result from an unrepaired break in DNA.

The events that occur when a cell passes through crisis are a paradigm for the generation of chromosomal alterations. Loss of telomeres induces DNA rearrangements (see *30.24 Telomere shortening causes cell senescence).* Failure of the normal protective mechanisms allows the damaged cells to survive.

Gross rearrangements can also be provoked by failure of protective mechanisms during a normal cell cycle. Checkpoint pathways respond to DNA damage by halting the cell cycle in its current phase (see Figure 29.21 in *29.13 DNA damage triggers a checkpoint).* The checkpoint triggers an effector pathway that repairs the damage, after which the cell cycle is allowed to proceed. Mutations in the S phase checkpoint pathway in *S. cerevisiae* result in an increase of more than 100 X in the rate of genome rearrangements. This happens because the cell cycle is allowed to proceed in spite of the presence of breaks in DNA. **Figure 30.46** shows that similar effects are produced by mutations of some recombination-repair pathways or pathways for telomere maintenance. Analogous events could be involved in creating genetic instabilities that lead to cancer.

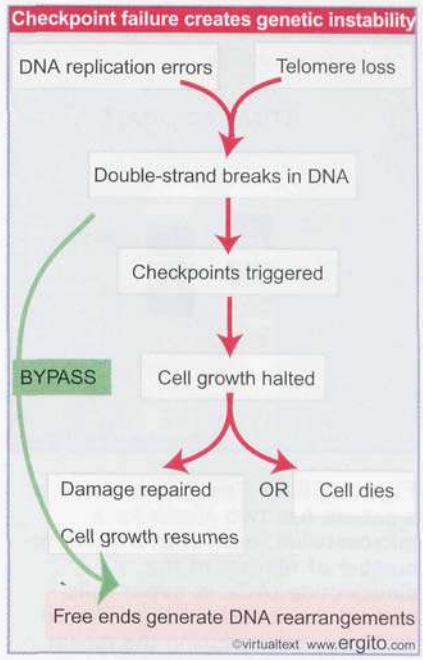


Figure 30.46 DNA replication errors or loss of telomeres may generate double-strand breaks in DNA. Checkpoints detect the breaks and protect the cell from perpetuating errors. If the checkpoints fail, DNA rearrangements occur.

30.29 Defects in repair systems cause mutations to accumulate in tumors

Key Concepts

- Loss of mismatch-repair systems generates a high mutation rate in HNPCC.

All cells have systems to protect themselves against damage from the environment or errors that may occur during replication (see *15 Recombination and repair).* The overall mutation rate is the result of the balance between the introduction of mutations and their removal by these systems. One means by which cancer cells increase the rate of mutation is to inactivate some of their repair systems, so that spontaneous mutations accumulate instead of being removed. In effect, a mutation that occurs in a mutator gene causes mutations to accumulate in other genes. (A mutator gene can be any type of gene—such as a DNA polymerase or a repair enzyme—whose function affects the integrity of DNA sequences.)

The MutSL system is a particularly important target. This system is responsible for removing mismatches in newly replicated bacterial DNA. Its homologues perform similar functions in eukaryotic cells. During replication of a microsatellite DNA, DNA polymerase may slip backward by one or more of the short repeating units. The additional unit(s) are extruded as a single-stranded region from the duplex. If not removed, they result in an increase in the length of the microsatellite in the next replication cycle (see Figure 4.27). This is averted when homologues of the MutSL system recognize the single-stranded extrusion and replace

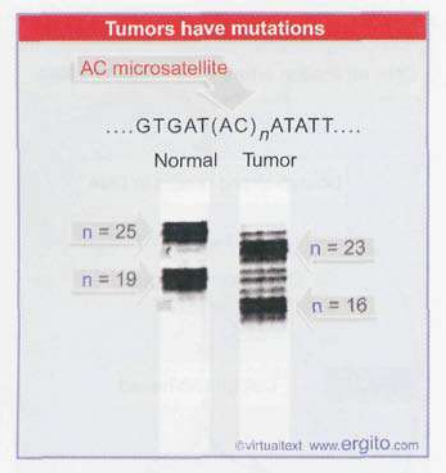


Figure 30.47 The normal tissue of a patient has two alleles for a microsatellite, each with a different number of repeats of the dinucleotide (AC). In tumor cells, both of these alleles have suffered deletions, one reducing the repeat number from 25 to 23, the other reducing it from 1 9 to 16. The repeat number of each allele in each situation is in fact probably unique, but some additional bands are generated as an artefact during the amplification procedure used to generate the samples. Data kindly provided by Bert Vogelstein. The bands remaining at the normal position in the tumor samples are due to contamination of the tumor sample with normal tissue.

the newly-synthesized material with a nucleotide sequence that properly matches the template (see Figure 15.47).

In the human disease of HNPCC (hereditary nonpolyposis colorectal cancer), new microsatellite sequences are found at a high frequency in tumor cells when their DNA sequences are compared with somatic cells of the same patient. **Figure 30.47** shows an example. This microsatellite has a repeat sequence of AC (reading just one strand of DNA). The length of the repeat varies from 14-27 copies in the population. Any particular individual shows two repeat lengths, one corresponding to each allele in the diploid cell. In many patients, the repeat length is changed. Most often it is reduced at both alleles, as shown in the example in the figure.

The idea that this type of change might be the result of loss of the mismatch-repair system was confirmed by showing that *mutS* and *mutL* homologues (*hMSH2*, *hMLH1*) are mutated in the tumors. As expected, the tumor cells are deficient in mismatch-repair. Change in the microsatellite sequences is of course only one of the types of mutation that result from the loss of the mismatch-repair system (it is especially easy to diagnose).

The case of HNPCC illustrates both the role of multiple mutations in malignancy and the contribution that is made by mutator genes. At least 7 independent genetic events are required to form a fully tumorigenic colorectal cancer. More than 90% of cases have mutations in the mismatch-repair system, and the tumor cells have mutation rates that are elevated by 2-3 orders of magnitude from normal somatic cells. The high mutation rate is responsible for creating new variants in the tumor that provide the raw material from which cells with more aggressive growth properties will arise.

Several human diseases are caused by mutations in the systems that execute checkpoints, including Ataxia telangiectasia (see *29.13 DNA damage triggers a checkpoint)*, Nijmegan breakage syndrome, and Bloom's syndrome, all of which are characterized by chromosomal rearrangements that are triggered by breaks in DNA.

30.30 Summary

A tumor cell is distinguished from a normal cell by its immortality, morphological transformation, and (sometimes) ability to metastasize. Oncogenes are identified by genetic changes that represent gain-of-functions associated with the acquisition of these properties. An oncogene may be derived from a proto-oncogene by mutations that affect its function or level of expression. Tumor suppressors are identified by loss-of-function mutations that allow increased cell proliferation. The mutations may either eliminate function of the tumor repressor or create a dominant negative version.

DNA tumor viruses carry oncogenes without cellular counterparts. Their oncogenes may work by inhibiting the activities of cellular tumor suppressors. RNA tumor viruses carry *v-onc* genes that are derived from the mRNA transcripts of cellular (*c-onc*) genes. Some *v-onc* oncogenes represent the full length of the *c-onc* proto-oncogene, but others are truncated at one or both ends. Most are expressed as fusion proteins with a retroviral product. Src is an exception in which the retrovirus (RSV) is replication-competent, and the protein is expressed as an independent entity.

Some *v-onc* genes are qualitatively different from their *c-onc* counterparts, since the *v-onc* gene is oncogenic at low levels of protein, while the *c-onc* gene is not active even at high levels. In such cases, proto-oncogenes are activated efficiently only by changes in the protein coding sequence. Other proto-oncogenes

can be activated by large (>10x) increases in the level of expression; *c-myc* is an example that can be activated quantitatively by a variety of means, including translocations with the Ig or TCR loci or insertion of retroviruses.

c-onc genes may have counterpart *v-onc* genes in retroviruses, but some proto-oncogenes have been identified only by their association with cellular tumors. The transfection assay detects some activated *c-onc* sequences by their ability to transform rodent fibroblasts. *ras* genes are the predominant type identified by this assay. The creation of transgenic mice directly demonstrates the transforming potential of certain oncogenes.

Cellular oncoproteins may be derived from several types of genes. The common feature is that each type of gene product is likely to be involved in pathways that regulate growth, and the oncoprotein has lack of regulation or increased activity.

Growth factor receptors located in the plasma membrane are represented by truncated versions in *v-onc* genes. The protein tyrosine kinase activity of the cellular receptor is activated only when ligand binds, but the oncogenic versions have constitutive activity or altered regulation. In the same way, mutation of genes for polypeptide growth factors gives rise to oncogenes, because a receptor becomes inappropriately activated.

Some oncoproteins are cytoplasmic tyrosine kinases; their targets are largely unknown. They may be activated in response to the autophosphorylation of tyrosine kinase receptors. The molecular basis for the difference between c-Src and v-Src lies in the phosphorylation states of two tyrosines. Phosphorylation of Tyr-527 in the C-terminal tail of c-Src suppresses phosphorylation of Tyr-416. The phosphorylated Tyr-527 binds to the SH2 domain of Src. However, when the SH2 domain recognizes the phosphopeptide sequence created by autophosphorylation of PDGF receptor; the PDGF receptor displaces the C-terminal region of Src, thus allowing dephosphorylation of Tyr-527, with the consequent phosphorylation of Tyr-416 and activation of the kinase activity. v-Src has lost the repressive C-terminus that includes Tyr-527, and therefore has permanently phosphorylated Tyr-416, and is constitutively active.

Ras proteins can bind GTP and are related to the α subunits of G proteins involved in signal transduction across the cell membrane. Oncogenic variants have reduced GTPase activity, and therefore are constitutively active. Activation of Ras is an obligatory step in a signal transduction cascade that is initiated by activation of a tyrosine kinase receptor such as the EGF receptor; the cascade passes to the ERK MAP kinase, which is a serine/threonine kinase, and terminates with the nuclear phosphorylation of transcription factors including Fos.

Nuclear oncoproteins may be involved directly in regulating gene expression, and include Jun and Fos, which are part of the AP1 transcription factor. v-ErbA is derived from another transcription factor, the thyroid hormone receptor, and is a dominant negative mutant that prevents the cellular factor from functioning. v-Rel is related to the common factor NF-κB, and influences the set of genes that are activated by transcription factors in this family.

Retinoblastoma (RB) arises when both copies of the *RB* gene are deleted or inactivated. The *RB* product is a nuclear phosphoprotein whose state of phosphorylation controls entry into S phase. Nonphosphorylated RB sequesters the transcription factor E2F. The RB-E2F complex represses certain target genes. E2F is released when RB is phosphorylated by cyclin/cdk complexes; E2F can then activate genes whose products are needed for S phase. Loss of RB prevents repression by RB-E2F, and means that E2F is constitutively available. The cell cannot be restrained from proceeding through the cycle. Adenovirus E1A and papova virus T antigens bind to nonphosphorylated RB, and thus prevent it from binding to E2F.

By Book_Crazy [IND]

p53 was originally classified as an oncogene because missense mutations in it are oncogenic. It is now classified as a tumor suppressor because the missense mutants in fact function by inhibiting the activity of wild-type p53. The same phenotype is produced by loss of both wild-type alleles. The level of p53 is usually low, but in response to damage to DNA, p53 activity increases, and triggers either of two pathways, depending upon the stage of the cell cycle and the ce\\ phenotype. Early in the cycle, it provides a checkpoint that prevents further progress; this allows damaged DNA to be repaired before replication. Later in the cycle, it causes apoptosis, so that the cell with damaged DNA dies instead of perpetuating itself. Loss of p53 function is common in established cell lines and may be important in immortalization *in vitro.* Absence of p53 is common in human tumors and may contribute to the progression of a wide variety of tumors, without specificity for cell type.

p53 is activated by binding to damaged DNA, for which it uses a (non-sequence-specific) DNA-binding domain. One important target that activates p53 is the single-stranded overhanging end that is generated at a shortened telomere. When it is activated, p53 uses another DNA-binding domain to recognize a palindromic ~10 bp sequence. Genes whose promoters have this sequence and which are activated by p53 include the cdk inhibitor p21 and the protein GADD45 (which is activated by several pathways for response to DNA damage). Activation of these and other genes (involving a transactivation domain that interacts directly with TBP) is probably the means by which p53 causes cell cycle arrest. p53 has a less well characterized ability to repress some genes. Mutant p53 lacks these activities, and therefore allows the perpetuation of cells with damaged DNA. Loss of p53 may be associated with increased amplification of DNA sequences.

p53 is bound by viral oncogenes such as SV40 T antigen, whose oncogenic properties result, at least in part, from the ability to block p53 function. It is also bound by the cellular proto-oncogene, Mdm2, which inhibits its activity. p53 and Mdm2 are mutual antagonists.

The locus INK4A contains two tumor suppressors that together control both major tumor suppressor pathways. $p19^{ARF}$ inhibits Mdm2, so that p19 in effect turns on p53. $p16^{INK4A}$ inhibits the cdk4/6 kinase, which phosphorylates RB. Deletion of INK4A therefore blocks both tumor suppressor pathways by leading to activation of Mdm2 (inhibiting p53) and activation of cdk4/6 (inhibiting RB).

Loss of p53 may be necessary for immortalization, because both the G1 checkpoint and the trigger for apoptosis are inactivated. Telomerase is usually turned off in differentiating cells, which provides a mechanism of tumor suppression by preventing indefinite growth. Reactivation of telomerase is usually necessary to allow continued proliferation of tumor cells. The crisis that is encountered by cells placed in culture results from shortening of telomeres to the point at which genetic instability is created by the chromosome ends. Loss of p53 is important in passing through crisis, because otherwise p53 is activated by the ends generated by telomere loss.

Several independent events are required to convert a normal cell into a cancer cell, typically involving both immortalizing and transforming functions. The required number of events is in the range of 4-10, and would not normally be expected to occur during the life span of a cell. Early events may increase the rate of occurrence of mutational change by damaging the repair or other systems that limit mutational damage. One important target is the MutSL system that is responsible for removing mismatches in replicated DNA.

References

30.1 Introduction

rev Balmain, A. (2002). Cancer as a complex genetic trait: tumor susceptibility in humans and mouse models. Cell 108, 145-152.

DePinho, R. A. (2000). The age of cancer. Nature 408, 248-254.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. Cell 100, 57-70.

Kinzler, K. W. and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell 87, 159-170.

Loeb, L. A. (2001). A mutator phenotype in cancer. Cancer Res. 61, 3230-3239.

Loeb, L. A., Springgate, C. F., and Battula, N. (1974). Errors in DNA replication as a basis of malignant changes. Cancer Res. 34, 231 1-2321.

ref Nowell, P. C. (1976). The clonal evolution of tumor cell populations. Science 194, 23-28.

30.2 Tumor cells are immortalized and transformed

ref Hayflick, L., and Moorhead, P. S. (1961). The serial cultivation of human diploid cell strains. Exp. Cell Res. 25, 585-621.

30.7 Retroviral oncogenes have cellular counterparts

exp Martin, G. S. (2002). Identification of a retroviral transforming gene
(www.ergito.com/lookup.jsp?expt = martin)

rev Bishop, J. M. (1983). Cellular oncogenes and retroviruses. Ann. Rev. Biochem. 52, 301-354.
Varmus, H. (1984). The molecular genetics of cellular oncogenes. Ann. Rev. Genet. 18, 553-612.

ref Brugge, J. and Erikson, R. L. (1977). Identification of a transformation-specific antigen induced by an avian sarcoma virus. Nature 269, 346-348.

30.8 Quantitative or qualitative changes can explain oncogenicity

rev Bishop, J. M. (1985). Viral oncogenes. Cell 42, 23-38.

30.9 Ras oncogenes can be detected in a transfection assay

exp Weinberg, R. (2002). The Discovery of Oncogenes in Human Tumors
(www.ergito.com/lookup.jsp?expt = Weinberg)

30.10 Ras proto-oncogenes can be activated by mutation at specific positions

rev Barbacid, M. (1987). Ras genes. Ann. Rev. Biochem. 56, 779-827.

Lowy, D. R. (1993). Function and regulation of Ras. Ann. Rev. Biochem. 62, 851-891.

30.12 Proto-oncogenes can be activated by translocation

rev Adams, J. M. and Cory, S. (1991). Transgenic models of tumor development. Science 254, 1161-1167.

Cory, S. and Adams, J. M. (1988). Transgenic mice and oncogenesis. Ann. Rev. Immunol. 6, 25-48.
Showe, L. C. and Croce, C. M. (1987). The role of chromosomal translocations in B- and T-cell neoplasia. Ann. Rev. Immunol. 5, 253-277.

30.14 Oncogenes code for components of signal transduction cascades

rev Cantley, L. C, Auger, K. R., Carpenter, C, Duckworth, B., Graziani, A., Kapeller, R., and Soltoff, S. (1991). Oncogenes and signal transduction. Cell 64, 281-302.

Cross, M. and Dexter, T. M. (1991). Growth factors in development, transformation, and tumorigenesis. Cell 64, 271-280.

Heldin, C.-H. and Westermark, B. (1984). Growth factors: mechanism of action and relation to oncogenes. Cell 37, 9-20.
Jove, R. and Hanafusa, H. (1987). Cell transformation by the viral src oncogene. Ann. Rev. Cell Biol. 3, 31-56.

ref Angel, P., Allegretto, E. A., Okino, S. T., Hattori, K., Boyle, W. J., Hunter, T., and Karin, M. (1988). Oncogene jun encodes a sequence-specific trans-activator similar to AP1. Nature 332, 166-171.
Bohmann, D., Bos, T. J., Admon, A., Nishimura, T., Vogt, P. K., and Tjian, R. (1987). Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP1. Science 238, 1386-1392.
Bos, T. J., Bohmann, D., Tsuchie, H., Tjian, R., and Vogt, P. K. (1988). v-jun encodes a nuclear protein with enhancer binding properties of AP1. Cell 52, 705-712.

Collet, M. S. and Erikson, R. L. (1978). Protein kinase activity associated with the avian sarcoma virus src gene product. Proc. Nat. Acad. Sci. USA 75, 2021-2024.

Hunter, T. and Sefton, B. (1980). Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. Proc. Nat. Acad. Sci. USA 77, 1311-1315.
Waterfield, M. D. et al. (1983). Platelet derived growth factors is structurally related to the putative transforming protein $p28^{sis}$ of simian sarcoma virus. Nature 304, 35-39.

30.15 Growth factor receptor kinases can be mutated to oncogenes

rev Deuel, T. F. (1987). Polypeptide growth factors: roles in normal and abnormal cell growth. Ann. Rev. Cell Biol. 3, 443-492.

Rettenmier, C. W., Roussel, M. F., and Sherr, C. J. (1988). The colony-stimulating factor 1 (CSF-1) receptor (c-fms proto-oncogene product) and its ligand. J. Cell Sci. Suppl. 9, 27-44.
Schlessinger, J. (2002). Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. Cell 110, 669-672.

Sherr, C. J., Rettenmier, C. W., and Roussel, M. F. (1988). Macrophage colony-stimulating factor, CSF-1, and its proto-oncogene-encoded receptor. Cold Spring Harb Symp Quant Biol 53 Pt 1, 521-530.

ref Bargmann, C. I., Hung, M. C, and Weinberg, R. A. (1986). Multiple independent activations of the neu oncogene by a point mutation altering the transmembrane domain of p185. Cell 45, 649-657.
Downward, J., Yarden, Y., Mayes, E., Scrace, G., Totty, N., Stockwell, P., Ullrich, A., Schlessinger, J., and Waterfield, M. D. (0). Close similarity of epidermal growth factor receptor and v-erb-B oncogene protein sequences. Nature 307, 521-527.
Sherr, C. J., Rettenmier, C. W., Sacca, R., Roussel, M. F., Look, A. T., and Stanley, E. R. (1985). The c-fms proto-oncogene product is related to the receptor for the mononuclear phagocyte growth factor, CSF-1. Cell 41, 665-676.

30.16 Src is the prototype for the proto-oncogenic cytoplasmic tyrosine kinases

rev Martin, G. S. (2001). The hunting of the Src. Nat. Rev. Mol. Cell Biol. 2, 467-475.

ref Hirai, H. and Varmus, H. E. (1990). Mutations in src homology regions 2 and 3 of activated chicken c-src that result in preferential transformation of mouse or chicken cells. Proc. Nat. Acad. Sci. USA 87, 8592-8596.

Parker, R. C, Varmus, H. E., and Bishop, J. M. (1984). Expression of v-src and chicken c-src in rat cells demonstrates qualitative differences between pp60v-src and pp60c-src. Cell 37, 131-139.

Swanstrom, R., Parker, R. C, Varmus, H. E., and Bishop, J. M. (1983). Transduction of a cellular oncogene: the genesis of Rous sarcoma virus. Proc. Nat. Acad. Sci. USA 80, 2519-2523.

Takeya, T. and Hanafusa, H. (1983). Structure and sequence of the cellular gene homologous to the RSV src gene and the mechanism for generating the transforming virus. Cell 32, 881-890.

30.17 Src activity is controlled by phosphorylation

rev Thomas, S. M. and Brugge, J. S. (1997). Cellular functions regulated by Src family kinases. Ann. Rev. Cell Dev. Biol. 13, 513-609.

ref Xu, W., Doshi, A., Lei, M., Eck, M. J., and Harrison, S. C. (1999). Crystal structures of c-Src reveal features of its autoinhibitory mechanism. Mol. Cell 3, 629-638.

30.18 Oncoproteins may regulate gene expression

rev Nevins, J. R. (1987). Regulation of early adenovirus gene expression. Microbiol. Rev. 51, 419-430.

Verma, I. M., Stevenson, J. K., Schwarz, E. M., Van Antwerp, D., and Miyamoto, S. (1995). Rel/NF-kappa B/I kappa B family: intimate tales of association and dissociation. Genes Dev. 9, 2723-2735.

ref Alani, R., Brown, P., Binetruy, B., Dosaka, H., Rosenberg, R. K., Angel, P., Karin, M., and Birrer, M. J. (1991). The transactivating domain of the c-Jun proto-oncoprotein is required for cotransformation of rat embryo cells. Mol. Cell. Biol. 11, 6286-6295.

Ballard, D. W., Walker, W. H., Doerre, S., Sista, P., Molitor, J. A., Dixon, E. P., Peffer, N. J., Hannink, M., and Greene, W. C. (1990). The v-rel oncogene encodes a kappa B enhancer binding protein that inhibits NF-kappa B function. Cell 63, 803-814.

Bohmann, D. and Tjian, R. (1989). Biochemical analysis of transcriptional activation by Jun: differential activity of c- and v-Jun. Cell 59, 709-717.

Bohmann, D., Bos, T. J., Admon, A., Nishimura, T., Vogt, P. K., and Tjian, R. (1987). Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP1. Science 238, 1386-1392.

Bos, T. J., Bohmann, D., Tsuchie, H., Tjian, R., and Vogt, P. K. (1988). v-jun encodes a nuclear protein with enhancer binding properties of AP1. Cell 52, 705-712.

Clark, E. A., Golub, T. R., Lander, E. S., and Hynes, R. O. (2000). Genomic analysis of metastasis reveals an essential role for RhoC. Nature 406, 532-535.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. Nature 412, 822-826.

Ghosh, S., Gifford, A. M., Riviere, L. R., Tempst, P., Nolan, G. P., and Baltimore, D. (1990). Cloning of the p50 DNA binding subunit of NF-kappa B: homology to rel and dorsal. Cell 62, 1019-1029.

Kieran, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., Le Bail, O., Urban, M. B., Kourilsky, P., Baeuerle, P. A., and Israel, A. (1990). The DNA binding subunit of NF-kappa B is identical to factor KBF1 and homologous to the rel oncogene product. Cell 62, 1007-1018.

Sylla, B. S. and Temin, H. M. (1986). Activation of oncogenicity of the c-rel proto-oncogene. Mol. Cell. Biol. 6, 4709-4716.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C, Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419, 624-629.

Zenke, M., Munoz, A., Sap, J., Vennstrom, B., and Beug, H. (1990). v-erbA oncogene activation entails the loss of hormone-dependent regulator activity of c-erbA. Cell 61, 1035-1049.

30.19 RB is a tumor suppressor that controls the cell cycle

ref Cavanee, W. K. et al. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 305, 779-784.

Sarnow, P. et al. (1982). Adenovirus E1b-58kd tumor antigen and SV40 large tumor antigen are physically associated with the same 54 kD cellular protein in transformed cells same 54 kd cellular protein in transformed cells. Cell 28, 387-394.

Wang, T. C, Cardiff, R. D., Zukerberg, L., Lees, E., Arnold, A., and Schmidt, E. V. (1994). Mammary hyperplasia and carcinoma in MMTV-cyclin D1 transgenic mice. Nature 369, 669-671.

Yu, Q., Geng, Y., and Sicinski, P. (2001). Specific protection against breast cancers by cyclin D1 ablation. Nature 411, 1017-1021.

30.20 Tumor suppressor p53 suppresses growth or triggers apoptosis

exp Levine, A. (2002). p53 is a Tumor Suppressor Gene (www.ergito.com/lookup.jsp?expt= levine)

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. Cell 88, 323-331.

Levine, A. J., Momand, J., and Finlay, C. A. (1991). The p53 tumor suppressor gene. Nature 351, 453-456.

Marshall, C. J. (1991). Tumor suppressor genes. Cell 64, 313-326.

ref Finlay, C. A., Hinds, P. W., and Levine, A. J. (1989). The p53 proto-oncogene can act as a suppressor of transformation. Cell 57, 1083-1093.

Linzer, D. I. H. and Levine, A. J. (1979). Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. Cell 17, 43-52.

Malkin, D. et al. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250, 1233-1238.

30.21 p53 is a DNA-binding protein

ref Seto, E., Usheva, A., Zambetti, G. P., Momand, J., Horikoshi, N., Weinmann, R., Levine, A. J., and Shenk, T. (1992). Wild-type p53 binds to the TATA-binding protein and represses transcription. Proc. Nat. Acad. Sci. USA 89, 12028-1245.

Vaux, D. L., Cory, S., and Adams, J. M. (1988). Bcl2 gene promotes hematopoietic cell survival and cooperates with c-myc to immortalize pre-B cells. Nature 335, 440-442.

30.22 p53 is controlled by other tumor suppressors and oncogenes

rev Sherr, C. J. (2001). The INK4a/ARF network in tumour suppression. Nat. Rev. Mol. Cell Biol. 2, 731-737.

ref Momand, J., Zambetti, G. P., Olson, D. C, George, D., and Levine, A. J. (1992). The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. Cell 69, 1237-1245.

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. Nature 408, 307-310.

30.23 p53 is activated by modifications of amino acids

ref Zacchi, P., Gostissa, M., Uchida, T., Salvagno, C, Avolio, F., Volinia, S., Ronai, Z., Blandino, G., Schneider, C, and Del Sal, G. (2002). The prolyl isomerase Pin1 reveals a mechanism to control p53 functions after genotoxic insults. Nature 419, 853-857.

Zheng, H., You, H., Zhou, X. Z., Murray, S. A., Uchida, T., Wulf, G., Gu, L., Tang, X., Lu, K. P., and Xiao, Z. X. (2002). The prolyl isomerase Pin1 is a regulator of p53 in genotoxic response. Nature 419, 849-853.

30.24 Telomere shortening causes cell senescence

exp Shay (2002). Immortalizing human cells with telomerase (www.ergito.com/lookup.jsp?expt = shay)

ref Blasco, M. A. et al. (1997). Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. Cell 91, 25-34.

Greenberg, R. A., Chin, L., Femino, A., Lee, K. H., Gottlieb, G. J., Singer, R. H., Greider, C. W., and DePinho, R. A. (1999). Short dysfunctional telomeres impair tumorigenesis in the INK4a(delta2/3) cancer-prone mouse. Cell 97, 515-525.

Kim N. W., Piatyszek M. A., Prowse K. R., Harley C. B., West M. D., Ho P. L., Coviello G. M., Wright W. E., Weinrich S. L., Shay J. W. (1994). Specific association of human telomerase activity with immortal cells and cancer. Science 266, 2011-2015.

Meyerson, M. et al. (1997). hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. Cell 90, 785-795.

Wright, W. E., Brasiskyte, D., Piatyszek, M. A., and Shay, J. W. (1996). Experimental elongation of telomeres extends the lifespan of immortal x normal cell hybrids. EMBO J. 15, 1734-1741.

30.25 Immortalization depends on loss of p53

rev Maser, R. S. and DePinho, R. A. (2002). Connecting chromosomes, crisis, and cancer. Science 297, 565-569.

ref Donehower, L. A. et al. (1992). Mice deficient for p53 are developmentally normal but susceptible for spontaneous tumors. Nature 356, 215-221.

Karlseder, J., Broccoli, D., Dai, Y., Hardy, S., and de Lange, T. (1999). p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. Science 283, 1321-1325.

Li, G. Z., Eller, M. S., Firoozabadi, R., and Gilchrest, B. A. (2003). Evidence that exposure of the telomere 3' overhang sequence induces senescence. Proc. Nat. Acad. Sci. USA 100, 527-531.

30.26 Different oncogenes are associated with immortalization and transformation

exp Green, H. (2002). The Story of 3T3 Cells: A Voyage of Discovery Without an Itinerary (www.ergito.com/lookup.jsp?expt = green)

rev Hanahan, D. (1988). Dissecting multistep tumorigenesis in transgenic mice. Ann. Rev. Genet. 22, 479-519.

Hunter, T. (1991). Cooperation between oncogenes. Cell 64, 249-270.

ref Brinster, R. L. et al. (1984). Transgenic mice harboring SV40 T-antigen genes develop characteristic brain tumors. Cell 37, 367-379.

Sinn, E., Muller, W., Pattengale, P., Tepler, I., Wallace, R., and Leder, P. (1987). Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo. Cell 49, 465-4.

Stewart, T. A., Pattengale, P. K., and Leder, P. (1984). Spontaneous mammary adenocarcinomas in transgenic mice that carry and express MTV/myc fusion genes. Cell 38, 627-637.

30.27 p53 may affect aging

ref Chin, L., Artandi, S. E., Shen, Q., Tarn, A., Lee, S. L., Gottlieb, G. J., Greider, C. W., and DePinho, R. A. (1999). p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. Cell 97, 527-538.

Rudolph, K. L., Chang, S., Lee, H. W., Blasco, M., Gottlieb, G. J., Greider, C, and DePinho, R. A. (1999). Longevity, stress response, and cancer in aging telomerase-deficient mice. Cell 96, 701-712.

Tyner, S. D., Venkatachalam, S., Choi, J., Jones, S., Ghebranious, N., Igelmann, H., Lu, X., Soron, G., Cooper, B., Brayton, C, Hee Park, S., Thompson, T., Karsenty, G., Bradley, A., and Donehower, L. A. (2002). p53 mutant mice that display early ageing-associated phenotypes. Nature 415, 45-53.

30.28 Genetic instability is a key event in cancer

rev Kolodner, R. D., Putnam, C. D., and Myung, K. (2002). Maintenance of genome stability in *S. cerevisiae.* Science 297, 552-557.

Maser, R. S. and DePinho, R. A. (2002). Connecting chromosomes, crisis, and cancer. Science 297, 565-569.

Schar, P. (2001). Spontaneous DNA damage, genome instability, and cancer—when DNA replication escapes control. Cell 104, 329-332.

ref Myung, K. and Kolodner, R. D. (2002). Suppression of genome instability by redundant S-phase checkpoint pathways in S. *cerevisiae.* Proc. Nat. Acad. Sci. USA 99, 4500-4507.

30.29 Defects in repair systems cause mutations to accumulate in tumors

rev Kinzler, K. W. and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell 87, 159-170.

ref Aaltonen, L. A., Peltomaki, P., Leach, F. S., Sistonen, P., Pylkkanen, L., Mecklin, J. P., Jarvinen, H., Powell, S. M., Jen, J., Hamilton, S. R., et al. (1993). Clues to the pathogenesis of familial colorectal cancer. Science 260, 812-816.

Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homologue MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 75, 1027-1038.

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 363, 558-561.

Leach, F. S., Nicolaides, N. C, Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomaki, P., Sistonen, P., Aaltonen, L. A., Nystrom-Lahti, M., et al. (1993). Mutations of a mutS homologue in hereditary nonpolyposis colorectal cancer. Cell 75, 1215-1225.

Chapter 31

Gradients, cascades, and signaling pathways

31.1 Introduction

Development begins with a single fertilized egg, but gives rise to cells that have different developmental fates. The problem of early development is to understand how this asymmetry is introduced: how does a single initial cell give rise within a few cell divisions to progeny cells that have different properties from one another?

The means by which asymmetry is generated varies with the type of organism. The egg itself may be homogeneous, with the acquisition of asymmetry depending on the process of the initial division cycles, as in the case of mammals. Or the egg may have an initial asymmetry in the distribution of its cytoplasmic components, which in turn gives rise to further differences as development proceeds, as in the case of *Drosophila*.

Early development is defined by the formation of axes. By whatever means are used to develop asymmetry, the early embryo develops differences along the anterior-posterior axis (head-tail) and along the dorsal-ventral axis (top-bottom). At the stage of interpreting the axial information, a relatively restricted set of signaling pathways is employed, and essentially the same pathways are found in flies and mammals.

The paradigm for considering the molecular basis for development is to suppose that each cell type may be characterized by its pattern of gene expression, that is, by the particular gene products that it produces. The principal level for controlling gene expression is at transcription, and components of pathways regulating transcription provide an important class of developmental regulators. We may include a variety of activities within the rubric of transcriptional regulators, which could act to change the structure of a promoter region, to initiate transcription at a promoter, to regulate the activity of an enhancer, or indeed sometimes to repress the action of transcription factors. However, the regulators of transcription most often prove to be DNA-binding proteins that activate transcription at particular promoters or enhancers.

By Book_Crazy [IND]

31.2 Fly development uses a cascade of transcription factors

Key Concepts

- The genes that control the early stages of fly development code for transcription factors.
- At each stage, the factors in one area of the egg control the synthesis of further factors that will define smaller areas.
- Maternal genes are expressed during oogenesis and act in the oocyte.
- Three successive groups of segmentation genes are expressed after fertilization to control the number or polarity of segments.
- Homeotic genes control the identity of a segment.

The systematic manner in which the regulators are turned on and off to form circuits that determine body parts has been worked out in detail in *D. melanogaster*. The basic principle is that a series of events resulting from the initial asymmetry of the egg is translated into the control of gene expression so that specific regions of the egg acquire different properties. The means by which asymmetry is translated into control of gene expression differ for each of four systems that have been characterized in the insect egg. It may involve localization of factors that control transcription or translation within the egg, or localized control of the activities of such factors. But the end result is the same: spatial and temporal regulation of gene expression.

Early in development, the identities of parts of the embryo are determined: regions are defined whose descendants will form particular body parts. The genes that regulate this process are identified by loci in which mutations cause a body part to be absent, to be duplicated, or to develop as another body part. Such loci are prime candidates for genes whose function is to provide regulatory "switches." Most of these genes code for regulators of transcription. They act upon one another in a hierarchical manner, but they act also upon other genes whose products are actually responsible for the formation of pattern. The ultimate targets are genes that code for kinases, cytoskeletal elements, secreted proteins, and transmembrane receptors.

The establishment of a specific pattern of transcription in a particular region of the embryo leads to a *cascade* of control, when regulatory events are connected so that a gene turned on (or off) at one stage itself controls expression of other genes at the next stage. Formally, such a cascade resembles those described previously for bacteriophages or for bacterial sporulation (as discussed in *10 The operon)*, although it is more complex in the case of eukaryotic development. The common feature of regulatory proteins is that they are transcription factors that regulate the expression of other transcription factors (as well as other target proteins). As in the case of prokaryotic regulation, the basic relationship between the regulator protein and the target gene is that the regulator recognizes a short sequence in the DNA of the promoter (or an enhancer) of a target gene. All of the targets for a particular regulator are identified by their possession of a copy of the appropriate consensus sequence.

The development of an adult organism from a fertilized egg follows a predetermined pathway, in which specific genes are turned on and off at particular times. From the perspective of mechanism, we have most information about the control of transcription. However, subsequent stages of gene expression are also targets for regulation. And, of course, the cascade of gene regulation is connected to other types of signaling, including cell-cell interactions that define boundaries between groups of cells.

The mechanics of development in terms of cellular events are different in different types of species, but we assume that the principle established with *Drosophila* will hold in all cases: that a regulatory cascade determines the appropriate pattern of gene expression in cells of the embryo and ultimately of the adult. Indeed, homologous genes in distantly related organisms play related roles in development. The same pathways are found in (for example) flies and mammals, although the consequences of their employment are rather different in terms of the structures that develop.

Genes involved in regulating development are identified by mutations that are lethal early in development or that cause the development of abnormal structures. A mutation that affects the development of a particular body part attracts our attention because a single body part is a complex structure, requiring expression of a particular set of many genes. Single mutations that influence the structure of the entire body part therefore identify potential regulator genes that switch or select between developmental pathways.

In *Drosophila*, the body part that is analyzed is the segment, the basic unit that can be seen looking at the adult fly. Mutations fall into (at least) three groups, defined by their effect on the segmental structure:

- Maternal genes are expressed during oogenesis by the mother. They may act upon or within the maturing oocyte.
- Segmentation genes are expressed after fertilization. Mutations in these genes alter the number or polarity of segments. Three groups of segmentation genes act sequentially to define increasingly smaller regions of the embryo.
- Homeotic genes control the identity of a segment, but do not affect the number, polarity, or size of segments. Mutations in these genes cause one body part to develop the phenotype of another part.

The genes in each group act successively to define the properties of increasingly more restricted parts of the embryo. The maternal genes define broad regions in the egg; differences in the distribution of maternal gene products control the expression of segmentation genes; and the homeotic genes determine the identities of individual segments.

31.3 A gradient must be converted into discrete compartments

Key Concepts

- During the first 13 division cycles, nuclei divide in a common cytoplasm; cells form only at blastoderm.
- Gradients define the polarity of the egg along both the anterior-posterior and dorsal-ventral axes.
- The gradients consist of RNAs or proteins that are differentially distributed in the common cytoplasm.
- The location of a nucleus in the cytoplasm with regard to the two axes determines the fate of the cells that descend from it.

The basic question of *Drosophila* development is illustrated in Figure 31.1 in terms of three stages of development: the egg; the larva; and the adult fly.

At the start of development, gradients are established in the egg along the anterior-posterior and dorsal-ventral axes. The anterior end of the egg becomes the head of the adult; the posterior end becomes the tail. The dorsal side is on top (looking down on a larva); the ventral side

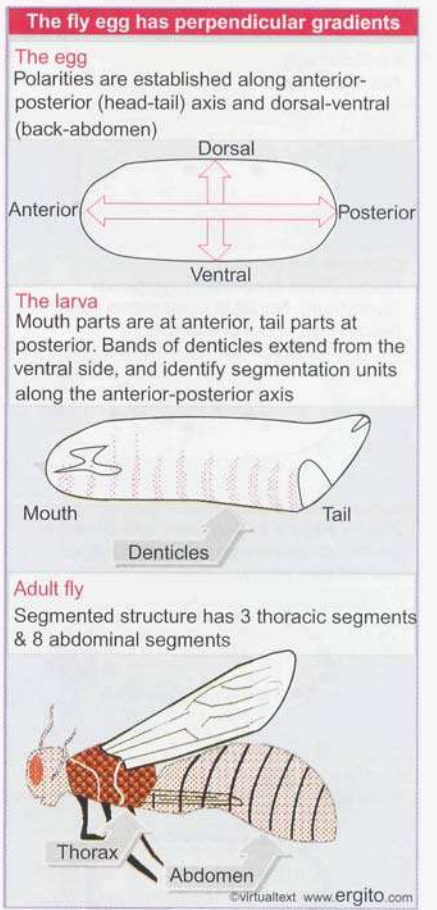


Figure 31.1 Gradients in the egg are translated into segments on the anterior-posterior axis and into specialized structures on the dorsal-ventral axis of the larva, and then into the segmented structure of the adult fly.

By Book_Crazy [IND]

is underneath. The gradients consist of molecules (proteins or RNAs) that are differentially distributed in the cytoplasm. The gradient responsible for anterior-posterior development is established soon after fertilization; the dorsal-ventral gradient is established a little later. It is only a modest oversimplification to say that the anterior-posterior systems control positional information along the larva, while the dorsal-ventral system regulates tissue differentiation (that is, the specification of distinct embryonic tissues, including mesoderm, neuroectoderm, and dorsal ectoderm).

Insect development involves two quite different types of structures. The first part of development is concerned with elaborating the larva; then the larva metamorphoses into the fly. This means that the structure of the embryo (the larva) is distinct from the structure of the adult (the fly), in contrast with development of (for example) mammals, where the embryo develops the same body parts that are found in the adult. As the larva develops, it forms some body parts that are exclusively larval (they will not give rise to adult tissues; often they are polyploid), while other body parts are the progenitors that will metamorphose into adult structures (usually they are diploid). In spite of the differences between insect development and vertebrate development, the same general principles appear to govern both processes, and we discover relationships between *Drosophila* regulators and mammalian regulators.

Discrete regions in the embryo correspond to parts of the adult body. They are shown in terms of the superficial organization of the larva in the middle panel of Figure **31.1.** Bands of denticles (small hairs) are found in a particular pattern on the surface (cuticle) of the larva. The cuticular pattern has features determined by both the anterior-posterior axis and the dorsal-ventral axis:

- Along the anterior-posterior axis, the denticles form discrete bands. Each band corresponds to a segment of the adult fly: in fact, the 11 bands of denticles correspond on a 1 : 1 basis with the 11 segments of the adult.
- Along the dorsal-ventral axis, the denticles that extend from the ventral surface are coarse; those that extend from the dorsal surface are much finer.

Although the cuticle represents only the surface body layer, its structure is diagnostic of the overall organization of the embryo in both axes. Much of the analysis of phenotypes of mutants in *Drosophila* development has therefore been performed in terms of the distortion of the denticle patterns along one axis or the other.

