

INFORMATION PATHWAYS

24 Genes and Chromosomes 979

25 DNA Metabolism 1009

26 RNA Metabolism 1057

27 Protein Metabolism 1103

28 Regulation of Gene Expression 1155

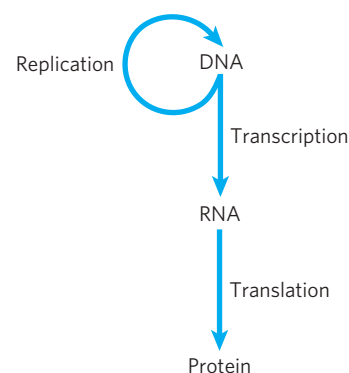
The third and final part of this book explores the biochemical mechanisms underlying the apparently contradictory requirements for both genetic continuity and the evolution of living organisms. What is the molecular nature of genetic material? How is genetic information transmitted from one generation to the next with high fidelity? How do the rare changes in genetic material that are the raw material of evolution arise? How is genetic information ultimately expressed in the amino acid sequences of the astonishing variety of protein molecules in a living cell?

Today's understanding of information pathways has arisen from the convergence of genetics, physics, and chemistry in modern biochemistry. This was epitomized by the discovery of the double-helical structure of DNA, postulated by James Watson and Francis Crick in 1953 (see Fig. 8–13). Genetic theory contributed the concept of coding by genes. Physics permitted the determination of molecular structure by x-ray diffraction analysis. Chemistry revealed the composition of DNA. The profound impact of the Watson-Crick hypothesis arose from its ability to account for a wide range of observations derived from studies in these diverse disciplines.

This revolutionized our understanding of the structure of DNA and inevitably stimulated questions about its function. The double-helical structure itself clearly suggested how DNA might be copied so that the information it contains can be transmitted from one generation to the next. Clarification of how the information in DNA is converted into functional proteins came with

the discovery of messenger RNA and transfer RNA and with the deciphering of the genetic code.

These and other major advances gave rise to the central dogma of molecular biology, comprising the three major processes in the cellular utilization of genetic information. The first is replication, the copying of parental DNA to form daughter DNA molecules with identical nucleotide sequences. The second is transcription, the process by which parts of the genetic message encoded in DNA are copied precisely into RNA. The third is translation, whereby the genetic message encoded in messenger RNA is translated on the ribosomes into a polypeptide with a particular sequence of amino acids.



The central dogma of molecular biology, showing the general pathways of information flow via replication, transcription, and translation. The term “dogma” is a misnomer and is retained for historical reasons only. Introduced by Francis Crick at a time when little evidence supported these ideas, the dogma has become a well-established principle.

Part III explores these and related processes. In Chapter 24 we examine the structure, topology, and packaging of chromosomes and genes. The processes underlying the central dogma are elaborated in Chapters 25 through 27. Finally, we turn to regulation, examining how the expression of genetic information is controlled (Chapter 28).

A major theme running through these chapters is the added complexity inherent in the biosynthesis of macromolecules that contain information. Assembling nucleic acids and proteins with particular sequences of nucleotides and amino acids represents nothing less than preserving the faithful expression of the template upon which life itself is based. We might expect the formation of phosphodiester bonds in DNA or peptide bonds in proteins to be a trivial feat for cells, given the arsenal of enzymatic and chemical tools described in Part II. However, the framework of patterns and rules established in our examination of metabolic pathways thus far must be enlarged considerably to take into account molecular information. Bonds must be formed between *particular* subunits in informational biopolymers, avoiding either the occurrence or the persistence of sequence errors. This has an enormous impact on the thermodynamics, chemistry, and enzymology of the biosynthetic processes. Formation of a peptide bond requires an energy input of only about 21 kJ/mol of bonds and can be catalyzed by relatively simple enzymes. But to synthesize a bond between two specific amino acids at a particular point in a polypeptide, the cell invests about 125 kJ/mol while making use of more than

200 enzymes, RNA molecules, and specialized proteins. The chemistry involved in peptide bond formation does not change because of this requirement, but additional processes are layered over the basic reaction to ensure that the peptide bond is formed between particular amino acids. Biological information is expensive.

The dynamic interaction between nucleic acids and proteins is another central theme of Part III. Regulatory and catalytic RNA molecules are gradually taking a more prominent place in our understanding of these pathways (discussed in Chapters 26 and 27). However, most of the processes that make up the pathways of cellular information flow are catalyzed and regulated by proteins. An understanding of these enzymes and other proteins can have practical as well as intellectual rewards, because they form the basis of recombinant DNA technology (introduced in Chapter 9).

Evolution again constitutes an overarching theme. Many of the processes outlined in Part III can be traced back billions of years, and a few can be traced to LUCA, the last universal common ancestor. The ribosome, most of the translational apparatus, and some parts of the transcriptional machinery are shared by every living organism on this planet. Genetic information is a kind of molecular clock that can help define ancestral relationships among species. Shared information pathways connect humans to every other species now living on Earth, and to all species that came before. Exploration of these pathways is allowing scientists to slowly open the curtain on the first act—the events that may have heralded the beginning of life on Earth.

Genes and Chromosomes

24.1 Chromosomal Elements 979

24.2 DNA Supercoiling 985

24.3 The Structure of Chromosomes 994

The size of DNA molecules presents an interesting biological puzzle. Given that these molecules are generally much longer than the cells or viral particles that contain them (**Fig. 24–1**), how do they fit



FIGURE 24–1 Bacteriophage T2 protein coat surrounded by its single, linear molecule of DNA. The DNA was released by lysing the bacteriophage particle in distilled water and allowing the DNA to spread on the water surface. An undamaged T2 bacteriophage particle consists of a head structure that tapers to a tail by which the bacteriophage attaches itself to the outer surface of a bacterial cell. All the DNA shown in this electron micrograph is normally packaged inside the phage head.

into their cellular or viral packages? To address this question, we shift our focus from the secondary structure of DNA, considered in Chapter 8, to the extraordinary degree of organization required for the tertiary packaging of DNA into **chromosomes**—the repositories of genetic information. The chapter begins with an examination of the elements that make up viral and cellular chromosomes, and then considers chromosomal size and organization. We then discuss DNA topology, describing the coiling and supercoiling of DNA molecules. Finally, we consider the protein-DNA interactions that organize chromosomes into compact structures.

24.1 Chromosomal Elements

Cellular DNA contains genes and intergenic regions, both of which may serve functions vital to the cell. The more complex genomes, such as those of eukaryotic cells, demand increased levels of chromosomal organization, and this is reflected in the chromosomes' structural features. We begin by considering the different types of DNA sequences and structural elements within chromosomes.

Genes Are Segments of DNA That Code for Polypeptide Chains and RNAs

Our understanding of genes has evolved tremendously over the last century. Classically, a gene was defined as a portion of a chromosome that determines or affects a single character or **phenotype** (visible property), such as eye color. George Beadle and Edward Tatum proposed a molecular definition of a gene in 1940. After exposing spores of the fungus *Neurospora crassa* to x rays and other agents now known to damage DNA and cause alterations in DNA sequence (**mutations**), they detected mutant fungal strains that lacked one or another specific enzyme, sometimes resulting in the failure of an entire metabolic pathway. Beadle and Tatum concluded that a gene is a segment of genetic material that determines, or codes for, one enzyme: the

one gene–one enzyme hypothesis. Later this concept was broadened to **one gene–one polypeptide**, because many genes code for a protein that is not an enzyme or for one polypeptide of a multisubunit protein.



George W. Beadle,
1903-1989



Edward L. Tatum,
1909-1975

The modern biochemical definition of a gene is even more precise. A **gene** is all the DNA that encodes the primary sequence of some final gene product, which can be either a polypeptide or an RNA with a structural or catalytic function. DNA also contains other segments or sequences that have a purely regulatory function. **Regulatory sequences** provide signals that may denote the beginning or the end of genes, or influence the transcription of genes, or function as initiation points for replication or recombination (Chapter 28). Some genes can be expressed in different ways to generate multiple gene products from a single segment of DNA. The special transcriptional and translational mechanisms that allow this are described in Chapters 26 through 28.

We can estimate directly the minimum overall size of genes that encode proteins. As described in detail in Chapter 27, each amino acid of a polypeptide chain is coded for by a sequence of three consecutive nucleotides in a single strand of DNA (**Fig. 24-2**), with these “codons” arranged in a sequence that corresponds to the sequence of amino acids in the polypeptide that the gene encodes. A polypeptide chain of 350 amino acid residues (an average-size chain) corresponds to 1,050 bp of DNA. Many genes in eukaryotes and a few in bacteria and archaea are interrupted by noncoding DNA segments and are therefore considerably longer than this simple calculation would suggest.

How many genes are in a single chromosome? The *Escherichia coli* chromosome, one of the bacterial genomes that have been completely sequenced, is a circular DNA molecule (in the sense of an endless loop rather than a perfect circle) with 4,639,675 bp. These base pairs encode about 4,300 genes for proteins and another 157 genes for structural or catalytic RNA molecules. Among eukaryotes, the approximately 3.1 billion base pairs of the human genome include approximately 25,000 genes on the 24 different chromosomes.

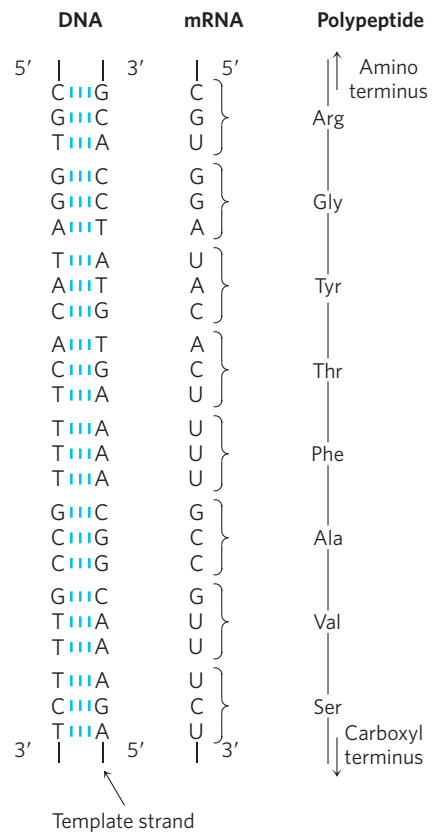


FIGURE 24-2 Colinearity of the coding nucleotide sequences of DNA and mRNA and the amino acid sequence of a polypeptide chain. The triplets of nucleotide units in DNA determine the amino acids in a protein through the intermediary mRNA. One of the DNA strands serves as a template for synthesis of mRNA, which has nucleotide triplets (codons) complementary to those of the DNA. In some bacterial and many eukaryotic genes, coding sequences are interrupted at intervals by regions of noncoding sequences (called introns).

DNA Molecules Are Much Longer Than the Cellular or Viral Packages That Contain Them

Chromosomal DNAs are often many orders of magnitude longer than the cells or viruses in which they are located (**Fig. 24-1**; **Table 24-1**). This is true of every class of organism or viral parasite.

Viruses Viruses are not free-living organisms; rather, they are infectious parasites that use the resources of a host cell to carry out many of the processes they require to propagate. Many viral particles consist of no more than a genome (usually a single RNA or DNA molecule) surrounded by a protein coat.

Almost all plant viruses and some bacterial and animal viruses have RNA genomes. These genomes tend to be particularly small. For example, the genomes of mammalian retroviruses such as HIV are about 9,000 nucleotides long, and the genome of the bacteriophage Q β has 4,220 nucleotides. Both types of virus have single-stranded RNA genomes.

TABLE 24-1 The Sizes of DNA and Viral Particles for Some Bacterial Viruses (Bacteriophages)

Virus	Size of viral DNA (bp)	Length of viral DNA (nm)	Long dimension of viral particle (nm)
ϕ X174	5,386	1,939	25
T7	39,936	14,377	78
λ (lambda)	48,502	17,460	190
T4	168,889	60,800	210

Note: Data on size of DNA are for the replicative form (double-stranded). The contour length is calculated assuming that each base pair occupies a length of 3.4 Å (see Fig. 8-13).

The genomes of DNA viruses vary greatly in size (Table 24-1). Many viral DNAs are circular for at least part of their life cycle. During viral replication within a host cell, specific types of viral DNA called **replicative forms** may appear; for example, many linear DNAs become circular and all single-stranded DNAs become double-stranded. A typical medium-size DNA virus is bacteriophage λ (lambda), which infects *E. coli*. In its replicative form inside cells, λ DNA is a circular double helix. This double-stranded DNA contains 48,502 bp and has a contour length of 17.5 μ m. Bacteriophage ϕ X174 is a much smaller DNA virus; the DNA in the viral particle is a single-stranded circle, and the double-stranded replicative form contains 5,386 bp. Although viral genomes are small, the contour lengths of their DNAs are typically hundreds of times longer than the long dimensions of the viral particles that contain them (Table 24-1).

Bacteria A single *E. coli* cell contains almost 100 times as much DNA as a bacteriophage λ particle. The chromosome of an *E. coli* cell is a single double-stranded circular DNA molecule. Its 4,639,675 bp have a contour length of about 1.7 mm, some 850 times the length of the *E. coli* cell (Fig. 24-3). In addition to the very large, circular DNA chromosome in their nucleoid, many bacteria contain one or more small circular DNA molecules that are free in the cytosol. These extrachromosomal elements are called **plasmids** (Fig. 24-4; see also p. 317). Most plasmids are only a few thousand base pairs long, but some contain more than 10,000 bp. They carry genetic information and undergo replication to yield daughter plasmids, which pass into the daughter cells at cell division. Plasmids have been found in yeast and other fungi as well as in bacteria.

In many cases plasmids confer no obvious advantage on their host, and their sole function seems to be self-propagation. However, some plasmids carry genes that are useful to the host bacterium. For example, some plasmid genes make a host bacterium resistant to antibacterial agents. Plasmids carrying the gene for the enzyme β -lactamase confer resistance to β -lactam anti-

biotics such as penicillin, ampicillin, and amoxicillin (see Fig. 6-31). These and similar plasmids may pass from an antibiotic-resistant cell to an antibiotic-sensitive cell of the same or another bacterial species, making the recipient cell antibiotic resistant. The extensive use of antibiotics in some human populations has served as a strong selective force, encouraging the spread of antibiotic resistance-coding plasmids (as well as transposable elements, described below, that harbor similar genes) in disease-causing bacteria. Physicians are becoming increasingly reluctant to prescribe antibiotics unless a clear clinical need is confirmed. For similar reasons, the widespread use of antibiotics in animal feeds is being curbed.

Eukaryotes A yeast cell, one of the simplest eukaryotes, has 2.6 times more DNA in its genome than an *E. coli* cell (Table 24-2). Cells of *Drosophila*, the fruit fly used in classical genetic studies, contain more than 35 times as much DNA as *E. coli* cells, and human cells have almost 700 times as much. The cells of many plants and amphibians contain even more. The genetic material of eukaryotic cells is apportioned into chromosomes, the diploid ($2n$) number depending on the species (Table 24-2). A human somatic cell, for example, has 46 chromosomes (Fig. 24-5). Each chromosome of a eukaryotic cell, such as that shown in Figure 24-5a, contains a single, very large, duplex DNA molecule. The DNA molecules in the 24 different types of human chromosomes (22 matching pairs plus the X and Y sex chromosomes) vary in length over a 25-fold range. Each type of chromosome in eukaryotes carries a characteristic set of genes.

The DNA molecules of one human genome (22 chromosomes plus X and Y or two X chromosomes), placed end to end, would extend for about a meter. Most human cells are diploid, and each cell contains a total of 2 m of DNA. An adult human body contains approximately 10^{14} cells and thus a total DNA length of 2×10^{11} km. Compare this with the circumference of the earth (4×10^4 km) or the distance between the earth and the sun (1.5×10^8 km)—a dramatic illustration of the extraordinary degree of DNA compaction in our cells.

FIGURE 24-3 The length of the *E. coli* chromosome (1.7 mm) depicted in linear form relative to the length of a typical *E. coli* cell (2 μm).

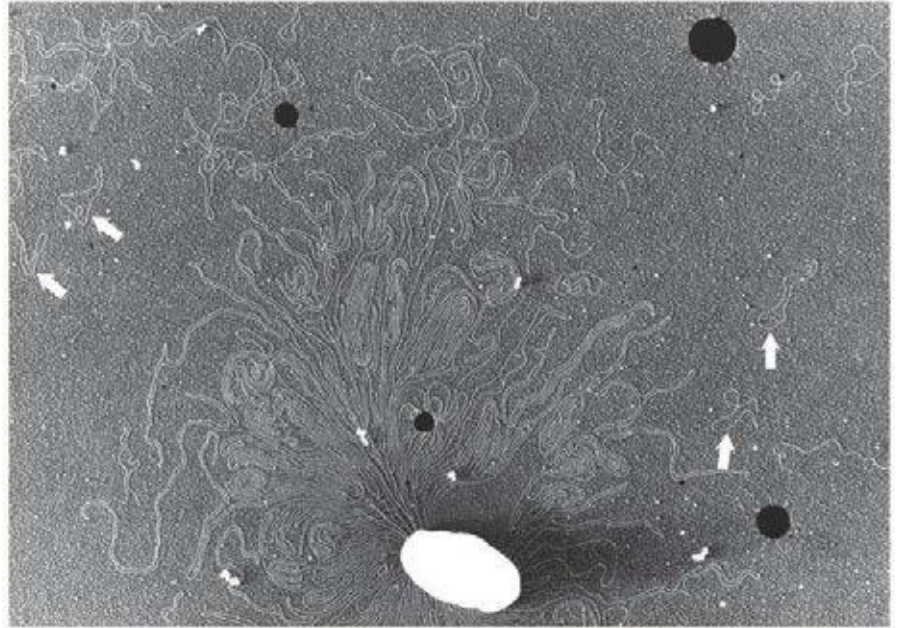
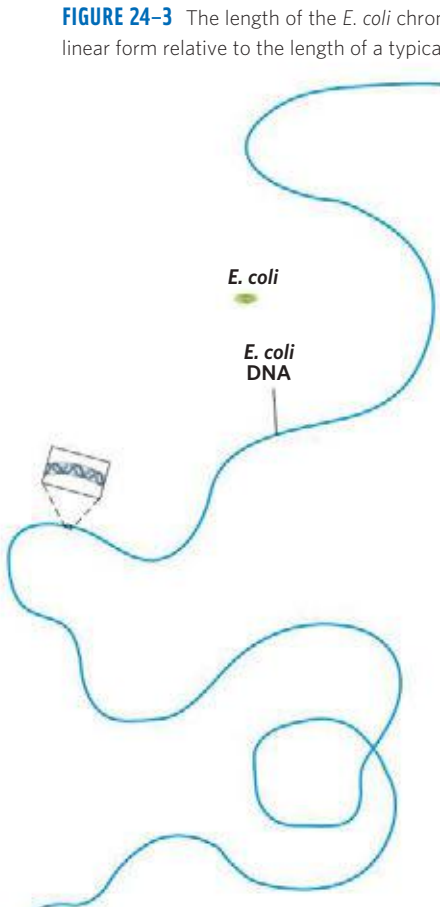


FIGURE 24-4 DNA from a lysed *E. coli* cell. In this electron micrograph several small, circular plasmid DNAs are indicated by white arrows. The black spots and white specks are artifacts of the preparation.

TABLE 24-2 DNA, Gene, and Chromosome Content in Some Genomes

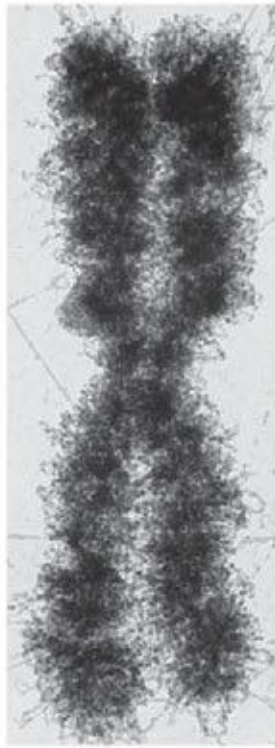
	Total DNA (bp)	Number of chromosomes*	Approximate number of genes
<i>Escherichia coli</i> K12 (bacterium)	4,639,675	1	4,435
<i>Saccharomyces cerevisiae</i> (yeast)	12,080,000	16 [†]	5,860
<i>Caenorhabditis elegans</i> (nematode)	90,269,800	12 [‡]	23,000
<i>Arabidopsis thaliana</i> (plant)	119,186,200	10	33,000
<i>Drosophila melanogaster</i> (fruit fly)	120,367,260	18	20,000
<i>Oryza sativa</i> (rice)	480,000,000	24	57,000
<i>Mus musculus</i> (mouse)	2,634,266,500	40	27,000
<i>Homo sapiens</i> (human)	3,070,128,600	46	29,000

Note: This information is constantly being refined. For the most current information, consult the websites for the individual genome projects.

*The diploid chromosome number is given for all eukaryotes except yeast.

[†]Haploid chromosome number. Wild yeast strains generally have eight (octoploid) or more sets of these chromosomes.

[‡]Number for females, with two X chromosomes. Males have an X but no Y, thus 11 chromosomes in all.



(a)



(b)

FIGURE 24-5 Eukaryotic chromosomes. (a) A pair of linked and condensed sister chromatids of a human chromosome. Eukaryotic chromosomes are in this state after replication at metaphase during mitosis. (b) A complete set of chromosomes from a leukocyte from one of the authors. There are 46 chromosomes in every normal human somatic cell.

Eukaryotic cells also have organelles, mitochondria (**Fig. 24-6**) and chloroplasts, that contain DNA. Mitochondrial DNA (mtDNA) molecules are much smaller than the nuclear chromosomes. In animal cells, mtDNA contains fewer than 20,000 bp (16,569 bp



FIGURE 24-6 A dividing mitochondrion. Some mitochondrial proteins and RNAs are encoded by one of the copies of the mitochondrial DNA (none of which are visible here). The DNA (mtDNA) is replicated each time the mitochondrion divides, before cell division.

in human mtDNA) and is a circular duplex. Each mitochondrion typically has 2 to 10 copies of this mtDNA molecule, and the number can rise to hundreds in certain cells of an embryo that is undergoing cell differentiation. In a few organisms (trypanosomes, for example) each mitochondrion contains thousands of copies of mtDNA, organized into a complex and inter-linked matrix known as a kinetoplast. Plant cell mtDNA ranges in size from 200,000 to 2,500,000 bp. Chloroplast DNA (cpDNA) also exists as circular duplexes and ranges in size from 120,000 to 160,000 bp. The evolutionary origin of mitochondrial and chloroplast DNAs has been the subject of much speculation. A widely accepted view is that they are vestiges of the chromosomes of ancient bacteria that gained access to the cytoplasm of host cells and became the precursors of these organelles (see Fig. 1-38). Mitochondrial DNA codes for the mitochondrial tRNAs and rRNAs and for a few mitochondrial proteins. More than 95% of mitochondrial proteins are encoded by nuclear DNA. Mitochondria and chloroplasts divide when the cell divides. Their DNA is replicated before and during division, and the daughter DNA molecules pass into the daughter organelles.

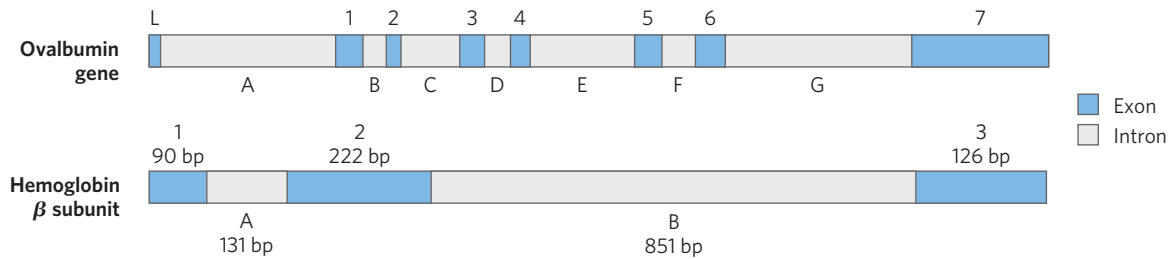


FIGURE 24-7 Introns in two eukaryotic genes. The gene for ovalbumin has seven introns (A to G), splitting the coding sequences into eight exons (L, and 1 to 7). The gene for the β subunit of hemoglobin has two

introns and three exons, including one intron that alone contains more than half the base pairs of the gene.

Eukaryotic Genes and Chromosomes Are Very Complex

Many bacterial species have only one chromosome per cell and, in nearly all cases, each chromosome contains only one copy of each gene. A very few genes, such as those for rRNAs, are repeated several times. Genes and regulatory sequences account for almost all the DNA in bacteria. Moreover, almost every gene is precisely colinear with the amino acid sequence (or RNA sequence) for which it codes (Fig. 24-2).

The organization of genes in eukaryotic DNA is structurally and functionally much more complex. The study of eukaryotic chromosome structure, and more recently the sequencing of entire eukaryotic genomes, has yielded many surprises. Many, if not most, eukaryotic genes have a distinctive and puzzling structural feature: their nucleotide sequences contain one or more intervening segments of DNA that do not code for the amino acid sequence of the polypeptide product. These nontranslated inserts interrupt the otherwise colinear relationship between the nucleotide sequence of the gene and the amino acid sequence of the polypeptide it encodes. Such nontranslated DNA segments in genes are called **intervening sequences** or **introns**, and the coding segments are called **exons**. Few bacterial genes contain introns. In higher eukaryotes, the typical gene has much more intron sequence than sequences devoted to exons. For example, in the gene coding for the single polypeptide chain of ovalbumin, an avian egg protein (Fig. 24-7), the introns are much longer than the exons; altogether, seven introns make up 85% of the gene's DNA. The gene for the muscle protein titin is the intron champion, with 178 introns. Genes for histones seem to have no introns. In most cases the function of introns is not clear. In total, only about 1.5% of human DNA is "coding" or exon DNA, carrying information for protein products. However, when the much larger introns are included in the count, as much as 30% of the human genome consists of genes. A great deal of work remains to be done to understand the other genomic sequences. Much of the DNA that is not within genes is made up of repeated sequences of several kinds. These include transposable elements (transposons), molecular parasites that account for nearly half of the DNA in the human genome (see Fig. 9-29 and Chapters 25 and 26).

Approximately 3% of the human genome consists of **highly repetitive** sequences, also referred to as **simple-sequence DNA** or **simple sequence repeats (SSR)**. These short sequences, generally less than 10 bp long, are sometimes repeated millions of times per cell. The simple-sequence DNA has also been called **satellite DNA**, so named because its unusual base composition often causes it to migrate as "satellite" bands (separated from the rest of the DNA) when fragmented cellular DNA samples are centrifuged in a cesium chloride density gradient. Studies suggest that simple-sequence DNA does not encode proteins or RNAs. Unlike the transposable elements, the highly repetitive DNA can have identifiable functional importance in human cellular metabolism, because much of it is associated with two defining features of eukaryotic chromosomes: centromeres and telomeres.

The **centromere** (Fig. 24-8) is a sequence of DNA that functions during cell division as an attachment point for proteins that link the chromosome to the mitotic spindle. This attachment is essential for the equal and orderly distribution of chromosome sets to daughter cells. The centromeres of *Saccharomyces cerevisiae* have been isolated and studied. The sequences essential to centromere function are about 130 bp long and are very rich in A=T pairs. The centromeric sequences of higher eukaryotes are much longer and, unlike those of yeast, generally contain simple-sequence DNA, which consists of thousands of tandem copies of one or a few short sequences of 5 to 10 bp, in the same orientation. The precise role of simple-sequence DNA in centromere function is not yet understood.

Telomeres (Greek *telos*, "end") are sequences at the ends of eukaryotic chromosomes that help stabilize the chromosome. Telomeres end with multiple repeated sequences of the form

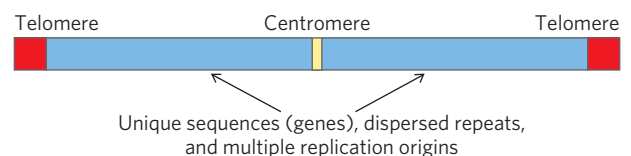
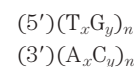


FIGURE 24-8 Important structural elements of a yeast chromosome.

TABLE 24-3 Telomere Sequences

Organism	Telomere repeat sequence
<i>Homo sapiens</i> (human)	(TTAGGG) _n
<i>Tetrahymena thermophila</i> (ciliated protozoan)	(TTGGGG) _n
<i>Saccharomyces cerevisiae</i> (yeast)	((TG) ₁₋₃ (TG) ₂₋₃) _n
<i>Arabidopsis thaliana</i> (plant)	(TTTAGGG) _n

where x and y are generally between 1 and 4 (Table 24-3). The number of telomere repeats, n , is in the range of 20 to 100 for most single-celled eukaryotes and is generally more than 1,500 in mammals. The ends of a linear DNA molecule cannot be routinely replicated by the cellular replication machinery (which may be one reason why bacterial DNA molecules are circular). Repeated telomeric sequences are added to eukaryotic chromosome ends primarily by the enzyme telomerase (see Fig. 26-38).

Artificial chromosomes (Chapter 9) have been constructed as a means of better understanding the functional significance of many structural features of eukaryotic chromosomes. A reasonably stable artificial linear chromosome requires only three components: a centromere, a telomere at each end, and sequences that allow the initiation of DNA replication. Yeast artificial chromosomes (YACs; see Fig. 9-6) have been developed as a research tool in biotechnology. Similarly, human artificial chromosomes (HACs) are being developed for the treatment of genetic diseases. These may eventually provide a new path to the intracellular replacement of missing or defective gene products or somatic gene therapy.

SUMMARY 24.1 Chromosomal Elements

- ▶ Genes are segments of a chromosome that contain the information for a functional polypeptide or RNA molecule. In addition to genes, chromosomes contain a variety of regulatory sequences involved in replication, transcription, and other processes.
- ▶ Genomic DNA and RNA molecules are generally orders of magnitude longer than the viral particles or cells that contain them.
- ▶ Many genes in eukaryotic cells (but few in bacteria and archaea) are interrupted by noncoding sequences, or introns. The coding segments separated by introns are called exons.
- ▶ Only about 1.5% of human genomic DNA encodes proteins. Even when introns are included, less than one-third of human genomic DNA consists of genes. Much of the remainder consists of repeated sequences of various types. Nucleic acid parasites known as transposons account for about half of the human genome.

- ▶ Eukaryotic chromosomes have two important special-function repetitive DNA sequences: centromeres, which are attachment points for the mitotic spindle, and telomeres, located at the ends of chromosomes.

24.2 DNA Supercoiling

Cellular DNA, as we have seen, is extremely compacted, implying a high degree of structural organization. The folding mechanism not only must pack the DNA but also must permit access to the information in the DNA. Before considering how this is accomplished in processes such as replication and transcription, we need to examine an important property of DNA structure known as **supercoiling**.

“Supercoiling” means the coiling of a coil. An old-fashioned telephone cord, for example, is typically a coiled wire. The path taken by the wire between the base of the phone and the receiver often includes one or more supercoils (**Fig. 24-9**). DNA is coiled in the form of a double helix, with both strands of the DNA coiling around an axis. The further coiling of that axis upon itself (**Fig. 24-10**) produces DNA supercoiling. As detailed below, DNA supercoiling is generally a manifestation of structural strain. When there is no net bending of the DNA axis upon itself, the DNA is said to be in a **relaxed** state.

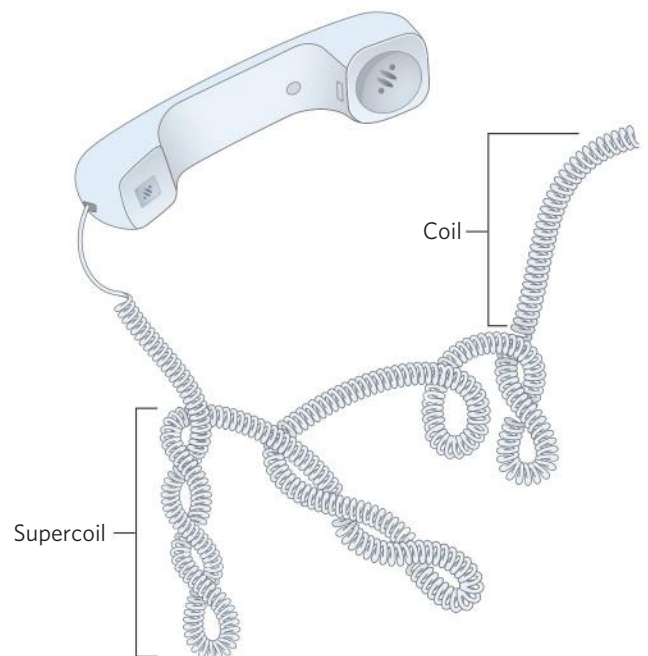


FIGURE 24-9 Supercoils. A typical phone cord is coiled like a DNA helix, and the coiled cord can itself coil in a supercoil. The illustration is especially appropriate because an examination of phone cords helped lead Jerome Vinograd and his colleagues to the insight that many properties of small circular DNAs can be explained by supercoiling. They first detected DNA supercoiling—in small circular viral DNAs—in 1965.

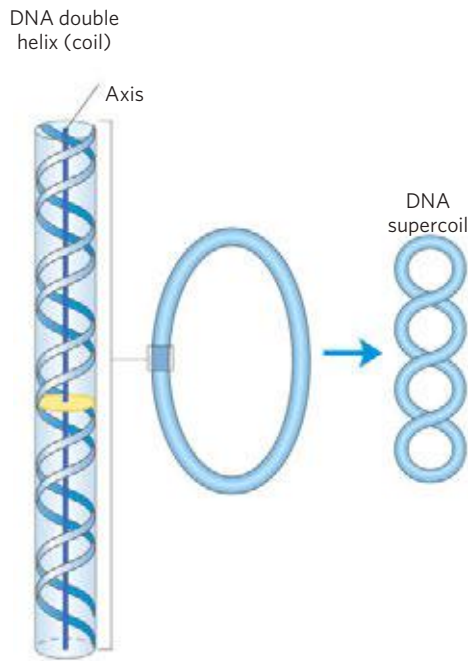


FIGURE 24-10 Supercoiling of DNA. When the axis of the DNA double helix is coiled on itself, it forms a new helix (superhelix). The DNA superhelix is usually called a supercoil.

We might have predicted that DNA compaction involved some form of supercoiling. Perhaps less predictable is that replication and transcription of DNA also affect and are affected by supercoiling. Both processes require a separation of DNA strands—a process complicated by the helical interwinding of the strands (as demonstrated in **Fig. 24-11**).

That a DNA molecule would bend on itself and become supercoiled in tightly packaged cellular DNA would seem logical, then, and perhaps even trivial, were it not for one additional fact: many circular DNA molecules remain highly supercoiled even after they are extracted and purified, freed from protein and other cellular components. This indicates that supercoiling is an intrinsic property of DNA tertiary structure. It occurs in all cellular DNAs and is highly regulated by each cell.

Several measurable properties of supercoiling have been established, and the study of supercoiling has provided many insights into DNA structure and function. This work has drawn heavily on concepts derived from a branch of mathematics called **topology**, the study of the properties of an object that do not change under continuous deformations. For DNA, continuous deformations include conformational changes due to thermal motion or an interaction with proteins or other molecules; discontinuous deformations involve DNA strand breakage. For circular DNA molecules, a topological property is one that is unaffected by deformations of the DNA strands as long as no breaks are introduced. Topological properties are changed only by breakage and rejoining of the backbone of one or both DNA strands.

We now examine the fundamental properties and physical basis of supercoiling.

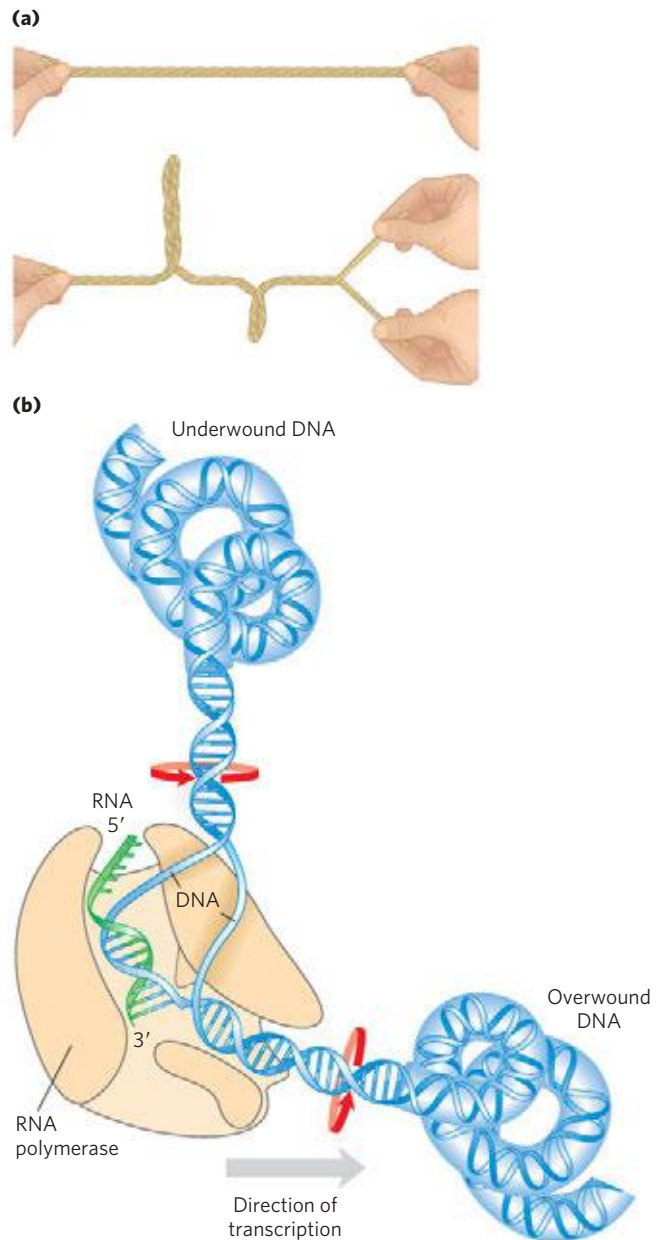


FIGURE 24-11 The effects of replication and transcription on DNA supercoiling. Because DNA is a double-helical structure, strand separation leads to added stress and supercoiling if the DNA is constrained (not free to rotate) ahead of the strand separation. **(a)** The general effect can be illustrated by twisting two strands of a rubber band about each other to form a double helix. If one end is constrained, separating the two strands at the other end will lead to twisting. **(b)** In a DNA molecule, the progress of a DNA polymerase or RNA polymerase (as shown here) along the DNA involves separation of the strands. As a result, the DNA becomes overwound ahead of the enzyme (upstream) and underwound behind it (downstream). Red arrows indicate the direction of winding.

Most Cellular DNA Is Underwound

To understand supercoiling, we must first focus on the properties of small circular DNAs such as plasmids and small viral DNAs. When these DNAs have no breaks in either strand, they are referred to as **closed-circular DNAs**. If the DNA of a closed-circular molecule conforms

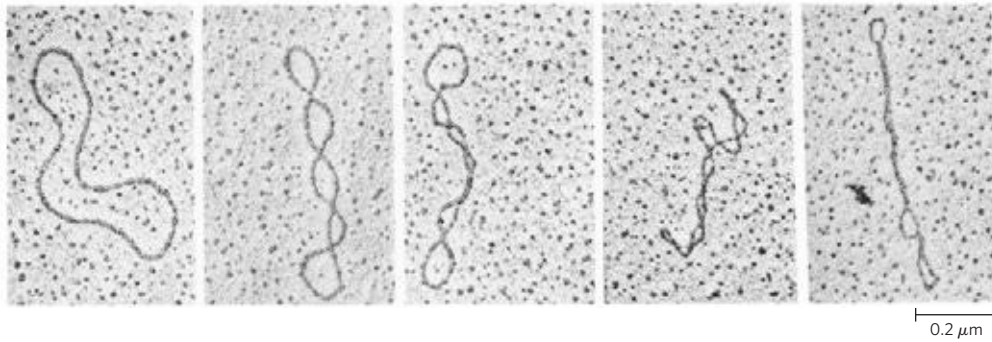


FIGURE 24-12 Relaxed and supercoiled plasmid DNAs. The molecule in the leftmost electron micrograph is relaxed; the degree of supercoiling increases from left to right.

closely to the B-form structure (the Watson-Crick structure; see Fig. 8-13), with one turn of the double helix per 10.5 bp, the DNA is relaxed rather than supercoiled (**Fig. 24-12**). Supercoiling results when DNA is subject to some form of structural strain. Purified closed-circular DNA is rarely relaxed, regardless of its biological origin. Furthermore, DNAs derived from a given cellular source have a characteristic degree of supercoiling. DNA structure is therefore strained in a manner that is regulated by the cell to induce the supercoiling.

In almost every instance, the strain is a result of **underwinding** of the DNA double helix in the closed circle. In other words, the DNA has *fewer* helical turns than would be expected for the B-form structure. The effects of underwinding are summarized in **Figure 24-13**. An 84 bp segment of a circular DNA in the relaxed state would contain eight double-helical turns, or one for every 10.5 bp. If one of these turns were removed, there would be $(84 \text{ bp})/7 = 12.0 \text{ bp per turn}$, rather than the 10.5 found in B-DNA (**Fig. 24-13b**). This is a deviation from the most stable DNA form, and the molecule is thermodynamically strained as a result. Generally, much of this strain would be accommodated by coiling the axis of the DNA on itself to form a supercoil (**Fig. 24-13c**; some of the strain in this 84 bp segment would simply become dispersed in the untwisted structure of the larger DNA molecule). In principle, the strain could also be accommodated by separating the two DNA strands over a distance of about 10 bp (**Fig. 24-13d**). In isolated closed-circular DNA, strain introduced by underwinding is generally accommodated by supercoiling rather than strand separation, because coiling the axis of the DNA usually requires less energy than breaking the hydrogen bonds that stabilize paired bases. Note, however, that the underwinding of DNA *in vivo* makes separation of the DNA strands easier, facilitating access to the information they contain.

Every cell actively underwinds its DNA with the aid of enzymatic processes (described below), and the resulting strained state represents a form of stored energy. Cells maintain DNA in an underwound state to facilitate its compaction by coiling. The underwinding of DNA is also important to enzymes of DNA metabolism

that must bring about strand separation as part of their function.

The underwound state can be maintained only if the DNA is a closed circle or if it is bound and stabilized by proteins so that the strands are not free to rotate about each other. If there is a break in one strand of an isolated, protein-free circular DNA, free rotation at that point will cause the underwound DNA to revert spontaneously to the relaxed state. In a closed-circular DNA molecule, however, the number of helical turns cannot be changed without at least transiently breaking one of

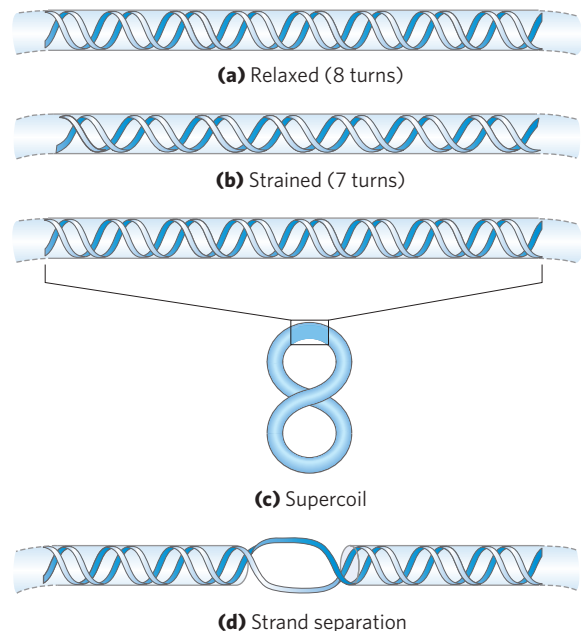


FIGURE 24-13 Effects of DNA underwinding. **(a)** A segment of DNA in a closed-circular molecule, 84 bp long, in its relaxed form with eight helical turns. **(b)** Removal of one turn induces structural strain. **(c)** The strain is generally accommodated by formation of a supercoil. **(d)** DNA underwinding also makes the separation of strands somewhat easier. In principle, each turn of underwinding should facilitate strand separation over about 10 bp, as shown. However, the hydrogen-bonded base pairs would generally preclude strand separation over such a short distance, and the effect becomes important only for longer DNAs and higher levels of DNA underwinding.

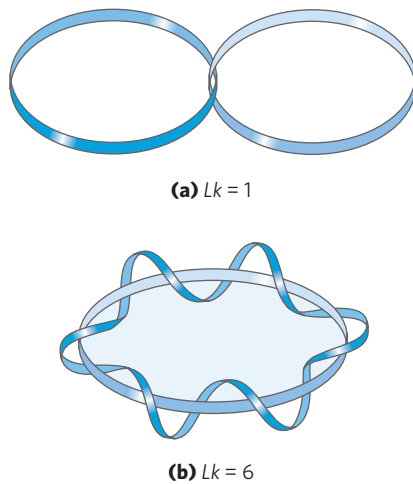


FIGURE 24-14 Linking number, Lk . Here, as usual, each blue ribbon represents one strand of a double-stranded DNA molecule. For the molecule in (a), $Lk = 1$. For the molecule in (b), $Lk = 6$. One of the strands in (b) is kept untwisted for illustrative purposes, to define the border of an imaginary surface (shaded blue). The number of times the twisting strand penetrates this surface provides a rigorous definition of linking number.

the DNA strands. The number of helical turns in a DNA molecule therefore provides a precise description of supercoiling.

DNA Underwinding Is Defined by Topological Linking Number

The field of topology provides some ideas that are useful to the discussion of DNA supercoiling, particularly the concept of **linking number**. Linking number is a topological property of double-stranded DNA, because it does not vary when the DNA is bent or deformed, as long as both DNA strands remain intact. Linking number (Lk) is illustrated in **Figure 24-14**.

Let's begin by visualizing the separation of the two strands of a double-stranded circular DNA. If the two strands are linked as shown in Figure 24-14a, they are effectively joined by what can be described as a topological bond. Even if all hydrogen bonds and base-stacking interactions were abolished such that the strands were not in physical contact, this topological bond would still link the two strands. Visualize one of the circular strands as the boundary of a surface (such as a soap film spanning the space framed by a circular wire before you blow a soap bubble). The linking number can be defined as the number of times the second strand pierces this surface. For the molecule in Figure 24-14a, $Lk = 1$; for that in Figure 24-14b, $Lk = 6$. The linking number for a closed-circular DNA is always an integer. By convention, if the links between two DNA strands are arranged so that the strands are interwound in a right-handed helix, the linking number is defined as positive (+); for strands interwound in a left-handed helix, the linking number is negative (-). Negative linking numbers are, for all practical purposes, not encountered in DNA.

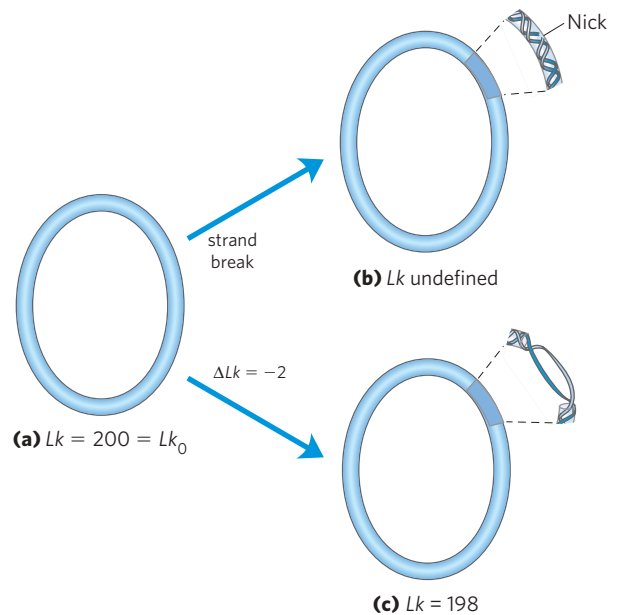


FIGURE 24-15 Linking number applied to closed-circular DNA molecules. A 2,100 bp circular DNA is shown in three forms: (a) relaxed, $Lk = 200$; (b) relaxed with a nick (break) in one strand, Lk undefined; and (c) underwound by two turns, $Lk = 198$. The underwound molecule generally exists as a supercoiled molecule, but underwinding also facilitates the separation of DNA strands.

We can now extend these ideas to a closed-circular DNA with 2,100 bp (**Fig. 24-15a**). When the molecule is relaxed, the linking number is simply the number of base pairs divided by the number of base pairs per turn, which is close to 10.5; so in this case, $Lk = 200$. For a circular DNA molecule to have a topological property such as linking number, neither strand may contain a break. If there is a break in either strand, the strands can, in principle, be unraveled and separated completely. In this case, no topological bond exists and Lk is undefined (Fig. 24-15b).

We can now describe DNA underwinding in terms of changes in the linking number. The linking number in relaxed DNA, Lk_0 , is used as a reference. For the molecule shown in Figure 24-15a, $Lk_0 = 200$; if two turns are removed from this molecule, $Lk = 198$. The change can be described by the equation

$$\begin{aligned}\Delta Lk &= Lk - Lk_0 \\ &= 198 - 200 = -2\end{aligned}\quad (24-1)$$

It is often convenient to express the change in linking number in terms of a quantity that is independent of the length of the DNA molecule. This quantity, called the **specific linking difference** or **superhelical density** (σ), is a measure of the number of turns removed relative to the number present in relaxed DNA:

$$\sigma = \frac{\Delta Lk}{Lk_0}\quad (24-2)$$

In the example in Figure 24-15c, $\sigma = -0.01$, which means that 1% (2 of 200) of the helical turns present in the DNA (in its B form) have been removed. The degree

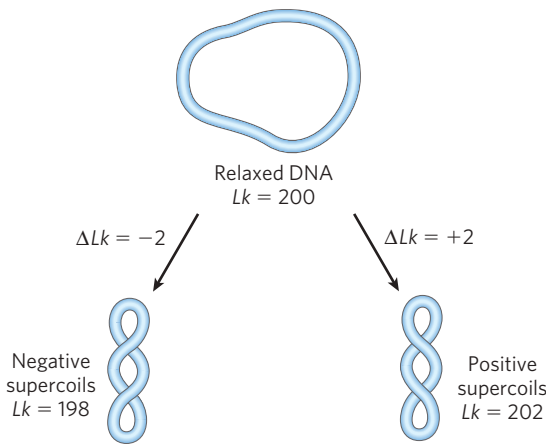


FIGURE 24-16 Negative and positive supercoils. For the relaxed DNA molecule of Figure 24-15a, underwinding or overwinding by two helical turns ($Lk = 198$ or 202) will produce negative or positive supercoiling, respectively. Note that the DNA axis twists in opposite directions in the two cases.

of underwinding in cellular DNAs generally falls in the range of 5% to 7%; that is, $\sigma = -0.05$ to -0.07 . The negative sign indicates that the change in linking number is due to underwinding of the DNA. The supercoiling induced by underwinding is therefore defined as negative supercoiling. Conversely, under some conditions DNA can be overwound, resulting in positive supercoiling. Note that the twisting path taken by the axis of the DNA helix when the DNA is underwound (negative supercoiling) is the mirror image of that taken when the DNA is overwound (positive supercoiling) (**Fig. 24-16**). Supercoiling is not a random process; the path of the supercoiling is largely prescribed by the torsional strain imparted to the DNA by decreasing or increasing the linking number relative to B-DNA.

Linking number can be changed by ± 1 by breaking one DNA strand, rotating one of the ends 360° about the unbroken strand, and rejoining the broken ends. This change has no effect on the number of base pairs or the number of atoms in the circular DNA molecule. Two forms of a circular DNA that differ only in a topological property such as linking number are referred to as **topoisomers**.

WORKED EXAMPLE 24-1 Calculation of Superhelical Density

What is the superhelical density (σ) of a closed-circular DNA with a length of 4,200 bp and a linking number (Lk) of 374? What is the superhelical density of the same DNA when $Lk = 412$? Are these molecules negatively or positively supercoiled?

Solution: First, calculate Lk_0 by dividing the length of the closed-circular DNA (in bp) by 10.5 bp/turn: $(4,200 \text{ bp}) / (10.5 \text{ bp/turn}) = 400$. We can now calculate ΔLk from Equation 24-1: $\Delta Lk = Lk - Lk_0 = 374 - 400 = -26$. Substituting the values for ΔLk and Lk_0 into Equation

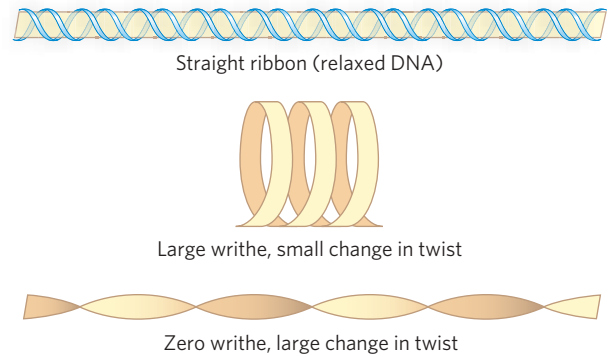


FIGURE 24-17 Ribbon model for illustrating twist and writhe. The tan ribbon represents the axis of a relaxed DNA molecule. Strain introduced by twisting the ribbon (underwinding the DNA) can be manifested as writhe or twist. Topological changes in linking number are usually accompanied by geometric changes in both writhe and twist.

24-2: $\sigma = \Delta Lk / Lk_0 = -26 / 400 = -0.065$. Since the superhelical density is negative, this DNA molecule is negatively supercoiled.

When the same DNA molecule has an Lk of 412, $\Delta Lk = 412 - 400 = 12$ and $\sigma = 12 / 400 = 0.03$. The superhelical density is positive, and the molecule is positively supercoiled.

Linking number can be broken down into two structural components, **twist (Tw)** and **writhe (Wr)** (**Fig. 24-17**). These are more difficult to describe than linking number, but writhe may be thought of as a measure of the coiling of the helix axis, and twist as determining the local twisting or spatial relationship of neighboring base pairs. When the linking number changes, some of the resulting strain is usually compensated for by writhe (supercoiling) and some by changes in twist, giving rise to the equation

$$Lk = Tw + Wr$$

Tw and Wr need not be integers. Twist and writhe are geometric rather than topological properties, because they may be changed by deformation of a closed-circular DNA molecule.

In addition to causing supercoiling and making strand separation somewhat easier, the underwinding of DNA facilitates structural changes in the molecule. These are of less physiological importance but help illustrate the effects of underwinding. Recall that a cruciform (see Fig. 8-19) generally contains a few unpaired bases; DNA underwinding helps to maintain the required strand separation (**Fig. 24-18**). Underwinding of a right-handed DNA helix also facilitates the formation of short stretches of left-handed Z-DNA in regions where the base sequence is consistent with the Z form (see Chapter 8).

Topoisomerases Catalyze Changes in the Linking Number of DNA

DNA supercoiling is a precisely regulated process that influences many aspects of DNA metabolism. Every cell

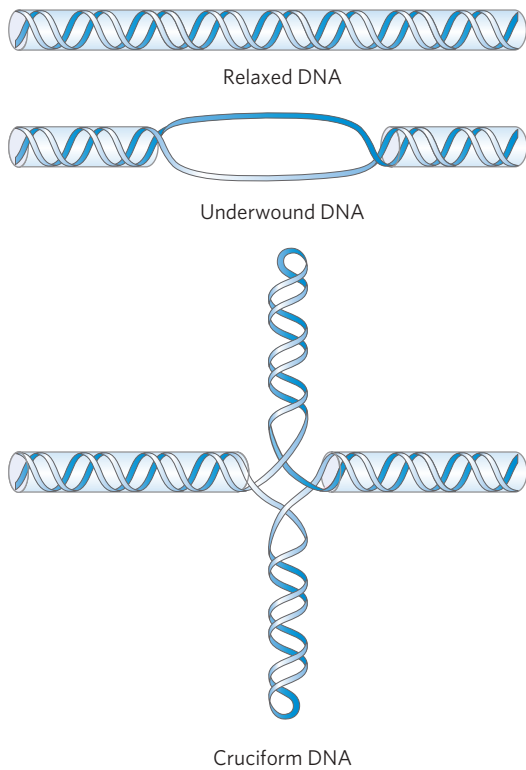


FIGURE 24-18 Promotion of cruciform structures by DNA underwinding. In principle, cruciforms can form at palindromic sequences (see Fig. 8-19), but they seldom occur in relaxed DNA because the linear DNA accommodates more paired bases than does the cruciform structure. Underwinding of the DNA facilitates the partial strand separation needed to promote cruciform formation at appropriate sequences.

has enzymes with the sole function of underwinding and/or relaxing DNA. The enzymes that increase or decrease the extent of DNA underwinding are **topoisomerases**; the property of DNA that they change is the linking number. These enzymes play an especially important role in processes such as replication and DNA packaging. There are two classes of topoisomerases. **Type I topoisomerases** act by transiently breaking one of the two DNA strands, passing the unbroken strand through the break and rejoining the broken ends; they change Lk in increments of 1. **Type II topoisomerases** break both DNA strands and change Lk in increments of 2.

The effects of these enzymes can be demonstrated with agarose gel electrophoresis (**Fig. 24-19**). A population of identical plasmid DNAs with the same linking number migrates as a discrete band during electrophoresis. Topoisomers with Lk values differing by as little as 1 can be separated by this method, so changes in linking number induced by topoisomerases are readily detected.

E. coli has at least four individual topoisomerases (I through IV). Those of type I (topoisomerases I and III) generally relax DNA by removing negative supercoils (increasing Lk). The way in which bacterial type I topoisomerases change linking number is illustrated in **Figure 24-20**. A bacterial type II enzyme, called either topoisomerase II or DNA gyrase, can introduce negative

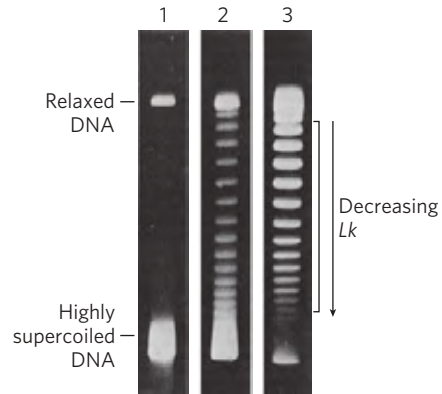


FIGURE 24-19 Visualization of topoisomers. In this experiment, all DNA molecules have the same number of base pairs but exhibit some range in the degree of supercoiling. Because supercoiled DNA molecules are more compact than relaxed molecules, they migrate more rapidly during gel electrophoresis. The gels shown here separate topoisomers (moving from top to bottom) over a limited range of superhelical density. In lane 1, highly supercoiled DNA migrates in a single band, even though different topoisomers are probably present. Lanes 2 and 3 illustrate the effect of treating the supercoiled DNA with a type I topoisomerase; the DNA in lane 3 was treated for a longer time than that in lane 2. As the superhelical density of the DNA is reduced to the point where it corresponds to the range in which the gel can resolve individual topoisomers, distinct bands appear. Individual bands in the region indicated by the bracket next to lane 3 each contain DNA circles with the same linking number; the linking number changes by 1 from one band to the next.

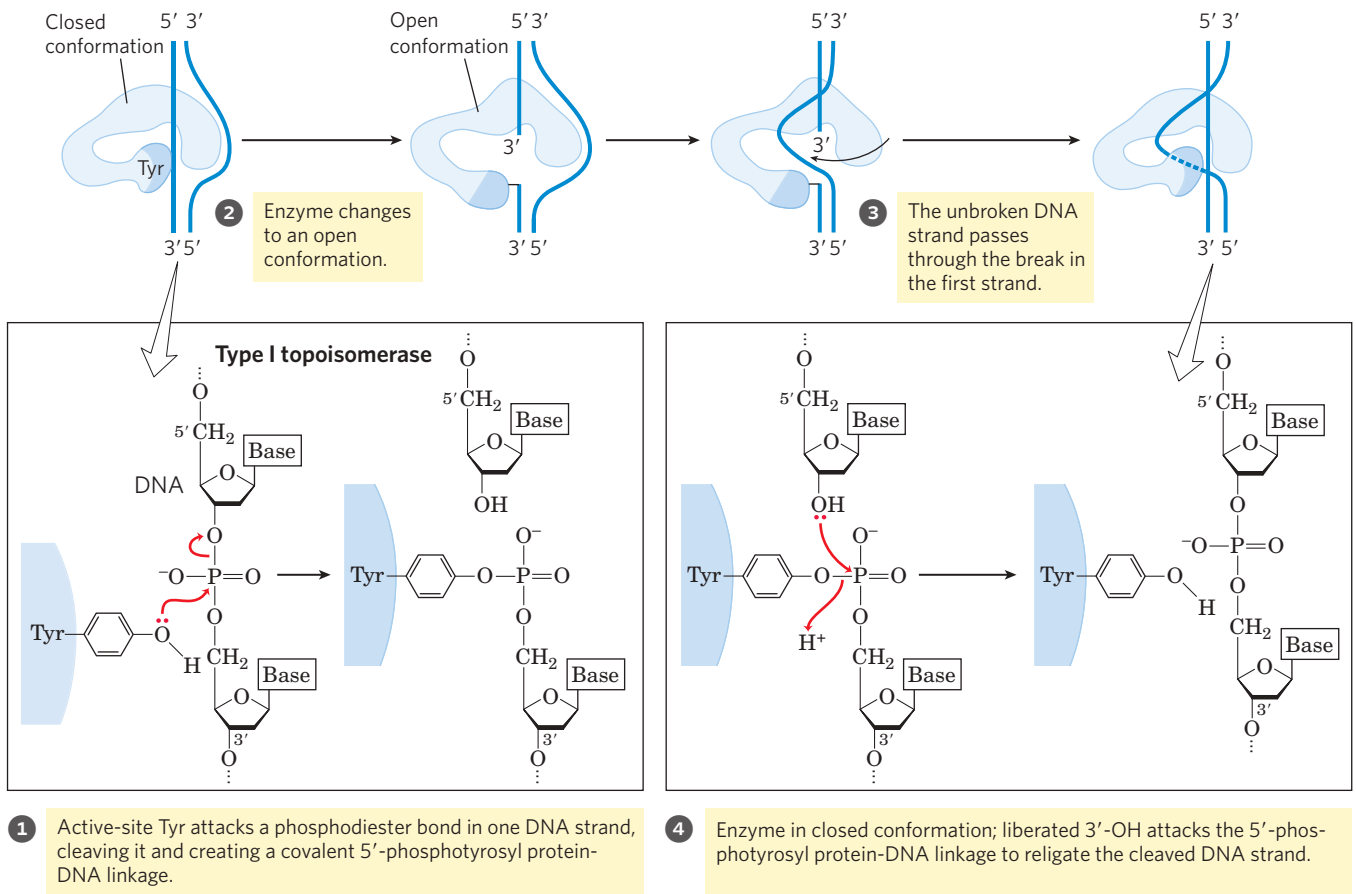
supercoils (decrease Lk). It uses the energy of ATP to accomplish this. To alter DNA linking number, type II topoisomerases cleave both strands of a DNA molecule and pass another duplex through the break. The degree of supercoiling of bacterial DNA is maintained by regulation of the net activity of topoisomerases I and II.

Eukaryotic cells also have type I and type II topoisomerases. The type I enzymes are topoisomerases I and III; the single type II enzyme has two isoforms in vertebrates, called $II\alpha$ and $II\beta$. Most of the type II enzymes, including a DNA gyrase in archaea, are related and define a family called type IIA. Archaea also have an unusual enzyme, topoisomerase VI, which alone defines the type IIB family. The eukaryotic type II topoisomerases cannot underwind DNA (introduce negative supercoils), but they can relax both positive and negative supercoils (**Fig. 24-21**).

As we will show in the next few chapters, topoisomerases play a critical role in every aspect of DNA metabolism. As a consequence, they are important drug targets for the treatment of bacterial infections and cancer (Box 24-1).

DNA Compaction Requires a Special Form of Supercoiling

Supercoiled DNA molecules are uniform in a number of respects. The supercoils are right-handed in a negatively supercoiled DNA molecule (**Fig. 24-16**), and they



MECHANISM FIGURE 24-20 The type I topoisomerase reaction. Bacterial topoisomerase I increases Lk by breaking one DNA strand, passing the unbroken strand through the break, then resealing the break.

Nucleophilic attack by the active-site Tyr residue breaks one DNA strand. The ends are ligated by a second nucleophilic attack. At each step, one high-energy bond replaces another.

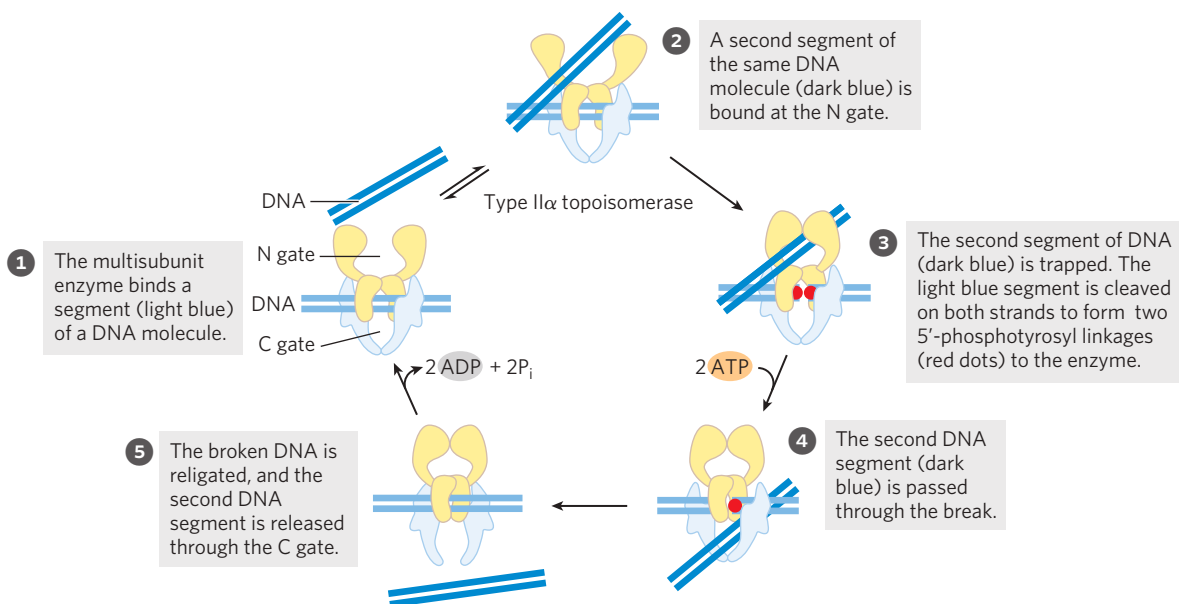


FIGURE 24-21 Alteration of the linking number by a eukaryotic type II α topoisomerase. The general mechanism features the passage of one intact duplex DNA segment through a transient double-strand break in another segment. The DNA segment enters and leaves the topoisomerase

through gated cavities above and below the bound DNA, which are called the N gate and the C gate. Two ATPs are bound and hydrolyzed during this cycle. The enzyme structure and use of ATP are specific to this reaction.

BOX 24-1 MEDICINE Curing Disease by Inhibiting Topoisomerases

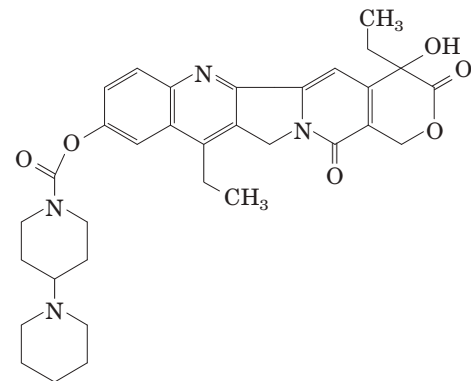
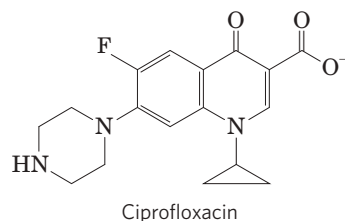
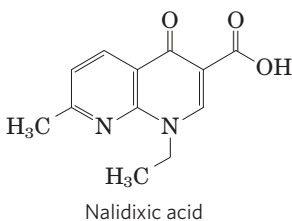
The topological state of cellular DNA is intimately connected with its function. Without topoisomerases, cells cannot replicate or package their DNA, or express their genes—and they die. Inhibitors of topoisomerases have therefore become important pharmaceutical agents, targeted at infectious agents and malignant cells.

Two classes of bacterial topoisomerase inhibitors have been developed as antibiotics. The coumarins, including novobiocin and coumermycin A1, are natural products derived from *Streptomyces* species. They inhibit the ATP binding of the bacterial type II topoisomerases, DNA gyrase and topoisomerase IV. These antibiotics are not often used to treat infections in humans, but research continues to identify clinically effective variants.

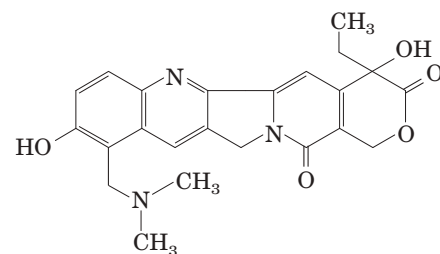
The quinolone antibiotics, also inhibitors of bacterial DNA gyrase and topoisomerase IV, first appeared in 1962 with the introduction of nalidixic acid. This compound had limited effectiveness and is no longer used clinically in the United States, but the continued development of this class of drugs eventually led to the introduction of the fluoroquinolones, exemplified by ciprofloxacin (Cipro). The quinolones act by blocking the last step of the topoisomerase reaction, the resealing of the DNA strand breaks. Ciprofloxacin is a wide-spectrum antibiotic. It is one of the few antibiot-

ics reliably effective in treating anthrax infections and is considered a valuable agent in protection against possible bioterrorism. Quinolones are selective for the bacterial topoisomerases, inhibiting the eukaryotic enzymes only at concentrations several orders of magnitude greater than the therapeutic doses.

Some of the most important chemotherapeutic agents used in cancer treatment are inhibitors of human topoisomerases. Topoisomerases are generally



Irinotecan



Topotecan

tend to be extended and narrow rather than compacted, often with multiple branches (**Fig. 24-22**). At the superhelical densities normally encountered in cells, the length of the supercoil axis, including branches, is about 40% of the length of the DNA. This type of super-

coiling is referred to as **plectonemic** (from the Greek *plektos*, “twisted,” and *nema*, “thread”). This term can be applied to any structure with strands intertwined in some simple and regular way, and it is a good description of the general structure of supercoiled DNA in solution.

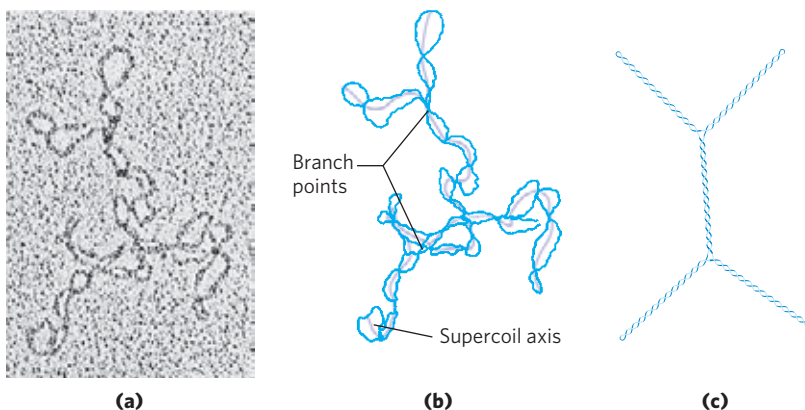


FIGURE 24-22 Plectonemic supercoiling. (a) Electron micrograph of plectonemically supercoiled plasmid DNA and (b) an interpretation of the observed structure. The purple lines show the axis of the supercoil; note the branching of the supercoil. (c) An idealized representation of this structure.

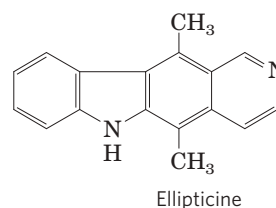
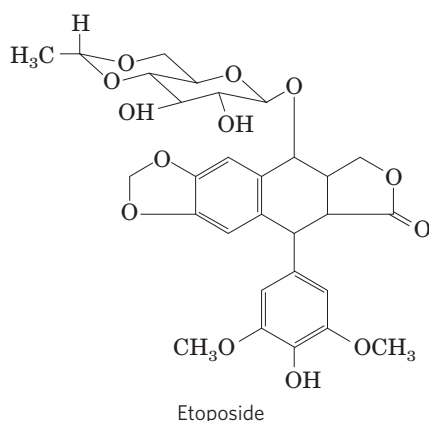
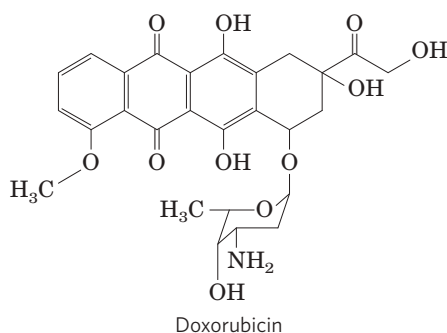
present at elevated levels in tumor cells, and agents targeted to these enzymes are much more toxic to the tumors than to most other tissue types. Inhibitors of both type I and type II topoisomerases have been developed as anticancer drugs.

Camptothecin, isolated from a Chinese ornamental tree and first tested clinically in the 1970s, is an inhibitor of eukaryotic type I topoisomerases. Clinical

trials indicated limited effectiveness, despite its early promise in preclinical work on mice. However, two effective derivatives, irinotecan (Campto) and topotecan (Hycamtin)—used to treat colorectal cancer and ovarian cancer, respectively—were developed in the 1990s. Additional derivatives are likely to be approved for clinical use in the coming years. All of these drugs act by trapping the topoisomerase-DNA complex in which the DNA is cleaved, inhibiting religation.

The human type II topoisomerases are targeted by a variety of antitumor drugs, including doxorubicin (Adriamycin), etoposide (Etopophos), and ellipticine. Doxorubicin, effective against several kinds of human tumors, is an anthracycline in clinical use. Most of these drugs stabilize the covalent topoisomerase-DNA (cleaved) complex.

All of these anticancer agents generally increase the levels of DNA damage in the targeted, rapidly growing tumor cells. However, noncancerous tissues can also be affected, leading to a more general toxicity and unpleasant side effects that must be managed during therapy. As cancer therapies become more effective and survival statistics for cancer patients improve, the independent appearance of new tumors is becoming a greater problem. In the continuing search for new cancer therapies, the topoisomerases are likely to remain prominent targets for research.



Plectonemic supercoiling, the form observed in isolated DNAs in the laboratory, does not produce sufficient compaction to package DNA in the cell. A second form of supercoiling, **solenoidal (Fig. 24–23)**, can be adopted by an underwound DNA. Instead of the extended right-handed supercoils characteristic of the plectonemic form, solenoidal supercoiling involves tight left-handed turns, similar to the shape taken up by a garden hose neatly wrapped on a reel. Although their structures are dramatically different, plectonemic and solenoidal supercoiling are two forms of negative supercoiling that can be taken up by the *same* segment of underwound DNA. The two forms are readily interconvertible. Although the plectonemic form is more stable in solution, the solenoidal form can be stabilized by protein binding as it is in eukaryotic chromosomes. It provides a much greater degree of compaction (Fig. 24–23). Solenoidal supercoiling is the mechanism by which underwinding contributes to DNA compaction.

SUMMARY 24.2 DNA Supercoiling

- ▶ Most cellular DNAs are supercoiled. Underwinding decreases the total number of helical turns in the DNA relative to the relaxed, B form. To maintain an underwound state, DNA must be either a closed circle or bound to protein. Underwinding is quantified by a topological parameter called linking number, Lk .
- ▶ Underwinding is measured in terms of specific linking difference, σ (also called superhelical density), which is $(Lk - Lk_0)/Lk_0$. For cellular DNAs, σ is typically -0.05 to -0.07 , which means that approximately 5% to 7% of the helical turns in the DNA have been removed. DNA underwinding facilitates strand separation by enzymes of DNA metabolism.
- ▶ DNAs that differ only in linking number are called topoisomers. Enzymes that underwind

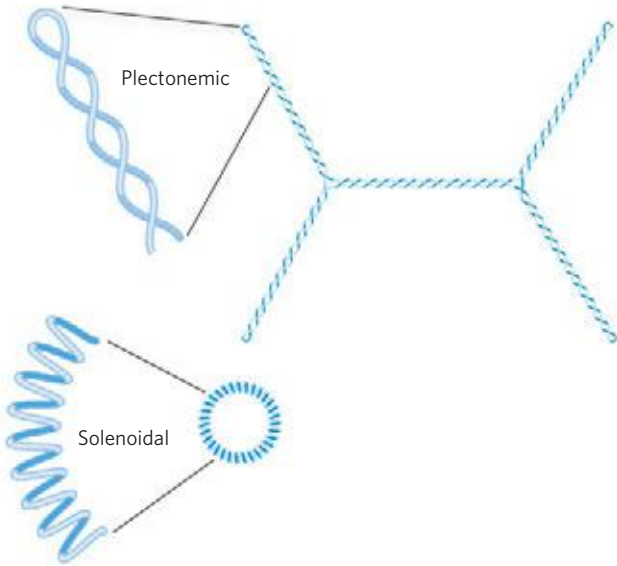


FIGURE 24-23 Plectonemic and solenoidal supercoiling of the same DNA molecule, drawn to scale. Plectonemic supercoiling takes the form of extended right-handed coils. Solenoidal negative supercoiling takes the form of tight left-handed turns about an imaginary tubelike structure. The two forms are readily interconverted, although the solenoidal form is generally not observed unless certain proteins are bound to the DNA. Solenoidal supercoiling provides a much greater degree of compaction.

and/or relax DNA, the topoisomerases, catalyze changes in linking number. The two classes of topoisomerases, type I and type II, change Lk in increments of 1 or 2, respectively, per catalytic event.

24.3 The Structure of Chromosomes

The term “chromosome” is used to refer to a nucleic acid molecule that is the repository of genetic information in a virus, a bacterium, a eukaryotic cell, or an organelle. It also refers to the densely colored bodies seen in the nuclei of dye-stained eukaryotic cells, as visualized using a light microscope.

Chromatin Consists of DNA and Proteins

The eukaryotic cell cycle (see Fig. 12-44) produces remarkable changes in the structure of chromosomes (**Fig. 24-24**). In nondividing eukaryotic cells (in G_0) and those in interphase (G_1 , S , and G_2), the chromosomal material, **chromatin**, is amorphous and seems to be randomly dispersed in certain parts of the nucleus. In the S phase of interphase the DNA in this amorphous state replicates, each chromosome producing two sister chromosomes (called sister chromatids) that remain associated with each other after replication is complete. The chromosomes become much more condensed

during prophase of mitosis, taking the form of a species-specific number of well-defined pairs of sister chromatids (**Fig. 24-5**).

Chromatin consists of fibers containing protein and DNA in approximately equal proportions (by mass), along with a small amount of RNA. The DNA in the chromatin is very tightly associated with proteins called **histones**, which package and order the DNA into structural units called **nucleosomes** (**Fig. 24-25**). Also found in chromatin are many nonhistone proteins, some of which help maintain chromosome structure

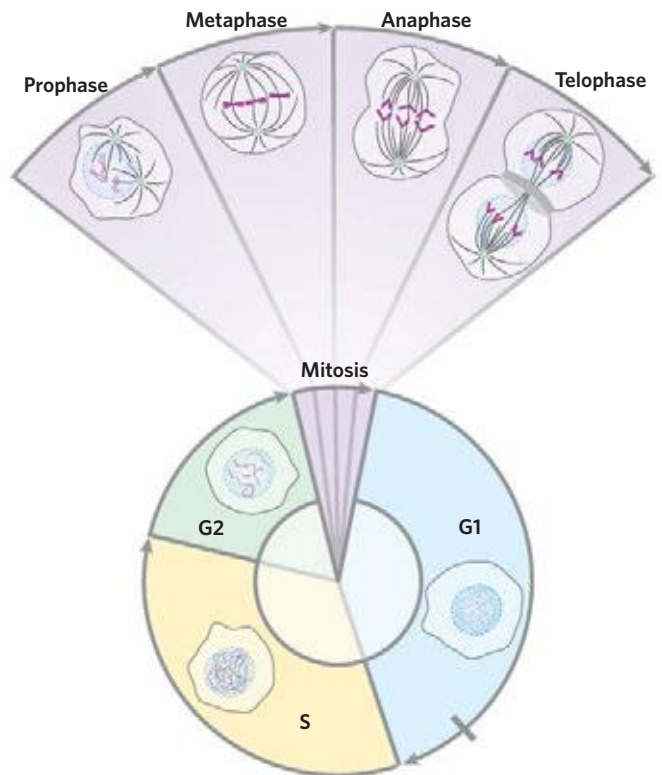


FIGURE 24-24 Changes in chromosome structure during the eukaryotic cell cycle. The relative lengths of the phases shown here are for convenience only. The duration of each phase varies with cell type and growth conditions (for single-celled organisms) or metabolic state (for multicellular organisms); mitosis is typically the shortest. Cellular DNA is uncondensed throughout interphase, as shown in the cartoons of the nucleus in the diagram. The interphase period can be divided (see Fig. 12-44) into the G_1 (gap) phase; the S (synthesis) phase, when the DNA is replicated; and the G_2 phase, throughout which the replicated chromosomes (chromatids) cohere to each other. Mitosis can be divided into four stages. The DNA undergoes condensation in prophase. During metaphase, the condensed chromosomes line up in pairs along the plane halfway between the spindle poles. The two chromosomes of each pair are linked to different spindle poles via microtubules that extend between the spindle and the centromere. The sister chromatids separate at anaphase, each drawn toward the spindle pole to which it is connected. The process is completed in telophase. After cell division is complete, the chromosomes decondense and the cycle begins anew.

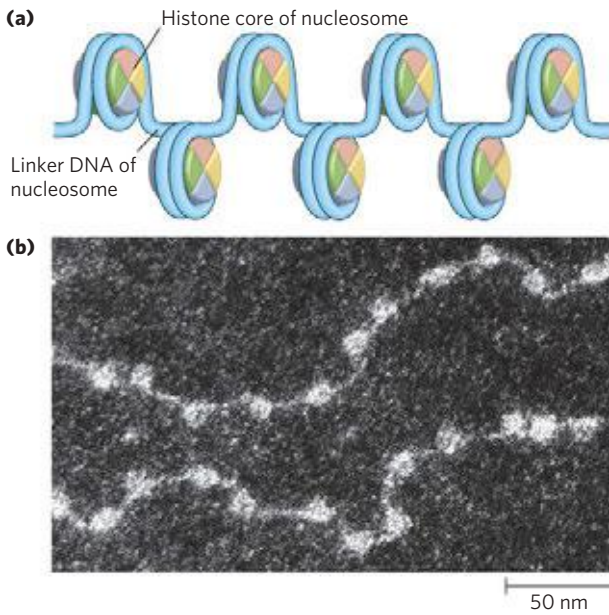


FIGURE 24-25 Nucleosomes. (a) Regularly spaced nucleosomes consist of core histone proteins bound to DNA. (b) In this electron micrograph, the DNA-wrapped histone octamer structures are clearly visible.

and others that regulate the expression of specific genes (Chapter 28). Beginning with nucleosomes, eukaryotic chromosomal DNA is packaged into a succession of higher-order structures that ultimately yield the compact chromosome seen with the light microscope. We now turn to a description of this structure in eukaryotes and compare it with the packaging of DNA in bacterial cells.

Histones Are Small, Basic Proteins

Found in the chromatin of all eukaryotic cells, histones have molecular weights between 11,000 and 21,000 and are very rich in the basic amino acids arginine and lysine (together these make up about one-fourth of the amino acid residues). All eukaryotic cells have five major classes of histones, differing in molecular weight and amino acid composition (Table 24-4). The H3 histones

are nearly identical in amino acid sequence in all eukaryotes, as are the H4 histones, suggesting strict conservation of their functions. For example, only 2 of 102 amino acid residues differ between the H4 histone molecules of peas and cows, and only 8 differ between the H4 histones of humans and yeast. Histones H1, H2A, and H2B show less sequence similarity among eukaryotic species.

Each type of histone is subject to enzymatic modification by methylation, acetylation, ADP-ribosylation, phosphorylation, glycosylation, sumoylation, or ubiquitination. Such modifications affect the net electric charge, shape, and other properties of histones, as well as the structural and functional properties of the chromatin, and they play a role in the regulation of transcription.

In addition, eukaryotes generally have several variant forms of certain histones, most notably histones H2A and H3, described in more detail below. The variant forms, along with their modifications, have specialized roles in DNA metabolism.

Nucleosomes Are the Fundamental Organizational Units of Chromatin

The eukaryotic chromosome depicted in Figure 24-5 represents the compaction of a DNA molecule about $10^5 \mu\text{m}$ long into a cell nucleus that is typically 5 to $10 \mu\text{m}$ in diameter. This compaction is achieved by means of several levels of highly organized folding. Subjection of chromosomes to treatments that partially unfold them reveals a structure in which the DNA is bound tightly to beads of protein, often regularly spaced. The beads in this “beads-on-a-string” arrangement are complexes of histones and DNA. The bead plus the connecting DNA that leads to the next bead form the nucleosome, the fundamental unit of organization on which the higher-order packing of chromatin is built (Fig. 24-26). The bead of each nucleosome contains eight histone molecules: two copies each of H2A, H2B, H3, and H4. The spacing of the nucleosome beads provides a repeating unit typically of about 200 bp, of which 146 bp are bound tightly around the eight-part histone core and the remainder serve as linker DNA between nucleosome beads.

Histone H1 binds to the linker DNA. Brief treatment of chromatin with enzymes that digest DNA causes the linker DNA to degrade preferentially, releasing histone particles containing 146 bp of bound DNA that have been protected from digestion. Researchers have crystallized nucleosome cores obtained in this way, and x-ray diffraction analysis reveals a particle made up of the eight histone molecules with the DNA wrapped around

TABLE 24-4 Types and Properties of the Common Histones

Histone	Molecular weight	Number of amino acid residues	Content of basic amino acids (% of total)	
			Lys	Arg
H1*	21,130	223	29.5	11.3
H2A*	13,960	129	10.9	19.3
H2B*	13,774	125	16.0	16.4
H3	15,273	135	19.6	13.3
H4	11,236	102	10.8	13.7

*The sizes of these histones vary somewhat from species to species. The numbers given here are for bovine histones.

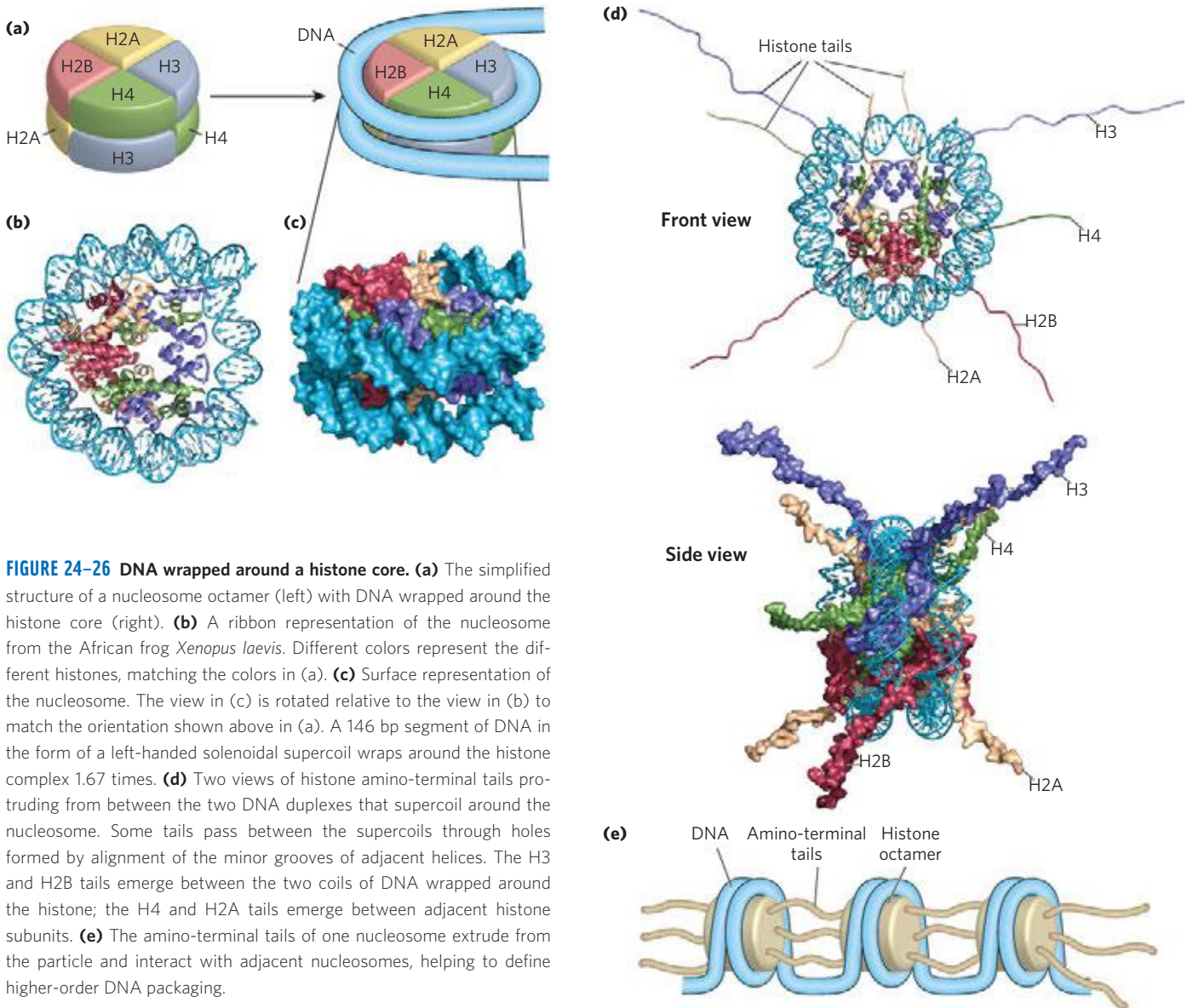


FIGURE 24-26 DNA wrapped around a histone core. (a) The simplified structure of a nucleosome octamer (left) with DNA wrapped around the histone core (right). (b) A ribbon representation of the nucleosome from the African frog *Xenopus laevis*. Different colors represent the different histones, matching the colors in (a). (c) Surface representation of the nucleosome. The view in (c) is rotated relative to the view in (b) to match the orientation shown above in (a). A 146 bp segment of DNA in the form of a left-handed solenoidal supercoil wraps around the histone complex 1.67 times. (d) Two views of histone amino-terminal tails protruding from between the two DNA duplexes that supercoil around the nucleosome. Some tails pass between the supercoils through holes formed by alignment of the minor grooves of adjacent helices. The H3 and H2B tails emerge between the two coils of DNA wrapped around the histone; the H4 and H2A tails emerge between adjacent histone subunits. (e) The amino-terminal tails of one nucleosome extrude from the particle and interact with adjacent nucleosomes, helping to define higher-order DNA packaging.

it in the form of a left-handed solenoidal supercoil (Fig. 24-26). Extending out from the nucleosome core are the amino-terminal tails of the histones, which are intrinsically disordered (Fig. 24-26d). Most of the histone modifications occur in these tails. The tails, in turn, play a key role in forming contacts between nucleosomes in the chromatin (Fig. 24-26e).

A close inspection of this structure reveals why eukaryotic DNA is underwound even though eukaryotic cells lack enzymes that underwind DNA. Recall that the solenoidal wrapping of DNA in nucleosomes is but one form of supercoiling that can be taken up by underwound (negatively supercoiled) DNA. The tight wrapping of DNA around the histone core requires the removal of about one helical turn in the DNA. When the protein core of a nucleosome binds *in vitro* to a relaxed, closed-circular DNA, the binding introduces a negative supercoil. Because this binding process does

not break the DNA or change the linking number, the formation of a negative solenoidal supercoil must be accompanied by a compensatory positive supercoil in the unbound region of the DNA (Fig. 24-27). As mentioned earlier, eukaryotic topoisomerases can relax positive supercoils. Relaxing the unbound positive supercoil leaves the negative supercoil fixed (through its binding to the nucleosome histone core) and results in an overall decrease in linking number. Indeed, topoisomerases have proved necessary for assembling chromatin from purified histones and closed-circular DNA *in vitro*.

Another factor that affects the binding of DNA to histones in nucleosome cores is the sequence of the bound DNA. Histone cores do not bind at random positions on the DNA; rather, some locations are more likely to be bound than others. This positioning is not fully understood but in some cases seems to depend on a

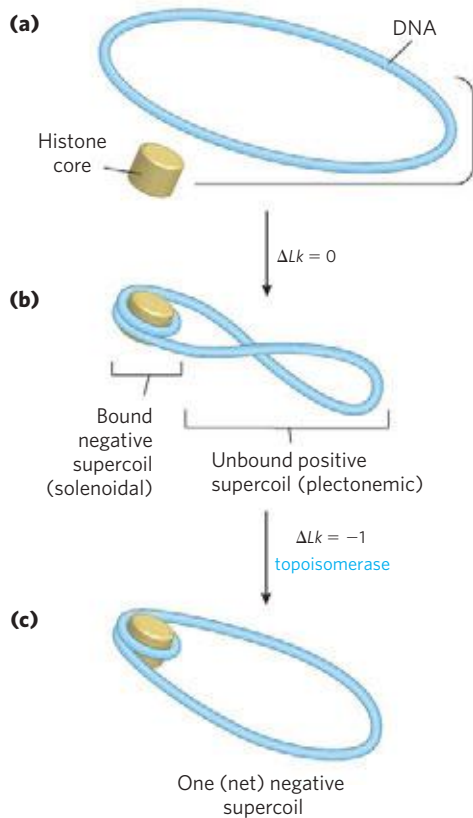


FIGURE 24-27 Chromatin assembly. (a) Relaxed, closed-circular DNA. (b) Binding of a histone core to form a nucleosome induces one negative supercoil. In the absence of any strand breaks, a positive supercoil must form elsewhere in the DNA ($\Delta Lk = 0$). (c) Relaxation of this positive supercoil by a topoisomerase leaves one net negative supercoil ($\Delta Lk = -1$).

local abundance of A=T base pairs in the DNA helix where it is in contact with the histones (**Fig. 24-28**). A cluster of two or three A=T base pairs facilitates the compression of the minor groove that is needed for the DNA to wrap tightly around the nucleosome's histone core. Nucleosomes bind particularly well to sequences where AA or AT or TT dinucleotides are staggered at 10 bp intervals, an arrangement that can account for up to 50% of the positions of bound histones in vivo.

Other proteins are required for the positioning of some nucleosome cores on DNA. In several organisms, certain proteins bind to a specific DNA sequence and facilitate the formation of a nucleosome core nearby. Nucleosomes are deposited on the DNA during replication, or following other processes that require a transient displacement of nucleosomes. Nucleosomes seem to be deposited in a stepwise manner. A tetramer of two H3 and two H4 histones binds first, followed by H2A-H2B dimers. The incorporation of nucleosomes into chromosomes after chromosomal replication is mediated by a complex of histone chaperones that include proteins known as chromatin assembly factor 1 (CAF1), RTT106 (*regulation of Ty1 transposition*), and anti-silencing factor 1 (ASF1).

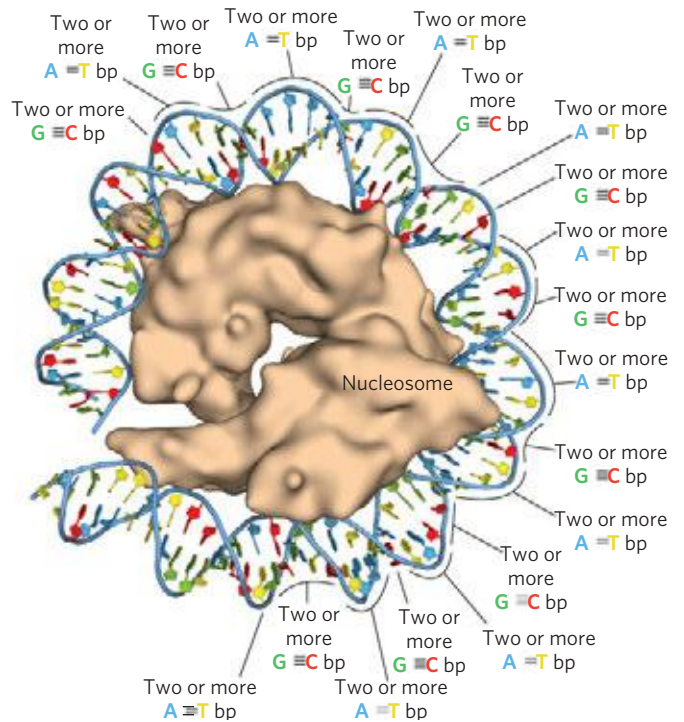


FIGURE 24-28 The effect of DNA sequence on nucleosome binding. (PDB ID 1AOI) Runs of two or more A=T base pairs facilitate the bending of DNA, while runs of two or more G=C base pairs have the opposite effect. When spaced at about 10 bp intervals, consecutive A=T base pairs help bend DNA into a circle. When consecutive G=C base pairs are spaced 10 bp apart, and offset by 5 bp from runs of A=T base pairs, DNA binding to the chromosome is facilitated.

These bind to acetylated variants of histones H3 and H4. The mechanism of nucleosome deposition is not understood in detail, although parts of this complex are known to directly interact with parts of the replication machinery. Some of the same histone chaperones, or different ones, may help assemble nucleosomes after DNA repair, transcription, or other processes. Histone exchange factors permit the substitution of histone variants for core histones in some contexts. Proper placement of these variant histones is important. Studies have shown that mice lacking one of these variant histones die as early embryos (Box 24-2). Precise positioning of nucleosome cores also plays a role in the expression of some eukaryotic genes (Chapter 28).

Nucleosomes Are Packed into Successively Higher-Order Structures

Wrapping of DNA around a nucleosome core compacts the DNA length about sevenfold. The overall compaction in a chromosome, however, is greater than 10,000-fold—ample evidence for even higher orders of structural organization. In chromosomes isolated by very gentle methods, nucleosome cores seem to be organized into a

BOX 24-2 METHODS Epigenetics, Nucleosome Structure, and Histone Variants

Information that is passed from one generation to the next—to daughter cells at cell division or from parent to offspring—but is not encoded in DNA sequences is referred to as **epigenetic** information. Much of it is in the form of covalent modification of histones and/or the placement of histone variants in chromosomes.

The chromatin regions where active gene expression (transcription) is occurring tend to be partially decondensed and are called **euchromatin**. In these regions, histones H3 and H2A are often replaced by the histone variants H3.3 and H2AZ, respectively (Fig. 1). The complexes that deposit nucleosomes containing histone variants on the DNA are similar to those that deposit nucleosomes with the more common histones. Nucleosomes containing histone H3.3 are deposited by a complex in which chromatin assembly factor 1 (CAF1) is replaced by the protein HIRA (the name is derived from a class of proteins called HIR, for *histone repressor*). Both CAF1 and HIRA can be considered histone chaperones, helping to ensure the proper assembly and placement of nucleosomes. Histone H3.3 differs in sequence from H3 by only four amino acid residues, but these residues all play key roles in histone deposition.

Like histone H3.3, H2AZ is associated with a distinct nucleosome deposition complex, and it is generally associated with chromatin regions involved in active transcription. Incorporation of H2AZ stabilizes the nucleosome octamer, but impedes some cooperative interactions between nucleosomes that are needed to compact the chromosome. This leads to a more open chromosome structure that facilitates the expression of genes in the region where H2AZ is located. The gene encoding H2AZ is essential in mammals. In fruit flies, loss of H2AZ prevents development beyond the larval stages.

Another H2A variant is H2AX, which is associated with DNA repair and genetic recombination. In mice, the absence of H2AX results in genome instability and male infertility. Modest amounts of H2AX seem to be scattered throughout the genome. When a double-strand break occurs, nearby molecules of H2AX become phosphorylated at Ser¹³⁹ in the carboxyl-terminal region. If this phosphorylation is blocked experimentally, formation of the protein complexes necessary for DNA repair is inhibited.

The H3 histone variant known as CENPA is associated with the repeated DNA sequences in centromeres. The chromatin in the centromere region contains the histone chaperones CAF1 and HIRA, and both proteins could be involved in the deposition of nucleosomes containing CENPA. Elimination of the gene for CENPA is lethal in mice.

The function and positioning of the histone variants can be studied by an application of technologies used in genomics. One useful technology is chromatin immunoprecipitation, or chromatin IP (ChIP). Nucleosomes containing a particular histone variant are precipitated by an antibody that binds specifically to this variant. These nucleosomes can be studied in isolation from their DNA, but more commonly the DNA associated with them is included in the study to determine where the nucleosomes of interest bind. The DNA can be labeled and used to probe a microarray (see Fig. 9-23), yielding a map of genomic sequences to which

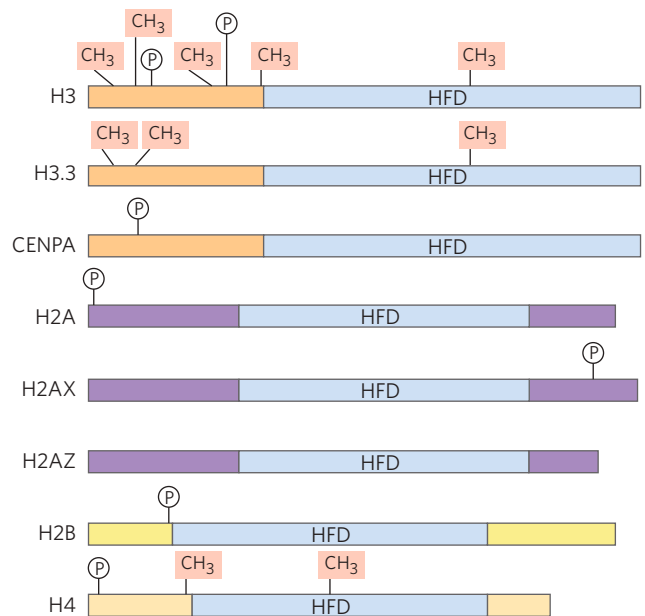


FIGURE 1 Several variants of histones H3, H2A, and H2B are known. Shown here are the core histones and a few of the known variants. Sites of Lys/Arg residue methylation and Ser phosphorylation are indicated. HFD denotes the histone-fold domain, a structural domain common to all core histones.

structure called the **30 nm fiber** (Fig. 24-29). This packing includes one molecule of histone H1 per nucleosome core. Two current models for the organization of histones and DNA in 30 nm fibers are presented in Figure 24-29. Organization into 30 nm fibers does not extend over the entire chromosome but is punctuated by regions bound by sequence-specific (nonhistone) DNA-binding proteins. The 30 nm structure also seems

to depend on the transcriptional activity of the particular region of DNA. Regions in which genes are being transcribed are apparently in a less-ordered state that contains little, if any, histone H1.

The 30 nm fiber—a second level of chromatin organization—provides an approximately 100-fold compaction of the DNA. The higher levels of folding are not yet understood, but certain regions of DNA

those particular nucleosomes bind. Because microarrays are often referred to as chips, this technique is called a ChIP-chip experiment (Fig. 2).

The histone variants, along with the many covalent modifications that histones undergo, help define and isolate the functions of chromatin. They mark the

chromatin, facilitating or suppressing specific functions such as chromosome segregation, transcription, and DNA repair. The histone modifications do not disappear at cell division or during meiosis, and thus they become part of the information transmitted from one generation to the next in all eukaryotic organisms.

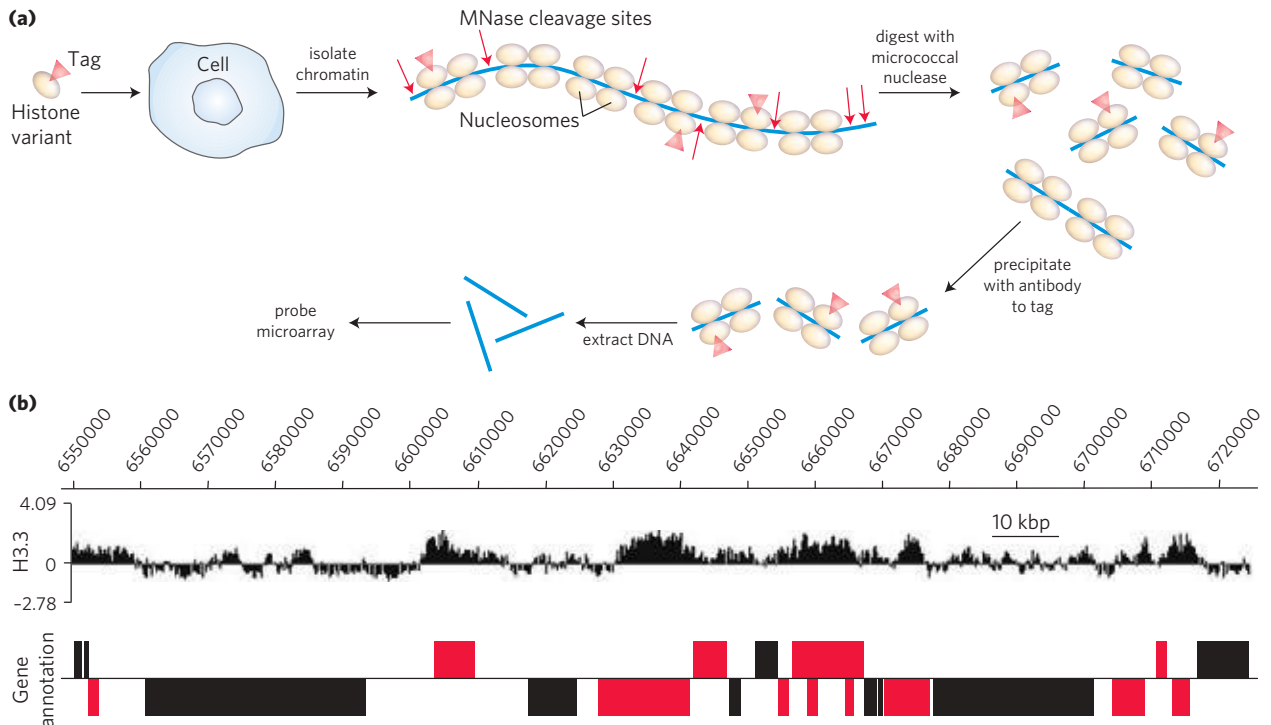


FIGURE 2 A ChIP-chip experiment is designed to reveal the genomic DNA sequences to which a particular histone variant binds. **(a)** A histone variant with an epitope tag (a protein or chemical structure recognized by an antibody; see Chapters 5 and 9) is introduced into a particular cell type, where it is incorporated into nucleosomes. (In some cases, an epitope tag is unnecessary because antibodies may be available that bind directly to the histone modification of interest.) Chromatin is isolated from the cells and digested briefly with micrococcal nuclease (MNase). The DNA bound by nucleosomes is protected from digestion, but the linker DNA is cleaved, releasing segments of DNA bound to one or two nucleosomes. An antibody that binds to the epitope tag is added, and the nucleosomes containing the epitope-tagged histone variant are selectively precipitated. The DNA in these nucleosomes is extracted from the precipitate, labeled, and used to probe a microarray representing all or selected parts of the genomic sequences of that particular cell type. **(b)** In this example, the binding of histone H3.3 is characterized in a short segment of chromosome 2L from

Drosophila melanogaster. Numbers at the top correspond to numbered nucleotide positions in this chromosome arm. Each spot in the microarray represents 100 bp of genomic sequence, so the data here represent more than 1700 separate spots in the microarray. At each spot, the signal from the labeled DNA that was precipitated with antibody to histone H3.3 is presented as a ratio of that signal relative to the control signal generated when total genomic DNA is isolated without immunoprecipitation, sheared, labeled with a different color of label, and used to probe the same microarray. Signals above the horizontal line indicate genomic positions where histone H3.3 binding is enriched relative to the control. Signals below the line are regions where histone H3.3 is relatively absent. Annotated (known) genes in this segment of the genome are shown in the bottom panel (thickened bars). Bars above the line are genes transcribed 5' to 3' left to right, and boxes below the line are genes transcribed right to left. Red bars are genes where RNA polymerase II is also abundant, indicating active transcription. The histone H3.3 binding is concentrated in and near these genes undergoing active transcription.

seem to associate with a chromosomal scaffold (**Fig. 24–30**). The scaffold-associated regions are separated by loops of DNA with perhaps 20 to 100 kbp. The DNA in a loop may contain a set of related genes. The scaffold itself may contain several proteins, notably topoisomerase II and SMC proteins, described below. The presence of topoisomerase II further emphasizes the relationship between DNA underwinding and chromatin structure.

Topoisomerase II is so important to the maintenance of chromatin structure that inhibitors of this enzyme can kill rapidly dividing cells. Several drugs used in cancer chemotherapy are topoisomerase II inhibitors that allow the enzyme to promote strand breakage but not the resealing of the breaks (see Box 24–1).

Evidence exists for additional layers of organization in eukaryotic chromosomes, each dramatically enhancing

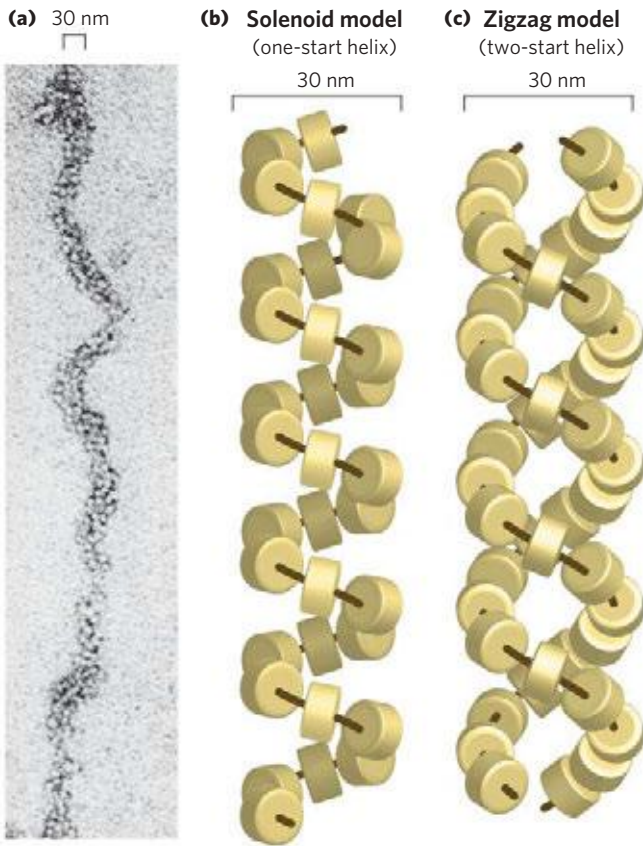


FIGURE 24-29 The 30 nm fiber, a higher-order organization of nucleosomes. The compact fiber is formed by the tight packing of nucleosomes. (a) The 30 nm fiber, as seen by electron microscopy. There are two proposed models for the structure that are consistent with available data: (b) the solenoid model featuring one helical array of nucleosomes and (c) the zigzag model featuring two helical arrays of nucleosomes wrapped about each other. The black line is intended only to trace the proposed general path of the organized structure.

the degree of compaction. One model for achieving this compaction is illustrated in **Figure 24-31**. Higher-order chromatin structure probably varies from chromosome to chromosome, from one region to the next in a single chromosome, and from moment to moment in the life of a cell. No single model can adequately describe these structures. Nevertheless, the principle is clear: DNA compaction in eukaryotic chromosomes is likely to involve coils upon coils upon coils . . . **Three-Dimensional Packaging of Nuclear Chromosomes**

Condensed Chromosome Structures Are Maintained by SMC Proteins

A third major class of chromatin proteins, in addition to the histones and topoisomerases, is the **SMC proteins** (structural maintenance of chromosomes). The primary structure of SMC proteins consists of five distinct domains (**Fig. 24-32a**). The amino- and carboxyl-terminal globular domains, N and C, each of which contains part of an ATP-hydrolytic site, are connected by two regions of α -helical coiled-coil motifs (see Fig. 4-11) that are joined by a hinge domain. The proteins are generally dimeric, forming a V-shaped complex that is thought to be tied together through the protein's hinge domains (**Fig. 24-32b**). One N and one C domain come together to form a complete ATP-hydrolytic site at each free end of the V.

Proteins in the SMC family are found in all types of organisms, from bacteria to humans. Eukaryotes have two major types, cohesins and condensins, both of which are bound by regulatory and accessory proteins (**Fig. 24-32c**). The **cohesins** play a substantial role in linking together sister chromatids immediately after replication and keeping them together as the chromosomes condense to metaphase. This linkage is essential if chromosomes are to segregate properly at cell division. The cohesins, along with a third protein, kleisin, are thought to form a ring around the replicated chromosomes that

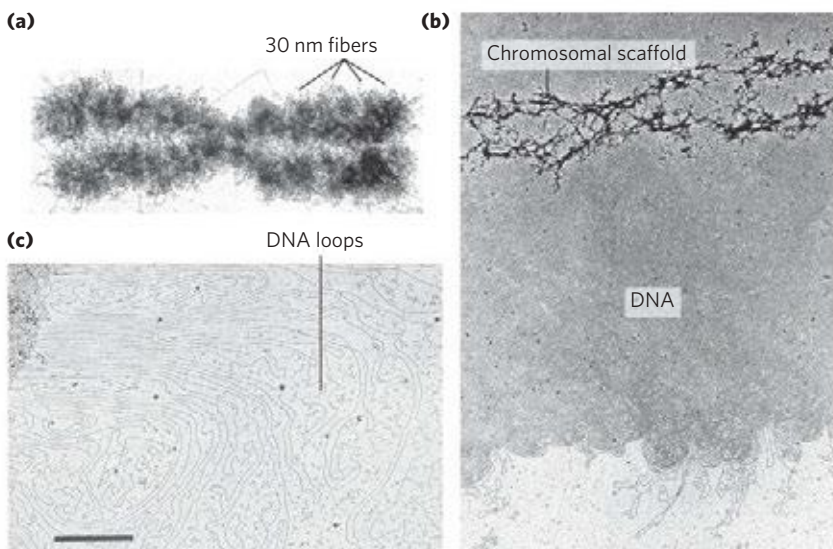


FIGURE 24-30 Loops of DNA attached to a chromosomal scaffold. (a) A swollen chromosome, produced in a buffer of low ionic strength, as seen in the electron microscope. Notice the appearance of 30 nm fibers (chromatin loops) at the margins. (b) Extraction of the histones leaves a proteinaceous chromosomal scaffold surrounded by naked DNA. (c) The DNA appears to be organized in loops attached at their base to the scaffold in the upper left corner. Scale bar = 1 μ m. The three images are at different magnifications.