The difference in form between the gradients of the egg and the segments of the adult poses some prime questions. How are gradients established in the egg? And how is a continuous gradient converted into discrete differences that define individual cell types? How can a large number of separate compartments develop from a single gradient?

The nature of the gradients, and their ability to affect the development of a variety of cell types located throughout the embryo, depend upon some idiosyncratic features of *Drosophila* development. The early stages are summarized in **Figure 31.2.**

At fertilization the egg possesses the two parental nuclei and is distinguished at the posterior end by the presence of a region called polar plasm. For the first 9 divisions, the nuclei divide in the common cytoplasm. Material can diffuse in this cytoplasm (although there are probably constraints imposed by cytoskeletal organization). At division 7, some nuclei migrate into the polar plasm, where they become precursors to germ cells. After division 9, nuclei migrate and divide to form a layer at the surface of the egg. Then they divide 4 times, after which membranes surround them to form somatic cells.

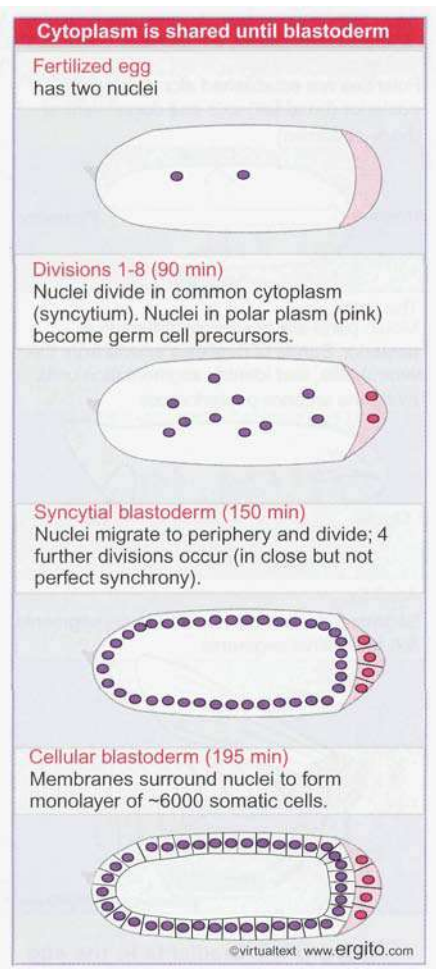


Figure 31.2 The early development of the *Drosophila* egg occurs in a common cytoplasm until the stage of cellular blastoderm.

Up to the point of cellularization, the nuclei effectively reside in a common cytoplasm. At the stage of the cellular blastoderm, the first discrete compartments become evident, and at this time particular regions of the egg are *determined* to become particular types of adult structures. (Determination is progressive and gradual; over the next few cell divisions, the fates of individual regions of the egg become increasingly restricted.) At the start of this process, nuclei migrate to the surface to form the monolayer of the blastoderm, but they do not do so in any predefined manner. It is therefore the location in which the nuclei find themselves at this stage that determines what types of cells their descendants will become. *A nucleus determines its position in the embryo by reference to the anterior-posterior and dorsal-ventral gradients, and behaves accordingly.*

31.4 Maternal gene products establish gradients in early embryogenesis

Key Concepts

- Four signaling pathways are initiated outside the egg and each leads to production of a morphogen in the egg.
- The anterior system is responsible for development of head and thorax.
- The posterior system is responsible for the segments of the abdomen.
- The terminal system is responsible for producing the acron (in the head) and the telson (at the tail).
- The dorsal-ventral system determines development of tissue types (mesoderm, neuroectoderm, and ectoderm).

An initial asymmetry is imposed on the *Drosophila* oocyte during oogenesis. **Figure 31.3** illustrates the structure of a follicle in the *Drosophila* ovary. A single progenitor undergoes four successive mitoses to generate 16 interconnected cells. The connections are known as "cytoplasmic bridges" or "ring canals." Individual cells have 2, 3, or 4 such connections. One of the two cells that has 4 connections undergoes meiosis to become the oocyte; the other 15 cells become "nurse cells." Cytoplasmic material, including protein and RNA, passes from the nurse cells to the oocyte; the accumulation of such material accounts for a considerable part of the volume of the egg. The cytoplasmic connections are made at one end of the oocyte, and this end becomes the anterior end of the egg.

Genes that are expressed within the mother fly are important for early development. These maternal genes are identified by female sterile mutations. They do not affect the mother itself, but are required in order to have progeny. Females with such mutations lay eggs that fail to develop into adults; the embryos can be recognized by defects in the cuticular pattern, and they die during development.

The common feature in all maternal genes is that they are expressed prior to fertilization (although their products may act either at the time of expression or be stored for later use). The maternal genes are divided into two classes, depending on their site of expression. Genes that are expressed in somatic cells of the mother that affect egg development are called *maternal somatic genes.* For example, they may act in the follicle cells. Genes that are expressed within the germline are called *maternal germline genes.* These genes may act either in the nurse cell or the oocyte. Some genes act at both stages.

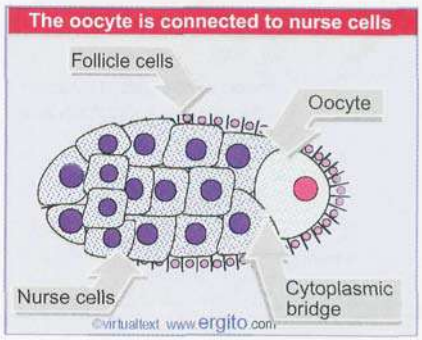


Figure 31.3 A *Drosophila* follicle contains an outer surface of follicle cells that surround nurse cells that are in close contact with the oocyte. Nurse cells are connected by cytoplasmic bridges to each other and to the anterior end of the oocyte. Follicle cells are somatic; nurse cells and the oocyte are germline in origin.

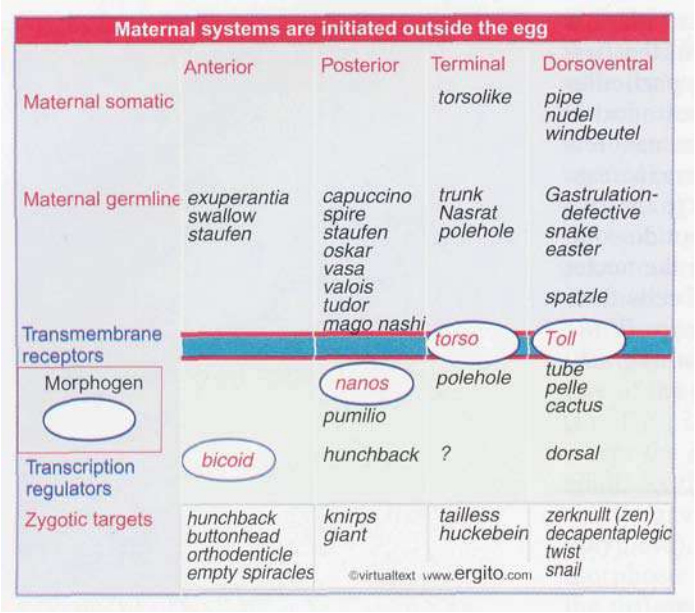


Figure 31.4 Each of the four maternal systems that functions in the egg is initiated outside the egg. The pathway is carried into the egg, where each pathway has a localized product that is the morphogen. This may be a receptor or a regulator of gene expression. The final component is a transcription factor, which acts on zygotic targets that are responsible for the next stage of development.

Four groups of genes concerned with the development of particular regions of the embryo can be identified by mutations in maternal genes. The genes in each group can be organized into a pathway that reflects their order of action, by conventional genetic tests (such as comparing the properties of double mutants with the individual mutants) or by biochemical assays.

The components of these pathways are summarized in **Figure 31.4,** which shows that there is a common principle to their operation. *Each pathway is initiated by localized events outside the egg; this results in the localization of a signal within the egg.* This signal takes the form of a protein with an asymmetric distribution; this is called a morphogen. Formally, we may define a morphogen as a protein whose local concentration (or activity) causes the surrounding region to take up a particular structure or fate. In each of these systems, the morphogen either is a transcriptional regulator or leads to the activation of a transcription factor in the localized region. Three systems are concerned with the anterior-posterior axis, and one with the dorsal-ventral axis:

- The anterior system is responsible for development of the head and thorax. The maternal germline products are required to localize the *bicoid* product at the anterior end of the egg. In fact, *bicoid* mRNA is transcribed in nurse cells and transported into the oocyte. Bicoid protein is the morphogen: it functions as a transcriptional regulator, and controls expression of the gene *hunchback* (and probably also other segmentation and homeotic genes).
- The posterior system is responsible for the segments of the abdomen. The nature of the initial asymmetric event is not clear. A large number of products act to cause the localization of the product of *nanos*, which is the morphogen. This leads to localized repression of expression of *hunchback* (via control of translation of the mRNA).
- The terminal system is responsible for development of the specialized structures at the unsegmented ends of the egg (the acron at the head, and the telson at the tail). As indicated by the dependence on maternal somatic genes, the initial events that create asymmetry occur in the follicle cells. They lead to localized activation of the transmembrane receptor coded by *torso;* the end product of the pathway has yet to be identified.
- The fourth system is responsible for dorsal-ventral development. The pathway is initiated by a signal from a follicle cell on the ventral side of the egg. It is transmitted through the transmembrane receptor coded by *Toll.* This leads to a gradient of activation of the transcription factor produced by *dorsal* (by controlling its localization within the cell).

About 30 maternal genes involved in pattern formation have been identified. All of the components of the four pathways are maternal, so we see that the systems for establishing the initial pattern formation all depend on events that occur prior to fertilization. The two body axes are established independently. Mutations that affect polarity cause posterior regions to develop as anterior structures, or ventral regions to develop in dorsal form. On the anterior-posterior axis, the anterior and posterior systems provide opposing gradients, with sources at the anterior and posterior ends of the embryo, respectively, that control development of the segments of the body. Defects in either system affect the body segments. The terminal and dorsal-ventral systems operate independently of the other systems.

31.5 Anterior development uses localized gene regulators

Key Concepts

- The anterior system localizes bicoid mRNA at the anterior end of the egg, generating a gradient of protein that extends along the anterior 40% of the egg.
- The concentration of bicoid protein determines the types of anterior (head) structures that are produced in each region.
- This system is instructive because it is required for development of the head structures.

Establishing asymmetry in an egg requires that some components—either RNAs or proteins—are *localized* instead of being diffused evenly through the cytosol. In anterior-posterior development in *Drosophila*, certain mRNAs are localized at the anterior or posterior end. **Figure 31.5** shows that when they are translated, their protein products diffuse away from the ends of the egg, generating a gradient along the anterior-posterior axis.

The existence of localized concentrations of materials needed for development can be tested by the rescue protocol summarized in **Figure 31.6.** Material is removed from a wild-type embryo and injected into the embryo of a mutant that is defective in early development. If the mutant embryo develops normally, we may conclude that the mutation causes a deficiency of material that is present in the wild-type embryo. This allows us to distinguish components that are necessary for morphogenesis, or that are upstream in the pathway, from the morphogen itself—only the morphogen has the property of localized rescue.

The rescue technique identifies bicoid as the morphogen required for anterior development. *bicoid* mutants do not develop heads; but the defect can be remedied by injecting mutant eggs with cytoplasm taken from the anterior tip of a wild-type embryo. Indeed, anterior structures develop elsewhere in the mutant embryo if wild-type anterior cytoplasm is injected! (This is called ectopic expression.) The extent of the rescue depends on the amount of wild-type cytoplasm injected. And the efficacy of the donor cytoplasm depends on the number of wild-type *bicoid* genes carried by the donor.

These results suggest that the anterior region of a wild-type embryo contains a concentration of some product that depends on the *bicoid* gene dosage. By purifying the active component in the preparation, it is possible to show that purified *bicoid* mRNA can substitute for the anterior cytoplasm. *This implies that the components on which* bicoid *acts are ubiquitous, and all that is required to trigger formation of anterior structures is an appropriate concentration of* bicoid *product.*

The product of *bicoid* establishes a gradient with its source (and therefore the highest concentration) at the anterior end of the embryo. The RNA is localized at the anterior tip of the embryo, but it is not translated during oogenesis. Translation begins soon after fertilization. The protein then establishes a gradient along the embryo, as indicated in **Figure 31.7.** The gradient could be produced by diffusion of the protein product from the localized source at the anterior tip. The gradient is established by division 7, and remains stable until after the blastoderm stage.

What is the consequence of establishing the bicoid gradient? The gradient can be increased or decreased by changing the number of functional gene copies in the mother. The concentration of bicoid protein is

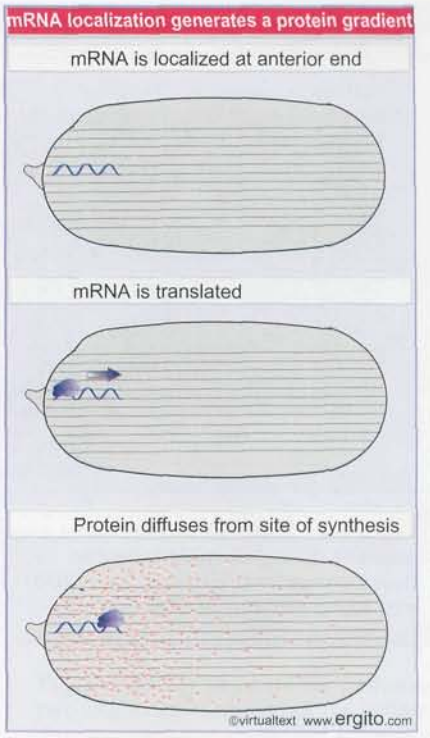


Figure 31.5 Translation of a localized mRNA generates a gradient of protein as the product diffuses away from the site of synthesis.

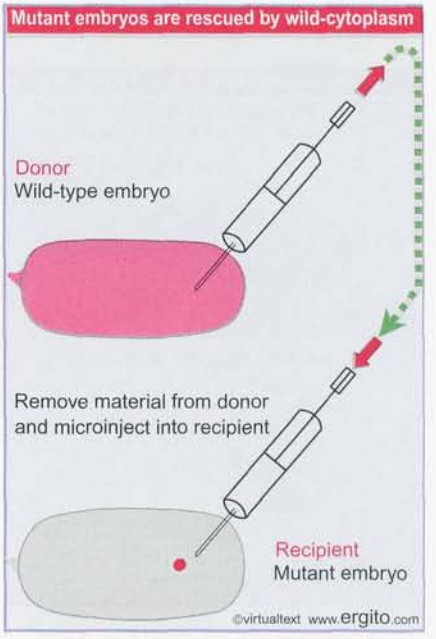


Figure 31.6 Mutant embryos that cannot develop can be rescued by injecting cytoplasm taken from a wild-type embryo. The donor can be tested for time of appearance and location of the rescuing activity; the recipient can be tested for time at which it is susceptible to rescue and the effects of injecting material at different locations.

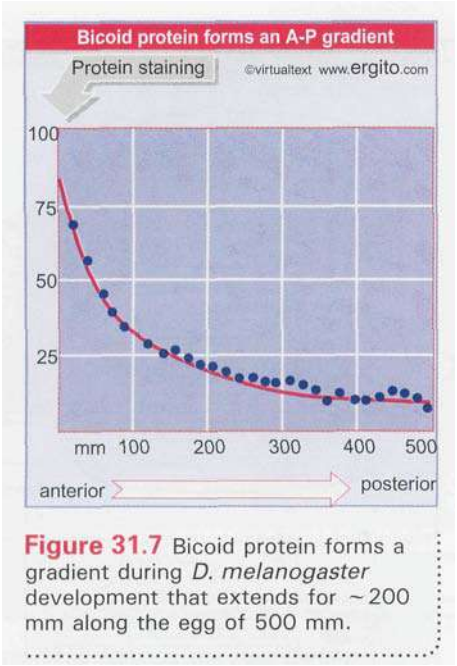


Figure 31.7 Bicoid protein forms a gradient during *D. melanogaster* development that extends for ~200 mm along the egg of 500 mm.

correlated with the development of anterior structures. Weakening the gradient causes anterior segments to develop more posterior-like characteristics; strengthening the gradient causes anterior-like structures to extend farther along the embryo. So the bicoid protein behaves as a morphogen that determines anterior-posterior position in the embryo in a concentration-dependent manner.

The fate of cells in the anterior part of the embryo is determined by the concentration of bicoid protein. The *bicoid* product is a sequence-specific DNA-binding protein that regulates transcription by binding to the promoters of its target genes. The immediate effect *of bicoid* is exercised on other genes that in turn regulate the development of yet further genes. A major target for *bicoid* is the gene *hunchback*. Transcription of *hunchback* is turned on by *bicoid* in a dose-dependent manner, that is, *hunchback* is activated above a certain threshold of bicoid protein. The effect of bicoid on hunchback is to produce a band of expression that occupies the anterior part of the embryo.

The relationship between *bicoid* and *hunchback* establishes the principle that a gradient can provide a spatial on-off switch that affects gene expression. In this way, quantitative differences in the amount of the morphogen (bicoid protein) are transformed into qualitatively different states (cell structures) during embryonic development. *bicoid* plays an instructive role in anterior development, since it is a positive regulator that is *needed for expression* of genes that in turn determine the synthesis of anterior structures.

31.6 Posterior development uses another localized regulator

Key Concepts

- The posterior system localizes nanos mRNA at the posterior end of the egg, generating a gradient of nanos protein that extends along the abdominal region.
- This system is permissive because its function is to repress genes whose products would interfere with posterior development.

Posterior development depends on the expression of a large group of genes. Embryos produced by females who are mutant for any one of these genes develop normal head and thoracic segments, but lack the entire abdomen. Some of these genes are concerned with exporting material from the nurse cells to the egg; others are required to transport or to localize the material within the egg.

The posterior pathway functions by a series of events in which one product is responsible for localizing the next. **Figure 31.8** correlates the order of genes in the genetic pathway with the activities of their products in the embryo. The functions *spir* and *capu* are needed for Staufen protein to be localized at the pole. Staufen protein in turn localizes *oskar* RNA; possibly a complex of Staufen protein and *oskar* RNA is assembled. These functions are needed to localize Vasa, which is an RNA-binding protein. Its specificity and targets are not known.

If *oskaris* overexpressed or mislocalized in the embryo, it induces germ cell formation at ectopic sites. It requires only the products of *vasa* and *tudor*. This implies that all of the activities that precede *oskar* in the pathway are needed only to localize *oskar* RNA. The ability both to form pole cells and to induce abdominal structures is possessed by *oskar*, in conjunction with *vasa* and *tudor* (and of course any components that are ubiquitously expressed in the egg). One effect of *oskar* function is to localize Vasa protein at the posterior end. The functions of

valois and *tudor* are not known, but it is possible that *valois* is off the main pathway.

Two types of pattern-determining event occur at the posterior pole, and the pathway branches at *tudor*. The polar plasm contains two morphogens: the posterior determinant (nanos) controls abdominal development; and another signal controls formation of the pole cells, which will give rise to the germline (see Figure 31.2). All of the posterior genes except *nanos* and *pumilio* are required for both processes, that is, they are defective in both abdominal development and pole cell formation. *nanos* and *pumilio* identify the abdominal branch. We do not know whether there are additional functions representing a separate branch for germ cell formation, or whether the pathway up to *tudor* is by itself sufficient.

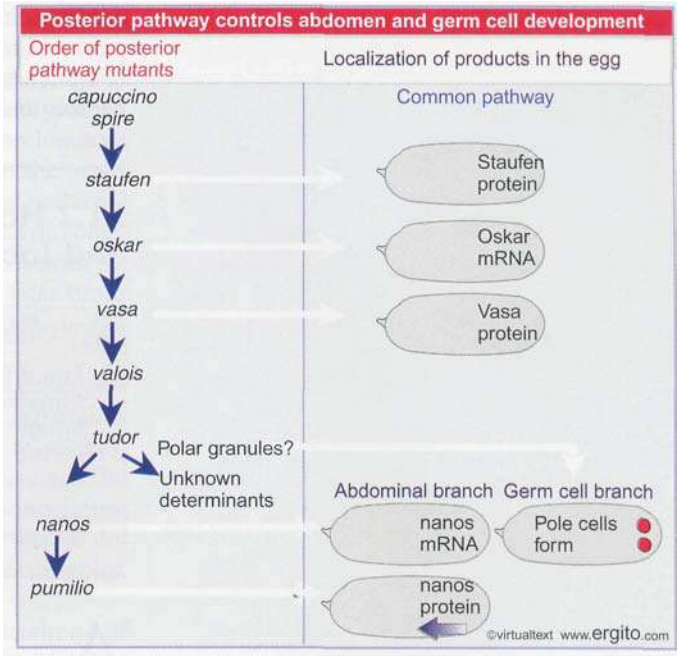


Figure 31.8 The posterior pathway has two branches responsible for abdominal development and germ cell formation.

The posterior system resembles the anterior system in the basic nature of the morphogenetic event: a maternal mRNA is localized at the posterior pole. This is the product of *nanos*, and provides the morphogen. There are two important differences between the systems. Localization is more complex than in the case of the anterior system, because posterior determinants that originate in the nurse cells must be transported the full length of the oocyte to the far pole. And nanos protein acts to *prevent* translation of a transcription factor (hunchback). Its role is said to be permissive, since it functions to repress genes whose products would interfere with posterior development.

How do we know that *nanos* is the morphogen at the end of the pathway? Rescue experiments (along the lines shown previously in Figure 31.6) with the mutants in the posterior group showed that in all but one case the cytoplasm of the nurse cell contained the posterior determinant (although it was absent from the posterior end of the oocyte itself). This indicates that these mutants all act in some subsidiary role, most probably concerned with transporting or localizing the morphogen in the egg. The exception was *nanos*, whose mutants did not contain any posterior-rescuing activity. Purified *nanos* RNA can rescue mutants in any of the other posterior genes, indicating that it is the last, or most downstream, component in the pathway. Indeed, injection *of nanos* RNA into ectopic locations in embryos can induce the formation of abdominal structures, showing that it provides the morphogen.

The upper part of **Figure 31.9** shows the localization of *nanos* mRNA at the posterior end of an early embryo. But the localization poses a dilemma: *nanos* activity is required for development of abdominal segments, that is, for structures occupying approximately the posterior half of the embryo. How does *nanos* RNA at the pole control abdominal development? The lower part of Figure 31.9 shows that nanos protein diffuses from the site of translation to form a gradient that extends along the abdominal region.

Both *bicoid* and *nanos* act on the expression of the *hunchback* gene. *hunchback* codes for a repressor of transcription: its presence is needed for formation of anterior structures (in the region of the thorax), and its absence is required for development of posterior structures. It has a complex pattern of expression. It is transcribed during oogenesis to give an mRNA that is uniformly distributed in the egg. After fertilization, the *hunchback* pattern is changed in two ways. The bicoid gradient activates synthesis of *hunchback* RNA in the anterior region. And nanos prevents translation of *hunchback* mRNA in the posterior region; a result of this inhibition is that the mRNA is degraded.

The anterior and posterior systems together therefore enhance hunchback levels in the anterior half of the egg, and remove it from the posterior half. The significance of this distribution lies with the genes

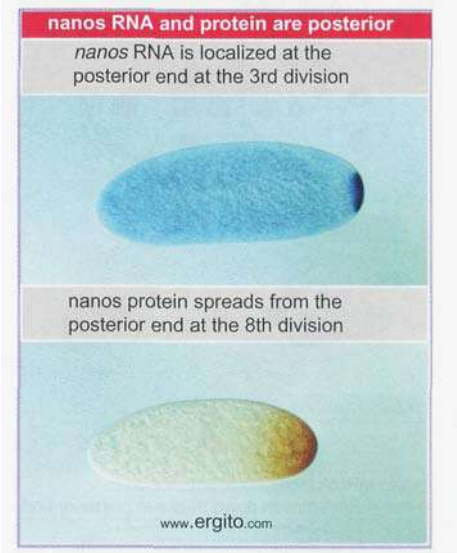


Figure 31.9 nanos products are localized at the posterior end of a *Drosophila* embryo. The upper photograph shows the tightly localized RNA in the very early embryo (at the time of the 3rd nuclear division). The lower photograph shows the spreading of nanos protein at the 8th nuclear division. Photographs kindly provided by Ruth Lehmann.

that hunchback regulates. It represses the genes *knirps* and (probably) *giant,* which are needed to form abdominal structures. So *the basic role of* hunchback *is to repress formation of abdominal structures by preventing the expression of* knirps *and* giant *in more anterior regions.*

31.7 How are mRNAs and proteins transported and localized?

Key Concepts

- The mRNAs that establish the anterior and posterior systems are transcribed in nurse cells and transported through cytoplasmic bridges into oocytes.
- *bicoid* mRNA is localized close to the point of entry, but *oskar* and *nanos* mRNAs are transported the length of the oocyte to the posterior end.
- Movement is accomplished by a motor attached to microtubules.

Anterior and posterior development both depend on the localization of an mRNA at one end of the egg. How does the mRNA reach the appropriate location and what is responsible for maintaining it there? Similar processes are involved for the anterior-posterior axis and for the dorsal-ventral axis. On the antero-posterior axis, *bicoid* and *oskar* mRNAs are localized at opposite ends of the egg. On the dorsal-ventral axis, *gurken* mRNA is initially localized at the posterior end and then becomes localized on the dorsal side of the anterior end. The principle is that the sites of transcription are distinct from the sites where the mRNAs are localized and translated, and an active transport process is required to localize the mRNAs.

bicoid, oskar, and *nanos* all are transcribed in nurse cells. Figure 31.10 shows that mRNA is transported through the cytoplasmic bridges into the oocyte. Within the oocyte, *bicoid* mRNA then remains at the anterior end, but *oskar* mRNA is transported the length of the oocyte to the posterior end. The typical means by which an mRNA is transported to a specific location in a cell involves movement along "tracks," which in principle can be either actin filaments or microtubules. This basically means that the mRNA is attached to the tracks by a motor protein that uses hydrolysis of ATP to drive movement (see Figure 31.10). In the example of the *Drosophila* egg, microtubules are the tracks used to transport these and other mRNAs. In fact, the microtubules form a continuous network that connects the oocyte to the nurse cells through the ring canals.

Genes whose products are needed to transport these mRNAs are identified by mutants in which the mRNAs are not properly localized. The most typical disruption of the pattern is for the mRNAs simply to be distributed throughout the egg. The best characterized of these transport genes are *exuperantia (exu)* and *swallow (swa).* Exu protein is part of a large ribonucleoprotein complex. This complex assembles in the nurse cell, where it uses microtubule tracks to move to the cytoplasmic bridge. Then it passes across the bridge into the oocyte in a way that is independent of microtubules. In the oocyte, it attaches to microtubules to move to its location.

The properties of *exu* and *swa* mutants show that there are common components for the transport and localization of different mRNAs. We do not yet know what differences exist between the complexes involved in transporting different mRNAs. However, we assume that there must be a component of each complex that is responsible for targeting it to the right location. By following different mRNAs, it seems that in each case the complex is transported to the anterior end of the oocyte, where

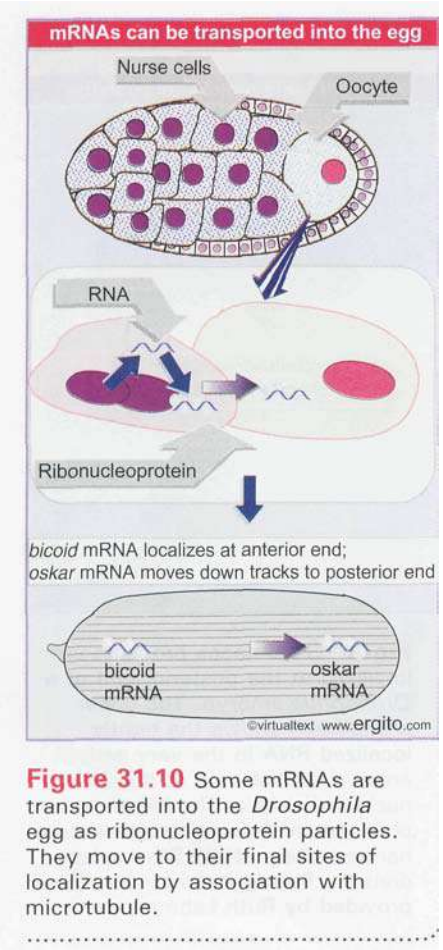


Figure 31.10 Some mRNAs are transported into the *Drosophila* egg as ribonucleoprotein particles. They move to their final sites of localization by association with microtubule.

it aggregates. Then a decision is made on further localization, and the complex is transported to the appropriate site.

Similar events occur at a later stage of development, in the syncytial blastoderm, when some mRNAs become localized on the apical side of the embryo. The same apparatus seems to be involved as in development of the oocyte. Usually it is responsible for apical localization of the products of several pair rule and segmentation genes. However, if the maternal transcripts of *gurken*, *bicoid*, or *nanos* are injected into the syncytial blastoderm, the apparatus localizes them on the apical side of the embryo. This suggests that the RNAs that are localized at early and at later times have the same set of signals to identify themselves as substrates to the localization apparatus.

mRNAs that are localized in the syncytial blastoderm are found in particles that are connected to microtubules by the motor dynein. The parallels between the transport systems of the oocyte and the blastoderm suggest that dynein also connects the maternal mRNAs to the microtubules. The proteins Egl (egalitarian) and BicD (bicoidD) associate with localizing transcripts and bind to dynein. In their absence, maternal transcripts do not localize properly in the oocyte. This suggests that an Egl/BicD complex may be the means of connecting the mRNA to the motor.

We know that the localization of the *bicoid* RNA to the anterior end of the oocyte depends upon sequences in the 3' untranslated region. This is a common theme, and localization of *oskar* and *nanos* mRNAs is controlled in the same way. We assume that the 3′ sequences provide binding sites for specific protein(s) that are involved in localization. Corresponding sequences in each mRNA will provide binding sites for the proteins that target the RNA to the appropriate sites in the oocyte. However, we are still missing the identification of the crucial protein that binds to the localizing sequence in the mRNA.

Localization of RNAs is not sufficient to ensure the pattern of expression. Translation is also controlled. The production of oskar and nanos proteins is controlled by repression of the mRNAs outside of the posterior region. In each case, translation is repressed by a protein that binds to the 3′ region. In the case of nanos, there is overlap between the elements required for localization and repression. The consequence of this overlap is to make localization and repression mutually exclusive, so that when a nanos mRNA is localized to the posterior end, it cannot be repressed.

31.8 How are gradients propagated?

Key Concepts

- Many morphogens form gradients that control differential expression of genes.
- A gradient in an egg or in the early *Drosophila* embryo is propagated by passive diffusion from a localized source.
- A gradient in a cellular tissue may be propagated by passive diffusion in the intercellular spaces or by an active process in which cells transmit the morphogen to other cells.
- A gradient may also be influenced by degradation of morphogen within cells.

The cytosol of an egg forms a single compartment, as indeed does the syncytium of the early *Drosophila* embryo. A protein may form a gradient simply by diffusing away from a localized source (see Figure 31.5). In the case *of Drosophila*, such sources are provided by localized mRNAs at either the anterior end (Figure 31.7) or posterior end (Figure 31.9).

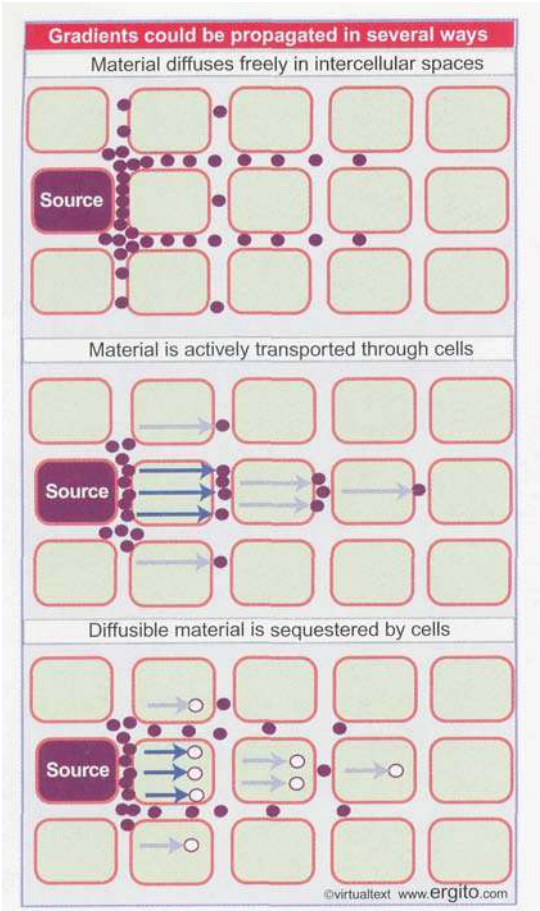


Figure 31.11 A gradient can form by passive diffusion or by active transport or may be affected by removing the diffusing material.

Gradients are also important in development of tissues consisting of cells. We know several cases in which a morphogen forms a gradient, and the cells in that gradient respond differently depending upon the local concentration of the morphogen. The differential response of cells can be seen by placing them in tissue culture on a medium that contains a gradient of morphogen. In the tissue, however, there may be more constraints upon the movement of morphogen. For a gradient to form, material must move in intercellular spaces between the cells or must be transported through the cells. **Figure 31.11** distinguishes some possible models.

The simplest situation is for passive diffusion. This appears to be responsible for the gradient of activin (a TGFβ homolog) that is secreted by cells in the amphibian embryo and induces formation of the mesoderm tissue layer (see *31.13 TGFβ/BMPs are diffusible morphogens*). **Figure 31.12** shows that the critical experiment is to create a tissue *(in vitro)* whose continuity is interrupted by a layer of cells that can neither respond to activin nor synthesize it. If the gradient is not stopped by these cells, the activin must be able either to pass freely through them or (more likely) diffuse past them. This turns out to be the case.

Formation of the anterior-posterior axis of the *Drosophila* wing is determined by a gradient of the TGFβ homolog, Dpp. Gradient formation may involve both of the mechanisms shown in Figure 31.11 for controlling distribution. The gradient cannot be propagated unless the cells in the issue have an active receptor for TGFβ and also the protein dynamin, which is involved in endocytosis (internalization) of Dpp. The basic means of propagating the gradient appears to be transcytosis, in which Dpp is taken up at one face of the cell, transported across the cell, and secreted through the membrane at the other side. Some of the Dpp may be degraded instead of being passed on to the next cell. **Figure 31.13** implies that the shape of the gradient may be influenced by the balance between these two processes.

31.9 Dorsal-ventral development uses localized receptor-ligand interactions

Key Concepts

- Gurken mRNA is localized on the dorsal side of the oocyte.
- It is translated into a TGFα-like growth factor that interacts with the Torpedo receptor on the adjoining follicle cell.
- The Torpedo receptor triggers a Ras/MAPK pathway that prevents the follicle cell from acquiring a ventral fate.

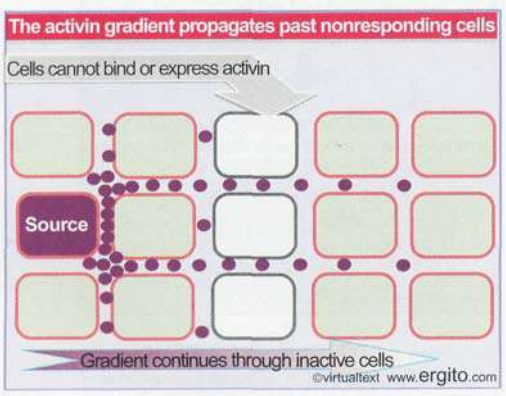


Figure 31.12 Insertion of a layer of cells that cannot interact with activin does not prevent propagation of the gradient.

Dorsal-ventral development displays a complex interplay between the oocyte and follicle cells, involving separate pathways that are required to develop ventral and dorsal structures. The formation of ventral pattern starts with the expression of genes in the oocyte that are needed for proper development of the follicle cells. And then expression of genes in the follicle cells transmits a signal to the oocyte that results in development of ventral structures. Another pathway is responsible for development of dorsal structures in the developing egg. Each of these systems functions by activating a localized ligand-receptor interaction that triggers a signal transduction pathway.

The localization of *gurken* mRNA initiates dorsal development, but also plays a role earlier in anterior-posterior patterning. These are key events that define the spatial asymmetry of the egg chamber. (The requirement of *gurken* for both pathways is the only feature that breaches their independence.)

First *gurken* mRNA is localized on the posterior side of the oocyte. This results in a signal that causes adjacent follicle cells to become posterior. The follicle cells signal back to the oocyte in a process that results in the establishment of a polarized network of microtubules. This is necessary for the localization of the maternal transcripts of *bicoid* and *oskar* to opposite poles (see *31.5 Anterior development uses localized gene regulators*).

Dorsal-ventral polarity is established later when *gurken* mRNA becomes localized on the dorsal side of the oocyte. **Figure 31.14** illustrates the pathway and its consequences. The products of *cornichon* and *brainiac* are needed for proper localization of the *gurken* mRNA or for activation of the protein. Of the group of loci that acts earlier, the products of *K10* and *squid* are needed to localize the RNA; and *cappuccino* and *spire* mutants have an array of defects that suggest their products have a general role in organizing the cytoskeleton of the oocyte. Accordingly, *cappuccino* and *spire* are required also for the earlier localization of *gurken* mRNA involved in anterior-posterior patterning.

gurken codes for a protein that resembles the growth factor TGFα. The next locus in the pathway is *torpedo*, which codes for the *Drosophila* EGF receptor. It is expressed in the follicle cells. So the pathway moves from oocyte to follicle cells when the ligand (Gurken), possibly in a transmembrane form that exposes the extracellular domain on the oocyte, interacts with the receptor (Torpedo) on the plasma membrane of a follicle cell.

An interesting and general principle emerges from the activation of Torpedo, which is a typical receptor tyrosine kinase. Activation of Torpedo leads to the activation of a Ras signaling pathway, which proceeds through Raf and D-mek (the equivalent of MAPKK), to activate a classic MAP kinase pathway. The ultimate readout of this pathway is not known, but its effect is to prevent activation on the dorsal side of the embryo of the ventral-determining pathway (see next section).

The utilization of this pathway shows that similar pathways may be employed in different circumstances to produce highly specific effects. The trigger to activate the pathway in the oocyte-follicle cell interaction is the specific localization of Gurken. The consequence is a change in the properties of follicle cells that prevents them from acquiring ventral fates. The basic components of the pathway, however, are the same as those employed in signal transduction of proliferation signals in vertebrate systems. The same pathway is employed again in the specific development of retinal cells in *Drosophila* itself, where another receptor-counter receptor interaction activates the Ras pathway, with specific, but very different effects on cell differentiation. So essentially the same pathway can be employed rij interpret an initial signal and produce a response that is predetermined by the cell phenotype.

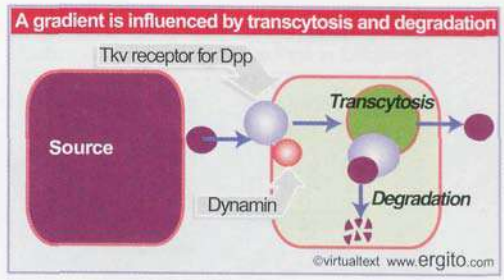


Figure 31.13 A gradient can be propagated by transcytosis, when a morphogen is endocytosed (internalized) at one face of a cell and secreted at the other face. The gradient will be sharpened if some of the morphogen is degraded within the cell.

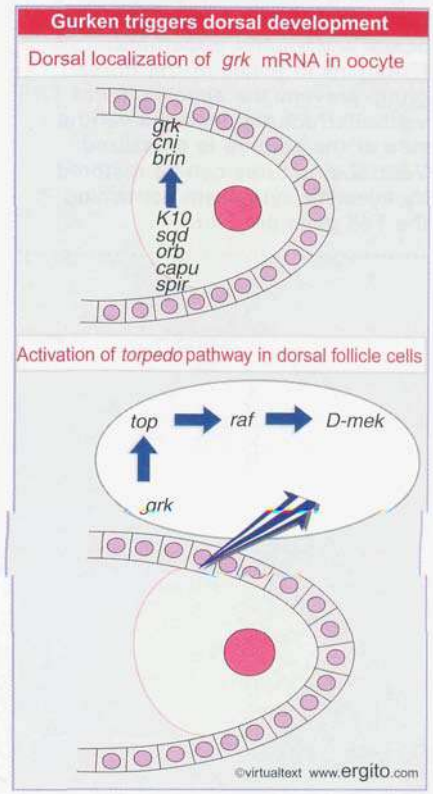


Figure 31.14 Dorsal and ventral identities are first distinguished when *grk* mRNA is localized on the dorsal side of the oocyte. Synthesis of Grk activates the receptor coded by *torpedo*, which triggers a MAPK pathway in the follicle cells.

31.10 Ventral development proceeds through Toll

Key Concepts

- The follicle cell on the ventral side produces an enzyme that modifies a proteoglycan.
- The proteoglycan triggers a series of proteolytic cleavages in the perivitelline space of the oocyte that activate the spatzle ligand.
- spatzle activates the receptor Toll, which is related to IL1 receptor, and triggers a pathway leading to activation of dorsal, which is related to the vertebrate transcription factor NF-κB.

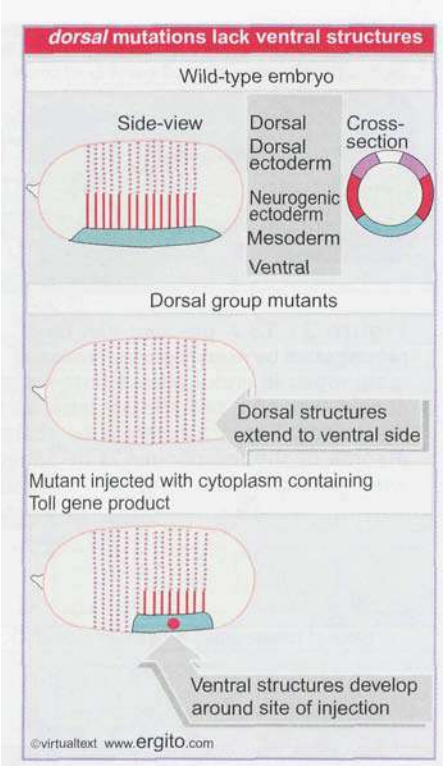


Figure 31.15 Wild-type *Drosophila* embryos have distinct dorsal and ventral structures. Mutations in genes of the *dorsal* group prevent the appearance of ventral structures, and the ventral side of the embryo is dorsalized. Ventral structures can be restored by injecting cytoplasm containing the Toll gene product.

Development of ventral structures requires a group of 11 maternal genes whose products establish the dorsal-ventral axis between the time of fertilization and cellular blastoderm (see Figure 31.4). **Figure 31.15** shows that the dorsal-ventral pattern can be viewed from the side by the phenotype of the cuticle, and can be seen in cross-section to represent the formation of different types of tissues. The *dorsal* system is necessary for the development of ventral structures including the mesoderm and neurogenic ectoderm. (The system was named for the effects of mutations [to dorsalize], rather than for the role of the gene products [to ventralize].) Mutants in any genes of the *dorsal* group lack ventral structures, and have dorsal structures on the ventral side, as indicated in the figure. But injecting wild-type cytoplasm into mutant embryos rescues the defect and allows ventral structures to develop.

The ventral-determining pathway also begins in the follicle cell and ends in the oocyte. The pathway is summarized in **Figure 31.16.** The initial steps are not well defined and require the expression of three loci in the follicle cells on the ventral side. These loci function before fertilization, but the egg does not receive the signal until after fertilization.

The three loci that act in the follicle cells are *nudel* (*ndl*), *windbeutel* (*wind*), and *pipe*. The roles of ndl and wind are not known, but *pipe* plays an interesting and novel role. *pipe* codes for an enzyme whose sequence suggests that it is similar to the enzyme heparan sulfate 2-0-

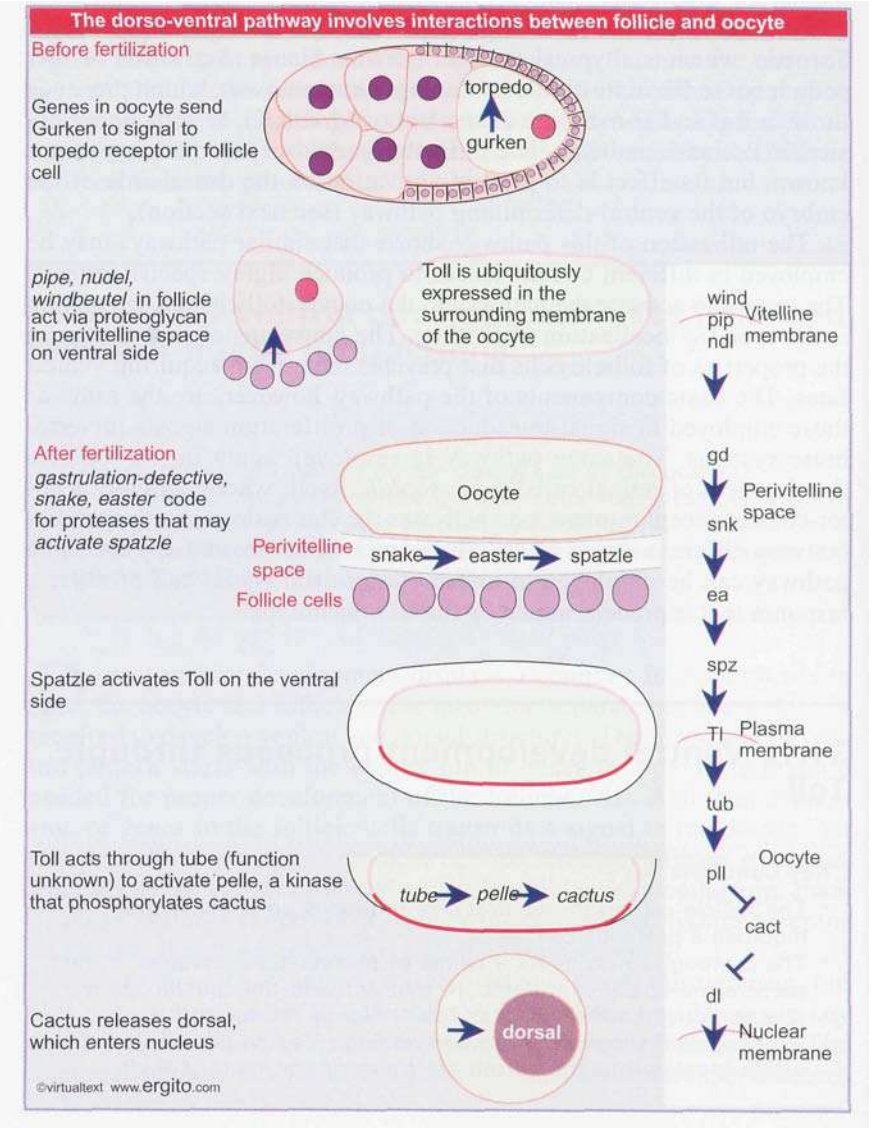


Figure 31.16 The dorsal-ventral pathway is summarized on the right and shown in detail on the left. It involves interactions between follicle cells and the oocyte. The pathway moves into the oocyte when spatzle binds to Toll and activates the morphogen. The pathway is completed by transporting the transcription factor dorsal into the nucleus.

sulfotransferase (HSST) that is involved in the synthesis of a class of proteins called proteoglycans. They are components of the extracellular matrix. These proteins have covalently attached carbohydrate side-chains called glycosaminoglycans that have a characteristic pattern of attached monosaccharides. HSST is an enzyme that adds sulfate to the 2-0 position of certain sugar residues in the monosaccharide.

The *pipe* gene is expressed in follicle cells on the ventral side of the embryo. We assume that, like HSST, it functions within the Golgi apparatus of the follicle cell. Its substrate is not known, but **Figure 31.17** shows that it is probably secreted from the follicle cell into the perivitelline space (the outermost layer of the oocyte). Because the proteoglycan is synthesized on the ventral side, this creates an asymmetry at the surface of the egg.

The presence of the proteoglycan in some unknown way triggers a series of proteolytic cleavages that occur in the perivitelline space. Several proteases act in succession, ending with the cleavage of the *spatzle* product. spatzle provides a ligand for a receptor coded by the *Toll* gene. Toll is the first component of the pathway that functions in the oocyte.

Rescue experiments identify *Toll* as the crucial gene that conveys the signal into the oocyte. *Toll*$^-$ mutants lack any dorsal-ventral gradient, and injection of Toll induces the formation of dorsal-ventral structures. The other genes of the dorsal group code for products that either regulate or are required for the action of *Toll,* but they do not establish the primary polarity.

There is a paradox in the distribution of Toll protein. *Toll* gene product activity is found in all parts of a donor embryo when cytoplasm is extracted and tested by injection. Yet it induces ventral structures only in the appropriate location in normal development. An initial general distribution of *Toll* gene product must therefore in some way be converted into a concentration of active product by local events.

Toll is a transmembrane protein homologous to the vertebrate interleukin-1 (IL1) receptor. It is located in the plasma membrane of the egg cell, with its ligand-binding domain extending into the perivitelline space. Binding of ligand is sufficient to activate the ventral-determining pathway. The reaction occurs on the ventral side of the perivitelline space. The spatzle ligand either cannot diffuse far from the site where it is generated, or perhaps it binds to Toll very rapidly, with the result that Toll is activated only on the ventral side of the embryo. Loss-of-function mutations in Toll are dorsalized, because the receptor cannot be activated. There are also dominant (*Toll*^{D}) mutations, which confer ventral properties on dorsal regions; these are gain-of-function mutations, which are ventralized because the receptor is constitutively active. Genetic analysis shows that *toll* acts via *tube* and *pelle.* Tube is probably an adaptor protein that recruits the kinase pelle to the activated receptor. The target for the pelle kinase is not proven, but its activation leads to phosphorylation of the product of *cactus,* which is the final regulator of the transcription factor coded by *dorsal.*

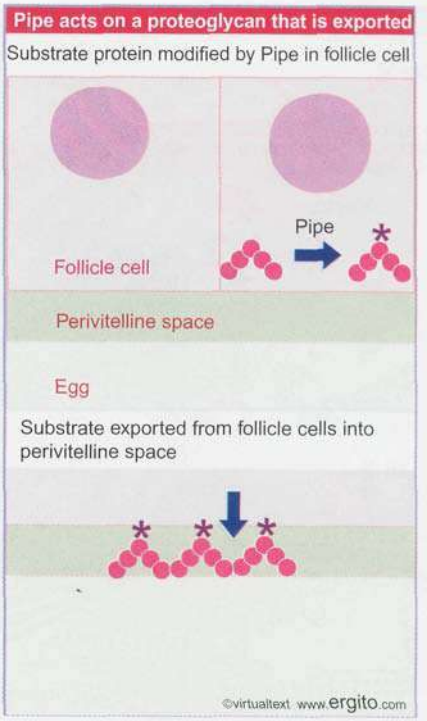


Figure 31.17 Pipe modifies a proteolglycan within the follicle cells. The modified protein is exported to the perivitelline space.

31.11 Dorsal protein forms a gradient of nuclear localization

Key Concepts

- The activation of dorsal is achieved by releasing it in the cytoplasm so that it can enter the nucleus.
- A gradient of dorsal with regard to nuclear localization is established along the ventral to dorsal axis.

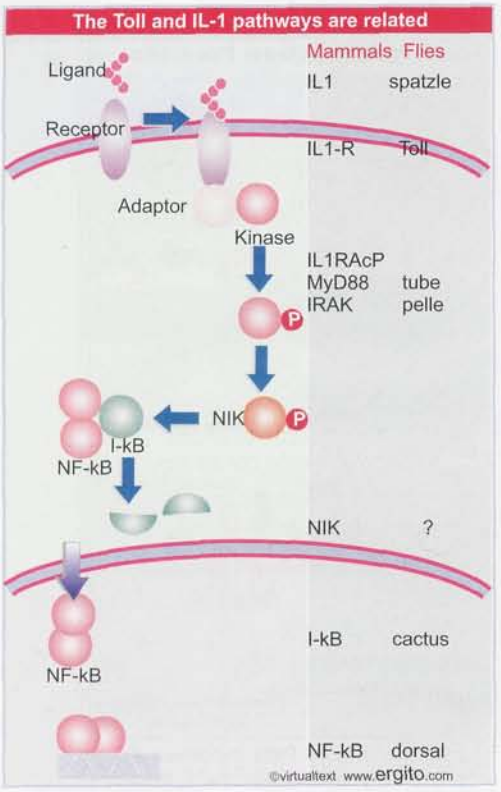


Figure 31.18 Activation of IL1 receptor triggers formation of a complex containing adaptor(s) and a kinase. The IRAK kinase activates NIK, which phosphorylates I-κB. This triggers degradation of I-κB, releasing NF-κB, which translocates to the nucleus to activate transcription.

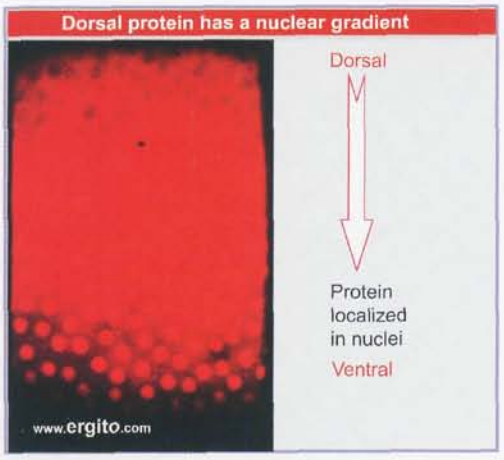


Figure 31.19 Dorsal protein forms a gradient of nuclear localization from ventral to dorsal side of the embryo. On the ventral side (lower) the protein identifies bright nuclei; on the dorsal side (upper) the nuclei lack protein and show as dark holes in the bright cytoplasm. Photograph kindly provided by Michael Levine.

Figure 31.18 shows the parallels between the *toll* signaling pathway in flies and the IL1 vertebrate pathway (where the biochemistry is well characterized). Activation of the receptor causes a complex to assemble that includes adaptor proteins (several in vertebrates), which bind a kinase. Activation of the vertebrate kinase (IRAK) in turn activates the kinase NIK, which phosphorylates I-κB. It is not clear whether the fly kinase (pelle) acts directly on cactus (the equivalent of Iκ-B) or through an intermediate.

At all events, dorsal and cactus form an interacting pair of proteins that are related to the transcription factor NF-KB and its regulator IκB. NF-KB consists of two subunits (related in sequence) which are bound by IKB in the cytoplasm. When IKB is phosphorylated, it releases NF-KB, which then moves into the nucleus, where it functions as a transcription factor of genes whose promoters have the κB sequence motif. (An example of the pathway is illustrated in Figure 22.12.) Cactus regulates dorsal in the same way that IKB regulates NF-KB. A cactus-dorsal complex is inert in the cytoplasm, but when cactus is phosphorylated, it releases dorsal protein, which enters the nucleus. The pathway is therefore conserved from receptor to effector, since activation of interleukin-1 receptor has as a principal effect the activation of NF-KB, and activation of Toll leads to activation of dorsal. A related pathway, triggered by a Toll-like receptor (TLR), is found in the system of innate immunity that is conserved from flies to mammals (see *26.21 Innate immunity utilizes conserved signaling pathways).*

As a result of the activation of Toll, a gradient of dorsal protein in the nucleus is established, from ventral to dorsal side of the embryo. On the ventral side, dorsal protein is released to the nucleus, but on the dorsal side of the embryo it remains in the cytoplasm. A steep gradient is established at the stage of syncytial blastoderm, and becomes sharper during the transition to cellular blastoderm. The proportion of dorsal protein that is in the nucleus correlates with the ventral phenotype that will be displayed by this region. An example of a gradient visualized by staining with antibody against dorsal protein is shown in **Figure 31.19.** The total amount of dorsal protein in the embryo does not change: the gradient is established solely by a redistribution of the protein between nucleus and cytoplasm.

Dorsal both activates and represses gene expression. It activates the genes *twist* and *snail*, which are required for the development of ventral structures. And it represses the genes *dpp* and *zen*, which are required for the development of dorsal structures; as a result, these genes are expressed only in the 40% most dorsal of the embryo (see *31.13 TGFβ/BMPs are diffusible morphogens*).

One of the crucial aspects of dorsal-ventral development is the relationship between the different pathways. This is summarized in **Figure 31.20.** The ability of one system to repress the next is responsible for restricting the localized activities to the appropriate part of the embryo. The initial interaction between gurken and torpedo leads to the repression of spatzle activity on the dorsal side of the embryo. This restricts the activation of dorsal protein to the ventral side of the embryo. Nuclear localization of dorsal protein in turn represses the expression of *dpp*, so that it forms a gradient diffusing from the dorsal side. In this way, ventral structures are formed in the nuclear gradient of dorsal protein, and dorsal structures are formed in the gradient of dpp protein.

The terminal system is initiated in a way that is similar to the dorsal-ventral system. A transmembrane receptor, coded by the *torso* gene, is produced by translation of a maternal RNA after fertilization. The receptor is localized throughout the embryo. It is activated at the poles by local production of an extracellular ligand. Torso protein has a kinase activity, which initiates a cascade that leads to local expression of the *tailless* and *huckebein* RNAs, which code for factors that regulate transcription.

31.12 Patterning systems have common features

The pattern of regulators at each stage of development for each of the systems is summarized in **Figure 31.21.** Two types of mechanism are used to create the initial asymmetry. For the anterior-posterior axis, an RNA is localized at one end of the egg (bicoid for the anterior system, nanos for the posterior system); localization depends upon the interaction of sequences in the 3' end of the RNA with maternal proteins. In the case of the dorsal-ventral and terminal systems, a receptor protein is specifically activated in a localized manner, as the result of the limited availability of its ligand. All of these interactions depend on RNAs and/or proteins expressed from maternal genes.

The local event leads to the production of a morphogen, which forms a gradient, either quantitatively (bicoid) or by nucleocytoplasmic distribution (dorsal), or is localized in a broad restricted region (nanos). The extent of the region in which the morphogen is active is ~50% across the egg for each of these systems. The morphogens are translated from maternal RNAs, and development is therefore still dependent on maternal genes up to this stage.

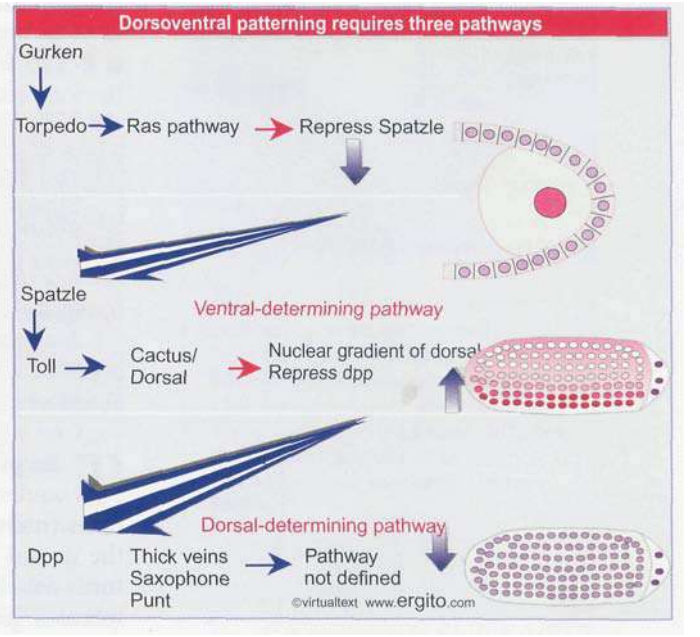


Figure 31.20 Dorsal-ventral patterning requires the successive actions of three localized systems.

Establishing anterior-posterior and dorsal-ventral gradients is the first step in determining orientation and spatial organization of the embryo. Under the direction of maternal genes, gradients form across the common cytoplasm and influence the behavior of the nuclei located in it. The next step is the development of discrete regions that will give rise to different body parts. This requires the expression of the zygotic genome, and the loci that now become active are called *zygotic genes.* Genes involved at this stage are identified by segmentation mutants.

The products of the segmentation genes form bands that distinguish individual regions on the anterior-posterior axis. When we consider the results of the anterior and posterior systems together, we see that there are several broad regions (the two regions generated by the anterior system are adjacent to the two regions defined by the posterior system; see Figure

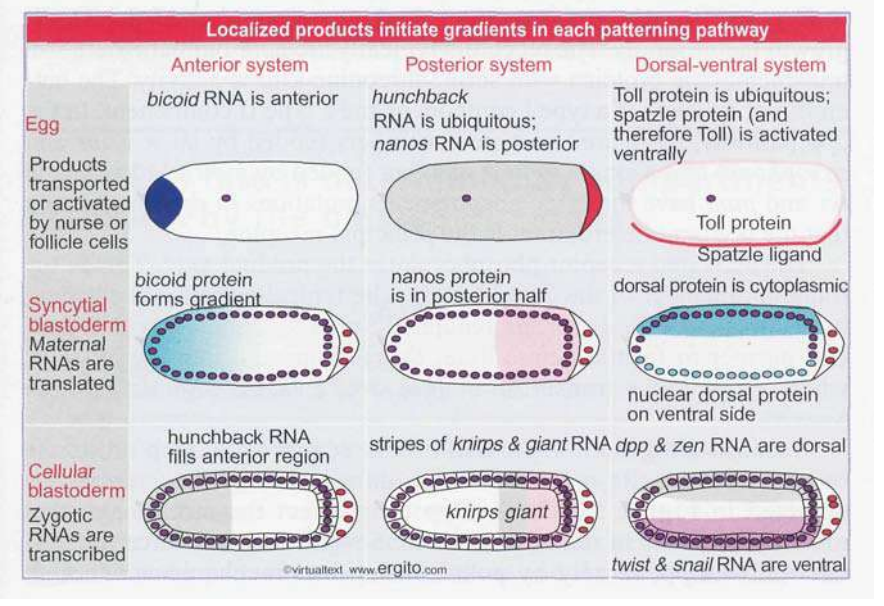


Figure 31.21 In each axis-determining system, localized products in the egg cause other maternal RNAs or proteins to be broadly localized at syncytial blastoderm, and zygotic RNAs are transcribed in bands at cellular blastoderm.

31.27). On the dorsal-ventral axis, there are three rather broad bands that define the regions in which the mesoderm, neuroectoderm, and dorsal ectoderm form (proceeding from the ventral to the dorsal side).

31.13 TGFβ/BMPs are diffusible morphogens

Key Concepts

- The TGFβ/BMP family provides ligands for receptors that activate Smads transcription factors.
- Synthesis of the Dpp member of this family is repressed on the ventral side of the fly embryo.
- It diffuses from a source on the ventral side and induces neural tissues.
- A similar pathway functions in vertebrates but is inverted with regard to the dorsal-ventral axis.

The principle of dorsal-ventral development in flies, amphibians, and mammals is the same. On one side of the animal, neural structures (including the CNS) develop. This is the ventral side in flies, and the dorsal side in vertebrates. On the other side, mesenchymal structures develop. This is the dorsal side in flies, and the ventral side in vertebrates. The important point here is that the same relative development is seen from one side of the animal to the other, but its absolute direction is reversed between flies and vertebrates. This must mean that the dorsal-ventral axis was inverted at some point during evolution, causing the CNS to be displaced from the ventral side to the dorsal side. This idea is supported by the fact that the same signaling pathway is initiated on the dorsal side of flies and on the ventral side of vertebrate embryos.

Mesenchymal (non-neural) structures are determined by diffusible factors in the TGFβ/BMP family. These factors are small polypeptide ligands for receptors that activate the Smads transcription factors (see Figure 28.46). Formation of neural structures requires counteracting activities that also diffuse from a center; they prevent the TGFβ/BMP ligands from activating the target receptors. (The names reflect the histories of their discoveries as transforming growth factor β and bone morphogenetic proteins; but in fact the most important role of these polypeptides is as morphogens in development.)

The involvement of this pathway in development was first described in *Drosophila*, where the product of *dpp* is a member of the TGFβ growth factor family. The receptors typically are heterodimers that form transmembrane proteins with serine/threonine kinase activity. The heterodimer consists of a type 1 component and a type II component. In the *dpp* pathway, there are two type I members (coded by *thick veins* and *saxophone)* and a single type II member (coded by *punt)*. Mutations in *tkv* and *punt* have the same phenotype as mutations in *dpp,* suggesting that the tkv/punt heterodimer is the principal receptor.

The activated receptor phosphorylates the product *mad*. This is the founding member of the Smad family. The typical pattern of activation in mammalian cells is for the regulated Smad to associate with a general partner to form a heterodimer that is imported into the nucleus, where it activates transcription (see *28.21 TGFβ signals through Smads)*.

Because the gene is repressed on the ventral side, Dpp protein is secreted from cells only across the dorsal side of the embryo, as depicted in **Figure 31.22.** So Dpp is in effect the morphogen that induces synthesis of dorsal structures. Several loci influence the production of Dpp, largely by post-translational mechanisms. The net

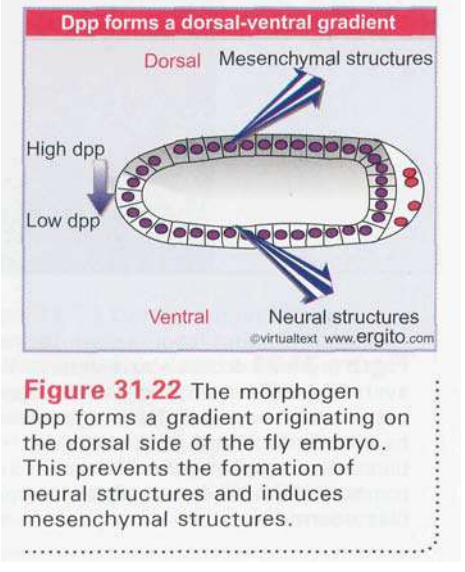


Figure 31.22 The morphogen Dpp forms a gradient originating on the dorsal side of the fly embryo. This prevents the formation of neural structures and induces mesenchymal structures.

result is to increase Dpp activity on the dorsal side, and to repress it on the ventral side, of the embryo. The concentration of Dpp directly affects the cell phenotype, the most dorsal phenotypes requiring the greatest concentration.

The same pathway is involved in inducing the analogous structures in frog or mouse, but it is inverted with regard to the Dorsal-ventral axis. **Figure 31.23** shows that Bmp4 is secreted from one side of the egg. It is antagonized by a variety of factors. Neural tissues develop in the (dorsal) regions which Bmp4 is prevented from reaching.

The crucial unifying feature is that neural tissues are induced when the activity of Dpp/Bmp4 is antagonized. Typically the Dpp/Bmp diffuses from a source, and different phenotypes may be produced by different concentrations of the morphogen. It is controversial whether the morphogen diffuses extracellularly or whether there may be a relay system that propagates it from cell to cell. Analogous pathways, triggered by different Bmps, are involved in the development of many organs.

Figure 31.24 compares the pathways in fly and frog. The basic principle is to control the availability of Dpp/Bmp. An antagonist binds to Dpp/Bmp and prevents it from binding to its receptor. The antagonists are large extracellular proteins. The antagonist is destroyed by a protease, releasing Dpp/Bmp. Neural tissue is formed in regions where Dpp/Bmp actions is prevented, whereas ectodermal tissue is formed in regions where Dpp/Bmp is activated.

The fly pathway is well characterized for dorsal-ventral development. There are two types of mutants. Mutations in *sog* and *tsg* identify genes whose products antagonize Dpp, whereas mutations in *tolloid* suggest that it activates Dpp. Sog fulfills the role of antagonist illustrated in Figure 31.24, and is destroyed by the protease tolloid. Tsg is a sort of co-antagonist, which enhances the effect of Sog.

The biochemical reactions actually have been better characterized for the corresponding frog proteins (Chordin is related to Sog). Frogs may have several such pathways, with a variety of Bmp ligands that interact in an overlapping manner with a family of receptors. The frog pathway shown in Figure 31.24 is for the ventralizing effects of Bmp4, but the others are similar, although their specific effects on morphogenetic determination are of course different. There can be variation in specificity at each stage of the pathway. The antagonists, ligands, and receptors may be expressed in different places and times, providing specificity with regard to local concentrations of the morphogen, but there may also be partial redundancy. The genes for two proteins (Noggin and Chordin) both must be knocked out in mouse to produce a phenotype. Each receptor has specificity for certain Smads, so that different target genes can be activated in different tissues.

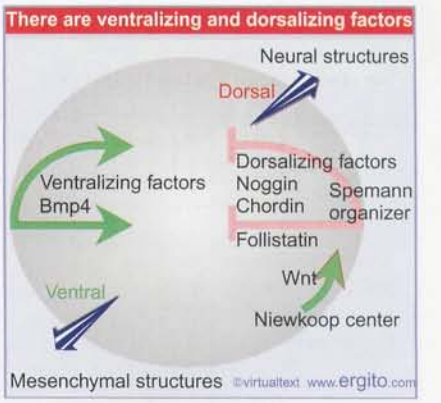


Figure 31.23 Two common pathways are used in early development of *Xenopus*. The Niewkoop center uses the Wnt pathway to induce the Spemann organizer. The organizer diffuses dorsalizing factors that counteract the effects of the ventralizing BMPs.

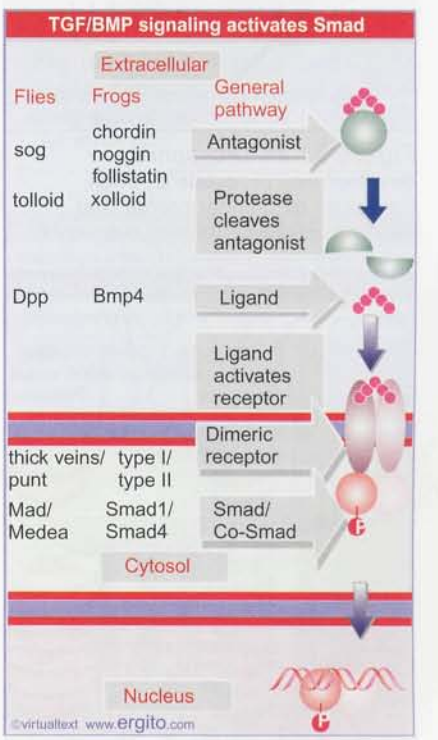


Figure 31.24 The TGFβ/Bmp signaling pathway is conserved in evolution. The ligand may be sequestered by an antagonist, which is cleaved by a protease. Ligand binds to a dimeric receptor, causing the phosphorylation of a specific Smad, which together with a Co-Smad translocates to the nucleus to activate gene expression.

31.14 Cell fate is determined by compartments that form by the blastoderm stage

Key Concepts

- A compartment of cells is defined at blastoderm and will give rise to a specific set of adult structures.
- Each segment consists of an anterior compartment and posterior compartment.
- Segmentation loci are divided into three groups of genes.

By the blastoderm stage, cells have begun to acquire information about the pathways they will follow and the structures they will

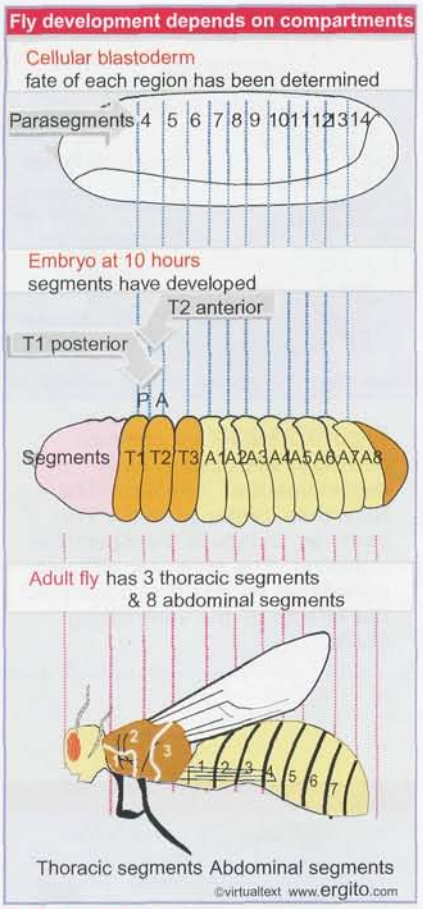


Figure 31.25 *Drosophila* development proceeds through formation of compartments that define parasegments and segments.

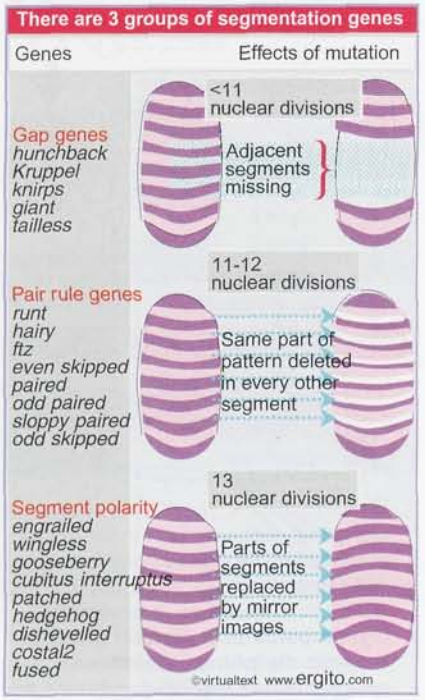


Figure 31.26 Segmentation genes affect the number of segments and fall into three groups.

therefore form. This information derives initially from the maternal regulators, and then it is further refined by the actions of zygotic genes. This makes it possible to draw a "fate map" of the blastoderm embryo to identify each region in terms of the adult segments that will develop from the descendants of the embryonic cells. The concept that is intrinsic in the fate map is that a region identified at blastoderm consists of a "compartment" of cells that will give rise specifically to a particular adult structure.

We can consider the development of *D. melanogaster* in terms of the two types of unit depicted in **Figure 31.25:** the segment and parasegment:

- The segment is a visible morphological structure. The adult fly consists of a series of clearly demarcated segments, and the larva has a series of corresponding segments separated by grooves. We are concerned primarily with the three thoracic (T) and eight abdominal (A) segments, about whose development most is known. The pattern of segmental units is determined by blastoderm, when the main mass of the embryo is divided into a series of alternating anterior (A) and posterior (P) compartments. So a segment consists of an A compartment succeeded by a P compartment; segment A3, for example, consists of compartments A3 A and A3P.
- Another type of classification originates earlier, when divisions can first be seen at gastrulation. The embryo can be divided into parasegments, each consisting of a P compartment succeeded by an A compartment. Parasegment 8, for example, consists of compartments A2P and A3A. In the 5-6 hour embryo, shallow grooves on the surface separate the adjacent parasegments. When segments form at around 9 hours, the grooves deepen and move, so that each segmental boundary represents the center of a parasegment. So the anterior part of the segment is derived from one parasegment, and the posterior part of the segment is derived from the next parasegment. In effect, the segmental units are initially evident as P-A pairs in parasegments, and then are recognized as A-P pairs in segments.

How are these compartments defined during embryogenesis? The general nature of segmentation mutants suggests that the functions of segmentation genes are to establish "rules" by which segments form; *a mutation changes a rule in such a way as to cause many or all segments to form improperly.* The drastic consequences of segment malformation make these mutants embryonic lethals—they die at various stages before metamorphosis into adults.

Probably ~30 loci are involved in segment formation. **Figure 31.26** shows that they can be classified according to the size of the unit that they affect:

- Gap gene mutants have a group of several adjacent segments deleted from the final pattern. Four gap genes are involved in formation of the major body segments, and others are concerned with the head and tail structures.
- Pair-rule mutants have corresponding parts of the pattern deleted in every other segment. The afflicted segments may be even-numbered or odd-numbered. There are 8 pair-rule genes.
- Segment polarity mutants most often lose part of the P compartment of each segment, and it is replaced by a mirror image duplication of the A compartment. Some mutants cause loss of A compartments or middle segments. There are ~16 segment polarity genes.

These groups of genes are expressed at successive periods during development; and they define increasingly restricted regions of the egg, as can be seen from **Figure 31.27.** The maternal genes establish gradients from the anterior and posterior ends. The maternal gradients either activate or repress the gap genes, which are amongst the earliest to be

transcribed following fertilization (following the 11th nuclear division); they divide the embryo into 4 broad regions. The gap genes regulate the pair-rule genes, which are transcribed slightly later; their target regions are restricted to *pairs* of segments. The pair-rule genes in turn regulate the segment polarity genes, which are expressed during the 13th nuclear division, and by now the target size is the *individual* segment.

Many of the maternal genes, the gap genes, and the pair-rule genes are regulators of transcription. Their effects may be either to activate or to repress transcription; in some cases, a given protein may activate some target genes and repress other target genes, depending on its level or the context. The genes in any one class regulate one another as well as regulating the genes of the next class. When we reach the level of segment polarity genes, the nature of the regulatory event changes, and many of the gene products act on communication between cells to maintain borders between compartments, for example, to control the secretion of a protein from one cell to influence its neighbor.

The principle that emerges from this analysis is that at each stage *a small number of maternal, gap, and pair-rule regulator proteins is used in combinatorial associations to specify the pattern of gene expression in a particular region of the embryo.*

31.15 Gap genes are controlled by bicoid and by one another

Key Concepts

- Gap genes affect a group of segments, and are controlled by bicoid and by interactions among themselves.
- Pair-rule genes are expressed in either even- or odd-numbered segments, and are controlled by the gap genes.
- Segment polarity mutants are controlled by the pair-rule genes, and are expressed in segments where they affect anterior or posterior identification of the compartments.

The gap genes are controlled in two ways: they may respond directly to the bicoid morphogen; and they regulate one another. The four bands shown in Figure 31.27 are created by the levels of the two proteins bicoid and hunchback. The synthesis of hunchback mRNA is activated by bicoid. Hunchback protein forms a gradient in the egg that in turn controls the expression of other genes.

The most anterior band in Figure 31.27 consists of hunchback protein. The next band consists of Kruppel protein; transcription of the *Kruppel* gene is activated by hunchback protein. The next two bands consist of knirps and giant proteins. Transcription of these genes is repressed by hunchback. They are expressed in the posterior part of the embryo because nanos has prevented the expression of hunchback there.

Figure 31.28 examines the transition from the 4 band to the 7 striped stage in more detail. The detailed interactions among the gap proteins are determined by examining the pattern of the distribution of other gap proteins in a mutant lacking one particular gap protein. Hunchback plays an especially important role. It is expressed in a broad anterior region, with a gradient of decline in the middle of the embryo. High levels of hunchback repress *Kruppel;* this determines the anterior boundary of Kruppel expression, which rises just as hunchback falls off, in parasegment 3. But some level of hunchback is needed for *Kruppel* expression, so when the level of hunchback decreases further, *Kruppel* is turned off, around parasegment 5. In the same way, expression of

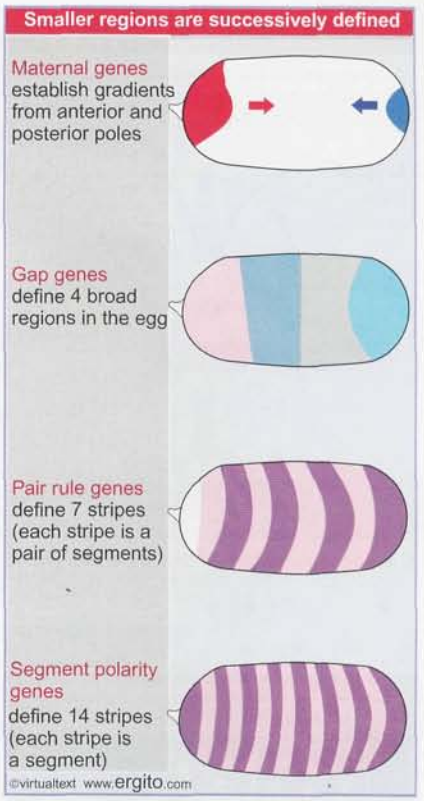


Figure 31.27 Maternal and segmentation genes act progressively on smaller regions of the embryo.

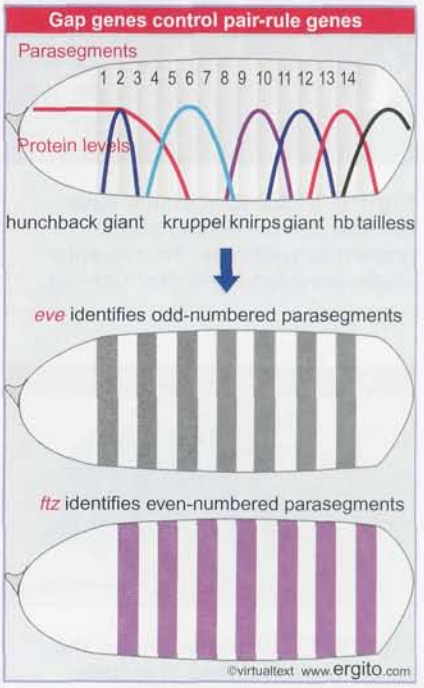


Figure 31.28 Expression of the gap genes defines adjacent regions of the embryo. The gap genes control the pair-rule genes, each of which is expressed in 7 stripes.

giant responds to successive changes in the level of hunchback; and *knirps* expression requires the absence of hunchback.

The control is refined further by interactions among the proteins. *The general principle is that one interaction may be required to express a protein in a particular region, and other interactions may be required to repress its expression at the boundaries.* The effects are worked out by examining pairwise interactions. For example, overexpression of *giant* causes the Kruppel band to become much narrower, suggesting that giant contributes to repressing the boundaries of Kruppel. The posterior margins of knirps and giant are determined by the operation of the terminal system. Altogether, these interactions mean that, as we proceed along the egg from anterior to posterior, any particular position can be defined by the levels of the various gap proteins.

31.16 Pair-rule genes are regulated by gap genes

Key Concepts

- Pair-rule genes are expressed in either even- or odd-numbered segments, and are controlled by the gap genes.

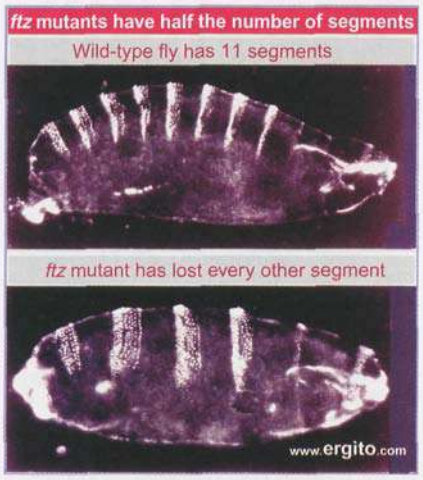


Figure 31.29 *ftz* mutants have half the number of segments present in wild-type. Photographs kindly provided by Walter Gehring.

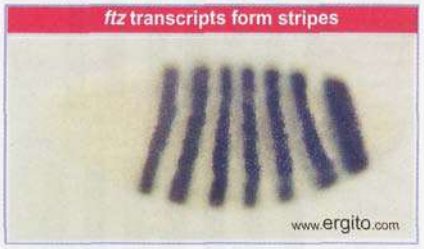


Figure 31.30 Transcripts of the *ftz* gene are localized in stripes corresponding to even numbered parasegments. The expressed regions correspond to the regions that are missing in the *ftz* mutant of the previous figure. Photograph kindly provided by Walter Gehring.

All the gap proteins are regulators of transcription, and in addition to regulating one another, they regulate the expression of the pair-rule genes that function at the next stage. Each pair-rule protein is found in a pattern of 7 "stripes" along the embryo, and Figure 31.28 shows the approximate positions that these stripes will take as the result of expression of the gap genes. (Of course, the parasegments have not developed yet, and are shown just to relate their positions to the protein distribution.) The 7 stripes of a pair-rule gene identify either all the odd-numbered parasegments (like *eve)* or all the even-numbered parasegments (like *ftz).* Two of the pair-rule genes, *hairy* and *eve,* are called primary pair-rule genes, because they are expressed first, and their pattern of expression influences the expression of the other pair-rule genes.