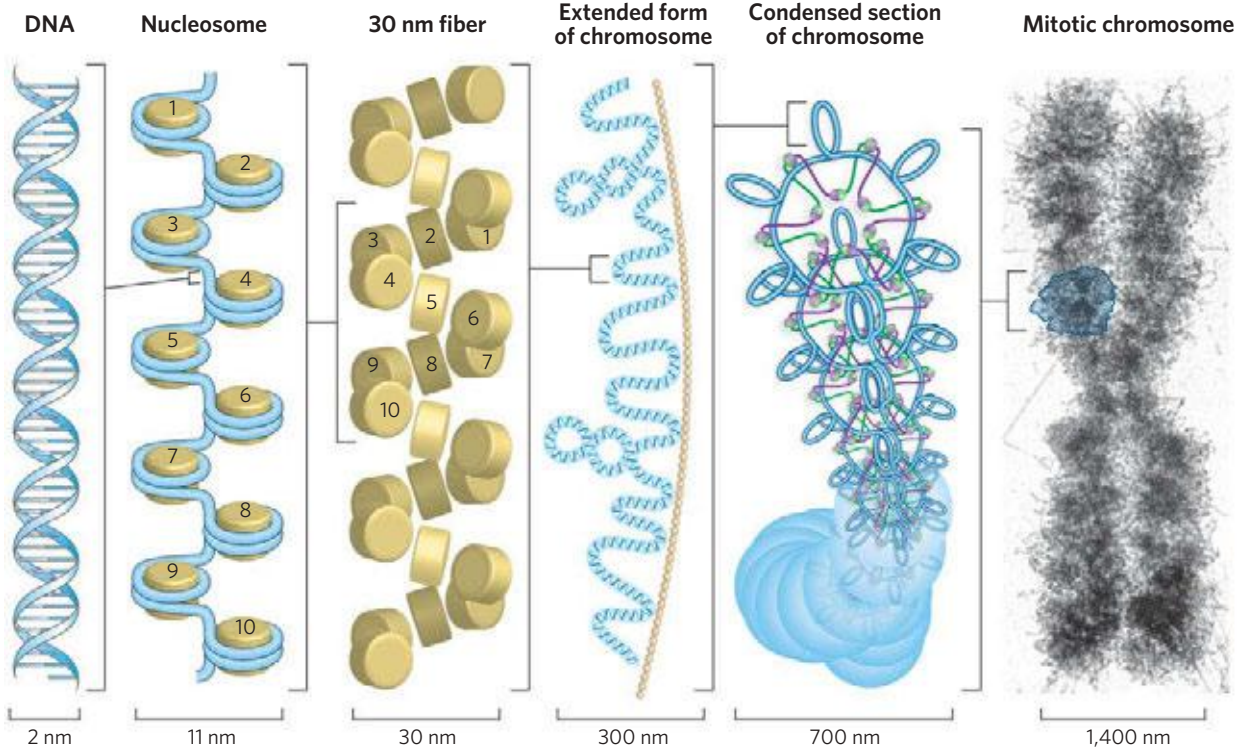


FIGURE 24-31 Compaction of DNA in a eukaryotic chromosome. This model shows the levels of organization that could provide the observed degree of DNA compaction in the chromosomes of eukaryotes. First the DNA is wrapped around histone octamers, then H1 stimulates formation of the 30 nm fiber. Further levels of organization are not well understood

but seem to involve further coiling and loops in the form of rosettes, which also coil into thicker structures. Overall, progressive levels of organization take the form of coils upon coils upon coils. It should be noted that in cells, the higher-order structures (above the 30 nm fiber) are unlikely to be as uniform as depicted here.

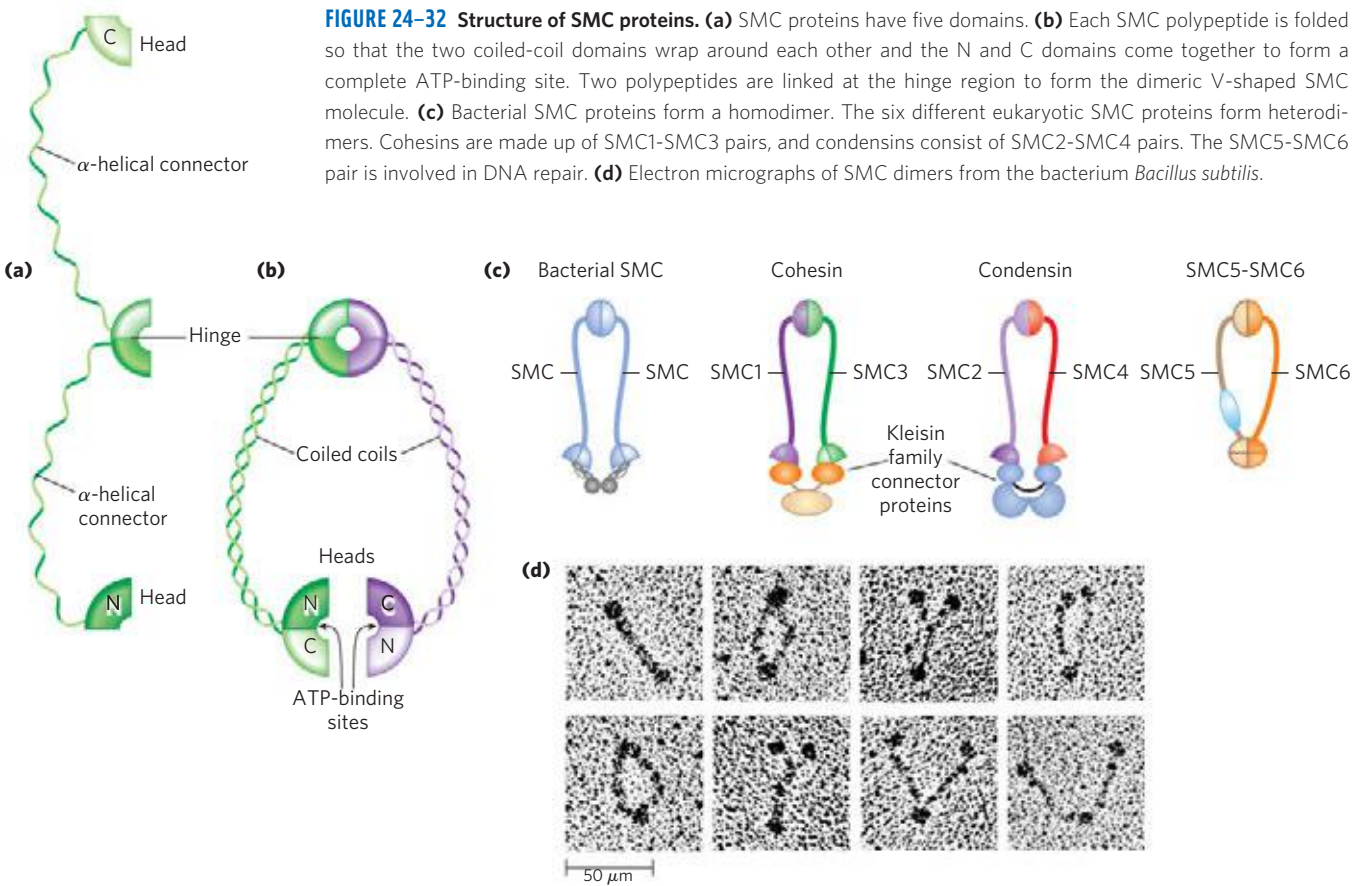


FIGURE 24-32 Structure of SMC proteins. (a) SMC proteins have five domains. (b) Each SMC polypeptide is folded so that the two coiled-coil domains wrap around each other and the N and C domains come together to form a complete ATP-binding site. Two polypeptides are linked at the hinge region to form the dimeric V-shaped SMC molecule. (c) Bacterial SMC proteins form a homodimer. The six different eukaryotic SMC proteins form heterodimers. Cohesins are made up of SMC1-SMC3 pairs, and condensins consist of SMC2-SMC4 pairs. The SMC5-SMC6 pair is involved in DNA repair. (d) Electron micrographs of SMC dimers from the bacterium *Bacillus subtilis*.

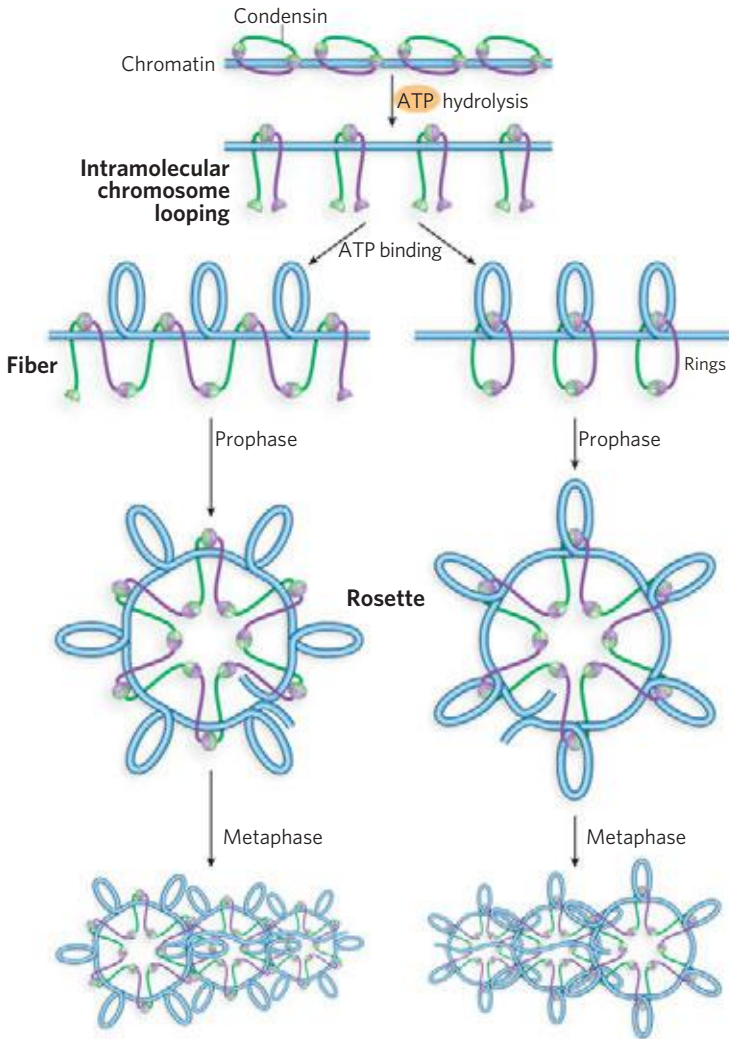


FIGURE 24-33 The possible role of condensins in chromatin condensation. Initially, the DNA is bound at the hinge region of the SMC protein, in the interior of what can become an intramolecular SMC ring. ATP binding leads to head-to-head association, forming supercoiled loops in the bound DNA. Subsequent rearrangement of the head-to-head interactions to form rosettes condenses the DNA. Condensins may organize the looping of the chromosome segments in a number of ways. Two current models are shown.

ties them together until separation is required at cell division. The ring may expand and contract in response to ATP hydrolysis. The **condensins** are essential to the condensation of chromosomes as cells enter mitosis. In the laboratory, condensins bind to DNA in a manner that creates positive supercoils; that is, condensin binding causes the DNA to become overwound, in contrast to the underwinding induced by the binding of nucleosomes. One model for the role of condensins in chromatin compaction is presented in **Figure 24-33**. The cohesins and condensins are essential in orchestrating the many changes in chromosome structure during the eukaryotic cell cycle (**Fig. 24-34**).

Bacterial DNA Is Also Highly Organized

We now turn briefly to the structure of bacterial chromosomes. Bacterial DNA is compacted in a structure called the **nucleoid**, which can occupy a significant fraction of the cell volume (**Fig. 24-35**). The DNA seems to be attached at one or more points to the inner surface of the plasma membrane. Much less is known about the structure of the nucleoid than of eukaryotic chromatin, but a complex organization is slowly being revealed. In *E. coli*, a scaffoldlike structure

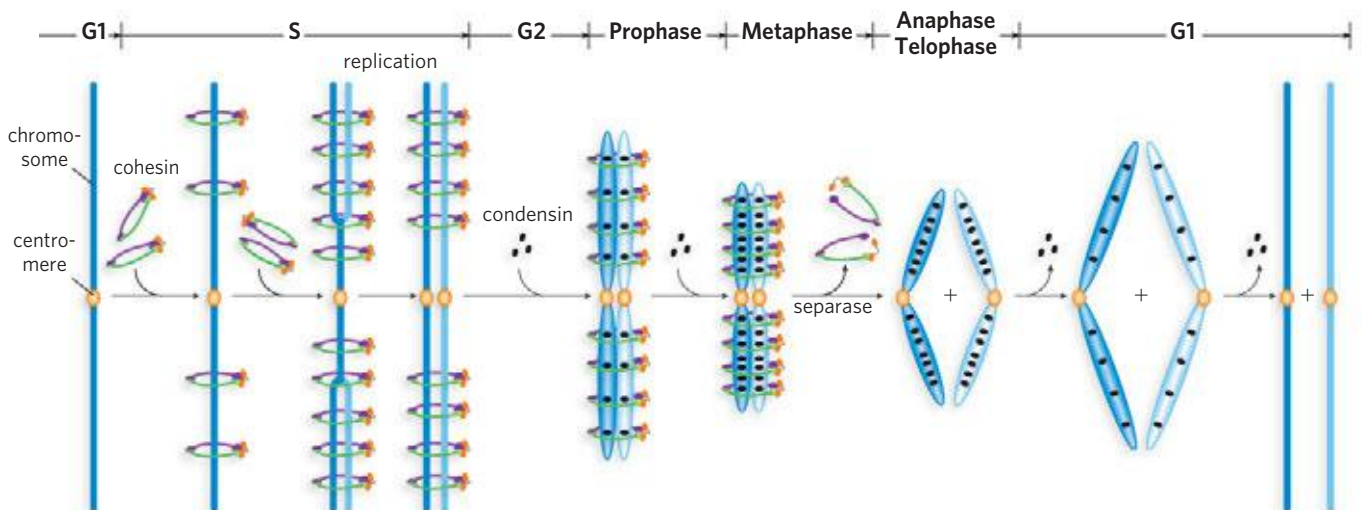


FIGURE 24-34 The roles of cohesins and condensins in the eukaryotic cell cycle. Cohesins are loaded onto the chromosomes during G1 (see Fig. 24-24), tying the sister chromatids together during replication. At the onset of mitosis, condensins bind and maintain the chromatids in a

condensed state. During anaphase, the enzyme separase removes the cohesin links. Once the chromatids separate, condensins begin to unload and the daughter chromosomes return to the uncondensed state.

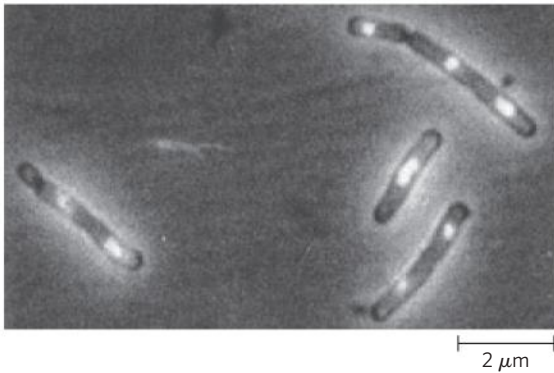


FIGURE 24–35 *E. coli* nucleoids. The DNA of these cells is stained with a dye that fluoresces when exposed to UV light. The light area defines the nucleoid. Note that some cells have replicated their DNA but have not yet undergone cell division and hence have multiple nucleoids.

seems to organize the *circular* chromosome into a series of about 500 looped domains, each encompassing 10,000 bp on average (Fig. 24–36), as described above for chromatin. The domains are topologically constrained; for example, if the DNA is cleaved in one domain, only the DNA within that domain will be relaxed. The domains do not have fixed end points. Instead, the boundaries are most likely in constant motion along the DNA, coordinated with DNA replication. Bacterial DNA does not seem to have any structure comparable to the local organization provided by nucleosomes in eukaryotes. Histone-like proteins are abundant in *E. coli*—the best-characterized example is a two-subunit protein called HU (M_r 19,000)—but these proteins bind and dissociate within minutes, and no regular, stable DNA-histone structure has been found. The dynamic structural changes in the bacterial chromosome may reflect a requirement for more ready access to its genetic information. The bacterial cell division cycle can be as short as 15 min, whereas a typical eukaryotic cell may not divide for hours or even months. In addition, a much greater fraction of bacterial DNA is used to encode RNA and/or protein products. Higher rates of cellular metabolism in bacteria mean that a much higher proportion of the DNA is being transcribed or replicated at a given time than in most eukaryotic cells.

With this overview of the complexity of DNA structure, we are now ready to turn, in the next chapter, to a discussion of DNA metabolism.

SUMMARY 24.3 The Structure of Chromosomes

- ▶ The fundamental unit of organization in the chromatin of eukaryotic cells is the nucleosome, which consists of histones and a 200 bp segment of DNA. A core protein particle containing eight histones (two copies each of histones H2A, H2B, H3, and H4) is encircled by a segment of DNA (about 146 bp) in the form of a left-handed solenoidal supercoil.

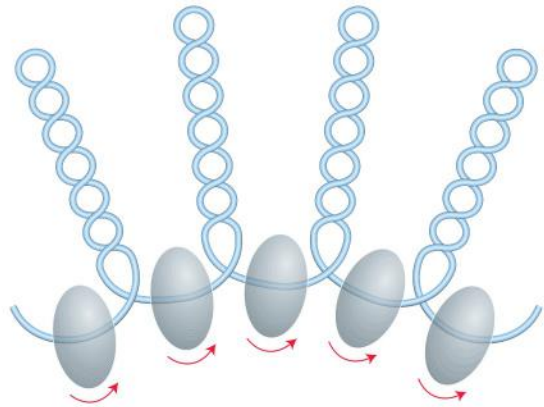


FIGURE 24–36 Looped domains of the *E. coli* chromosome. Each domain is about 10,000 bp in length. The domains are not static, but move along the DNA as replication proceeds. Barriers at the boundaries of the domains, of unknown composition, prevent the relaxation of DNA beyond the boundaries of the domain where a strand break occurs. The putative boundary complexes are shown as gray-shaded ovoids. The arrows denote movement of DNA through the boundary complexes.

- ▶ Nucleosomes are organized into 30 nm fibers, and the fibers are extensively folded to provide the 10,000-fold compaction required to fit a typical eukaryotic chromosome into a cell nucleus. The higher-order folding involves attachment to a chromosomal scaffold that contains histone H1, topoisomerase II, and SMC proteins. The SMC proteins, principally cohesins and condensins, play important roles in keeping the chromosomes organized during each stage of the cell cycle.
- ▶ Bacterial chromosomes are extensively compacted into the nucleoid, but the chromosome seems to be much more dynamic and irregular in structure than eukaryotic chromatin, reflecting the shorter cell cycle and very active metabolism of a bacterial cell.

Key Terms

Terms in bold are defined in the glossary.

chromosome 979	centromere 984
phenotype 979	telomere 984
mutation 979	supercoil 985
gene 980	relaxed DNA 985
regulatory	topology 986
sequence 980	underwinding 987
plasmid 981	linking number 988
intron 984	specific linking
exon 984	difference 988
simple-sequence	superhelical density
DNA 984	(σ) 988
satellite	topoisomers 989
DNA 984	twist 989

writhe 989
topoisomerases 990
plectonemic 992
 solenoidal 993
chromatin 994
histones 994
nucleosome 994

epigenetic 998
euchromatin 998
 30 nm fiber 998
 SMC proteins 1000
 cohesins 1000
 condensins 1002
nucleoid 1002

Further Reading

General

Cox, M.M., Doudna, J., & O'Donnell, M. (2012) *Molecular Biology: Principles and Practice*, W. H. Freeman & Company, New York.

Cozzarelli, N.R. & Wang, J.C. (eds). (1990) *DNA Topology and Its Biological Effects*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kornberg, A. & Baker, T.A. (1991) *DNA Replication*, 2nd edn, W. H. Freeman & Company, New York.

A good place to start for further information on the structure and function of DNA.

Genes and Chromosomes

Campbell, A., Lichten, M., & Schupbach, G. (2010) Telomeric strategies: means to an end. *Annu. Rev. Genet.* **44**, 243–269.

Levin, H.L. & Moran, J.V. (2011) Dynamic interactions between transposons and their hosts. *Nat. Rev. Genet.* **12**, 615–627.

McEachern, M.J., Krauskopf, A., & Blackburn, E.H. (2000) Telomeres and their control. *Annu. Rev. Genet.* **34**, 331–358.

Roy, S.W. & Gilbert, W. (2006) The evolution of spliceosomal introns: patterns, puzzles, and progress. *Nat. Rev. Genet.* **7**, 211–221.

Verdaasdonk, J.S. & Bloom, K. (2011) Centromeres: unique chromatin structures that drive chromosome segregation. *Nat. Rev. Mol. Cell Biol.* **12**, 320–332.

Supercoiling and Topoisomerases

Boles, T.C., White, J.H., & Cozzarelli, N.R. (1990) Structure of plectonemically supercoiled DNA. *J. Mol. Biol.* **213**, 931–951.

A study that defines several fundamental features of supercoiled DNA.

Garcia H.G., Grayson, P., Han, L., Inamdar, M., Kondev, J., Nelson, P.C., Phillips, R., Widom, W., & Wiggins, P.A. (2007) Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers* **85**, 115–130.

A nice description of the physics of bent DNA.

Kohanski, M.A., Dwyer, D.J., & Collins, J.J. (2010) How antibiotics kill bacteria: from targets to networks. *Nat. Rev. Microbiol.* **8**, 423–435.

Lebowitz, J. (1990) Through the looking glass: the discovery of supercoiled DNA. *Trends Biochem. Sci.* **15**, 202–207.

A short and interesting historical note.

Pommier, Y. (2006) Topoisomerase I inhibitors: camptothecins and beyond. *Nat. Rev. Cancer* **6**, 789–802.

Vos, S.M., Tretter, E.M., Schmidt, B.H., & Berger, J.M. (2011) All tangled up: how cells direct, manage, and exploit topoisomerase function. *Nat. Rev. Mol. Cell Biol.* **12**, 827–841.

Chromatin and Nucleosomes

Campos, E.I. & Reinberg, D. (2009) Histones: annotating chromatin. *Annu. Rev. Genet.* **43**, 559–599.

Carter, S.D. & Sjogren, C. (2012) The SMC complexes, DNA and chromosome topology: right or knot? *Crit. Rev. Biochem. Mol. Biol.* **47**, 1–16.

Dillon, S.C. & Dorman, C.J. (2010) Bacterial nucleoid-associated proteins: nucleoid structure and gene expression. *Nat. Rev. Microbiol.* **8**, 185–195.

Kornberg, R.D. (1974) Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871.

A classic paper that introduced the subunit model for chromatin.

Losada, A. & Hirano, T. (2005) Dynamic molecular linkers of the genome: the first decade of SMC proteins. *Genes Dev.* **19**, 1269–1287.

Luijsterburg, M.S., White, M.F., van Driel, R., & Remus, T.D. (2008) The major architects of chromatin: architectural proteins in bacteria, archaea, and eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* **43**, 393–418.

Margueron, R. & Reinberg, D. (2010) Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.* **11**, 285–296.

Rando, O.J. (2007) Chromatin structure in the genomics era. *Trends Genet.* **23**, 67–73.

A description of the imaginative methods being employed to study nucleosome modification patterns, nucleosome positioning, and other aspects of chromosome structure on a genomic scale.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.Z., & Widom, J. (2006) A genomic code for nucleosome positioning. *Nature* **442**, 772–778.

Problems

1. Packaging of DNA in a Virus Bacteriophage T2 has a DNA of molecular weight 120×10^6 contained in a head about 210 nm long. Calculate the length of the DNA (assume the molecular weight of a nucleotide pair is 650) and compare it with the length of the T2 head.

2. The DNA of Phage M13 The base composition of phage M13 DNA is A, 23%; T, 36%; G, 21%; C, 20%. What does this tell you about the DNA of phage M13?

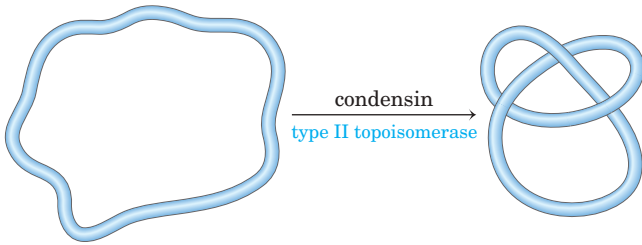
3. The *Mycoplasma* Genome The complete genome of the simplest bacterium known, *Mycoplasma genitalium*, is a circular DNA molecule with 580,070 bp. Calculate the molecular weight and contour length (when relaxed) of this molecule. What is Lk_0 for the *Mycoplasma* chromosome? If $\sigma = -0.06$, what is Lk ?

4. Size of Eukaryotic Genes An enzyme isolated from rat liver has 192 amino acid residues and is coded for by a gene with 1,440 bp. Explain the relationship between the number of amino acid residues in the enzyme and the number of nucleotide pairs in its gene.

5. Linking Number A closed-circular DNA molecule in its relaxed form has an Lk of 500. Approximately how many base pairs are in this DNA? How is the linking number altered (increases, decreases, doesn't change, becomes undefined) when (a) a protein complex binds to form a nucleosome, (b) one DNA strand is broken, (c) DNA gyrase and ATP are added to the DNA solution, or (d) the double helix is denatured by heat?

6. DNA Topology In the presence of a eukaryotic condensin and a type II topoisomerase, the Lk of a relaxed closed-

circular DNA molecule does not change. However, the DNA becomes highly knotted.

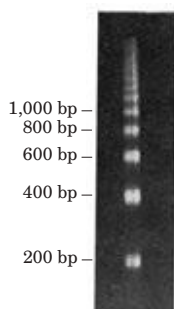


Formation of the knots requires breakage of the DNA, passage of a segment of DNA through the break, and religation by the topoisomerase. Given that every reaction of the topoisomerase would be expected to result in a change in linking number, how can Lk remain the same?

7. Superhelical Density Bacteriophage λ infects *E. coli* by integrating its DNA into the bacterial chromosome. The success of this recombination depends on the topology of the *E. coli* DNA. When the superhelical density (σ) of the *E. coli* DNA is greater than -0.045 , the probability of integration is $<20\%$; when σ is less than -0.06 , the probability is $\sim 70\%$. Plasmid DNA isolated from an *E. coli* culture is found to have a length of 13,800 bp and an Lk of 1,222. Calculate σ for this DNA and predict the likelihood that bacteriophage λ will be able to infect this culture.

8. Altering Linking Number (a) What is the Lk of a 5,000 bp circular duplex DNA molecule with a nick in one strand? (b) What is the Lk of the molecule in (a) when the nick is sealed (relaxed)? (c) How would the Lk of the molecule in (b) be affected by the action of a single molecule of *E. coli* topoisomerase I? (d) What is the Lk of the molecule in (b) after eight enzymatic turnovers by a single molecule of DNA gyrase in the presence of ATP? (e) What is the Lk of the molecule in (d) after four enzymatic turnovers by a single molecule of bacterial type I topoisomerase? (f) What is the Lk of the molecule in (d) after binding of one nucleosome?

9. Chromatin Early evidence that helped researchers define nucleosome structure is illustrated by the agarose gel below, in which the thick bands represent DNA. It was generated by briefly treating chromatin with an enzyme that degrades DNA, then removing all protein and subjecting the purified DNA to electrophoresis. Numbers at the side of the gel denote the position to which a linear DNA of the indicated size would migrate. What does this gel tell you about chromatin structure? Why are the DNA bands thick and spread out rather than sharply defined?



10. DNA Structure Explain how the underwinding of a B-DNA helix might facilitate or stabilize the formation of Z-DNA.

11. Maintaining DNA Structure (a) Describe two structural features required for a DNA molecule to maintain a negatively supercoiled state. (b) List three structural changes that become more favorable when a DNA molecule is negatively supercoiled. (c) What enzyme, with the aid of ATP, can generate negative superhelicity in DNA? (d) Describe the physical mechanism by which this enzyme acts.

12. Yeast Artificial Chromosomes (YACs) YACs are used to clone large pieces of DNA in yeast cells. What three types of DNA sequence are required to ensure proper replication and propagation of a YAC in a yeast cell?

13. Nucleoid Structure in Bacteria In bacteria, the transcription of a subset of genes is affected by DNA topology, with expression increasing or (more often) decreasing when the DNA is relaxed. When a bacterial chromosome is cleaved at a specific site by a restriction enzyme (one that cuts at a long, and thus rare, sequence), only nearby genes (within 10,000 bp) exhibit either an increase or decrease in expression. The transcription of genes elsewhere in the chromosome is unaffected. Explain. (Hint: See Fig. 24–36.)

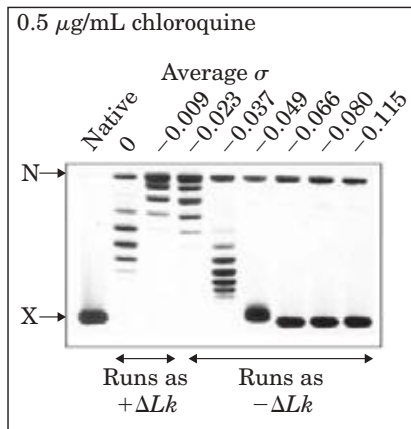
14. DNA Topology When DNA is subjected to electrophoresis in an agarose gel, shorter molecules migrate faster than longer ones. Closed-circular DNAs of the same size but with different linking numbers also can be separated on an agarose gel: topoisomers that are more supercoiled, and thus more condensed, migrate faster through the gel. In the gel shown below, purified plasmid DNA has migrated from top to bottom. There are two bands, with the faster band much more prominent.

(a) What are the DNA species in the two bands? (b) If topoisomerase I is added to a solution of this DNA, what will happen to the upper and lower bands after electrophoresis? (c) If DNA ligase is added to the DNA, will the appearance of the bands change? Explain your answer. (d) If DNA gyrase plus ATP is added to the DNA after the addition of DNA ligase, how will the band pattern change?

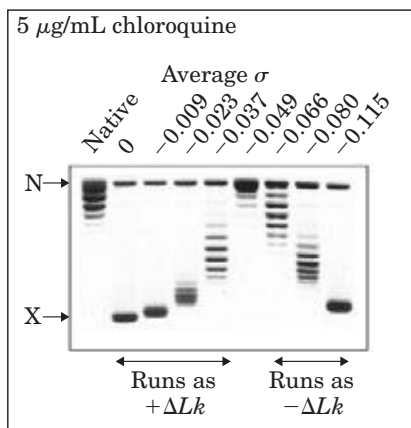


15. DNA Topoisomers When DNA is subjected to electrophoresis in an agarose gel, shorter molecules migrate faster than longer ones. Closed-circular DNAs of the same size but different linking number also can be separated on an agarose gel: topoisomers that are more supercoiled, and thus more condensed, migrate faster through the gel—from top to bottom in the gels shown on the right. A dye, chloroquine, was

added to these gels. Chloroquine intercalates between base pairs and stabilizes a more underwound DNA structure. When the dye binds to a relaxed closed-circular DNA, the DNA is underwound where the dye binds, and unbound regions take on positive supercoils to compensate. In the experiment shown here, topoisomerases were used to make preparations of the same DNA circle with different superhelical densities (σ). Completely relaxed DNA migrated to the position labeled N (for nicked), and highly supercoiled DNA (above the limit where individual topoisomers can be distinguished) to the position labeled X.



Gel A



Gel B

(a) In gel A, why does the $\sigma = 0$ lane (i.e., DNA prepared so that $\sigma = 0$, on average) have multiple bands?

(b) In gel B, is the DNA from the $\sigma = 0$ preparation positively or negatively supercoiled in the presence of the intercalating dye?

(c) In both gels, the $\sigma = -0.115$ lane has two bands, one a highly supercoiled DNA and one relaxed. Propose a reason for the presence of relaxed DNA in these lanes (and others).

(d) The native DNA (leftmost lane in each gel) is the same DNA circle isolated from bacterial cells and untreated. What is the approximate superhelical density of this native DNA?

16. Nucleosomes The human genome contains just over 3.1 billion base pairs. Assuming it is covered with nucleosomes that are spaced as described in this chapter, how many molecules of histone H2A are present in one somatic human cell?

(Do not consider reductions in H2A due to its replacement in some regions by H2A variants.) How would the number change after DNA replication but before cell division?

Data Analysis Problem

17. Defining the Functional Elements of Yeast Chromosomes Figure 24–8 shows the major structural elements of a chromosome of baker's yeast (*Saccharomyces cerevisiae*). Heiter, Mann, Snyder, and Davis (1985) determined the properties of some of these elements. They based their study on the finding that in yeast cells, plasmids (which have genes and an origin of replication) act differently from chromosomes (which have these elements plus centromeres and telomeres) during mitosis. The plasmids are not manipulated by the mitotic apparatus and segregate randomly between daughter cells. Without a selectable marker to force the host cells to retain them (see Fig. 9–4), these plasmids are rapidly lost. In contrast, chromosomes, even without a selectable marker, are manipulated by the mitotic apparatus and are lost at a very low rate (about 10^{-5} per cell division).

Heiter and colleagues set out to determine the important components of yeast chromosomes by constructing plasmids with various parts of chromosomes and observing whether these “synthetic chromosomes” segregated properly during mitosis. To measure the rates of different types of failed chromosome segregation, the researchers needed a rapid assay to determine the number of copies of synthetic chromosomes present in different cells. This assay took advantage of the fact that wild-type yeast colonies are white whereas certain adenine-requiring (ade^-) mutants yield red colonies on nutrient media. Specifically, $ade2^-$ cells lack functional AIR carboxylase (the enzyme of step 6a in Figure 22–35) and accumulate AIR (5-aminoimidazole ribonucleotide) in their cytoplasm. This excess AIR is converted to a conspicuous red pigment. The other part of the assay involved the gene *SUP11*, which encodes an ochre suppressor (a type of nonsense suppressor; see Box 27–4) that suppresses the phenotype of some $ade2^-$ mutants.

Heiter and coworkers started with a diploid strain of yeast homozygous for $ade2^-$; these cells are red. When the mutant cells contain one copy of *SUP11*, the metabolic defect is partly suppressed and the cells are pink. When the cells contain two or more copies of *SUP11*, the defect is completely suppressed and the cells are white.

The researchers inserted one copy of *SUP11* into synthetic chromosomes containing various elements thought to be important in chromosome function, and then observed how well these chromosomes were passed from one generation to the next. These pink cells were plated on nonselective media, and the behavior of the synthetic chromosomes was observed. Specifically, Heiter and coworkers looked for colonies in which the synthetic chromosomes segregated improperly at the first division after plating, giving rise to a colony that is half one genotype and half the other. Because yeast cells are nonmotile, this will be a sectored colony, with one half one color and the other half another color.

(a) One way for the mitotic process to fail is *nondisjunction*: the chromosome replicates but the sister chromatids fail to separate, so both copies of the chromosome end up in the same daughter cell. Explain how nondisjunction of the synthetic chromosome would give rise to a colony that is half red and half white.

(b) Another way for the mitotic process to fail is *chromosome loss*: the chromosome does not enter the daughter nucleus or is not replicated. Explain how loss of the synthetic chromosome would give rise to a colony that is half red and half pink.

By counting the frequency of the different colony types, Heiter and colleagues could estimate the frequency of these aberrant mitotic events with different types of synthetic chromosome. First, they explored the requirement for centromeric sequences by constructing synthetic chromosomes with different-sized DNA fragments containing a known centromere. Their results are shown below.

Synthetic chromosome	Size of centromere-containing fragment (kbp)	Chromosome loss (%)	Nondisjunction (%)
1	none	—	>50
2	0.63	1.6	1.1
3	1.6	1.9	0.4
4	3.0	1.7	0.35
5	6.0	1.6	0.35

(c) Based on these data, what can you conclude about the size of the centromere required for normal mitotic segregation? Explain your reasoning.

(d) Interestingly, all the synthetic chromosomes created in these experiments were circular and lacked telomeres. Explain how they could be replicated more-or-less properly.

Heiter and colleagues next constructed a series of linear synthetic chromosomes that included the functional centromeric sequence and telomeres, and measured the total mitotic error frequency (% loss + % nondisjunction) as a function of size:

Synthetic chromosome	Size (kbp)	Total error frequency (%)
6	15	11.0
7	55	1.5
8	95	0.44
9	137	0.14

(e) Based on these data, what can you conclude about the chromosome size required for normal mitotic segregation? Explain your reasoning.

(f) Normal yeast chromosomes are linear, range from 250 kbp to 2,000 kbp in length, and have a mitotic error rate of about 10^{-5} per cell division. Extrapolating the results from (e), do the centromeric and telomeric sequences used in these experiments explain the mitotic stability of normal yeast chromosomes, or must other elements be involved? Explain your reasoning. (Hint: A plot of log (error rate) vs. length will be helpful.)

Reference

Heiter, P., Mann, C., Snyder, M., & Davis, R.W. (1985) Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. *Cell* **40**, 381–392.

this page left intentionally blank

DNA Metabolism

25.1 DNA Replication 1011

25.2 DNA Repair 1027

25.3 DNA Recombination 1038

As the repository of genetic information, DNA occupies a unique and central place among biological macromolecules. The nucleotide sequences of DNA encode the primary structures of all cellular RNAs and proteins and, through enzymes, indirectly affect the synthesis of all other cellular constituents. This passage of information from DNA to RNA and protein guides the size, shape, and functioning of every living thing.

DNA is a marvelous device for the stable storage of genetic information. The phrase “stable storage,” however, conveys a static and misleading picture. It fails to capture the complexity of processes by which genetic information is preserved in an uncorrupted state and then transmitted from one generation of cells to the next. DNA metabolism comprises both the process that gives rise to faithful copies of DNA molecules (replication) and the processes that affect the inherent structure of the information (repair and recombination). Together, these activities are the focus of this chapter.

The metabolism of DNA is shaped by the requirement for an exquisite degree of accuracy. The chemistry of joining one nucleotide to the next in DNA replication is elegant and simple, almost deceptively so. However, as is the case with all information-containing polymers, forming a covalent link between two monomeric units is just a small part of the biochemical process. As we shall see, complexity arises in the form of enzymatic devices to ensure that the *correct* nucleotide is added and that genetic information is transmitted intact. Uncorrected errors that arise during DNA synthesis can have dire consequences, not only because they can permanently affect or eliminate the function of a gene but also because the change is inheritable.

The enzymes that synthesize DNA may copy DNA molecules that contain millions of bases. They do so with extraordinary fidelity and speed, even though the

DNA substrate is highly compacted and bound with other proteins. Formation of phosphodiester bonds to link nucleotides in the backbone of a growing DNA strand is therefore only one part of an elaborate process that requires myriad proteins and enzymes.

Maintaining the integrity of genetic information lies at the heart of DNA repair. As detailed in Chapter 8, DNA is susceptible to many types of damaging reactions. Such reactions are infrequent but significant nevertheless, because of the very low biological tolerance for changes in DNA sequence. DNA is the only macromolecule for which repair systems exist; the number, diversity, and complexity of DNA repair mechanisms reflect the wide range of insults that can harm DNA.

Cells can rearrange their genetic information by processes collectively called recombination—seemingly undermining the principle that the stability and integrity of genetic information are paramount. However, most DNA rearrangements in fact play constructive roles in maintaining genomic integrity, contributing in special ways to DNA replication, DNA repair, and chromosome segregation.

Special emphasis is given in this chapter to the *enzymes* of DNA metabolism. They merit careful study not only because of their intrinsic biological importance and interest but also for their increasing importance in medicine and for their everyday use as reagents in a wide range of modern biochemical technologies. Many of the seminal discoveries in DNA metabolism have been made with *Escherichia coli*, so its well-understood enzymes are generally used to illustrate the ground rules. A quick look at some relevant genes on the *E. coli* genetic map (**Fig. 25–1**) provides just a hint of the complexity of the enzymatic systems involved in DNA metabolism.

Before taking a closer look at replication, we must make a short digression into the use of abbreviations in naming bacterial genes and proteins—you will encounter many of these in this and later chapters. Similar conventions exist for naming eukaryotic genes, although the exact form of the abbreviations may vary with the species and no single convention applies to all eukaryotic systems.

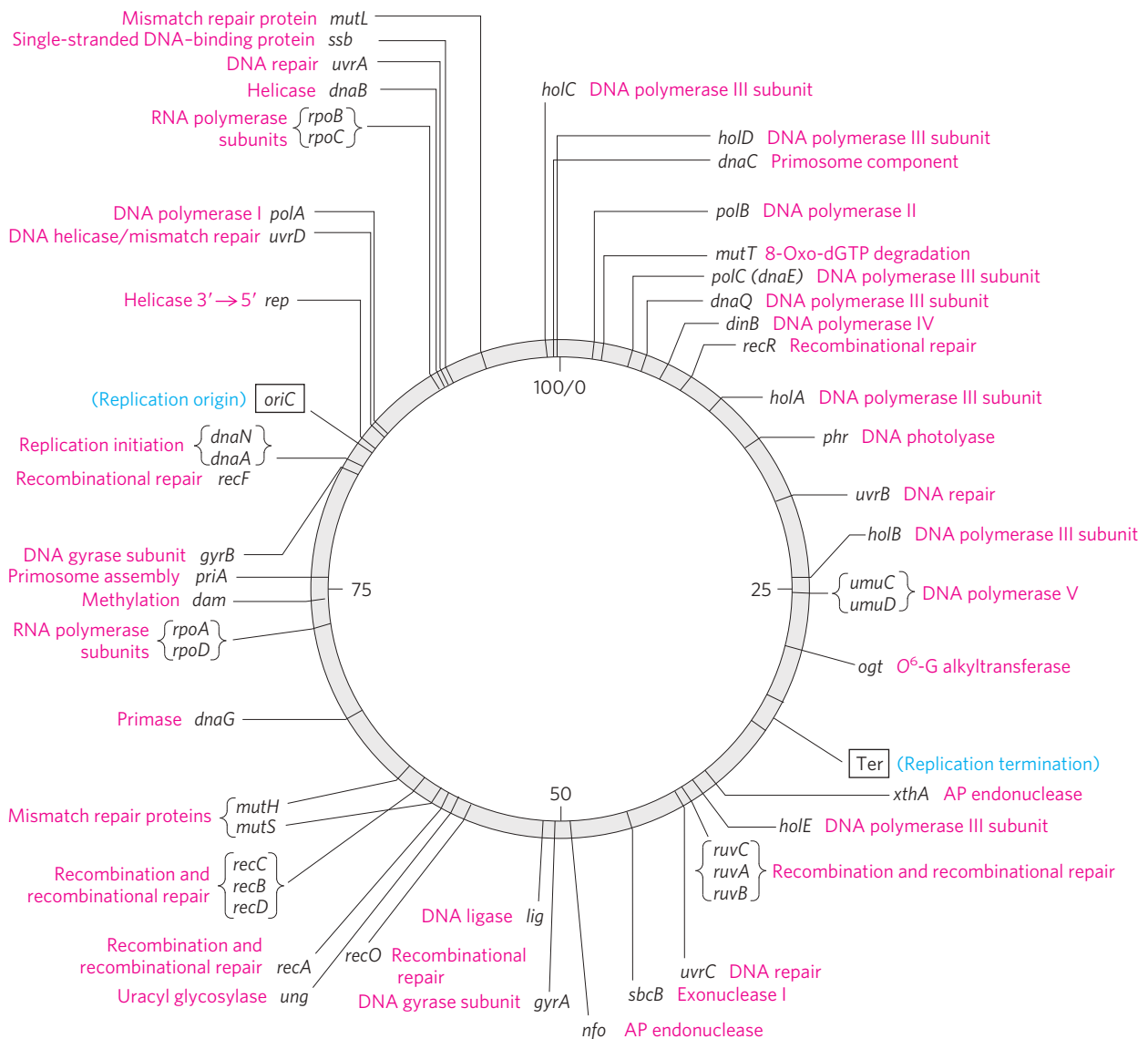


FIGURE 25-1 Map of the *E. coli* chromosome. The map shows the relative positions of genes encoding many of the proteins important in DNA metabolism. The number of genes known to be involved provides a hint of the complexity of these processes. The numbers 0 to 100 inside the circular chromosome denote a genetic measurement called minutes. Each minute corresponds to ~40,000 bp along the DNA molecule. The

three-letter names of genes and other elements generally reflect some aspect of their function. These include *mut*, *mutagenesis*; *dna*, *DNA replication*; *pol*, *DNA polymerase*; *rpo*, *RNA polymerase*; *uvr*, *UV resistance*; *rec*, *recombination*; *dam*, *DNA adenine methylation*; *lig*, *DNA ligase*; *Ter*, *termination of replication*; and *ori*, *origin of replication* (*oriC* in *E. coli*, as shown here).

KEY CONVENTION: Bacterial genes generally are named using three lowercase, italicized letters that often reflect a gene's apparent function. For example, the *dna*, *uvr*, and *rec* genes affect *DNA replication*, *resistance to the damaging effects of UV radiation*, and *recombination*, respectively. Where several genes affect the same process, the letters *A*, *B*, *C*, and so forth, are added—as in *dnaA*, *dnaB*, *dnaQ*, for example—usually reflecting their order of discovery rather than their order in a reaction sequence. ■

The use of abbreviations in naming proteins is less straightforward. During genetic investigations, the protein product of each gene is usually isolated and charac-

terized. Many bacterial genes have been identified and named before the roles of their protein products are understood in detail. Sometimes the gene product is found to be a previously isolated protein, and some renaming occurs. Often, however, the product turns out to be an as yet unknown protein, with an activity not easily described by a simple enzyme name.

KEY CONVENTION: Bacterial proteins often retain the name of their genes. When referring to the protein product of an *E. coli* gene, roman type is used and the first letter is capitalized: for example, the *dnaA* and *recA* gene products are the DnaA and RecA proteins, respectively. ■

25.1 DNA Replication

Long before the structure of DNA was known, scientists wondered at the ability of organisms to create faithful copies of themselves and, later, at the ability of cells to produce many identical copies of large, complex macromolecules. Speculation about these problems centered around the concept of a **template**, a structure that would allow molecules to be lined up in a specific order and joined to create a macromolecule with a unique sequence and function. The 1940s brought the revelation that DNA was the genetic molecule, but not until James Watson and Francis Crick deduced its structure did the way in which DNA could act as a template for the replication and transmission of genetic information become clear: *one strand is the complement of the other*. The strict base-pairing rules mean that each strand provides the template for a new strand with a predictable and complementary sequence (see Figs 8–14, 8–15).

Nucleotides: Building Blocks of Nucleic Acids

The fundamental properties of the DNA replication process and the mechanisms used by the enzymes that catalyze it have proved to be essentially identical in all species. This mechanistic unity is a major theme as we proceed from general properties of the replication process, to *E. coli* replication enzymes, and, finally, to replication in eukaryotes.

DNA Replication Follows a Set of Fundamental Rules

Early research on bacterial DNA replication and its enzymes helped to establish several basic properties that have proven applicable to DNA synthesis in every organism.

DNA Replication Is Semiconservative Each DNA strand serves as a template for the synthesis of a new strand, producing two new DNA molecules, each with one new strand and one old strand. This is **semiconservative replication**.

Watson and Crick proposed the hypothesis of semi-conservative replication soon after publication of their 1953 paper on the structure of DNA, and the hypothesis was proved by ingeniously designed experiments carried out by Matthew Meselson and Franklin Stahl in 1957. Meselson and Stahl grew *E. coli* cells for many generations in a medium in which the sole nitrogen source (NH_4Cl) contained ^{15}N , the “heavy” isotope of nitrogen, instead of the normal, more abundant “light” isotope, ^{14}N . The DNA isolated from these cells had a density about 1% greater than that of normal [^{14}N]DNA (**Fig. 25–2a**). Although this is only a small difference, a mixture of heavy [^{15}N]DNA and light [^{14}N]DNA can be separated by centrifugation to equilibrium in a cesium chloride density gradient.

The *E. coli* cells grown in the ^{15}N medium were transferred to a fresh medium containing only the ^{14}N

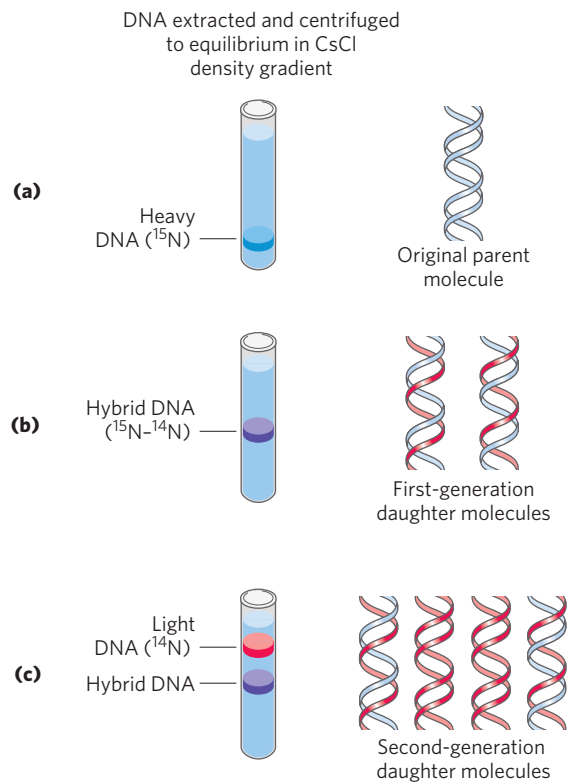


FIGURE 25–2 The Meselson-Stahl experiment. **(a)** Cells were grown for many generations in a medium containing only heavy nitrogen, ^{15}N , so that all the nitrogen in their DNA was ^{15}N , as shown by a single band (blue) when centrifuged in a CsCl density gradient. **(b)** Once the cells had been transferred to a medium containing only light nitrogen, ^{14}N , cellular DNA isolated after one generation equilibrated at a higher position in the density gradient (purple band). **(c)** A second cycle of replication yielded a hybrid DNA band (purple) and another band (red), containing only [^{14}N]DNA, confirming semiconservative replication.

isotope, where they were allowed to grow until the cell population had just doubled. The DNA isolated from these first-generation cells formed a *single* band in the CsCl gradient at a position indicating that the double-helical DNA molecules of the daughter cells were hybrids containing one new ^{14}N strand and one parent ^{15}N strand (Fig. 25–2b).

This result argued against conservative replication, an alternative hypothesis in which one progeny DNA molecule would consist of two newly synthesized DNA strands and the other would contain the two parent strands; this would not yield hybrid DNA molecules in the Meselson-Stahl experiment. The semiconservative replication hypothesis was further supported in the next step of the experiment (Fig. 25–2c). Cells were again allowed to double in number in the ^{14}N medium. The isolated DNA product of this second cycle of replication exhibited *two* bands in the density gradient, one with a density equal to that of light DNA and the other with the density of the hybrid DNA observed after the first cell doubling.

Replication Begins at an Origin and Usually Proceeds Bidirectionally Following the confirmation of a semiconservative mechanism of replication, a host of questions arose. Are the parent DNA strands completely unwound before each is replicated? Does replication begin at random places or at a unique point? After initiation at any point in the DNA, does replication proceed in one direction or both?

An early indication that replication is a highly coordinated process in which the parent strands are simultaneously unwound and replicated was provided by John Cairns, using autoradiography. He made *E. coli* DNA radioactive by growing cells in a medium containing thymidine labeled with tritium (^3H). When the DNA was carefully isolated, spread, and overlaid with a photographic emulsion for several weeks, the radioactive thymidine residues generated “tracks” of silver grains in the emulsion, producing an image of the DNA molecule. These tracks revealed that the intact chromosome of *E. coli* is a single huge circle, 1.7 mm long. Radioactive DNA isolated from cells during replication showed an extra loop (Fig. 25-3). Cairns concluded that the loop resulted from the formation of two radioactive daughter strands, each complementary to a parent strand. One or both ends of the loop are dynamic points, termed **replication forks**, where parent DNA is being unwound and the separated strands quickly replicated. Cairns’s results demonstrated that both DNA strands are replicated simultaneously, and variations on his experiment indicated that replication of bacterial chromosomes is bidirectional: both ends of the loop have active replication forks.

The determination of whether the replication loops originate at a unique point in the DNA required landmarks along the DNA molecule. These were provided by a technique called **denaturation mapping**, developed by Ross Inman and colleagues. Using the 48,502 bp chromosome of bacteriophage λ , Inman showed that DNA could be selectively denatured at sequences unusually rich in A=T base pairs, generating a reproducible pattern of single-strand bubbles (see Fig. 8-28). Isolated DNA containing replication loops can be partially denatured in the same way. This allows the position and progress of the replication forks to be measured and mapped, using the denatured regions as points of reference. The technique revealed that in this system the replication loops always initiate at a unique point, which was termed an **origin**. It also confirmed the earlier observation that replication is usually bidirectional. For circular DNA molecules, the two replication forks meet at a point on the side of the circle opposite to the origin. Specific origins of replication have since been identified and characterized in bacteria and lower eukaryotes.

DNA Synthesis Proceeds in a 5'→3' Direction and Is Semidiscontinuous A new strand of DNA is always synthesized in the 5'→3' direction, with the free 3' OH as the point at which the DNA is elongated (the 5' and 3' ends of a

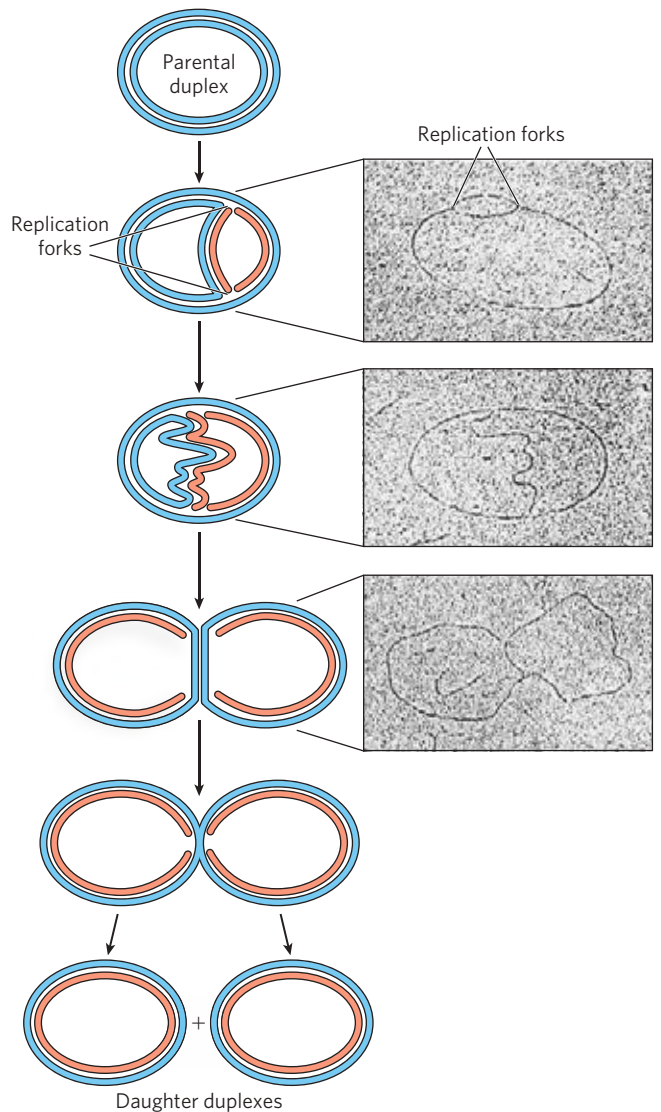


FIGURE 25-3 Visualization of DNA replication. Stages in the replication of circular DNA molecules have been visualized by electron microscopy. Replication of a circular chromosome produces a structure resembling the Greek letter theta, θ , as both strands are replicated simultaneously (new strands shown in light red). The electron micrographs show images of plasmid DNA being replicated from a single replication origin.

DNA strand are defined in Fig. 8-7). Because the two DNA strands are antiparallel, the strand serving as the template is read from its 3' end toward its 5' end.