Recall that mutations in pair-rule genes delete half the segments. **Figure 31.29** compares the segmentation patterns of wild-type and *fushi tarazu (ftz)* larvae. The mutant has only half the number of segments, because every other segment is missing.

The ftz mRNA is present from early blastoderm to gastrula stages of development. **Figure 31.30** shows the locations of the transcripts, visualized *in situ* at blastoderm in wild type. *The gene is expressed in 7 stripes, each 3-4 cells wide, running across the embryo.* As shown previously in Figure 31.28, the stripes correspond to even-numbered parasegments (4 = T1P/T2A, 6 = T3P/A1A, 8 = A2P/A3A, etc.).

This pattern suggests a function for the *ftz* gene: *it must be expressed at blastoderm for the structures that will be descended from the even-numbered parasegments to develop.* Mutants in which *ftz* is defective lack these parasegments because the gene product is absent during the period when they must be formed. In other words, expression of *ftz* is required for survival of the cells in the regions in which it is expressed.

The expression of *ftz* is an example of the general rule that the stripes in which a pair-rule gene is expressed correspond to the regions that are missing from the embryo when the gene is mutated. *Compartments are therefore determined by the pattern of expression of segmentation genes.* The width of the stripe in which a gene is expressed

corresponds to the size of the segmental unit that it affects. Different mechanisms are used to specify the expression patterns of different pair-rule genes; we have the most information about *ftz* and *eve*.

In the early embryo, *ftz* is uniformly expressed. If protein synthesis is blocked before the stripes develop, the embryo retains the initial pattern. So the development of stripes depends on the specific degradation of *ftz* RNA in the regions between the bands and at the anterior and posterior ends of the embryo. Once the stripes have developed, transcription of *ftz* ceases in the interbands and at the ends of the embryo. The specificity of transcription depends on regions upstream of the *ftz* promoter, and also on the function of several other segmentation genes. The transcription of *ftz* responds to other pair-rule genes (and perhaps gap genes) through elements that act on all stripes.

The expression pattern of *eve* is complementary to *ftz*, but has a different basis: it is controlled separately in each stripe. A detailed reconstruction using subregions of the *eve* promoter shows that the information for localization in each stripe is coded in a separate part of the promoter; the promoter can be divided into regions that respond to the local levels of gap gene products in particular parasegments. For example, the promoter region that is responsible for *eve* expression in parasegment 3 has binding sites for the gap proteins bicoid, hunchback, giant, and Kruppel. **Figure 31.31** shows that this part of the promoter extends for 480 bp. It works in the following way. *eve* transcription is activated by hunchback and bicoid. The two boundaries are determined because the promoter is repressed by giant on the anterior side and by Kruppel on the posterior side (see also Figure 31.28). Other parts of the promoter respond to the protein levels in other parts of the embryo. So the different stripes of the primary pair-rule gene products are regulated by separate pathways, each of which is susceptible to activation by a particular combination of gap gene products and other regulators.

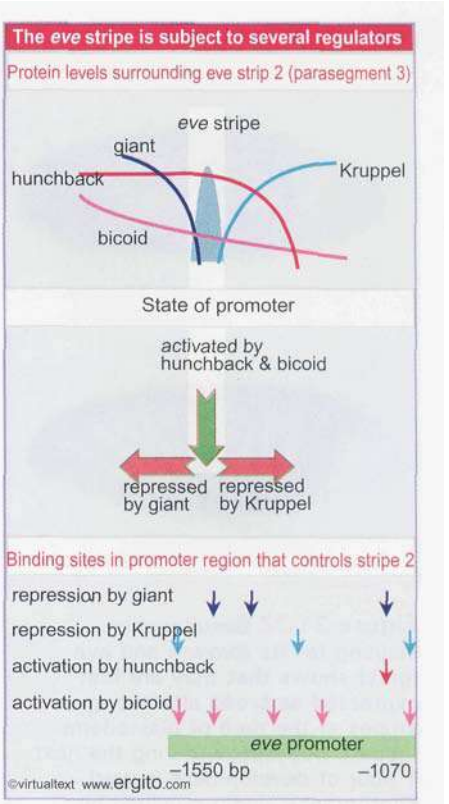


Figure 31.31 The *eve* stripe in parasegment 3 is activated by hunchback and bicoid. Repression by giant sets the anterior boundary; repression by Kruppel sets the posterior boundary. Multiple binding sites for these proteins in a 480 bp region of the promoter control expression of the gene.

This illustrates in miniature the principle that combinations of proteins control gene expression in local areas. The general principle is that generally distributed proteins (such as bicoid or hunchback) are needed for activation, whereas the borders are formed by selective repression (by giant and Kruppel in this particular example). We should emphasize that the hierarchy of gene control is not exclusively restricted to interactions between successive stages of control (maternal gap pair-rule). For example, the involvement of bicoid protein in regulating *eve* transcription in parasegment 3 shows that a maternal gene may have a direct effect on a pair-rule gene.

The stripes of *eve* and *ftz* are fuzzy to begin with, and become sharper as development proceeds, corresponding to more finely defined units. **Figure 31.32** shows an example of an embryo simultaneously stained for expression of *ftz* and *eve*. Initially there is a series of alternating fuzzy stripes, but the stripes narrow from the posterior margin and sharpen on the anterior side as they intensify during development. This may depend on an autoregulatory loop, in which the expression of the gene is regulated by its own product.

31.17 Segment polarity genes are controlled by pair-rule genes

Key Concepts

- Segment polarity genes are expressed in segments where they affect anterior or posterior identification of the compartments.

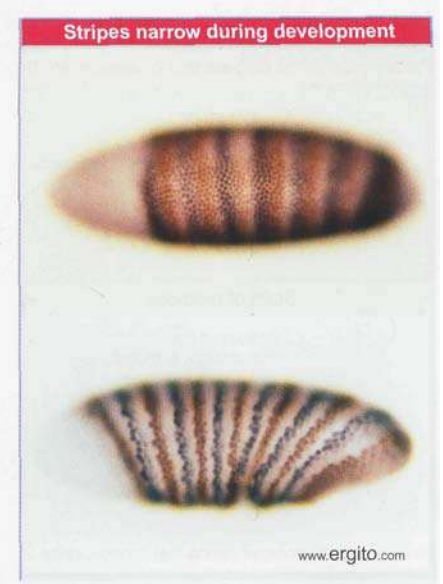


Figure 31.32 Simultaneous staining for ftz (brown) and eve (grey) shows that they are first expressed as broad alternating stripes at the time of blastoderm (upper), but narrow during the next 1 hour of development (lower). Photographs kindly provided by Peter Lawrence.

The pair-rule genes control the expression of the segment polarity genes, which are expressed in 14 stripes. Each stripe identifies a segment. The compartmental pattern in which segment polarity genes are expressed is exceedingly precise. Perhaps the ultimate demonstration of precision is provided by the pattern gene *engrailed*. The function coded by *engrailed* is needed in all segments and is concerned with the distinction between the A and P compartments. *engrailed* is expressed in every P compartment, but not in A compartments. Mutants in this gene do not distinguish between anterior and posterior compartments of the segments.

Antibodies against the protein coded by *engrailed* react against the nucleus of cells expressing it. The regions in which *engrailed* is expressed form a pattern of stripes. When the stripes of engrailed protein first become apparent, *they are only one cell wide.* **Figure 31.33** shows the pattern at a stage when each segment has a stripe just 1 cell in width, with the stripe beginning to widen into several cells.

Actually, the pattern of stripes becomes established over a 30 minute period, moving along the embryo from anterior to posterior. Initially one stripe is apparent; then every other segment has a stripe; and finally the complete pattern has a stripe 3-4 cells wide corresponding to the P compartment of every segment.

The expression of *engrailed* is of particular importance, because it defines the boundaries of the actual compartments from which adult structures will be derived. The initial 1-cell-wide stripes of engrailed protein form at the anterior boundaries of both the *ftz* and *eve* stripes, and delineate what will become the anterior boundary of every P compartment. Why is *engrailed* initially transcribed exclusively in this anterior edge, within the broader stripes of *ftz* and *eve* expression? This question is a specific example of a more general question: how can a broad stripe be subdivided into more restricted, narrower stripes? We can consider two general types of model:

- A combinatorial model supposes that different genes are expressed in overlapping patterns of stripes. A pattern of stripes develops for each of the pair-rule genes. The different pair-rule gene stripes overlap, because they are out of phase with one another. As a result of these patterns, different cells in the cellular blastoderm express different combinations of pair-rule genes. Each compartment is defined by the particular combination of the genes that are expressed, and these combinations determine the responses of the cells at the next stage of development. In other words, the segmentation genes are controlled by the pair-rule genes in the same general manner that the pair-rule genes are controlled by the gap genes.
- A boundary model supposes that a compartment is defined by the striped pattern of expression, but that interactions involving cell-cell communication at the boundaries cause subdivisions to arise within the compartment. In the case of *engrailed*, we would suppose that some unique event is triggered by the juxtaposition of cells possessing ftz (or eve) with cells that do not, and this is necessary to trigger *engrailed* expression.

Figure 31.33 Engrailed protein is localized in nuclei and forms stripes as precisely delineated as 1 cell in width. Photograph kindly provided by Patrick O'Farrell.

Each of the 14 segments is subdivided further into anterior and posterior compartments by the activities of the segment polarity genes. The actions of the segment polarity genes are the same in every segment. For example, *engrailed* distinguishes the A and P compartments.

engrailed is a transcription factor, but other segment polarity genes have different types of functions. The products of the segment polarity genes include secreted proteins, transmembrane proteins, kinases,

cytoskeletal proteins, as well as transcription factors. Cell-cell interactions become important at this stage for defining and maintaining the nature of the compartments (see next section).

31.18 Wingless and engrailed expression alternate in adjacent cells

Key Concepts

- Wingless and engrailed have a mutually reinforcing interaction between neighboring cells.
- Wingless is secreted at the posterior boundary of a cell.
- It activates the Fz (or dFz2) receptor in the adjacent cell, which triggers the translocation of Armadillo to the nucleus.
- Armadillo causes engrailed to be expressed, and engrailed causes secretion of hedgehog at the anterior boundary, where it acts on the neighboring cell to maintain wingless expression.

The circuit that defines the boundaries between anterior and posterior compartments is based on mutual interactions between segment polarity genes. *wingless* codes for a protein that is secreted and taken up by the adjacent row of cells. It is initially expressed in a row of cells immediately adjacent to the anterior side of the cells expressing *engrailed;* so *wingless* comes to identify the posterior boundary of the preceding parasegment. The initial expression of *engrailed* in response to ftz and eve is shortly replaced by an autoregulatory loop in which secretion of wingless protein from the adjacent cells is needed for expression of *engrailed;* and expression of *engrailed* is needed for expression of *wingless.* This keeps the boundary sharp.

The wingless signaling pathway is one of the most interesting, and has close parallels in all animal development. Like other signaling pathways utilized in development, it is initiated by an extracellular ligand, and results in the expression of a transcription factor, although the interactions between components of the pathway are somewhat unusual.

In fly embryonic development at the stage of segmental definition, the cells that define the boundaries of the A and P compartments express wingless (Wg) and engrailed (En) in a reciprocal relationship. **Figure 31.34** shows that wingless protein is secreted from a cell at a boundary, and acts upon the cell on its posterior side. The *wingless* signaling pathway causes the *engrailed* gene to be expressed. Engrailed causes the production of hedgehog (Hh) protein, which in turn is secreted. Hedgehog acts on the cell on its anterior side to maintain *wingless* expression. Wg is also required for patterning of adult eyes, legs, and wings (hence its name).

The identification of the receptor for Wg on the posterior cell has actually been very difficult. Wg interacts with frizzled *in vitro,* but mutational analysis suggests that the related protein, DFz2 (*Drosophila frizzled-2*) is the receptor. It is possible that these may play redundant roles. Another protein that is required for reception/signaling is the product of *arrow,* which is a single-pass membrane protein and is classified as a coreceptor. The frizzled family members are 7-membrane pass proteins, with the appearance of classical receptors (although the major pathway does not appear to involve G proteins).

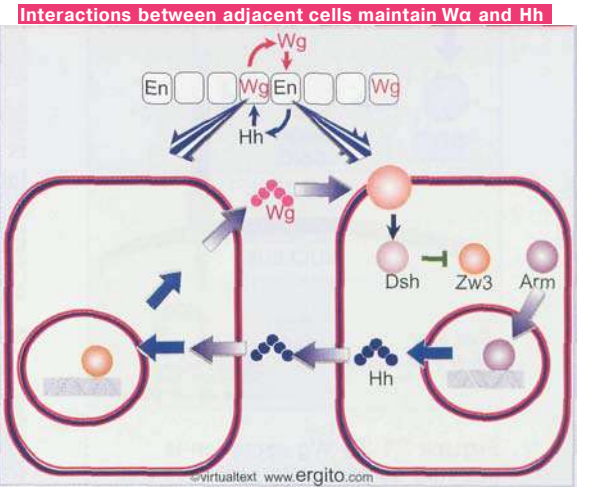


Figure 31.34 Reciprocal interactions maintain Wg and Hh signaling between adjacent cells. Wg activates a receptor, which activates a pathway leading to translocation of Arm to the nucleus. This activates engrailed, which leads to expression of Hedgehog protein, which is secreted to act on the neighboring cell, where it maintains Wg expression.

31.19 The wingless/wnt pathway signals to the nucleus

Key Concepts

- Wingless in Drosophila) and Wnt (in vertebrates) activate a receptor that blocks the action of a cytosolic Ser/Thr kinase.
- The kinase phosphorylates Armadillo/β-catenin that is localized in cytosolic complexes.
- In the absence of phosphorylation, Armadillo/β-catenin is stabilized, and translocates to the nucleus where it activates transcription.
- A separate pool of Armadillo/β-catenin is present in complexes at the cell surface, but is not a target for the pathway.
- The function of the cancer-causing gene APC is to destabilize β-catenin, and colon cancers caused by mutation in APC have elevated levels of β-catenin.

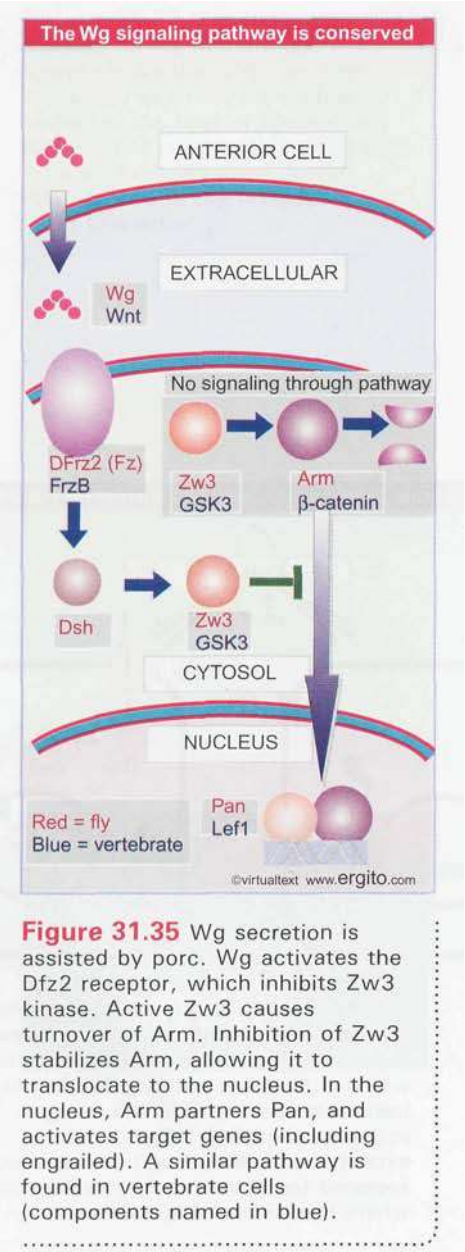


Figure 31.35 Wg secretion is assisted by porc. Wg activates the Dfz2 receptor, which inhibits Zw3 kinase. Active Zw3 causes turnover of Arm. Inhibition of Zw3 stabilizes Arm, allowing it to translocate to the nucleus. In the nucleus, Arm partners Pan, and activates target genes (including engrailed). A similar pathway is found in vertebrate cells (components named in blue).

The interaction between wingless and its receptor activates a signaling pathway. Mutants in other genes that have segment polarity defects similar to *wg* mutants identify the other components of the pathway. They signal positively to execute the pathway. However, mutants in *zw3* have an opposite phenotype; *zw3* functions to block the pathway. **Figure 31.35** shows the results of ordering the genes genetically, and defining the biochemical interactions between their products.

The fly and vertebrate transduction pathways have homologous components. The components in the order of their function in the pathway are as follows:

- Dsh (coded by *Dishevelled)* is a phosphoprotein that responds to the interaction of wingless/Wnt with the frizzled receptor.
- Dsh signals to a Ser/Thr kinase, called Zw3 in Drosophila, but called GSK3 in vertebrates (named for its historical identification as glycogen synthase kinase, but is in fact a homologue of Zw3). Zw3/GSK3 is constitutively active, unless and until it is inactivated by Dsh.
- Zw3/GSK phosphorylates a target called Arm (Armadillo) in *Drosophila* and β-catenin in vertebrate cells.

Arm/β-catenin is the effector of the pathway. Any free Arm/β-catenin is constitutively degraded. Its degradation is triggered by the phosphorylation of its N-terminus by the kinase Zw3/GSK. The phosphoserine in Arm/β-catenin is recognized as a target for addition of the small polypeptide ubiquitin, which causes the protein to be degraded by the proteasome (see *8.31 Ubiquitination targets proteins for degradation)*.

When its degradation is inhibited and Arm/β-catenin accumulates, it translocates to the nucleus. There it binds to a partner, called Pan in Drosophila, and called Tcf/LEF1 in vertebrates, (depending on the system). The complex activates transcription at promoters that are bound by the Pan/Tcf subunit. When Tcf1 binds to DNA, β-catenin can activate transcription at the target promoters.

So wingless/Wnt signaling controls the availability of Arm/β-catenin by causing its degradation to be inhibited. The most surprising feature of this pathway is the nature of the Arm/β-catenin protein. It has two unconnected activities:

- It is a component of a complex that links the cytoskeleton at adhesion complexes. β-catenin binds to cadherin. Mutations *of armadillo* that disrupt the cadherin-binding site show a defect in cell adhesion.
- A separate domain of Arm/β-catenin has a transactivation function when the protein translocates to the nucleus.

By Book_Crazy [IND]

How does Arm/β-catenin participate in two so very different activities? It is in fact bound by a large number of potential partners. Most of them recognize a series of repeats in the central sequence of the protein, with the result that most of these complexes are mutually exclusive. The various complexes are localized in different places in the cell. When Arm/β-catenin binds to cadherins or certain other proteins of the plasma membrane, it forms a complex that participates in cell-cell adhesion. This complex is not a target for the wingless/Wnt pathway, which acts on Arm/β-catenin that is free in the cytosol.

This signaling pathway is also implicated in colon cancer. Mutations in *APC* (adenomatous polyposis coli) are common in colon cancer. APC binds to β-catenin, and its usual effect is to destabilize it. The mutant proteins found in colon cancer allow levels of β-catenin to increase. Mutations in β-catenin that increase its stability have the same effect. One possibility is that APC is the direct target for GSK3, and that its phosphorylation causes it to destabilize β-catenin. However, it is not yet clear whether APC is required for morphogenetic pathways in flies and vertebrates.

31.20 Complex loci are extremely large and involved in regulation

Key Concepts

- Complex loci were identified by interallelic interactions that did not fit the usual complementation behavior.
- They are extremely large and may include multiple protein-coding units as well as *cis*-acting regulatory sites.
- The order of mutations from upstream to downstream corresponds to the order of the body parts that are affected from anterior to posterior.
- *ANT-C* includes several genes that affect head segments, whereas *BX-C* has only three genes and many regulatory sites.
- A segment is distinguished from the preceding (anterior-side) segment by expression of an additional protein-coding unit.
- Loss of a protein-coding unit causes a homeotic transformation in which a segment has the identity of the segment on its anterior-side.

Segment polarity genes control the anterior-posterior pattern within each segment. Homeotic genes impose the program that determines the unique differentiation of each segment. Most homeotic genes are expressed in a spatially restricted manner that corresponds to parasegments.

Homeotic genes interact in complicated interlocking patterns. Many homeotic genes code for transcription factors that act upon other homeotic genes as well as upon other target loci. As a result, a mutation in one homeotic gene influences the expression of other homeotic genes. The consequence is that *the final appearance of a mutant depends not only on the loss of one homeotic gene function, but also on how other homeotic genes change their spatial patterns in response to the loss.*

Homeotic genes act during embryogenesis. Their expression depends on the prior expression of the segmentation genes; we might regard the homeotic genes as integrating the pattern of signals established by the segmentation genes. Homeotic mutants "transform" part of a segment or an entire segment into another type of segment; they may cause one segment of the abdomen to develop as another, legs to develop in place of antennae, or wings to develop in place of eyes. Note

that homeotic genes do not *create* patterns *de novo;* they modify cell fates that are determined by genes such as the segment polarity genes, by switching the set of genes that functions in a particular place. Indeed, the segment polarity genes are active at about the same time as the peak of expression of the homeotic genes.

The genetic properties of some homeotic mutations are unusual and led to the identification of complex loci. A conventional gene—even an interrupted one—is identified at the level of the genetic map by a cluster of noncomplementing mutations. In the case of a large gene, the mutations might map into individual clusters corresponding to the exons. A hallmark of a complex locus is that, in addition to rather well-spaced groups of mutations, extending over a relatively large map distance, there are complex patterns of complementation, in which some pairwise combinations complement but others do not. The individual mutations may have different and complex morphological effects on the phenotype. These relationships are caused by the existence of an array of regulatory elements. Many of the bizarre results that are obtained in complementation assays turn out to result from mutations in promoters or enhancers that affect expression in one cell type but not another. We now recognize that complex loci do not have any novel features of genetic organization, apart from the fact that they have many regulatory elements that control expression in different parts of the embryo.

Two of the complex loci are involved in regulating development of the adult insect body. The *ANT-C* and *BX-C* complex loci together provide a continuum of functions that specify the identities of all of the segmented units of the fly. Each of these complexes contains several homeotic genes. The two separate complexes may have evolved from a split in a single ancestral complex, as suggested by the evolution of the corresponding genes in other species. In the beetle *Tribolium*, the *ANT-C* and *BX-C* complexes are found together at a single chromosomal location. The individual genes may have been derived from duplications and mutations of an original ancestral gene. And in mammals, there are arrays of related genes whose individual members are related sequentially to the genes of the *ANT-C* and *BX-C* complexes (see *31.22 The homeobox is a common coding motif in homeotic genes).*

The homeotic genes clustered at the *ANT-C* and *BX-C* complexes show a relationship between genetic order and the position in which they are expressed in the body of the fly. *Proceeding from left to right, each homeotic gene in the complex acts upon a more posterior region of the fly. The basic principle is that formation of a compartment requires the gene product(s) expressed in the previous compartment, plus a new function coded by the next gene along the cluster.* So loss-of-function mutations usually cause one compartment to have the phenotype of the corresponding compartment on its anterior side. The individual genes code for a set of transcription factors that have related DNA-binding domains (see next section).

The identities of the most anterior parts of the fly (parasegments 1-4) are specified by *ANT-C,* which contains several homeotic genes, including *labial (lab), proboscipedia (pb), Deformed (Dfd), Sex combs reduced (Scr),* and *Antennapedia (Antp).* The homeotic genes lie in a cluster over a region of ~350 kb, but several other genes are interspersed; most of these genes are regulators that function at different stages of development.

Figure 31.36 correlates the organization of *ANT-C* with its effects upon body parts. Adjacent genes are expressed in successively more posterior parts of the embryo, ranging from the leftmost gene *labial* (the most anterior acting, which affects the head) to the rightmost gene *Antp* (the most posterior acting, which affects segments T2-T3).

The *Antp* gene gave its name to the complex, and among the mutations in it are alleles that change antennae into second legs, or second and third legs into first legs. *Antp* usually functions in the thorax; it is needed

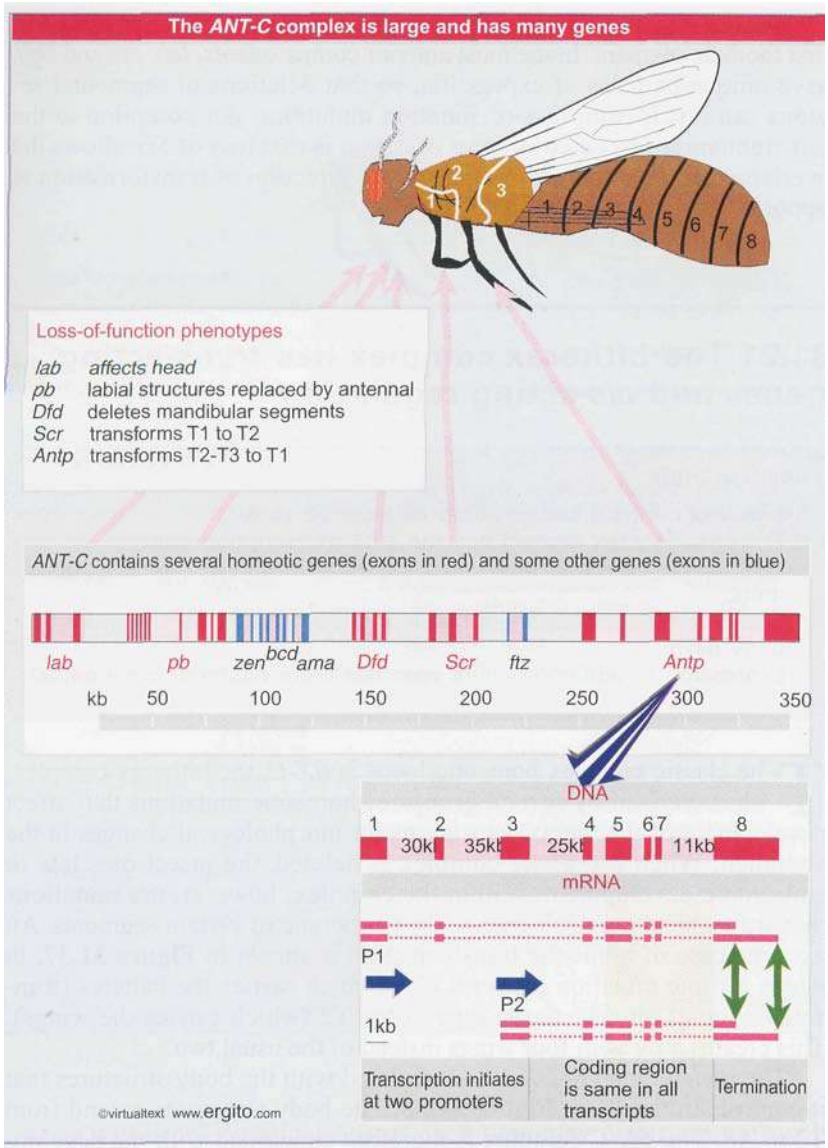


Figure 31.36 The homeotic genes of the *ANT-C* complex confer identity on the most anterior segments of the fly. The genes vary in size, and are interspersed with other genes. The *antp* gene is very large and has alternative forms of expression.

both to promote formation of segments T2-T3 and to suppress formation of head structures. Loss of function therefore causes T2-T3 to resemble the more anterior structure of T1; gain of function, for example, by over-expression in the head, causes the anterior region to develop structures of the thorax. (The molecular action of *Antp* is to prevent the action of genes *hth* and *exd* that promote formation of antennal structures. Hth causes exd to be imported into the nucleus, where it switches on the genes that make the antenna.)

Figure 31.36 summarizes the organization of the gene. It has 8 exons, separated by very large introns, and altogether spanning ~103 kb. The single open reading frame begins only in exon 5, and apparently gives rise to a protein of 43 kD. The discrepancy between the length of the locus and the size of the protein means that only 1% of its DNA codes for protein.

Transcription starts at either of two promoters, located ~70 kb apart! One promoter is located upstream of exon 1, the other upstream of exon 3. Use of the first promoter is associated with omission of exon 3. The transcripts generated from either promoter end either within or after exon 8. All the transcripts appear to code for the same protein. Each promoter has its own tissue-specific expression pattern. We do not know if there is any significance to the difference in the structure of the two types of transcript.

By Book_Crazy [IND]

The other genes of the *ANT-C* complex are expressed in the head and first thoracic segment. In the most anterior compartments, *lab*, *pb*, and *Dfd*, have unique patterns of expression, so that deletions of segmental regions can result from loss-of-function mutations. An exception to the left-right/anterior-posterior order of action is that loss of *Scr* allows the overlapping *Antp* to function, that is, the direction of transformation is opposite from usual.

31.21 The *bithorax* complex has trans-acting genes and c/s-acting regulators

Key Concepts

- *bithorax* controls body structures from T2 to A8.
- The ultrabithorax domain has the *Ubx* transcription unit.
- The infraabdominal domain has the *AbdA* and *AbdB* transcription units.
- The order of units on the genetic map coincides with the order of body parts.
- Expression of additional units specifies more posterior body parts.

The classic complex homeotic locus is *BX-C*, the bithorax complex, characterized by several groups of homeotic mutations that affect development of the thorax, causing major morphological changes in the abdomen. When the whole complex is deleted, the insect dies late in embryonic development. Within the complex, however, are mutations that are viable, but which change the phenotype of certain segments. An extreme case of homeotic transformation is shown in **Figure 31.37,** in which a triple mutation converts T3A (which carries the halteres [truncated wings]) into the tissue type of the T2 (which carries the wings). This creates a fly with four wings instead of the usual two.

The genetic map of *BX-C* is correlated with the body structures that it controls in the fly in **Figure 31.38.** The body structures extend from T2 to A8. The *BX-C* complex is therefore concerned with the development of the major part of the body of the fly. Like *ANT-C*, a crucial feature of this complex is also that mutations affecting particular segments lie in the same order on the genetic map as the corresponding segments in the body of the fly. Proceeding from left to right along the genetic map, mutations affect segments in the fly that become successively more posterior.

A difference between *ANT-C* and *BX-C* is that *ANT-C* functions largely or exclusively via its protein-coding loci, but *BX-C* displays a complex pattern of c/s-acting interactions in addition to the effects of mutations in protein-coding regions. The *BX-C* occupies 315 kb, of which only 1.4% codes for protein. The individual mutations fall into two classes:

- Three transcription units *(Ubx, abdA, AbdB)* produce mRNAs that code for proteins. The transcription units are large (>75 kb for *Ubx*, and >20 kb each for *abdA* and *AbdB).* Each contains several large introns. (The *bxd* and *iab4* regions produce RNAs that do not code for proteins; again, the transcription units are large, and the RNA products are spliced. Their functions are unknown.)
- There are c/s-acting mutations at intervals throughout the entire cluster. They control expression of the transcription units. *Cis*-acting mutations of any particular type may occur in a large region. The locations shown on the map are only approximate, and the bound-

A *bithorax* fly has 4 wings

Figure 31.37 A four-winged fly is produced by a triple mutation in *abx, bx,* and *pbx* at the *BX-C* complex. Photograph kindly provided by Ed Lewis.

By Book_Crazy [IND]

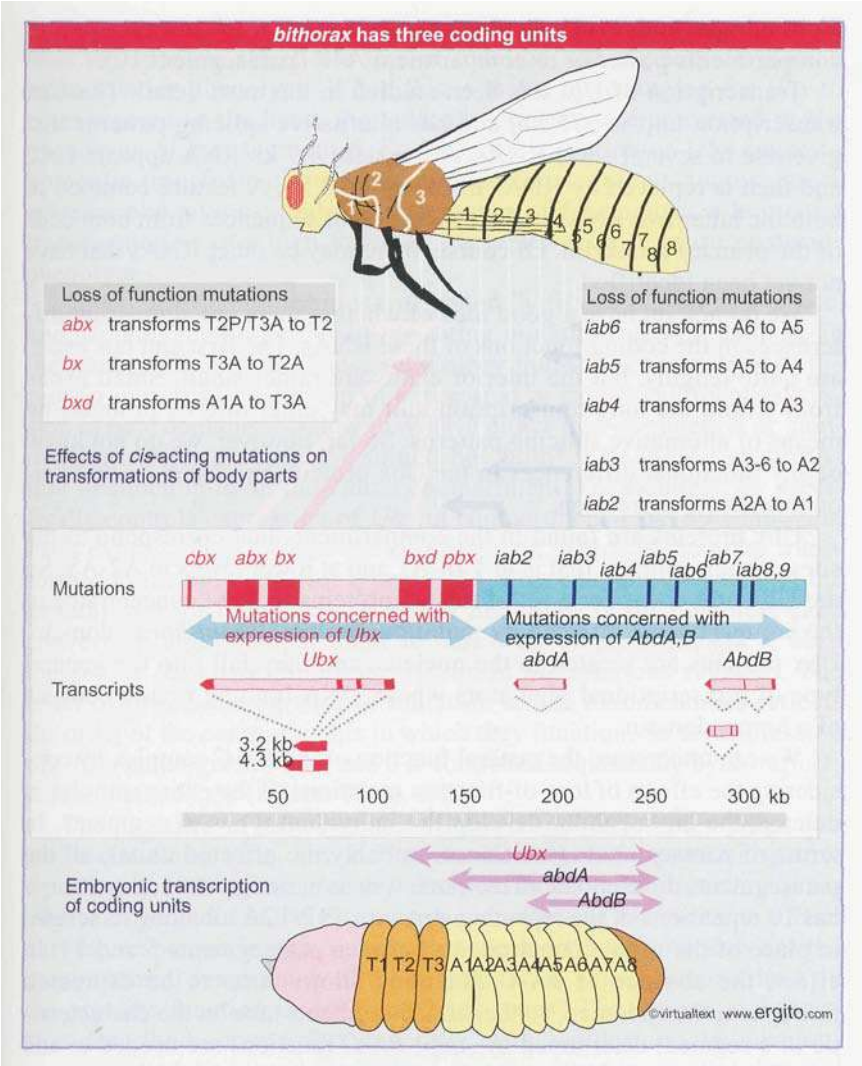


Figure 31.38 The bithorax *(BX-C)* locus has 3 coding units. A series of regulatory mutations affects successive segments of the fly. The sites of the regulatory mutations show the regions within which deletions, insertions, and translocations confer a given phenotype.

aries within which mutations of each type may occur are not well defined.

As a historical note, the complex was originally defined in terms of two "domains." Mutations in the *Ultrabithorax* domain were characterized first; they have the thoracic segments T2P-T3P and the abdominal compartment A1A as their targets (this corresponds to parasegments 5-6). These mutations lie either in the *Ubx* transcription unit or in the *cis*-acting sites that control it. The mutations within the ultrabithorax domain are named for their phenotypes. The *bx* and *bxd* types are identified by a series of mutations, in each case dispersed over ~10 kb. The *abx* and *pbx* mutations are caused by deletions, which vary from 1-10 kb.

Mutations in the *Infraabdominal* domain were found later; they have the abdominal segments A1P-A8P (parasegments 7-14) as targets. These mutations lie either in the *AbdA,B* transcription units, or in the *cis*-acting sites that control them. Within the infraabdominal domain, *cis*-acting mutations are named systematically as *iab2-9*. These mutations affect individual compartments, or sometimes adjacent sets of compartments, as shown at the top of the figure.

Proceeding from left to right along the cluster, transcripts are found in increasingly posterior parts of the embryo, as shown at the bottom of the figure. The patterns overlap. *Ubx* has an anterior boundary of expression in compartment T2P (parasegment 5), *abdA* is expressed

from compartment A1P (parasegment 7), and *AbdB* is expressed in compartments posterior to compartment A4P (parasegment 10).

Transcription of *Ubx* has been studied in the most detail. The *Ubx* transcription unit is ~75 kb, and has alternative splicing patterns that give rise to several short RNAs. A transient 4.7 kb RNA appears first, and then is replaced by RNAs of 3.2 and 4.3 kb. A feature common to both the latter two RNAs is their inclusion of sequences from both ends of the primary transcript. Of course, there may be other RNAs that have not yet been identified.

We do not yet have a good idea of whether there are significant differences in the coding functions of these RNAs. The first and last exons are quite lengthy, but the interior exons are rather small. Small exons from within the long transcription unit may enter mRNA products by means of alternative splicing patterns. So far, however, we do not know of any functional differences in the Ubx proteins produced by the various modes of expression.

Ubx proteins are found in the compartments that correspond to the sites of transcription, that is in T2P-A1 and at lower levels in A2-A8. So the Ubx unit codes for a set of related proteins that are concentrated in the compartments affected by mutations in the *Ultrabithorax* domain. Ubx proteins are located in the nucleus, and they fall into the general type of transcriptional regulators whose DNA-binding region consists of a homeodomain.

We can understand the general function of the *BX-C* complex by considering the effects of loss-of-function mutations. If the entire complex is deleted, the larva cannot develop the individual types of segments. In terms of parasegments (which are probably the affected units), all the parasegments differentiate in the same way as parasegment 4; the embryo has 10 repetitions of the repeating structure T1P/T2A all along its length, in place of the usual compartments between parasegments 5 and 14. In effect, the absence of *BX-C* functions allows *Antp* to be expressed throughout the abdomen, so that all the segments take on the characteristic of a segment determined by *Antp; BX-C* functions are needed to add more posterior-type information.

Each of the transcription units affects successive segments, according to its pattern of expression. So if *Ubx* alone is present, the larva has parasegment 4 (T1P/T2A), parasegment 5 (T2P/T3A), and then 8 copies of parasegment 6 (T3P/A1A). This suggests that the expression of *Ubx* is needed for the compartments anterior to A1A. *Ubx* is also expressed in the more posterior segments, but in the wild type, *abdA* and *AbdB* are also present. If they are removed, the expression of *Ubx* alone in all the posterior segments has the same effect that it usually has in parasegment 6 (T3P/A1A).

The addition of *abdA* to *Ubx* adds the wild-type pattern to parasegments 7, 8, and 9. In other words, *Ubx* plus *abdA* can specify up to compartments A3P/A4A, and in the absence of *AbdB*, this continues to be the default pattern for all the more posterior compartments. The addition of *AbdB* is needed to specify parasegments 10-14.

The general model for the function of the *ANT-C* and *BX-C* complexes is to suppose that additional functions are added to define successive segments proceeding in the posterior direction. It functions by reliance on a *combinatorial pattern* in which the addition of successive gene products confers new specificities. This explains the rule that a loss-of-function mutation in one of the genes of the *ANT-C/BX-C* complexes generally allows the gene on the more anterior side of the mutated gene to determine phenotype, that is, loss-of-function results in homeotic transformation of posterior regions into more anterior phenotypes.

Expression of *Ubx* in a more anterior segment than usual should have the opposite effect to a loss-of-function: the segment develops a more anterior phenotype. When this is tested by arranging for *Ubx* to be expressed in the head, the anterior segments are converted to the phenotype of parasegment 6. So lack of expression of *Ubx* causes a homeotic transformation in which posterior segments acquire more anterior phenotypes; and overexpression of *Ubx* causes a homeotic transformation in which anterior segments acquire more posterior phenotypes.

This type of relationship is true generally for the cluster as a whole, and explains the properties of *cis*-acting mutations as well as those in the transcription units. These regulatory mutations cause loss of the protein in part of its domain of expression or cause additional expression in new domains. So they may have either loss-of-function or gain-of-function phenotypes (or sometimes both). The most common is loss-of-function in an individual compartment. For example, *bx* specifically controls expression of *Ubx* in compartment T3A; a *bx* mutation loses expression of *Ubx* in that compartment, which is therefore transformed to the more anterior type of T2A. This example is typical of the general rule for individual *cis*-acting mutations in the complex; each converts a target compartment *so that it develops as though it were located at the corresponding position in the previous segment.* The order of the *cis*-acting sites of mutation on the chromosomes reflects the order of the compartments in which they function. So the expression of *Ubx* in parasegments 4, 5, and 6 is controlled sequentially by *abx* (affects parasegment 5), *bx* (affects T3A), etc.

The presence of only 3 genes within the *BX-C* complex poses two major questions. First, how do the combinations of 3 proteins specify the identity of 10 parasegments? One possibility is that there are quantitative differences in the various regions, allowing for the same sort of varying responses in target genes that we described previously for the combinatorial functioning of the segmentation genes. Second, how do the proteins function in different tissue types? The pattern of expression described above refers generally to the epidermis; the development of other tissues is controlled in a way that is parallel, but not identical. For example, although *Ubx* is expressed in all posterior segments up to A8 in the epidermis, in mesoderm, it is repressed posterior of segment A7. The posterior boundary reflects repression by *abdA,* since in *abdA* mutants, *Ubx* expression extends posterior in the mesoderm.

Why are loci involved in regulating development of the adult insect from the embryonic larva different from genes coding for the everyday proteins of the organism? Is their enormous length necessary to generate the alternative products? Could it be connected with some timing mechanism, determined by how long it takes to transcribe the unit? At a typical rate of transcription, it would take ~100 minutes to transcribe *Antp,* which is a significant proportion of the 22 hour duration of *D. melanogaster* embryogenesis.

Proceeding from anterior to posterior along the embryo, we encounter the changing patterns of expression of the genes of the *ANT-C* and *BC-C* loci. What controls their transcription? As in the case of the segmentation loci, the homeotic loci are controlled *partially by the genes that were expressed at the previous stage of development, and partially by interactions among themselves.* For example, the expression of *Ubx* is changed by mutations in *bicoid, hunchback,* or *Kruppel.* The anterior boundary of expression respects the parasegment border defined by *ftz* and *eve.* The general principle is that all of these regulatory genes function by controlling transcription, either by activating it or by repressing it, and that the gene products may exert specific effects by both qualitative and quantitative combinations.