If synthesis always proceeds in the 5'→3' direction, how can both strands be synthesized simultaneously? If both strands were synthesized *continuously* while the replication fork moved, one strand would have to undergo 3'→5' synthesis. This problem was resolved by Reiji Okazaki and colleagues in the 1960s. Okazaki found that one of the new DNA strands is synthesized in short pieces, now called **Okazaki fragments**. This work ultimately led to the conclusion that one strand is synthesized continuously and the other discontinuously (Fig. 25-4). The continuous strand, or **leading strand**, is the one in which 5'→3' synthesis proceeds in the *same* direction as

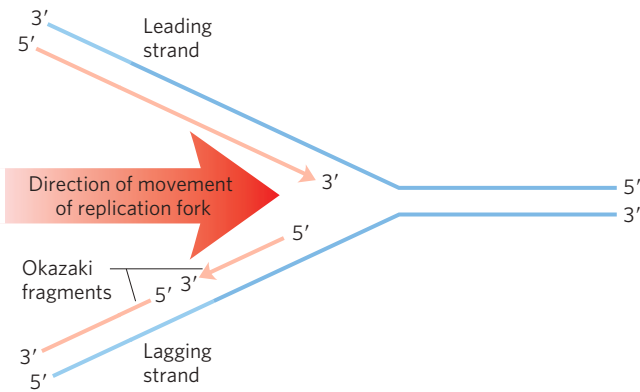


FIGURE 25-4 Defining DNA strands at the replication fork. A new DNA strand (light red) is always synthesized in the 5'→3' direction. The template is read in the opposite direction, 3'→5'. The leading strand is continuously synthesized in the direction taken by the replication fork. The other strand, the lagging strand, is synthesized discontinuously in short pieces (Okazaki fragments) in a direction opposite to that in which the replication fork moves. The Okazaki fragments are spliced together by DNA ligase. In bacteria, Okazaki fragments are ~1,000 to 2,000 nucleotides long. In eukaryotic cells, they are 150 to 200 nucleotides long.

replication fork movement. The discontinuous strand, or **lagging strand**, is the one in which 5'→3' synthesis proceeds in the direction *opposite* to the direction of fork movement. Okazaki fragments range in length from a few hundred to a few thousand nucleotides, depending on the cell type. As we shall see later, leading and lagging strand syntheses are tightly coordinated.

DNA Is Degraded by Nucleases

To explain the enzymology of DNA replication, we first introduce the enzymes that degrade DNA rather than synthesize it. These enzymes are known as **nucleases**, or **DNases** if they are specific for DNA rather than RNA. Every cell contains several different nucleases, belonging to two broad classes: exonucleases and endonucleases. **Exonucleases** degrade nucleic acids from one end of the molecule. Many operate in only the 5'→3' or the 3'→5' direction, removing nucleotides only from the 5' or the 3' end, respectively, of one strand of a double-stranded nucleic acid or of a single-stranded DNA. **Endonucleases** can begin to degrade at specific internal sites in a nucleic acid strand or molecule, reducing it to smaller and smaller fragments. A few exonucleases and endonucleases degrade only single-stranded DNA. There are a few important classes of endonucleases that cleave only at specific nucleotide sequences (such as the restriction endonucleases that are so important in biotechnology; see Chapter 9, Fig. 9-2). You will encounter many types of nucleases in this and subsequent chapters.

DNA Is Synthesized by DNA Polymerases

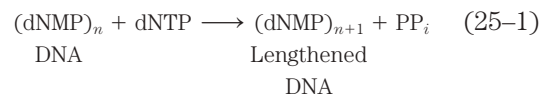
The search for an enzyme that could synthesize DNA began in 1955. Work by Arthur Kornberg and colleagues



Arthur Kornberg,
1918-2007

led to the purification and characterization of a DNA polymerase from *E. coli* cells, a single-polypeptide enzyme now called **DNA polymerase I** (M_r 103,000; encoded by the *polA* gene). Much later, investigators found that *E. coli* contains at least four other distinct DNA polymerases, described below.

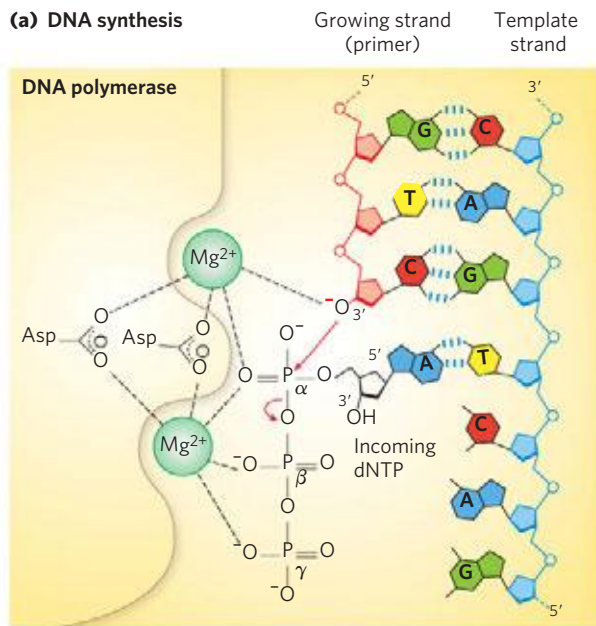
Detailed studies of DNA polymerase I revealed features of the DNA synthetic process that are now known to be common to all DNA polymerases. The fundamental reaction is a phosphoryl group transfer. The nucleophile is the 3'-hydroxyl group of the nucleotide at the 3' end of the growing strand. Nucleophilic attack occurs at the α phosphorus of the incoming deoxynucleoside 5'-triphosphate (**Fig. 25-5a**). Inorganic pyrophosphate is released in the reaction. The general reaction is



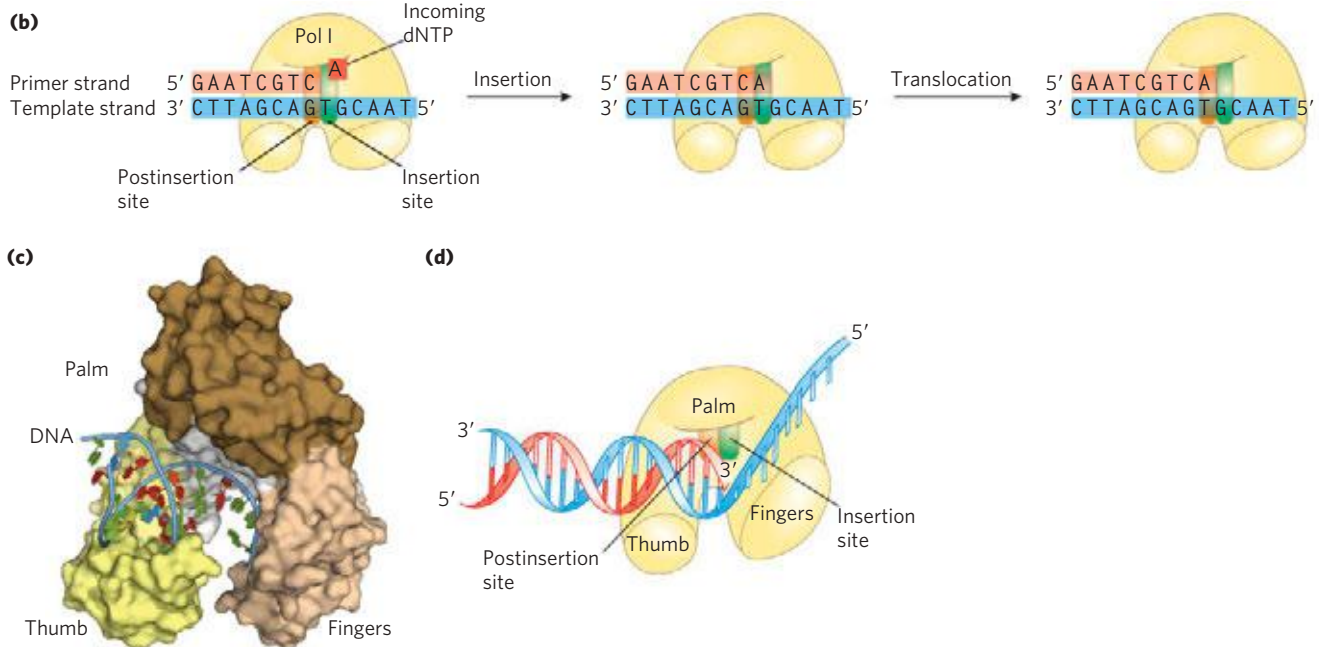
where dNMP and dNTP are deoxynucleoside 5'-monophosphate and 5'-triphosphate, respectively. Catalysis by virtually all DNA polymerases prominently involves two Mg^{2+} ions at the active site (**Fig. 25-5a**). One of these helps to deprotonate the 3'-hydroxyl group, rendering it a more effective nucleophile. The other binds to the incoming dNTP and facilitates departure of the pyrophosphate.

The reaction seems to proceed with only a minimal change in free energy, given that one phosphodiester bond is formed at the expense of a somewhat less stable phosphate anhydride. However, noncovalent base-stacking and base-pairing interactions provide additional stabilization to the lengthened DNA product relative to the free nucleotide. Also, the formation of products is facilitated in the cell by the 19 kJ/mol generated in the subsequent hydrolysis of the pyrophosphate product by the enzyme pyrophosphatase (p. 524).

Early work on DNA polymerase I led to the definition of two central requirements for DNA polymerization (**Fig. 25-5**). First, all DNA polymerases require a template. The polymerization reaction is guided by a template DNA strand according to the base-pairing rules predicted by Watson and Crick: where a guanine is present in the template, a cytosine deoxynucleotide is added to the new strand, and so on. This was a particularly important discovery, not only because it provided a chemical basis for accurate semiconservative DNA replication but also because it represented the first example of the use of a template to guide a biosynthetic reaction.

(a) DNA synthesis

MECHANISM FIGURE 25-5 Elongation of a DNA chain. **(a)** The catalytic mechanism for addition of a new nucleotide by DNA polymerase involves two Mg^{2+} ions, coordinated to the phosphate groups of the incoming nucleotide triphosphate, the 3'-hydroxyl group that will act as a nucleophile, and three Asp residues, two of which are highly conserved in all DNA polymerases. The Mg^{2+} ion depicted at the top facilitates attack of the 3'-hydroxyl group of the primer on the α phosphate of the nucleotide triphosphate; the other Mg^{2+} ion facilitates displacement of the pyrophosphate. Both ions stabilize the structure of the pentacovalent transition state. RNA polymerases use a similar mechanism (see Fig. 26-1a). **(b)** DNA polymerase I activity also requires a single unpaired strand to act as template and a primer strand to provide the free hydroxyl group at the 3' end, to which the new nucleotide unit is added. Each incoming nucleotide is selected in part by base-pairing to the appropriate nucleotide in the template strand. The reaction product has a new free 3' hydroxyl, allowing the addition of another nucleotide. The newly formed base pair migrates to make the active site available to the next pair to be formed. **(c)** The core of most DNA polymerases is shaped like a human hand that wraps around the active site. The structure shown is the DNA polymerase I of *Thermus aquaticus*, bound to DNA (PDB ID 4KTQ). **(d)** A cartoon interpretation of the polymerase structure shows the insertion and postinsertion parts of the active site. The insertion site is where the nucleotide addition occurs, and the postinsertion site is where the newly formed base pair migrates after it appears. **Nucleotide Polymerization by DNA Polymerase**



Second, the polymerases require a **primer**. A primer is a strand segment (complementary to the template) with a free 3'-hydroxyl group to which a nucleotide can be added; the free 3' end of the primer is called the **primer terminus**. In other words, part of the new strand must already be in place: all DNA polymerases can only add nucleotides to a preexisting strand. Many primers are oligonucleotides of RNA rather than DNA, and specialized enzymes synthesize primers when and where they are required.

A DNA polymerase active site has two parts (Fig. 25-5b). The incoming nucleotide is initially positioned in the **insertion site**. Once the phosphodiester bond is formed, the polymerase slides forward on the DNA and

the new base pair is positioned in the **postinsertion site**. These elements are located in a pocket that resembles the palm of a hand (Fig. 25-5c).

After adding a nucleotide to a growing DNA strand, a DNA polymerase either dissociates or moves along the template and adds another nucleotide. Dissociation and reassociation of the polymerase can limit the overall polymerization rate—the process is generally faster when a polymerase adds more nucleotides without dissociating from the template. The average number of nucleotides added before a polymerase dissociates defines its **processivity**. DNA polymerases vary greatly in processivity; some add just a few nucleotides before dissociating, others add many thousands. **Nucleotide Polymerization by DNA Polymerase**

Replication Is Very Accurate

Replication proceeds with an extraordinary degree of fidelity. In *E. coli*, a mistake is made only once for every 10^9 to 10^{10} nucleotides added. For the *E. coli* chromosome of $\sim 4.6 \times 10^6$ bp, this means that an error occurs only once per 1,000 to 10,000 replications. During polymerization, discrimination between correct and incorrect nucleotides relies not just on the hydrogen bonds that specify the correct pairing between complementary bases but also on the common geometry of the standard A=T and G≡C base pairs (Fig. 25-6). The active site of DNA polymerase I accommodates only base pairs with this geometry. An incorrect nucleotide may be able to hydrogen-bond with a base in the template, but it generally will not fit into the active site. Incorrect bases can be rejected before the phosphodiester bond is formed.

The accuracy of the polymerization reaction itself, however, is insufficient to account for the high degree of fidelity in replication. Careful measurements *in vitro* have shown that DNA polymerases insert one incorrect nucleotide for every 10^4 to 10^5 correct ones. These mistakes sometimes occur because a base is briefly in an unusual tautomeric form (see Fig. 8-9), allowing it to hydrogen-bond with an incorrect partner. *In vivo*, the error rate is reduced by additional enzymatic mechanisms.

One mechanism intrinsic to virtually all DNA polymerases is a separate 3'→5' exonuclease activity that

double-checks each nucleotide after it is added. This nuclease activity permits the enzyme to remove a newly added nucleotide and is highly specific for mismatched base pairs (Fig. 25-7). If the polymerase has added the wrong nucleotide, translocation of the enzyme to the position where the next nucleotide is to be added is inhibited. This kinetic pause provides the opportunity for a correction. The 3'→5' exonuclease activity removes the mispaired nucleotide, and the polymerase begins again. This activity, known as **proofreading**, is not simply the reverse of the polymerization reaction (Eqn 25-1), because pyrophosphate is not involved. The polymerizing and proofreading activities of a DNA polymerase can be measured separately. Proofreading improves the inherent accuracy of the polymerization reaction 10^2 - to 10^3 -fold. In the monomeric DNA polymerase I, the polymerizing and proofreading activities have separate active sites within the same polypeptide.

When base selection and proofreading are combined, DNA polymerase leaves behind one net error for every 10^6 to 10^8 bases added. Yet the measured accuracy of replication in *E. coli* is higher still. The additional accuracy is provided by a separate enzyme system that repairs the mismatched base pairs remaining after replication. We describe this mismatch repair, along with other DNA repair processes, in Section 25.2.

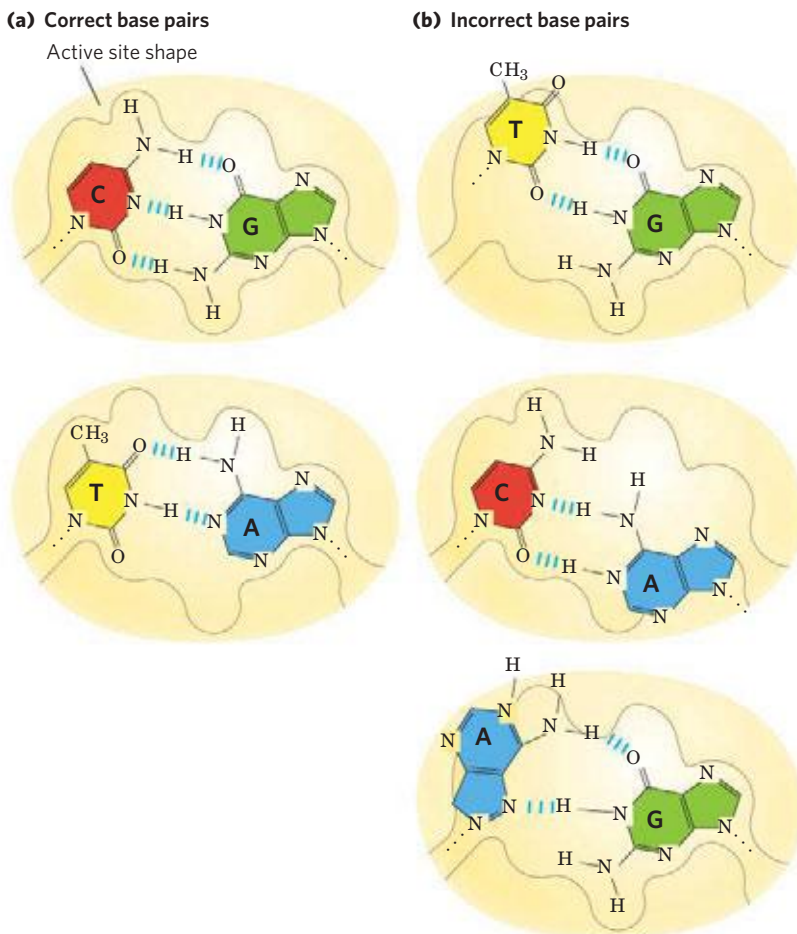


FIGURE 25-6 Contribution of base-pair geometry to the fidelity of DNA replication. (a) The standard A=T and G≡C base pairs have very similar geometries, and an active site sized to fit one will generally accommodate the other. (b) The geometry of incorrectly paired bases can exclude them from the active site, as occurs on DNA polymerase.

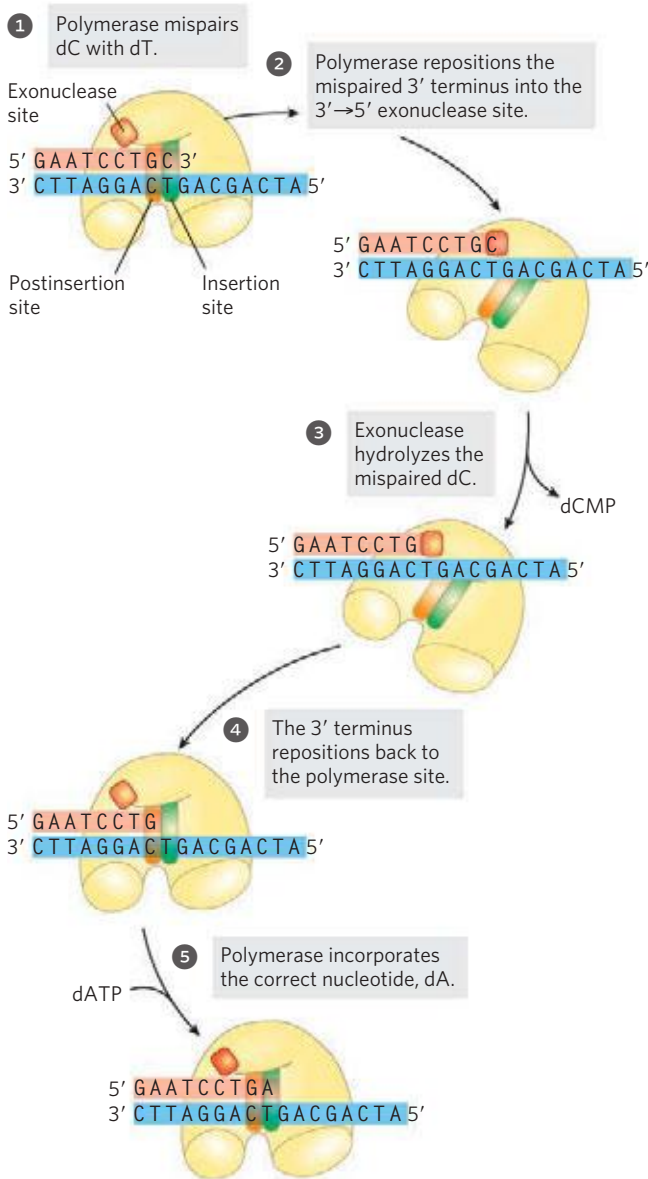


FIGURE 25-7 An example of error correction by the 3'→5' exonuclease activity of DNA polymerase I. Structural analysis has located the exonuclease activity behind the polymerase activity as the enzyme is oriented in its movement along the DNA. A mismatched base (here, a C-T mismatch) impedes translocation of DNA polymerase I to the next site. The DNA bound to the enzyme slides backward into the exonuclease site, and the enzyme corrects the mistake with its 3'→5' exonuclease activity. The enzyme then resumes its polymerase activity in the 5'→3' direction.

E. coli Has at Least Five DNA Polymerases

More than 90% of the DNA polymerase activity observed in *E. coli* extracts can be accounted for by DNA polymerase I. Soon after the isolation of this enzyme in 1955, however, evidence began to accumulate that it is not suited for replication of the large *E. coli* chromosome. First, the rate at which it adds nucleotides (600 nucleotides/min) is too slow (by a factor of 100 or more) to account for the rates at which the replication fork moves in the bacterial cell. Second, DNA polymerase I has a relatively low processivity. Third, genetic studies have demonstrated that many genes, and therefore many proteins, are involved in replication: DNA polymerase I clearly does not act alone. Fourth, and most important, in 1969 John Cairns isolated a bacterial strain with an altered gene for DNA polymerase I that produced an inactive enzyme. Although this strain was abnormally sensitive to agents that damaged DNA, it was nevertheless viable!

A search for other DNA polymerases led to the discovery of *E. coli* **DNA polymerase II** and **DNA polymerase III** in the early 1970s. DNA polymerase II is an enzyme involved in one type of DNA repair (Section 25.3). DNA polymerase III is the principal replication enzyme in *E. coli*. The properties of these three DNA polymerases are compared in Table 25-1. DNA polymerases IV and V, identified in 1999, are involved in an unusual form of DNA repair (Section 25.2).

TABLE 25-1 Comparison of Three DNA Polymerases of *E. coli*

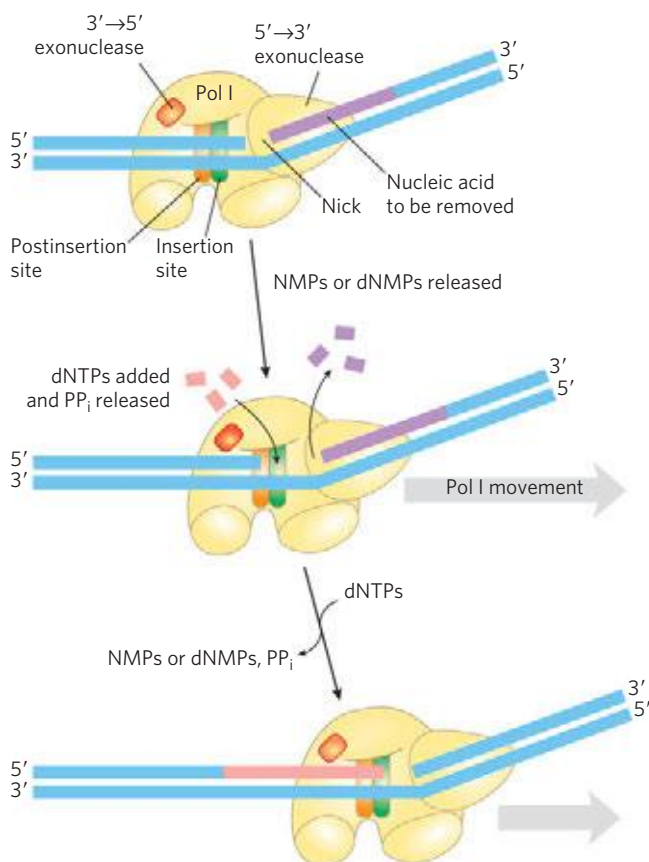
	DNA polymerase		
	I	II	III
Structural gene*	<i>polA</i>	<i>polB</i>	<i>polC (dnaE)</i>
Subunits (number of different types)	1	7	≥10
M_r	103,000	88,000 [†]	791,500
3'→5' Exonuclease (proofreading)	Yes	Yes	Yes
5'→3' Exonuclease	Yes	No	No
Polymerization rate (nucleotides/s)	10–20	40	250–1,000
Processivity (nucleotides added before polymerase dissociates)	3–200	1,500	≥500,000

*For enzymes with more than one subunit, the gene listed here encodes the subunit with polymerization activity. Note that *dnaE* is an earlier designation for the gene now referred to as *polC*.

[†]Polymerization subunit only. DNA polymerase II shares several subunits with DNA polymerase III, including the β , γ , δ , δ' , χ , and Ψ subunits (see Table 25-2).

DNA polymerase I, then, is not the primary enzyme of replication; instead it performs a host of cleanup functions during replication, recombination, and repair. The polymerase's special functions are enhanced by its 5'→3' exonuclease activity. This activity, distinct from the 3'→5' proofreading exonuclease (Fig. 25-7), is located in a structural domain that can be separated from the enzyme by mild protease treatment. When the 5'→3' exonuclease domain is removed, the remaining fragment (M_r 68,000), the **large fragment** or **Klenow fragment**, retains the polymerization and proofreading activities. The 5'→3' exonuclease activity of intact DNA polymerase I can replace a segment of DNA (or RNA) paired to the template strand, in a process known as nick translation (Fig. 25-8). Most other DNA polymerases lack a 5'→3' exonuclease activity.

DNA polymerase III is much more complex than DNA polymerase I (Table 25-2). Its polymerization and proofreading activities reside in its α and ϵ (epsilon) subunits, respectively. The θ subunit associates with α and ϵ to form a core polymerase, which can polymerize DNA but with limited processivity. Two core polymerases can be linked by another set of subunits, a clamp-loading complex, or γ complex, consisting of five subunits of four different types, $\tau_2\gamma\delta\delta'$. The core polymerases are linked through the τ (tau) subunits.



Two additional subunits, χ (chi) and ψ (psi), are bound to the clamp-loading complex. The entire assembly of 13 protein subunits (nine different types) is called DNA polymerase III* (Fig. 25-9a).

DNA polymerase III* can polymerize DNA, but with a much lower processivity than one would expect for the organized replication of an entire chromosome. The necessary increase in processivity is provided by the addition of the β subunits, four of which complete the DNA polymerase III holoenzyme. The β subunits associate in pairs to form donut-shaped structures that encircle the DNA and act like clamps (Fig. 25-9b). Each dimer associates with a core subassembly of polymerase III* (one dimeric clamp per core subassembly) and slides along the DNA as replication proceeds. The β sliding clamp prevents the dissociation of DNA polymerase III from DNA, dramatically increasing processivity—to greater than 500,000 (Table 25-1).

DNA Replication Requires Many Enzymes and Protein Factors

Replication in *E. coli* requires not just a single DNA polymerase but 20 or more different enzymes and proteins, each performing a specific task. The entire complex has been termed the **DNA replicase system** or **replisome**. The enzymatic complexity of replication reflects the constraints imposed by the structure of DNA and by the requirements for accuracy. The main classes of replication enzymes are considered here in terms of the problems they overcome.

Access to the DNA strands that are to act as templates requires separation of the two parent strands. This is generally accomplished by **helicases**, enzymes that move along the DNA and separate the strands, using chemical energy from ATP. Strand separation creates topological stress in the helical DNA structure (see Fig. 24-11), which is relieved by the action of **topoisomerases**. The separated strands are stabilized by **DNA-binding proteins**. As noted earlier, before DNA polymerases can begin synthesizing DNA, primers must be present on the template—generally, short

FIGURE 25-8 Nick translation. The bacterial DNA polymerase I has three domains, catalyzing its DNA polymerase, 5'→3' exonuclease, and 3'→5' exonuclease activities. The 5'→3' exonuclease domain is in front of the enzyme as it moves along the DNA and is not shown in Figure 25-5. By degrading the DNA strand ahead of the enzyme and synthesizing a new strand behind, DNA polymerase I can promote a reaction called nick translation, where a break or nick in the DNA is effectively moved along with the enzyme. This process has a role in DNA repair and in the removal of RNA primers during replication (both described later). The strand of nucleic acid to be removed (either DNA or RNA) is shown in purple, the replacement strand in red. DNA synthesis begins at a nick (a broken phosphodiester bond, leaving a free 3' hydroxyl and a free 5' phosphate). A nick remains where DNA polymerase I eventually dissociates, and the nick is later sealed by another enzyme.

TABLE 25-2 Subunits of DNA Polymerase III of *E. coli*

Subunit	Number of subunits per holoenzyme	M_r of subunit	Gene	Function of subunit	
α	2	129,900	<i>polC (dnaE)</i>	Polymerization activity	Core polymerase
ϵ	2	27,500	<i>dnaQ (mutD)</i>	3'→5' Proofreading exonuclease	
θ	2	8,600	<i>holE</i>	Stabilization of ϵ subunit	
τ	2	71,100	<i>dnaX</i>	Stable template binding; core enzyme dimerization	Clamp-loading (γ) complex that loads β subunits on lagging strand at each Okazaki fragment
γ	1	47,500	<i>dnaX*</i>	Clamp loader	
δ	1	38,700	<i>holA</i>	Clamp opener	
δ'	1	36,900	<i>holB</i>	Clamp loader	
χ	1	16,600	<i>holC</i>	Interaction with SSB	
ψ	1	15,200	<i>holD</i>	Interaction with γ and χ	
β	4	40,600	<i>dnaN</i>	DNA clamp required for optimal processivity	

*The γ subunit is encoded by a portion of the gene for the τ subunit, such that the amino-terminal 66% of the τ subunit has the same amino acid sequence as the γ subunit. The γ subunit is generated by a translational frameshifting mechanism (p. 1111) that leads to premature translational termination.

segments of RNA synthesized by enzymes known as **primases**. Ultimately, the RNA primers are removed and replaced by DNA; in *E. coli*, this is one of the many functions of DNA polymerase I. After an RNA primer is removed and the gap is filled in with DNA, a nick

remains in the DNA backbone in the form of a broken phosphodiester bond. These nicks are sealed by **DNA ligases**. All these processes require coordination and regulation, an interplay best characterized in the *E. coli* system.

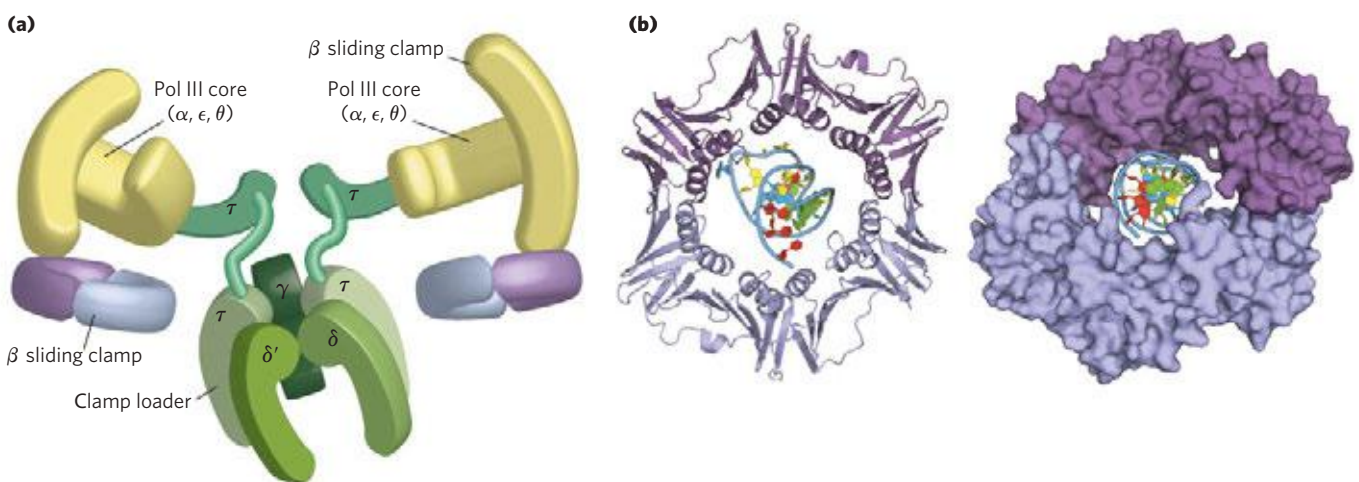


FIGURE 25-9 DNA polymerase III. (a) Architecture of bacterial DNA polymerase III (Pol III). Two core domains, composed of subunits α , ϵ , and θ , are linked by a five-subunit clamp-loading complex (also known as the γ complex) with the composition $\tau_2\gamma\delta\delta'$. The core subunits and clamp-loader complex constitute DNA polymerase III*. The γ and τ subunits are encoded by the same gene. The γ subunit is a shortened version of the τ subunit; τ thus contains a domain identical to γ , along with an additional segment that interacts with the core polymerase. The other two subunits of DNA polymerase III*, χ and ψ (not shown), also bind to the clamp-loading

complex. Two β clamps interact with the two-core subassembly, each clamp a dimer of the β subunit. The complex interacts with the DnaB helicase (described later in the text) through the τ subunits. (b) Two β subunits of *E. coli* polymerase III form a circular clamp that surrounds the DNA. The clamp slides along the DNA molecule, increasing the processivity of the polymerase III holoenzyme to greater than 500,000 nucleotides by preventing its dissociation from the DNA. The two β subunits are shown in two shades of purple as ribbon structures (left) and surface images (right), surrounding the DNA (derived from PDB ID 2POL).

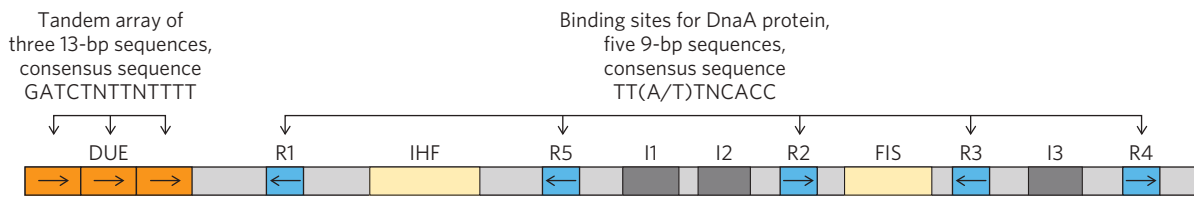


FIGURE 25-10 Arrangement of sequences in the *E. coli* replication origin, *oriC*. Consensus sequences (p. 104) for key repeated elements are shown. N represents any of the four nucleotides. The horizontal arrows indicate the orientations of the nucleotide sequences (left-to-right arrow denotes

sequence in top strand; right-to-left, bottom strand). FIS and IHF are binding sites for proteins described in the text. R sites are bound by DnaA. I sites are additional DnaA-binding sites (with different sequences), bound by DnaA only when the protein is complexed with ATP.

Replication of the *E. coli* Chromosome Proceeds in Stages

The synthesis of a DNA molecule can be divided into three stages: initiation, elongation, and termination, distinguished both by the reactions taking place and by the enzymes required. As you will find here and in the next two chapters, synthesis of the major information-containing biological polymers—DNAs, RNAs, and proteins—can be understood in terms of these same three stages, with the stages of each pathway having unique characteristics. The events described below reflect information derived primarily from *in vitro* experiments using purified *E. coli* proteins, although the principles are highly conserved in all replication systems.

Initiation The *E. coli* replication origin, *oriC*, consists of 245 bp and contains DNA sequence elements that are highly conserved among bacterial replication origins. The general arrangement of the conserved sequences is illustrated in **Figure 25-10**. Two types of sequences

are of special interest: five repeats of a 9 bp sequence (R sites) that serve as binding sites for the key initiator protein DnaA, and a region rich in A=T base pairs called the **DNA unwinding element (DUE)**. There are three additional DnaA-binding sites (I sites), and binding sites for the proteins IHF (integration host factor) and FIS (factor for inversion stimulation). These two proteins were discovered as required components of certain recombination reactions described later in this chapter, and their names reflect those roles. Another DNA-binding protein, HU (a histonelike bacterial protein originally dubbed factor U), also participates but does not have a specific binding site.

At least 10 different enzymes or proteins (summarized in Table 25-3) participate in the initiation phase of replication. They open the DNA helix at the origin and establish a prepriming complex for subsequent reactions. The crucial component in the initiation process is the DnaA protein, a member of the **AAA+ ATPase** protein family (ATPases associated with diverse cellular activities). Many AAA+ ATPases, including DnaA,

TABLE 25-3 Proteins Required to Initiate Replication at the *E. coli* Origin

Protein	M_r	Number of subunits	Function
DnaA protein	52,000	1	Recognizes <i>ori</i> sequence; opens duplex at specific sites in origin
DnaB protein (helicase)	300,000	6*	Unwinds DNA
DnaC protein	174,000	6*	Required for DnaB binding at origin
HU	19,000	2	Histonelike protein; DNA-binding protein; stimulates initiation
FIS	22,500	2*	DNA-binding protein; stimulates initiation
IHF	22,000	2	DNA-binding protein; stimulates initiation
Primase (DnaG protein)	60,000	1	Synthesizes RNA primers
Single-stranded DNA-binding protein (SSB)	75,600	4*	Binds single-stranded DNA
DNA gyrase (DNA topoisomerase II)	400,000	4	Relieves torsional strain generated by DNA unwinding
Dam methylase	32,000	1	Methylates (5')GATC sequences at <i>oriC</i>

*Subunits in these cases are identical.

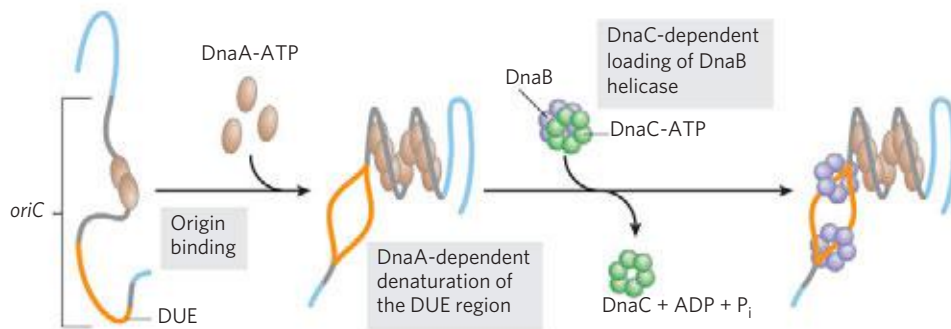


FIGURE 25-11 Model for initiation of replication at the *E. coli* origin, *oriC*.

Eight DnaA protein molecules, each with a bound ATP, bind at the R and I sites in the origin (see Fig. 25-10). The DNA is wrapped around this complex, which forms a right-handed helical structure. The A=T-rich DUE region is denatured as a result of the strain imparted by the adjacent DnaA binding.

form oligomers and hydrolyze ATP relatively slowly. This ATP hydrolysis acts as a switch mediating interconversion of the protein between two states. In the case of DnaA, the ATP-bound form is active and the ADP-bound form is inactive.

Eight DnaA protein molecules, all in the ATP-bound state, assemble to form a helical complex encompassing the R and I sites in *oriC* (Fig. 25-11). DnaA has a higher affinity for the R sites than I sites, and binds R sites equally well in its ATP- or ADP-bound form. The I sites, which bind only the ATP-bound DnaA, allow discrimination between the active and inactive forms of DnaA. The tight right-handed wrapping of the DNA around this complex introduces an effective positive supercoil (see Chapter 24). The associated strain in the nearby DNA leads to denaturation in the A=T-rich DUE region. The complex formed at the replication origin also includes several DNA-binding proteins—HU, IHF, and FIS—that facilitate DNA bending.

The DnaC protein, another AAA+ ATPase, then loads the DnaB protein onto the separated DNA strands in the denatured region. A hexamer of DnaC, each subunit bound to ATP, forms a tight complex with the hexameric, ring-shaped DnaB helicase. This DnaC-DnaB interaction opens the DnaB ring, the process being aided by a further interaction between DnaB and DnaA. Two of the ring-shaped DnaB hexamers are loaded in the DUE, one onto each DNA strand. The ATP bound to DnaC is hydrolyzed, releasing the DnaC and leaving the DnaB bound to the DNA.

Loading of the DnaB helicase is the key step in replication initiation. As a replicative helicase, DnaB migrates along the single-stranded DNA in the 5'→3' direction, unwinding the DNA as it travels. The DnaB helicases loaded onto the two DNA strands thus travel in opposite directions, creating two potential replication forks. All other proteins at the replication fork are linked directly or indirectly to DnaB. The DNA polymerase III

Formation of the helical DnaA complex is facilitated by the proteins HU, IHF, and FIS, which are not shown here because their detailed structural roles have not yet been defined. Hexamers of the DnaB protein bind to each strand, with the aid of DnaC protein. The DnaB helicase activity further unwinds the DNA in preparation for priming and DNA synthesis.

holoenzyme is linked through the τ subunits; additional DnaB interactions are described below. As replication begins and the DNA strands are separated at the fork, many molecules of single-stranded DNA-binding protein (SSB) bind to and stabilize the separated strands, and DNA gyrase (DNA topoisomerase II) relieves the topological stress induced ahead of the fork by the unwinding reaction.

Initiation is the only phase of DNA replication that is known to be regulated, and it is regulated such that replication occurs only once in each cell cycle. The mechanism of regulation is not yet entirely understood, but genetic and biochemical studies have provided insights into several separate regulatory mechanisms.

Once DNA polymerase III has been loaded onto the DNA, along with the β subunits (signaling completion of the initiation phase), the protein Hda binds to the β subunits and interacts with DnaA to stimulate hydrolysis of its bound ATP. Hda is yet another AAA+ ATPase closely related to DnaA (its name is derived from *homologous to DnaA*). This ATP hydrolysis leads to disassembly of the DnaA complex at the origin. Slow release of ADP by DnaA and rebinding of ATP cycles the protein between its inactive (with bound ADP) and active (with bound ATP) forms on a time scale of 20 to 40 minutes.

The timing of replication initiation is affected by DNA methylation and interactions with the bacterial plasma membrane. The *oriC* DNA is methylated by the Dam methylase (Table 25-3), which methylates the N^6 position of adenine within the palindromic sequence (5')GATC. (Dam is not a biochemical expletive; it stands for *DNA adenine methylation*.) The *oriC* region of *E. coli* is highly enriched in GATC sequences—it has 11 of them in its 245 bp, whereas the average frequency of GATC in the *E. coli* chromosome as a whole is 1 in 256 bp.

Immediately after replication, the DNA is hemimethylated: the parent strands have methylated *oriC* sequences

but the newly synthesized strands do not. The hemimethylated *oriC* sequences are now sequestered by interaction with the plasma membrane (the mechanism is unknown) and by binding of the protein SeqA. After a time, *oriC* is released from the plasma membrane, SeqA dissociates, and the DNA must be fully methylated by Dam methylase before it can again bind DnaA and initiate a new round of replication.

Elongation The elongation phase of replication includes two distinct but related operations: leading strand synthesis and lagging strand synthesis. Several enzymes at the replication fork are important to the synthesis of both strands. Parent DNA is first unwound by DNA helicases, and the resulting topological stress is relieved by topoisomerases. Each separated strand is then stabilized by SSB. From this point, synthesis of leading and lagging strands is sharply different.

Leading strand synthesis, the more straightforward of the two, begins with the synthesis by primase (DnaG protein) of a short (10 to 60 nucleotide) RNA primer at the replication origin. DnaG interacts with DnaB helicase to carry out this reaction, and the primer is synthesized in the direction opposite to that in which the DnaB helicase is moving. In effect, the DnaB helicase moves along the strand that becomes the lagging strand in DNA synthesis; however, the first

primer laid down in the first DnaG-DnaB interaction serves to prime leading strand DNA synthesis in the opposite direction. Deoxyribonucleotides are added to this primer by a DNA polymerase III complex linked to the DnaB helicase tethered to the opposite DNA strand. Leading strand synthesis then proceeds continuously, keeping pace with the unwinding of DNA at the replication fork.

Lagging strand synthesis, as we have noted, is accomplished in short Okazaki fragments (**Fig. 25-12a**). First, an RNA primer is synthesized by primase and, as in leading strand synthesis, DNA polymerase III binds to the RNA primer and adds deoxyribonucleotides (Fig. 25-12b). On this level, the synthesis of each Okazaki fragment seems straightforward, but the reality is quite complex. The complexity lies in the *coordination* of leading and lagging strand synthesis. Both strands are produced by a *single* asymmetric DNA polymerase III dimer; this is accomplished by looping the DNA of the lagging strand as shown in **Figure 25-13**, bringing together the two points of polymerization.

The synthesis of Okazaki fragments on the lagging strand entails some elegant enzymatic choreography. DnaB helicase and DnaG primase constitute a functional unit within the replication complex, the **primosome**. DNA polymerase III uses one set of its core

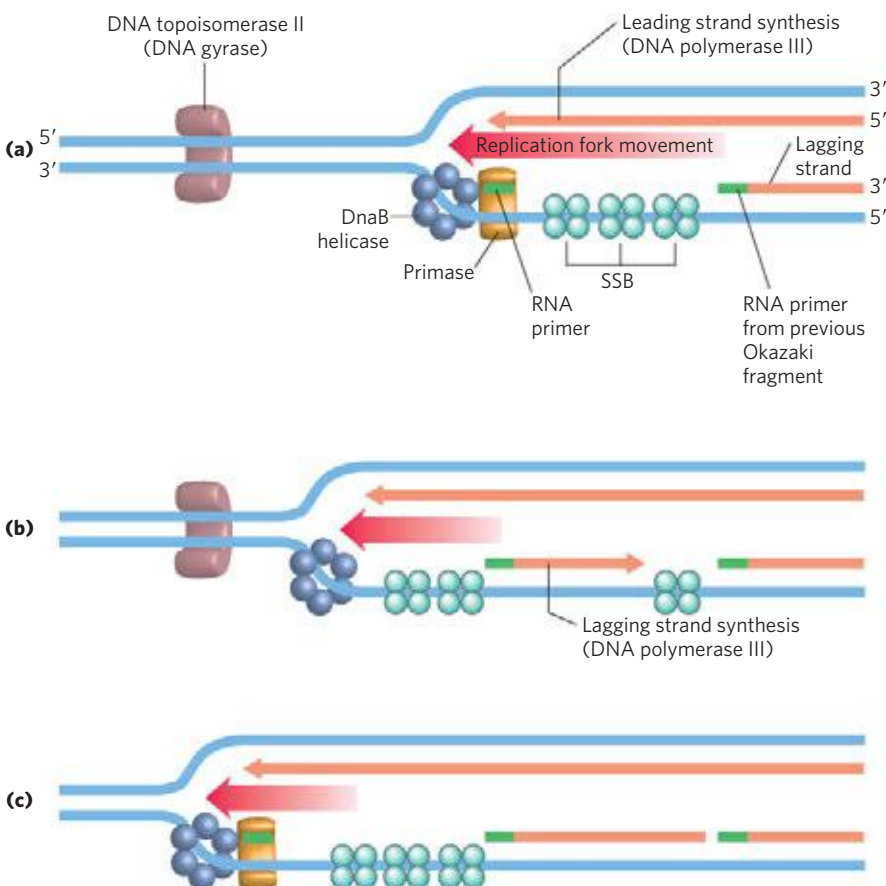


FIGURE 25-12 Synthesis of Okazaki fragments.

(a) At intervals, primase synthesizes an RNA primer for a new Okazaki fragment. Note that if we consider the two template strands as lying side by side, lagging strand synthesis formally proceeds in the opposite direction from fork movement. **(b)** Each primer is extended by DNA polymerase III. **(c)** DNA synthesis continues until the fragment extends as far as the primer of the previously added Okazaki fragment. A new primer is synthesized near the replication fork to begin the process again.

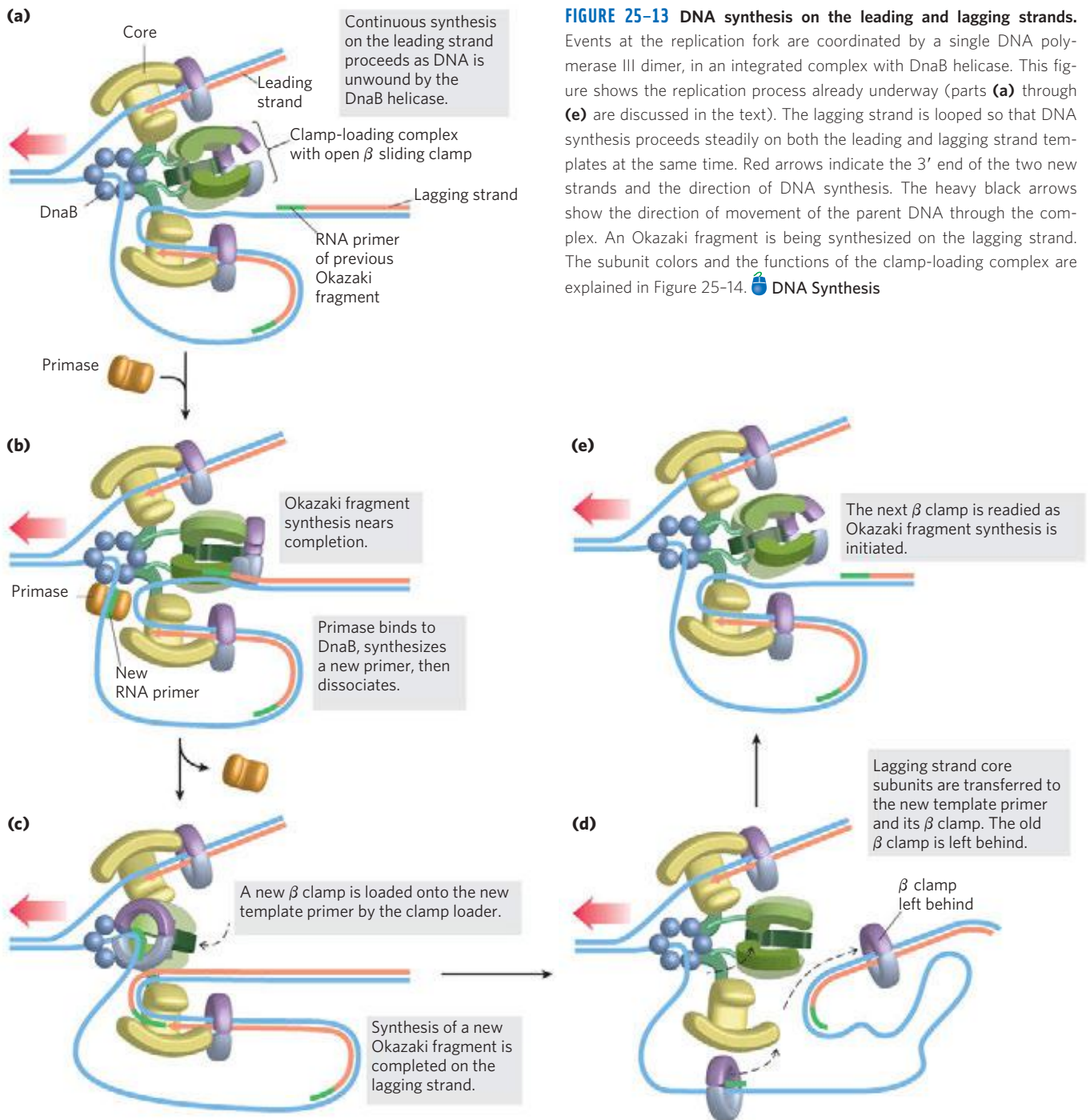


FIGURE 25-13 DNA synthesis on the leading and lagging strands. Events at the replication fork are coordinated by a single DNA polymerase III dimer, in an integrated complex with DnaB helicase. This figure shows the replication process already underway (parts **(a)** through **(e)** are discussed in the text). The lagging strand is looped so that DNA synthesis proceeds steadily on both the leading and lagging strand templates at the same time. Red arrows indicate the 3' end of the two new strands and the direction of DNA synthesis. The heavy black arrows show the direction of movement of the parent DNA through the complex. An Okazaki fragment is being synthesized on the lagging strand. The subunit colors and the functions of the clamp-loading complex are explained in Figure 25-14. DNA Synthesis

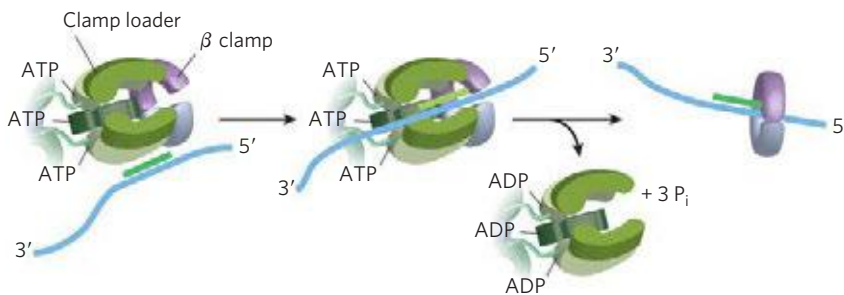
subunits (the core polymerase) to synthesize the leading strand continuously, while the other set of core subunits cycles from one Okazaki fragment to the next on the looped lagging strand. DnaB helicase, bound in front of DNA polymerase III, unwinds the DNA at the replication fork (Fig. 25-13a) as it travels along the lagging strand template in the 5'→3' direction. DnaG primase occasionally associates with DnaB helicase and synthesizes a short RNA primer (Fig. 25-13b). A new β sliding clamp is then positioned at the primer by the clamp-loading complex of DNA

polymerase III (Fig. 25-13c). When synthesis of an Okazaki fragment has been completed, replication halts, and the core subunits of DNA polymerase III dissociate from their β sliding clamp (and from the completed Okazaki fragment) and associate with the new clamp (Fig. 25-13d, e). This initiates synthesis of a new Okazaki fragment. As noted earlier, the entire complex responsible for coordinated DNA synthesis at a replication fork is known as the replisome. The proteins acting at the replication fork are summarized in Table 25-4.

TABLE 25-4 Proteins of the *E. coli* Replisome

Protein	M_r	Number of subunits	Function
SSB	75,600	4	Binding to single-stranded DNA
DnaB protein (helicase)	300,000	6	DNA unwinding; primosome constituent
Primase (DnaG protein)	60,000	1	RNA primer synthesis; primosome constituent
DNA polymerase III	791,500	17	New strand elongation
DNA polymerase I	103,000	1	Filling of gaps; excision of primers
DNA ligase	74,000	1	Ligation
DNA gyrase (DNA topoisomerase II)	400,000	4	Supercoiling

Source: Modified from Kornberg, A. (1982) *Supplement to DNA Replication*, Table 511-2, W. H. Freeman and Company, New York.

**FIGURE 25-14** The DNA polymerase III clamp loader.

The five subunits of the clamp-loading complex are the γ , δ , and δ' subunits and the amino-terminal domain of each τ subunit (see Fig. 25-9). The complex binds to three molecules of ATP and to a dimeric β clamp. This binding forces the β clamp open at one of its two subunit interfaces. Hydrolysis of the bound ATP allows the β clamp to close again around the DNA.

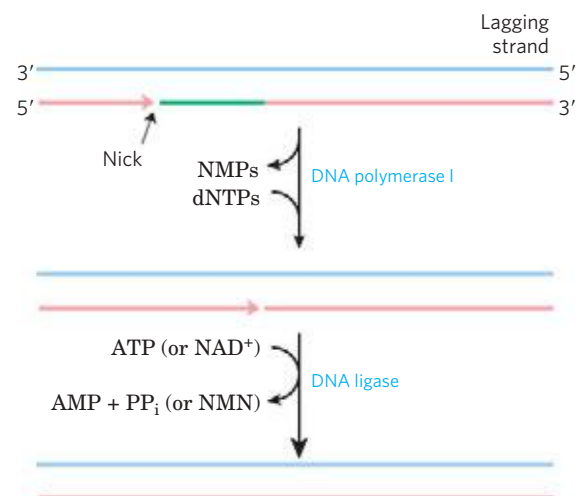
The clamp-loading complex of DNA polymerase III, consisting of parts of the two τ subunits along with the γ , δ , and δ' subunits, is also an AAA+ ATPase. This complex binds to ATP and to the new β sliding clamp. The binding imparts strain on the dimeric clamp, opening up the ring at one subunit interface (Fig. 25-14). The newly primed lagging strand is slipped into the ring through the resulting break. The clamp loader then hydrolyzes ATP, releasing the β sliding clamp and allowing it to close around the DNA.

The replisome promotes rapid DNA synthesis, adding $\sim 1,000$ nucleotides/s to each strand (leading and lagging). Once an Okazaki fragment has been completed, its RNA primer is removed and replaced with DNA by DNA polymerase I, and the remaining nick is sealed by DNA ligase (Fig. 25-15).

DNA ligase catalyzes the formation of a phosphodiester bond between a 3' hydroxyl at the end of one DNA strand and a 5' phosphate at the end of another strand. The phosphate must be activated by adenylation. DNA ligases isolated from viruses and eukaryotes use ATP for this purpose. DNA ligases from bacteria are unusual in that many use NAD^+ —a cofactor that usually functions in hydride transfer reactions (see Fig. 13-24)—as the source of the AMP activating group (Fig. 25-16). DNA ligase is another enzyme of DNA metabolism that has become an

important reagent in recombinant DNA experiments (see Fig. 9-1).

Termination Eventually, the two replication forks of the circular *E. coli* chromosome meet at a terminus region

**FIGURE 25-15** Final steps in the synthesis of lagging strand segments.

RNA primers in the lagging strand are removed by the 5'→3' exonuclease activity of DNA polymerase I and are replaced with DNA by the same enzyme. The remaining nick is sealed by DNA ligase. The role of ATP or NAD^+ is shown in Figure 25-16.

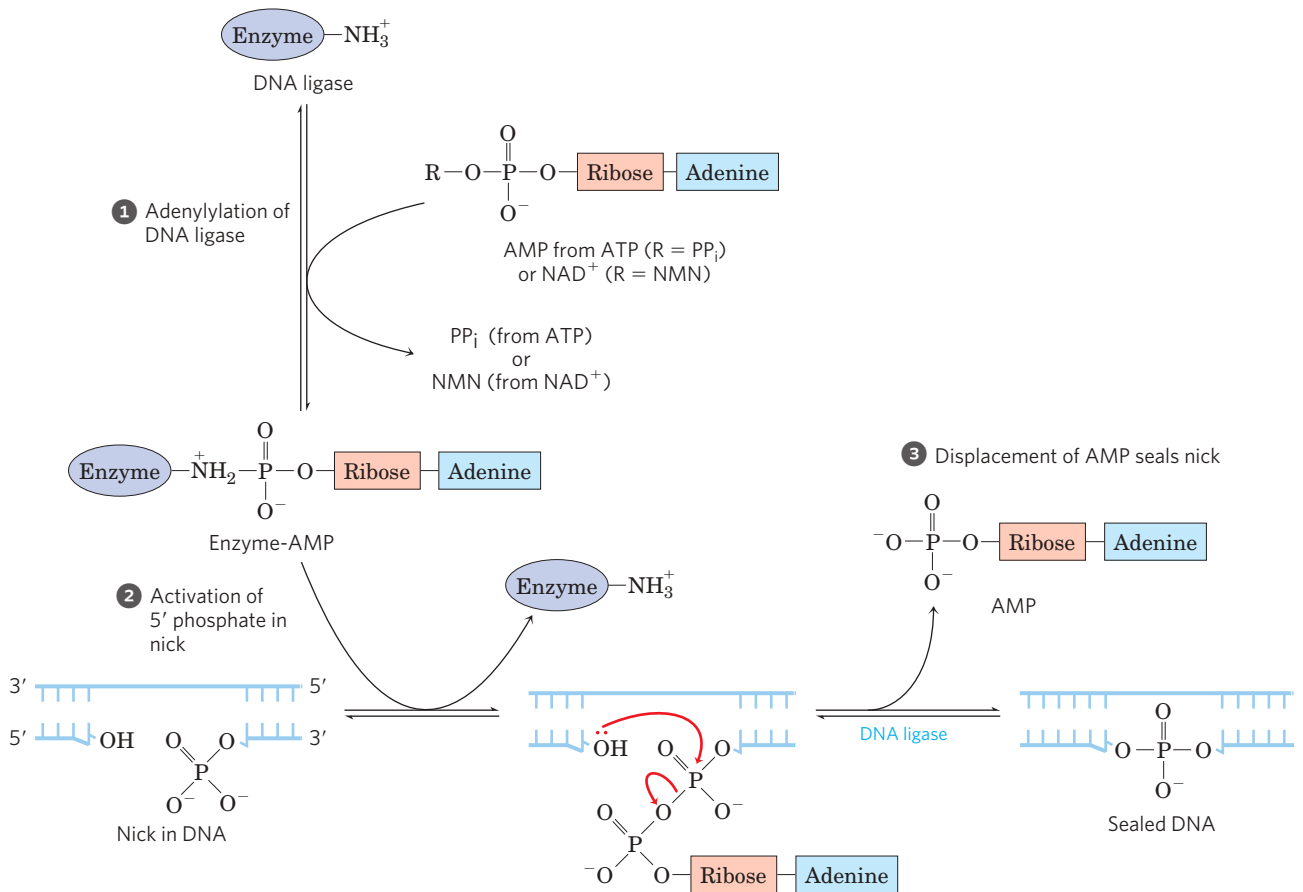


FIGURE 25-16 Mechanism of the DNA ligase reaction. In each of the three steps, one phosphodiester bond is formed at the expense of another. Steps 1 and 2 lead to activation of the 5' phosphate in the nick. An AMP group is transferred first to a Lys residue on the enzyme and then to the 5' phosphate in the nick. In step 3, the 3'-hydroxyl

group attacks this phosphate and displaces AMP, producing a phosphodiester bond to seal the nick. In the *E. coli* DNA ligase reaction, AMP is derived from NAD^+ . The DNA ligases isolated from some viral and eukaryotic sources use ATP rather than NAD^+ and they release pyrophosphate rather than nicotinamide mononucleotide (NMN) in step 1.

containing multiple copies of a 20 bp sequence called Ter (Fig. 25-17). The Ter sequences are arranged on the chromosome to create a trap that a replication fork

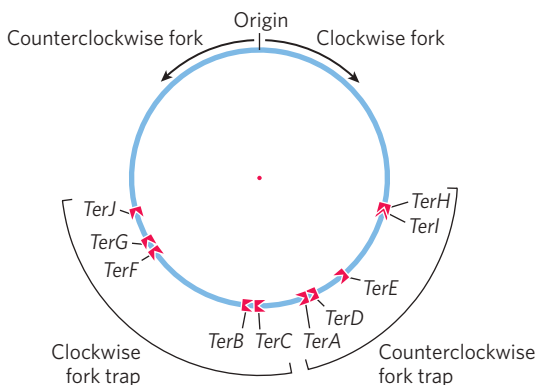


FIGURE 25-17 Termination of chromosome replication in *E. coli*. The Ter sequences (*TerA* through *TerJ*) are positioned on the chromosome in two clusters with opposite orientations.

can enter but cannot leave. The Ter sequences function as binding sites for the protein Tus (terminus utilization substance). The Tus-Ter complex can arrest a replication fork from only one direction. Only one Tus-Ter complex functions per replication cycle—the complex first encountered by either replication fork. Given that opposing replication forks generally halt when they collide, Ter sequences would not seem to be essential, but they may prevent overreplication by one fork in the event that the other is delayed or halted by an encounter with DNA damage or some other obstacle.

So, when either replication fork encounters a functional Tus-Ter complex, it halts; the other fork halts when it meets the first (arrested) fork. The final few hundred base pairs of DNA between these large protein complexes are then replicated (by an as yet unknown mechanism), completing two topologically interlinked (catenated) circular chromosomes (Fig. 25-18). DNA circles linked in this way are known as **catenanes**. Separation of the catenated circles in *E. coli* requires

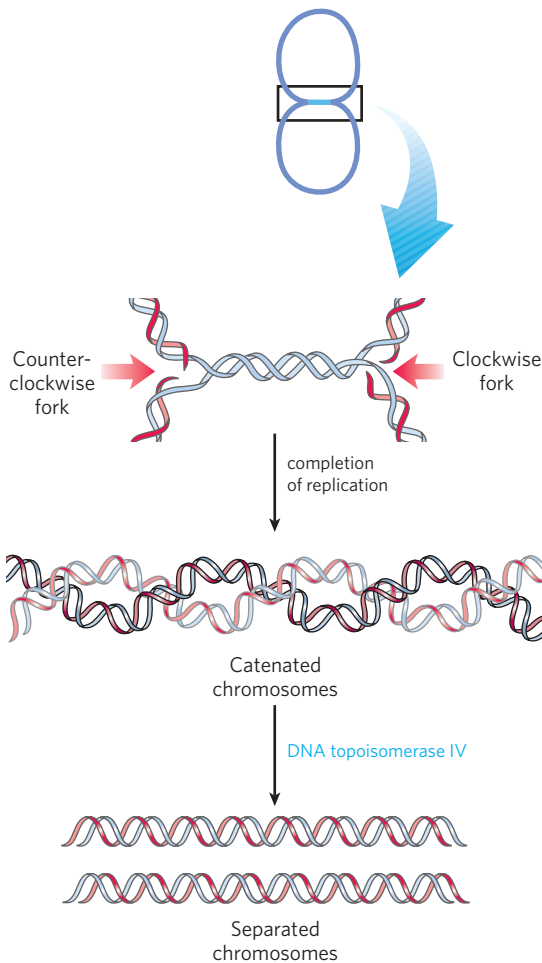


FIGURE 25-18 Role of topoisomerases in replication termination. Replication of the DNA separating opposing replication forks leaves the completed chromosomes joined as catenanes, or topologically interlinked circles. The circles are not covalently linked, but because they are interwound and each is covalently closed, they cannot be separated—except by the action of topoisomerases. In *E. coli*, a type II topoisomerase known as DNA topoisomerase IV plays the primary role in the separation of catenated chromosomes, transiently breaking both DNA strands of one chromosome and allowing the other chromosome to pass through the break.

topoisomerase IV (a type II topoisomerase). The separated chromosomes then segregate into daughter cells at cell division. The terminal phase of replication of other circular chromosomes, including many of the DNA viruses that infect eukaryotic cells, is similar.

Replication in Eukaryotic Cells Is Similar but More Complex

The DNA molecules in eukaryotic cells are considerably larger than those in bacteria and are organized into complex nucleoprotein structures (chromatin; p. 994). The essential features of DNA replication are the same

in eukaryotes and bacteria, and many of the protein complexes are functionally and structurally conserved. However, eukaryotic replication is regulated and coordinated with the cell cycle, introducing some additional complexities.

Origins of replication have a well-characterized structure in some lower eukaryotes, but they are much less defined in higher eukaryotes. In vertebrates, a variety of A=T-rich sequences may be used for replication initiation, and the sites may vary from one cell division to the next. Yeast (*Saccharomyces cerevisiae*) has defined replication origins called autonomously replicating sequences (ARS), or **replicators**. Yeast replicators span ~150 bp and contain several essential, conserved sequences. About 400 replicators are distributed among the 16 chromosomes of the haploid yeast genome.

Regulation ensures that all cellular DNA is replicated once per cell cycle. Much of this regulation involves proteins called cyclins and the cyclin-dependent kinases (CDKs) with which they form complexes (p. 484). The cyclins are rapidly destroyed by ubiquitin-dependent proteolysis at the end of the M phase (mitosis), and the absence of cyclins allows the establishment of **pre-replicative complexes (pre-RCs)** on replication initiation sites. In rapidly growing cells, the pre-RC forms at the end of M phase. In slow-growing cells, it does not form until the end of G1. Formation of the pre-RC renders the cell competent for replication, an event sometimes called **licensing**.

As in bacteria, the key event in the initiation of replication in all eukaryotes is the loading of the replicative helicase, a heterohexameric complex of **mini-chromosome maintenance (MCM) proteins** (MCM2 to MCM7). The ring-shaped MCM2–7 helicase, functioning much like the bacterial DnaB helicase, is loaded onto the DNA by another six-protein complex called **ORC (origin recognition complex)** (Fig. 25-19). ORC has five AAA+ ATPase domains among its subunits and is functionally analogous to the bacterial DnaA. Two other proteins, CDC6 (cell division cycle) and CDT1 (CDC10-dependent transcript 1), are also required to load the MCM2–7 complex, and the yeast CDC6 is another AAA+ ATPase.

Commitment to replication requires the synthesis and activity of S-phase cyclin-CDK complexes (such as the cyclin E–CDK2 complex; see Fig. 12-46) and CDC7-DBF4. Both types of complexes help to activate replication by binding to and phosphorylating several proteins in the pre-RC. Other cyclins and CDKs function to inhibit the formation of more pre-RC complexes once replication has been initiated. For example, CDK2 binds to cyclin A as cyclin E levels decline during S phase, inhibiting CDK2 and preventing the licensing of additional pre-RC complexes.

The rate of movement of the replication fork in eukaryotes (~50 nucleotides/s) is only one-twentieth that observed in *E. coli*. At this rate, replication of an

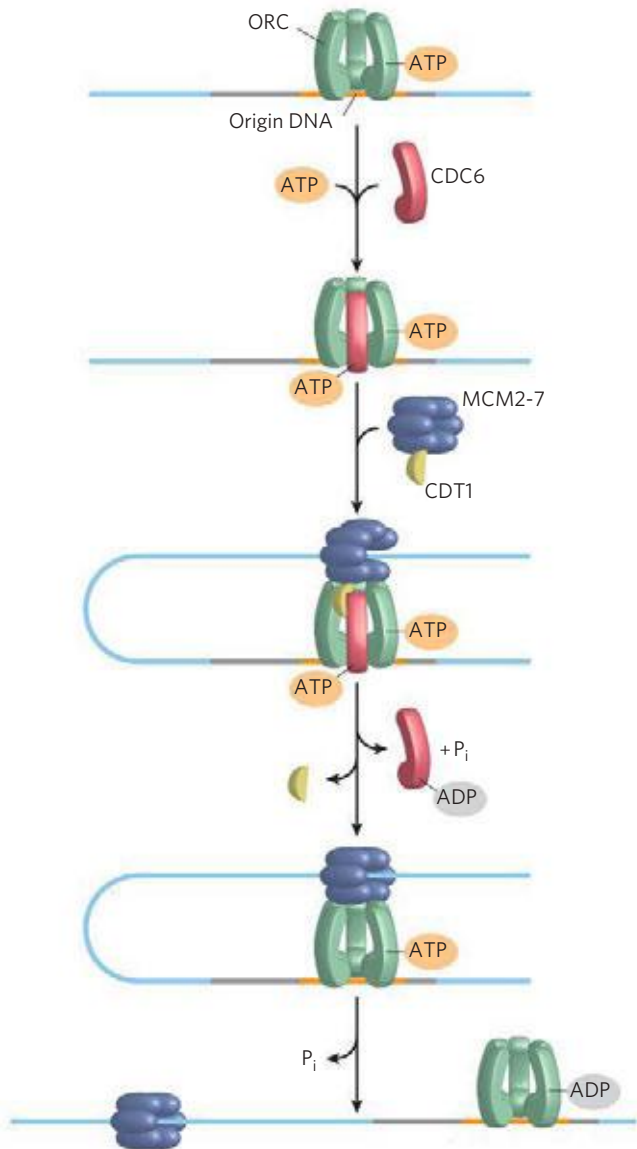


FIGURE 25-19 Assembly of a pre-replicative complex at a eukaryotic replication origin. The initiation site (origin) is bound by ORC, CDC6, and CDT1. These proteins, many of them AAA+ ATPases, promote loading of the replicative helicase, MCM2-7, in a reaction that is analogous to the loading of the bacterial DnaB helicase by DnaC protein. Loading of the MCM helicase complex onto the DNA forms the pre-replicative complex, or pre-RC, and is the key step in the initiation of replication.

average human chromosome proceeding from a single origin would take more than 500 hours. Replication of human chromosomes in fact proceeds bidirectionally from many origins, spaced 30 to 300 kbp apart. Eukaryotic chromosomes are almost always much larger than bacterial chromosomes, so multiple origins are probably a universal feature of eukaryotic cells.

Like bacteria, eukaryotes have several types of DNA polymerases. Some have been linked to particular


functions, such as the replication of mitochondrial DNA. The replication of nuclear chromosomes involves DNA polymerase α , in association with DNA polymerase δ . **DNA polymerase α** is typically a multisubunit enzyme with similar structure and properties in all eukaryotic cells. One subunit has a primase activity, and the largest subunit ($M_r \sim 180,000$) contains the polymerization activity. However, this polymerase has no proofreading 3'→5' exonuclease activity, making it unsuitable for high-fidelity DNA replication. DNA polymerase α is believed to function only in the synthesis of short primers (either RNA or DNA) for Okazaki fragments on the lagging strand. These primers are then extended by the multi-subunit **DNA polymerase δ** . This enzyme is associated with and stimulated by proliferating cell nuclear antigen (PCNA; M_r 29,000), a protein found in large amounts in the nuclei of proliferating cells. The three-dimensional structure of PCNA is remarkably similar to that of the β subunit of *E. coli* DNA polymerase III (Fig. 25-9b), although primary sequence homology is not evident. PCNA has a function analogous to that of the β subunit, forming a circular clamp that greatly enhances the processivity of the polymerase. DNA polymerase δ has a 3'→5' proofreading exonuclease activity and seems to carry out both leading and lagging strand synthesis in a complex comparable to the dimeric bacterial DNA polymerase III.

Yet another polymerase, **DNA polymerase ϵ** , replaces DNA polymerase δ in some situations, such as in DNA repair. DNA polymerase ϵ may also function at the replication fork, perhaps playing a role analogous to that of the bacterial DNA polymerase I, removing the primers of Okazaki fragments on the lagging strand.

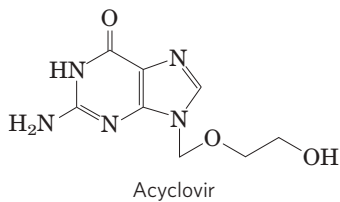
Two other protein complexes also function in eukaryotic DNA replication. RPA (replication protein A) is a eukaryotic single-stranded DNA-binding protein, equivalent in function to the *E. coli* SSB protein. RFC (replication factor C) is a clamp loader for PCNA and facilitates the assembly of active replication complexes. The subunits of the RFC complex have significant sequence similarity to the subunits of the bacterial clamp-loading (γ) complex.

The termination of replication on linear eukaryotic chromosomes involves the synthesis of special structures called **telomeres** at the ends of each chromosome, as discussed in the next chapter.

Viral DNA Polymerases Provide Targets for Antiviral Therapy

 Many DNA viruses encode their own DNA polymerases, and some of these have become targets for pharmaceuticals. For example, the DNA polymerase of the herpes simplex virus is inhibited by acyclovir, a compound developed by Gertrude Elion and

George Hitchings (p. 923). Acyclovir consists of guanine attached to an incomplete ribose ring.



It is phosphorylated by a virally encoded thymidine kinase; acyclovir binds to this viral enzyme with an affinity 200-fold greater than its binding to the cellular thymidine kinase. This ensures that phosphorylation occurs mainly in virus-infected cells. Cellular kinases convert the resulting acyclo-GMP to acyclo-GTP, which is both an inhibitor and a substrate of DNA polymerases; acyclo-GTP competitively inhibits the herpes DNA polymerase more strongly than cellular DNA polymerases. Because it lacks a 3' hydroxyl, acyclo-GTP also acts as a chain terminator when incorporated into DNA. Thus viral replication is inhibited at several steps. ■

SUMMARY 25.1 DNA Replication

- ▶ Replication of DNA occurs with very high fidelity and at a designated time in the cell cycle. Replication is semiconservative, each strand acting as template for a new daughter strand. It is carried out in three identifiable phases: initiation, elongation, and termination. The process starts at a single origin in bacteria and usually proceeds bidirectionally.
- ▶ DNA is synthesized in the 5'→3' direction by DNA polymerases. At the replication fork, the leading strand is synthesized continuously in the same direction as replication fork movement; the lagging strand is synthesized discontinuously as Okazaki fragments, which are subsequently ligated.
- ▶ The fidelity of DNA replication is maintained by (1) base selection by the polymerase, (2) a 3'→5' proofreading exonuclease activity that is part of most DNA polymerases, and (3) specific repair systems for mismatches left behind after replication.
- ▶ Most cells have several DNA polymerases. In *E. coli*, DNA polymerase III is the primary replication enzyme. DNA polymerase I is responsible for special functions during replication, recombination, and repair.
- ▶ The separate initiation, elongation, and termination phases of DNA replication involve an array of enzymes and protein factors, many belonging to the AAA+ ATPase family.

- ▶ The major replicative DNA polymerase in eukaryotes is DNA polymerase δ . DNA polymerase α functions to synthesize primers. DNA polymerase ϵ functions in DNA repair.

25.2 DNA Repair

Most cells have only one or two sets of genomic DNA. Damaged proteins and RNA molecules can be quickly replaced by using information encoded in the DNA, but DNA molecules themselves are irreplaceable. Maintaining the integrity of the information in DNA is a cellular imperative, supported by an elaborate set of DNA repair systems. DNA can become damaged by a variety of processes, some spontaneous, others catalyzed by environmental agents (Chapter 8). Replication itself can very occasionally damage the information content in DNA when errors introduce mismatched base pairs (such as G paired with T).

The chemistry of DNA damage is diverse and complex. The cellular response to this damage includes a wide range of enzymatic systems that catalyze some of the most interesting chemical transformations in DNA metabolism. We first examine the effects of alterations in DNA sequence and then consider specific repair systems.

Mutations Are Linked to Cancer



The best way to illustrate the importance of DNA repair is to consider the effects of *unrepaired* DNA damage (a lesion). The most serious outcome is a change in the base sequence of the DNA, which, if replicated and transmitted to future generations of cells, becomes permanent. A permanent change in the nucleotide sequence of DNA is called a **mutation**. Mutations can involve the replacement of one base pair with another (substitution mutation) or the addition or deletion of one or more base pairs (insertion or deletion mutations). If the mutation affects nonessential DNA or if it has a negligible effect on the function of a gene, it is known as a **silent mutation**. Rarely, a mutation confers some biological advantage. Most nonsilent mutations, however, are neutral or deleterious.

In mammals there is a strong correlation between the accumulation of mutations and cancer. A simple test developed by Bruce Ames measures the potential of a given chemical compound to promote certain easily detected mutations in a specialized bacterial strain (**Fig. 25–20**). Few of the chemicals that we encounter in daily life score as mutagens in this test. However, of the compounds known to be carcinogenic from extensive animal trials, more than 90% are also found to be mutagenic in the Ames test. Because of this strong correlation between mutagenesis and carcinogenesis, the Ames test for bacterial mutagens is widely used as a

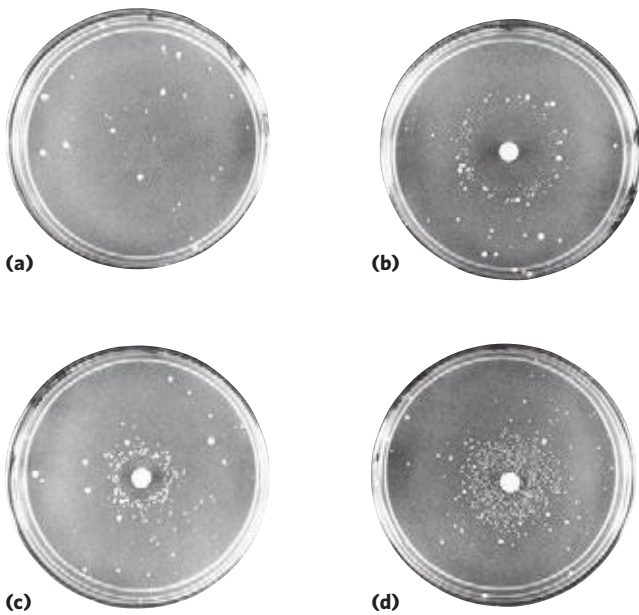


FIGURE 25-20 Ames test for carcinogens, based on their mutagenicity.

A strain of *Salmonella typhimurium* having a mutation that inactivates an enzyme of the histidine biosynthetic pathway is plated on a histidine-free medium. Few cells grow. (a) The few small colonies of *S. typhimurium* that do grow on a histidine-free medium carry spontaneous back-mutations that permit the histidine biosynthetic pathway to operate. Three identical nutrient plates (b), (c), and (d) have been inoculated with an equal number of cells. Each plate then receives a disk of filter paper containing progressively lower concentrations of a mutagen. The mutagen greatly increases the rate of back-mutation and hence the number of colonies. The clear areas around the filter paper indicate where the concentration of mutagen is so high that it is lethal to the cells. As the mutagen diffuses away from the filter paper, it is diluted to sublethal concentrations that promote back-mutation. Mutagens can be compared on the basis of their effect on mutation rate. Because many compounds undergo a variety of chemical transformations after entering cells, compounds are sometimes tested for mutagenicity after first incubating them with a liver extract. Some substances have been found to be mutagenic only after this treatment.

rapid and inexpensive screen for potential human carcinogens.

The genomic DNA in a typical mammalian cell accumulates many thousands of lesions during a 24-hour period. However, as a result of DNA repair, fewer than 1 in 1,000 become a mutation. DNA is a relatively stable molecule, but in the absence of repair systems, the cumulative effect of many infrequent but damaging reactions would make life impossible. ■

All Cells Have Multiple DNA Repair Systems

The number and diversity of repair systems reflect both the importance of DNA repair to cell survival and the diverse sources of DNA damage (Table 25-5). Some common types of lesions, such as pyrimidine dimers (see Fig. 8-31), can be repaired by several distinct

TABLE 25-5 Types of DNA Repair Systems in *E. coli*

Enzymes/proteins	Type of damage		
Mismatch repair			
Dam methylase MutH, MutL, MutS proteins DNA helicase II SSB DNA polymerase III Exonuclease I Exonuclease VII RecJ nuclease Exonuclease X DNA ligase	} Mismatches		
Base-excision repair			
DNA glycosylases AP endonucleases DNA polymerase I DNA ligase		} Abnormal bases (uracil, hypoxanthine, xanthine); alkylated bases; in some other organisms, pyrimidine dimers	
Nucleotide-excision repair			
ABC excinuclease DNA polymerase I DNA ligase			} DNA lesions that cause large structural changes (e.g., pyrimidine dimers)
Direct repair			
DNA photolyases		Pyrimidine dimers	
<i>O</i> ⁶ -Methylguanine-DNA methyltransferase		<i>O</i> ⁶ -Methylguanine	
AlkB protein		1-Methylguanine, 3-methylcytosine	

systems. Many DNA repair processes also seem to be extraordinarily inefficient energetically—an exception to the pattern observed in the vast majority of metabolic pathways, where every ATP is generally accounted for and used optimally. When the integrity of the genetic information is at stake, the amount of chemical energy invested in a repair process seems almost irrelevant.

DNA repair is possible largely because the DNA molecule consists of two complementary strands. DNA damage in one strand can be removed and accurately replaced by using the undamaged complementary strand as a template. We consider here the principal types of repair systems, beginning with those that repair the rare nucleotide mismatches that are left behind by replication.

Mismatch Repair Correction of the rare mismatches left after replication in *E. coli* improves the overall fidelity

of replication by an additional factor of 10^2 to 10^3 . The mismatches are nearly always corrected to reflect the information in the old (template) strand, so the repair system must somehow discriminate between the template and the newly synthesized strand. The cell accomplishes this by tagging the template DNA with methyl groups to distinguish it from newly synthesized strands. The mismatch repair system of *E. coli* includes at least 12 protein components (Table 25–5) that function either in strand discrimination or in the repair process itself.

The strand discrimination mechanism has not been worked out for most bacteria or eukaryotes, but is well understood for *E. coli* and some closely related bacterial species. In these bacteria, strand discrimination is based on the action of Dam methylase, which, as you will recall, methylates DNA at the N^6 position of all adenines within (5')GATC sequences. Immediately after passage of the replication fork, there is a short period (a few seconds or minutes) during which the template strand is methylated but the newly synthesized strand is not (Fig. 25–21). The transient unmethylated state of GATC sequences in the newly synthesized strand permits the new strand to be distinguished from the template strand. Replication mismatches in the vicinity of a hemimethylated GATC sequence are then repaired according to the information in the methylated parent (template) strand. Tests in vitro show that if both strands are methylated at a GATC sequence, few mismatches are repaired; if neither strand is methylated, repair occurs but does not favor either strand. The cell's methyl-directed mismatch repair system efficiently repairs mismatches up to 1,000 bp from a hemimethylated GATC sequence.

How is the mismatch correction process directed by relatively distant GATC sequences? A mechanism is illustrated in Figure 25–22. MutL protein forms a complex with MutS protein, and the complex binds to all mismatched base pairs (except C–C). MutH protein binds to MutL and to GATC sequences encountered by the MutL–MutS complex. DNA on both sides of the mismatch is threaded through the MutL–MutS complex, creating a DNA loop; simultaneous movement of both legs of the loop through the complex is equivalent to the complex moving in both directions at once along the DNA. MutH has a site-specific endonuclease activity that is inactive until the complex encounters a hemimethylated GATC sequence. At this site, MutH

catalyzes cleavage of the unmethylated strand on the 5' side of the G in GATC, which marks the strand for repair. Further steps in the pathway depend on where

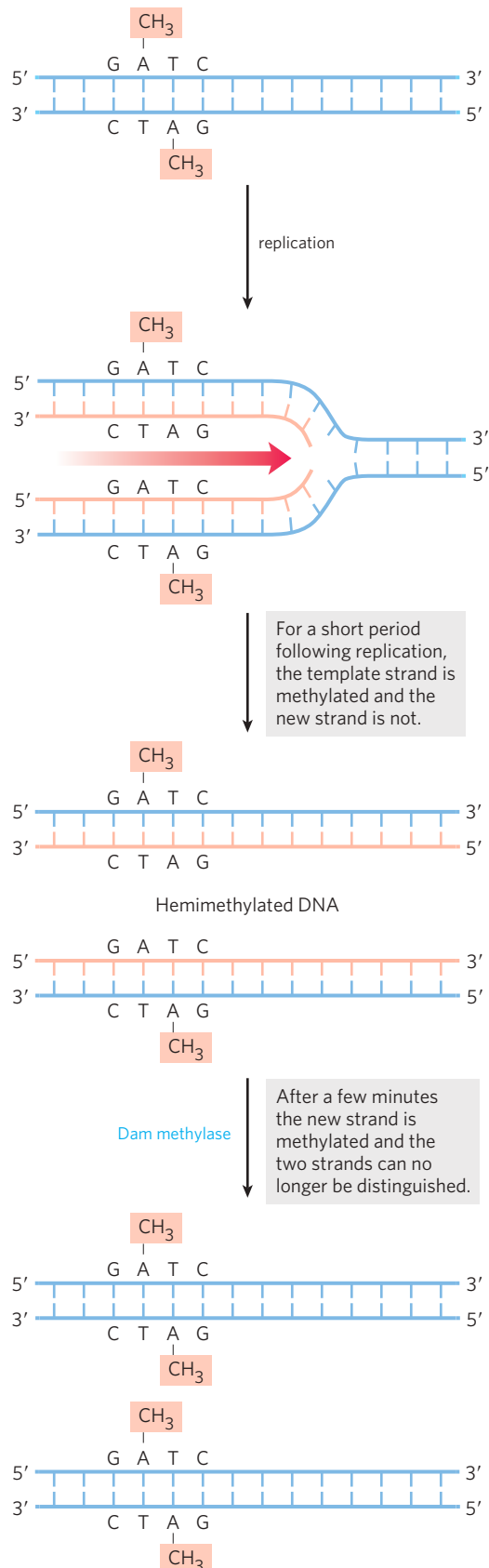


FIGURE 25–21 Methylation and mismatch repair. Methylation of DNA strands can serve to distinguish parent (template) strands from newly synthesized strands in *E. coli* DNA, a function that is critical to mismatch repair (see Fig. 25–22). The methylation occurs at the N^6 of adenines in (5')GATC sequences. This sequence is a palindrome (see Fig. 8–18), present in opposite orientations on the two strands.

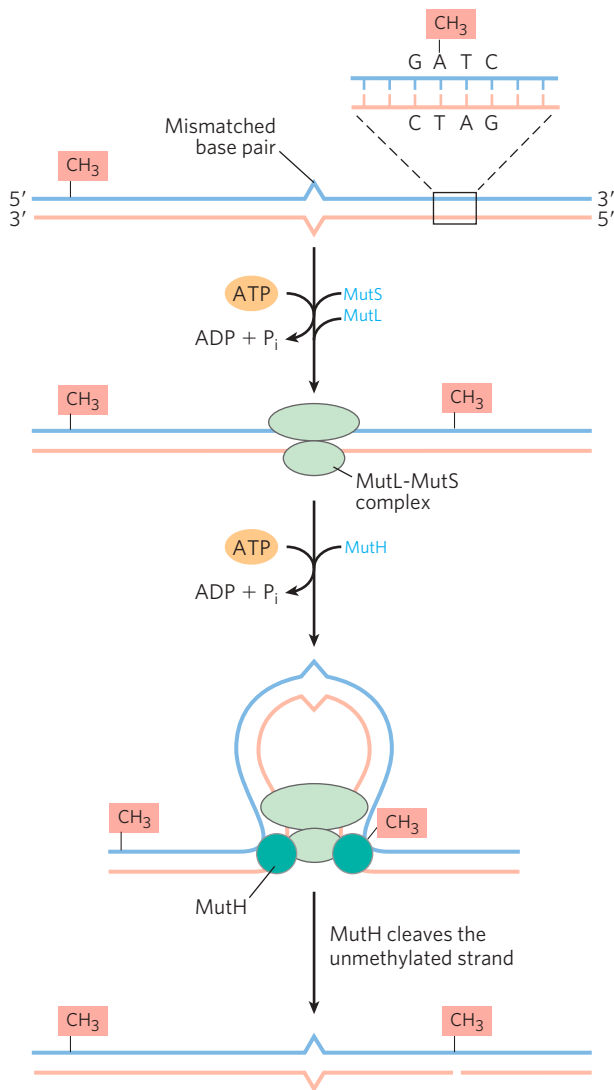


FIGURE 25-22 A model for the early steps of methyl-directed mismatch repair. The proteins involved in this process in *E. coli* have been purified (see Table 25-5). Recognition of the sequence (5')GATC and of the mismatch are specialized functions of the MutH and MutS proteins, respectively. The MutL protein forms a complex with MutS at the mismatch. DNA is threaded through this complex such that the complex moves simultaneously in both directions along the DNA until it encounters a MutH protein bound at a hemimethylated GATC sequence. MutH cleaves the unmethylated strand on the 5' side of the G in this sequence. A complex consisting of DNA helicase II and one of several exonucleases then degrades the unmethylated DNA strand from that point toward the mismatch (see Fig. 25-23).

the mismatch is located relative to this cleavage site (**Fig. 25-23**).

When the mismatch is on the 5' side of the cleavage site (**Fig. 25-23**, right side), the unmethylated strand is unwound and degraded in the 3'→5' direction from the cleavage site through the mismatch, and this segment is replaced with new DNA. This process requires the combined action of DNA helicase II (also called UvrD helicase), SSB, exonuclease I or exonuclease X (both of which degrade strands of DNA in the 3'→5' direction),

DNA polymerase III, and DNA ligase. The pathway for repair of mismatches on the 3' side of the cleavage site is similar (**Fig. 25-23**, left), except that the exonuclease is either exonuclease VII (which degrades single-stranded DNA in the 5'→3' or 3'→5' direction) or RecJ nuclease (which degrades single-stranded DNA in the 5'→3' direction).

Mismatch repair is a particularly expensive process for *E. coli* in terms of energy expended. The mismatch may be 1,000 bp or more from the GATC sequence. The degradation and replacement of a strand segment of this length require an enormous investment in activated deoxynucleotide precursors to repair a *single* mismatched base. This again underscores the importance to the cell of genomic integrity.

All eukaryotic cells have several proteins structurally and functionally analogous to the bacterial MutS and MutL (but not MutH) proteins. Alterations in human genes encoding proteins of this type produce some of the most common inherited cancer-susceptibility syndromes (see Box 25-1, p. 1037), further demonstrating the value to the organism of DNA repair systems. The main MutS homologs in most eukaryotes, from yeast to humans, are MSH2 (*MutS* homolog), MSH3, and MSH6. Heterodimers of MSH2 and MSH6 generally bind to single base-pair mismatches, and bind less well to slightly longer mispaired loops. In many organisms the longer mismatches (2 to 6 bp) may be bound instead by a heterodimer of MSH2 and MSH3, or are bound by both types of heterodimers in tandem. Homologs of MutL, predominantly a heterodimer of MLH1 (*MutL* homolog) and PMS1 (*post-meiotic segregation*), bind to and stabilize the MSH complexes. Many details of the subsequent events in eukaryotic mismatch repair remain to be worked out. In particular, we do not know the mechanism by which newly synthesized DNA strands are identified, although research has revealed that this strand identification does not involve GATC sequences.

Base-Excision Repair Every cell has a class of enzymes called **DNA glycosylases** that recognize particularly common DNA lesions (such as the products of cytosine and adenine deamination; see Fig. 8-30a) and remove the affected base by cleaving the *N*-glycosyl bond. This cleavage creates an apurinic or apyrimidinic site in the DNA, commonly referred to as an **AP site** or **abasic site**. Each DNA glycosylase is generally specific for one type of lesion.

Uracil DNA glycosylases, for example, found in most cells, specifically remove from DNA the uracil that results from spontaneous deamination of cytosine. Mutant cells that lack this enzyme have a high rate of G≡C to A=T mutations. This glycosylase does not remove uracil residues from RNA or thymine residues from DNA. The capacity to distinguish thymine from uracil, the product of cytosine deamination—necessary for the selective repair of the latter—may be one reason

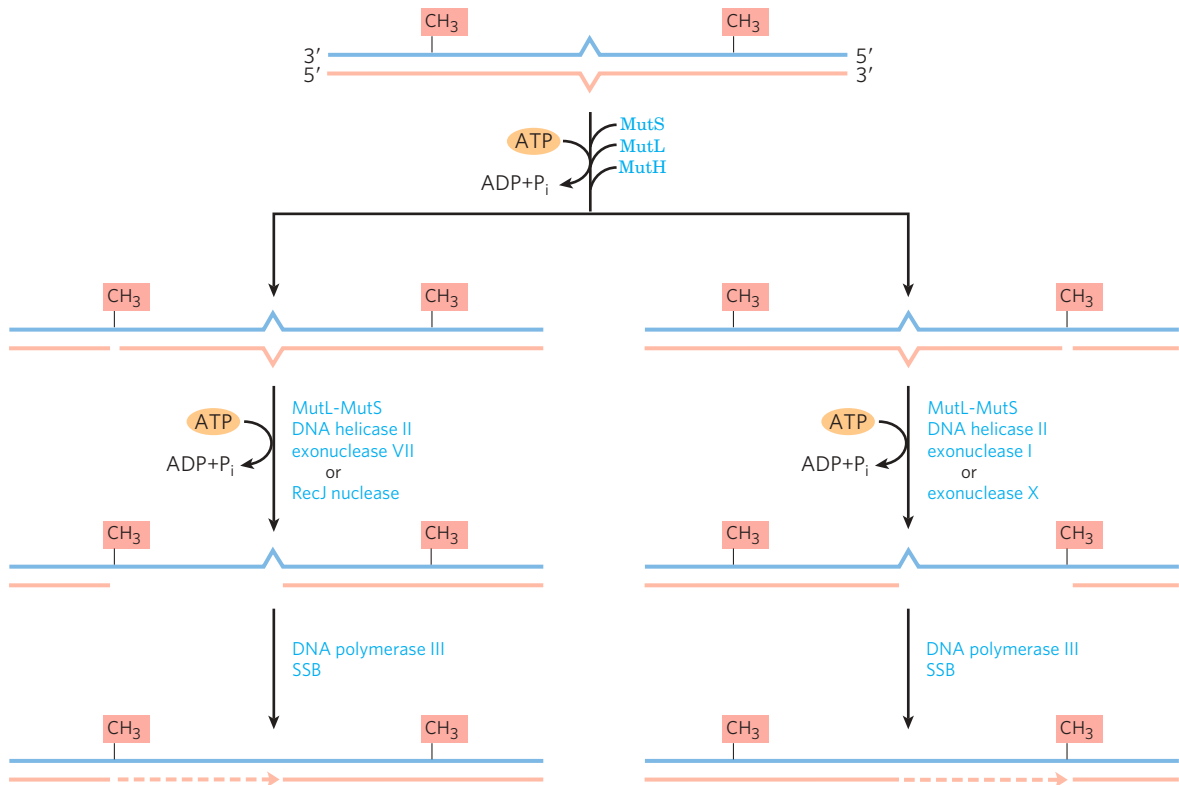


FIGURE 25-23 Completing methyl-directed mismatch repair. The combined action of DNA helicase II, SSB, and one of four different exonucleases removes a segment of the new strand between the MutH cleavage site and a point just beyond the mismatch. The exonuclease

that is used depends on the location of the cleavage site relative to the mismatch, as shown by the alternative pathways here. The resulting gap is filled in (dashed line) by DNA polymerase III, and the nick is sealed by DNA ligase (not shown).

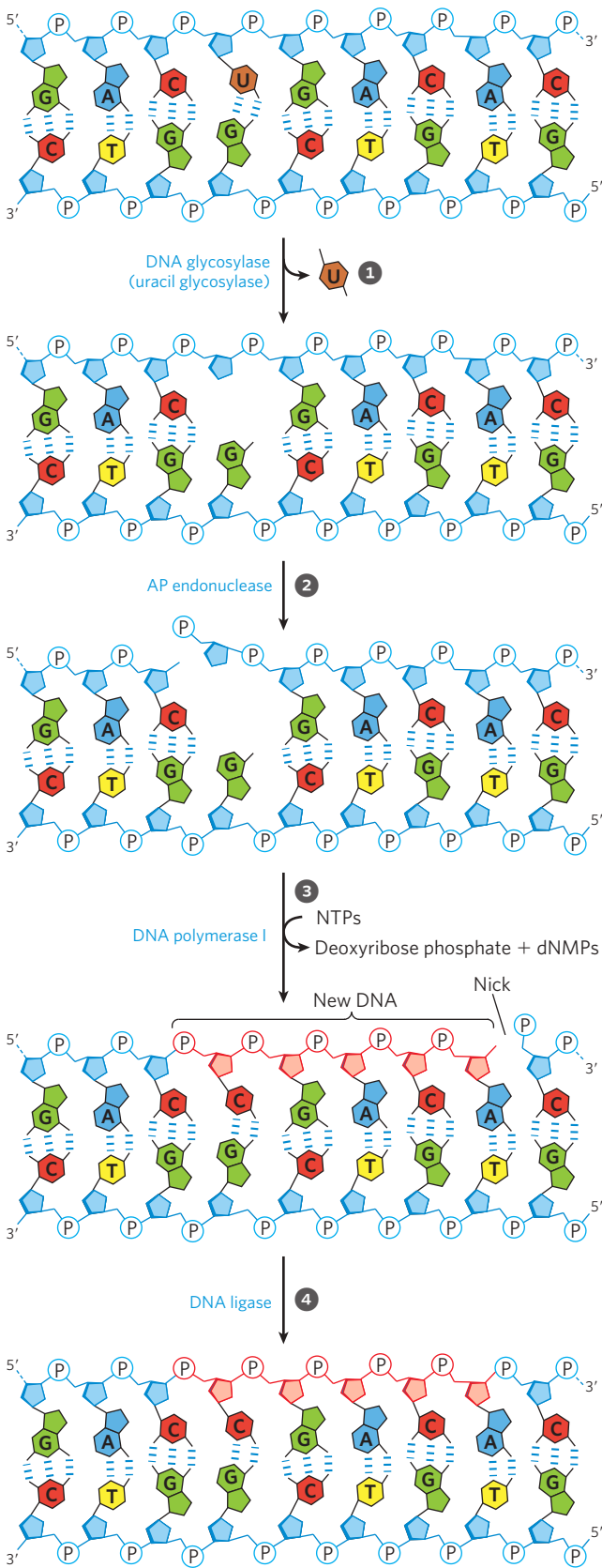
why DNA evolved to contain thymine instead of uracil (p. 299).

Most bacteria have just one type of uracil DNA glycosylase, whereas humans have at least four types, with different specificities—an indicator of the importance of uracil removal from DNA. The most abundant human uracil glycosylase, UNG, is associated with the human replisome, where it eliminates the occasional U residue inserted in place of a T during replication. The deamination of C residues is 100-fold faster in single-stranded DNA than in double-stranded DNA, and humans have the enzyme hSMUG1, which removes any U residues that occur in single-stranded DNA during replication or transcription. Two other human DNA glycosylases, TDG and MBD4, remove either U or T residues paired with G, generated by deamination of cytosine or 5-methylcytosine, respectively.

Other DNA glycosylases recognize and remove a variety of damaged bases, including formamidopyrimidine and 8-hydroxyguanine (both arising from purine oxidation), hypoxanthine (arising from adenine deamination), and alkylated bases such as 3-methyladenine and 7-methylguanine. Glycosylases that recognize other lesions, including pyrimidine dimers, have also been identified in some classes of organisms. Remember that AP sites also arise from the slow, spontaneous hydrolysis of the *N*-glycosyl bonds in DNA (see Fig. 8–30b).

Once an AP site has been formed by a DNA glycosylase, another type of enzyme must repair it. The repair is *not* made by simply inserting a new base and re-forming the *N*-glycosyl bond. Instead, the deoxyribose 5'-phosphate left behind is removed and replaced with a new nucleotide. This process begins with one of the **AP endonucleases**, enzymes that cut the DNA strand containing the AP site. The position of the incision relative to the AP site (5' or 3' to the site) varies with the type of AP endonuclease. A segment of DNA including the AP site is then removed, DNA polymerase I replaces the DNA, and DNA ligase seals the remaining nick (**Fig. 25-24**). In eukaryotes, nucleotide replacement is carried out by specialized polymerases, as described below.

Nucleotide-Excision Repair DNA lesions that cause large distortions in the helical structure of DNA generally are repaired by the nucleotide-excision system, a repair pathway critical to the survival of all free-living organisms. In nucleotide-excision repair (**Fig. 25-25**), a multisubunit enzyme (excinuclease) hydrolyzes two phosphodiester bonds, one on either side of the distortion caused by the lesion. In *E. coli* and other bacteria, the enzyme system hydrolyzes the fifth phosphodiester bond on the 3' side and the eighth phosphodiester bond on the 5' side to generate a fragment of 12 to 13 nucleotides (depending



on whether the lesion involves one or two bases). In humans and other eukaryotes, the enzyme system hydrolyzes the sixth phosphodiester bond on the 3' side and the twenty-second phosphodiester bond on the 5' side,

FIGURE 25-24 DNA repair by the base-excision repair pathway. ① A DNA glycosylase recognizes a damaged base (in this case, a uracil) and cleaves between the base and deoxyribose in the backbone. ② An AP endonuclease cleaves the phosphodiester backbone near the AP site. ③ DNA polymerase I initiates repair synthesis from the free 3' hydroxyl at the nick, removing (with its 5'→3' exonuclease activity) and replacing a portion of the damaged strand. ④ The nick remaining after DNA polymerase I has dissociated is sealed by DNA ligase.

producing a fragment of 27 to 29 nucleotides. Following the dual incision, the excised oligonucleotides are released from the duplex and the resulting gap is filled—by DNA polymerase I in *E. coli* and DNA polymerase ϵ in humans. DNA ligase seals the nick.

In *E. coli*, the key enzymatic complex is the ABC excinuclease, which has three subunits, UvrA (M_r 104,000), UvrB (M_r 78,000), and UvrC (M_r 68,000). The term “excinuclease” is used to describe the unique capacity of this enzyme complex to catalyze two specific endonucleolytic cleavages, distinguishing this activity from that of standard endonucleases. A complex of the UvrA and UvrB proteins (A_2B) scans the DNA and binds to the site of a lesion. The UvrA dimer then dissociates, leaving a tight UvrB-DNA complex. UvrC protein then binds to UvrB, and UvrB makes an incision at the fifth phosphodiester bond on the 3' side of the lesion. This is followed by a UvrC-mediated incision at the eighth phosphodiester bond on the 5' side. The resulting 12 to 13 nucleotide fragment is removed by UvrD helicase. The short gap thus created is filled in by DNA polymerase I and DNA ligase. This pathway (Fig. 25-25, left) is a primary repair route for many types of lesions, including cyclobutane pyrimidine dimers, 6-4 photoproducts (see Fig. 8-31), and several other types of base adducts including benzo[*a*]pyrene-guanine, which is formed in DNA by exposure to cigarette smoke. The nucleolytic activity of the ABC excinuclease is novel in the sense that two cuts are made in the DNA.

The mechanism of eukaryotic excinucleases is quite similar to that of the bacterial enzyme, although 16 polypeptides with no similarity to the *E. coli* excinuclease subunits are required for the dual incision. As described in Chapter 26, some of the nucleotide-excision repair and base-excision repair in eukaryotes is closely tied to transcription. Genetic deficiencies in nucleotide-excision repair in humans give rise to a variety of serious diseases (see Box 25-1).

Direct Repair Several types of damage are repaired without removing a base or nucleotide. The best-characterized example is direct photoreactivation of cyclobutane pyrimidine dimers, a reaction promoted by **DNA photolyases**. Pyrimidine dimers result from a UV-induced reaction, and photolyases use energy derived from absorbed light to reverse the damage (**Fig. 25-26**). Photolyases generally contain two cofactors that serve as light-absorbing agents, or chromophores.

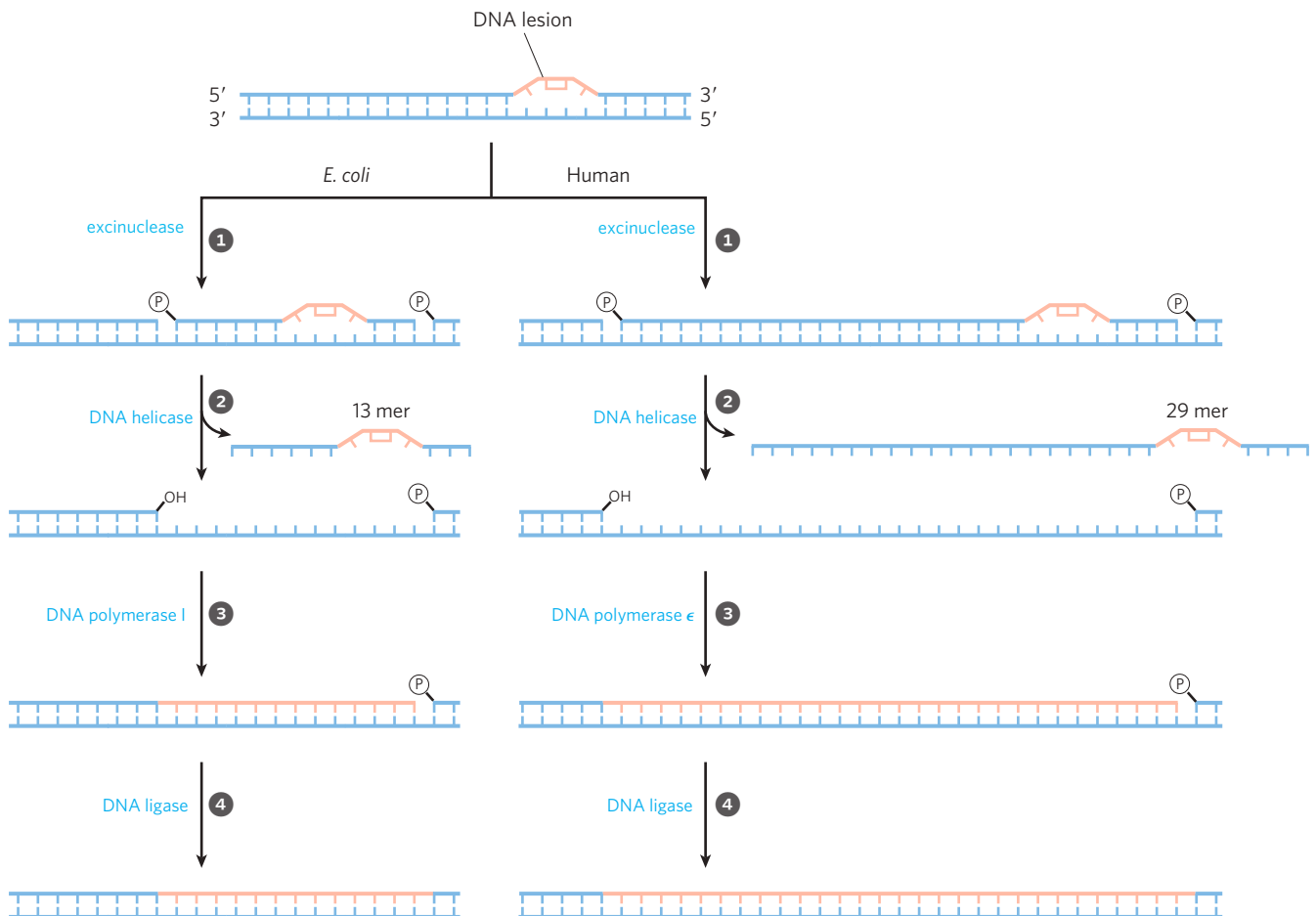


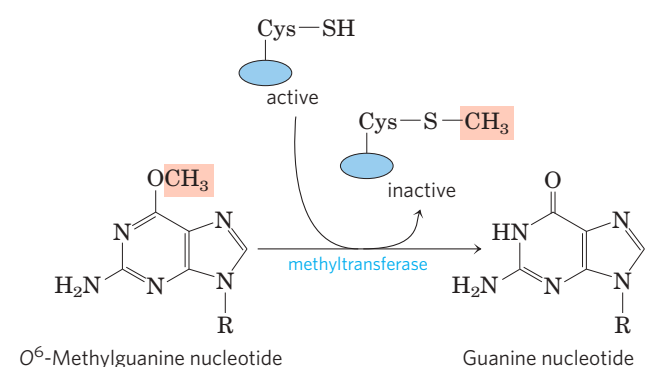
FIGURE 25-25 Nucleotide-excision repair in *E. coli* and humans.

The general pathway of nucleotide-excision repair is similar in all organisms. **1** An excinuclease binds to DNA at the site of a bulky lesion and cleaves the damaged DNA strand on either side of the lesion. **2** The

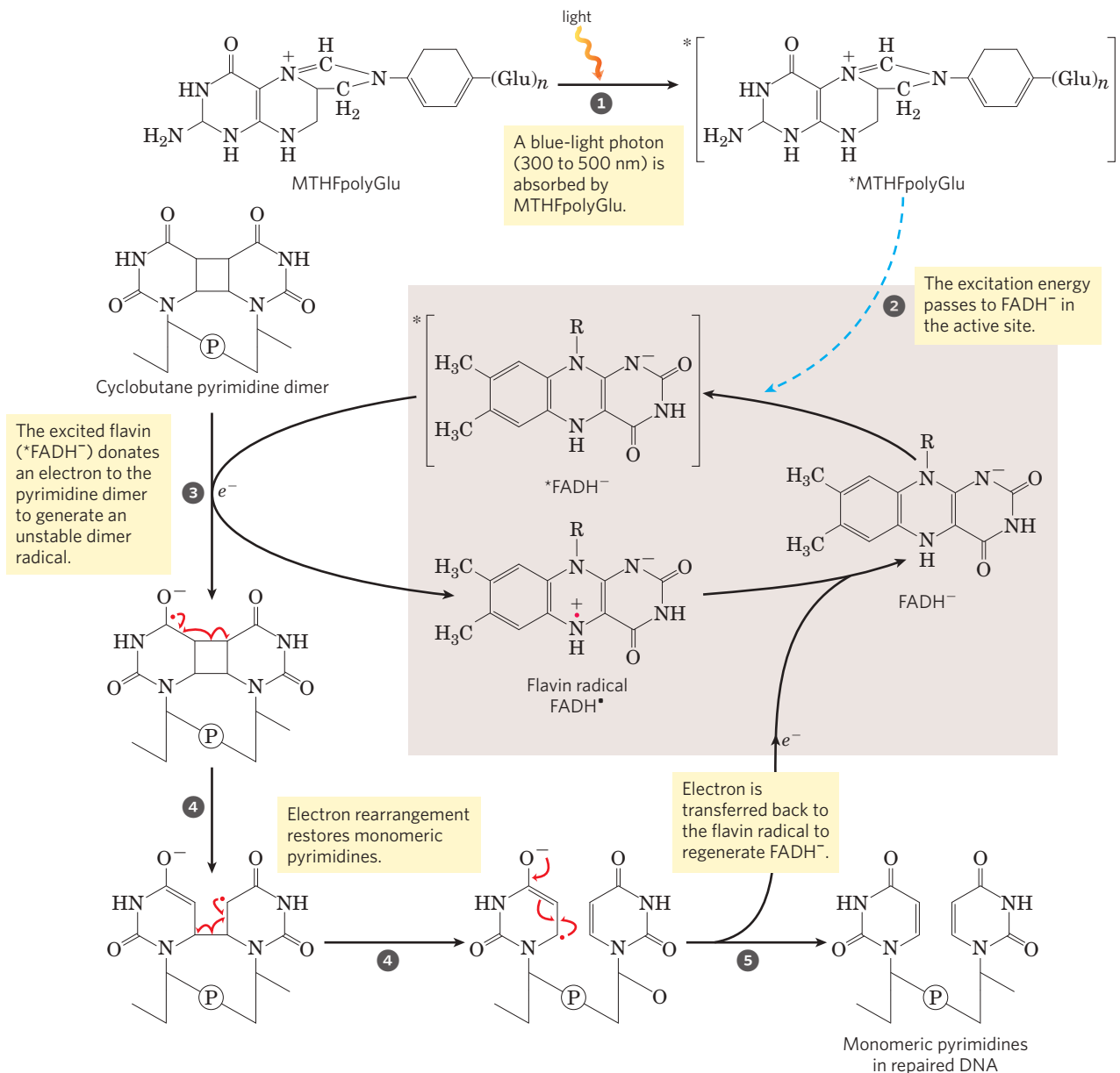
DNA segment—of 13 nucleotides (13 mer) or 29 nucleotides (29 mer)—is removed with the aid of a helicase. **3** The gap is filled in by DNA polymerase, and **4** the remaining nick is sealed with DNA ligase.

One of the chromophores is always FADH₂. In *E. coli* and yeast, the other chromophore is a folate. The reaction mechanism entails the generation of free radicals. DNA photolyases are not present in the cells of placental mammals (which include humans).

Additional examples can be seen in the repair of nucleotides with alkylation damage. The modified nucleotide *O*⁶-methylguanine forms in the presence of alkylating agents and is a common and highly mutagenic lesion (p. 302). It tends to pair with thymine rather than cytosine during replication, and therefore causes G≡C to A=T mutations (Fig. 25-27). Direct repair of *O*⁶-methylguanine is carried out by *O*⁶-methylguanine-DNA methyltransferase, a protein that catalyzes transfer of the methyl group of *O*⁶-methylguanine to one of its own Cys residues. This methyltransferase is not strictly an enzyme, because a single methyl transfer event permanently methylates the protein, making it inactive in this pathway. The consumption of an entire protein molecule to correct a single damaged base is another vivid illustration of the priority given to maintaining the integrity of cellular DNA.



A very different but equally direct mechanism is used to repair 1-methyladenine and 3-methylcytosine. The amino groups of A and C residues are sometimes methylated when the DNA is single-stranded, and the methylation directly affects proper base pairing. In *E. coli*, oxidative demethylation of these alkylated nucleotides is mediated by the AlkB protein, a member of the α -ketoglutarate-Fe²⁺-dependent dioxygenase superfamily (Fig. 25-28). (See Box 4-3 for a description of another member of this enzyme family.)