31.22 The homeobox is a common coding motif in homeotic genes

Key Concepts

- The homeobox codes for a 60 amino acid protein domain that is a DNA-binding motif.
- Many Drosophila homeotic and segmentation genes code for transcription factors that use a homeodomain to bind DNA.
- The homeodomain does not fully specify the target DNA site, which is influenced by protein-protein combinatorial interactions.
- Homeodomains are found in important regulators of development in a wide range of organisms.
- A vertebrate Hox cluster contains several genes that have homeoboxes, related to the genes in the ANT-C and BX-C fly loci.
- The Hox genes play roles in vertebrate development that are analogous to the roles of the fly genes.

The three groups of genes that control *D. melanogaster* development—maternal genes, segmentation genes, and homeotic genes—regulate one another and (presumably) target genes that code for structural proteins. Interactions between the regulator genes have been defined by analyses that show defects in expression of one gene in mutants of another. However, we have identified rather few of the structural targets on which these groups of genes act to cause differentiation of individual body parts.

Consistent with the idea that the segmentation genes code for proteins that regulate transcription, the genes of 3 gap loci *(hb, Kr,* and *kni)* contain zinc finger motifs. As first identified in the transcription factors TFIIIA and Sp1 (see Figure 22.13), these motifs are responsible for making contacts with DNA. The products of other loci in the gap class also have DNA-binding motifs; *giant* encodes a protein with a basic zipper motif, and *tailless* encodes a protein that resembles the steroid receptors. This suggests that the general function of gap genes is to function as transcriptional regulators.

Conserved motifs are found in many of the homeotic and segmentation genes. The most common of the conserved motifs is the homeobox, a 180 bp region located near the 3' end in several segmentation and homeotic genes. There are ~40 genes in *Drosophila* that contain a homeobox, and almost all are known to be involved in developmental regulation. (The homeobox was first identified by its predominance in the homeotic genes, from which it took its name.) The protein sequence coded by the homeobox is called the homeodomain; it is a DNA-binding motif in transcription factors (see Figure 22.24 in *22.14 Homeodomains bind related targets in DNA).*

The fly homeodomains fall into several groups. A major group in *Drosophila* consists of the homeotic genes in the *BX-C/ANT-C* complexes; they are called the *Antennapedia* group. Their homeodomains are 70-80% conserved, and usually occur at the C-terminal end of the protein (see Figure 22.22). A distinct homeodomain sequence is found in the related genes *engrailed* and *invected;* it has only 45% sequence conservation with the *Antennapedia* group (see Figure 22.23). Other types of homeodomain sequences are represented in 2-4 genes each.

Many of the *Drosophila* genes that contain homeoboxes are organized into clusters. Three of the homeotic genes in the *BX-C* cluster have homeoboxes, the *ANT-C* complex contains a group of 5 homeotic genes with homeoboxes, and 4 other genes at *ANT-C* also contain homeoboxes. The homeotic genes at *BX-C* and *ANT-C* are sometimes described under the general heading of *HOM-C* genes.

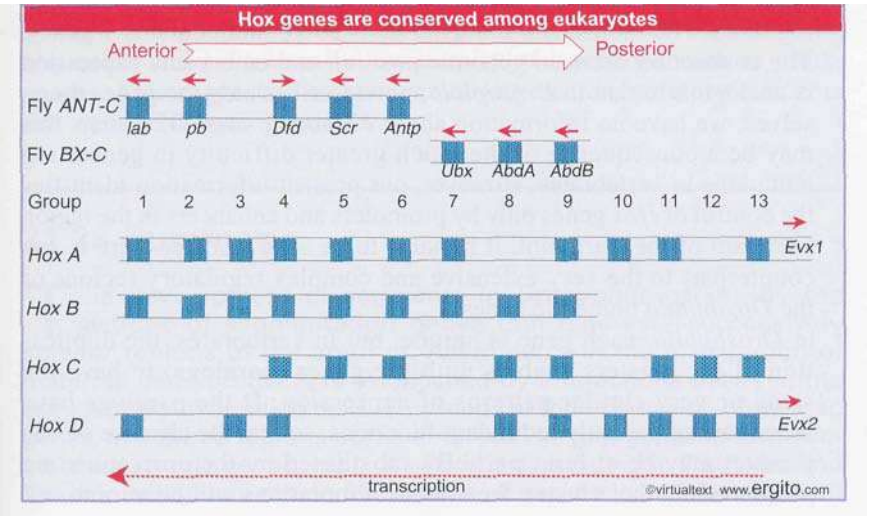
What is the basic function of the *HOM-C* genes in determining identity on the anterior-posterior axis? We assume that homeodomains with different amino acid sequences recognize different target sequences in DNA. Experiments in which regions have been swapped between different proteins suggest that a major part of the specificity of these proteins rests with the homeodomain. However, the ability to bind to a particular DNA target site may not account entirely for their properties. For example, some of these proteins can either activate or repress transcription in response to the context, that is, their actions depend on the set of other proteins that are bound, not just on recognition of the DNA-binding site.

The similarities between the homeodomains of the more closely related members of the group suggest that they could recognize overlapping patterns of target sites. This would open the way for combinatorial effects that could be based on quantitative as well as qualitative differences, that is, there could be competition between proteins with related homeodomains for the same sites. In some cases, different homeoproteins recognize the same target sites on DNA, which poses a puzzle with regard to defining their specificity of action; we assume that there are subtle differences in DNA-binding yet to be discovered, or there are other interactions, such as protein-protein interactions, that play a role.

The homeobox motif is extensively represented in evolution. A striking extension of the significance of homeoboxes is provided by the discovery that a DNA probe representing the homeobox hybridizes with the genomes of many eukaryotes. Genes containing homeoboxes have been characterized in detail in frog, mouse, and human DNA. The frog and mammalian genes are expressed in early embryogenesis, which strengthens the parallel with the fly genes, and suggests the possibility that genes containing homeoboxes are involved in regulation of embryogenesis in a variety of species.

Genes in mammals (and possibly all animals) that are related to the *HOM-C* group have a striking property: like those of the *BX-C/ANT-C* complexes, they are organized in clusters. The individual mammalian genes are called *Hox* genes. A cluster of *Hox* genes may extend 20-100 kb and contain up to 10 genes. Four *Hox* clusters of genes containing homeoboxes have been characterized in the mouse and human genomes. Their organization is compared with the two large fly clusters in **Figure 31.39.**

By comparing the sequences of the homeoboxes (and sometimes other short regions), the mammalian genes can be placed into groups that correspond with the fly genes. This is shown by vertical alignment in the figure. For example, *HoxA4* and *HoxB4* are best related to *Dfd*.



By Book_Crazy [IND]

Figure 31.39 Mouse and human genomes each contain 4 clusters of genes that have homeoboxes. The order of genes reflects the regions in which they are expressed on the anterior-posterior axis. The Hox genes are aligned with the fly genes according to homology, which is strong for groups 1, 2, 4, and 9. The genes are named according to the group and the cluster, e.g., HoxA1 is the most anterior gene in the HoxA group. All Hox genes are present in both man and mice except for some mouse genes missing from cluster C.

When these relationships are defined for the cluster as a whole, it appears that within each cluster we can recognize a series of genes that are related to the genes in the *ANT-C* and *BX-C* clusters. Groups 1-9 in the mammalian loci are defined as corresponding to the genes of the *ANT-C* and *BX-C* loci organized end to end in anterior-posterior orientation. Groups 10-13 appear to have arisen by tandem duplications and divergence of group 9 (the *AbdB* homologue). The corresponding loci in each cluster are sometimes called paralogs (for example, *HoxA4* and *HoxB4* are paralogous).

This situation could have arisen if the fly and mammalian loci diverged at a point when there was only a single complex, containing all of the genes that define anterior-posterior polarity. The organism *Amphioxus,* which corresponds to a line of evolution parallel to the vertebrates, has a single *Hox* cluster containing one member of each paralogous group; this appears to be a direct representative of the original cluster. During evolution, the *Drosophila* genes broke into two separate clusters, while the entire group of mammalian genes became duplicated, some individual members being lost from each complex after the duplication.

The parallel between the mouse and fly genes extends to their pattern of spatial expression. The genes within a *Hox* cluster are expressed in the embryo in a manner that matches their organization in the genome. Progressing from the left toward the right end of the cluster drawn in Figure 31.39, genes are expressed in the embryo in locations progressively more restricted to the posterior end. The patterns of expression for fly and mouse are compared schematically in **Figure 31.40.** The domain of expression extends strongly to the posterior boundary shown in the figure, and then tails off into more posterior segments.

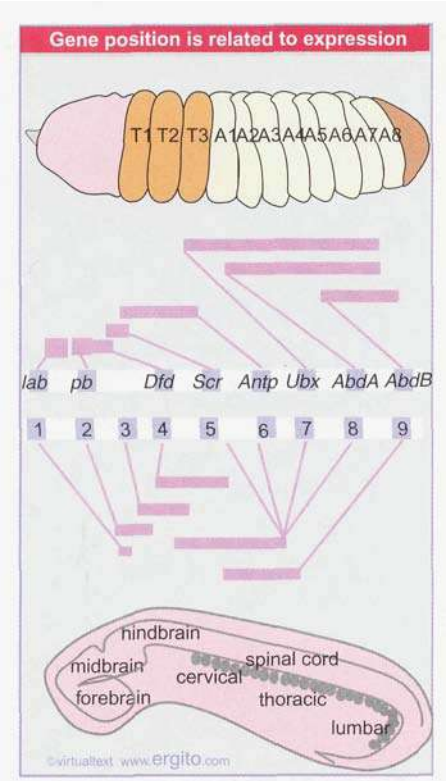


Figure 31.40 A comparison of *ANT-C/BX-C* and HoxB expression patterns shows that the individual gene products share a progressive localization of expression towards the more posterior of the animal proceeding along the gene cluster from left to right. Expression patterns show the regions of transcription in the fly epidermis at 10 hours, and in the central nervous system of the mouse embryo at 12 days.

These results raise the extraordinary possibility that the clusters of genes not only share a common evolution, but also have maintained a common general function in which genome organization is related to spatial expression in fly and mouse, and there is some correspondence between the homologous genes. The idea of such a relationship is strengthened by the observation that ectopic expression of mouse *HoxD4* or *HoxB6* in *Drosophila* cause homeotic transformations virtually identical to those caused by homeotic expression of *Dfd* or *Antp*, respectively! Since the homology between these mouse and fly proteins rests almost exclusively with their homeodomains, this reinforces the view that these domains determine specificity.

There are some differences in the apparent behavior of the vertebrate Hox clusters and the *ANT-C/BX-C* fly clusters:

- The *Hox* genes are small, and there is a greater number of protein-coding units. The mouse *HoxB* cluster is ~120 kb and contains 9 genes. The connection between genomic position and embryonic expression is analogous to that in *Drosophila,* but describes only the genes themselves; we have no information about *cis*-acting sites. Of course, this may be a consequence of the much greater difficulty in generating mutations in vertebrates. However, our present information identifies the control of *Hox* genes only by promoters and enhancers in the region upstream of the startpoint. It remains to be seen whether there is any counterpart to the very extensive and complex regulatory regions of the *Drosophila* homeotic genes.
- In *Drosophila,* each gene is unique; but in vertebrates, the duplication of the clusters enables multiple genes (paralogs) to have the same or very similar patterns of expression. If the paralogs have redundant or partially redundant functions, so that the absence of one product may be at least partially substituted by the corresponding protein of another cluster, the effects of mutations will be minimized.

Disruptions of *Hox* genes in mice often generate recessive lethals. In the examples of *HoxA1* and *HoxA3* various structures of the head and thorax are absent. Not all of the structures that usually express the mutant genes are missing, suggesting that there is indeed some functional redundancy, that is, other *Hox* genes of group 1 or group 3 can substitute in some but not all other tissues for the absence of the *HoxA* gene.

Homeotic transformations are less common with mutants in mice than in *Drosophila*, but sometimes occur. Loss *of HoxC8*, for example, causes some skeletal segments to show more anterior phenotypes. It remains to be seen whether this is a general rule.

Ectopic expression of *Hox* genes has been used successfully to demonstrate that gain-of-function can transform the identity of a segment towards the identity usually conferred by the gene. The most common type of effect in *Drosophila* is to transform a segment into a phenotype that is usually more posterior; in effect, the expression of the homeotic gene has added additional information that confers a more posterior identity on the segment. Similar effects are observed in some cases in the mouse. However, the pattern is not completely consistent.

Taken together, these results make it clear that the *Hox* genes resemble their counterparts in *Drosophila* in determining patterning along the anterior-posterior axis. It may be the case that there is a combinatorial code *of Hox* gene expression, or there may be differences in degree of functional redundancy between paralogs, but we cannot yet provide a systematic model for their role in determining pattern.

The most striking feature of organization of the *Hox* loci still defies explanation: why has the organization of the cluster, in which genomic position correlates with embryonic expression, been maintained in evolution? The obvious explanation is that there is some overall control of gene expression to ensure that it proceeds through the cluster, with the result that a gene could be properly expressed *only* when it is within the cluster. But this does not appear to be true, at least for those individual cases in which genes have been removed from the cluster. Analysis of promoter regions suggests that a *Hox* gene may be controlled by a series of promoter or enhancer elements that together ensure its overall pattern of expression. Usually these elements are in the region upstream of the startpoint. For example, *HoxB4* expression can be reconstructed as the sum of the properties of a series of such elements, tested by introducing appropriate constructs to make transgenic mice. But then why should there have been evolutionary pressure to retain genes in an ordered cluster? One possibility is that an enhancer for one gene might be embedded within another gene, in such a way that, even if an individual gene could function when translocated elsewhere, its removal would impede the expression of other gene(s). An indication that there may be something special about the organization of the region is given by the fact that it has an unusually high density of conserved noncoding sequences and an unusually low density of insertions such as transposons. This suggests the existence of largescale regulatory elements that we have not yet identified.

31.23 Summary

The development of segments in *Drosophila* occurs by the actions of segmentation genes that delineate successively smaller regions of the embryo. Asymmetry in the distribution of maternal gene products is established by interactions between the oocyte and surrounding cells. This leads to the expression of the gap genes, in 4 broad regions of the embryo. The gap genes in turn control the pair-rule genes, each of which is distributed in 7 stripes; and the pair-rule genes define the pattern of expression of the segment

polarity genes, which delineate individual compartments. At each stage of expression, the relevant genes are controlled both by the products of genes that were expressed at the previous stage, and by interactions among themselves. The segmentation genes act upon the homeotic genes, which determine the identities of the individual compartments.

Each of the 4 maternal systems consists of a cascade which generates a locally distributed or locally active morphogen. The morphogen either is a transcription factor or causes the activation of a transcription factor. The transcription factor is the last component in each pathway.

The major anterior-posterior axis is determined by two systems: the anterior system establishes a gradient of bicoid from the anterior pole; and the posterior system produces nanos protein in the posterior half of the egg. These systems function to define a gradient of hunchback protein from the anterior end, with broad bands of knirps and giant in the posterior half. The terminal system acts to produce localized events at both termini. The dorsal-ventral system produces a gradient of nuclear localization of dorsal protein on the ventral side, which represses expression of *dpp* and *zen;* this leads to the ventral activation of *twist* and *snail,* and the dorsal-side activation of *dpp* and *zen.*

Each system is initiated by localization of a morphogen in the egg as a result of its interaction with the surrounding cells. For the anterior and posterior systems, this takes the form of localizing an RNA; *bicoid* mRNA is transported into and localized at the anterior end, and *nanos* mRNA is transported to the posterior end. For the dorsal-ventral system, the Toll receptor is located ubiquitously on the oocyte membrane, but the spatzle ligand is activated ventrally and therefore triggers the pathway on the ventral side. The pathway resembles the mammalian IL-1 signal transduction pathway and culminates in the phosphorylation of cactus, which regulates the dorsal transcription factor. On the dorsal side of the embryo, the morphogen Dpp is released; it is a member of the TGFβ family that diffuses to interact with its receptor. A ligand-receptor interaction involving related members of the TGFβ/receptor families is also employed in a comparable role in vertebrate development.

The early embryo consists of a syncytium, in which nuclei are exposed to common cytoplasm. It is this feature that allows all 4 maternal systems to control the function of a nucleus according to the coordinates of its position on the anterior-posterior and dorsal-ventral axes. At cellular blastoderm, zygotic RNAs are transcribed, and the developing embryo becomes dependent upon its own genes. Cells form at the blastoderm stage, after which successive interactions involve a cascade of transcriptional regulators.

Three gap genes are zinc-finger proteins, and one is a basic zipper protein. Their concentrations control expression of the pair-rule genes, which are also transcription factors. In particular, the expression of *eve* and *ftz* controls the boundaries of compartments, functioning in every other segment. The segment polarity genes represent the first step in the developmental cascade that involves functions other than transcription factors. Interactions between the segmentation gene products define unique combinations of gene expression for each segment.

The segment polarity genes include proteins involved in cell-cell interactions as well as transcription factors. The basic circuitry that determines the anterior and posterior polarities of compartments is maintained by an autoregulatory interaction between the cells at the boundary. An anterior compartment secretes wingless protein, which acts upon the cell on the posterior side. This causes engrailed to be expressed in the posterior cell, which in turn causes secretion of hedgehog on the anterior side. Hedgehog causes the anterior cell to express wingless.

Homeotic genes impose the program that determines the unique differentiation of each segment. The complex loci *ANT-C* and *BX-C*

each contain a cluster of functions, whose spatial expression on the anterior-posterior axis reflects genetic position in the cluster. Each cluster contains one exceedingly large transcription unit as well as other, shorter units. Many of the transcription units (including the largest genes, *Ubx* and *Antp*) have patterns of alternative splicing, but no significance has been attributed to this yet. Proceeding from left to right in each cluster, genes are expressed in more posterior tissues. The genes are expressed in overlapping patterns in such a way that addition of a function confers new features of posterior identity; thus loss of a function results in a homeotic transformation from posterior to more anterior phenotype. The genes are controlled in a complex manner by a series of regulatory sites that extend over large regions; mutations in these sites are c/s-acting, and may cause either loss-of-function or gain-of-function. The c/s-acting mutations tend to act on successive segments of the fly, by controlling expression of the homeotic proteins.

The genes of the *ANT-C* and *BX-C* loci, and many segmentation genes (including the maternal gene *bicoid* and most of the pair-rule genes) contain a conserved motif, the homeobox. Homeoboxes are also found in genes of other eukaryotes, including worms, frogs, and mammals. In each case, these genes are expressed during early embryogenesis. In mammals, the *Hox* genes (which specify homeodomains in the *Antennapedia* class) are organized in clusters. There are 4 *Hox* clusters in both man and mouse. These clusters can be aligned with the *ANT-C/BX-C* clusters in such a way as to recognize homologies between genes at corresponding positions. Proceeding towards the right in a *Hox* cluster, a gene is expressed more towards the posterior of the embryo. The Hox genes have roles in conferring identity on segments of the brain and skeleton (and other tissues). The analogous clusters represent regulators of embryogenesis in mammals and flies. Hox clusters may be a characteristic of all animals.

Drosophila genes containing homeoboxes form an intricate regulatory network, in which one gene may activate or repress another. The relationship between the sequence of the homeodomain, the DNA target it recognizes, and the regulatory consequences, remains to be fully elucidated. Specificity in target choice appears to reside largely in the homeodomain; we have yet to explain the abilities of a particular homeoprotein to activate or to repress gene transcription at its various targets. The general principle is that segmentation and homeotic genes act in a transcriptional cascade, in which a series of hierarchical interactions between the regulatory proteins is succeeded by the activation of structural genes coding for body parts.

References

31.2 Fly development uses a cascade of transcription factors

rev Lawrence, P. (1992). The Making of a Fly. Blackwell Scientific, Oxford.

Mahowald, A. P. and Hardy, P. A. (1985). Genetics of *Drosophila* embryogenesis. Ann. Rev. Genet. 19, 149-177.

31.4 Maternal gene products establish gradients in early embryogenesis

rev Ingham, P. W. (1988). The molecular genetics of embryonic pattern formation in *Drosophila.* Nature 335, 25-34.

Lawrence, P. (1992). The Making of a Fly. Blackwell Scientific, Oxford.

Mahowald, A. P. and Hardy, P. A. (1985). Genetics of *Drosophila* embryogenesis. Ann. Rev. Genet. 19, 149-177.

31.5 Anterior development uses localized gene regulators

rev Lawrence, P. A. and Struhl, G. (1996). Morphogens, compartments, and pattern: lessons from *Drosophila?* Cell 85, 951-961.

McGinnis, W. and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell 68, 283-302.

ref Driever, W. and Nusslein-Volhard, C. (1988). The bicoid protein determines position in the *Drosophila* embryo in a concentration dependent manner. Cell 54, 95-104.

31.6 Posterior development uses another localized regulator

ref Sprenger, F. and Nusslein-Volhard, C. (1992). Cellular terminal regions of the *Drosophila* egg. Cell 71, 987-1001.

Struhl, G., Johnston, P., and Lawrence, P. A. (1992). Control of *Drosophila* body pattern by the hunchback morphogen gradient. Cell 69, 237-249.

31.7 How are mRNAs and proteins transported and localized?

rev Johnstone, 0. and Lasko, P. (2001). Translational regulation and RNA localization in *Drosophila oocytes and* embryos. Ann. Rev. Genet. 35, 365-406.

Palacios, I. M. and Johnston, D. S. (2001). Getting the message across: the intracellular localization of mRNAs in higher eukaryotes. Ann. Rev. Cell Dev. Biol. 17, 569-614.

ref Bergsten, S. E. and Gavis, E. R. (1999). Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. Development 126, 659-669.

Bullock, S. L. and Ish-Horowicz, D. (2001). Conserved signals and machinery for RNA transport in *Drosophila* oogenesis and embryogenesis. Nature 414, 611-616.

Crucs, S., Chatterjee, S., and Gavis, E. R. (2000). Overlapping but distinct RNA elements control repression and activation of nanos translation. Mol Cell 5, 457-467.

Dahanukar, A. and Wharton, R. (1966). The nanos gradient in *Drosophila* embryos is generated by translation regulation. Genes Dev. 10, 2610-2620.

Driever, W. and Nusslein-Volhard, C. (1988). A gradient of bicoid protein in *Drosophila* embryos. Cell 54, 83-93.

Driever, W. and Nusslein-Volhard, C. (1988). The bicoid protein determines position in the *Drosophila* embryo in a concentration dependent manner. Cell 54, 95-104.

Gavis, E. R., Curtis, D., and Lehmann, R. (1996). Identification of *cis*-acting sequences that control *nanos* localization. Dev. Biol. 176, 36-50.

Gavis, E. R., Curtis, D., and Lehmann, R. (1996). A conserved 90 nucleotide element mediates translational repression of *nanos* RNA. Development 122, 2791-2800.

Kim-Ha, J., Kerr, K., and MacDonald, P. M. (1995). Translational regulation of *oskar* mRNA by bruno, an ovarian RNA-binding protein, is essential. Cell 81, 403-412.

Pokrywka, N. J. and Stephenson, E. C. (1995). Microtubules are a general component of mRNA localization systems in *Drosophila* oocytes. Dev. Biol. 167, 363-370.

Theurkauf, W. E. and Hazelrigg, T. I. (1998). *In vivo* analyses of cytoplasmic transport and cytoskeletal organization during *Drosophila* oogenesis: characterization of a multi-step anterior localization pathway. Development 125, 3655-3666.

Theurkauf, W. E., Alberts, B. M., Jan, Y. N., and Jongens, T. A. (1993). A central role for microtubules in the differentiation of *Drosophila* oocytes. Development 118, 1169-1180.

Wilhelm, J. E. et al. (2000). Isolation of a ribonucleoprotein complex involved in mRNA localization in *Drosophila* oocytes. J. Cell Biol. 148, 427-439.

Wilkie, G. S. and Davis, I. (2001). Drosophila wingless and pair-rule transcripts localize apically by dynein-mediated transport of RNA particles. Cell 105, 209-219.

31.8 How are gradients propagated?

ref Entchev, E. V., Schwabedissen, A., and Gonzalez-Gaitan, M. (2000). Gradient formation of the TGF-beta homologue Dpp. Cell 103, 981-991.

Gurdon, J. B., Harger, P., Mitchell, A., and Lemaire, P. (1994). Activin signalling and response to a morphogen gradient. Nature 371, 487-492.

Nellen, D., Burke, R., Struhl, G., and Basler, K. (1996). Direct and long-range action of a DPP morphogen gradient. Cell 85, 357-368.

31.9 *Dorsal-ventral development uses localized* receptor-ligand interactions

rev Wylie, A. A., Murphy, S. K., Orton, T. C, and Jirtle, R. L. (1996). Intercellular signaling and the polarization of body axes during *Drosophila* oogenesis. Genes Dev. 10, 1711-1723.

ref Price, J. V., Clifford, R. J., and Schupbach, T. (1989). The maternal ventralizing locus *torpedo* is allelic to *faint little ball,* an embryonic lethal, and encodes the *Drosophila* EGF receptor homolog. Cell 56, 1085-1092.

Schejter, E. D. and Shilo, B. Z. (1989). The *Drosophila* EGF receptor homologue (*DER*) gene is allelic to *faint little ball,* a locus essential for embryonic development. Cell 56, 1093-1104.

Schupbach, T. (1987). Germ line and soma cooperate during oogenesis to establish the Dorsal-ventral pattern of egg shell and embryo in *D. melanogaster.* Cell 49, 699-707.

31.10 Ventral development proceeds through Toll

rev Morisato, D. and Anderson, K. V. (1995). Signaling pathways that establish the dorsal-ventral pattern of the *Drosophila* embryo. Ann. Rev. Genet. 29, 371-399.

ref Anderson, K. V., Bokla, L., and Nusslein-Volhard, C. (1985). Establishment of dorsal-ventral polarity in the *Drosophila* embryo: the induction of polarity by the Toll gene product. Cell 42, 791-798.

Hashimoto, C, Hudson, K. L, and Anderson, K. V. (1988). The *Toll* gene of *Drosophila,* required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. Cell 52, 269-279.

Morisato, D. and Anderson, K. V. (1994). The *spatzle* gene encodes a component of the extracellular signaling pathway establishing the dorsal-ventral pattern of the *Drosophila* embryo. Cell 76, 677-688.

Sen, J., Goltz, J. S., Stevens, L., and Stein, D. (1998). Spatially restricted expression of pipe in the *Drosophila* egg chamber defines embryonic dorsal-ventral polarity. Cell 95, 471-481.

Stein, D., Roth, S., Vogelsang, E., Nusslein-Volhard, C, and Vogelsang, C. (1991). The polarity of the dorsoventral axis in the *Drosophila* embryo is defined by an extracellular signal. Cell 65, 725-735.

31.11 Dorsal protein forms a gradient of nuclear localization

rev Belvin, M. P. and Anderson, K. V. (1996). A conserved signaling pathway: the *Drosophila* Toll-Dorsal pathway. Ann. Rev. Cell Dev. Biol. 12, 393-416.

ref Ghosh, S., Gifford, A. M., Riviere, L. R., Tempst, P., Nolan, G. P., and Baltimore, D. (1990). Cloning of the p50 DNA binding subunit of NF-kappa B: homology to rel and dorsal. Cell 62, 1019-1029.

Ip, Y. T., Kraut, R., Levine, M., and Rushlow, C. A. (1991). The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in *Drosophila.* Cell 64, 439-446.

Kidd, S. (1992). Characterization of the *Drosophila* cactus locus and analysis of interactions between cactus and dorsal proteins. Cell 71, 623-635.

Kieran, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., Le Bail, 0., Urban, M. B., Kourilsky, P., Baeuerle, P. A., and Israel, A. (1990). The DNA binding subunit of NF-kappa B is identical to factor KBF1 and homologous to the rel oncogene product. Cell 62, 1007-1018.

Roth, S., Stein, D., and Nusslein-Volhard, C. (1989). A gradient of nuclear localization of the dorsal protein determines Dorsal-ventral pattern in the *Drosophila* embryo. Cell 59, 1189-202.
Rushlow, C. A., Han, K., Manley, J. L., and Levine, M. (1989). The graded distribution of the dorsal morphogen is initiated by selective nuclear transport in *Drosophila.* Cell 59, 1165-1177.

31.13 TGFβ/BMPs are diffusible morphogens

rev De Robertis, E. M. and Sasi, Y. (1996). A common plan for Dorsal-ventral patterning in Bilateria. Nature 380, 37-40.

ref Chang, C, Holtzman, D. A., Chau, S., Chickering, T., Woolf, E. A., Holmgren, L. M., Bodorova, J., Gearing, D. P., Holmes, W. E., and Brivanlou, A. H. (2001). Twisted gastrulation can function as a BMP antagonist. Nature 410, 483-487.
Ferguson, E. L. and Anderson, K. V. (1992). decapentaplegic acts as a morphogen to organize dorsal-ventral pattern in the *Drosophila* embryo. Cell 71, 451-461.

Nellen, D., Burke, R., Struhl, G., and Basler, K. (1996). Direct and long-range action of a DPP morphogen gradient. Cell 85, 357-368.
Nellen, D., Affolter, M., and Basler, K. (1994). Receptor serine/threonine kinases implicated in the control of *Drosophila* body pattern by decapentaplegic. Cell 78, 225-237.
Ross, J. J., Shimmi, J. J., Vilmos, P., Petryk, A., Kim, H., Gaudenz, K., Hermanson, S., Ekker, S. C, O'Connor, M. B., and Marsh, J. L. (2001). Twisted gastrulation is a conserved extracellular BMP antagonist. Nature 410, 479-483.
Ruberte, E., Marty, T., Nellen, D., Affolter, M., and Basler, K. (1995). An absolute requirement for both the type II and type I receptors, punt and thick veins, for dpp signaling in vivo. Cell 80, 889-897.

Scott, I. C, Scott, I. C, Scott, I. C, Pappano, W. N., Maas, S. A., Cho, K. W., and Greenspan, D. S. (2001). Homologues of Twisted gastrulation are extracellular cofactors in antagonism of BMP signalling. Nature 410, 475-478.

31.14 Cell fate is determined by compartments that form by the blastoderm stage

rev Ingham, P. W. and Martinez-Arias, A. (1992). Boundaries and fields in early embryos. Cell 68, 221-235.

Scott, M. P. and Carroll, S. B. (1987). The segmentation and homeotic gene network in early *Drosophila* development. Cell 51, 689-698.

ref Lehmann, R. and Frohnhofer, H. G. (1989). Segmental polarity and identity in the abdomen of *Drosophila* is controlled by the relative position of gap gene expression. Development 107 Suppl, 21-29.

Nusslein-Vollhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila.* Nature 287, 795-801.

31.15 Gap genes are controlled by bicoid and by one another

ref Simpson-Brose, M., Treisman, J., and Desplan, C. (1994). Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila.* Cell 78, 855-865.
Tabata, T., Schwartz, C, Gustavson, E., Ali, Z., and Kornberg, T. B. (1995). Creating a *Drosophila* wing de novo, the role of engrailed, and the compartment border hypothesis. Development 121, 3359-3369.

31.16 Pair-rule genes are regulated by gap genes

ref Lawrence, P. A. and Johnston, P. (1989). Pattern formation in the *Drosophila* embryo: allocation of cells to parasegments by even-skipped and fushi tarazu. Development 105, 761-767.

31.17 Segment polarity genes are controlled by pair-rule genes

ref Kornberg, T. (1981). Engrailed: a gene controlling compartment and segment formation in *Drosophila.* Proc. Nat. Acad. Sci. USA 78, 1095-1099.
Kornberg, T., Siden, I., O'Farrell, P., and Simon, M. (1985). The engrailed locus of *Drosophila: in situ* localization of transcripts reveals compartment-specific expression. Cell 40, 45-53.
Tabata, T., Schwartz, C, Gustavson, E., Ali, Z., and Kornberg, T. B. (1995). Creating a *Drosophila* wing de novo, the role of engrailed, and the compartment border hypothesis. Development 121, 3359-3369.

31.18 Wingless and engrailed expression alternate in adjacent cells

ref Bhanot, P. et al. (1995). A new member of the frizzled family from *Drosophila* functions as a wingless receptor. Nature 382, 225-230.
Muller, H., Samanta, R., and Wieschaus, E. (1999). Wingless signaling in the *Drosophila* embryo: zygotic requirements and the role of the *frizzled* genes. Development 126, 577-586.

Rijsewijk, F. M. (1987). The *Drosophila* homologue of the mouse mammary oncogene int-1 is identical to the segment polarity gene wingless. Cell 50, 649-657.

Wehrli, M., Dougan, S. T., Caldwell, K., O'Keefe, L, Schwartz, S., Vaizel-Ohayon, D., Schejter, E., Tomlinson, A., and DiNardo, S. (2000). arrow encodes an LDL-receptor-related protein essential for Wingless signalling. Nature 407, 527-530.

31.19 The wingless/wnt pathway signals to the nucleus

rev Zhurinsky, J., Shtutman, M., and Ben-Ze'ev, A. (2000). Plakoglobin and beta-catenin: protein interactions, regulation and biological roles. J. Cell Sci. 113, 3127-3139.

ref Brunner, E. (1997). *pangolin* encodes a Lef-1 homologue that acts downstream of Armadillo to transduce the Wingless signal in *Drosophila.* Nature 385, 829-833.
Graham, T. A., Weaver, C, Mao, F., Kimelman, D., and Xu, W. (2000). Crystal structure of a beta-catenin/Tcf complex. Cell 103, 885-896.
Molenaar, M. et al. (1996). XTcf-3 transcription factor mediates β-catenin-induced axis formation in *Xenopus* embryos. Cell 86, 391-399.

31.20 Complex loci are extremely large and involved in regulation

rev Montgomery, G. (2002). E. B. Lewis and the bithorax complex. Part I. Genetics 160, 1265-1272.
Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M., and McGinnis, W. (1985). Homeo box genes of the *Antennapedia* and *bithorax* complexes of *Drosophila.* Cell 43, 71-80.
Scott, M. P. (1987). Complex loci of *Drosophila.* Ann. Rev. Biochem. 56, 195-227.

ref Duncan, I. and Montgomery, G. (2002). E. B. Lewis and the Bithorax Complex. Part ii. from cis-trans test to the genetic control of development. Genetics 161, 1-10.

Scott, M. P. et al. (1983). The molecular organization of the Antennapedia locus of *Drosophila.* Cell 35, 763-766.

31.21 The *bithorax* complex has *trans*-acting genes and *cis*-acting regulators

rev Lewis, E. B. (1985). Regulation of the genes of the bithorax complex in *Drosophila.* Cold Spring Harb Symp Quant Biol 50, 155-164.

ref Beachy, P. A., Helfand, S. L., and Hogness, D. S. (1985). Segmental distribution of bithorax complex proteins during *Drosophila* development. Nature 313, 545-551.

Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., and Lewis, E. B. (1985). The abdominal region of the bithorax complex. Cell 43, 81-96.

Lewis, E. B. (1978). A gene complex controlling segmentation in *Drosophila.* Nature 276, 565-570.

Martin, C. H. et al. (1995). Complete sequence of the bithorax complex of *Drosophila.* Proc. Nat. Acad. Sci. USA 92, 8398-8402.

31.22 The homeobox is a common coding motif in homeotic genes

rev Hunt, P. and Krumlauf, R. (1992). Hox codes and positional specification in vertebrate embryonic axes. Ann. Rev. Cell Biol. 8, 227-256.

Krumlauf, R. (1994). Hox genes in vertebrate development. Cell 78, 191-201.

McGinnis, W. and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell 68, 283-302.

Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M., and McGinnis, W. (1985). Homeo box genes of the *Antennapedia* and *bithorax* complexes of *Drosophila.* Cell 43, 71-80.

Scott, M. P (1989). The structure and function of the homeodomain. Biochim. Biophys. Acta 989, 25-48.

ref Garcia-Fernandez, J. and Holland, P. W. H. (1994). Archetypal organization of the amphioxus Hox gene cluster. Nature 370, 563-566.

Graham, A., Papalopulu, N., and Krumlauf, R. (1989). The murine and *Drosophila* homeobox gene complexes have common features of organization and expression. Cell 57, 367-378.

Malicki, J., Schughart, K., and McGinnis, W. (1990). Mouse hox-22 specifies thoracic segmental identity in *Drosophila* embryos and larvae. Cell 63, 961-967.

McGinnis, W. et al. (1984). A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell 37, 403-408.

Glossary

The **A domain** is the conserved 11 bp sequence of A-T base pairs in the yeast ARS element that comprises the replication origin.

The **A site** of the ribosome is the site that an aminoacyl-tRNA enters to base pair with the codon.

Abortive initiation describes a process in which RNA polymerase starts transcription but terminates before it has left the promoter. It then reinitiates. Several cycles may occur before the elongation stage begins.

The **abundance** of an mRNA is the average number of molecules per cell.

Abundant mRNAs consist of a small number of individual species, each present in a large number of copies per cell.

The **acceptor arm** of tRNA is a short duplex that terminates in the CCA sequence to which an amino acid is linked.

An **acentric fragment** of a chromosome (generated by breakage) lacks a centromere and is lost at cell division.

Acquired immunity is another term for adaptive immunity.

Acridines are mutagens that act on DNA to cause the insertion or deletion of a single base pair. They were useful in defining the triplet nature of the genetic code.

An **activator** is a protein that stimulates the expression of a gene, typically by acting at a promoter to stimulate RNA polymerase. In eukaryotes, the sequence to which it binds in the promoter is called a response element.

An **active site** is the restricted part of an enzyme to which a substrate binds.

Active transport is an energy-consuming process that moves molecules against an electrochemical gradient. Energy for the movement is provided by hydrolysis of ATR

An **acute transforming virus** carries a gene(s) that originated in a cellular genome. Its transforming capacity is the result of expression of that gene. Because the gene replaced viral sequences, the virus does not have the capacity to replicate independently.

Adaptin is a subunit of the cytosolic adaptor proteins that mediate formation of clathrin-coated vesicles. There are several types of adaptin subunits.

Adaptive immunity is the response mediated by lymphocytes that are activated by their specific interaction with antigen. The adaptive immune response develops over several days as lymphocytes with antigen-specific receptors are stimulated to proliferate and become effector cells. It is responsible for immunological memory.

Adaptor proteins bind to signals in the cytoplasmic tails of transmembrane cargo proteins and recruit clathrin molecules in the assembly of clathrin-coated pits and vesicles. Different types of adaptor proteins function at different compartments. Each adaptor protein contains four different subunits.

An **addiction system** is a survival mechanism used by plasmids. The mechanism kills the bacterium upon loss of the plasmid.

Adenylate cyclase is an enzyme that uses ATP as a substrate to generate cyclic AMP, in which 5' and 3' positions of the sugar ring are connected via a phosphate group.

Agropine plasmids carry genes coding for the synthesis of opines of the agropine type. The tumors usually die early.

An **alarmone** is a small molecule in bacteria that is produced as a result of stress and which acts to alter the state of gene expression. The unusual nucleotides ppGpp and pppGpp are examples.

An **allele** is one of several alternative forms of a gene occupying a given locus on a chromosome.

Allelic exclusion describes the expression in any particular lymphocyte of only one allele coding for the expressed immunoglobulin. This is caused by feedback from the first immunoglobulin allele to be expressed that prevents activation of a copy on the other chromosome.

Allosteric regulation describes the ability of a protein to change its conformation (and therefore activity) at one site as the result of binding a small molecule to a second site located elsewhere on the protein.

Alternative splicing describes the production of different RNA products from a single product by changes in the usage of splicing junctions.

The **Alu domain** comprises the parts of the 7S RNA of the SRP that are related to Alu RNA.

The **Alu family** is a set of dispersed, related sequences, each ~300 bp long, in the human genome. The individual members have Alu cleavage sites at each end (hence the name).

Amanitin (more fully α-amanitin) is a bicyclic octapeptide derived from the poisonous mushroom Amanita phalloides; it inhibits transcription by certain eukaryotic RNA polymerases, especially RNA polymerase **II.**

The **amber** codon is the triplet UAG, one of the three termination codons that end protein synthesis.

An **aminoacyl-tRNA** is a tRNA linked to an amino acid. The COOH group of the amino acid is linked to the 3'- or 2'-OH group of the terminal base of the tRNA.

Aminoacyl-tRNA synthetases are enzymes responsible for covalently linking amino acids to the 2'- or 3′-OH position of tRNA.

Amphipathic structures have two surfaces, one hydrophilic and one hydrophobic. Lipids are amphipathic; and some protein regions may form amphipathic helices, with one charged face and one neutral face.

Amplification refers to the production of additional copies of a chromosomal sequence, found as intrachromosomal or extrachromosomal DNA.

The **anaphase promoting complex** (APC) is a set of proteins that triggers proteolysis or targets required to allow chromosomes to separate.

An **anchor** (stop-transfer) (often referred to as a "transmembrane anchor") is a segment of a transmembrane protein which resides in the membrane.

Anchorage dependence describes the need of normal eukaryotic cells for a surface to attach to in order to grow in culture.

An **aneuploid** set of chromosomes differs from the usual diploid constitution by loss or duplication of chromosomes or chromosomal segments.

Annealing of DNA describes the renaturation of a duplex structure from single strands that were obtained by denaturing duplex DNA.

In *Drosophila*, the **anterior system** is one of the maternal systems that establishes the polarity of the oocyte. The set of genes in the anterior system play a role in the proper formation of the head and the thorax.

The **anterior-posterior axis** is the line running from the head to the tail of an animal.

Anterograde transport is the direction of membrane transport specified by the movement of macromolecules through the secretory pathway (from the rough endoplasmic reticulum, through the Golgi complex, and to the plasma membrane). It is also called forward transport.

An anti-insulator is a sequence that allows an enhancer to overcome the effect of an insulator.

Anti-Sm is an autoimmune antiserum that defines the Sm epitope that is common to a group of proteins found in snRNPs that are involved in RNA splicing.

An **antibody** is a protein (immunoglobulin) produced by B lymphocyte cells that recognizes a particular 'foreign antigen', and thus triggers the immune response.

The **anticodon** is a trinucleotide sequence in tRNA which is complementary to the codon in mRNA and enables the tRNA to place the appropriate amino acid in response to the codon.