MECHANISM FIGURE 25-26 Repair of pyrimidine dimers with photolyase. Energy derived from absorbed light is used to reverse the photo-reaction that caused the lesion. The two chromophores in *E. coli* photolyase (M_r 54,000), N^5,N^{10} -methylene-tetrahydrofolylpolyglutamate (MTHFpolyGlu) and FADH^- , perform complementary functions. MTHF-

polyGlu functions as a photoantenna to absorb blue-light photons. The excitation energy passes to FADH^- , and the excited flavin ($*\text{FADH}^-$) donates an electron to the pyrimidine dimer, leading to the rearrangement as shown.

The Interaction of Replication Forks with DNA Damage Can Lead to Error-Prone Translesion DNA Synthesis

The repair pathways considered to this point generally work only for lesions in double-stranded DNA, the undamaged strand providing the correct genetic information to restore the damaged strand to its original state. However, in certain types of lesions, such as double-strand breaks, double-strand cross-links, or lesions in a single-stranded DNA, the complementary strand is itself damaged or is absent. Double-strand breaks and lesions in single-stranded DNA most often

arise when a replication fork encounters an unrepaired DNA lesion (**Fig. 25-29**). Such lesions and DNA cross-links can also result from ionizing radiation and oxidative reactions.

At a stalled bacterial replication fork, there are two avenues for repair. In the absence of a second strand, the information required for accurate repair must come from a separate, homologous chromosome. The repair system thus involves homologous genetic recombination. This recombinational DNA repair is considered in detail in Section 25.3. Under some conditions, a second repair pathway, **error-prone translesion DNA**

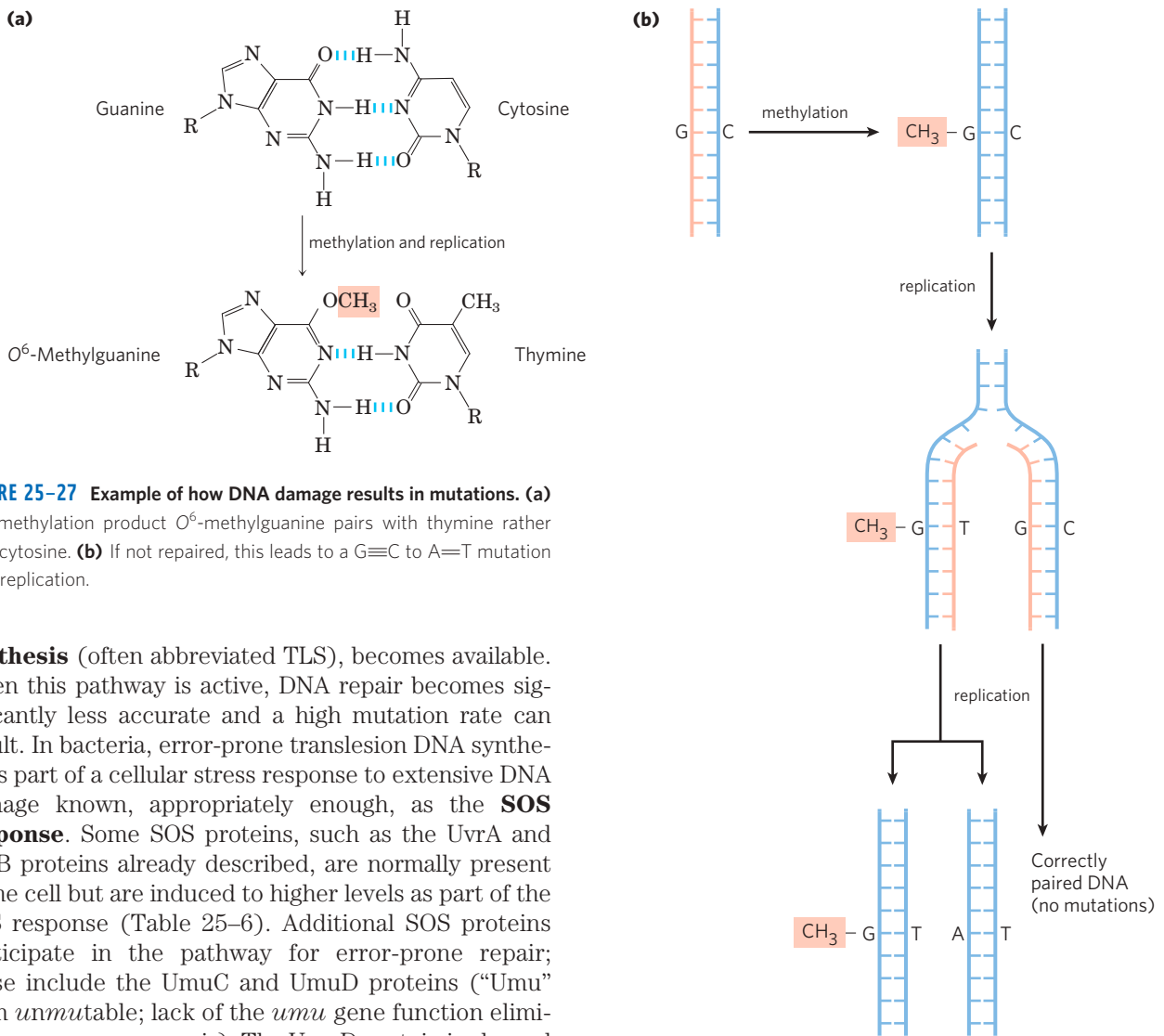


FIGURE 25-27 Example of how DNA damage results in mutations. **(a)**

The methylation product O^6 -methylguanine pairs with thymine rather than cytosine. **(b)** If not repaired, this leads to a $G \equiv C$ to $A = T$ mutation after replication.

synthesis (often abbreviated TLS), becomes available. When this pathway is active, DNA repair becomes significantly less accurate and a high mutation rate can result. In bacteria, error-prone translesion DNA synthesis is part of a cellular stress response to extensive DNA damage known, appropriately enough, as the **SOS response**. Some SOS proteins, such as the UvrA and UvrB proteins already described, are normally present in the cell but are induced to higher levels as part of the SOS response (Table 25-6). Additional SOS proteins participate in the pathway for error-prone repair; these include the UmuC and UmuD proteins (“Umu” from *unmutable*; lack of the *umu* gene function eliminates error-prone repair). The UmuD protein is cleaved

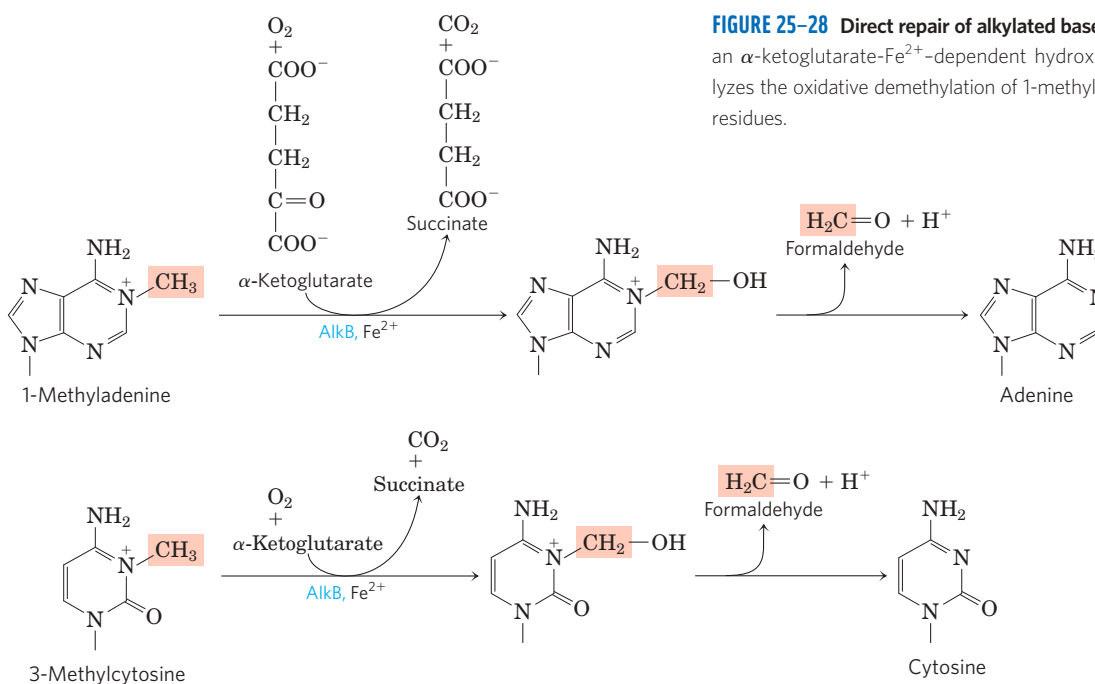


FIGURE 25-28 Direct repair of alkylated bases by AlkB. The AlkB protein is an α -ketoglutarate- Fe^{2+} -dependent hydroxylase (see Box 4-3). It catalyzes the oxidative demethylation of 1-methyladenine and 3-methylcytosine residues.

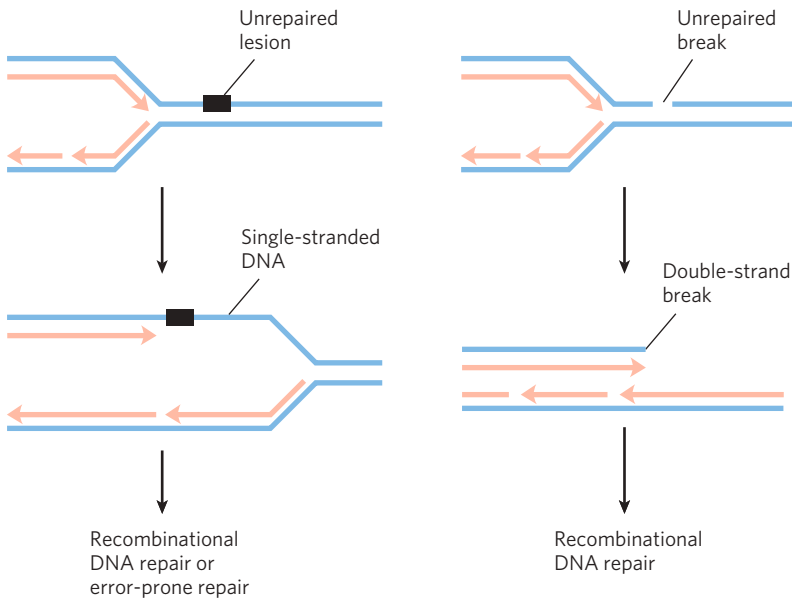


FIGURE 25-29 DNA damage and its effect on DNA replication. If the replication fork encounters an unrepaired lesion or strand break, replication generally halts and the fork may collapse. A lesion is left behind in an unreplicated, single-stranded segment of the DNA (left); a strand break becomes a double-strand break (right). In each case the damage to one strand cannot be repaired by mechanisms described earlier in this chapter, because the complementary strand required for direct accurate repair is damaged or absent. There are two possible avenues for repair: recombinational DNA repair (one pathway is described in Fig. 25-30) or, when lesions are unusually numerous, error-prone repair. The latter mechanism involves a novel DNA polymerase (DNA polymerase V, encoded by the *umuC* and *umuD* genes) that can replicate, albeit inaccurately, over many types of lesions. The repair mechanism is “error-prone” because mutations often result.

in an SOS-regulated process to a shorter form called UmuD', which forms a complex with UmuC and a protein called RecA (described in Section 25.3) to create a specialized DNA polymerase, DNA polymerase V (UmuD'₂UmuCRecA), that can replicate past many of the DNA lesions that would normally block replication. Proper base pairing is often impossible at

the site of such a lesion, so this translesion replication is error-prone.

Given the emphasis on the importance of genomic integrity throughout this chapter, the existence of a system that increases the rate of mutation may seem incongruous. However, we can think of this system as a desperation strategy. The *umuC* and *umuD* genes are fully

TABLE 25-6 Genes Induced as Part of the SOS response in *E. coli*

Gene name	Protein encoded and/or role in DNA repair
Genes of known function	
<i>polB</i> (<i>dinA</i>)	Encodes polymerization subunit of DNA polymerase II, required for replication restart in recombinational DNA repair
<i>uvrA</i> } <i>uvrB</i> }	Encode ABC excinuclease subunits UvrA and UvrB
<i>umuC</i> } <i>umuD</i> }	Encode core and polymerization subunits of DNA polymerase V
<i>sulA</i>	Encodes protein that inhibits cell division, possibly to allow time for DNA repair
<i>recA</i>	Encodes RecA protein, required for error-prone repair and recombinational repair
<i>dinB</i>	Encodes DNA polymerase IV
<i>ssb</i>	Encodes single-stranded DNA-binding protein (SSB)
<i>himA</i>	Encodes subunit of integration host factor (IHF), involved in site-specific recombination, replication, transposition, regulation of gene expression
Genes involved in DNA metabolism, but role in DNA repair unknown	
<i>uvrD</i>	Encodes DNA helicase II (DNA-unwinding protein)
<i>recN</i>	Required for recombinational repair
Genes of unknown function	
<i>dinD</i>	
<i>dinF</i>	

Note: Some of these genes and their functions are further discussed in Chapter 28.

induced only late in the SOS response, and they are not activated for translesion synthesis initiated by UmuD cleavage unless the levels of DNA damage are particularly high and all replication forks are blocked. The mutations resulting from DNA polymerase V-mediated replication kill some cells and create deleterious mutations in others, but this is the biological price a species pays to overcome an otherwise insurmountable barrier to replication, as it permits at least a few mutant daughter cells to survive.

Yet another DNA polymerase, DNA polymerase IV, is also induced during the SOS response. Replication by DNA polymerase IV, a product of the *dinB* gene, is also highly error-prone. The bacterial DNA polymerases IV and V are part of a family of TLS polymerases found in all organisms. These enzymes lack a proofreading exonuclease and possess a more open active site that accommodates damaged template nucleotides. The fidelity of base selection during replication can be reduced by a factor of 10^2 , lowering overall replication fidelity to one error in $\sim 1,000$ nucleotides.

Mammals have many low-fidelity DNA polymerases of the TLS polymerase family. However, the presence of these enzymes does not necessarily translate into an unacceptable mutational burden, because most of these enzymes also have specialized functions in DNA repair. DNA polymerase η (eta), for example,

found in all eukaryotes, promotes translesion synthesis primarily across cyclobutane T–T dimers. Few mutations result in this case, because the enzyme preferentially inserts two A residues across from the linked T residues. Several other low-fidelity polymerases, including DNA polymerases β , ι (iota), and λ , have specialized roles in eukaryotic base-excision repair. Each of these enzymes has a 5'-deoxyribose phosphate lyase activity in addition to its polymerase activity. After base removal by a glycosylase and backbone cleavage by an AP endonuclease, these polymerases remove the abasic site (a 5'-deoxyribose phosphate) and fill in the very short gap. The frequency of mutation due to DNA polymerase η activity is minimized by the very short lengths (often one nucleotide) of DNA synthesized.

What emerges from research into cellular DNA repair systems is a picture of a DNA metabolism that maintains genomic integrity with multiple and often redundant systems. In the human genome, more than 130 genes encode proteins dedicated to the repair of DNA. In many cases, the loss of function of one of these proteins results in genomic instability and an increased occurrence of oncogenesis (Box 25–1). These repair systems are often integrated with the DNA replication systems and are complemented by recombination systems, which we turn to next.

BOX 25–1 MEDICINE DNA Repair and Cancer

Human cancers develop when genes that regulate normal cell division (oncogenes and tumor suppressor genes; Chapter 12) fail to function, are activated at the wrong time, or are altered. As a consequence, cells may grow out of control and form a tumor. The genes controlling cell division can be damaged by spontaneous mutation or overridden by the invasion of a tumor virus (Chapter 26). Not surprisingly, alterations in DNA repair genes that result in an increased rate of mutation can greatly increase an individual's susceptibility to cancer. Defects in the genes encoding the proteins involved in nucleotide-excision repair, mismatch repair, recombinational repair, and error-prone translesion DNA synthesis have all been linked to human cancers. Clearly, DNA repair can be a matter of life and death.

Nucleotide-excision repair requires a larger number of proteins in humans than in bacteria, although the overall pathways are very similar. Genetic defects that inactivate nucleotide-excision repair have been associated with several genetic diseases, the best-studied of which is xeroderma pigmentosum (XP). Because nucleotide-excision repair is the sole repair pathway for pyrimidine dimers in humans, people with XP are extremely sensitive to light and readily develop sunlight-induced skin cancers. Most people

with XP also have neurological abnormalities, presumably because of their inability to repair certain lesions caused by the high rate of oxidative metabolism in neurons. Defects in the genes encoding any of at least seven different protein components of the nucleotide-excision repair system can result in XP, giving rise to seven different genetic groups denoted XPA to XPG. Several of these proteins (notably those defective in XPB, XPD, and XPG) also play roles in transcription-coupled base-excision repair of oxidative lesions, described in Chapter 26.

Most microorganisms have redundant pathways for the repair of cyclobutane pyrimidine dimers—making use of DNA photolyases and sometimes base-excision repair as alternatives to nucleotide-excision repair—but humans and other placental mammals do not. This lack of a backup for nucleotide-excision repair for removing pyrimidine dimers has led to speculation that early mammalian evolution involved small, furry, nocturnal animals with little need to repair UV damage. However, mammals do have a pathway for the translesion bypass of cyclobutane pyrimidine dimers, which involves DNA polymerase η . This enzyme preferentially inserts two A residues opposite a T–T pyrimidine dimer, minimizing mutations. People with a genetic condition in which DNA polymerase η

(continued on next page)

BOX 25-1 MEDICINE DNA Repair and Cancer (Continued)

function is missing exhibit an XP-like illness known as XP-variant or XP-V. Clinical manifestations of XP-V are similar to those of the classic XP diseases, although mutation levels are higher in XP-V when cells are exposed to UV light. Apparently, the nucleotide-excision repair system works in concert with DNA polymerase η in normal human cells, repairing and/or bypassing pyrimidine dimers as needed to keep cell growth and DNA replication going. Exposure to UV light introduces a heavy load of pyrimidine dimers, and some must be bypassed by translesion synthesis to keep replication on track. When one system is missing, it is partly compensated for by the other. A loss of DNA polymerase η activity leads to stalled replication forks and bypass of UV lesions by different, more mutagenic, translesion synthesis (TLS) polymerases. As when other DNA repair systems are absent, the resulting increase in mutations often leads to cancer.

One of the most common inherited cancer-susceptibility syndromes is hereditary nonpolyposis colon cancer (HNPCC). This syndrome has been traced to defects in mismatch repair. Human and

other eukaryotic cells have several proteins analogous to the bacterial MutL and MutS proteins (see Fig. 25-22). Defects in at least five different mismatch repair genes can give rise to HNPCC. The most prevalent are defects in the *hMLH1* (human MutL homolog 1) and *hMSH2* (human MutS homolog 2) genes. In individuals with HNPCC, cancer generally develops at an early age, with colon cancers being most common.

Most human breast cancer occurs in women with no known predisposition. However, about 10% of cases are associated with inherited defects in two genes, *BRCA1* and *BRCA2*. Human *BRCA1* and *BRCA2* are large proteins (1,834 and 3,418 amino acid residues, respectively) that interact with a wide range of other proteins involved in transcription, chromosome maintenance, DNA repair, and control of the cell cycle. *BRCA2* has been implicated in the recombinational DNA repair of double-strand breaks. However, the precise molecular function of *BRCA1* and *BRCA2* in these various cellular processes is not yet clear. Women with defects in either the *BRCA1* or *BRCA2* gene have a greater than 80% chance of developing breast cancer.

SUMMARY 25.2 DNA Repair

- ▶ Cells have many systems for DNA repair. Mismatch repair in *E. coli* is directed by transient nonmethylation of (5')GATC sequences on the newly synthesized strand.
- ▶ Base-excision repair systems recognize and repair damage caused by environmental agents (such as radiation and alkylating agents) and spontaneous reactions of nucleotides. Some repair systems recognize and excise only damaged or incorrect bases, leaving an AP (abasic) site in the DNA. This is repaired by excision and replacement of the DNA segment containing the AP site.
- ▶ Nucleotide-excision repair systems recognize and remove a variety of bulky lesions and pyrimidine dimers. They excise a segment of the DNA strand including the lesion, leaving a gap that is filled in by DNA polymerase and ligase activities.
- ▶ Some DNA damage is repaired by direct reversal of the reaction causing the damage: pyrimidine dimers are directly converted to monomeric pyrimidines by a photolyase, and the methyl group of *O*⁶-methylguanine is removed by a methyltransferase.
- ▶ In bacteria, error-prone translesion DNA synthesis, involving TLS DNA polymerases, occurs in response to very heavy DNA damage. In

eukaryotes, similar polymerases have specialized roles in DNA repair that minimize the introduction of mutations.

25.3 DNA Recombination

The rearrangement of genetic information within and among DNA molecules encompasses a variety of processes, collectively placed under the heading of genetic recombination. The practical applications of DNA rearrangements in altering the genomes of increasing numbers of organisms are now being explored (Chapter 9).

Genetic recombination events fall into at least three general classes. **Homologous genetic recombination** (also called general recombination) involves genetic exchanges between any two DNA molecules (or segments of the same molecule) that share an extended region of nearly identical sequence. The actual sequence of bases is irrelevant, as long as it is similar in the two DNAs. In **site-specific recombination** the exchanges occur only at a *particular* DNA sequence. **DNA transposition** is distinct from both other classes in that it usually involves a short segment of DNA with the remarkable capacity to move from one



Barbara McClintock,
1902-1992

location in a chromosome to another. These “jumping genes” were first observed in maize in the 1940s by Barbara McClintock. There is in addition a wide range of unusual genetic rearrangements for which no mechanism or purpose has yet been proposed. Here we focus on the three general classes.

The functions of genetic recombination systems are as varied as their mechanisms. They include roles in specialized DNA repair systems, specialized activities in DNA replication, regulation of expression of certain genes, facilitation of proper chromosome segregation during eukaryotic cell division, maintenance of genetic diversity, and implementation of programmed genetic rearrangements during embryonic development. In most cases, genetic recombination is closely integrated with other processes in DNA metabolism, and this becomes a theme of our discussion.

Bacterial Homologous Recombination Is a DNA Repair Function

In bacteria, homologous genetic recombination is primarily a DNA repair process and in this context (as noted in Section 25.2) is referred to as **recombinational DNA repair**. It is usually directed at the reconstruction of replication forks that have stalled or collapsed at the site of DNA damage. Homologous genetic recombination can also occur during conjugation (mating), when chromosomal DNA is transferred from one bacterial cell (donor) to another (recipient). Recombination during conjugation, although rare in wild bacterial populations, contributes to genetic diversity.

An example of what happens when a replication fork encounters DNA damage is shown in **Figure 25-30**. A common feature of the DNA repair pathways

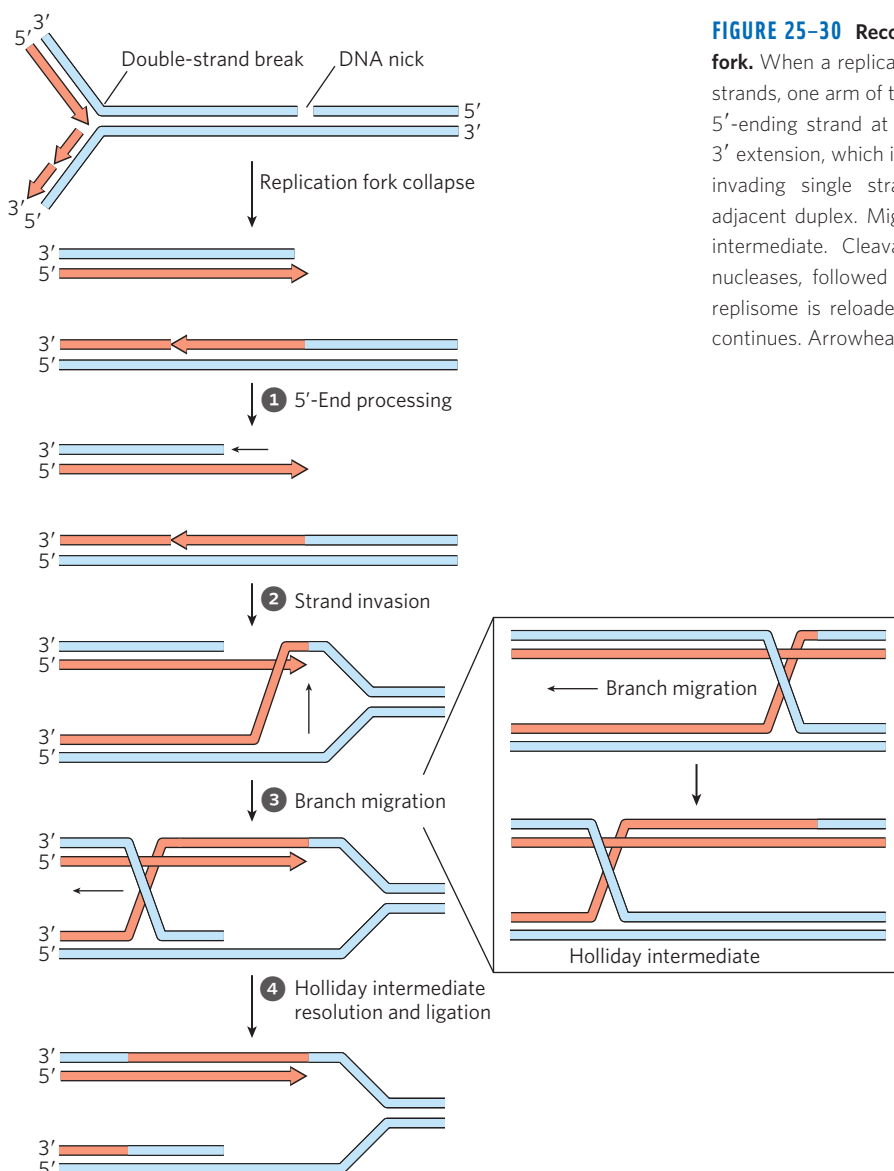


FIGURE 25-30 Recombinational DNA repair at a collapsed replication fork. When a replication fork encounters a break in one of the template strands, one arm of the fork is lost and the replication fork collapses. The 5'-ending strand at the break is degraded to create a single-stranded 3' extension, which is then used in a strand invasion process, pairing the invading single strand with its complementary strand within the adjacent duplex. Migration of the branch (inset) can create a Holliday intermediate. Cleavage of the Holliday intermediate by specialized nucleases, followed by ligation, restores a viable replication fork. The replisome is reloaded onto this structure (not shown), and replication continues. Arrowheads represent 3' ends.

illustrated in Figures 25–22 to 25–25 is that they introduce a transient break into one of the DNA strands. If a replication fork encounters a damaged site under repair near a break in one of the template strands, one arm of the replication fork becomes disconnected by a double-strand break and the fork collapses. The end of that break is processed by degrading the 5'-ending strand. The resulting 3' single-strand extension is bound by a recombinase that uses it to promote strand invasion: the 3' end invades the intact duplex DNA connected to the other arm of the fork and pairs with its complementary sequence. This creates a branched DNA structure (a point where three DNA segments come together). The DNA branch can be moved in a process called **branch migration** to create an X-like crossover structure known as a **Holliday intermediate**, named after researcher Robin Holliday, who first postulated its existence. The Holliday intermediate is then resolved by a special class of nuclease. The overall process reconstructs the replication fork.

In *E. coli*, the DNA end-processing is promoted by the RecBCD nuclease/helicase. The RecBCD enzyme binds to linear DNA at a free (broken) end and moves inward along the double helix, unwinding and degrading the DNA in a reaction coupled to ATP hydrolysis (**Fig. 25–31**). The RecB and RecD subunits are helicase motors, with RecB moving 3'→5' along one strand and RecD moving 5'→3' along the other strand. The activity of the enzyme is altered when it interacts with a sequence referred to as **chi**, (5')GCTGGTGG, which binds tightly to a site on the RecC subunit. From that point, degradation of the strand with a 3' terminus is greatly reduced, but degradation of the 5'-terminal strand is increased. This process creates a single-stranded DNA with a 3' end, which is used during subsequent steps in recombination. The 1,009 chi sequences scattered throughout the *E. coli* genome enhance the frequency of recombination about 5- to 10-fold within 1,000 bp of the chi site. The enhancement declines as the distance from

the chi site increases. Sequences that enhance recombination frequency have also been identified in several other organisms.

The bacterial recombinase is the RecA protein. RecA is unusual among the proteins of DNA metabolism

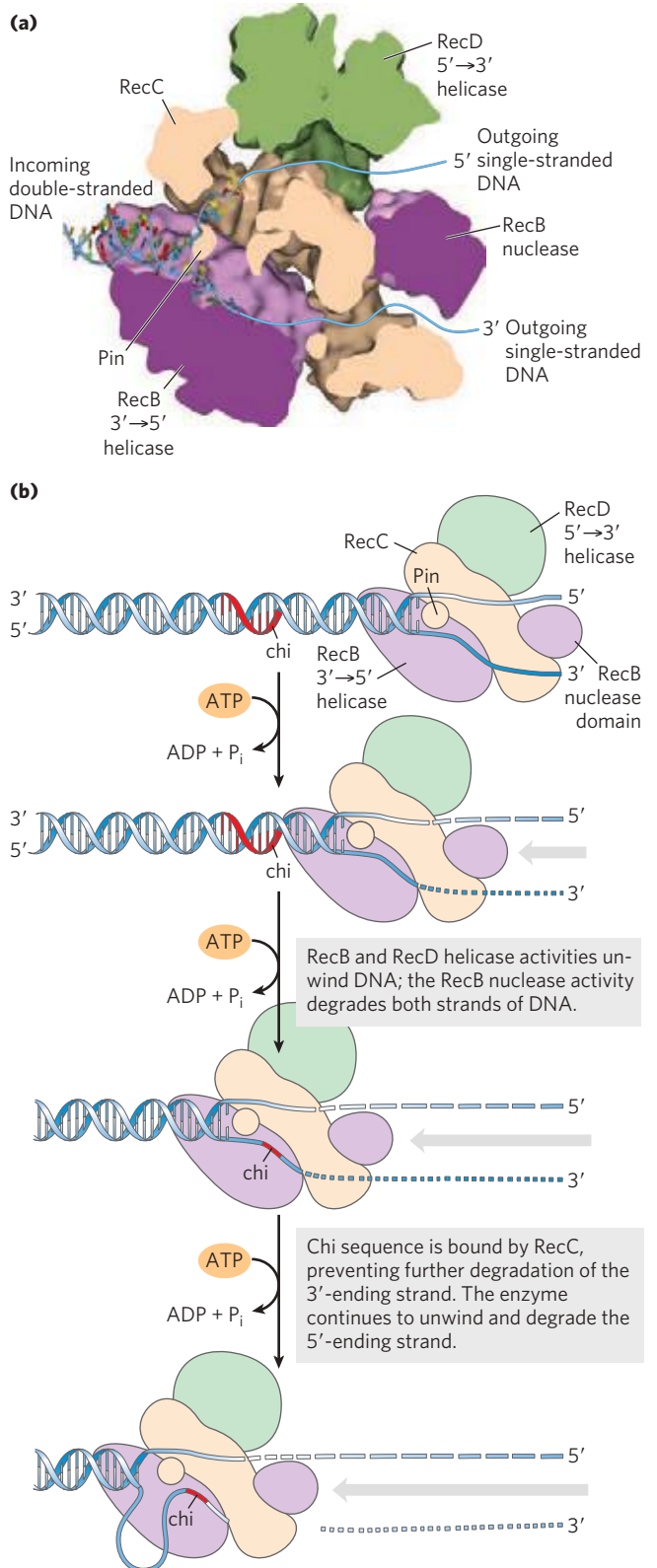
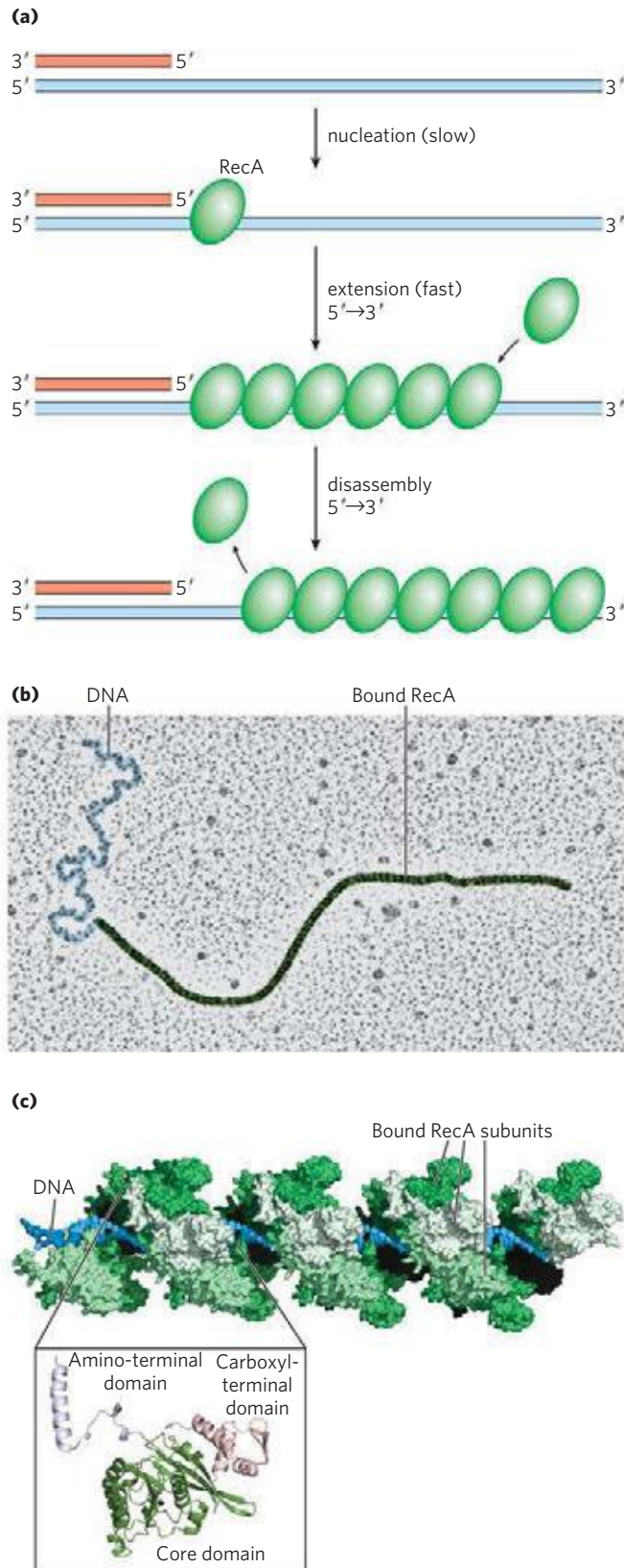


FIGURE 25–31 The RecBCD helicase/nuclease. (a) A cutaway view of the RecBCD enzyme structure as it is bound to DNA (PDB ID 1W36). The subunits are shown in different colors, with the DNA entering the structure at left and the unwound DNA strands (not part of the solved structure) are modeled as exiting to the right. A bulbous protein structure called a pin, part of the RecC subunit, facilitates the separation of strands. (b) Activities of the RecBCD enzyme at a DNA end. The RecB and RecD subunits are helicases, molecular motors that propel the complex along the DNA, a process that requires ATP; RecB degrades both strands as the complex travels, cleaving the 3'-ending strand more often than the 5'-ending strand. When a chi site is encountered on the 3'-ending strand, the RecC subunit binds to it, halting the advance of this strand through the complex. The 5'-ending strand continues to be degraded as the 3'-ending strand is looped out, eventually creating a 3' single-stranded extension. RecA protein (not shown) is finally loaded onto the processed DNA by the RecBCD enzyme.

in that its active form is an ordered, helical filament of up to several thousand subunits that assemble cooperatively on DNA (**Fig. 25–32**). This filament usually forms on single-stranded DNA, such as that produced



by the RecBCD enzyme. Its formation is not as straightforward as shown in Figure 25–32, since the single-stranded DNA-binding protein (SSB) is normally present and specifically impedes the binding of the first few subunits to DNA (filament nucleation). The RecBCD enzyme acts directly as a RecA loader, facilitating the nucleation of a RecA filament on single-stranded DNA that is coated with SSB. The filaments assemble and disassemble in a 5′→3′ direction. Many other bacterial proteins regulate the formation and disassembly of RecA filaments. RecA protein promotes the central steps of homologous recombination, including the DNA strand invasion step of Figure 25–31 and a number of other strand exchange reactions *in vitro*.

After strand invasion has occurred, branch migration is promoted by a complex called RuvAB (**Fig. 25–33a**). Once a Holliday intermediate has been created, it can be cleaved by a specialized nuclease called RuvC (**Fig. 25–33b**), and nicks are sealed with DNA ligase. A viable replication fork structure is thus reconstructed, as outlined in Figure 25–31.

After the recombination steps are completed, the replication fork reassembles in a process called **origin-independent restart of replication**. Four proteins (PriA, PriB, PriC, and DnaT) act with DnaC to load the DnaB helicase onto the reconstructed replication fork. The DnaG primase then synthesizes an RNA primer, and DNA polymerase reassembles on DnaB to restart DNA synthesis. Originally discovered as a component required for the replication of ϕ X174 DNA *in vitro*, a complex of PriA, PriB, PriC, and DnaT, along with DnaB, DnaC, and DnaG, is now termed the **replication restart primosome**. Restart of the replication fork also requires DNA polymerase II, in a role not yet defined; this polymerase II activity gives way to DNA polymerase III activity for the extensive replication generally required to complete the chromosome. In this way, the process of recombination is tightly intertwined with replication. One process of DNA metabolism supports the other.

Eukaryotic Homologous Recombination Is Required for Proper Chromosome Segregation during Meiosis

In eukaryotes, homologous genetic recombination can have several roles in replication and cell division, including the repair of stalled replication forks. Recombination

FIGURE 25–32 RecA protein filaments. RecA and other recombinases in this class function as filaments of nucleoprotein. **(a)** Filament formation proceeds in discrete nucleation and extension steps. Nucleation is the addition of the first few RecA subunits. Extension occurs by adding RecA subunits so that the filament grows in the 5′→3′ direction. When disassembly occurs, subunits are subtracted from the trailing end. **(b)** Colorized electron micrograph of a RecA filament bound to DNA. **(c)** Segment of a RecA filament with four helical turns (24 RecA subunits; derived from PDB ID 3CMX). Notice the bound double-stranded DNA in the center. The core domain of RecA is structurally related to the domains in helicases.

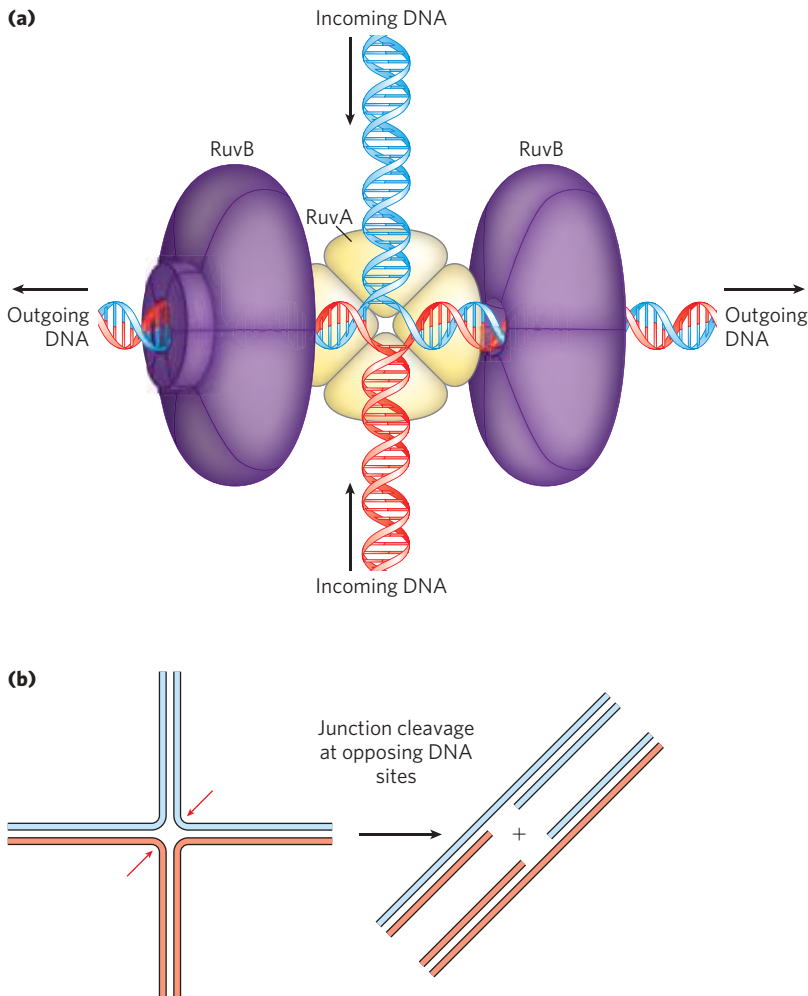
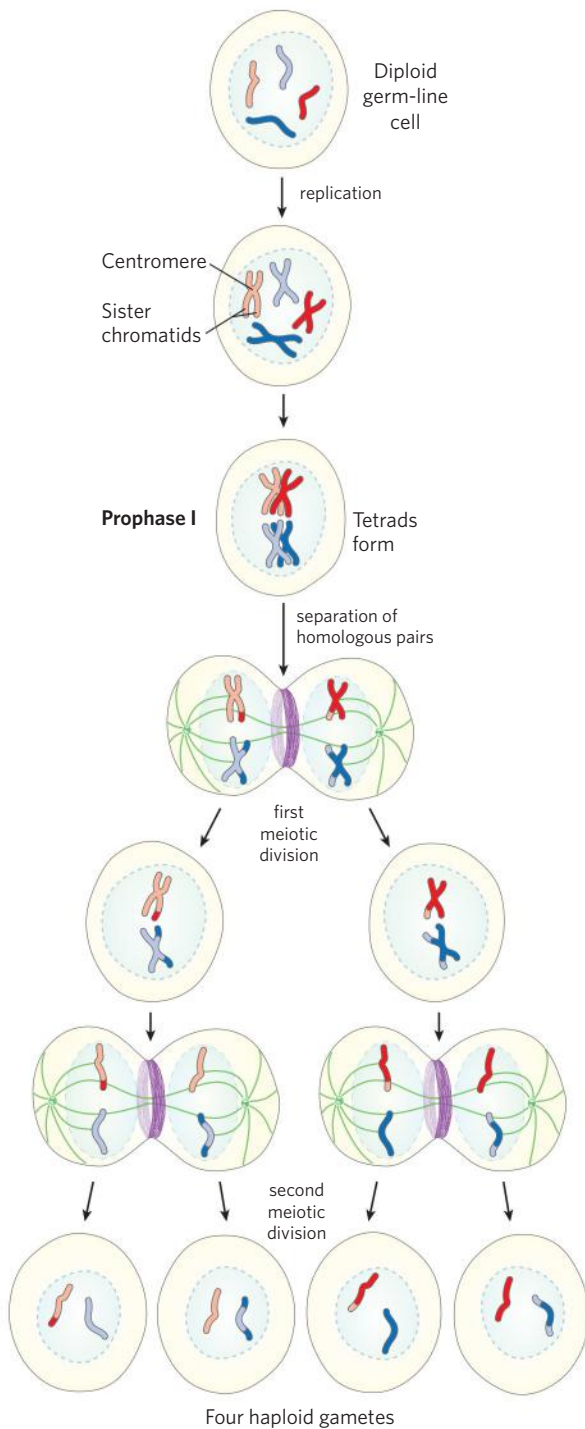


FIGURE 25-33 Catalysis of DNA branch migration and Holliday intermediate resolution by the RuvA, RuvB, and RuvC proteins. **(a)** The RuvA protein binds directly to a Holliday intermediate where the four DNA arms come together. The hexameric RuvB protein is a DNA translocase. Two hexamers bind to opposing arms of the Holliday intermediate and propel the DNA outward in a reaction coupled to ATP hydrolysis. The branch thus moves. **(b)** RuvC is a specialized nuclease that binds to the RuvAB complex and cleaves the Holliday intermediate on opposing sides of the junction (red arrows) so that two contiguous DNA arms remain in each product.

occurs with the highest frequency during **meiosis**, the process by which diploid germ-line cells with two sets of chromosomes divide to produce haploid gametes (sperm cells or ova) in animals (haploid spores in plants)—each gamete having only one member of each chromosome pair (**Fig. 25-34**). Meiosis begins with replication of the DNA in the germ-line cell so that each DNA molecule is present in four copies. Each set of four homologous chromosomes (tetrad) exists as two pairs of sister chromatids, and the sister chromatids remain associated at their centromeres. The cell then goes through two rounds of cell division without an intervening round of DNA replication. In the first cell division, the two pairs of sister chromatids are segregated into daughter cells. In the second cell division, the two chromosomes in each sister chromatid pair are segregated into new daughter cells. In each division, the chromosomes to be segregated are drawn into the daughter cells by spindle fibers attached to opposite poles of the dividing cell. The two successive divisions reduce the DNA content to the haploid level in each gamete. Proper chromosome segregation into daughter cells requires that physical links exist between the homologous chromosomes to be segregated. As the

spindle fibers attach to the centromeres of chromosomes and start to pull, the links between homologous chromosomes create tension. This tension, sensed by cellular mechanisms not yet understood, signals that this pair of chromosomes or sister chromatids is properly aligned for segregation. Once the tension is sensed, the links are gradually dissolved and segregation proceeds. If improper spindle fiber attachment occurs (e.g., if the centromeres of a chromosome pair are attached to the same cellular pole), a cellular kinase senses the lack of tension and activates a system that removes the spindle attachments, allowing the cell to try again.

During the second meiotic division, the centromeric attachments between the sister chromatids, augmented by cohesins deposited during replication (see **Fig. 24-34**), provide the needed physical links to guide segregation. However, during the first meiotic cell division, the two pairs of sister chromatids to be segregated are not related by a recent replication event and are not linked by cohesins or any other physical association. Instead, the homologous pairs of sister chromatids are aligned and new links are created by recombination, a process involving the breakage and rejoining of DNA



(Fig. 25–35). This exchange, also referred to as crossing over, can be observed with the light microscope. Crossing over links the two pairs of sister chromatids together at points called chiasmata (singular, chiasma). Also during crossovers, genetic material is exchanged between the pairs of sister chromatids. These exchanges also increase genetic diversity in the resulting gametes. The importance of meiotic recombination to proper chromosomal segregation is well illustrated by the physiological and societal consequences of their failure (Box 25–2).

FIGURE 25–34 Meiosis in animal germ-line cells. The chromosomes of a hypothetical diploid germ-line cell (four chromosomes; two homologous pairs) replicate and are held together at their centromeres. Each replicated double-stranded DNA molecule is called a chromatid (sister chromatid). In prophase I, just before the first meiotic division, the two homologous sets of chromatids align to form tetrads, held together by covalent links at homologous junctions (chiasmata). Crossovers occur within the chiasmata (see Fig. 25–35). These transient associations between homologs ensure that the two tethered chromosomes segregate properly in the next step, when attached spindle fibers pull them toward opposite poles of the dividing cell in the first meiotic division. The products of this division are two daughter cells each with two pairs of different sister chromatids. The pairs now line up across the equator of the cell in preparation for separation of the chromatids (now called chromosomes). The second meiotic division produces four haploid daughter cells that can serve as gametes. Each has two chromosomes, half the number of the diploid germ-line cell. The chromosomes have re-sorted and recombined.

Crossing over is not an entirely random process, and “hot spots” have been identified on many eukaryotic chromosomes. However, the assumption that crossing over can occur with equal probability at almost any point along the length of two homologous chromosomes remains a reasonable approximation in many cases, and it is this assumption that permits the genetic mapping of genes on a particular chromosome. The frequency of homologous recombination in any region separating two points on a chromosome is roughly proportional to the distance between the points, and this allows determination of the relative positions of and distances between different genes. The independent assortment of unlinked genes on different chromosomes (Fig. 25–36) makes another major contribution to the genetic diversity of the gametes. These genetic realities guide many of the modern applications of genomics, such as defining haplotypes (see Fig. 9–30) or searching for disease genes in the human genome (see Fig. 9–34).

Homologous recombination thus serves at least three identifiable functions in eukaryotes: (1) it contributes to the repair of several types of DNA damage; (2) it provides, in eukaryotic cells, a transient physical link between chromatids that promotes the orderly segregation of chromosomes at the first meiotic cell division; and (3) it enhances genetic diversity in a population.

Recombination during Meiosis Is Initiated with Double-Strand Breaks

A likely pathway for homologous recombination during meiosis is outlined in Figure 25–35a. The model has four key features. First, homologous chromosomes are aligned. Second, a double-strand break in a DNA molecule is created and then the exposed ends are processed by an exonuclease, leaving a single-strand extension with a free 3′-hydroxyl group at the broken end (step 1). Third, the exposed 3′ ends invade the intact duplex DNA of the homolog, and this is followed by

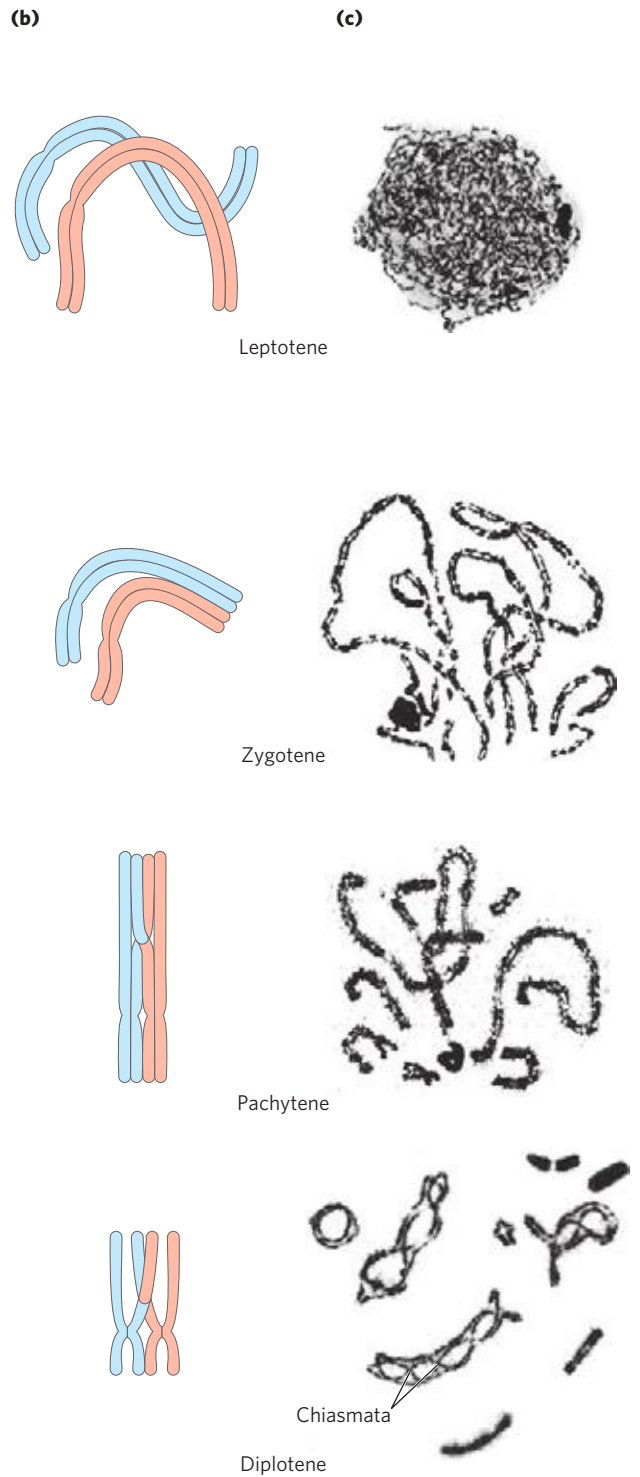
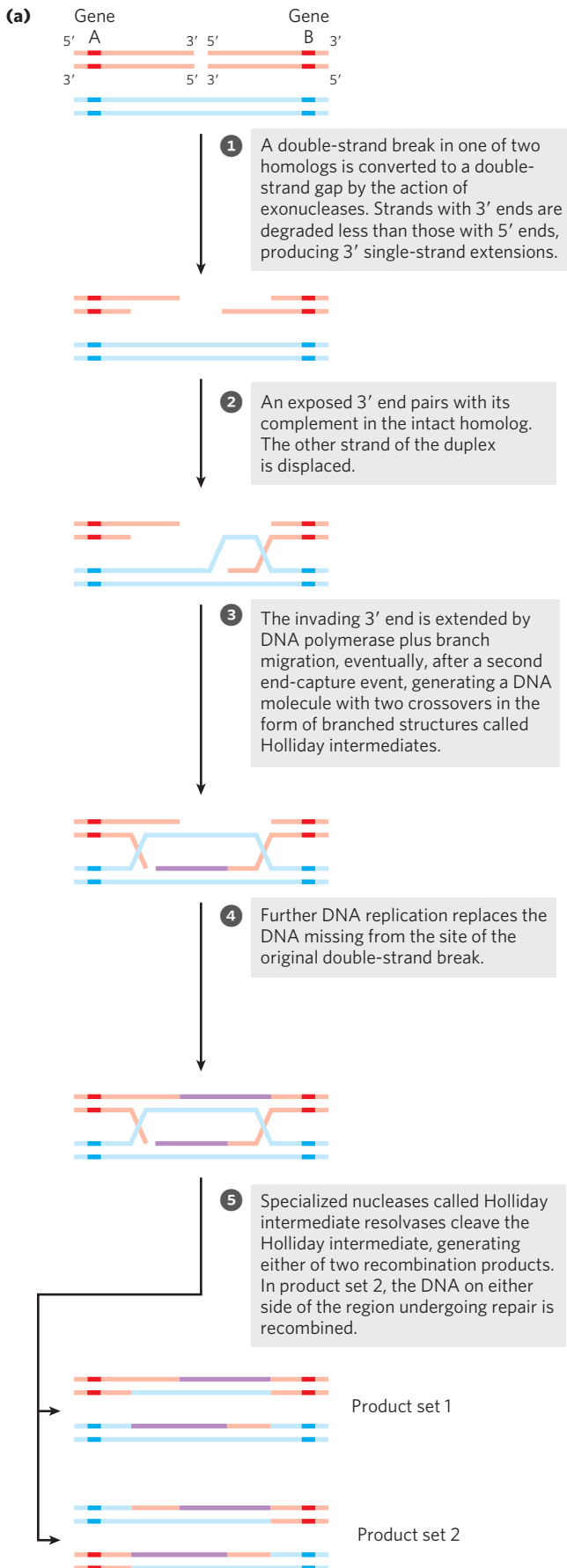


FIGURE 25–35 Recombination during prophase I in meiosis. (a) Model of double-strand break repair for homologous genetic recombination. The two homologous chromosomes (one shown in red, the other blue) involved in this recombination event have identical or very nearly identical sequences. Each of the two genes shown has different alleles on the two chromosomes. The steps are described in the text. (b) Crossing over occurs during prophase of meiosis I. The several stages of prophase I are aligned with the recombination processes that are occurring in part (a).

Double-strand breaks are introduced and processed in the leptotene stage. The strand invasion and completion of crossover occur later. As homologous sequences in the two pairs of sister chromatids are aligned in the zygotene stage, synaptonemal complexes form and strand invasion occurs. The homologous chromosomes are tightly aligned by the pachytene stage. (c) Homologous chromosomes of a grasshopper are viewed at successive stages of meiotic prophase I. The chiasmata become visible in the diplotene stage.

BOX 25–2 MEDICINE Why Proper Chromosomal Segregation Matters

When chromosomal alignment and recombination are not correct and complete in meiosis I, segregation of chromosomes can go awry. One result may be aneuploidy, a condition in which a cell has the wrong number of chromosomes. The haploid products of meiosis (gametes or spores) may have no copies or two copies of a chromosome. When a gamete with two copies of a chromosome joins with a gamete with one copy of a chromosome during fertilization, cells in the resulting embryo have three copies of that chromosome (they are trisomic).

In *S. cerevisiae*, aneuploidy resulting from errors in meiosis occurs at a rate of about 1 in 10,000 meiotic events. In fruit flies, the rate is about 1 in a few thousand. Rates of aneuploidy in mammals are considerably higher. In mice, the rate is 1 in 100, and it is even higher in other mammals. The rate of aneuploidy in fertilized human eggs has been estimated as 10% to 30%; most of these aneuploid cells are monosomies (they have a single copy of a chromosome) or trisomies. This is almost certainly an underestimate. Most trisomies are lethal and many result in miscarriage long before the pregnancy is detected. Aneuploidy is the leading cause of pregnancy loss. The few trisomic fetuses that survive to birth generally have three copies of chromosome 13, 18, or 21 (trisomy 21 is Down syndrome). Abnormal complements of the sex chromosomes are also found in the human population. Almost all monosomies are fatal in the early stages of fetal development. The societal consequences of aneuploidy in humans are considerable. Aneuploidy is the leading genetic cause of developmental and mental disabilities. At the heart of these high rates is a feature of meiosis in female mammals that has special significance for the human species.