The **anticodon arm** of tRNA is a stem loop structure that exposes the anticodon triplet at one end.

An **antigen** is any foreign substance whose entry into an organism provokes an immune response by stimulating the synthesis of an antibody (an immunoglobulin protein that can bind to the antigen).

An **antigenic determinant** is the portion of an antigen that is recognized by the antigen receptor on lymphocytes. It is also called an epitope.

Antigenic variation describes the ability of a trypanosome to change its surface protein, so that the host is challenged with a different antigen.

Antiparallel strands of the double helix are organized in opposite orientation, so that the 5' end of one strand is aligned with the 3' end of the other strand.

An **antiporter** is a type of carrier protein that simultaneously moves two different types of solutes in opposite directions across the plasma membrane.

An **antisense gene** codes for an (antisense) RNA that has a complementary sequence to an RNA that is its target.

The **antisense strand** (template strand) of DNA is complementary to the sense strand, and is the one that acts as the template for synthesis of mRNA.

Antitermination is a mechanism of transcriptional control in which termination is prevented at a specific terminator site, allowing RNA polymerase to read into the genes beyond it.

Antitermination proteins allow RNA polymerase to transcribe through certain terminator sites.

Anucleate bacteria lack nuclei, but are of similar shape to wild-type bacteria.

Apoptosis (programmed cell death) is the capacity of a cell to respond to a stimulus by initiating a pathway that leads to its death by a characteristic set of reactions.

An **arm** of tRNA is one of the four (or in some cases five) stem-loop structures that make up the secondary structure.

The arms of a lambda phage attachment site are the sequences flanking the core region where the recombination event occurs.

ARS (autonomous replication sequence) is an origin for replication in yeast. The common feature among different *ARS* sequences is a conserved 11 bp sequence called the A-domain.

An **assembly factor** is a protein that is required for formation of a macromolecular structure but is not itself part of that structure.

att sites are the loci on a phage and the bacterial chromosome at which recombination integrates the phage into, or excises it from, the bacterial chromosome.

Attenuation describes the regulation of bacterial operons by controlling termination of transcription at a site located before the first structural gene.

An **attenuator** is a terminator sequence at which attenuation occurs.

Autogenous control describes the action of a gene product that either inhibits (negative autogenous control) or activates (positive autogenous control) expression of the gene coding for it.

An **autoimmune disease** is a pathological condition in which the immune response is directed to self antigen.

An **autonomous controlling element** in maize is an active transposon with the ability to transpose *{compare with* nonautonomous controlling element).

The ability of a species of kinase to phosphorylate itself is referred to as **autophosphorylation.** Autophosphorylation does not necessarily occur on the same polypeptide chain as the catalytic site; for example, in a dimer, each subunit may phosphorylate the other.

Autosplicing (self-splicing) describes the ability of an intron to excise itself from an RNA by a catalytic action that depends only on the sequence of RNA in the intron.

Avirulent mutants of a bacterium or virus have lost the capacity to infect a host productively, that is, to make more bacterium or virus.

Axes are straight lines passing through an organism, around which the organism is symmetrically arranged.

An **axial element** is a proteinaceous structure around which the chromosomes condense at the start of synapsis.

A **B cell** is a lymphocyte that produces antibodies. B cells development occurs primarily in bone marrow.

B cell memory is responsible for rapid antibody production during a secondary immune response and subsequent responses. Memory B cells produce antibodies of higher affinity than naive B cells.

The **B cell receptor** (BCR) is the antigen receptor complex on the cell surface of B lymphocytes. It consists of membrane-bound immunoglobulin bound noncovalently to Igα and Igβ chains.

B-form DNA is a right-handed double helix with 10 base pairs per complete turn (360°) of the helix. This is the form found under physiological conditions whose structure was proposed by Crick and Watson.

A **back mutation** reverses the effect of a mutation that had inactivated a gene; thus it restores wild type.

A **backcross** (also known as a testcross) describes a genetic cross in which a hybrid strain is crossed to one of its two parental strains.

The **background level** of mutation describes the rate at which sequence changes accumulate in the genome of an organism. It reflects the balance between the occurrence of spontaneous mutations and their removal by repair systems, and is characteristic for any species.

A **bacterial artificial chromosome** (BAC) is a synthetic DNA molecule that contains the sequences needed for replication and segregation in bacteria. This is used in genomic cloning to amplify sequences typically 100-200 kb long. They are usually derived from derived from the naturally-occurring F factor episome.

Bam islands are a series of short, repeated sequences found in the nontranscribed spacer of *Xenopus* rDNA genes. The name reflects their isolation by use of the BamI restriction enzyme.

Bands of polytene chromosomes are visible as dense regions that contain the majority of DNA. They include active genes.

A **basal factor** is a transcription factor required by RNA polymerase II to form the initiation complex at all promoters. Factors are identified as TFIIX, where X is a number.

The level of response from a system in the absence of a stimulus is its **basal level.** (The basal level of transcription of a gene is the level that occurs in the absence of any specific activation.)

The **basal transcription apparatus** is the complex of transcription factors that assembles at the promoter before RNA polymerase is bound.

Base mispairing is a coupling between two bases that does not conform to the Watson-Crick rule, e.g., adenine with cytosine, thymine with guanine.

Base pairing describes the specific (complementary) interactions of adenine with thymine or of guanine with cytosine in a DNA double helix (thymine is replaced by uracil in double helical RNA).

Each VSG (variable surface glycoprotein) of a trypanosome is coded by a **basic copy** gene.

A bHLH protein has a basic DNA-binding region adjacent to the helix-loop-helix motif.

Bidirectional replication describes a system in which an origin generates two replication forks that proceed away from the origin in opposite directions.

The **bithorax complex** is a group of homeotic genes which are responsible for the diversification of the different segments of the fly.

A **bivalent** is the structure containing all four chromatids (two representing each homologue) at the start of meiosis.

A **blocked** reading frame cannot be translated into protein because of the occurrence of termination codons.

Branch migration describes the ability of a DNA strand partially paired with its complement in a duplex to extend its pairing by displacing the resident strand with which it is homologous.

The **branch site** is a short sequence just before the end of an intron at which the lariat intermediate is formed in splicing by joining the 5' nucleotide of the intron to the 2' position of an adenosine.

Breakage and reunion describes the mode of genetic recombination, in which two DNA duplex molecules are broken at corresponding points and then rejoined crosswise (involving formation of a length of heteroduplex DNA around the site of joining).

The **breakage-fusion-bridge** cycle is a type of chromosomal behavior in which a broken chromatid fuses to its sister, forming a "bridge". When the centromeres separate at mitosis, the chromosome breaks again (not necessarily at the bridge), thereby restarting the cycle.

Some species of yeast, the most well known of which is *Saccharomyces cerevisiae*, reproduce by forming a **bud.** The bud is formed off the side of the mother cell and gradually enlarges over the course of the cell cycle. Its interior is initially continuous with the cytoplasm of the mother cell, but after a copy of the genome is segregated into the bud during mitosis a wall is constructed between the two and the bud breaks free to become a separate cell.

A **bZIP** protein has a basic DNA-binding region adjacent to a leucine zipper dimerization motif.

C genes code for the constant regions of immunoglobulin protein chains.

C-bands are generated by staining techniques that react with centromeres. The centromere appears as a darkly-staining dot.

The **C-value** is the total amount of DNA in the genome (per haploid set of chromosomes).

The **C-value paradox** describes the lack of relationship between the DNA content (C-value) of an organism and its coding potential.

A CAAT **box** is part of a conserved sequence located upstream of the startpoints of eukaryotic transcription units; it is recognized by a large group of transcription factors.

A **cap** is the structure at the 5' end of eukaryotic mRNA, introduced after transcription by linking the terminal phosphate of 5' GTP to the terminal base of the mRNA. The added G (and sometimes some other bases) are methylated, giving a structure of the form 7MeG5'ppp5'Np . . .

A **cap** 0 at the 5' end of mRNA has only a methyl group on 7-guanine.

A **cap** 1 at the 5' end of mRNA has methyl groups on the terminal 7-guanine and the 2′-O position of the next base.

A **cap** 2 has three methyl groups (7-guanine, 2′-O position of next base, and N^6 adenine) at the 5' end of mRNA.

A capsid is the external protein coat of a virus particle.

The **carboxy terminal domain** (CTD) of eukaryotic RNA polymerase is phosphorylated at initiation and is involved in coordinating several activities with transcription.

A **carcinogen** is a chemical that increases the frequency with which cells are converted to a cancerous condition.

Cargo describes any macromolecule (e.g., RNA, soluble or membrane proteins) that is transported from one compartment to another. Cargo may contain sequences or modifications that specify their destination.

A **carrier protein** moves a solute directly from one side of the plasma membrane to the other. In the process, the protein undergoes a conformational change.

A **cascade** is a sequence of events, each of which is stimulated by the previous one. In transcriptional regulation, as seen in sporulation and phage lytic development, it means that regulation is divided into stages, and at each stage, one of the genes that are expressed codes for a regulator needed to express the genes of the next stage.

Caspases comprise a family of proteases some of whose members are involved in apoptosis (programmed cell death).

The **cassette** model for yeast mating type proposes that there is a single active locus (the active cassette) and two inactive copies of the locus (the silent cassettes). Mating type is changed when an active cassette of one type is replaced by a silent cassette of the other type.

To **catenate** is to link together two circular molecules as in a chain.

The **CD region** (common docking) is a C-terminal region in a MAP kinase (separate from the active site) that is involved in binding to a target protein.

CD3 is a complex of proteins that associates with the T cell antigen receptor's α and β chains. Each complex consists of one each of the 8, ε, γ chains and two ζ chains.

cdc is an abbreviation for "cell division cycle". It is most frequently used as part of the names of a large collection of yeast mutants isolated in the 1970s in which the cell cycle arrested at a specific point in each type of mutant.

cDNA is a single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription *in vitro.*

The **cell cycle** is the set of stages through which a cell progresses from one division to the next.

The **cell division cycle** is the entire sequence of events required to reliably replicate the cell's genetic material and separate the two copies into new cells. The term "cell division cycle" has been largely replaced by the term "cell cycle".

The **cell-mediated response** is the immune response that is mediated primarily by T lymphocytes. It is defined based on immunity that cannot be transferred from one organism to another by serum antibody.

The **central dogma** describes the basic nature of genetic information: sequences of nucleic acid can be perpetuated and interconverted by replication, transcription, and reverse transcription, but translation from nucleic acid to protein is unidirectional, because nucleic acid sequences cannot be retrieved from protein sequences.

The **central element** is a structure that lies in the middle of the synaptonemal complex, along which the lateral elements of homologous chromosomes align.

A **centriole** is a small hollow cylinder consisting of microtubules. It occurs in the centrosome (a type of microtubule organizing center) and is thought to play a role in organizing the microtubules.

The **centromere** is a constricted region of a chromosome that includes the site of attachment (the kinetochore) to the mitotic or meiotic spindle.

Centrosomes are the regions from which microtubules are organized at the poles of a mitotic cell. In animal cells, each centrosome contains a pair of centrioles surrounded by a dense amorphous region to which the microtubules attach.

Chaperones are a class of proteins which bind to incompletely folded or assembled proteins in order to assist their folding or prevent them from aggregating.

A **checkpoint** is an event in the cell cycle that can only proceed if some earlier event has been completed.

Chemical proofreading describes a proofreading mechanism in which the correction event occurs after addition of an incorrect subunit to a polymeric chain, by reversing the addition reaction.

A chiasma (*pl.* chiasmata) is a site at which two homologous chromosomes appear to have exchanged material during meiosis.
Chloroplast DNA (ctDNA) is an independent genome (usually circular) found in a plant chloroplast.
Chromatids are the copies of a chromosome produced by replication. The name is usually used to describe the copies in the period before they separate at the subsequent cell division.
Chromatin describes the state of nuclear DNA and its associated proteins during the interphase (between mitoses) of the eukaryotic cell cycle.
Chromatin remodeling describes the energy-dependent displacement or reorganization of nucleosomes that occurs in conjunction with activation of genes for transcription.
The **chromocenter** is an aggregate of heterochromatin from different chromosomes.
Chromomeres are densely staining granules visible in chromosomes under certain conditions, especially early in meiosis, when a chromosome may appear to consist of a series of chromomeres.
A **chromosome** is a discrete unit of the genome carrying many genes. Each chromosome consists of a very long molecule of duplex DNA and an approximately equal mass of proteins. It is visible as a morphological entity only during cell division.
Chromosome pairing is the coupling of the homologous chromosomes at the start of meiosis.
cis configuration describes two sites on the same molecule of DNA.
The *cis* **face** of the Golgi is the side juxtaposed to the nucleus.
A ***cis*-acting** site affects the activity only of sequences on its own molecule of DNA (or RNA); this property usually implies that the site does not code for protein.
A ***cis*-dominant** site or mutation affects the properties only of its own molecule of DNA. *cis*-dominance is taken to indicate that a site does not code for a diffusible product. (A rare exception is that a protein is cis-dominant when it is constrained to act only on the DNA or RNA from which it was synthesized.)
The **cisternae** of the Golgi apparatus are the successive stacks, each bounded by a membrane, that make up individual compartments.
Cisternal maturation is a model for the mechanism for cargo transport through the Golgi stack. It is also called cisternal migration or cisternal progression. In this model, a new Golgi cisterna forms at the cis face, then moves forward in the stack as the protein content of the cisterna changes from *cis* to medial to *trans.* Proteins that belong in earlier cisternae are retrieved by retrograde transport vesicles.
A **cistron** is the genetic unit defined by the *cis/trans* test; equivalent to gene.
Class switching describes a change in Ig gene organization in which the C region of the heavy chain is changed but the V region remains the same.
Clathrin proteins interact with adaptor proteins to form the coat on some of the vesicles that bud from the cytoplasmic face of the plasma membrane and the *trans-Golgi* network. Clathrin is composed of heavy and light chains that form triskelions, which then assemble into polyhedral curved lattices during the formation of clathrin-coated pits and vesicles.
A **clathrin-coated vesicle** is a membrane-bounded compartment that mediates endocytosis, formation of secretory granules at the trans-Golgi network, and transport from the trans-Golgi network to the endocytic pathway. In addition to clathrin, its major constituents include cargo and adaptor proteins.
A **clear plaque** is a type of plaque that contains only lysed bacterial cells.
The constriction in the cell cortex that separates newly reformed nuclei after mitosis and results in the formation of two cells is the **cleavage furrow.**
The fertilized eggs of some species are very large and initially undergo several rounds of cell division without any growth of the cells between successive mitoses. As a result each embryo is progressively divided into into smaller and smaller cells. This process is the **cleavage stage** of embryogenesis.
Clonal deletion describes the elimination of a clonal population of lymphocytes. At certain stages of lymphocyte development, clonal deletion can be induced when lymphocyte antigen receptors bind to their cognate antigen.
The **clonal selection** theory proposed that each lymphocyte expresses a single antigen receptor specificity and that only those lymphocytes that bind to a given antigen are stimulated to proliferate and to function in eliminating that antigen. Thus, the antigen "selects" the lymphocytes to be activated. Clonal selection is now an established principle in immunology.
The **cloverleaf** describes the structure of tRNA drawn in two dimensions, forming four distinct arm-loops.
Co-translational translocation describes the movement of a protein across a membrane as the protein is being synthesized. The term is usually restricted to cases in which the ribosome binds to the channel. This form of translocation may be restricted to the endoplasmic reticulum.
Coactivators are factors required for transcription that do not bind DNA but are required for (DNA-binding) activators to interact with the basal transcription factors.
A **coated pit** is an infolding of membrane formed during clathrin-mediated endocytosis. It is pinched off to form a clathrin-coated vesicle.
Coated vesicles are vesicles whose membrane has on its surface a layer of a protein such as clathrin, cop-I or COP-II.
Coatomer is another name for the complex of COPI coat proteins.
A **coding end** is produced during recombination of immunoglobulin and T cell receptor genes. Coding ends are at the termini of the cleaved V and (D)J coding regions. The subsequent joining of the coding ends yields a coding joint.
A **coding region** is a part of the gene that represents a protein sequence.
The **coding strand** (sense strand) of DNA has the same sequence as the mRNA and is related to the protein sequence that it represents by the genetic code.
Two alleles are said to be **codominant** when they are each equally evident in the phenotype of the heterozygote.
A **codon** is a triplet of nucleotides that represents an amino acid or a termination signal.
A **cofactor** is a small inorganic component (often a metal ion) that is required for the proper structure or function of an enzyme.
Cognate tRNAs (isoaccepting tRNAs) are those recognized by a particular aminoacyl-tRNA synthetase. They all are charged with the same amino acid.
Cohesin proteins form a complex that holds sister chromatids together. They include some SMC proteins.
Coincidental evolution (coevolution) describes a situation in which two genes evolve together as a single unit.
A **cointegrate** structure is produced by fusion of two replicons, one originally possessing a transposon, the other lacking it; the cointegrate has copies of the transposon present at both junctions of the replicons, oriented as direct repeats.
A **colinear** relationship describes the 1:1 representation of a sequence of triplet nucleotides in a sequence of amino acids.
A **compatibility group** of plasmids contains members unable to coexist in the same bacterial cell.
Two mutants are said to **complement** each other when a diploid that is heterozygous for each mutation produces the wild type phenotype.
Complementary base pairs are defined by the pairing reactions in double helical nucleic acids (A with T in DNA or with U in RNA, and C with G).
A **complementation group** is a series of mutations unable to complement when tested in pairwise combinations in *trans;* defines a genetic unit (the cistron).
A **complementation test** determines whether two mutations are alleles of the same gene. It is accomplished by crossing two dif-

ferent recessive mutations that have the same phenotype and determining whether the wild-type phenotype can be produced. If so, the mutations are said to complement each other and are probably not mutations in the same gene.

Complete dominance is the state in which the phenotype is the same when the dominant allele is homozygous or heterozygous.

A **complex locus** (of *D. melanogaster)* has genetic properties inconsistent with the function of a gene representing a single protein. Complex loci are usually very large (> 100 kb) at the molecular level.

A **complex oligosaccharide** is an N-linked oligosaccharide that is made during transit through the Golgi apparatus. Mannose residues are trimmed from the high mannose precursor in the rough endoplasmic reticulum and *cis* Golgi, and other sugars are added by enzymes in the medial and trans Golgi cisternae to form a complex oligosaccharide.

Complexity is the total length of different sequences of DNA present in a given preparation.

Composite transposons (composite elements) have a central region flanked on each side by insertion sequences, either or both of which may enable the entire element to transpose.

A **concentration gradient** is a change in the concentration of a molecule or ion from one point to another. The gradient might be gradual (as in a solution that is not homogenous) or abrupt (created by a membrane).

Concerted evolution describes the ability of two related genes to evolve together as though constituting a single locus.

Condensin proteins are components of a complex that binds to chromosomes to cause condensation for meiosis or mitosis. They are members of the SMC family of proteins.

Conjugation is a process in which two cells come in contact and exchange genetic material. In bacteria, DNA is transferred from a donor to a recipient cell. In protozoa, DNA passes from each cell to the other.

A **consensus sequence** is an idealized sequence in which each position represents the base most often found when many actual sequences are compared.

Conservative transposition refers to the movement of large elements, originally classified as transposons, but now considered to be episomes. The mechanism of movement resembles that of phage excision and integration.

Conserved positions are defined when many examples of a particular nucleic acid or protein are compared and the same individual bases or amino acids are always found at particular locations.

Constant regions (C regions) of immunoglobulins are coded by C gene segments and are the parts of the chain that vary least. Those of heavy chains identify the type of immunoglobulin.

A **constitutive** process is one that occurs all the time, unchanged by any form of stimulus or external condition.

Constitutive heterochromatin describes the inert state of permanently nonexpressed sequences, usually satellite DNA.

Constitutively secreted macromolecules are transported to the plasma membrane or secreted at a relatively constant rate. They include lipids and soluble and membrane proteins. They are not secreted by regulated exocytosis and they exit to the plasma membrane from the trans-Golgi network.

The **context** of a codon in mRNA refers to the fact that neighboring sequences may change the efficiency with which a codon is recognized by its aminoacyl-tRNA or is used to terminate protein synthesis.

A **contig** is a continuous stretch of genomic DNA generated by assembling cloned fragments by means of their overlaps.

The **contractile ring** is a ring of actin filaments that forms around the equator at the end of mitosis and is responsible for pinching the daughter cells apart.

Controlling elements of maize are transposable units originally identified solely by their genetic properties. They may be autonomous (able to transpose independently) or nonautonomous (able to transpose only in the presence of an autonomous element).

Cooperativity in protein binding describes an effect in which binding of the first protein enhances binding of a second protein (or another copy of the same protein).

Coordinate regulation refers to the common control of a group of genes.

A **COP-I**-coated vesicle is a membrane-bounded compartment that buds from the cytoplasmic face of the Golgi complex and mediate retrograde transport from the Golgi complex to the rough endoplasmic reticulum. COP-I-coated vesicles may also mediate transport between Golgi cisternae.

A **COP-II** coat consists of a protein complex containing 5 major proteins. COPII-coated vesicles are membrane-bounded vesicles that bud from the cytoplasmic face of the rough endoplasmic reticulum and mediate anterograde transport from the rough ER to the Golgi.

Copy choice is a type of recombination used by RNA viruses, in which the RNA polymerase switches from one template to another during synthesis.

The **copy number** is the number of copies of a plasmid that is maintained in a bacterium (relative to the number of copies of the origin of the bacterial chromosome).

Cordycepin is 3' deoxyadenosine, an inhibitor of polyadenylation of RNA.

The **core** sequence is the segment of DNA that is common to the attachment sites on both the phage lambda and bacterial genomes. It is the location of the recombination event that allows phage lambda to integrate.

Core DNA is the 146 bp of DNA contained on a core particle.

The **core enzyme** is the complex of RNA polymerase subunits that undertakes elongation. It does not include additional subunits or factors that may needed for initiation or termination.

A **core histone** is one of the four types (H2A, H2B, H3, H4) found in the core particle derived from the nucleosome (this excludes histone H1).

The **core particle** is a digestion product of the nucleosome that retains the histone octamer and has 146 bp of DNA; its structure appears similar to that of the nucleosome itself.

The **core promoter** of RNA polymerase I is the region immediately surrounding the startpoint. It is necessary and sufficient to initiate transcription, but only at a low level.

A **corepressor** is a small molecule that triggers repression of transcription by binding to a regulator protein.

Cosuppression describes the ability of a transgene (usually in plants) to inhibit expression of the corresponding endogenous gene.

A **Cot curve** is a plot of the extent of renaturation of DNA against time.

The $\mathbf{Cot^{1}/_{2}}$ is the midpoint of a Cot curve. It is proportional to the complexity of the DNA sequences in the renaturation reaction.

A **countertranscript** is an RNA molecule that prevents an RNA primer from initiating transcription by base pairing with the primer.

CpG island is a stretch of 1-2 kb in a mammalian genome that is rich in unmethylated CpG doublets.

Crisis is a state reached when primary cells placed into culture are unable to replicate their DNA because their telomeres have become too short. Most cells die, but a few emerge by a process of immortalization that usually involves changes to bypass the limitations of telomeric length.

Crossing-over describes the reciprocal exchange of material between chromosomes that occurs during prophase I of meiosis and is responsible for genetic recombination.

Crossover control limits the number of recombination events between meiotic chromosomes to 1-2 crossovers per pair of homologs.

Crossover fixation refers to a possible consequence of unequal crossing-over that allows a mutation in one member of a tandem cluster to spread through the whole cluster (or to be eliminated).

Crown gall disease is a tumor that can be induced in many plants by infection with the bacterium Agrobacterium tumefaciens.

CRP activator (CAP activator) is a positive regulator protein activated by cyclic AMP. It is needed for RNA polymerase to initiate transcription of certain (catabolite-sensitive) operons of *E. coli*.

Cryptic satellite is a satellite DNA sequence not identified as such by a separate peak on a density gradient; that is, it remains present in main-band DNA.

The **cutting periodicity** is the spacing between cleavages on each strand when a duplex DNA immobilized on a flat surface is attacked by a DNAase that makes single-strand cuts.

Cyclins are proteins that accumulate continuously throughout the cell cycle and are then destroyed by proteolysis during mitosis. A cyclin is one of the two subunits of the M-phase kinase.

A **cyclin-dependent kinase** (cdk) is one of a family of kinases which are inactive unless bound to a cyclin molecule. Most cyclin-dependent kinases participate in some aspect of cell cycle control.

Cyclin-dependent kinase inhibitors (cki) are a class of proteins which inhibit cyclin-dependent kinases by binding to them. Inhibition lasts until the cki is inactivated, often in response to a signal for the cell cycle to progress.

The **cyclosome** is a multisubunit complex which initiates anaphase and the exit of cells from mitosis by promoting the ubiquitination and proteolysis of a variety of proteins. These include the mitotic cyclins, several proteins required to hold sister chromatids together, and other proteins which control the dynamics of the mitotic spindle.

A **cytokine** is a small polypeptide that affects the growth of particular types of cells.

Cytokinesis is the process involved in separation and movement apart of daughter cells. Cytokinesis occurs at the end of mitosis.

The **cytoplasmic domain** is the part of a transmembrane protein that is exposed to the cytosol.

The side of the plasma membrane, or of the membrane of an organelle, which faces the cytoplasm is its **cytoplasmic face.**

Cytoplasmic inheritance is a property of genes located in mitochondria or chloroplasts.

A **cytotoxic T cell** is a T lymphocyte (usually $CD8^+$) that can be stimulated to kill cells containing intracellular pathogens, such as viruses.

Cytotype is a cytoplasmic condition that affects P element activity. The effect of cytotype is due to the presence or absence of transposition repressors, which are provided by the mother to the

The D **arm** of tRNA has a high content of the base dihydrouridine.

A **D loop** is a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region. The same term is used also to describe the displacement of a region of one strand of duplex DNA by a complementary single-stranded invader.

The D **segment** is an additional sequence that is found between the V and J regions of an immunoglobulin heavy chain.

A **daughter** strand or duplex of DNA refers to the newly synthesized DNA.

The two cells that result from a cell division are referred to as **daughter cells.** In budding yeast only the cell derived from the bud is called the daughter cell.

A **de novo methylase** adds a methyl group to an unmethylated target sequence on DNA.

A **deacetylase** is an enzyme that removes acetyl groups from proteins.

Deacylated tRNA has no amino acid or polypeptide chain attached because it has completed its role in protein synthesis and is ready to be released from the ribosome.

The **death domain** is a protein-protein interaction motif found in certain proteins of the apoptotic pathway.

The **degradosome** is a complex of bacterial enzymes, including RNAases, a helicase, and enolase (a glycolytic enzyme), which may be involved in degrading mRNA.

Delayed early genes in phage lambda are equivalent to the middle genes of other phages. They cannot be transcribed until regulator protein(s) coded by the immediate early genes have been synthesized.

Deletions are generated by removal of a sequence of DNA, the regions on either side being joined together.

A **demethylase** is a casual name for an enzyme that removes a methyl group, typically from DNA, RNA, or protein.

Denaturation of protein describes its conversion from the physiological conformation to some other (inactive) conformation.

A **density gradient** is used to separate macromolecules on the basis of differences in their density. It is prepared from a heavy soluble compound such as CsCl.

Density-dependent inhibition describes the limitation that eukaryotic cells in culture grow only to a limited density, because growth is inhibited, by processes involving cell-cell contacts.

A **denticle** is a pigmented, hardened spike of cuticle protruding from the ventral epidermis of a *Drosophila* embryo.

A **deoxyribonuclease** (DNAase) is an enzyme that specifically digests DNA. It may cut only one strand or may cut both strands.

Deoxyribonucleic acid (DNA) is a nucleic acid molecule consisting of long chains of polymerized (deoxyribo)nucleotides. In double-stranded DNA the two strands are held together by hydrogen bonds between complementary nucleotide base pairs.

The **derepressed** state describes a gene that is turned on because a small molecule corepressor is absent. It has the same effect as the induced state that is produced by a small molecule inducer for a gene that is regulated by induction. In describing the effect of a mutation, derepressed and constitutive have the same meaning.

A **dicentric chromosome** is the product of fusing two chromosome fragments, each of which has a centromere. It is unstable and may be broken when the two centromeres are pulled to opposite poles in mitosis.

Direct repeats are identical (or closely related) sequences present in two or more copies in the same orientation in the same molecule of DNA; they are not necessarily adjacent.

Divergence is the percent difference in nucleotide sequence between two related DNA sequences or in amino acid sequences between two proteins.

DNA fingerprinting analyzes the differences between individuals of the fragments generated by using restriction enzymes to cleave regions that contain short repeated sequences. Because these are unique to every individual, the presence of a particular subset in any two individuals can be used to define their common inheritance (e.g. a parent-child relationship).

DNA ligase makes a bond between an adjacent 3'-OH and 5'-phosphate end where there is a nick in one strand of duplex DNA.

A **dna mutant** of bacteria is temperature-sensitive; it cannot synthesize DNA at 42°C, but can do so at 37°C.

A **DNA polymerase** is an enzyme that synthesizes a daughter strand(s) of DNA (under direction from a DNA template). Any particular enzyme may be involved in repair or replication (or both).

A **DNA replicase** is a DNA-synthesizing enzyme required specifically for replication.

DNA topoisomerase is an enzyme that changes the number of times the two strands in a closed DNA molecule cross each other. It does this by cutting the DNA, passing DNA through the break, and resealing the DNA.

DNAases are enzymes that attack bonds in DNA.

The initial association of a translating ribosome with the translocation channel in the membrane of the ER is called **docking.**

The **docking groove** is a region near to, but distinct from the active site of a MAP kinase that is involved in binding to a target protein.

The **docking site** (D domain) is a region in a target protein that used by a MAP kinase to bind to it. The docking site has a high

concentration of hydrophobic residues separated from two basic residues.

Dolichol is a lipid that consists of a long chain of isoprenoid units and is present in the membrane of the rough endoplasmic reticulum. It is part of the precursor in the synthesis of N-linked oligosaccharides. An oligosaccharide is assembled onto dolichol via a pyrophosphoryl linkage, then transferred to particular asparagine residues of a nascent polypeptide.

A **domain** of a chromosome may refer *either* to a discrete structural entity defined as a region within which supercoiling is independent of other domains; *or* to an extensive region including an expressed gene that has heightened sensitivity to degradation by the enzyme DNAase I.

A **domain** of a protein is a discrete continuous part of the amino acid sequence that can be equated with a particular function.

A **dominant** allele determines the phenotype displayed in a heterozygote with another (recessive) allele. An allele is one of several alternative forms of a gene occupying a given locus on a chromosome.

Dominant negative mutations are frans-acting and are a hallmark of negative complementation occurring in multimeric proteins where one mutant subunit may poison the whole multimer even though the other subunits are wild-type.

The **dorsal-ventral axis** is the line running from the back to the belly of an animal.

Dosage compensation describes mechanisms employed to compensate for the discrepancy between the presence of two X chromosomes in one sex but only one X chromosome in the other sex.

Double-minute chromosomes are extrachromosomal elements formed by amplification of DHFR genes in response to methotrexate treatment. They are large enough to be visible in the light microscope.

A **double-strand break** (DSB) occurs when both strands of a DNA duplex are cleaved at the same site. Genetic recombination is initiated by double-strand breaks. The cell also has repair systems that act on double-strand breaks created at other times.

The **doubling time** is the period (usually measured in minutes) that it takes for a bacterial cell to reproduce.

A **down mutation** in a promoter decreases the rate of transcription.

Downstream identifies sequences proceeding farther in the direction of expression; for example, the coding region is downstream of the initiation codon.

A DP **thymocyte** is a double positive thymocyte. It is an immature T cell that expresses cell surface CD4 and CD8. Selection of DP thymocytes in the thymus yields mature T cells expressing either CD4 or CD8.

A **dual specificity kinase** is a protein kinase that can phosphorylate tyrosine or threonine or serine amino acids.

Dynamin is a cytosolic protein that is a GTPase and is required for clathrin-mediated vesicle formation. Although the exact role of dynamin is debated, dynamin polymers are involved in the scission of clathrin-coated pits from membranes. A variant of dynamin functions in mitochondrial septation.

An **early endosome** is the part of the endosomal compartment in which endocytosed molecules appear after a minute or so. Early endosomes are located near the plasma membrane, function in sorting of endocytosed molecules, and have a pH of about 6.

Early genes are transcribed before the replication of phage DNA. They code for regulators and other proteins needed for later stages of infection.

Early infection is the part of the phage lytic cycle between entry and replication of the phage DNA. During this time, the phage synthesizes the enzymes needed to replicate its DNA.

Ectopic refers to something being out of place.

EF-G is an elongation factor needed for the translocation stage of bacterial protein synthesis.

An **effector** is the target protein for the activated G protein.

The **effector site** is the site that is bound by a small molecule on an allosteric protein. The result of binding is to change the activity of the active site, which is located elsewhere on the protein.

An **electrical gradient** is a change in the amount of charge from one point to another.

A change in the concentration of ions from one point to another produces an **electrochemical gradient.** The term indicates that there is a change in the concentration of both electrical charge and of a chemical species.

Elongation is the stage in a macromolecular synthesis reaction (replication, transcription, or translation) when the nucleotide or polypeptide chain is being extended by the addition of individual subunits.

Elongation factors (EF in prokaryotes, eEF in eukaryotes) are proteins that associate with ribosomes cyclically, during addition of each amino acid to the polypeptide chain.

End labeling describes the addition of a radioactively labeled group to one end (5' or 3') of a DNA strand.

Endocytic vesicles are membranous particles that transport proteins through endocytosis; also known as clathrin-coated vesicles.

Endocytosis describes the process by which material at the surface of the cell is internalized. The process involves the formation of a membranous vesicle.

Endonucleases cleave bonds within a nucleic acid chain; they may be specific for RNA or for single-stranded or double-stranded DNA.

The **endoplasmic reticulum** is an organelle involved in the synthesis of lipids, membrane proteins, and secretory proteins. It consists of a highly convoluted sheet of membranes, extending from the outer layer of the nuclear envelope into the cytoplasm.

An **endosome** is an organelle that functions to sort endocytosed molecules and molecules delivered from the trans-Golgi network and deliver them to other compartments, such as lysosomes. It consists of membrane-bounded tubules and vesicles.

An **endotoxin** is a toxin that is present on the surface of Gram-negative bacteria (as opposed to exotoxins, which are secreted). LPS is an example of an endotoxin.

An **enhanceosome** is a complex of transcription factors that assembles cooperatively at an enhancer.

An **enhancer** is a *cis*-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter.

Enzyme turnover is the process through which the enzyme returns to its original shape, enabling the enzyme to catalyze another reaction.

Epigenetic changes influence the phenotype without altering the genotype. They consist of changes in the properties of a cell that are inherited but that do not represent a change in genetic information.

An **episome** is a plasmid able to integrate into bacterial DNA.

An **epitope** is the portion of an antigen that is recognized by the antigen receptor on lymphocytes. It is also called an antigenic determinant.

Error-prone synthesis occurs when DNA incorporates noncomplementary bases into the daughter strand.

An **established** cell line consists of cells that can be grown indefinitely in culture (they are said to be immortalized).. The cells usually have had chromosomal changes in order to adapt to culture conditions.

Euchromatin comprises all of the genome in the interphase nucleus except for the heterochromatin. The euchromatin is less tightly coiled than heterochromatin, and contains the active or potentially active genes.

The **evolutionary clock** is defined by the rate at which mutations accumulate in a given gene.

The **excision** of phage or episome or other sequence describes its release from the host chromosome as an autonomous DNA molecule.

Excision repair describes a type of repair system in which one strand of DNA is directly excised and then replaced by resynthesis using the complementary strand as template.

The **exocyst** is a complex of 8 proteins that is found at sites on the plasma membrane where secretion occurs. It tethers secretory vesicles to the membrane as the first step in the process of membrane fusion.

Exocytosis is the process of secreting proteins from a cell into the medium, by the fusion of the secretory vesicle with the plasma membrane.

An **exon** is any segment of an interrupted gene that is represented in the mature RNA product.

Exon definition describes the process when a pair of splicing sites are recognized by interactions involving the 5' site of the intron and also the 5' of the next intron downstream.

Exon trapping inserts a genomic fragment into a vector whose function depends on the provision of splicing junctions by the fragment.

Exonucleases cleave nucleotides one at a time from the end of a polynucleotide chain; they may be specific for either the 5' or 3' end of DNA or RNA.

The **exosome** is a complex of several exonucleases involved in degrading RNA.

Exportins are transport receptors that bind their cargo and associate with RanGTP in the nucleus. The trimeric complex translocates across the nuclear envelope into the cytoplasm, where hydrolysis of GTP bound to Ran results in release of cargo.

An **expressed sequence tag** (EST) is a short sequence of DNA taken from a cDNA copy of an mRNA. The EST is complementary to the mRNA and can be used to identify genes corresponding to the mRNA.

An **expression site** in a trypanosome genome is a locus near a telomere that can express the VSG gene that is located there.

The **expression-linked copy** (ELC)in a trypanosome genome is the one copy of a VSG gene that is expressed.

Extein sequences remain in the mature protein that is produced by processing a precursor via protein splicing.

The **external domain** is the part of a plasma membrane protein that extends outside of the cell. Upon internalization, the protein's external domain extends into the lumen (the topological equivalent of the outside of the cell) of an organelle.

The **extra arm** of tRNA lies between the TyC and anticodon arms. It is the most variable in length in tRNA, from 3-21 bases. tRNAs are called class 1 if they lack it, and class 2 if they have it.

The **extracellular matrix** (ECM) is a relatively rigid layer of insoluble glycoproteins that fill the spaces between cells in multicellular organisms. These glycoproteins connect to plasma membrane proteins.

An **extrachromosomal genome** in a bacterium is a self-replicating set of genes that is not part of the bacterial chromosome. In many cases, the genes are necessary for bacterial growth under certain environmental conditions.

Extranuclear genes reside outside the nucleus in organelles such as mitochondria and chloroplasts.

The **F plasmid** is an episome that can be free or integrated in *E. coli*, and which in either form can sponsor conjugation.

Facultative heterochromatin describes the inert state of sequences that also exist in active copies-for example, one mammalian X chromosome in females.

The **fast component** of a reassociation reaction is the first to renature and contains highly repetitive DNA.

Feedback inhibition describes the ability of a small molecule product of a metabolic pathway to inhibit the activity of an enzyme that catalyzes an earlier step in the pathway.

In *Drosophila,* a **female sterile** mutation is one in that causes sterility in the female, often because of abnormalities in oogenesis.

The **10 nm fiber** is a linear array of nucleosomes, generated by unfolding from the natural condition of chromatin.

The **30 nm fiber** is a coiled coil of nucleosomes. It is the basic level of organization of nucleosomes in chromatin.

Fixation is the process by which a new allele replaces the allele that was previously predominant in a population.

Fluidity is a property of membranes; it indicates the ability of lipids to move laterally within their particular monolayer.

Transformed cells grow as a compact mass of rounded-up cells that grows in dense clusters, piled up on one another. They appear as a distinct **focus** on a culture plate, contrasted with normal cells that grow as a spread-out monolayer attached to the substratum.

Footprinting is a technique for identifying the site on DNA bound by some protein by virtue of the protection of bonds in this region against attack by nucleases.

Forward mutations inactivate a wild-type gene.

Frameshift mutations arise by deletions or insertions that are not a multiple of 3 bp; they change the frame in which triplets are translated into protein.

A **fully methylated** site is a palindromic sequence that is methylated on both strands of DNA.

Functionally redundant genes fulfill the same function in the same time and place, so that mutation of every member of the set is necessary to show a deficient phenotype.

G proteins are guanine nucleotide-binding proteins. Trimeric G proteins are associated with the plasma membrane. When bound by GDP the trimer remains intact and is inert. When the GDP is replaced by GTP, the a subunit is released from the $\beta\gamma$ dimer. Either the a monomer or the $\beta\gamma$ dimer then activates or represses a target protein. Monomeric G proteins are cytosolic and work on the same principle that the form bound to GDP is inactive, but the form bound to GTP is active.

G-bands are generated on eukaryotic chromosomes by staining techniques and appear as a series of lateral striations. They are used for karyotyping (identifying chromosomal regions by the banding pattern).

GO is a noncycling state in which a cell has ceased to divide.

G1 is the period of the eukaryotic cell cycle between the last mitosis and the start of DNA replication.

G2 phase is the period of the cell cycle separating the replication of a cell's chromosomes (S phase) from the following mitosis (M phase).

A **gain-of-function** mutation represents acquisition of a new activity. It is dominant.

In *Drosophila* the **gap genes** are a set of genes that help set up the segmentation of the embryo. Gap genes encode transcription factors that are expressed in broad regions of the embryo. Gap genes activate transcription of the pair-rule genes.

A channel which only allows passage of its substrate under certain conditions is referred to as "gated". Gated channels can exist in at least two conformations, one of which is open and the other closed.

The GC **box** is a common pol II promoter element consisting of the sequence GGGCGG.

A **gene** (cistron) is the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

A **gene cluster** is a group of adjacent genes that are identical or related.

Gene conversion is the alteration of one strand of a heteroduplex DNA to make it complementary with the other strand at any position(s) where there were mispaired bases.

A **gene family** consists of a set of genes whose exons are related; the members were derived by duplication and variation from some ancestral gene.

The **genetic code** is the correspondence between triplets in DNA (or RNA) and amino acids in protein.

Genetic instability (genome instability) refers to a state in which there is large increase (X 100-fold) in the frequency of changes in the genome as seen by chromosomal rearrangements or other events that affect the genetic content. This is a key occurrence in the generation of cancer cells.

The **genome** is the complete set of sequences in the genetic material of an organism. It includes the sequence of each chromosome plus any DNA in organelles.

The **glucocorticoid response element** (GRE) is a sequence in a promoter or enhancer that is recognized by the glucocorticoid receptor, which is activated by glucocorticoid steroids.

Glucose repression (catabolite repression) describes the decreased expression of many bacterial operons that results from addition of glucose.