In a human male, germ-line cells begin to undergo meiosis at puberty, and each meiotic event requires a relatively short time. In contrast, meiosis in the germ-line cells of human females is a highly protracted process. The production of an egg begins before a female is born—with the onset of meiosis in the fetus, at 12 to 13 weeks of gestation. Meiosis is initiated in all the developing fetal germ-line cells over a period of a few weeks. The cells proceed through much of meiosis I. Chromosomes line up and generate crossovers, continuing just beyond the pachytene phase (see Fig. 25–35)—and

then the process stops. The chromosomes enter an arrested phase called the dictyate stage, with the crossovers in place, a kind of suspended animation where they remain as the female matures—typically from 13 to 50 years. It is not until sexual maturity that individual germ-line cells continue through the two meiotic cell divisions to produce egg cells.

Between the onset of the dictyate stage and the final completion of meiosis, something may happen that disrupts or damages the crossovers linking homologous chromosomes in the germ-line cells. As a woman ages, the rate of trisomy in the egg cells she produces increases, dramatically so as she approaches menopause (Fig. 1). There are many hypotheses on why this occurs, and several different factors may play a role. However, most of the hypotheses are centered on recombination crossovers in meiosis I and their stability over the protracted dictyate stage.

It is not yet clear what medical steps could be taken to reduce the incidence of aneuploidy in females of child-bearing age. What is revealed is the inherent importance of recombination and crossover generation in human meiosis.

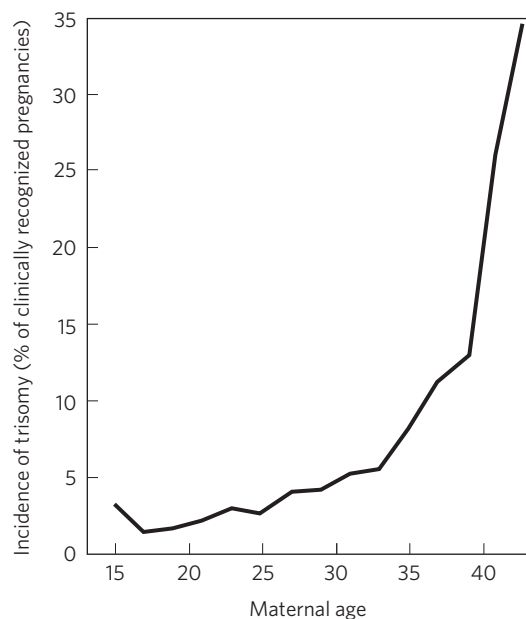


FIGURE 1 The increasing incidence of human trisomy with increasing age of the mother.

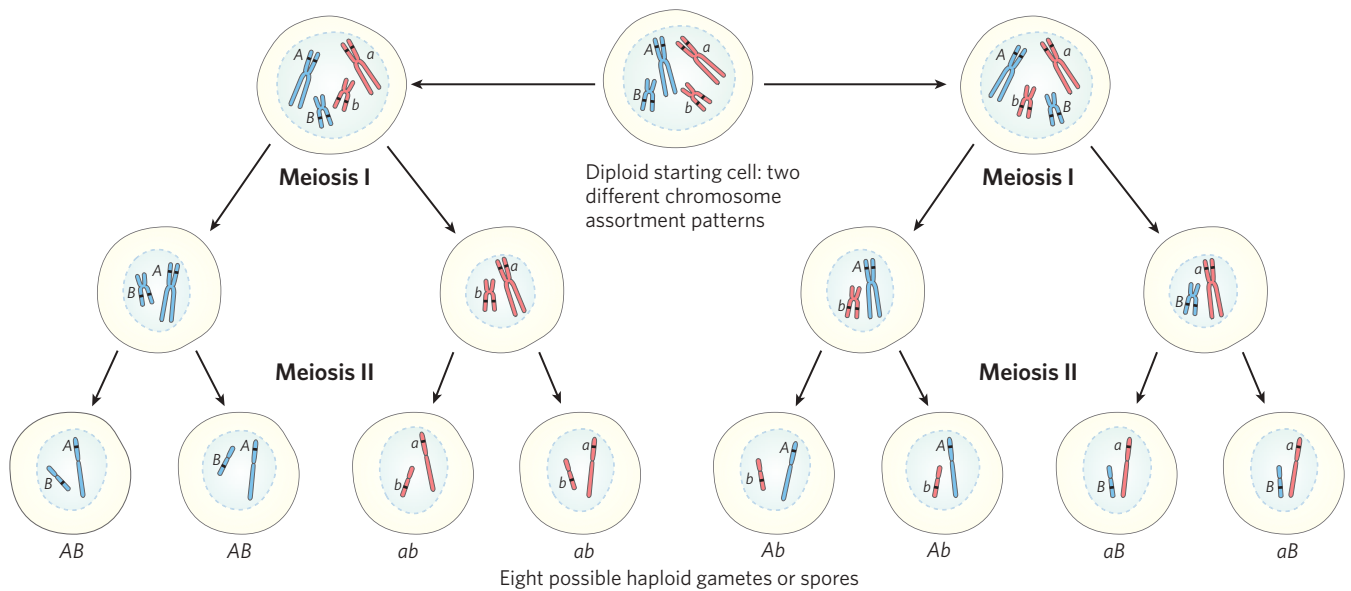


FIGURE 25-36 The contribution of independent assortment to genetic diversity. In this example, the chromosomes have already been replicated to create two pairs of sister chromatids for two chromosomes. Blue and red distinguish the sister chromatids of each pair. One gene on each chromosome is highlighted, with different alleles (A or a , B or b) in

the homologs. Independent assortment can lead to gametes with any combination of the alleles present on the two different chromosomes. Crossing over (not shown here; see Fig. 25-34) would also contribute to genetic diversity in a typical meiotic sequence.

branch migration and/or replication to create a pair of Holliday intermediates (steps 2 to 4). Fourth, cleavage of the two crossovers creates either of two pairs of complete recombinant products (step 5). Note the similarity of these steps to the bacterial recombinational repair processes outlined in Figure 25-30.

In this **double-strand break repair model** for recombination, the 3' ends are used to initiate the genetic exchange. Once paired with the complementary strand on the intact homolog, a region of hybrid DNA is created that contains complementary strands from two different parent DNAs (the product of step 2 in Fig. 25-35a). Each of the 3' ends can then act as a primer for DNA replication. Meiotic homologous recombination can vary in many details from one species to another, but most of the steps outlined above are generally present in some form. There are two ways to cleave, or “resolve,” the Holliday intermediate with a RuvC-like nuclease so that the two products carry genes in the same linear order as in the substrates—the original, un-recombined chromosomes (step 5). If cleaved one way, the DNA flanking the region containing the hybrid DNA is not recombined; if cleaved the other way, the flanking DNA is recombined. Both outcomes are observed in vivo.

The homologous recombination illustrated in Figure 25-35 is a very elaborate process that is essential to accurate chromosome segregation. Its molecular consequences for the generation of genetic diversity are subtle. To understand how this process contributes to diversity, we should keep in mind that the two homologous chromosomes that undergo recombination are not

necessarily *identical*. The linear array of genes may be the same, but the base sequences in some of the genes may differ slightly (in different alleles). In a human, for example, one chromosome may contain the allele for hemoglobin A (normal hemoglobin) while the other contains the allele for hemoglobin S (the sickle-cell mutation). The difference may consist of no more than one base pair among millions. Homologous recombination does not change the linear array of genes, but it can determine which alleles become linked on a single chromosome and are thereby passed to the next generation together. The independent assortment of different chromosomes (Fig. 25-36) determines which gene alleles from different chromosomes are inherited together.

Site-Specific Recombination Results in Precise DNA Rearrangements

Homologous genetic recombination, the type we have discussed so far, can involve any two homologous sequences. The second general type of recombination, site-specific recombination, is a very different type of process: recombination is limited to specific sequences. Recombination reactions of this type occur in virtually every cell, filling specialized roles that vary greatly from one species to another. Examples include regulation of the expression of certain genes and promotion of programmed DNA rearrangements in embryonic development or in the replication cycles of some viral and plasmid DNAs. Each site-specific recombination system consists of an enzyme called a recombinase and a short

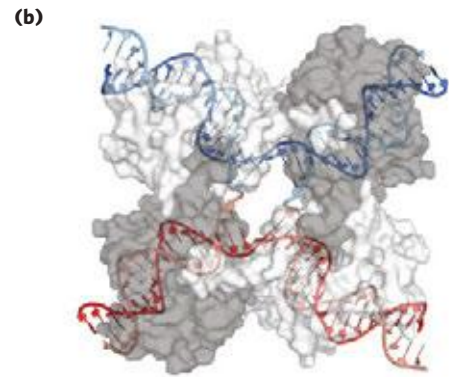
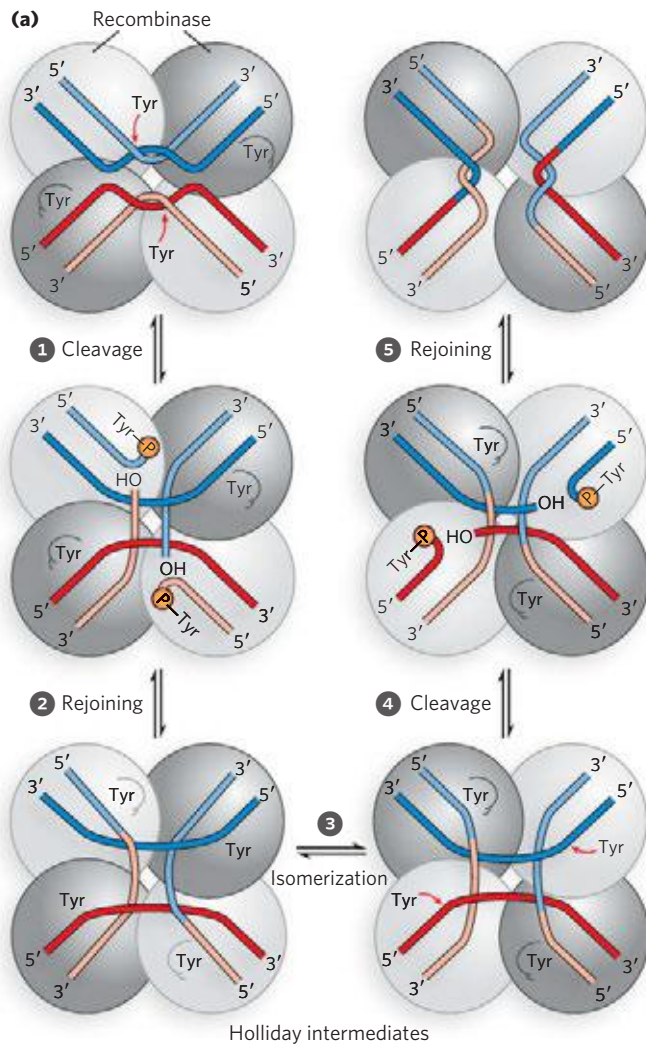


FIGURE 25-37 A site-specific recombination reaction. (a) The reaction shown here is for a common class of site-specific recombinases called integrase-class recombinases (named after bacteriophage λ integrase, the first recombinase characterized). These enzymes use Tyr residues as nucleophiles at the active site. The reaction is carried out within a tetramer of identical subunits. Recombinase subunits bind to a specific sequence, the recombination site. **1** One strand in each DNA is cleaved at particular points in the sequence. The nucleophile is the —OH group of an active-site Tyr residue, and the product is a covalent phosphotyrosine link between protein and DNA **2**. After isomerization **3**, the cleaved strands join to new partners, producing a Holliday intermediate. Steps **4** and **5** complete the reaction by a process similar to the first two steps. The original sequence of the recombination site is regenerated after recombining the DNA flanking the site. These steps occur within a complex of multiple recombinase subunits that sometimes includes other proteins not shown here. (b) Surface contour model of a four-subunit integrase-class recombinase called the Cre recombinase, bound to a Holliday intermediate (shown with light blue and dark blue helix strands). The protein has been rendered transparent so that the bound DNA is visible (derived from PDB ID 3CRX). Another group of recombinases, called the resolvase/invertase family, use a Ser residue as nucleophile at the active site.

(20 to 200 bp), unique DNA sequence where the recombinase acts (the recombination site). One or more auxiliary proteins may regulate the timing or outcome of the reaction.

There are two general classes of site-specific recombination systems, which rely on either Tyr or Ser residues in the active site. In vitro studies of many site-specific recombination systems in the tyrosine class have elucidated some general principles, including the fundamental reaction pathway

strands are rejoined to new partners to form a Holliday intermediate, with new phosphodiester bonds created at the expense of the protein-DNA linkage (step **2**). An isomerization then occurs (step **3**), and the process is repeated at a second point within each of the two recombination sites (steps **4** and **5**). In the systems that employ an active-site Ser residue, both strands of each recombination site are cut concurrently and rejoined to new partners without the Holliday intermediate. In both types of system, the exchange is always reciprocal and precise, regenerating the recombination sites when the reaction is complete. We can view a recombinase as a site-specific endonuclease and ligase in one package.

The sequences of the recombination sites recognized by site-specific recombinases are partially asymmetric (nonpalindromic), and the two recombining sites align in the same orientation during the recombinase reaction. The outcome depends on the location and orientation of the recombination sites. If the two sites are on the same DNA molecule, the reaction either inverts or deletes the

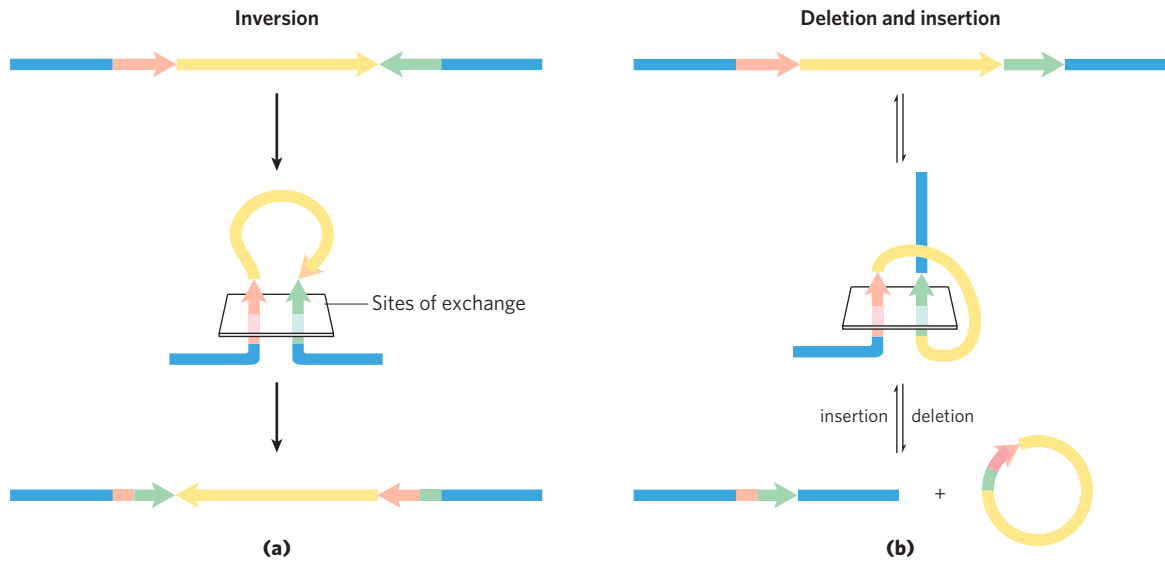


FIGURE 25-38 Effects of site-specific recombination. The outcome of site-specific recombination depends on the location and orientation of the recombination sites (red and green) in a double-stranded DNA molecule. Orientation here (shown by arrowheads) refers to the order of nucleotides in the recombination site, not the 5'→3' direction.

(a) Recombination sites with opposite orientation in the same DNA molecule. The result is an inversion. **(b)** Recombination sites with the same orientation, either on one DNA molecule, producing a deletion, or on two DNA molecules, producing an insertion.

intervening DNA, determined by whether the recombination sites have the opposite or the same orientation, respectively. If the sites are on different DNAs, the recombination is intermolecular; if one or both DNAs are circular, the result is an insertion. Some recombinase systems are highly specific for one of these reaction types and act only on sites with particular orientations.

Complete chromosomal replication can require site-specific recombination. Recombinational DNA repair of a circular bacterial chromosome, while essential, sometimes generates deleterious byproducts. The resolution of a Holliday intermediate at a replication fork by a nuclease such as RuvC, followed by completion of replication, can give rise to one of two products: the usual two monomeric chromosomes or a contiguous dimeric chromosome (Fig. 25-39). In the latter case, the covalently linked chromosomes cannot be segregated to daughter cells at cell division and the dividing cells become “stuck.” A specialized site-specific recombination system in *E. coli*, the XerCD system, converts the dimeric chromosomes to monomeric chromosomes so that cell division can proceed. The reaction is a site-specific deletion (Fig. 25-38b). This

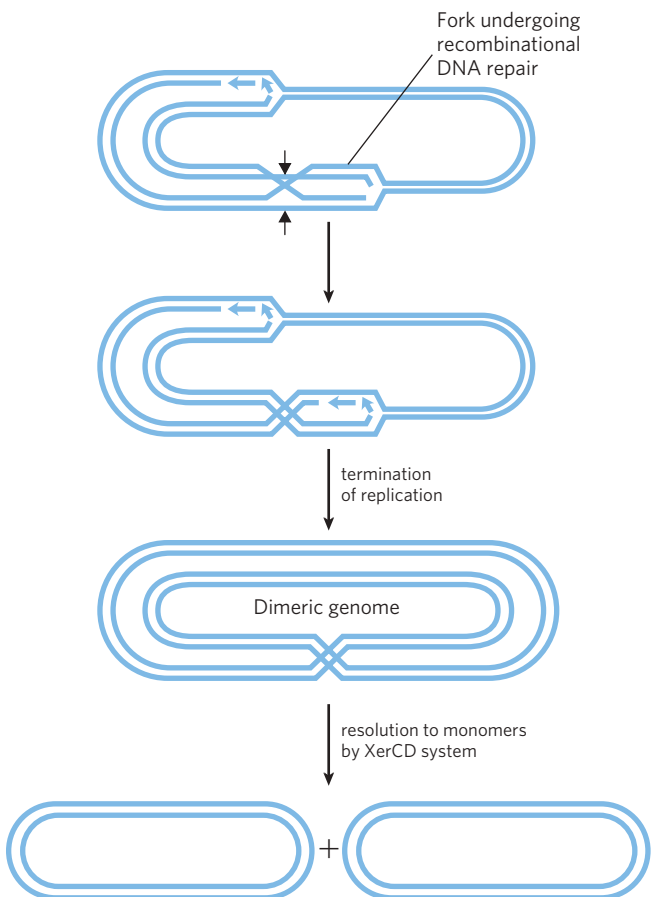



FIGURE 25-39 DNA deletion to undo a deleterious effect of recombinational DNA repair. The resolution of a Holliday intermediate during recombination (if cut at the points indicated by red arrows) can generate a contiguous dimeric chromosome. A specialized site-specific recombinase in *E. coli*, XerCD, converts the dimer to monomers, allowing chromosome segregation and cell division to proceed.

is another example of the close coordination between DNA recombination processes and other aspects of DNA metabolism.

Transposable Genetic Elements Move from One Location to Another

We now consider the third general type of recombination system: recombination that allows the movement of transposable elements, or **transposons**. These segments of DNA, found in virtually all cells, move, or “jump,” from one place on a chromosome (the donor site) to another on the same or a different chromosome (the target site). DNA sequence homology is not usually required for this movement, called **transposition**; the new location is determined more or less randomly. Insertion of a transposon in an essential gene could kill the cell, so transposition is tightly regulated and usually very infrequent. Transposons are perhaps the simplest of molecular parasites, adapted to replicate passively within the chromosomes of host cells. In some cases they carry genes that are useful to the host cell, and thus exist in a kind of symbiosis with the host.

 Bacteria have two classes of transposons. **Insertion sequences** (simple transposons) contain only the sequences required for transposition and the genes for the proteins (transposases) that promote the process. **Complex transposons** contain one or more genes in addition to those needed for transposition. These extra genes might, for example, confer resistance to antibiotics and thus enhance the survival chances of the host cell. The spread of antibiotic-resistance elements among disease-causing bacterial populations that is rendering some antibiotics ineffectual (p. 981) is mediated in part by transposition. ■

Bacterial transposons vary in structure, but most have short repeated sequences at each end that serve as binding sites for the transposase. When transposition occurs, a short sequence at the target site (5 to 10 bp) is duplicated to form an additional short repeated sequence that flanks each end of the inserted transposon (**Fig. 25–40**). These duplicated segments result from the cutting mechanism used to insert a transposon into the DNA at a new location.

There are two general pathways for transposition in bacteria. In direct (or simple) transposition (**Fig. 25–41**, left), cuts on each side of the transposon excise it, and the transposon moves to a new location. This leaves a double-strand break in the donor DNA that must be repaired. At the target site, a staggered cut is made (as in **Fig. 25–40**), the transposon is inserted into the break, and DNA replication fills in the gaps to duplicate the target site sequence. In replicative transposition (**Fig. 25–41**, right), the entire transposon is replicated, leaving a copy behind at the donor location. A **cointegrate** is an intermediate in this process, consisting of the donor region covalently linked to DNA at the

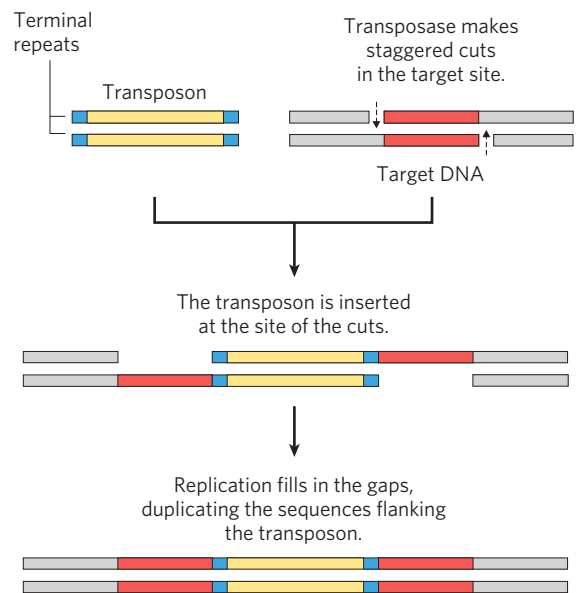


FIGURE 25–40 Duplication of the DNA sequence at a target site when a transposon is inserted. The sequences that are duplicated following transposon insertion are shown in red. These sequences are generally only a few base pairs long, so their size relative to that of a typical transposon is greatly exaggerated in this drawing.

target site. Two complete copies of the transposon are present in the cointegrate, both having the same relative orientation in the DNA. In some well-characterized transposons, the cointegrate intermediate is converted to products by site-specific recombination, in which specialized recombinases promote the required deletion reaction.

Eukaryotes also have transposons, structurally similar to bacterial transposons, and some use similar transposition mechanisms. In other cases, however, the mechanism of transposition seems to involve an RNA intermediate. Evolution of these transposons is intertwined with the evolution of certain classes of RNA viruses. Both are described in the next chapter.

Immunoglobulin Genes Assemble by Recombination

Some DNA rearrangements are a programmed part of development in eukaryotic organisms. An important example is the generation of complete immunoglobulin genes from separate gene segments in vertebrate genomes. A human (like other mammals) is capable of producing *millions* of different immunoglobulins (antibodies) with distinct binding specificities, even though the human genome contains only ~29,000 genes. Recombination allows an organism to produce an extraordinary diversity of antibodies from a limited DNA-coding capacity. Studies of the recombination mechanism reveal a close relationship to DNA transposition and suggest that this system for generating antibody diversity may have evolved from an ancient cellular invasion of transposons.

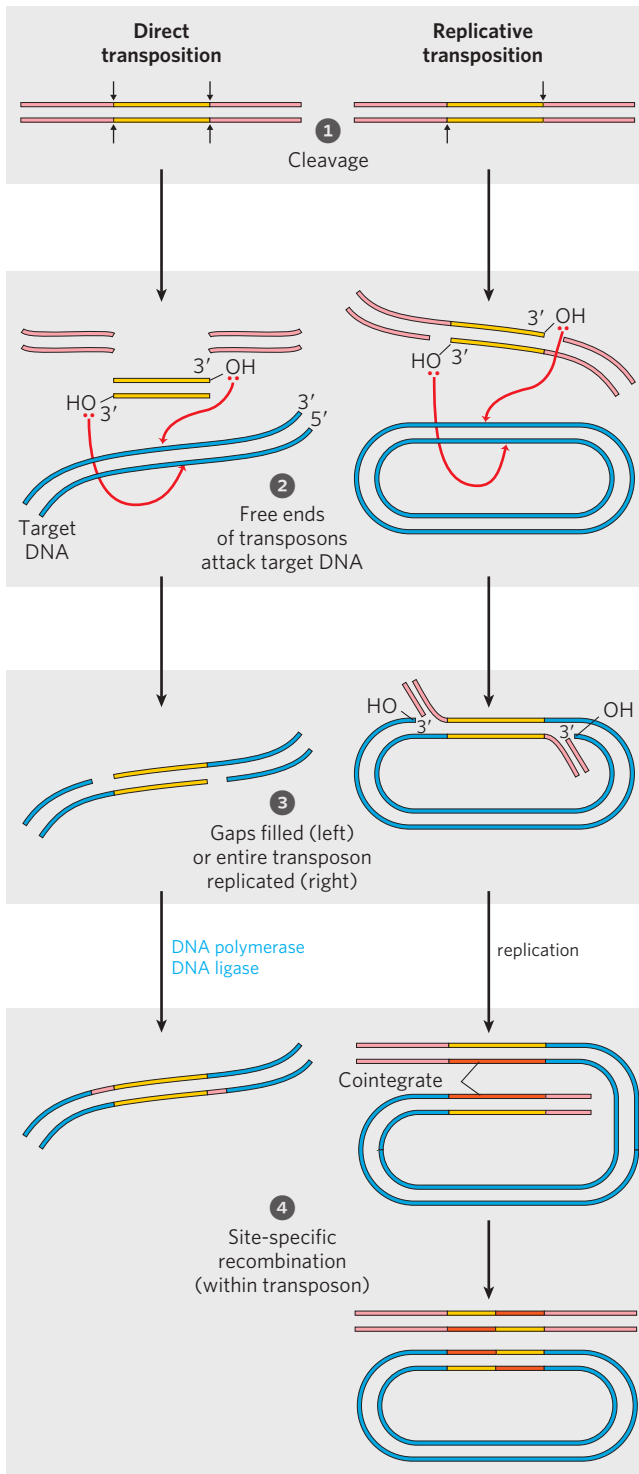


FIGURE 25-41 Two general pathways for transposition: direct (simple) and replicative. 1 The DNA is first cleaved on each side of the transposon, at the sites indicated by arrows. 2 The liberated 3'-hydroxyl groups at the ends of the transposon act as nucleophiles in a direct attack on phosphodiester bonds in the target DNA. The target phosphodiester bonds are staggered (not directly across from each other) in the two DNA strands. 3 The transposon is now linked to the target DNA. In direct transposition (left), replication fills in gaps at each end to complete the process. In replicative transposition (right), the entire transposon is replicated to create a cointegrate intermediate. 4 The cointegrate is often resolved later, with the aid of a separate site-specific recombination system. The cleaved host DNA left behind after direct transposition is either repaired by DNA end-joining or degraded (not shown). The latter outcome can be lethal to an organism.

lambda, which differ somewhat in the sequences of their constant regions. For all three types of polypeptide chain (heavy chain, and kappa and lambda light chains), diversity in the variable regions is generated by a similar mechanism. The genes for these polypeptides are divided into segments, and the genome contains clusters with multiple versions of each segment. The joining of one version of each gene segment creates a complete gene.

Figure 25-42 depicts the organization of the DNA encoding the kappa light chains of human IgG and shows how a mature kappa light chain is generated. In undifferentiated cells, the coding information for this polypeptide chain is separated into three segments. The V (variable) segment encodes the first 95 amino acid residues of the variable region, the J (joining) segment encodes the remaining 12 residues of the variable region, and the C segment encodes the constant region. The genome contains ~300 different V segments, 4 different J segments, and 1 C segment.

As a stem cell in the bone marrow differentiates to form a mature B lymphocyte, one V segment and one J segment are brought together by a specialized recombination system (Fig. 25-42). During this programmed DNA deletion, the intervening DNA is discarded. There are about $300 \times 4 = 1,200$ possible V-J combinations. The recombination process is not as precise as the site-specific recombination described earlier, so additional variation occurs in the sequence at the V-J junction. This increases the overall variation by a factor of at least 2.5, so the cells can generate about $2.5 \times 1,200 = 3,000$ different V-J combinations. The final joining of the V-J combination to the C region is accomplished by an RNA-splicing reaction after transcription, a process described in Chapter 26.

The recombination mechanism for joining the V and J segments is illustrated in **Figure 25-43**. Just beyond each V segment and just before each J segment lie recombination signal sequences (RSS). These are bound by proteins called RAG1 and RAG2 (products of the *recombination activating gene*). The RAG proteins catalyze the formation of a double-strand break between

We can use the human genes that encode proteins of the immunoglobulin G (IgG) class to illustrate how antibody diversity is generated. Immunoglobulins consist of two heavy and two light polypeptide chains (see Fig. 5-21). Each chain has two regions, a variable region, with a sequence that differs greatly from one immunoglobulin to another, and a region that is virtually constant within a class of immunoglobulins. There are also two distinct families of light chains, kappa and

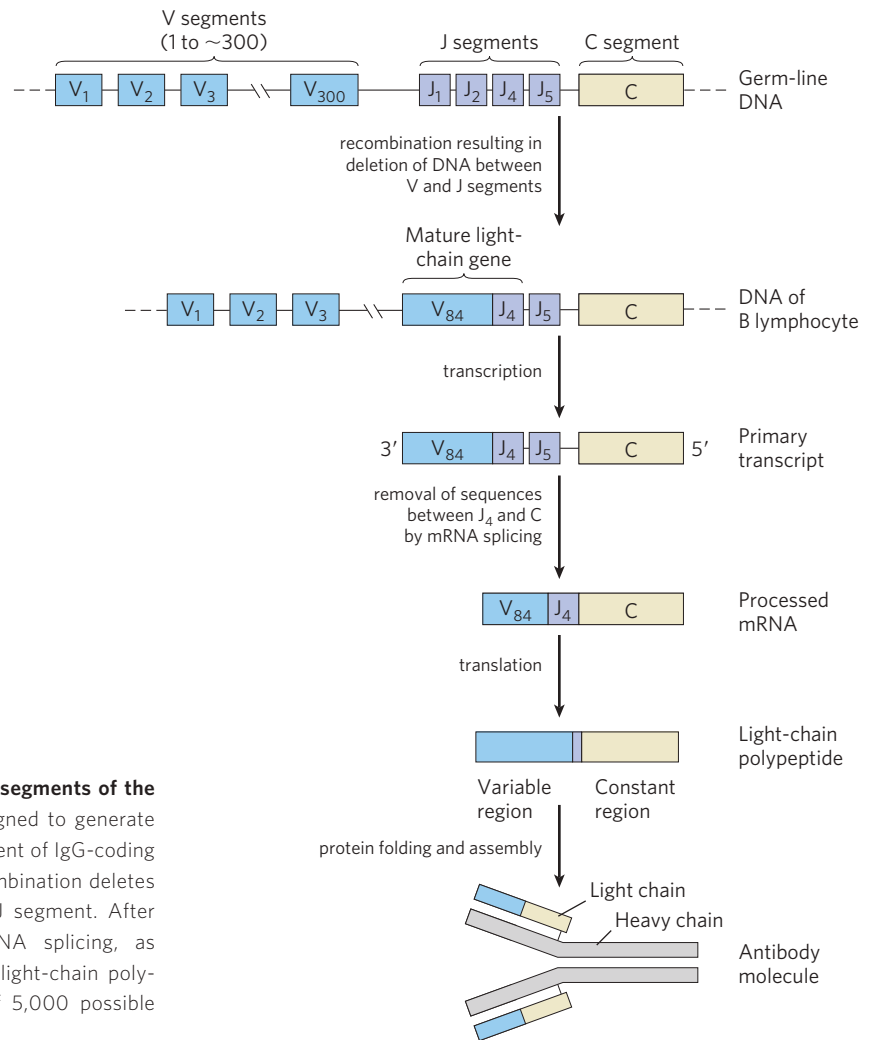


FIGURE 25–42 Recombination of the V and J gene segments of the human IgG kappa light chain. This process is designed to generate antibody diversity. At the top is shown the arrangement of IgG-coding sequences in a stem cell of the bone marrow. Recombination deletes the DNA between a particular V segment and a J segment. After transcription, the transcript is processed by RNA splicing, as described in Chapter 26; translation produces the light-chain polypeptide. The light chain can combine with any of 5,000 possible heavy chains to produce an antibody molecule.

the signal sequences and the V (or J) segments to be joined. The V and J segments are then joined with the aid of a second complex of proteins.

The genes for the heavy chains and the lambda light chains form by similar processes. Heavy chains have more gene segments than light chains, with more than 5,000 possible combinations. Because any heavy chain can combine with any light chain to generate an immunoglobulin, each human has at least $3,000 \times 5,000 = 1.5 \times 10^7$ possible IgGs. And additional diversity is generated by high mutation rates (of unknown mechanism) in the V sequences during B-lymphocyte differentiation. Each mature B lymphocyte produces only one type of antibody, but the range of antibodies produced by the B lymphocytes of an individual organism is clearly enormous.

Did the immune system evolve in part from ancient transposons? The mechanism for generation of the double-strand breaks by RAG1 and RAG2 does mirror several reaction steps in transposition (Fig. 25–43). In addition, the deleted DNA, with its terminal RSS, has a sequence structure found in most transposons. In the test tube, RAG1 and RAG2 can associate with this

deleted DNA and insert it, transposonlike, into other DNA molecules (probably a rare reaction in B lymphocytes). Although we cannot know for certain, the properties of the immunoglobulin gene rearrangement system suggest an intriguing origin in which the distinction between host and parasite has become blurred by evolution.

SUMMARY 25.3 DNA Recombination

- ▶ DNA sequences are rearranged in recombination reactions, usually in processes tightly coordinated with DNA replication or repair.
- ▶ Homologous genetic recombination can take place between any two DNA molecules that share sequence homology. In bacteria, recombination serves mainly as a DNA repair process, focused on reactivating stalled or collapsed replication forks or on the general repair of double-strand breaks. In eukaryotes, recombination is essential to ensure accurate chromosomal segregation during the first meiotic cell division. It also helps to create genetic diversity in the resulting gametes.

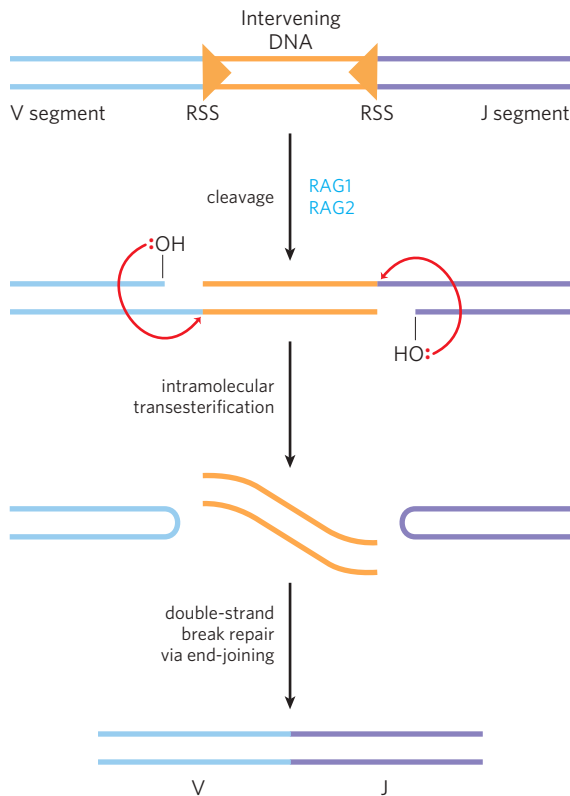


FIGURE 25–43 Mechanism of immunoglobulin gene rearrangement.

The RAG1 and RAG2 proteins bind to the recombination signal sequences (RSS) and cleave one DNA strand between the RSS and the V (or J) segments to be joined. The liberated 3' hydroxyl then acts as a nucleophile, attacking a phosphodiester bond in the other strand to create a double-strand break. The resulting hairpin bends on the V and J segments are cleaved, and the ends are covalently linked by a complex of proteins specialized for end-joining repair of double-strand breaks. The steps in the generation of the double-strand break catalyzed by RAG1 and RAG2 are chemically related to steps in transposition reactions.

- ▶ Site-specific recombination occurs only at specific target sequences, and this process can also involve a Holliday intermediate. Recombinases cleave the DNA at specific points and ligate the strands to new partners. This type of recombination is found in virtually all cells, and its many functions include DNA integration and regulation of gene expression.
- ▶ In virtually all cells, transposons use recombination to move within or between chromosomes. In vertebrates, a programmed recombination reaction related to transposition joins immunoglobulin gene segments to form immunoglobulin genes during B-lymphocyte differentiation.

Key Terms

Terms in bold are defined in the glossary.

template 1011	replication fork 1012
semiconservative	origin 1012
replication 1011	Okazaki fragment 1012

leading strand 1012	DNA polymerase α 1026
lagging strand 1013	DNA polymerase δ 1026
nucleases 1013	DNA polymerase ϵ 1026
exonucleases 1013	mutation 1027
endonucleases 1013	base-excision repair 1030
DNA polymerase I 1013	DNA glycosylases 1030
primer 1014	AP site 1030
primer terminus 1014	AP endonucleases 1031
processivity 1014	DNA photolyases 1032
proofreading 1015	error-prone translesion DNA synthesis 1035
DNA polymerase III 1016	SOS response 1035
replisome 1017	homologous genetic recombination 1038
helicases 1017	site-specific recombination 1038
topoisomerases 1017	recombination 1038
primases 1018	DNA transposition 1038
DNA ligases 1018	recombinational DNA repair 1039
DNA unwinding element (DUE) 1019	branch migration 1040
AAA+ ATPases 1019	Holliday intermediate 1040
primosome 1021	meiosis 1042
catenane 1025	double-strand break repair model 1046
pre-replicative complex (pre-RC) 1025	transposon 1049
licensing 1025	transposition 1049
minichromosome maintenance (MCM) protein 1025	insertion sequence 1049
ORC (origin recognition complex) 1025	cointegrate 1049

Further Reading

General

Cox, M.M., Doudna, J.A., & O'Donnell, M.E. (2012) *Molecular Biology: Principles and Practice*, W. H. Freeman and Company, New York, NY.

An excellent introduction to the science and how it is done.

Friedberg, E.C., Walker, G.C., Siede, W., Wood, R.D., Schultz, R.A., & Ellenberger, T. (2006) *DNA Repair and Mutagenesis*, 2nd edn, American Society for Microbiology, Washington, DC.

A thorough treatment of DNA metabolism and a good place to start exploring this field.

DNA Replication

Bloom, L.B. (2006) Dynamics of loading the *Escherichia coli* DNA polymerase processivity clamp. *Crit. Rev. Biochem. Mol. Biol.* **41**, 179–208.

Heller, R.C. & Marians, K.J. (2006) Replisome assembly and the direct restart of stalled replication forks. *Nat. Rev. Mol. Cell Biol.* **7**, 932–943.

Mechanisms for the restart of replication forks before the repair of DNA damage.

Hübscher, U., Maga, G., & Spadari, S. (2002) Eukaryotic DNA polymerases. *Annu. Rev. Biochem.* **71**, 133–163.

Good summary of the properties and roles of the more than one dozen known eukaryotic DNA polymerases.

Indiani, C. & O'Donnell, M. (2006) The replication clamp-loading machine at work in the three domains of life. *Nat. Rev. Mol. Cell Biol.* **7**, 751–761.

- Kaguni, J.** (2011) Replication initiation at the *Escherichia coli* chromosomal origin. *Curr. Opin. Chem. Biol.* **15**, 606–613.
- Kool, E.T.** (2002) Active site tightness and substrate fit in DNA replication. *Annu. Rev. Biochem.* **71**, 191–219.
Excellent summary of the molecular basis of replication fidelity by a DNA polymerase—base-pair geometry as well as hydrogen bonding.
- Kunkel, T.A. & Burgers, P.M.** (2008) Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol.* **18**, 521–527.
- Masai, H., Matsumoto, S., You, Z.Y., Yoshizawa-Sugata, N., & Oda, M.** (2010) Eukaryotic chromosome DNA replication: where, when, and how? *Annu. Rev. Biochem.* **79**, 89–130.
- O'Donnell, M.** (2006) Replisome architecture and dynamics in *Escherichia coli*. *J. Biol. Chem.* **281**, 10,653–10,656.
An excellent summary of what goes on at a replication fork.
- Stillman, B.** (2005) Origin recognition and the chromosome cycle. *FEBS Lett.* **579**, 877–884.
Good summary of the initiation of eukaryotic DNA replication.

DNA Repair

- Erzberger, J.P. & Berger, J.M.** (2006) Evolutionary relationships and structural mechanisms of AAA+ proteins. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 93–114.
- Lynch, M.** (2010) Evolution of the mutation rate. *Trends Genet.* **26**, 345–352.
- Kunkel, T.A. & Erie, D.A.** (2005) DNA mismatch repair. *Annu. Rev. Biochem.* **74**, 681–710.
- Schlacher, K. & Goodman, M.J.** (2007) Lessons from 50 years of SOS DNA-damage-induced mutagenesis. *Nat. Rev. Mol. Cell Biol.* **8**, 587–594.
- Sutton, M.D., Smith, B.T., Godoy, V.G., & Walker, G.C.** (2000) The SOS response: recent insights into umuDC-dependent mutagenesis and DNA damage tolerance. *Annu. Rev. Genet.* **34**, 479–497.
- Wilson, D.M. III & Bohr, V.A.** (2007) The mechanics of base excision repair, and its relation to aging and disease. *DNA Repair* **6**, 544–559.

DNA Recombination

- Cox, M.M.** (2001) Historical overview: searching for replication help in all of the rec places. *Proc. Natl. Acad. Sci. USA* **98**, 8173–8180.
A review of how recombination was shown to be a replication-fork repair process.
- Cox, M.M.** (2007) Regulation of bacterial RecA protein function. *Crit. Rev. Biochem. Mol. Biol.* **42**, 41–63.
- Grindley, N.D.F., Whiteson, K.L., & Rice, P.A.** (2006) Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* **75**, 567–605.
- Haniford, D.B.** (2006) Transposome dynamics and regulation in Tn10 transposition. *Crit. Rev. Biochem. Mol. Biol.* **41**, 407–424.
A detailed look at one well-studied bacterial transposon.
- Heyer, W.-D., Ehmsen, K.T., & Liu, J.** (2010) Regulation of homologous recombination in eukaryotes. *Annu. Rev. Genet.* **44**, 113–139.
- Levin, H.L. & Moran, J.V.** (2011) Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627.
- Lusetti, S.L. & Cox, M.M.** (2002) The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. *Annu. Rev. Biochem.* **71**, 71–100.
- Mimitou, E.P. & Symington, L.S.** (2009) Nucleases and helicases take center stage in homologous recombination. *Trends Biochem. Sci.* **34**, 264–272.
- Montano, S.P. & Rice, P.A.** (2011) Moving DNA around: DNA transposition and retroviral integration. *Curr. Opin. Struct. Biol.* **21**, 370–378.

- San Filippo, J., Sung, P., & Klein, H.** (2008) Mechanism of eukaryotic homologous recombination. *Annu. Rev. Biochem.* **77**, 229–257.

- Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C., & Wigley, D.B.** (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**, 187–193.

Problems

1. Conclusions from the Meselson-Stahl Experiment

The Meselson-Stahl experiment (see Fig. 25–2) proved that DNA undergoes semiconservative replication in *E. coli*. In the “dispersive” model of DNA replication, the parent DNA strands are cleaved into pieces of random size, then joined with pieces of newly replicated DNA to yield daughter duplexes. Explain how the results of Meselson and Stahl’s experiment ruled out such a model.

2. Heavy Isotope Analysis of DNA Replication

A culture of *E. coli* growing in a medium containing $^{15}\text{NH}_4\text{Cl}$ is switched to a medium containing $^{14}\text{NH}_4\text{Cl}$ for three generations (an eightfold increase in population). What is the molar ratio of hybrid DNA (^{15}N – ^{14}N) to light DNA (^{14}N – ^{14}N) at this point?

3. Replication of the *E. coli* Chromosome

The *E. coli* chromosome contains 4,639,221 bp.

(a) How many turns of the double helix must be unwound during replication of the *E. coli* chromosome?

(b) From the data in this chapter, how long would it take to replicate the *E. coli* chromosome at 37°C if two replication forks proceeded from the origin? Assume replication occurs at a rate of 1,000 bp/s. Under some conditions *E. coli* cells can divide every 20 min. How might this be possible?

(c) In the replication of the *E. coli* chromosome, about how many Okazaki fragments would be formed? What factors guarantee that the numerous Okazaki fragments are assembled in the correct order in the new DNA?

4. Base Composition of DNAs Made from Single-Stranded Templates

Predict the base composition of the total DNA synthesized by DNA polymerase on templates provided by an equimolar mixture of the two complementary strands of bacteriophage ϕX174 DNA (a circular DNA molecule). The base composition of one strand is A, 24.7%; G, 24.1%; C, 18.5%; and T, 32.7%. What assumption is necessary to answer this problem?

5. DNA Replication

Kornberg and his colleagues incubated soluble extracts of *E. coli* with a mixture of dATP, dTTP, dGTP, and dCTP, all labeled with ^{32}P in the α -phosphate group. After a time, the incubation mixture was treated with trichloroacetic acid, which precipitates the DNA but not the nucleotide precursors. The precipitate was collected, and the extent of precursor incorporation into DNA was determined from the amount of radioactivity present in the precipitate.

(a) If any one of the four nucleotide precursors were omitted from the incubation mixture, would radioactivity be found in the precipitate? Explain.

(b) Would ^{32}P be incorporated into the DNA if only dTTP were labeled? Explain.

(c) Would radioactivity be found in the precipitate if ^{32}P labeled the β or γ phosphate rather than the α phosphate of the deoxyribonucleotides? Explain.

6. The Chemistry of DNA Replication All DNA polymerases synthesize new DNA strands in the $5' \rightarrow 3'$ direction. In some respects, replication of the antiparallel strands of duplex DNA would be simpler if there were also a second type of polymerase, one that synthesized DNA in the $3' \rightarrow 5'$ direction. The two types of polymerase could, in principle, coordinate DNA synthesis without the complicated mechanics required for lagging strand replication. However, no such $3' \rightarrow 5'$ -synthesizing enzyme has been found. Suggest two possible mechanisms for $3' \rightarrow 5'$ DNA synthesis. Pyrophosphate should be one product of both proposed reactions. Could one or both mechanisms be supported in a cell? Why or why not? (Hint: You may suggest the use of DNA precursors not actually present in extant cells.)

7. Activities of DNA Polymerases You are characterizing a new DNA polymerase. When the enzyme is incubated with [^{32}P]-labeled DNA and no dNTPs, you observe the release of [^{32}P]dNMPs. This release is prevented by adding unlabeled dNTPs. Explain the reactions that most likely underlie these observations. What would you expect to observe if you added pyrophosphate instead of dNTPs?

8. Leading and Lagging Strands Prepare a table that lists the names and compares the functions of the precursors, enzymes, and other proteins needed to make the leading strand versus the lagging strand during DNA replication in *E. coli*.

9. Function of DNA Ligase Some *E. coli* mutants contain defective DNA ligase. When these mutants are exposed to ^3H -labeled thymine and the DNA produced is sedimented on an alkaline sucrose density gradient, two radioactive bands appear. One corresponds to a high molecular weight fraction, the other to a low molecular weight fraction. Explain.

10. Fidelity of Replication of DNA What factors promote the fidelity of replication during synthesis of the leading strand of DNA? Would you expect the lagging strand to be made with the same fidelity? Give reasons for your answers.

11. Importance of DNA Topoisomerases in DNA Replication DNA unwinding, such as that occurring in replication, affects the superhelical density of DNA. In the absence of topoisomerases, the DNA would become overwound ahead of a replication fork as the DNA is unwound behind it. A bacterial replication fork will stall when the superhelical density (σ) of the DNA ahead of the fork reaches +0.14 (see Chapter 24).

Bidirectional replication is initiated at the origin of a 6,000 bp plasmid *in vitro*, in the absence of topoisomerases. The plasmid initially has a σ of -0.06 . How many base pairs will be unwound and replicated by each replication fork before the forks stall? Assume that both forks travel at the same rate and that each includes all components necessary for elongation except topoisomerase.

12. The Ames Test In a nutrient medium that lacks histidine, a thin layer of agar containing $\sim 10^9$ *Salmonella typhimurium* histidine auxotrophs (mutant cells that require histidine to survive) produces ~ 13 colonies over a two-day incubation period at 37°C (see Fig. 25–20). How do these

colonies arise in the absence of histidine? The experiment is repeated in the presence of $0.4 \mu\text{g}$ of 2-aminoanthracene. The number of colonies produced over two days exceeds 10,000. What does this indicate about 2-aminoanthracene? What can you surmise about its carcinogenicity?

13. DNA Repair Mechanisms Vertebrate and plant cells often methylate cytosine in DNA to form 5-methylcytosine (see Fig. 8–5a). In these same cells, a specialized repair system recognizes G–T mismatches and repairs them to G=C base pairs. How might this repair system be advantageous to the cell? (Explain in terms of the presence of 5-methylcytosine in the DNA.)



14. DNA Repair in People with Xeroderma Pigmentosum

The condition known as xeroderma pigmentosum (XP) arises from mutations in at least seven different human genes (see Box 25–1). The deficiencies are generally in genes encoding enzymes involved in some part of the pathway for human nucleotide-excision repair. The various types of XP are denoted A through G (XPA, XPB, etc.), with a few additional variants lumped under the label XP-V.

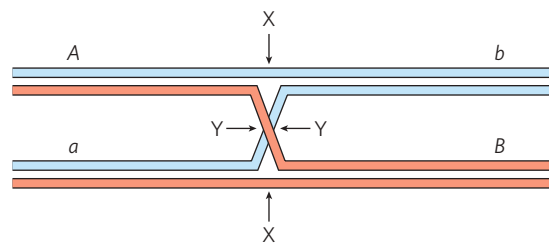
Cultures of fibroblasts from healthy individuals and from patients with XPG are irradiated with ultraviolet light. The DNA is isolated and denatured, and the resulting single-stranded DNA is characterized by analytical ultracentrifugation.

(a) Samples from the normal fibroblasts show a significant reduction in the average molecular weight of the single-stranded DNA after irradiation, but samples from the XPG fibroblasts show no such reduction. Why might this be?

(b) If you assume that a nucleotide-excision repair system is operative in fibroblasts, which step might be defective in the cells from the patients with XPG? Explain.

15. Holliday Intermediates How does the formation of Holliday intermediates in homologous genetic recombination differ from their formation in site-specific recombination?

16. Cleavage of Holliday Intermediates A Holliday intermediate is formed between two homologous chromosomes, at a point between genes *A* and *B*, as shown below. The chromosomes have different alleles of the two genes (*A* and *a*, *B* and *b*). Where would the Holliday intermediate have to be cleaved (points *X* and/or *Y*) to generate a chromosome that would convey (a) an *Ab* genotype or (b) an *ab* genotype?

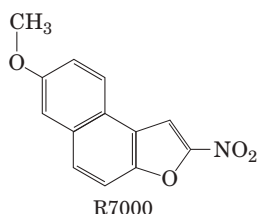


17. A Connection between Replication and Site-Specific Recombination Most wild strains of *Saccharomyces cerevisiae* have multiple copies of the circular plasmid 2μ (named for its contour length of about $2 \mu\text{m}$), which has $\sim 6,300$ bp of DNA. For its replication the plasmid uses the host replication

system, under the same strict control as the host cell chromosomes, replicating only once per cell cycle. Replication of the plasmid is bidirectional, with both replication forks initiating at a single, well-defined origin. However, one replication cycle of a 2μ plasmid can result in more than two copies of the plasmid, allowing amplification of the plasmid copy number (number of plasmid copies per cell) whenever plasmid segregation at cell division leaves one daughter cell with fewer than the normal complement of plasmid copies. Amplification requires a site-specific recombination system encoded by the plasmid, which serves to invert one part of the plasmid relative to the other. Explain how a site-specific inversion event could result in amplification of the plasmid copy number. (Hint: Consider the situation when replication forks have duplicated one recombination site but not the other.)

Data Analysis Problem

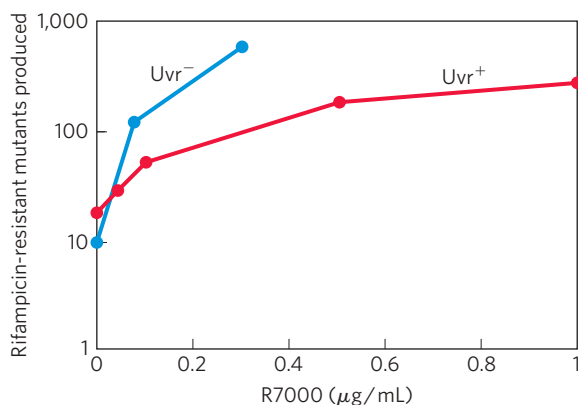
18. Mutagenesis in *Escherichia coli* Many mutagenic compounds act by alkylating the bases in DNA. The alkylating agent R7000 (7-methoxy-2-nitronaphtho[2,1-*b*]furan) is an extremely potent mutagen.



In vivo, R7000 is activated by the enzyme nitroreductase, and this more reactive form covalently attaches to DNA—primarily, but not exclusively, to G≡C base pairs.

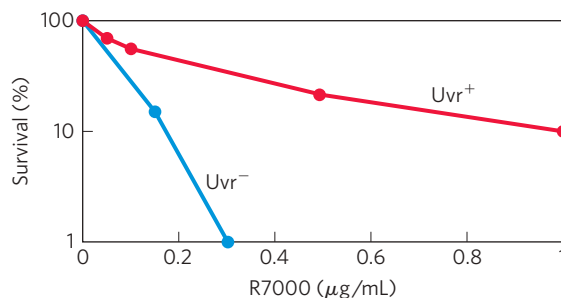
In a 1996 study, Quillardet, Touati, and Hofnung explored the mechanisms by which R7000 causes mutations in *E. coli*. They compared the genotoxic activity of R7000 in two strains of *E. coli*: the wild-type (Uvr^+) and mutants lacking *uvrA* activity (Uvr^- ; see Table 25–6). They first measured rates of mutagenesis. Rifampicin is an inhibitor of RNA polymerase (see Chapter 26). In its presence, cells will not grow unless certain mutations occur in the gene encoding RNA polymerase; the appearance of rifampicin-resistant colonies thus provides a useful measure of mutagenesis rates.

The effects of different concentrations of R7000 were determined, with the results shown in the graph below.



(a) Why are some mutants produced even when no R7000 is present?

Quillardet and colleagues also measured the survival rate of bacteria treated with different concentrations of R7000 with the following results.



(b) Explain how treatment with R7000 is lethal to cells.

(c) Explain the differences in the mutagenesis curves and in the survival curves for the two types of bacteria, Uvr^+ and Uvr^- , as shown in the graphs.

The researchers then went on to measure the amount of R7000 covalently attached to the DNA in Uvr^+ and Uvr^- *E. coli*. They incubated bacteria with [^3H]R7000 for 10 or 70 minutes, extracted the DNA, and measured its ^3H content in counts per minute (cpm) per μg of DNA.

Time (min)	^3H in DNA (cpm/ μg)	
	Uvr^+	Uvr^-
10	76	159
70	69	228

(d) Explain why the amount of ^3H drops over time in the Uvr^+ strain and rises over time in the Uvr^- strain.

Quillardet and colleagues then examined the particular DNA sequence changes caused by R7000 in the Uvr^+ and Uvr^- bacteria. For this, they used six different strains of *E. coli*, each with a different point mutation in the *lacZ* gene, which encodes β -galactosidase (this enzyme catalyzes the same reaction as lactase; see Fig. 14–11). Cells with any of these mutations have a nonfunctional β -galactosidase and are unable to metabolize lactose (i.e., a Lac^- phenotype). Each type of point mutation required a specific reverse mutation to restore *lacZ* gene function and Lac^+ phenotype. By plating cells on a medium containing lactose as the sole carbon source, it was possible to select for these reverse-mutated, Lac^+ cells. And by counting the number of Lac^+ cells following mutagenesis of a particular strain, the researchers could measure the frequency of each type of mutation.

First, they looked at the mutation spectrum in Uvr^- cells. The following table shows the results for the six strains, CC101 through CC106 (with the point mutation required to produce Lac^+ cells indicated (μg in parentheses).

Number of Lac⁺ cells (average ± SD)

R7000 (μg/mL)	CC101	CC102	CC103	CC104	CC105	CC106
	(A=T to C≡G)	(G≡C to A=T)	(G≡C to C≡G)	(G≡C To T=A)	(A=T to T=A)	(A=T to G≡C)
0	6 ± 3	11 ± 9	2 ± 1	5 ± 3	2 ± 1	1 ± 1
0.075	24 ± 19	34 ± 3	8 ± 4	82 ± 23	40 ± 14	4 ± 2
0.15	24 ± 4	26 ± 2	9 ± 5	180 ± 71	130 ± 50	3 ± 2

(e) Which types of mutation show significant increases above the background rate due to treatment with R7000? Provide a plausible explanation for why some have higher frequencies than others.