A **glycolipid** has a head consisting of an oligosaccharide, linked to a fatty acid tail.

GMP-PCP is an analog of GTP that cannot be hydrolyzed. It is used to test which stage in a reaction requires hydrolysis of GTP.

Golgi apparatus consists of individual stacks of membranes near the endoplasmic reticulum; involved in glycosylating proteins and sorting them for transport to different cellular locations.

Gratuitous inducers resemble authentic inducers of transcription but are not substrates for the induced enzymes.

A **growth factor** is a ligand, usually a small polypeptide, that activates a receptor in the plasma membrane to stimulate growth of the target cell. Growth factors were originally isolated as the components of serum that enabled cells to grow in culture.

GT-AG rule describes the presence of these constant dinucleotides at the first two and last two positions of introns of nuclear genes.

A **guide RNA** is a small RNA whose sequence is complementary to the sequence of a correctly edited RNA. It is used as a template for the insertion of nucleotides into the pre-edited RNA.

The **H2 locus** is the mouse major histocompatibility complex, a cluster of genes on chromosome 17. The genes encode proteins for antigen presentation, cytokines, and complement proteins.

The **haplotype** is the particular combination of alleles in a defined region of some chromosome, in effect the genotype in miniature. Originally used to described combinations of MHC alleles, it now may be used to describe particular combinations of RFLPs, SNPs, or other markers.

A **hapten** is a small molecule that acts as an antigen when conjugated to a protein.

Hb anti-Lepore is a fusion gene produced by unequal crossing-over that has the N-terminal part of β globin and the C-terminal part of δ globin.

Hb Kenya is a fusion gene produced by unequal crossing-over between the between $^{A}\gamma$ and β globin genes.

Hb Lepore is an unusual globin protein that results from unequal crossing-over between the β and δ genes. The genes become fused together to produce a single β-like chain that consists of the N-terminal sequence of δ joined to the C-terminal sequence of β.

HbH disease results from a condition in which there is a disproportionate amount of the abnormal tetramer β_4 relative to the amount of normal hemoglobin ($\alpha_2\beta_2$).

The **headpiece** is the DNA-binding domain of the *lac* repressor.

Heat shock genes are a set of loci that are activated in response to an increase in temperature (and other abuses to the cell). They occur in all organisms. They usually include chaperones that act on denatured proteins.

The **heat shock response element** (HSE) is a sequence in a promoter or enhancer that is used to activate a gene by an activator induced by heat shock.

The immunoglobulin **heavy chain** is one of two types of polypeptides in an antibody. Each antibody contains two heavy chains. The N-terminus of the heavy chain forms part of the antigen recognition site, whereas the C-terminus determines the subclass (isotype).

Heavy strands and light strands of a DNA duplex refer to the density differences that result when there is an asymmetry between base representation in the two strands such that one strand is rich in T and G bases and the other is rich in C and A bases. This occurs in some satellite and mitochondrial DNAs.

A **helicase** is an enzyme that uses energy provided by ATP hydrolysis to separate the strands of a nucleic acid duplex.

The **helix-loop-helix** (HLH) motif is responsible for dimerization of a class of transcription factors called HLH proteins. A bHLH protein has a basic sequence close to the dimerization motif that binds to DNA.

The **helix-turn-helix** motif describes an arrangement of two α helices that form a site that binds to DNA, one fitting into the major groove of DNA and other lying across it.

A **helper T cell** is a T lymphocyte that activates macrophages and stimulates B cell proliferation and antibody production. Helper T cells usually express cell surface CD4 but not CD8.

A **helper virus** provides functions absent from a defective virus, enabling the latter to complete the infective cycle during a mixed infection.

A **hemi-methylated** site is a palindromic sequence that is methylated on only one strand of DNA.

Hemimethylated DNA is methylated on one strand of a target sequence that has a cytosine on each strand.

Heterochromatin describes regions of the genome that are highly condensed, are not transcribed, and are late-replicating. Heterochromatin is divided into two types, which are called constitutive and facultative.

Heteroduplex DNA (hybrid DNA) is generated by base pairing between complementary single strands derived from the different parental duplex molecules; it occurs during genetic recombination.

Heterogeneous nuclear RNA (hnRNA) comprises transcripts of nuclear genes made by RNA polymerase II; it has a wide size distribution and low stability.

A **heterokaryon** is a cell containing two (or more) nuclei in a common cytoplasm, generated by fusing somatic cells.

A **heteromultimer** is a protein that is composed of nonidentical subunits (coded by different genes).

An individual is said to be **heterozygous** when it has different alleles of a given gene on each of its homologous chromosomes.

An **Hfr** cell is a bacterium that has an integrated F plasmid within its chromosome. Hfr stands for high frequency recombination, referring to the fact that chromosomal genes are transferred from an Hfr cell to an F' cell much more frequently than from an F^+ cell.

A **high mannose oligosaccharide** is an N-linked oligosaccharide that contains N-acetylglucosamine linked only to mannose residues. It is covalently added to transmembrane proteins in the rough endoplasmic reticulum and is trimmed and modified in the Golgi apparatus.

Highly repetitive DNA (simple sequence DNA) is the first component to reassociate and is equated with satellite DNA.

Histones are conserved DNA-binding proteins that form the basic subunit of chromatin in eukaryotes. Histones H2A, H2B, H3, H4 form an octameric core around which DNA coils to form a nucleosome. Histone H1 is external to the nucleosome.

Histone acetyltransferase (HAT) enzymes modify histones by addition of acetyl groups; some transcriptional coactivators have HAT activity.

Histone deacetyltransferase (HDAC) enzymes remove acetyl groups from histones; they may be associated with repressors of transcription.

The **histone fold** is a motif found in all four core histones in which three α-helices are connected by two loops.

The **HLA** locus is the human major histocompatibility complex, a cluster of genes on chromosome 6. The genes encode proteins for antigen presentation, cytokines, and complement proteins.

An **hnRNP** is the ribonucleoprotein form of hnRNA (heterogeneous nuclear RNA), in which the hnRNA is complexed with proteins.

A Holliday structure is an intermediate structure in homologous recombination, where the two duplexes of DNA are connected by the genetic material exchanged between two of the four strands, one from each duplex. A joint molecule is said to be resolved when nicks in the structure restore two separate DNA duplexes.
The **holoenzyme** (complete enzyme) is the complex of five subunits including core enzyme ($\alpha_2\beta\beta'$) and cr factor that is competent to initiate bacterial transcription.
The **homeobox** describes the conserved sequence that is part of the coding region of *D. melanogaster* homeotic genes; it is also found in amphibian and mammalian genes expressed in early embryonic development.
The **homeodomain** is a DNA-binding motif that typifies a class of transcription factors. The DNA sequence that codes for it is called the homeobox.
Homeotic genes are defined by mutations that convert one body part into another; for example, an insect leg may replace an antenna.
A **homogeneously staining region** (HSR) is produced by the tandem amplification of a chromosomal sequence. As a result, it does not have a banded pattern.
Homologous recombination (generalized recombination) involves a reciprocal exchange of sequences of DNA, e.g. between two chromosomes that carry the same genetic loci.
A **homomultimer** is a protein composed of identical subunits.
An individual is said to be **homozygous** when it has identical alleles of a given gene.
A **hotspot** describes a site in the genome at which the frequency of mutation (or recombination) is very much increased.
Housekeeping genes (constitutive genes) are those (theoretically) expressed in all cells because they provide basic functions needed for sustenance of all cell types.
The **humoral response** is an immune response that is mediated primarily by antibodies. It is defined as immunity that can be transferred from one organism to another by serum antibody.
Hybrid dysgenesis describes the inability of certain strains of *D. melanogaster* to interbreed, because the hybrids are sterile (although otherwise they may be phenotypically normal).
Hybridization describes the pairing of complementary RNA and DNA strands to give an RNA-DNA hybrid.
Hybridoma is a cell line produced by fusing a myeloma with a lymphocyte; it continues indefinitely to express the immunoglobulins of both parents.
A **hydropathy plot** is a measure of the hydrophobicity of a protein region and therefore of the likelihood that it will reside in a membrane.
Hydrops fetalis is a fatal disease resulting from the absence of the hemoglobin a gene.
Hypermutation describes the introduction of somatic mutations in a rearranged immunoglobulin gene. The mutations can change the sequence of the corresponding antibody, especially in its antigen-binding site.
A **hypersensitive site** is a short region of chromatin detected by its extreme sensitivity to cleavage by DNAase I and other nucleases; it comprises an area from which nucleosomes are excluded.

IAPs are inhibitors of apoptosis. They function by antagonizing the actions of caspases.
Icosahedral symmetry is typical of viruses that have capsids that are polyhedrons.
The **idling reaction** results in the production of pppGpp and ppGpp by ribosomes when an uncharged tRNA is present in the A site; triggers the stringent response.
IF-1 is a bacterial initiation factor that stabilizes the initiation complex.
IF-2 is a bacterial initiation factor that binds the initiator tRNA to the initiation complex.
IF-3 is a bacterial initiation factor required for 30S subunits to bind to initiation sites in mRNA. It also prevents 30S subunits from binding to 50S subunits.
Immediate early phage genes in phage lambda are equivalent to the early class of other phages. They are transcribed immediately upon infection by the host RNA polymerase.
An **immune response** is an organism's reaction, mediated by components of the immune system, to an antigen.
Immunity in phages refers to the ability of a prophage to prevent another phage of the same type from infecting a cell. It results from the synthesis of phage repressor by the prophage genome.
The **immunity region** is a segment of the phage genome that enables a prophage to inhibit additional phage of the same type from infecting the bacterium. This region has a gene that encodes for the repressor, as well as the sites to which the repressor binds.
An **immunoglobulin** (antibody) is a class of protein that is produced by B cells in response to antigen.
Importins are transport receptors that bind cargo molecules in the cytoplasm and translocate into the nucleus, where they release the cargo.
Imprecise excision occurs when the transposon removes itself from the original insertion site, but leaves behind some of its sequence.
Imprinting describes a change in a gene that occurs during passage through the sperm or egg with the result that the paternal and maternal alleles have different properties in the very early embryo. May be caused by methylation of DNA.
***In situ* hybridization** (cytological hybridization) is performed by denaturing the DNA of cells squashed on a microscope slide so that reaction is possible with an added single-stranded RNA or DNA; the added preparation is radioactively labeled and its hybridization is followed by autoradiography.
***In vitro* complementation** is a functional assay used to identify components of a process. The reaction is reconstructed using extracts from a mutant cell. Fractions from wild-type cells are then tested for restoration of activity.
Incision is a step in a mismatch excision repair system. An endonuclease recognizes the damaged area in the DNA, and isolates it by cutting the DNA strand on both sides of the damage.
Incomplete dominance is a state in which the heterozygote has a phenotype in between that of each of the homozygotes.
Mendel's law of **independent assortment** states that the assortment of one gene does not influence the assortment of another.
Indirect end labeling is a technique for examining the organization of DNA by making a cut at a specific site and isolating all fragments containing the sequence adjacent to one side of the cut; it reveals the distance from the cut to the next break(s) in DNA.
Induced mutations result from the action of a mutagen. The mutagen may act directly on the bases in DNA or it may act indirectly to trigger a pathway that leads to a change in DNA sequence.
An **inducer** is a small molecule that triggers gene transcription by binding to a regulator protein.
Inducer exclusion describes the inhibition of uptake of other carbon sources into the cell that is caused by uptake of glucose.
An **inducible** operon is expressed only in the presence of a specific small molecule (the inducer).
Induction of prophage describes its entry into the lytic (infective) cycle as a result of destruction of the lysogenic repressor, which leads to excision of free phage DNA from the bacterial chromosome.
Induction refers to the ability of bacteria (or yeast) to synthesize certain enzymes only when their substrates are present; applied to gene expression, it refers to switching on transcription as a result of interaction of the inducer with the regulator protein.
Initiation describes the stages of transcription up to synthesis of the first bond in RNA. This includes binding of RNA polymerase to the promoter and melting a short region of DNA into single strands.

The **initiation codon** is a special codon (usually AUG) used to start synthesis of a protein.

An **initiation complex** in bacterial protein synthesis contains a small ribosome subunit, initiation factors, and initiator aminoacyl-tRNA bound to mRNA at an AUG initiation codon.

Initiation factors (IF in prokaryotes, eIF in eukaryotes) are proteins that associate with the small subunit of the ribosome specifically at the stage of initiation of protein synthesis.

Innate immunity is the rapid response mediated by cells with nonvarying (germline-encoded) receptors that recognize pathogen. The cells of the innate immune response act to eliminate the pathogen and initiate the adaptive immune response.

The **inner core** is an intermediate in the synthesis of N-linked oligosaccharides. It is produced upon the removal of mannose residues from a high mannose oligosaccharide in the *cis* Golgi and is resistant to degradation by endoglycosidase H.

The **Inr** is the sequence of a pol II promoter between -3 and +5 and has the general sequence Py_2CAPy_5. It is the simplest possible pol II promoter.

An **insertion** is identified by the presence of an additional stretch of base pairs in DNA.

An **insertion sequence** (IS) is a small bacterial transposon that carries only the genes needed for its own transposition.

A gene or protein that plays an **instructive** role in development is one that gives a signal telling the cell what to do.

An **insulator** is a sequence that prevents an activating or inactivating effect passing from one side to the other.

An **intasome** is a protein-DNA complex between the phage lambda integrase (Int) and the phage lambda attachment site (*attP*).

An **integrant** (stable transfectant) is a cell line in which a gene introduced by transfection has become integrated into the genome.

An **integrase** is an enzyme that is responsible for a site-specific recombination that inserts one molecule of DNA into another.

Integration of viral or another DNA sequence describes its insertion into a host genome as a region covalently linked on either side to the host sequences.

An **intein** is the part that is removed from a protein that is processed by protein splicing.

Interallelic complementation (intragenic complementation) describes the change in the properties of a heteromultimeric protein brought about by the interaction of subunits coded by two different mutant alleles; the mixed protein may be more or less active than the protein consisting of subunits only of one or the other type.

Interbands are the relatively dispersed regions of polytene chromosomes that lie between the bands.

The **intercistronic region** is the distance between the termination codon of one gene and the initiation codon of the next gene.

Intermediate component(s) of a reassociation reaction are those reacting between the fast (satellite DNA) and slow (nonrepetitive DNA) components; contain moderately repetitive DNA.

Internalization is a process through which a ligand-receptor complex is brought into the cell.

Interphase is the period between mitotic cell divisions; divided into G1, S, and G2.

Interspersed repeats were originally defined as short sequences that are common and widely distributed in the genome. They are now known to consist of transposable elements.

Intrinsic terminators are able to terminate transcription by bacterial RNA polymerase in the absence of any additional factors.

An **intron** (intervening sequence) is a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it.

Intron definition describes the process when a pair of splicing sites are recognized by interactions involving only the 5' site and the branchpoint/3' site.

Intron homing describes the ability of certain introns to insert themselves into a target DNA. The reaction is specific for a single target sequence.

Invariant base positions in tRNA have the same nucleotide in virtually all (>95%) of tRNAs.

Inverted terminal repeats are the short related or identical sequences present in reverse orientation at the ends of some transposons.

An **ion channel** is a transmembrane protein which selectively allows the passage of one type of ion across the membrane. Ion channels are usually oligomers with a central aqueous pore through which the ion passes.

Ion selectivity refers to the specificity of an ion channel for a particular type of ion.

Isoschizomers are restriction enzymes that cleave the same DNA sequence but are affected differently by its state of methylation.

The J **segment** is a polypeptide that is integral to the assembly of dimeric IgA and pentameric IgM. It forms disulfide bonds with the immunoglobulin heavy chain.

A **joint molecule** is a pair of DNA duplexes that are connected together through a reciprocal exchange of genetic material.

A **kilobase** (kb) is a measure of length and may be used to refer to DNA (1000 base pairs) or to RNA (1000 bases).

Kinetic proofreading describes a proofreading mechanism that depends on incorrect events proceeding more slowly than correct events, so that incorrect events are reversed before a subunit is added to a polymeric chain.

The **kinetochore** is the structural feature of the chromosome to which microtubules of the mitotic spindle attach. Its location determines the centromeric region.

Kirromycin is an antibiotic that inhibits protein synthesis by acting on EF-Tu.

A **knot** in the DNA is an entangled region that cannot be resolved without cutting and rearranging the DNA.

Kuru is a human neurological disease caused by prions. It may be caused by eating infected brains.

The **lagging strand** of DNA must grow overall in the 3'-5' direction and is synthesized discontinuously in the form of short fragments (5′-3′) that are later connected covalently.

Lampbrush chromosomes are the extremely extended meiotic bivalents of certain amphibian oocytes.

The **large subunit** of the ribosome (50S in bacteria, 60S in eukaryotes) has the peptidyl transferase active site that synthesizes the peptide bond.

The **lariat** is an intermediate in RNA splicing in which a circular structure with a tail is created by a 5'-2' bond.

A **late endosome** is the part of the endosomal compartment in which endocytosed molecules appear after 5 to 10 minutes. Late endosomes are located close to the nucleus, function in delivering molecules to lysosomes, and are more acidic than early endosomes.

Late genes are transcribed when phage DNA is being replicated. They code for components of the phage particle.

Late infection is the part of the phage lytic cycle from DNA replication to lysis of the cell. During this time, the DNA is replicated and structural components of the phage particle are synthesized.

A **lateral element** is a structure in the synaptonemal complex. It is an axial element that is aligned with the axial elements of other chromosomes.

The **leader** of a protein is a short N-terminal sequence responsible for initiating passage into or through a membrane.

The **leader** (5' UTR) of an mRNA is the nontranslated sequence at the 5' end that precedes the initiation codon.

The **leader peptide** is the product that would result from translation of a short coding sequence used to regulate expression of the tryptophan by controlling ribosome movement.

The **leading strand** of DNA is synthesized continuously in the 5′-3′ direction.

Leaky mutations leave some residual function, either because the mutant protein is partially active (in the case of a missense mutation), or because a small amount of wild-type protein is made (in the case of a nonsense mutation).

The **leucine zipper** is a dimerization motif adjacent to a basic DNA-binding region that is found in a class of transcription factors.

The **leucine-rich region** (LLR) is a motif found in the extracellular domains of some surface receptor proteins in animal and plant cells.

A **library** is a set of cloned fragments together representing the entire genome (genomic library) or all the expressed genes (cDNA library).

A **licensing factor** is something in the nucleus that is necessary for replication, and is inactivated or destroyed after one round of replication. New licensing factors must be provided for further rounds of replication to occur.

A **ligand** is an extracellular molecule that binds to the receptor on the plasma membrane of a cell, thereby effecting a change in the cytoplasm.

Ligand-gated channels open or close in response to the binding of a specific molecule.

The immunoglobulin **light chain** is one of two types of polypeptides in an antibody. Each antibody contains two light chains. The N-terminus of the light chain forms part of the antigen recognition site.

Linkage describes the tendency of genes to be inherited together as a result of their location on the same chromosome; measured by percent recombination between loci.

A **linkage group** includes all loci that can be connected (directly or indirectly) by linkage relationships; equivalent to a chromosome.

A **linkage map** is a map showing the linear order of genes on a chromosome and the relative distances between them in recombinational units.

Linker DNA is all DNA contained on a nucleosome in excess of the 146 bp core DNA.

The **linking number** is the number of times the two strands of a closed DNA duplex cross over each other.

The **linking number paradox** describes the discrepancy between the existence of -2 supercoils in the path of DNA on the nucleosome compared with the measurement of -1 supercoil released when histones are removed.

A **lipid bilayer** is a structure formed by phospholipids in an aqueous solution. The structure consists of two sheets of phospholipids, in which the hydrophilic phosphate groups face the aqueous solution and the hydrophobic tails face each other.

Lipid trafficking is the movement of lipids among the various membranes of a eukaryotic cell.

A **lipopolysaccharide** (LPS) is a molecule containing both lipid and sugar components. It is present in the outer membrane of Gram-negative bacteria. It is also an endotoxin responsible for inducing septic shock during an infection.

A **locus** is the position on a chromosome at which the gene for a particular trait resides; a locus may be occupied by any one of the alleles for the gene.

The **locus control region** (LCR) that is required for the expression of several genes in a domain.

The **long terminal repeat** (LTR) is the sequence that is repeated at each end of the integrated retroviral genome.

A **loop** is a single-stranded region at the end of a hairpin in RNA (or single-stranded DNA); it corresponds to the sequence between inverted repeats in duplex DNA.

A **loose binding site** is any random sequence of DNA that is bound by the core RNA polymerase when it is not engaged in transcription.

A **loss-of-function** mutation inactivates a gene. It is recessive.

The **lumen** describes the interior of a compartment bounded by a membrane, usually the endoplasmic reticulum or the Golgi apparatus.

Luxury genes are those coding for specialized functions synthesized (usually) in large amounts in particular cell types.

Lysis describes the death of bacteria at the end of a phage infective cycle when they burst open to release the progeny of an infecting phage (because phage enzymes disrupt the bacterium's cytoplasmic membrane or cell wall). The same term also applies to eukaryotic cells; for example, when infected cells are attacked by the immune system.

Lysogeny describes the ability of a phage to survive in a bacterium as a stable prophage component of the bacterial genome.

Lysosomes are organelles that contain hydrolytic enzymes. Their primary function is the degradation of ingested materials for recycling.

Lytic infection of a bacterium by a phage ends in the destruction of the bacterium with release of progeny phage.

M phase kinase (MPF) was originally called the maturation promoting factor (or M phase-promoting factor). It is a dimeric kinase, containing the p34 catalytic subunit and a cyclin regulatory subunit, whose activation triggers the onset of mitosis.

A **maintenance methylase** adds a methyl group to a target site that is already hemimethylated.

The **major groove** of DNA is 22A across.

The **major histocompatibility complex** (MHC) is a chromosomal region containing genes that are involved in the immune response. The genes encode proteins for antigen presentation, cytokines, and complement proteins. The MHC is highly polymorphic.

Map distance is measured as cM (centiMorgans) = percent recombination (sometimes subject to adjustments).

A **MAP kinase** (MAPK) is a Ser/Thr protein kinase named for its original identification as a mitogen-activated kinase. There is a large group of cytosolic Thr/Ser protein kinases that form several signaling pathways. The name reflects their original isolation as mitogen-activated protein kinases.

A **map unit** is the distance between two genes that recombine with a frequency of 1%.

A **marker** is any allele of interest in an experiment.

A **maternal gene** is expressed by the mother during oogenesis. A maternal somatic gene is expressed in a somatic cell of the mother, whereas a maternal germline gene is expressed in the germline (e.g. the oocyte).

Maternal inheritance describes the preferential survival in the progeny of genetic markers provided by one parent.

The **mating type** is a property of haploid yeast that makes it able to fuse to form a diploid only with a cell of the opposite mating type.

A **matrix attachment site** (MAR) is a region of DNA that attaches to the nuclear matrix. It is also known as a scaffold attachment site **(SAR).**

Mediator is a large protein complex associated with yeast bacterial RNA polymerase II. It contains factors that are necessary for transcription from many or most promoters.

A **megabase** (Mb) is 1 million base pairs of DNA.

A **memory cell** is a lymphocyte that has been stimulated during the primary immune response to antigen and that is rapidly activated upon subsequent exposure to that antigen. Memory cells respond more rapidly to antigen than naive cells.

Messenger RNA (mRNA) is the intermediate that represents one strand of a gene coding for protein. Its coding region is related to the protein sequence by the triplet genetic code.

Metastasis describes the ability of tumor cells to leave their site of origin and migrate to other locations in the body, where a new colony is established.

Methotrexate is a drug that inhibits the enzyme DHFR (dihydrofolate reductase).

A **methyltransferase** (methylase) is an enzyme that adds a methyl group to a substrate, which can be a small molecule, a protein, or a nucleic acid.

An **MHC class I** protein mostly presents, to $CD8^+$ T cells, peptides that are produced by proteolytic degradation in the cytosol.

An **MHC class II** protein mostly presents, to $CD4^+$ T cells, peptides that are produced by proteolytic degradation in the endocytic pathway.

Micrococcal nuclease is an endonuclease that cleaves DNA; in chromatin, DNA is cleaved preferentially between nucleosomes.

MicroRNAs are very short RNAs that may regulate gene expression.

Microsatellite DNAs consist of repetitions of extremely short (typically <10 bp) units.

A **microtubule** is a filament consisting of dimers of tubulin. It serves as a track for the movement of chromosomes during cell division and for the movements of vesicles and organelles in nondividing cells. It is also a dynamic component of cilia and flagella.

A **microtubule organizing center** (MTOC) is a region from which microtubules emanate. The major MTOCs in a mitotic cell are the centrosomes.

The **midbody** is the last connection between two cells as they separate at the end of cytokinesis. A parallel array of microtubules derived from the mitotic spindle enters the midbody from each cell, and the two arrays interdigitate across its whole width.

Middle genes are phage genes that are regulated by the proteins coded by early genes. Some proteins coded by middle genes catalyze replication of the phage DNA; others regulate the expression of a later set of genes.

A **minicell** is an anucleate bacterial *(E. coli)* cell produced by a division that generates a cytoplasm without a nucleus.

The **minichromosome** of SV40 or polyoma is the nucleosomal form of the viral circular DNA.

Minisatellite DNAs consist of ~10 copies of a short repeating sequence. the length of the repeating unit is measured in 10s of base pairs. The number of repeats varies between individual genomes.

The **minor groove** of DNA is 12A across.

Minus strand DNA is the single-stranded DNA sequence that is complementary to the viral RNA genome of a plus strand virus.

A **mismatch** describes a site in DNA where the pair of bases does not conform to the usual G-C or A-T pairs. It may be caused by incorporation of the wrong base during replication or by mutation of a base.

Mismatch repair is a DNA repair mechanism that fixes bases that do not pair properly. The mechanism preferentially corrects the sequence of the daughter strand. by distingishing the daughter strand and parental strand on the basis of their states of methylation.

Missense mutations change a single codon and so may cause the replacement of one amino acid by another in a protein sequence.

A **missense suppressor** codes for a tRNA that has been mutated so as to recognize a different codon. By inserting a different amino acid at a mutant codon, the tRNA suppresses the effect of the original mutation.

Mitochondrial DNA (mtDNA) is an independent DNA genome, usually circular, that is located in the mitochondrion.

Mitosis (M phase) is the process by which the cell nucleus divides, resulting in daughter cells that contain the same amount of genetic material as the parent cell.

A **mitotic cyclin** (G2 cyclin) is a regulatory subunit that partners a kinase subunit to form the M phase inducer.

Modification of DNA or RNA includes all changes made to the nucleotides after their initial incorporation into the polynucleotide chain.

Modified bases are all those except the usual four from which DNA (T, C, A, G) or RNA (U, C, A, G) are synthesized; they result from postsynthetic changes in the nucleic acid.

Monocistronic mRNA codes for one protein.

A **monolayer** describes the growth of eukaryotic cells in culture as a layer only one cell deep.

Monster particles of bacteriophages form as the result of an assembly defect in which the capsid proteins form a head that is much longer than usual.

A **morphogen** is a factor that induces development of particular cell types in a manner that depends on its concentration.

A plasmid is said to be under **multicopy control** when the control system allows the plasmid to exist in more than one copy per individual bacterial cell.

Multiforked chromosome (in bacterium) has more than one replication fork, because a second initiation has occurred before the first cycle of replication has been completed.

A locus is said to have **multiple alleles** when more than two allelic forms have been found. Each allele may cause a different phenotype.

Mutagens increase the rate of mutation by inducing changes in DNA sequence, directly or indirectly.

A **mutator** is a gene whose mutation results in an increase in the basal level of mutation of the genome. Such genes are often code for proteins that are involved in repairing damaged DNA.

An N nucleotide sequence is a short non-templated sequence that is added randomly by the enzyme at coding joints during rearrangement of immunoglobulin and T cell receptor genes. N nucleotides augment the diversity of antigen receptors.

The **n-1 rule** states that only one X chromosome is active in female mammalian cells; any other(s) are inactivated.

N-formyl-methionyl-tRNA ($tRNA_f^{Met}$) is the aminoacyl-tRNA that initiates bacterial protein synthesis. The amino group of the methionine is formylated.

Nascent RNA is a ribonucleotide chain that is still being synthesized, so that its 3' end is paired with DNA where RNA polymerase is elongating.

Negative complementation occurs when interallelic complementation allows a mutant subunit to suppress the activity of a wild-type subunit in a multimeric protein.

The default state of genes that are controlled by **negative regulation** is to be expressed. A specific intervention is required to turn them off.

A **neutral** mutation has no effect on the genotype and does not affect natural selection.

Neutral substitutions in a protein cause changes in amino acids that do not affect activity.

Nick translation describes the ability of *E. coli* DNA polymerase I to use a nick as a starting point from which one strand of a duplex DNA can be degraded and replaced by resynthesis of new material; is used to introduce radioactively labeled nucleotides into DNA *in vitro.*

Non-homologous end-joining (NHEJ) ligates blunt ends. It is common to many repair pathways and to certain recombination pathways (such as immunoglobulin recombination).

Nonallelic genes are two (or more) copies of the same gene that are present at *different* locations in the genome (contrasted with alleles which are copies of the same gene derived from different parents and present at the same location on the homologous chromosomes).

A **nonautonomous controlling element** is a transposon in maize that encodes a non-functional transposase; it can transpose only in the presence of a *trans-acting* autonomous member of the same family.

A **nondefective virus** describes a transforming retrovirus that has all the normal capabilities in replication etc. Its ability to transform depends on its effects on expression of host genes at its site of insertion into the cellular genome.

A **nonhistone** is any structural protein found in a chromosome except one of the histones.

Nonpermissive conditions do not allow conditional lethal mutants to survive.

The recombination of V, (D), J gene segments results in a **nonproductive rearrangement** if the rearranged gene segments are not in the correct reading frame. A nonproductive rearrangement occurs when nucleotide addition or subtraction disrupts the reading frame or when.

Nonreciprocal recombination results from an error in pairing and crossing-over in which nonequivalent sites are involved in the two reacting genomes. It produces one recombinant with a deletion of material and one with a duplication.

Nonrepetitive DNA shows reassociation kinetics expected of unique sequences.

Nonreplicative transposition describes the movement of a transposon that leaves a donor site (usually generating a double-strand break) and moves to a new site.

A **nonsense mutation** is any change in DNA that causes a (termination) codon to replace a codon representing an amino acid.

A **nonsense suppressor** is a gene coding for a mutant tRNA able to respond to one or more of the termination codons.

Nonsense-mediated mRNA decay is a pathway that degrades an mRNA that has a nonsense codon prior to the last exon.

Nontranscribed spacer is the region between transcription units in a tandem gene cluster.

The **nonviral superfamily** of transposons originated independently of retroviruses.

Nopaline plasmids are Ti plasmids of *Agrobacterium tumefaciens* that carry genes for synthesizing the opine, nopaline. They retain the ability to differentiate into early embryonic structures.

The **nuclear envelope** is a layer of two concentric membranes (inner and outer nuclear membranes) that surrounds the nucleus and its underlying intermediate filament lattice, the nuclear lamina. The nuclear envelope is penetrated by nuclear pores. The outer membrane is continuous with the membrane of the rough endoplasmic reticulum.

A **nuclear export signal** (NES) is a domain of a protein, usually a short amino acid sequence, which interacts with an exportin, resulting in the transport of the protein from the nucleus to the cytoplasm.

A **nuclear localization signal** (NLS) is a domain of a protein, usually a short amino acid sequence, that interacts with an importin, allowing the protein to be transported into the nucleus.

A **nuclear pore complex** (NPC) is a very large, proteinaceous structure that extends through the nuclear envelope, providing a channel for bidirectional transport of molecules and macromolecules between the nucleus and the cytosol.

The **nucleation center** of TMV (tobacco mosaic virus) is a duplex hairpin where assembly of coat protein with RNA is initiated.

Nucleic acids are molecules that encode genetic information. They consist of a series of nitrogenous bases connected to ribose molecules that are linked by phosphodiester bonds. DNA is deoxyribonucleic acid, and RNA is ribonucleic acid.

The **nucleoid** is the region in a prokaryotic cell that contains the genome. The DNA is bound to proteins and is not enclosed by a membrane.

The **nucleolar organizer** is the region of a chromosome carrying genes coding for rRNA.

The **nucleolus** (*pl.* nucleoli) is a discrete region of the nucleus where ribosomes are produced.

Nucleoporin was originally defined to describe the components of the nuclear pore complex that bind to the inhibitory lectins, but now is used to mean any component of the basic nuclear pore complex.

The **nucleosome** is the basic structural subunit of chromatin, consisting of ~200 bp of DNA and an octamer of histone proteins.

Nucleosome positioning describes the placement of nucleosomes at defined sequences of DNA instead of at random locations with regards to sequence.

A **null** mutation completely eliminates the function of a gene, usually because it has been physically deleted.

The **ochre** codon is the triplet UAA, one of the three termination codons that end protein synthesis.

Octopine plasmids of *Agrobacterium tumefaciens* carry genes coding the synthesis of opines of the octopine type. The tumors are undifferentiated.

Okazaki fragments are the short stretches of 1000-2000 bases produced during discontinuous replication; they are later joined into a covalently intact strand.

Oncogenes are genes whose products have the ability to transform eukaryotic cells so that they grow in a manner analogous to tumor cells. Oncogenes carried by retroviruses have names of the form *v-onc.*

The **opal** codon is the triplet UGA, one of the three termination codons that end protein synthesis. It has evolved to code for an amino acid in a small number of organisms or organelles.

An **open complex** describes the stage of initiation of transcription when RNA polymerase causes the two strands of DNA to separate to form the "transcription bubble".

An **open reading frame** (ORF) is a sequence of DNA consisting of triplets that can be translated into amino acids starting with an initiation codon and ending with a termination codon.

The **operator** is the site on DNA at which a repressor protein binds to prevent transcription from initiating at the adjacent promoter.

An **operon** is a unit of bacterial gene expression and regulation, including structural genes and control elements in DNA recognized by regulator gene product(s).

An **opine** is a derivative of arginine that is synthesized by plant cells infected with crown gall disease.

The **origin** is a sequence of DNA at which replication is initiated.

Orthologs are corresponding proteins in two species as defined by sequence homologies.

A stretch of **overwound** DNA has more base pairs per turn than the usual average (10 bp = 1 turn). This means that the two strands of DNA are more tightly wound around each other, creating tension.

A **P element** is a type of transposon in *D. melanogaster.*

A **P nucleotide** sequence is a short palindromic (inverted repeat) sequence that is generated during rearrangement of immunoglobulin and T cell receptor V, (D), J gene segments. P nucleotides are generated at coding joints when RAG proteins cleave the hairpin ends generated during rearrangement.

The P **site** of the ribosome is the site that is occupied by peptidyl-tRNA, the tRNA carrying the nascent polypeptide chain, still paired with the codon to which it bound in the A site.

The **packing ratio** is the ratio of the length of DNA to the unit length of the fiber containing it.

In *Drosophila* the **pair-rule** genes are a set of genes that help set up the segmentation of the embryo. They are expressed in a striped pattern with one stripe in every other future segment.

A **palindrome** is a DNA sequence that reads the same on each strand of DNA when the strand is read in the 5' to 3' direction. It consists of adjacent inverted repeats.

Paralogs are highly similar proteins that are coded by the same genome.

A **paranemic joint** describes a region in which two complementary sequences of DNA are associated side by side instead of being intertwined in a double helical structure.

In *Drosophila,* a **parasegment** is a unit composed of the rear of one segment and the front of the adjacent segment.

A **parental** strand or duplex of DNA refers to the DNA that will be replicated.

A **parental genotype** is one that is identical to the genotype of one of the contributing parents.

Passive transport describes movement of molecules along their electrochemical gradient; no energy is required.

Patch recombinant DNA results from a Holliday junction being resolved by cutting the exchanged strands. The duplex is largely unchanged, except for a DNA sequence on one strand that came from the homologous chromosome.

A **pathogen-associated molecular pattern** (PAMP) is a molecular structure on the surface of a pathogen. A given PAMP may be conserved across a large number of pathogens. During an immune response, PAMPs may be recognized by receptors on cells that mediate innate immunity.

Peptidyl transferase is the activity of the ribosomal 50S subunit that synthesizes a peptide bond when an amino acid is added to a growing polypeptide chain. The actual catalytic activity is a property of the rRNA.

Peptidyl-tRNA is the tRNA to which the nascent polypeptide chain has been transferred following peptide bond synthesis during protein synthesis.

The **periplasm** (or periplasmic space) is the region between the inner and outer membranes in the bacterial envelope.

A **periseptal annulus** is an ring-like area where inner and outer membrane appear fused. Formed around the circumference of the bacterium, the periseptal annulus determines the location of the septum.

A protein that plays a **permissive** role in development is one that sets up a situation where a certain activity can occur, but does not cause the occurrence itself.

Peroxins are the protein components of the peroxisome.

The **peroxisome** is an organelle in the cytoplasm enclosed by a single membrane. It contains oxidizing enzymes.

Phage T4 is a virus that infects E. coli causing lysis of the bacterium.

A **pheromone** is a small molecule secreted by one mating type of an organism in order to interact with a member of the opposite mating type.

A **phospholipid** is a lipid that has a positively charged head that is linked by a phosphate group to the fatty acid tails.

A **phosphorelay** describes a pathway in which a phosphate group is passed along a series of proteins.

Photoreactivation is a type of direct repair that involves a light-dependent enzyme.

Pilin is the subunit that is polymerized into the pilus in bacteria.

A **pilus** (pili) is a surface appendage on a bacterium that allows the bacterium to attach to other bacterial cells. It appears like a short, thin, flexible rod. During conjugation, pili are used to transfer DNA from one bacterium to another.

A **plaque** is an area of clearing in a bacterial lawn. It is created by a single phage particle that has undergone multiple rounds of lytic growth.

A **plasmid** is a circular, extrachromosomal DNA. It is autonomous and can replicate itself.

A **plectonemic** joint is a region that consists of one molecule wound around another molecule, e.g. the DNA strands in a double helix.

Plus strand DNA is the strand of the duplex sequence representing a retrovirus that has the same sequence as that of the RNA.

A **plus strand virus** has a single-stranded nucleic acid genome whose sequence directly codes for the protein products.

Point mutation is a change in the sequence of DNA involving a single base pair.

Polarity refers to the effect of a mutation in one gene in influencing the expression (at transcription or translation) of subsequent genes in the same transcription unit.

Poly(A) is a stretch of ~200 bases of adenylic acid that is added to the 3' end of mRNA following its synthesis.

Poly(A) polymerase is the enzyme that adds the stretch of polyadenylic acid to the 3' of eukaryotic mRNA. It does not use a template.

Poly(A)-binding protein (PABP) is the protein that binds to the 3' stretch of poly(A) on a eukaryotic mRNA.

poly(A)$^+$ mRNA is mRNA that has a 3' terminal stretch of poly(A).

Polycistronic mRNA includes coding regions representing more than one gene.

Polymorphism (more fully genetic polymorphism) refers to the simultaneous occurrence in the population of genomes showing variations at a given position. The original definition applied to alleles producing different phenotypes. Now it is also used to describe changes in DNA affecting the restriction pattern or even the sequence. For practical purposes, to be considered as an example of a polymorphism, an allele should be found at a frequency > 1 % in the population.

A **polyribosome** (polysome) is an mRNA that is simultaneously being translated by several ribosomes.

Polytene chromosomes are generated by successive replications of a chromosome set without separation of the replicas.

Position effect variegation (PEV) is silencing of gene expression that occurs as the result of proximity to heterochromatin.

Positional information describes the localization of macromolecules at particular places in an embryo. The localization may itself be a form of information that is inherited.

Post-translational translocation (posttranslational) is the movement of a protein across a membrane after the synthesis of the protein is completed and it has been released from the ribosome.

In *Drosophila*, the **posterior system** is one of the maternal systems that establishes the polarity of the oocyte. The set of genes in the posterior system play a role in the proper formation of the pole plasm and the abdomen.

Postmeiotic segregation describes the segregation of two strands of a duplex DNA that bear different information (created by heteroduplex formation during meiosis) when a subsequent replication allows the strands to separate.

The **postreplication complex** is a protein-DNA complex in *S. cerevisiae* that consists of the ORC complex bound to the origin.

ppGpp is guanosine tetraphosphate. Diphosphate groups are attached to both the 5' and 3' positions.

pppGpp is a guanosine pentaphosphate, with a triphosphate attached to the 5' position and a diphosphate attached to the 3' position.

Pre-mRNA is used to describe the nuclear transcript that is processed by modification and splicing to give an mRNA.

Precise excision describes the removal of a transposon plus one of the duplicated target sequences from the chromosome. Such an event can restore function at the site where the transposon inserted.

Preinitiation complex in eukaryotic transcription describes the assembly of transcription factors at the promoter before RNA polymerase binds.

Premature termination describes the termination of protein or of RNA synthesis before the chain has been completed. In protein synthesis it can be caused by mutations that create termination codons within the coding region. In RNA synthesis it is caused by various events that act on RNA polymerase.

A protein to be imported into an organelle or secreted from bacteria is called a **"preprotein"** until its signal sequence has been removed.

The **prereplication complex** is a protein-DNA complex at the origin in *S. cerevisiae* that is required for DNA replication. The complex contains the ORC complex, Cdc6, and the MCM proteins.

Primary cells are eukaryotic cells taken into culture directly from the animal.

The **primary immune response** is an organism's immune response upon first exposure to a given antigen. It is characterized by a relatively shorter duration and lower affinity antibodies than in the secondary immune response.

A **primary transcript** is the original unmodified RNA product corresponding to a transcription unit.

The **primase** is a type of RNA polymerase that synthesizes short segments of RNA that will be used as primers for DNA replication.

A primer is a short sequence (often of RNA) that is paired with one strand of DNA and provides a free 3′-OH end at which a DNA polymerase starts synthesis of a deoxyribonucleotide chain.
The **primosome** describes the complex of proteins involved in the priming action that initiates replication on ϕX-type origins. It is also involved in restarting stalled replication forks.
A **prion** is a proteinaceous infectious agent, which behaves as an inheritable trait, although it contains no nucleic acid. Examples are PrP^{Sc}, the agent of scrapie in sheep and bovine spongiform encephalopathy, and Psi, which confers an inherited state in yeast.
The **procentriole** is an immature centriole, formed in the vicinity of a mature centriole.
A **processed pseudogene** is an inactive gene copy that lacks introns, contrasted with the interrupted structure of the active gene. Such genes originate by reverse transcription of mRNA and insertion of a duplex copy into the genome.
Processing of RNA describes changes that occur after its transcription, including modification of the 5′ and 3′ ends, internal methylation, splicing, or cleavage.
Processivity describes the ability of an enzyme to perform multiple catalytic cycles with a single template instead of dissociating after each cycle.
The recombination of V, (D), J gene segments results in a **productive rearrangement** if all the rearranged gene segments are in the correct reading frame.
Programmed frameshifting is required for expression of the protein sequences coded beyond a specific site at which a +1 or -1 frameshift occurs at some typical frequency.
A **promoter** is a region of DNA where RNA polymerase binds to initiate transcription.
Proofreading refers to any mechanism for correcting errors in protein or nucleic acid synthesis that involves scrutiny of individual units *after* they have been added to the chain.
Prophage is a phage genome covalently integrated as a linear part of the bacterial chromosome.
The **proteasome** is a large complex with an interior cavity that degrades cytosolic proteins previously marked by covalent addition of ubiquitin.
A **protein kinase** is a protein which transfers the terminal phosphate group from ATP onto another protein.
A **protein serine/threonine kinase** phosphorylates cytosolic proteins on either their serine or threonine residues.
Protein sorting (targeting) is the direction of different types of proteins for transport into or between specific organelles.
Protein splicing is the autocatalytic process by which an intein is removed from a protein and the exteins on either side become connected by a standard peptide bond.
A **protein tyrosine kinase** is a kinase enzyme whose target is a tyrosine amino acid in a protein.
The **proteome** is the complete set of proteins that is expressed by the entire genome. Because some genes code for multiple proteins, the size of the proteome is greater than the number of genes. Sometimes the term is used to describe complement of proteins expressed by a cell at any one time.
Proto-oncogenes are the normal counterparts in the eukaryotic genome to the *(v-onc)* oncogenes carried by some retroviruses. They are given names of the form *c-onc.*
Provirus is a duplex sequence of DNA integrated into a eukaryotic genome that represents the sequence of the RNA genome of a retrovirus.
PrP is the protein that is the active component of the prion that causes scrapie and related diseases. The form involved in the disease is called PrP^{Sc}.
Pseudogenes are inactive but stable components of the genome derived by mutation of an ancestral active gene. Usually they are inactive because of mutations that block transcription or translation or both.
A **puff** is an expansion of a band of a polytene chromosome associated with the synthesis of RNA at some locus in the band.
Puromycin is an antibiotic that terminates protein synthesis by mimicking a tRNA and becoming linked to the nascent protein chain.
A **pyrimidine dimer** is formed when ultraviolet irradiation generates a covalent link directly between two adjacent pyrimidine bases in DNA. It blocks DNA replication.

A **quick-stop mutant** is a type of DNA replication temperature-sensitive mutant *(dna)* in *E. coli* that immediately stops DNA replication when the temperature is increased to 42°C.

The R segments are the sequences that are repeated at the ends of a retroviral RNA. They are called R-U5 and U3-R.
An **r-protein** is one of the proteins of the ribosome.
Rab proteins make up a family of about 30 small Ras-like GTPases. Different Rabs are required for protein trafficking in different membrane systems. Although their exact role is not clear, Rabs appear to regulate membrane targeting and fusion.
Random drift describes the chance fluctuation (without selective pressure) of the levels of two alleles in a population.
Rapid lysis (r) mutants display a change in the pattern of lysis of *E. coli* at the end of an infection by a T-even phage.
A **reading frame** is one of the three possible ways of reading a nucleotide sequence. Each reading frame divides the sequence into a series of successive triplets. There are three possible reading frames in any sequence, depending on the starting point. If the first frame starts at position 1, the second frame starts at position 2, and the third frame starts at position 3.
Readthrough at transcription or translation occurs when RNA polymerase or the ribosome, respectively, ignores a termination signal because of a mutation of the template or the behavior of an accessory factor.
rec mutations of *E. coli* cannot undertake general recombination.
A **receptor** is a transmembrane protein, located in the plasma membrane, that binds a ligand in a domain on the extracellular side, and as a result has a change in activity of the cytoplasmic domain. (The same term is sometimes used also for the steroid receptors, which are transcription factors that are activated by binding ligands that are steroids or other small molecules.)
A **recessive** allele is obscured in the phenotype of a heterozygote by the dominant allele, often due to inactivity or absence of the product of the recessive allele.
A **reciprocal translocation** exchanges part of one chromosome with part of another chromosome.
Receding events occur when the meaning of a codon or series of codons is changed from that predicted by the genetic code. It may involve altered interactions between aminoacyl-tRNA and mRNA that are influenced by the ribosome.
The **recognition helix** is the one of the two helices of the helix-turn-helix motif that makes contacts with DNA that are specific for particular bases. This determines the specificity of the DNA sequence that is bound.
Recombinant progeny have a different genotype from that of either parent.
A **recombinant genotype** is one that consists of a new combination of genes produced by crossing over.
A **recombinant joint** is the point at which two recombining molecules of duplex DNA are connected (the edge of the heteroduplex region).
Recombination nodules (nodes) are dense objects present on the synaptonemal complex; they may represent protein complexes involved in crossing-over.
Recombination-repair is a mode of filling a gap in one strand of duplex DNA by retrieving a homologous single strand from another duplex.
Redundancy describes the concept that two or more genes may fulfill the same function, so that no single one of them is essential.

A regulator gene codes for a product (typically protein) that controls the expression of other genes (usually at the level of transcription).

A **relaxase** is an enzyme that cuts one strand of DNA, and binds to the free 5' end.

Relaxed mutants of *E. coli* do not display the stringent response to starvation for amino acids (or other nutritional deprivation).

A **release factor** (RF) is required to terminate protein synthesis to cause release of the completed polypeptide chain and the ribosome from mRNA. Individual factors are numbered. Eukaryotic factors are called eRF.

Renaturation describes the reassociation of denatured complementary single strands of a DNA double helix.

Repair of damaged DNA can take place by repair synthesis, when a strand that has been damaged is excised and replaced by the synthesis of a new stretch. It can also take place by recombination reactions, when the duplex region containing the damaged is replaced by an undamaged region from another copy of the genome.

The **repetition frequency** is the (integral) number of copies of a given sequence present in the haploid genome; equals 1 for nonrepetitive DNA, >2 for repetitive DNA.

Repetitive DNA behaves in a reassociation reaction as though many (related or identical) sequences are present in a component, allowing any pair of complementary sequences to reassociate.

Replacement sites in a gene are those at which mutations alter the amino acid that is coded.

Replication of duplex DNA takes place by synthesis of two new strands that are complementary to the parental strands. The parental duplex is replaced by two identical daughter duplexes, each of which has one parental strand and one newly synthesized strand. It is called semiconservative because the conserved units are the single strands of the parental duplex.

A **replication eye** is a region in which DNA has been replicated within a longer, unreplicated region.

A **replication fork** (growing point) is the point at which strands of parental duplex DNA are separated so that replication can proceed. A complex of proteins including DNA polymerase is found at the fork.

A **replication-defective** virus cannot perpetuate an infective cycle because some of the necessary genes are absent (replaced by host DNA in a transducing virus) or mutated.

Replicative transposition describes the movement of a transposon by a mechanism in which first it is replicated, and then one copy is transferred to a new site.

The **replicon** is a unit of the genome in which DNA is replicated. Each replicon contains an origin for initiation of replication.

The **replisome** is the multiprotein structure that assembles at the bacterial replicating fork to undertake synthesis of DNA. It contains DNA polymerase and other enzymes.

A **repressible** operon is expressed unless the small molecule co-repressor is present.

Repression describes the ability of bacteria to prevent synthesis of certain enzymes when their products are present; more generally, refers to inhibition of transcription (or translation) by binding of repressor protein to a specific site on DNA (or mRNA).

A **repressor** is a protein that inhibits expression of a gene. It may act to prevent transcription by binding to an operator site in DNA, or to prevent translation by binding to RNA.

Resolution occurs by a homologous recombination reaction between the two copies of the transposon in a cointegrate. The reaction generates the donor and target replicons, each with a copy of the transposon.

Resolvase is the enzyme activity involved in site-specific recombination between two transposons present as direct repeats in a cointegrate structure.

A **response element** is a sequence in a eukaryotic promoter that is recognized by a specific transcription factor.

Restriction endonucleases recognize specific short sequences of DNA and cleave the duplex (sometimes at target site, sometimes elsewhere, depending on type).

Restriction fragment length polymorphism (RFLP) refers to inherited differences in sites for restriction enzymes (for example, caused by base changes in the target site) that result in differences in the lengths of the fragments produced by cleavage with the relevant restriction enzyme. RFLPs are used for genetic mapping to link the genome directly to a conventional genetic marker.

A **restriction map** is a linear array of sites on DNA cleaved by various restriction enzymes.

A **retention signal** is the part of a protein that prevents it from leaving a compartment. An example is the transmembrane domain of Golgi resident proteins.

Retrograde translocation (reverse translocation) is the translocation of a protein from the lumen of the ER to the cytoplasm. It usually occurs to allow misfolded or damaged proteins to be degraded by the proteasome.

Retrograde transport describes movement of proteins in the reverse direction in the reticuloendothelial system, typically from Golgi to endoplasmic reticulum.

A **retroposon** (retrotransposon) is a transposon that mobilizes via an RNA form; the DNA element is transcribed into RNA, and then reverse-transcribed into DNA, which is inserted at a new site in the genome. The difference from retroviruses is that the retroposon does not have an infective (viral) form.

A **retrovirus** is an RNA virus with the ability to convert its sequence into DNA by reverse transcription.

Reverse transcriptase is an enzyme that uses a template of single-stranded RNA to generate a double-stranded DNA copy.

Reverse transcription is synthesis of DNA on a template of RNA; accomplished by reverse transcriptase enzyme.

Revertants are derived by reversion of a mutant cell or organism.

RF1 is the bacterial release factor that recognizes UAA and UAG as signals to terminate protein synthesis.

RF2 is the bacterial release factor that recognizes UAA and UGA as signals to terminate protein synthesis.

RF3 is a protein synthesis termination factor related to the elongation factor EF-G. It functions to release the factors RF1 or RF2 from the ribosome when they act to terminate protein synthesis.

Rho factor is a protein involved in assisting *E. coli* RNA polymerase to terminate transcription at certain terminators (called rho-dependent terminators).

Rho-dependent terminators are sequences that terminate transcription by bacterial RNA polymerase in the presence of the rho factor.

Ri plasmids are found in *Agrobacterium tumefaciens.* Like Ti plasmids, they carry genes that cause disease in infected plants. The disease may take the form of either hairy root disease or crown gall disease.

Ribonucleases (RNAases) are enzymes that cleave RNA. They may be specific for single-stranded or for double-stranded RNA, and may be either endonucleases or exonucleases.

A **ribonucleoprotein** is a complex of RNA with proteins.

Ribosomal DNA (rDNA) is usually a tandemly repeated series of genes coding for a precursor to the two large rRNAs.

Ribosomal RNA (rRNA) is a major component of the ribosome. Each of the two subunits of the ribosome has a major rRNA as well as many proteins.

The **ribosome** is a large assembly of RNA and proteins that synthesizes proteins under direction from an mRNA template. Bacterial ribosomes sediment at 70S, eukaryotic ribosomes at 80S. A ribosome can be dissociated into two subunits.

Ribosome stalling describes the inhibition of movement that occurs when a ribosome reaches a codon for which there is no corresponding charged aminoacyl-tRNA.

A **ribosome-binding site** is a sequence on bacterial mRNA that includes an initiation codon that is bound by a 30S subunit in the initiation phase of protein synthesis.

A ribozyme is an RNA that has catalytic activity.
A helix is said to be **right-handed** if the turns runs clockwise along the helical axis.
45S **RNA** is a precursor that contains the sequences of both major ribosomal RNAs (28S and 18S rRNAs).
5.8S **RNA** is an independent small RNA present on the large subunit of eukaryotic ribosomes. It is homologous to the 5' end of bacterial 23 S rRNA.
5S **RNA** is a 120 base RNA that is a component of the large subunit of the ribosome.
RNA editing describes a change of sequence at the level of RNA following transcription.
RNA interference describes situations in which antisense and sense RNAs apparently are equally effective in inhibiting expression of a target gene. It is caused by the ability of double-stranded sequences to cause degradation of sequences that are complementary to them.
An **RNA ligase** is an enzyme that functions in tRNA splicing to make a phosphodiester bond between the two exon sequences that are generated by cleavage of the intron.
RNA polymerases are enzymes that synthesize RNA using a DNA template (formally described as DNA-dependent RNA polymerases).
RNA silencing describes the ability of a dsRNA to suppress expression of the corresponding gene systemically in a plant.
RNA splicing is the process of excising the sequences in RNA that correspond to introns, so that the sequences corresponding to exons are connected into a continuous mRNA.
The **rolling circle** is a mode of replication in which a replication fork proceeds around a circular template for an indefinite number of revolutions; the DNA strand newly synthesized in each revolution displaces the strand synthesized in the previous revolution, giving a tail containing a linear series of sequences complementary to the circular template strand.
Rotational positioning describes the location of the histone octamer relative to turns of the double helix, which determines which face of DNA is exposed on the nucleosome surface.
Rough endoplasmic reticulum (rough ER) refers to the region of the endoplasmic reticulum to which ribosomes are bound. It is the site of synthesis of membrane proteins and secretory proteins.
RTK is an abbreviation for "receptor tyrosine kinase". These kinases are membrane-bound proteins with large cytoplasmic and extracellular domains. Specific binding of a ligand, such as a growth factor, to the extracellular domain causes the cytoplasmic domain to phosphorylate other proteins on tyrosine residues.

The S **domain** is the sequence of 7S RNA of the SRP that is not related to Alu RNA.
S **phase** is the restricted part of the eukaryotic cell cycle during which synthesis of DNA occurs.
The **S phase activator** is the cdk-cyclin complex that is responsible for initiating S phase.
An **S region** is an intron sequence involved in immunoglobulin class switching. S regions consist of repetitive sequences at the 5' end of gene segments encoding the heavy chain constant regions.
Satellite DNA (simple-sequence DNA) consists of many tandem repeats (identical or related) of a short basic repeating unit.
A **saturated** fatty acid only has single carbon-carbon bonds in its backbone.
A chromosome **scaffold** is a proteinaceous structure in the shape of a sister chromatid pair, generated when chromosomes are depleted of histones.
Scarce mRNA (complex mRNA) consists of a large number of individual mRNA species, each present in very few copies per cell. This accounts for most of the sequence complexity in RNA.
Scrapie is a infective agent made of protein.
scRNPs (scyrps) are small cytoplasmic ribonucleoproteins (scRNAs associated with proteins).
A **second messenger** is a small molecule that is generated when a signal transduction pathway is activated. The classic second messenger is cyclic AMP, which is generated when adenylate cyclase is activated by a G protein (when the G protein itself was activated by a transmembrane receptor).
A **second messenger gated channel** is an ion channel whose activity is controlled by small signaling molecules inside the cell.
Second-site reversion describes the occurrence of a second mutation that suppresses the effect of a first mutation.
A **secondary attachment site** is a locus on the bacterial chromosome into which phage lambda integrate inefficiently because the site resembles the *att* site.
The **secondary immune response** is an organism's immune response upon a second exposure to a given antigen. This second exposure is also referred to as a "booster". The secondary immune response is characterized by a more rapid induction, greater magnitude, and higher affinity antibodies than the primary immune response.
A **secretory granule** is a membrane-bounded compartment that contains molecules to be released from cells by regulated exocytosis (that is, the molecules are concentrated and stored in secretory granules, and are released only in response to a signal). It is also called a secretory vesicle.
A **secretory vesicle** is a membrane-bounded compartment that contains molecules to be released from cells by regulated exocytosis (that is, the molecules are concentrated in secretory vesicles and are released only in response to a signal). It is also called a secretory granule.
A **sector** is a patch of cells made up of a single altered cell and its progeny.
Securins are a class of proteins that prevent the initiation of anaphase by binding to and inhibiting separin, a protease which cleaves the structural component required for holding sister chromatids together. Inhibition of separin by securin ends when securin is itself proteolyzed as a result of activation of the anaphase promoting complex (APC).
Many organisms have a segmented body plan that divides the body into a number of repeating units, called segments, along the anterior-posterior axis.
In *Drosophila*, **segment polarity** genes are a set of genes that help set up the segmentation of the embryo. They are expressed in a striped pattern with one stripe in every future segment. Each stripe indicates the posterior margin of a segment.
Segmentation genes are concerned with controlling the number or polarity of body segments in insects.
Self-assembly refers to the ability of a protein (or of a complex of proteins) to form its final structure without the intervention of any additional components (such as chaperones). The term can also refer to the spontaneous formation of any biological structure that occurs when molecules collide and bind to each other.
Selfish DNA describes sequences that do not contribute to the genotype of the organism but have self-perpetuation within the genome as their sole function.
Semiconservative replication is accomplished by separation of the strands of a parental duplex, each then acting as a template for synthesis of a complementary strand.
A **semiconserved** (semiinvariant) position is one where comparison of many individual sequences finds the same type of base (pyrimidine or purine) always present.
Semidiscontinuous replication is mode in which one new strand is synthesized continuously while the other is synthesized discontinuously.
Senescent cells show visible changes in the appearance of a culture as the result of limitations posed by telomere shortening on the number of chromosomal replications that can occur.
Separins are proteins which play a direct role in initiating anaphase by cleaving and inactivating a component (a cohesin) that holds sister chromatids together.

The **septal ring** (Z-ring) is a complex of several proteins coded by *fis* genes of *E. coli* that forms at the mid-point of the cell. It gives rise to the septum at cell division. The first of the proteins to be incorporated is FtsZ, which gave rise to the original name of the Z-ring.

A **septum** is the structure that forms in the center of a dividing bacterium, providing the site at which the daughter bacteria will separate. The same term is used to describe the cell wall that forms between plant cells at the end of mitosis.

The -10 **sequence** is the consensus sequence centered about 10 bp before the startpoint of a bacterial gene. It is involved in melting DNA during the initiation reaction.

The -35 **sequence** is the consensus sequence centered about 35 bp before the startpoint of a bacterial gene. It is involved in initial recognition by RNA polymerase.

A **serpentine** receptor has 7 transmembrane segments. Typically it activates a trimeric G protein.

Serum dependence describes the need of eukaryotic cells for factors contained in serum in order to grow in culture.

The **serum response element** (SRE) is a sequence in a promoter or enhancer that is activated by transcription factor(s) induced by treatment with serum. This activates genes that stimulate cell growth.

An **SH2** domain (named originally as the *Src Homology* domain because it was identified in the Src product of the Rous sarcoma virus) is a region of ~100 amino acids that is bound by the SH2-binding domain of the protein upstream in a signal transduction cascade.

An **SH2-binding site** is an area on a protein that interacts with the SH2 domain of another protein.

An **SH3** domain is used by some proteins that contain SH2 domains to enable them to bind to the next component downstream in a signal transduction cascade.

The **Shine-Dalgarno** sequence is part or all of the polypurine sequence TATAATG centered about 10 bp before the AUG initiation codon on bacterial mRNA. It is complementary to the sequence at the 3' end of 16S rRNA.

Shotgun cloning analyzes an entire genome in the form of randomly generated fragments.

Sigma factor is the subunit of bacterial RNA polymerase needed for initiation; it is the major influence on selection of promoters.

The **sign inversion** model describes the mechanism of DNA gyrase. DNA gyrase binds a positive supercoil (inducing a compensatory negative supercoil elsewhere on the closed circular DNA), breaks both strands in one duplex, passes the other duplex through, and reseals the strands.

A **signal end** is produced during recombination of immunoglobulin and T cell receptor genes. The signal ends are at the termini of the cleaved fragment containing the recombination signal sequences. The subsequent joining of the signal ends yields a signal joint.

Signal peptidase is an enzyme within the membrane of the ER that specifically removes the signal sequences from proteins as they are translocated. Analogous activities are present in bacteria, archaebacteria, and in each organelle in a eukaryotic cell into which proteins are targeted and translocated by means of removable targeting sequences. Signal peptidase is one component of a larger protein complex.

The **signal recognition particle** (SRP) is a ribonucleoprotein complex that recognizes signal sequences during translation and guides the ribosome to the translocation channel. SRPs from different organisms may have different compositions, but all contain related proteins and RNAs.

A **signal sequence** is a short region of a protein that directs it to one of the cell's membranous organelles.

Signal transduction describes the process by which a receptor interacts with a ligand at the surface of the cell and then transmits a signal to trigger a pathway within the cell.

A **silencer** is a short sequence of DNA that can inactivate expression of a gene in its vicinity.

Silencing describes the repression of gene expression in a localized region, usually as the result of a structural change in chromatin.

Silent mutations do not change the product of a gene.

A **silent site** in a coding region is one where mutation does not change the sequence of the protein.

A **single copy** plasmid replicates under a control system analogous to the bacterial chromosome that allows only one copy to exist in an individual bacterial cell.

Single nucleotide polymorphism (SNP) describes a polymorphism (variation in sequence between individuals) caused by a change in a single nucleotide. This is responsible for most of the genetic variation between individuals.

The **single X hypothesis** describes the inactivation of one X chromosome in female mammals.

Single-strand assimilation (single-strand uptake) describes the ability of RecA protein to cause a single strand of DNA to displace its homologous strand in a duplex; that is, the single strand is assimilated into the duplex.

The **single-strand binding protein** (SSB) attaches to single-stranded DNA, thereby preventing the DNA from forming a duplex.

Single-strand exchange is a reaction in which one of the strands of a duplex of DNA leaves its former partner and instead pairs with the complementary strand in another molecule, displacing its homologue in the second duplex.

Single-strand passage is a reaction catalyzed by type I topoisomerase in which one section of single-stranded DNA is passed through another strand.

Site-specific recombination (specialized recombination) occurs between two specific (not necessarily homologous) sequences, as in phage integration/excision or resolution of cointegrate structures during transposition.

SL **RNA** (spliced leader RNA) is a small RNA that donates an exon in the *trans*-splicing reaction of trypanosomes and nematodes.

The **slow component** of a reassociation reaction is the last to reassociate; usually consists of nonrepetitive DNA.

A **slow-stop mutant** is a type of DNA replication temperature-sensitive mutant in E. coli that can finish a round of replication at the unpermissive temperature, but cannot start another.

Small cytoplasmic RNAs (scRNAs) are present in the cytoplasm and (sometimes are also found in the nucleus).

A **small nuclear RNA** (snRNA) is one of many small RNA species confined to the nucleus; several of the snRNAs are involved in splicing or other RNA processing reactions.

The **small subunit** of the ribosome (30S in bacteria, 40S in eukaryotes) binds the mRNA.

Smooth ER consists of a regions of endoplasmic reticulum devoid of ribosomes.

The **SNARE hypothesis** proposes that the specificity of a transport vesicle for its target membrane is mediated by the interaction of SNARE proteins. In this hypothesis, a SNARE on the vesicle (v-SNARE) binds specifically to its cognate SNARE on the target membrane (t-SNARE).

A **snoRNA** is a small nuclear RNA that is localized in the nucleolus.

snRNPs (snurps) are small nuclear ribonucleoproteins (snRNAs associated with proteins).

A **somatic mutation** is a mutation occurring in a somatic cell, and therefore affecting only its daughter cells; it is not inherited by descendants of the organism. Somatic mutations are generally spontaneous, but a case of directed mutation occurs in the immune system where more diversity is generated in rearranged immunoglobulin genes by somatic mutation.

Somatic recombination describes the process of joining a C gene to a C gene in a lymphocyte to generate an immunoglobulin or T cell receptor.

Sorting signal is a motif in a protein (either a short sequence of amino acids or a covalent modification) that is required for it to be incorporated into vesicles that carry it to a specific destination.

The SOS box is the DNA sequence (operator) of ~20bp recognized by LexA repressor protein.

An **SOS response** in *E. coli* describes the coordinate induction of many enzymes, including repair activities, in response to irradiation or other damage to DNA; results from activation of protease activity by RecA to cleave LexA repressor.

A **spacer** is a sequence in a gene cluster that separates the repeated copies of the transcription unit.

The **spindle** guides the movement of chromosomes during cell division. The structure is made up of microtubules.

Splice recombinant DNA results from a Holliday junction being resolved by cutting the non-exchanged strands. Both strands of DNA before the exchange point come from one chromosome; the DNA after the exchange point come from the homologous chromosome.

Splice sites are the sequences immediately surrounding the exon-intron boundaries.

The **spliceosome** is a complex formed by the snRNPs that are required for splicing together with additional protein factors.

Spontaneous mutations occur in the absence of any added reagent to increase the mutation rate, as the result of errors in replication (or other events involved in the reproduction of DNA) or by environmental damage.

Sporulation is the generation of a spore by a bacterium (by morphological conversion) or by a yeast (as the product of meiosis).

An **SR protein** has a variable length of n Arg-Ser-rich region and is involved in splicing.

An **sRNA** is a small bacterial RNA that functions as a regulator of gene expression.

START (restriction point) is the point during G1 at which a cell becomes committed to division.

Startpoint (startsite) refers to the position on DNA corresponding to the first base incorporated into RNA.

A **stem** is the base-paired segment of a hairpin structure in RNA.

Steroid receptors are transcription factors that are activated by binding of a steroid ligand.

A **sterol** is a compound containing a planar steroid ring.

A **stop codon** (termination codon) is one of three triplets (UAG, UAA, UGA) that causes protein synthesis to terminate. They are also known historically as *nonsense codons*. The UAA codon is called ochre, and the UAA codon as called amber, after the names of the nonsense mutations by which they were originally identified.

Strand displacement is a mode of replication of some viruses in which a new DNA strand grows by displacing the previous (homologous) strand of the duplex.

The **stringency** of a hybridization describes the effect of conditions on the degree of complementarity that is required for reaction. At the most stringent conditions, only exact complements can hybridize. As the stringency is lowered, an increasing number of mismatches can be tolerated between the two strands that are hybridizing.

The **stringent factor** is the protein RelA, which is associated with ribosomes. It synthesizes ppGpp and pppGpp when uncharged aminoacyl-tRNA enters the A site.

Stringent response refers to the ability of a bacterium to shut down synthesis of tRNA and ribosomes in a poor-growth medium.

A **structural distortion** is a change in the shape of a molecule.

A **structural gene** codes for any RNA or protein product other than a regulator.

Structural maintenance of chromosomes (SMC) describes a group of proteins that include the cohesins, which hold sister chromatids together, and the condensins, which are involved in chromosome condensation.

The **structural periodicity** is the number of base pairs per turn of the double helix of DNA.

A **subviral pathogen** is an infectious agent that is smaller than a virus, such as a virusoid.

Super-repressed is a mutant condition in which a repressible operon cannot be de-repressed, so it is always turned off.

Supercoiling describes the coiling of a closed duplex DNA in space so that it crosses over its own axis.

A **superfamily** is a set of genes all related by presumed descent from a common ancestor, but now showing considerable variation.

Suppression describes the occurrence of changes that eliminate the effects of a mutation without reversing the original change in DNA.

A frameshift **suppressor** is an insertion or deletion of a base that restores the original reading frame in a gene that has had a base deletion or insertion.

A **suppressor** is a second mutation that compensates the effects of a primary mutation, so that the combination of the two mutations restores wild phenotype.

Surveillance systems check nucleic acids for errors. The term is used in several different contexts. One example is the system that degrades mRNAs that have nonsense mutations. Another is the set of systems that react to damage in the double helix. The common feature is that the system recognizes an invalid sequence or structure and triggers a response.

SW1/SNF is a chromatin remodeling complex; it uses hydrolysis of ATP to change the organization of nucleosomes.

A **symporter** is a type of carrier protein that moves two different solutes across the plasma membrane in the same direction. The two solutes can be transported simultaneously or sequentially.

A **synapse** is a connection between a neuron and a target cell at which chemical information or electrical impulses may be transmitted.

Synapsis describes the association of the two pairs of sister chromatids (representing homologous chromosomes) that occurs at the start of meiosis; the resulting structure is called a bivalent.

The **synaptonemal complex** describes the morphological structure of synapsed chromosomes.

Synonym codons have the same meaning in the genetic code. Synonym tRNAs bear the same amino acid and respond to the same codon.

Synteny describes a relationship between chromosomal regions of different species where homologous genes occur in the same order.

T cells are lymphocytes of the T (thymic) lineage; may be subdivided into several functional types. They carry TcR (T-cell receptor) and are involved in the cell-mediated immune response.

The **T cell receptor** (TCR) is the antigen receptor on T lymphocytes. It is clonally expressed and binds to a complex of MHC class I or class II protein and antigen-derived peptide.

T-DNA is the segment of the Ti plasmid of *Agrobacterium tumefaciens* that is transferred to the plant cell nucleus during infection. It carries genes that transform the plant cell.

TAFs are the subunits of $TF_{II}D$ that assist TBP in binding to DNA. They also provide points of contact for other components of the transcription apparatus.

TATA box is a conserved A·T-rich septamer found about 25 bp before the startpoint of each eukaryotic RNA polymerase II transcription unit; may be involved in positioning the enzyme for correct initiation.

The **TATA-binding protein** (TBP) is the subunit of transcription factor $TF_{II}D$ that binds to DNA.

A **TATA-less promoter** does not have a TATA box in the sequence upstream of its startpoint.

Telomerase is the ribonucleoprotein enzyme that creates repeating units of one strand at the telomere, by adding individual bases to the DNA 3' end, as directed by an RNA sequence in the RNA component of the enzyme.

A **telomere** is the natural end of a chromosome; the DNA sequence consists of a simple repeating unit with a protruding single-stranded end that may fold into a hairpin.

Telomeric silencing describes the repression of gene activity that occurs in the vicinity of a telomere.

A **teratoma** is a growth in which many differentiated cell types - including skin, teeth, bone and others - grow in disorganized manner after an early embryo is transplanted into one of the tissues of an adult animal.

A terminal protein allows replication of a linear phage genome to start at the very end. The protein attaches to the 5' end of the genome through a covalent bond, is associated with a DNA polymerase, and contains a cytosine residue that serves as a primer.

The **terminal region** is the part of an N-linked oligosaccharide that consists of all the sugar residues added subsequent to formation of the inner core.

In *Drosophila,* the **terminal system** is one of the maternal systems that establishes the polarity of the oocyte. The set of genes in the terminal system play a role in the proper formation of the terminal structures at both ends of the fly.

A **terminase** enzyme cleaves multimers of a viral genome and then uses hydrolysis of ATP to provide the energy to translocate the DNA into an empty viral capsid starting with the cleaved end.

Termination is a separate reaction that ends a macromolecular synthesis reaction (replication, transcription, or translation), by stopping the addition of subunits, and (typically) causing dissolution of the synthetic apparatus.

A **terminator** is a sequence of DNA that causes RNA polymerase to terminate transcription.

A **terminus** is a segment of DNA at which replication ends.

The **ternary complex** in initiation of transcription consists of RNA polymerase and DNA and a dinucleotide that represents the first two bases in the RNA product.

$TF_{II}D$ is the transcription factor that binds to the TATA sequence upstream of the startpoint of promoters for RNA polymerase II. It consists of TBP (TATA binding protein) and the TAF subunits that bind to TBP.

Thalassemia is disease of red blood cells resulting from lack of either a or β globin.

Third base degeneracy describes the lesser effect on codon meaning of the nucleotide present in the third codon position.

The **Ti plasmid** is an episome of the bacterium *Agrobacterium tumefaciens* that carries the genes responsible for the induction of crown gall disease in infected plants.

Tight binding of RNA polymerase to DNA describes the formation of an open complex (when the strands of DNA have separated).

The **TIM complex** resides in the inner membrane of mitochondria and is responsible for transporting proteins from the intermembrane space into the interior of the organelle.

Bacterial transposons that contain markers that are not related to their function, e.g. drug resistance, are named as **Tn** followed by a number.

Tolerance is the lack of an immune response to an antigen (either self antigen or foreign antigen) due to clonal deletion.

A **Toll-like receptor** (TLR) is a plasma membrane receptor that is expressed on phagocytes and other cells and is involved in signaling during the innate immune response. TLRs are related to IL-1 receptors.

The **TOM complex** resides in the outer membrane of the mitochondrion and is responsible for importing proteins from the cytosol into the space between the membranes.

Topological isomers are molecules of DNA that are identical except for a difference in linking number.

A **tracer** is a radioactively labeled nucleic acid component included in a reassociation reaction in amounts too small to influence the progress of reaction.

Trailer (3′ UTR) is a nontranslated sequence at the 3′ end of an mRNA following the termination codon.

trans configuration of two sites refers to their presence on two different molecules of DNA (chromosomes).

The *trans* **face** of the Golgi is juxtaposed to the plasma membrane.

A ***trans*-acting** product can function on any copy of its target DNA. This implies that it is a diffusible protein or RNA.

A **transcript** is the RNA product produced by copying one strand of DNA. It may require processing to generate a mature RNA.

Transcription describes synthesis of RNA on a DNA template.

A **transcription factor** is required for RNA polymerase to initiate transcription at specific promoter(s), but is not itself part of the enzyme.

A **transcription unit** is the distance between sites of initiation and termination by RNA polymerase; may include more than one gene.

The **transcriptome** is the complete set of RNAs present in a cell, tissue, or organism. Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.

Transcytosis is the transport of molecules from one type of plasma membrane domain to another in polarized cells, such as neurons and epithelial cells.

A **transducing virus** carries part of the host genome in place of part of its own sequence. The best known examples are retroviruses in eukaryotes and DNA phages in *E. coli.*

A **transesterification** reaction breaks and makes chemical bonds in a coordinated transfer so that no energy is required.

Transfection of eukaryotic cells is the acquisition of new genetic markers by incorporation of added DNA.

The **transfer region** is a segment on the F plasmid that is required for bacterial conjugation.

Transfer RNA (tRNA) is the intermediate in protein synthesis that interprets the genetic code. Each tRNA can be linked to an amino acid. The tRNA has an anticodon sequence that complementary to a triplet codon representing the amino acid.

Transformation of bacteria describes the acquisition of new genetic markers by incorporation of added DNA.

Transformation (oncogenesis) of eukaryotic cells refers to their conversion to a state of unrestrained growth in culture, resembling or identical with the tumorigenic condition.

Transformed cells are cultured cells that have acquired many of the properties of cancer cells.

The **transforming principle** is DNA that is taken up by a bacterium and whose expression then changes the properties of the recipient cell.

A **transgene** is a gene that is introduced into a cell or animal from an external source.

Transgenic animals are created by introducing new DNA sequences into the germline via addition to the egg.

Transient transfectants have foreign DNA in an unstable, i.e. extrachromosomal form.

A **transition** is a mutation in which one pyrimidine is substituted by the other or in which one purine is substituted for the other.

A **transition vesicle** is a small membrane-bounded compartment that mediates transport between organelles, especially the rough endoplasmic reticulum and Golgi complex. It is also known as a transport vesicle. COPI- and COPII-coated vesicles are transition vesicles.

Translation is synthesis of protein on the mRNA template.

Translational positioning describes the location of a histone octamer at successive turns of the double helix, which determines which sequences are located in linker regions.

Protein **translocation** describes the movement of a protein across a membrane. This occurs across the membranes of organelles in eukaryotes, or across the plasma membrane in bacteria. Each membrane across which proteins are translocated has a channel specialized for the purpose.

Translocation describes a rearrangement in which part of a chromosome is detached by breakage and then becomes attached to some other chromosome.

Translocation describes the stage of nuclear import or export when a protein or RNA substrate moves through the nuclear pore.

Translocation is the movement of the ribosome one codon along mRNA after the addition of each amino acid to the polypeptide chain.

A **translocon** is a discrete structure in a membrane that forms a channel through which (hydrophilic) proteins may pass.

A **transmembrane protein** (integral membrane protein) extends across a lipid bilayer. A hydrophobic region (typically consisting of a stretch of 20-25 hydrophobic and/or uncharged amino acids) or regions of the protein resides in the membrane. Hydrophilic regions are exposed on one or both sides of the membrane.

The **transmembrane region** (transmembrane domain) is the part of a protein that spans the membrane bilayer. It is hydrophobic and in many cases contains approximately 20 amino acids that form an α-helix. It is also called the transmembrane domain.

Transplantation antigen is protein coded by a major histocompatibility locus, present on all mammalian cells, involved in interactions between lymphocytes.

A **transport signal** is the part of a cargo molecule that is recognized by coat proteins or cargo receptors for incorporation into budding transport vesicles. Examples of transport signals are short amino acid sequences, secondary structure, and a protein modification such as phosphorylation.

A **transporter** is a type of receptor that moves small molecules across the plasma membrane. It binds the molecules on its extracellular surface, and releases them into the cytoplasm.

Transposase is the enzyme activity involved in insertion of transposon at a new site.

Transposition refers to the movement of a transposon to a new site in the genome.

A **transposon** (transposable element) is a DNA sequence able to insert itself at a new location in the genome, without having any sequence relationship with the target locus.

A **transversion** is a mutation in which a purine is replaced by a pyrimidine or vice versa.

A **triskelion** is formed by the interaction of three heavy chains and three light chains of clathrin.

$tRNA_f^{Met}$ is the special RNA that is to initiate protein synthesis in bacteria. It mostly uses AUG, but can also respond to GUG and CUG.

$tRNA_i^{Met}$ is the special tRNA used to respond to initiation codons in eukaryotes.

$tRNA_m^{Met}$ inserts methionine at internal AUG codons.

A **true reversion** is a mutation that restores the original sequence of the DNA.

A **tumor suppressor** is identified by a loss-of-function mutation that contributes to cancer formation. They usually function to prevent cell division or to cause death of abnormal cells. The two most important are p53 and RB.

A **tumor virus** has the ability to transform an animal cell into a cancerous state.

The **twisting number** of a DNA is the number of base pairs divided by the number of base pairs per turn of the double helix.

Two hybrid assay detects interaction between two proteins by means of their ability to bring together a DNA-binding domain and a transcription-activating domain. The assay is performed in yeast using a reporter gene that responds to the interaction.

Ty stands for transposon yeast, the first transposable element to be identified in yeast.

A **type I topoisomerase** is an enzyme that changes the topology of DNA by nicking and resealing one strand of DNA.

A **type II topoisomerase** is an enzyme that changes the topology of DNA by nicking and resealing both strands of DNA.

U3 is the repeated sequence at the 3' end of a retroviral RNA.

U5 is the repeated sequence at the 5' end of a retroviral RNA.

Ubiquitin has a highly conserved sequence of 76 amino acids. It is linked via its COOH group to the e NH_2 group of a lysine residue in a target protein.

A stretch of **underwound** DNA has fewer base pairs per turn than the usual average (10 bp = 1 turn). This means that the two strands of DNA are less tightly wound around each other; ultimately this can lead to strand separation.

Unequal crossing-over describes a recombination event in which the two recombining sites lie at nonidentical locations in the two parental DNA molecules.

Unidirectional replication refers to the movement of a single replication fork from a given origin.

An **uninducible** mutant is one where the affected gene(s) cannot be expressed.

A **uniporter** is a type of carrier protein that moves only one type of solute across the plasma membrane.

The **unit cell** describes the state of an *E. coli* bacterium generated by a new division. It is 1.7 μm long and has a single replication origin.

An **unsaturated** fatty acid has some double carbon-carbon bonds in its backbone.

An **up mutation** in a promoter increases the rate of transcription.

Upstream identifies sequences proceeding in the opposite direction from expression; for example, the bacterial promoter is upstream of the transcription unit, the initiation codon is upstream of the coding region.

An **upstream activator sequence** (UAS) is the equivalent in yeast of the enhancer in higher eukaryotes.

A V **gene** is sequence coding for the major part of the variable (N-terminal) region of an immunoglobulin chain.

The **variable region** (V region) of an immunoglobulin chain is coded by the V gene and varies extensively when different chains are compared, as the result of multiple (different) genomic copies and changes introduced during construction of an active immunoglobulin.

The **variable surface glycoprotein** (VSG) is the protein on the surface of a trypanosome that changes during an infection so as to prevent the infected host from mounting an immune reaction to it.

Variegation of phenotype is produced by a change in genotype during somatic development.

The **vegetative phase** describes the period of normal growth and division of a bacterium. For a bacterium that can sporulate, this contrasts with the sporulation phase, when spores are being formed.

The **viral superfamily** comprises transposons that are related to retroviruses. They are defined by sequences that code for reverse transcriptase or integrase.

Virion is the physical virus particle (irrespective of its ability to infect cells and reproduce).

A **viroid** is a small infectious nucleic acid that does not have a protein coat.

Virulent phage mutants are unable to establish lysogeny.

A **virusoid** (satellite RNA) is a small infectious nucleic acid that is encapsidated by a plant virus together with its own genome.

VNTR (variable number tandem repeat) regions describe very short repeated sequences, including microsatellites and minisatellites.

Voltage-gated channels are open or closed depending on the voltage across the membrane.

Wobble hypothesis accounts for the ability of a tRNA to recognize more than one codon by unusual (non-G·C, non-A·T) pairing with the third base of a codon.

The **writhing number** is the number of times a duplex axis crosses over itself in space.

A **yeast artificial chromosome** (YAC) is a synthetic DNA molecule that contains an origin for replication, a centromere to support segregation, and telomeres to seal the ends. It is used as a means to propagate whatever genes it carries in yeast cells.

The **zinc finger** is a DNA-binding motif that typifies a class of transcription factor.

A **zoo blot** describes the use of Southern blotting to test the ability of a DNA probe from one species to hybridize with the DNA from the genomes of a variety of other species.