(f) Can all of the mutations you listed in (e) be explained as resulting from covalent attachment of R7000 to a G≡C base pair? Explain your reasoning.

(g) Figure 25–27b shows how methylation of guanine residues can lead to a G≡C to A=T mutation. Using a similar pathway, show how a G–R7000 adduct could lead to the G≡C to A=T or T=A mutations shown above. Which base pairs with the G–R7000 adduct?

The results for the Uvr⁺ bacteria are shown in the table below.

Number of Lac⁺ cells (average ± SD)

R7000 (μg/mL)	CC101	CC102	CC103	CC104	CC105	CC106
	(A=T to C≡G)	(G≡C to A=T)	(G≡C to C≡G)	(G≡C to T=A)	(A=T to T=A)	(A=T to G≡C)
0	2 ± 2	10 ± 9	3 ± 3	4 ± 2	6 ± 1	0.5 ± 1
1	7 ± 6	21 ± 9	8 ± 3	23 ± 15	13 ± 1	1 ± 1
5	4 ± 3	15 ± 7	22 ± 2	68 ± 25	67 ± 14	1 ± 1

(h) Do these results show that all mutation types are repaired with equal fidelity? Provide a plausible explanation for your answer.

Reference

Quillardet, P., Touati, E., & Hofnung, M. (1996) Influence of the *uvr*-dependent nucleotide excision repair on DNA adducts formation and mutagenic spectrum of a potent genotoxic agent: 7-methoxy-2-nitronaphtho[2,1-*b*]furan (R7000). *Mutat. Res.* **358**, 113–122.

RNA Metabolism

26.1 DNA-Dependent Synthesis of RNA 1058

26.2 RNA Processing 1069

26.3 RNA-Dependent Synthesis of RNA and DNA 1085

Expression of the information in a gene generally involves production of an RNA molecule transcribed from a DNA template. Strands of RNA and DNA may seem quite similar at first glance, differing only in that RNA has a hydroxyl group at the 2' position of the aldopentose, and uracil instead of thymine. However, unlike DNA, most RNAs carry out their functions as single strands, strands that fold back on themselves and have the potential for much greater structural diversity than DNA (Chapter 8). RNA is thus suited to a variety of cellular functions.

RNA is the only macromolecule known to have a role both in the storage and transmission of information and in catalysis, which has led to much speculation about its possible role as an essential chemical intermediate in the development of life on this planet. The discovery of catalytic RNAs, or ribozymes, has changed the very definition of an enzyme, extending it beyond the domain of proteins. Proteins nevertheless remain essential to RNA and its functions. In the modern cell, all nucleic acids, including RNAs, are complexed with proteins. Some of these complexes are quite elaborate, and RNA can assume both structural and catalytic roles within complicated biochemical machines.

All RNA molecules except the RNA genomes of certain viruses are derived from information permanently stored in DNA. During **transcription**, an enzyme system converts the genetic information in a segment of double-stranded DNA into an RNA strand with a base sequence complementary to one of the DNA strands. Three major kinds of RNA are produced. **Messenger RNAs (mRNAs)** encode the amino acid sequence of one or more polypeptides specified by a gene or set of genes. **Transfer RNAs (tRNAs)** read the information encoded in the mRNA and transfer the appropriate amino acid to a growing polypeptide chain during protein

synthesis. **Ribosomal RNAs (rRNAs)** are constituents of ribosomes, the intricate cellular machines that synthesize proteins. Many additional, specialized RNAs have regulatory or catalytic functions or are precursors to the three main classes of RNA. These special-function RNAs are no longer thought of as minor species in the catalog of cellular RNAs. In vertebrates, RNAs that do not fit into one of the classical categories (mRNA, tRNA, rRNA) seem to vastly outnumber those that do.

During replication the entire chromosome is usually copied, but transcription is more selective. Only particular genes or groups of genes are transcribed at any one time, and some portions of the DNA genome are never transcribed. The cell restricts the expression of genetic information to the formation of gene products needed at any particular moment. Specific regulatory sequences mark the beginning and end of the DNA segments to be transcribed and designate which strand in duplex DNA is to be used as the template. The transcript itself may interact with other RNA molecules as part of the overall regulatory program. The regulation of transcription is described in detail in Chapter 28.

The sum of all the RNA molecules produced in a cell under a given set of conditions is called the cellular **transcriptome**. Given the relatively small fraction of the human genome devoted to protein-encoding genes, we might expect that only a small part of the human genome is transcribed. This is not the case. Modern microarray analysis of transcription patterns has revealed that much of the genome of humans and other mammals is transcribed into RNA. The products are predominantly not mRNAs, tRNAs, or rRNAs, but rather special-function RNAs, a host of which are being discovered. Many of these seem to be involved in regulation of gene expression; however, the rapid pace of discovery has forced the realization that we do not yet know what many of these RNAs do.

In this chapter we examine the synthesis of RNA on a DNA template and the postsynthetic processing and turnover of RNA molecules. In doing so, we encounter many of the specialized functions of RNA, including

catalytic functions. Interestingly, the substrates for RNA enzymes are often other RNA molecules. We also describe systems in which RNA is the template and DNA the product, rather than vice versa. The information pathways thus come full circle, and reveal that template-dependent nucleic acid synthesis has standard rules, regardless of the nature of template or product (RNA or DNA). This examination of the biological interconversion of DNA and RNA as information carriers leads to a discussion of the evolutionary origin of biological information.

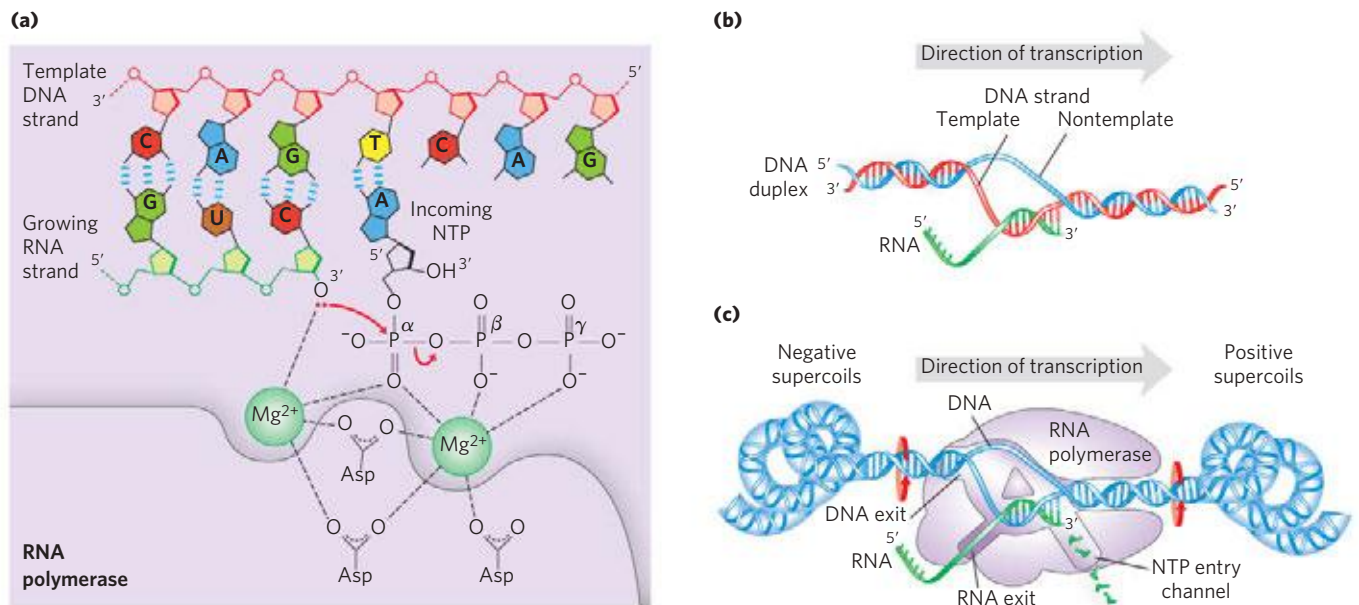
26.1 DNA-Dependent Synthesis of RNA

Our discussion of RNA synthesis begins with a comparison between transcription and DNA replication (Chapter 25). Transcription resembles replication in its fundamental chemical mechanism, its polarity (direction of synthesis), and its use of a template. And like replication, transcription has initiation, elongation, and termination phases—though in the literature on transcription, initiation is further divided into discrete

phases of DNA binding and initiation of RNA synthesis. Transcription differs from replication in that it does not require a primer and, generally, involves only limited segments of a DNA molecule. Additionally, within transcribed segments, only one DNA strand serves as a template for a particular RNA molecule.

RNA Is Synthesized by RNA Polymerases

The discovery of DNA polymerase and its dependence on a DNA template spurred a search for an enzyme that synthesizes RNA complementary to a DNA strand. By 1960, four research groups had independently detected an enzyme in cellular extracts that could form an RNA polymer from ribonucleoside 5'-triphosphates. Subsequent work on the purified *Escherichia coli* RNA polymerase helped to define the fundamental properties of transcription (**Fig. 26-1**). **DNA-dependent RNA polymerase** requires, in addition to a DNA template, all four ribonucleoside 5'-triphosphates (ATP, GTP, UTP, and CTP) as precursors of the nucleotide units of RNA, as well as Mg^{2+} . The protein also binds one Zn^{2+} . The chemistry and mechanism of RNA synthesis closely



MECHANISM FIGURE 26-1 Transcription by RNA polymerase in *E. coli*. For synthesis of an RNA strand complementary to one of two DNA strands in a double helix, the DNA is transiently unwound. **(a)** Catalytic mechanism of RNA synthesis by RNA polymerase. Note that this is essentially the same mechanism used by DNA polymerases (see Fig. 25-5a). The reaction involves two Mg^{2+} ions, coordinated to the phosphate groups of the incoming nucleoside triphosphates (NTPs) and to three Asp residues, which are highly conserved in the RNA polymerases of all species. One Mg^{2+} ion facilitates attack by the 3'-hydroxyl group on the α phosphate of the NTP; the other Mg^{2+} ion facilitates displacement of the pyrophosphate, and both metal ions stabilize the pentacovalent transition state.

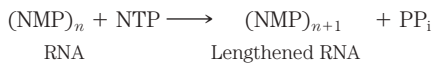
(b) About 17 bp of DNA are unwound at any given time. RNA polymerase and the transcription bubble move from left to right along the

DNA as shown, facilitating RNA synthesis. The DNA is unwound ahead and rewound behind as RNA is transcribed. As the DNA is rewound, the RNA-DNA hybrid is displaced and the RNA strand is extruded.

(c) Movement of an RNA polymerase along DNA tends to create positive supercoils (overwound DNA) ahead of the transcription bubble and negative supercoils (underwound DNA) behind it. The RNA polymerase is in close contact with the DNA ahead of the transcription bubble as well as with the separated DNA strands and the RNA within and immediately behind the bubble. A channel in the protein funnels new NTPs to the polymerase active site. The polymerase footprint encompasses about 35 bp of DNA during elongation.

(5') CGCTATAGCGTTT (3')	DNA nontemplate (coding) strand
(3') GCGATATCGCAA (5')	DNA template strand
(5') CGCUAUAGCGUUU (3')	RNA transcript

resemble those used by DNA polymerases (see Fig. 25–5a). RNA polymerase elongates an RNA strand by adding ribonucleotide units to the 3'-hydroxyl end, building RNA in the 5'→3' direction. The 3'-hydroxyl group acts as a nucleophile, attacking the α phosphate of the incoming ribonucleoside triphosphate (Fig. 26–1a) and releasing pyrophosphate. The overall reaction is



RNA polymerase requires DNA for activity and is most active when bound to a double-stranded DNA. As noted above, only one of the two DNA strands serves as a template. The template DNA strand is copied in the 3'→5' direction (antiparallel to the new RNA strand), just as in DNA replication. Each nucleotide in the newly formed RNA is selected by Watson-Crick base-pairing interactions: U residues are inserted in the RNA to pair with A residues in the DNA template, G residues are inserted to pair with C residues, and so on. Base-pair geometry (see Fig. 25–6) may also play a role in base selection.

Unlike DNA polymerase, RNA polymerase does not require a primer to initiate synthesis. Initiation occurs when RNA polymerase binds at specific DNA sequences called promoters (described below). The 5'-triphosphate group of the first residue in a nascent (newly formed) RNA molecule is not cleaved to release PP_i, but instead remains intact throughout the transcription process. During the elongation phase of transcription, the growing end of the new RNA strand base-pairs temporarily with the DNA template to form a short hybrid RNA-DNA double helix, estimated to be 8 bp long (Fig. 26–1b). The RNA in this hybrid duplex “peels off” shortly after its formation, and the DNA duplex re-forms.

FIGURE 26–2 Template and nontemplate (coding) DNA strands.

The two complementary strands of DNA are defined by their function in transcription. The RNA transcript is synthesized on the template strand and is identical in sequence (with U in place of T) to the nontemplate strand, or coding strand.

To enable RNA polymerase to synthesize an RNA strand complementary to one of the DNA strands, the DNA duplex must unwind over a short distance, forming a transcription “bubble.” During transcription, the *E. coli* RNA polymerase generally keeps about 17 bp unwound. The 8 bp RNA-DNA hybrid occurs in this unwound region. Elongation of a transcript by *E. coli* RNA polymerase proceeds at a rate of 50 to 90 nucleotides/s. Because DNA is a helix, movement of a transcription bubble requires considerable strand rotation of the nucleic acid molecules. DNA strand rotation is restricted in most DNAs by DNA-binding proteins and other structural barriers. As a result, a moving RNA polymerase generates waves of positive supercoils ahead of the transcription bubble and negative supercoils behind (Fig. 26–1c). This has been observed both in vitro and in vivo (in bacteria). In the cell, the topological problems caused by transcription are relieved through the action of topoisomerases (Chapter 24).

KEY CONVENTION: The two complementary DNA strands have different roles in transcription. The strand that serves as template for RNA synthesis is called the **template strand**. The DNA strand complementary to the template, the **nontemplate strand**, or **coding strand**, is identical in base sequence to the RNA transcribed from the gene, with U in the RNA in place of T in the DNA (Fig. 26–2). The coding strand for a particular gene may be located in either strand of a given chromosome (as shown in Fig. 26–3 for a virus). By convention, the regulatory sequences that control transcription (described later in this chapter) are designated by the sequences in the coding strand. ■

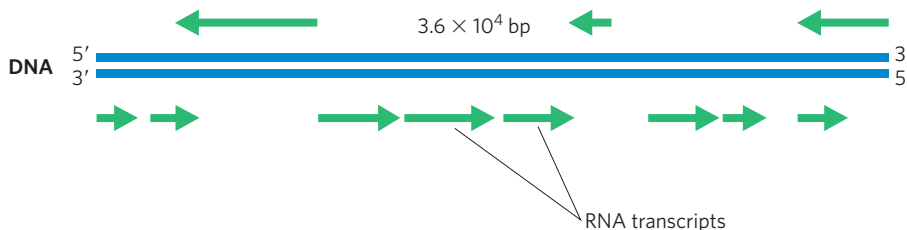


FIGURE 26–3 Organization of coding information in the adenovirus genome. The genetic information of the adenovirus genome (a conveniently simple example) is encoded by a double-stranded DNA molecule of 36,000 bp, both strands of which encode proteins. The information for most proteins is encoded by (that is, identical to) the top strand—by convention, the strand oriented 5' to 3' from left to right. The bottom strand acts as template for these transcripts. However, a few proteins are en-

coded by the bottom strand, which is transcribed in the opposite direction (and uses the top strand as template). Synthesis of mRNAs in adenovirus is actually much more complex than shown here. Many of the mRNAs derived using the upper strand as template are initially synthesized as a single, long transcript (25,000 nucleotides), which is then extensively processed to produce the separate mRNAs. Adenovirus causes upper respiratory tract infections in some vertebrates.

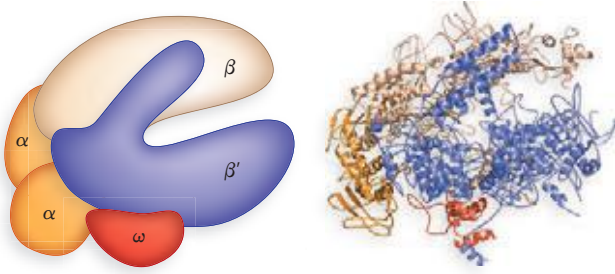


FIGURE 26-4 Structure of the RNA polymerase holoenzyme of the bacterium *Thermus aquaticus*. (Derived from PDB ID 1HQM) The overall structure of this enzyme is very similar to that of the *E. coli* RNA polymerase; no DNA or RNA is shown here. The several subunits of the bacterial RNA polymerase give the enzyme the shape of a crab claw. The pincers are formed from the large β and β' subunits. The subunits are shown in the same colors in the schematic and the ribbon structure.

The DNA-dependent RNA polymerase of *E. coli* is a large, complex enzyme with five core subunits ($\alpha_2\beta\beta'\omega$; M_r 390,000) and a sixth subunit, one of a group designated σ , with variants designated by size (molecular weight). The σ subunit binds transiently to the core and directs the enzyme to specific binding sites on the DNA (described below). These six subunits constitute the RNA polymerase holoenzyme (**Fig. 26-4**). The RNA polymerase holoenzyme of *E. coli* thus exists in several forms, depending on the type of σ subunit. The most common subunit is σ^{70} (M_r 70,000), and the upcoming discussion focuses on the corresponding RNA polymerase holoenzyme.

RNA polymerases lack a separate proofreading $3' \rightarrow 5'$ exonuclease active site (such as that of many DNA polymerases), and the error rate for transcription is higher than that for chromosomal DNA replication—approximately one error for every 10^4 to 10^5 ribonucleotides incorporated into RNA. Because many copies of an RNA are generally produced from a single gene and all RNAs are eventually degraded and replaced, a mistake in an RNA molecule is of less consequence to the cell than a mistake in the permanent information stored in DNA. Many RNA polymerases, including bacterial RNA polymerase and the eukaryotic RNA polymerase II (discussed below), do pause when a mismatched base is added during transcription, and they can remove mismatched nucleotides from the $3'$ end of a transcript by direct reversal of the polymerase reaction. But we do not yet know whether this activity is a true proofreading function and to what extent it may contribute to the fidelity of transcription.

RNA Synthesis Begins at Promoters

Initiation of RNA synthesis at random points in a DNA molecule would be an extraordinarily wasteful process. Instead, an RNA polymerase binds to specific sequences

in the DNA called **promoters**, which direct the transcription of adjacent segments of DNA (genes). The sequences where RNA polymerases bind can be quite variable, and much research has focused on identifying the particular sequences that are critical to promoter function.

In *E. coli*, RNA polymerase binding occurs within a region stretching from about 70 bp before the transcription start site to about 30 bp beyond it. By convention, the DNA base pairs that correspond to the beginning of an RNA molecule are given positive numbers, and those preceding the RNA start site are given negative numbers. The promoter region thus extends between positions -70 and $+30$. Analyses and comparisons of the most common class of bacterial promoters (those recognized by an RNA polymerase holoenzyme containing σ^{70}) have revealed similarities in two short sequences centered about positions -10 and -35 (**Fig. 26-5**). These sequences are important interaction sites for the σ^{70} subunit. Although the sequences are not identical for all bacterial promoters in this class, certain nucleotides that are particularly common at each position form a **consensus sequence** (recall the *E. coli* *oriC* consensus sequence; see Fig. 25-10). The consensus sequence at the -10 region is (5')TATAAT(3'); the consensus sequence at the -35 region is (5')TTGACA(3'). A third AT-rich recognition element, called the UP (upstream promoter) element, occurs between positions -40 and -60 in the promoters of certain highly expressed genes. The UP element is bound by the α subunit of RNA polymerase. The efficiency with which an RNA polymerase containing σ^{70} binds to a promoter and initiates transcription is determined in large measure by these sequences, the spacing between them, and their distance from the transcription start site.

Many independent lines of evidence attest to the functional importance of the sequences in the -35 and -10 regions. Mutations that affect the function of a given promoter often involve a base pair in these regions. Variations in the consensus sequence also affect the efficiency of RNA polymerase binding and transcription initiation. A change in only one base pair can decrease the rate of binding by several orders of magnitude. The promoter sequence thus establishes a basal level of expression that can vary greatly from one *E. coli* gene to the next. A method that provides information about the interaction between RNA polymerase and promoters is illustrated in Box 26-1.

The pathway of transcription initiation and the fate of the σ subunit are becoming much better defined (**Fig. 26-6**). The pathway consists of two major parts, binding and initiation, each with multiple steps. First, the polymerase, directed by its bound σ factor, binds to the promoter. A closed complex (in which the bound DNA is intact) and an open complex (in which the bound DNA is intact and partially unwound near the -10 sequence) form in succession. Second, transcription

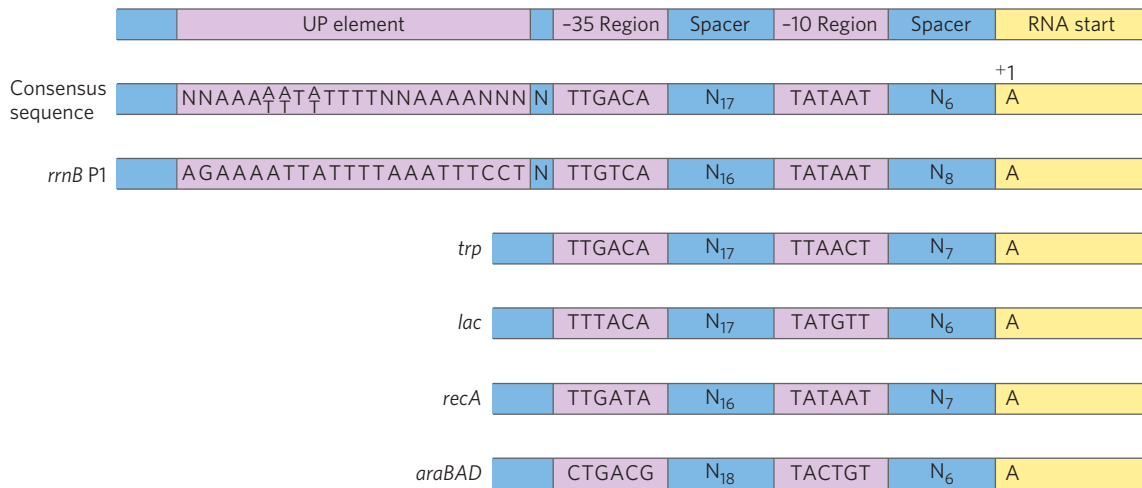


FIGURE 26-5 Typical *E. coli* promoters recognized by an RNA polymerase holoenzyme containing σ^{70} . Sequences of the nontemplate strand are shown, read in the 5'→3' direction, as is the convention for representations of this kind. The sequences differ from one promoter to the next, but comparisons of many promoters reveal similarities, particularly in the -10 and -35 regions. The sequence element UP, not present in all *E. coli* promoters, is shown in the P1 promoter for the highly expressed rRNA

gene *rrnB*. UP elements, generally occurring in the region between -40 and -60, strongly stimulate transcription at the promoters that contain them. The UP element in the *rrnB* P1 promoter encompasses the region between -38 and -59. The consensus sequence for *E. coli* promoters recognized by σ^{70} is shown second from the top. Spacer regions contain slightly variable numbers of nucleotides (N). Only the first nucleotide coding the RNA transcript (at position +1) is shown.

is initiated within the complex, leading to a conformational change that converts the complex to the elongation form, followed by movement of the transcription complex away from the promoter (promoter clearance). Any of these steps can be affected by the specific makeup of the promoter sequences. The σ subunit dissociates stochastically (at random) as the polymerase enters the elongation phase of transcription. The protein NusA (M_r 54,430) binds to the elongating RNA polymerase, competitively with the σ subunit. Once transcription is complete, NusA dissociates from the enzyme, the RNA polymerase dissociates from the DNA, and a σ factor (σ^{70} or another) can again bind to the enzyme to initiate transcription.

E. coli has other classes of promoters, bound by RNA polymerase holoenzymes with different σ subunits (Table 26-1). An example is the promoters of the heat shock genes. The products of this set of genes are made at higher levels when the cell has received an insult, such as a sudden increase in temperature. RNA polymerase binds to the promoters of these genes only when σ^{70} is replaced with the σ^{32} (M_r 32,000) subunit, which is specific for the heat shock promoters (see Fig. 28-3). By using different σ subunits, the cell can coordinate the expression of sets of genes, permitting major changes in cell physiology. Which sets of genes are expressed is determined by the availability of the various σ subunits, which is determined by several factors: regulated rates of synthesis and degradation, postsynthetic modifications that switch individual σ subunits between active

and inactive forms, and a specialized class of anti- σ proteins, each type binding to and sequestering a particular σ subunit (rendering it unavailable for transcription initiation).

Transcription Is Regulated at Several Levels

Requirements for any gene product vary with cellular conditions or developmental stage, and transcription of each gene is carefully regulated to form gene products only in the proportions needed. Regulation can occur at any step in transcription, including elongation and termination. However, much of the regulation is directed at the polymerase binding and transcription initiation steps outlined in Figure 26-6. Differences in promoter sequences are just one of several levels of control.

The binding of proteins to sequences both near to and distant from the promoter can also affect levels of gene expression. Protein binding can *activate* transcription by facilitating either RNA polymerase binding or steps further along in the initiation process, or it can *repress* transcription by blocking the activity of the polymerase. In *E. coli*, one protein that activates transcription is the **cAMP receptor protein (CRP)**, which increases the transcription of genes coding for enzymes that metabolize sugars other than glucose when cells are grown in the absence of glucose. **Repressors** are proteins that block the synthesis of RNA at specific genes. In the case of the Lac repressor (Chapter 28), transcription of the genes for the enzymes of lactose metabolism is blocked when lactose is unavailable.

BOX 26-1 METHODS RNA Polymerase Leaves Its Footprint on a Promoter

Footprinting, a technique derived from principles used in DNA sequencing, identifies the DNA sequences bound by a particular protein. Researchers isolate a DNA fragment thought to contain sequences recognized by a DNA-binding protein and radiolabel one end of one strand (Fig. 1). They then use chemical or enzymatic reagents to introduce random breaks in the DNA fragment (averaging about one per molecule).

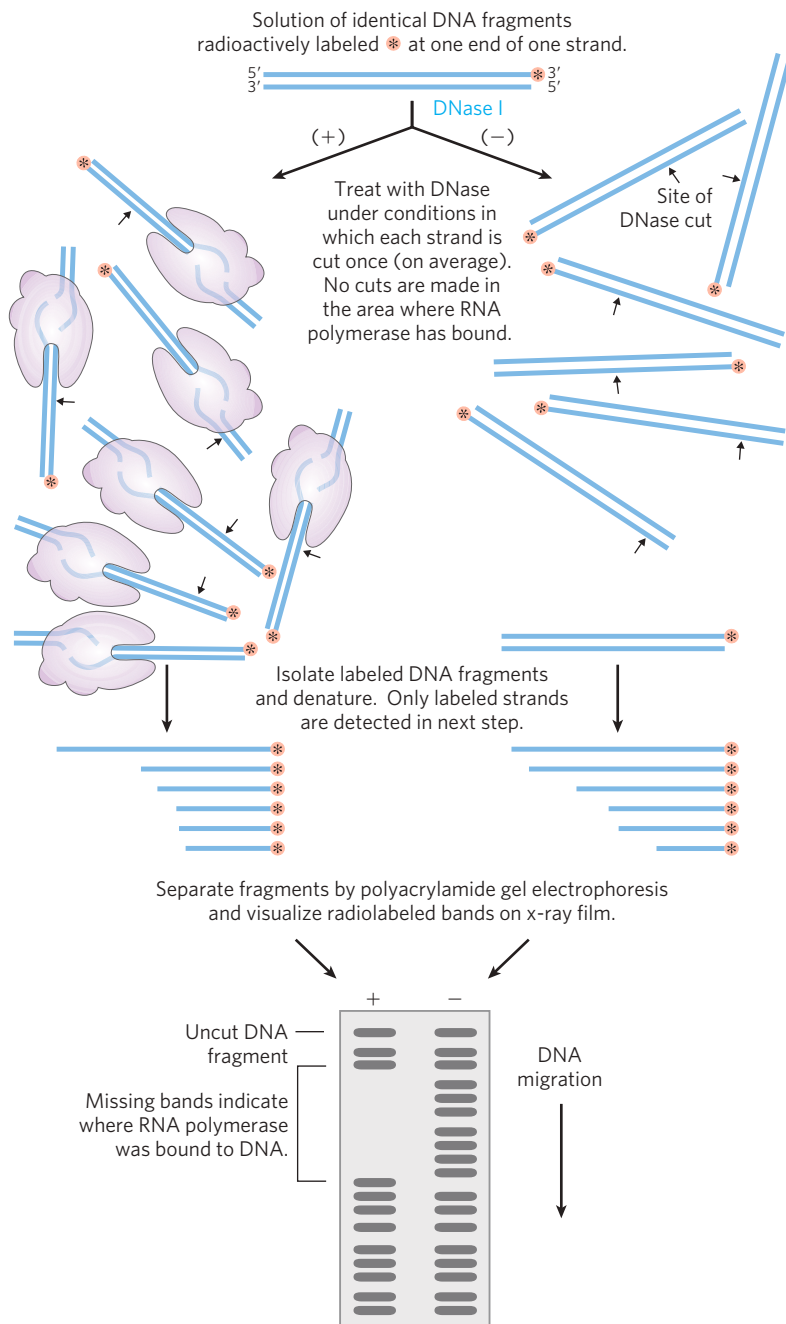


FIGURE 1 Footprint analysis of the RNA polymerase-binding site on a DNA fragment. Separate experiments are carried out in the presence (+) and absence (-) of the polymerase.

Separation of the labeled cleavage products (broken fragments of various lengths) by high-resolution electrophoresis produces a ladder of radioactive bands. In a separate tube, the cleavage procedure is repeated on copies of the same DNA fragment in the presence of the DNA-binding protein. The researchers then subject the two sets of cleavage products to electrophoresis and compare them side by side. A gap (“footprint”) in the series of radioactive bands derived from the DNA-protein sample, attributable to protection of the DNA by the bound protein, identifies the sequences that the protein binds.

The precise location of the protein-binding site can be determined by directly sequencing (see Fig. 8-34) copies of the same DNA fragment and including the sequencing lanes (not shown here) on the same gel with the footprint. Figure 2 shows footprinting results for the binding of RNA polymerase to a DNA fragment containing a promoter. The polymerase covers 60 to 80 bp; protection by the bound enzyme includes the -10 and -35 regions.

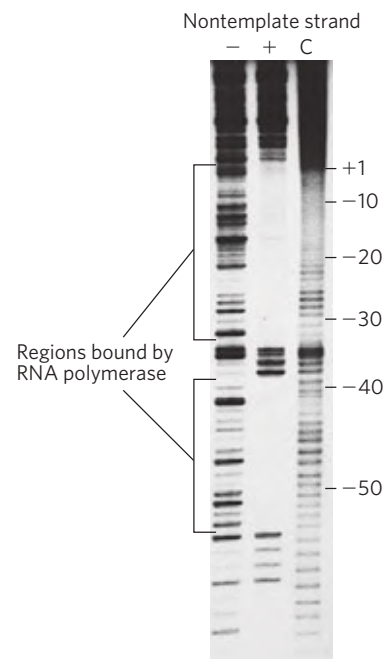


FIGURE 2 Footprinting results of RNA polymerase binding to the *lac* promoter (see Fig. 26-5). In this experiment, the 5' end of the nontemplate strand was radioactively labeled. Lane C is a control in which the labeled DNA fragments were cleaved with a chemical reagent that produces a more uniform banding pattern.

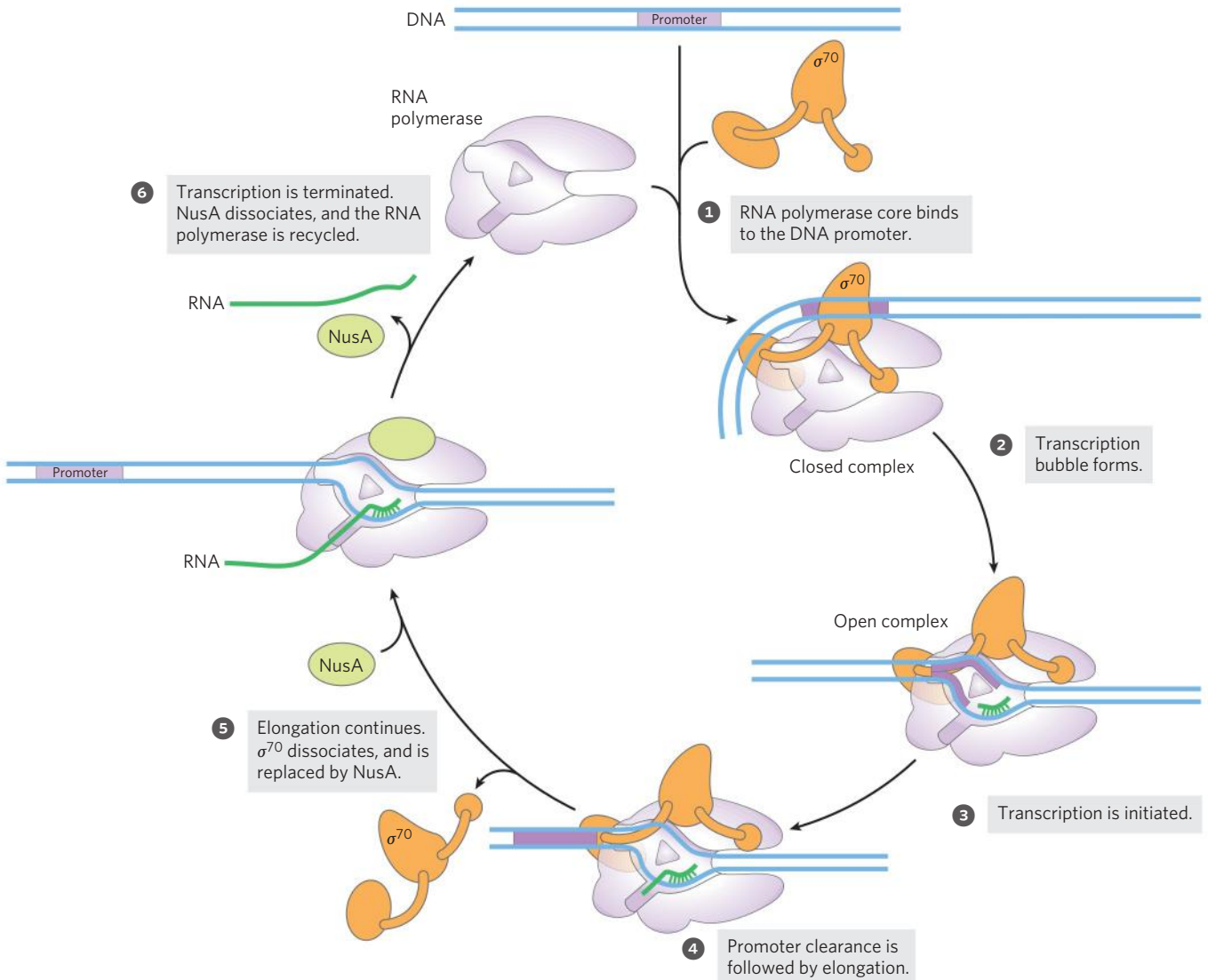


FIGURE 26-6 Transcription initiation and elongation by *E. coli* RNA polymerase. Initiation of transcription requires several steps generally divided into two phases, binding and initiation. In the binding phase, the initial interaction of the RNA polymerase with the promoter leads to formation of a closed complex, in which the promoter DNA is stably bound but not unwound. A 12 to 15 bp region of DNA—from within the -10 region to position $+2$ or $+3$ —is then unwound to form an open complex. Additional intermediates (not shown) have been detected in the pathways leading to the closed and open complexes, along with several

Transcription is the first step in the complicated and energy-intensive pathway of protein synthesis, so much of the regulation of protein levels in both bacterial and eukaryotic cells is directed at transcription, particularly its early stages. In Chapter 28 we describe many mechanisms by which this regulation is accomplished.

Specific Sequences Signal Termination of RNA Synthesis

RNA synthesis is processive; that is, the RNA polymerase will introduce a large number of nucleotides

changes in protein conformation. The initiation phase encompasses transcription initiation and promoter clearance (steps 1 through 4 here). Once elongation commences, the σ subunit is released and it is replaced by the protein NusA. The polymerase leaves the promoter and becomes committed to elongation of the RNA (step 5). When transcription is complete, the RNA is released, the NusA protein dissociates, and the RNA polymerase dissociates from the DNA (step 6). Another σ subunit binds to the RNA polymerase and the process begins again.

into a growing RNA molecule before dissociating (p. 1014). This is necessary because, if an RNA polymerase released an RNA transcript prematurely, it could not resume synthesis of the same RNA but instead would have to start again. However, an encounter with certain DNA sequences results in a pause in RNA synthesis, and at some of these sequences transcription is terminated. Our focus here is again on the well-studied systems in bacteria. *E. coli* has at least two classes of termination signals: one class relies on a protein factor called ρ (rho) and the other is ρ -independent.

TABLE 26-1 The Seven σ Subunits of *Escherichia coli*

σ subunit	K_d (nM)	Molecules/cell*	Holoenzyme ratio (%)*	Function
σ^{70}	0.26	700	78	Housekeeping
σ^{54}	0.30	110	8	Modulation of cellular nitrogen levels
σ^{38}	4.26	<1	0	Stationary phase genes
σ^{32}	1.24	<10	0	Heat shock genes
σ^{28}	0.74	370	14	Flagella and chemotaxis genes
σ^{24}	2.43	<10	0	Extracytoplasmic functions; some heat shock functions
σ^{18}	1.73	<1	0	Extracytoplasmic functions, including ferric citrate transport

Source: Adapted from Maeda, H., Fujita, N., & Ishihama, A. (2000) *Nucleic Acids Res.* 28, 3500.

Note: σ factors are widely distributed in bacteria; the number varies from a single σ factor in *Mycoplasma genitalium* to 63 distinct σ factors in *Streptomyces coelicolor*.

*Approximate number of each σ subunit per cell and the fraction of RNA polymerase holoenzyme complexed with each σ subunit during exponential growth. The numbers change as growth conditions change. The fraction of RNA polymerase complexed with each σ subunit reflects both the amount of the particular subunit and its affinity for the enzyme.

Most ρ -independent terminators have two distinguishing features. The first is a region that produces an RNA transcript with self-complementary sequences, permitting the formation of a hairpin structure (see Fig. 8-19a) centered 15 to 20 nucleotides before the projected end of the RNA strand. The second feature is a highly conserved string of three A residues in the template strand that are transcribed into U residues near the 3' end of the hairpin. When a polymerase arrives at a termination site with this structure, it pauses (Fig. 26-7a). Formation of the hairpin structure in the RNA disrupts several A=U base pairs in the RNA-DNA hybrid segment and may disrupt important interactions between RNA and the RNA polymerase, facilitating dissociation of the transcript.

The ρ -dependent terminators lack the sequence of repeated A residues in the template strand but usually include a CA-rich sequence called a *rut* (*rho* utilization) element. The ρ protein associates with the RNA at specific binding sites and migrates in the 5'→3' direction until it reaches the transcription complex that is paused at a termination site (Fig. 26-7b). Here it contributes to release of the RNA transcript. The ρ protein has an ATP-dependent RNA-DNA helicase activity that promotes translocation of the protein along the RNA, and ATP is hydrolyzed by ρ protein during the termination process. The detailed mechanism by which the protein promotes the release of the RNA transcript is not known.

Eukaryotic Cells Have Three Kinds of Nuclear RNA Polymerases

The transcriptional machinery in the nucleus of a eukaryotic cell is much more complex than that in bacteria. Eukaryotes have three RNA polymerases, designated I, II, and III, which are distinct complexes but

have certain subunits in common. Each polymerase has a specific function and is recruited to a specific promoter sequence.

RNA polymerase I (Pol I) is responsible for the synthesis of only one type of RNA, a transcript called preribosomal RNA (or pre-rRNA), which contains the precursor for the 18S, 5.8S, and 28S rRNAs (see Fig. 26-24). Pol I promoters differ greatly in sequence from one species to another. The principal function of RNA polymerase II (Pol II) is synthesis of mRNAs and some specialized RNAs. This enzyme can recognize thousands of promoters that vary greatly in sequence. Some Pol II promoters have a few sequence features in common, including a TATA box (eukaryotic consensus sequence TATA(A/T)A(A/T)(A/G)) near base pair -30 and an Inr sequence (initiator) near the RNA start site at +1 (Fig. 26-8). However, such promoters are in the minority, and Pol II operates at many promoters that lack these features.

RNA polymerase III (Pol III) makes tRNAs, the 5S rRNA, and some other small specialized RNAs. The promoters recognized by Pol III are well characterized. Interestingly, some of the sequences required for the regulated initiation of transcription by Pol III are located within the gene itself, whereas others are in more conventional locations upstream of the RNA start site (Chapter 28).

RNA Polymerase II Requires Many Other Protein Factors for Its Activity

RNA polymerase II is central to eukaryotic gene expression and has been studied extensively. Although this polymerase is strikingly more complex than its bacterial counterpart, the complexity masks a remarkable conservation of structure, function, and mechanism. Pol II isolated from yeast is a huge enzyme with 12 subunits. The largest subunit (RBP1) exhibits a high degree of homology to the β' subunit of bacterial RNA polymerase.

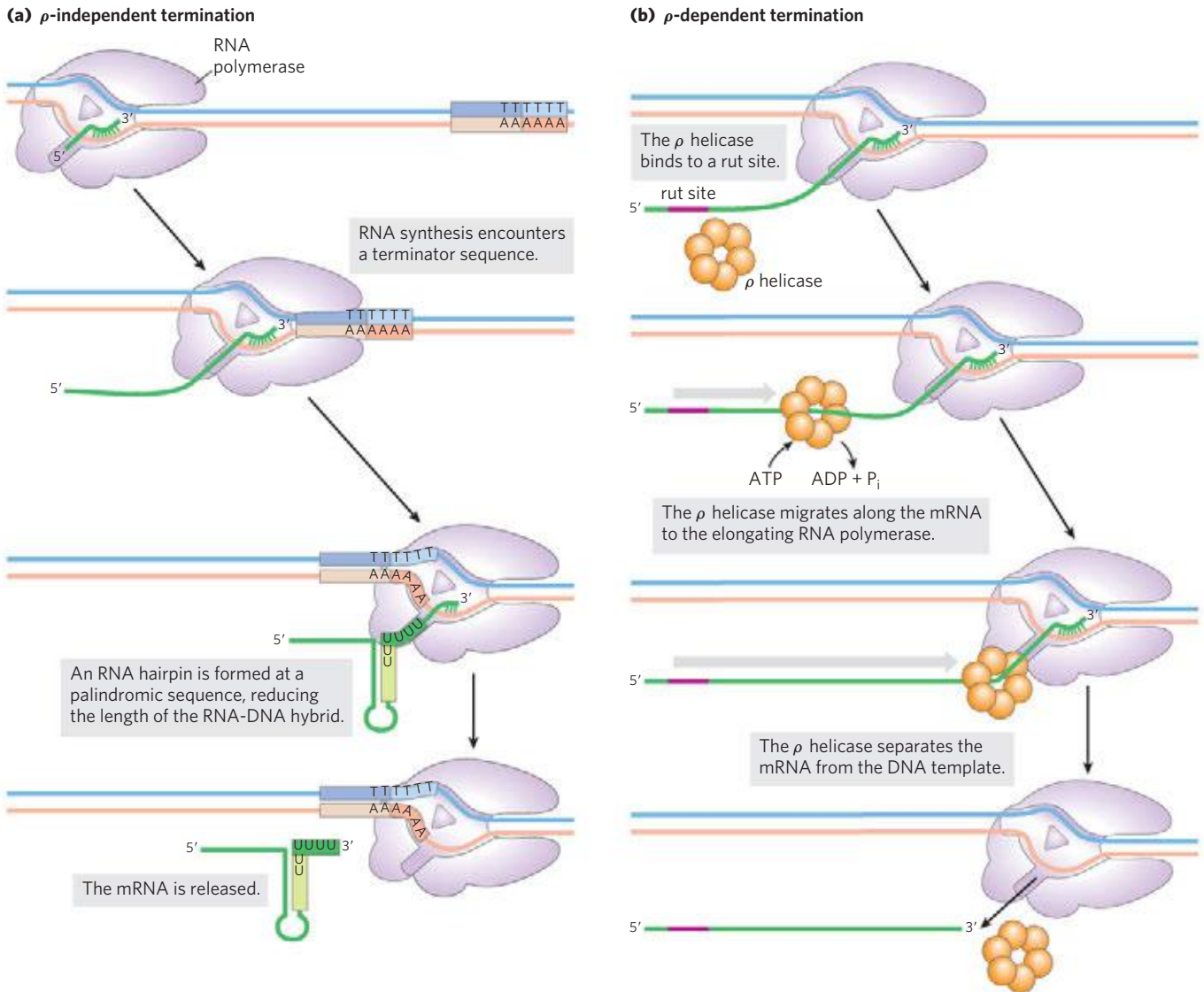


FIGURE 26-7 Termination of transcription in *E. coli*. (a) ρ -Independent termination. RNA polymerase pauses at a variety of DNA sequences, some of which are terminators. One of two outcomes is then possible: the polymerase bypasses the site and continues on its way, or the complex undergoes a conformational change (isomerization). In the latter case, intramolecular pairing of complementary sequences in the newly formed RNA transcript may form a hairpin that disrupts the RNA-DNA hybrid or the interactions between RNA and polymerase, or

both, resulting in isomerization. An A=U hybrid region at the 3' end of the new transcript is relatively unstable, and the RNA dissociates from the complex completely, leading to termination. This is the usual outcome at terminators. At other pause sites, the complex may escape after the isomerization step to continue RNA synthesis. (b) ρ -Dependent termination. RNAs that include a rut site (purple) recruit the ρ helicase. The ρ helicase migrates along the mRNA in the 5'→3' direction and separates it from the polymerase.

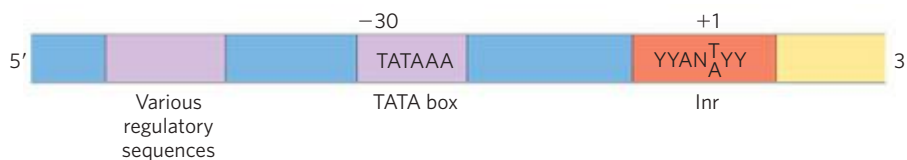


FIGURE 26-8 Some common sequences in promoters recognized by eukaryotic RNA polymerase II. The TATA box is the major assembly point for the proteins of the preinitiation complexes of Pol II. The DNA is unwound at the initiator sequence (Inr), and the transcription start site is usually within or very near this sequence. In the Inr consensus sequence shown here, N represents any nucleotide; Y, a pyrimidine nucleotide. Many additional sequences serve as binding sites for a wide variety of proteins that affect the activity of Pol II. These sequences are important in regulating Pol II promoters and differ greatly in type and number, and

in general the eukaryotic promoter is much more complex than suggested here (see Fig. 15-25). Many of the sequences are located within a few hundred base pairs of the TATA box on the 5' side; others may be thousands of base pairs away. The sequence elements summarized here are more variable among the Pol II promoters of eukaryotes than among the *E. coli* promoters (see Fig. 26-5). The majority of Pol II promoters lack a TATA box or a consensus Inr element or both. Additional sequences around the TATA box and downstream (to the right as drawn) of Inr may be recognized by one or more transcription factors.

TABLE 26–2 Proteins Required for Initiation of Transcription at the RNA Polymerase II (Pol II) Promoters of Eukaryotes

Transcription protein	Number of subunits	Subunit(s) M_r	Function(s)
Initiation			
Pol II	12	10,000–220,000	Catalyzes RNA synthesis
TBP (TATA-binding protein)	1	38,000	Specifically recognizes the TATA box
TFIIA	3	12,000, 19,000, 35,000	Stabilizes binding of TFIIB and TBP to the promoter
TFIIB	1	35,000	Binds to TBP; recruits Pol II–TFIIF complex
TFIIE	2	34,000, 57,000	Recruits TFIIH; has ATPase and helicase activities
TFIIF	2	30,000, 74,000	Binds tightly to Pol II; binds to TFIIB and prevents binding of Pol II to nonspecific DNA sequences
TFIIH	12	35,000–89,000	Unwinds DNA at promoter (helicase activity); phosphorylates Pol II (within the CTD); recruits nucleotide-excision repair proteins
Elongation*			
ELL [†]	1	80,000	
pTEFb	2	43,000, 124,000	Phosphorylates Pol II (within the CTD)
SII (TFIIS)	1	38,000	
Elongin (SIII)	3	15,000, 18,000, 110,000	

*The function of all elongation factors is to suppress the pausing or arrest of transcription by the Pol II–TFIIF complex.

[†]Name derived from eleven–nineteen /lysine–rich /leukemia. The gene for ELL is the site of chromosomal recombination events frequently associated with acute myeloid leukemia.

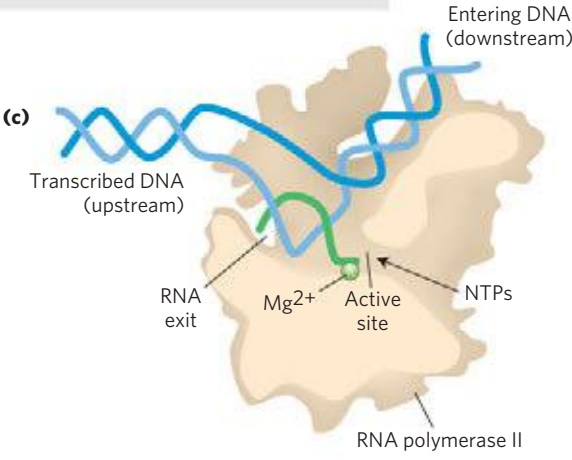
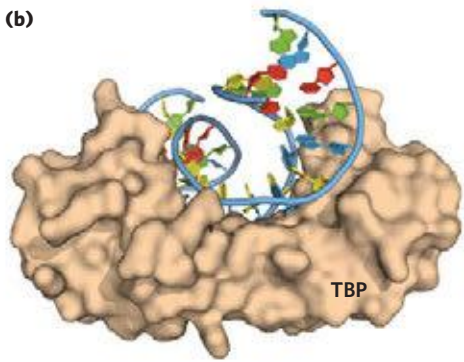
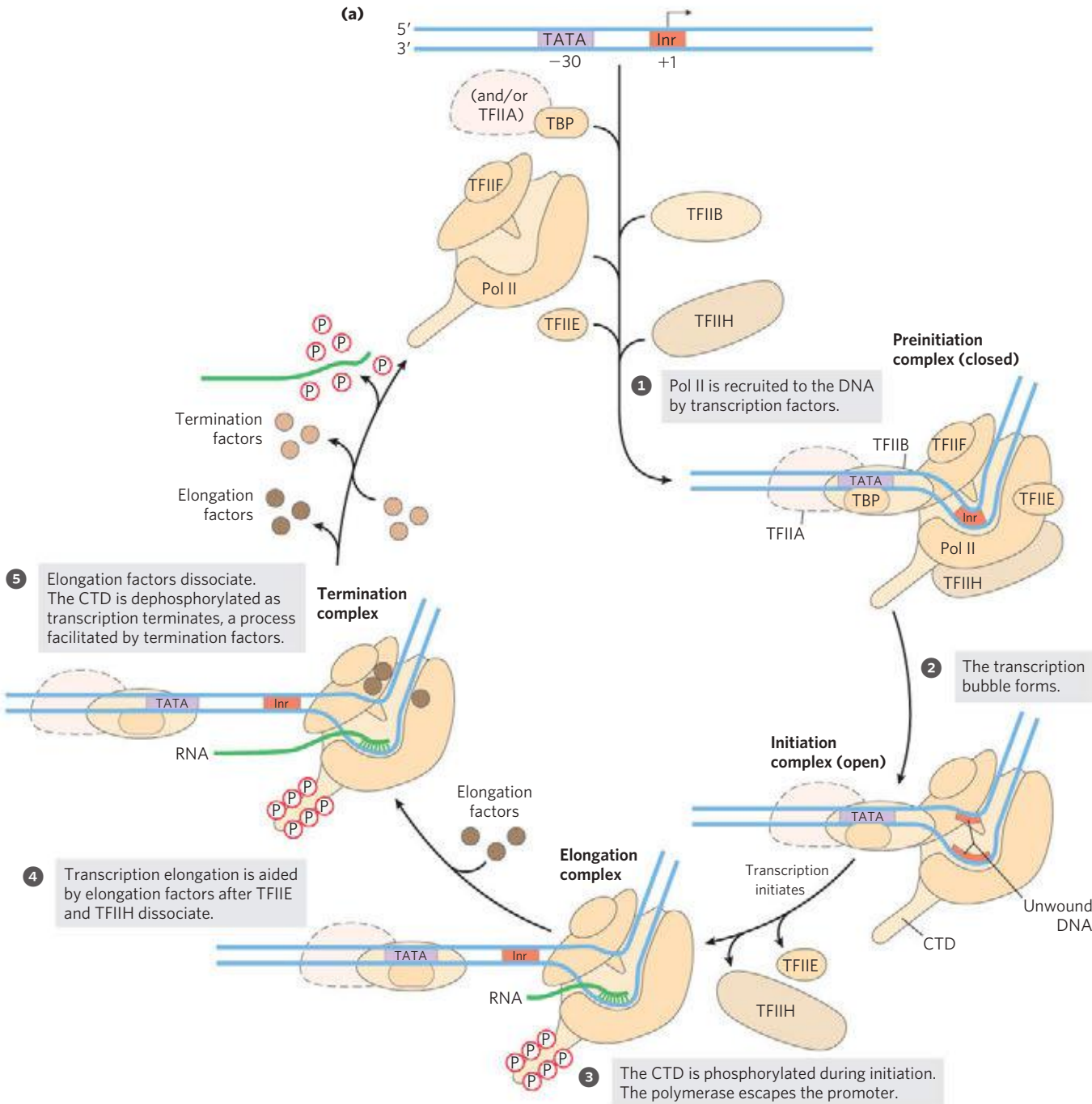
Another subunit (RBP2) is structurally similar to the bacterial β subunit, and two others (RBP3 and RBP11) show some structural homology to the two bacterial α subunits. Pol II must function with genomes that are more complex and with DNA molecules more elaborately packaged than in bacteria. The need for protein–protein contacts with the numerous other protein factors required to navigate this labyrinth accounts in large measure for the added complexity of the eukaryotic polymerase.

The largest subunit of Pol II (RBP1) also has an unusual feature, a long carboxyl-terminal tail consisting of many repeats of a consensus heptad amino acid sequence –YSPTSPS–. There are 27 repeats in the yeast enzyme (18 exactly matching the consensus) and 52 (21 exact) in the mouse and human enzymes. This carboxyl-terminal domain (CTD) is separated from the main body of the enzyme by an inherently unstructured linker sequence. The CTD has many important roles in Pol II function, as outlined below.

RNA polymerase II requires an array of other proteins, called **transcription factors**, in order to form the active transcription complex. The **general transcription factors** required at every Pol II promoter (factors usually designated TFII with an additional iden-

tifier) are highly conserved in all eukaryotes (Table 26–2). The process of transcription by Pol II can be described in terms of several phases—assembly, initiation, elongation, termination—each associated with characteristic proteins (**Fig. 26–9**). The step-by-step pathway described below leads to active transcription in vitro. In the cell, many of the proteins may be present in larger, preassembled complexes, simplifying the pathways for assembly on promoters. As you read about this process, consult Figure 26–9 and Table 26–2 to help keep track of the many participants.

FIGURE 26–9 Transcription at RNA polymerase II promoters. **(a)** The sequential assembly of TBP (often with TFIIA), TFIIB, TFIIF plus Pol II, TFIIE, and TFIIH results in a closed complex. Within the complex, the DNA is unwound at the Inr region by the helicase activity of TFIIH and perhaps of TFIIE, creating an open complex. The carboxyl-terminal domain of the largest Pol II subunit is phosphorylated by TFIIH, and the polymerase then escapes the promoter and begins transcription. Elongation is accompanied by the release of many transcription factors and is also enhanced by elongation factors (see Table 26–2). After termination, Pol II is released, dephosphorylated, and recycled. **(b)** Human TBP bound to DNA. The DNA is bent in this complex, opening the minor groove to allow specific hydrogen-bonding between protein and DNA (PDB ID 1TGH). **(c)** A cut-away view of transcription elongation promoted by the Pol II core enzyme.



Assembly of RNA Polymerase and Transcription Factors at a Promoter The formation of a closed complex begins when the TATA-binding protein (TBP) binds to the TATA box (Figs 26–9a, step ❶, and 26–9b). At promoters lacking a TATA box, TBP arrives as part of a multisubunit complex called TFIID (not shown in Fig. 26–9). The sequence elements that direct the binding of TFIID at TATA-less promoters are poorly understood. TBP is bound in turn by the transcription factor TFIIB, which also binds to DNA on either side of TBP. TFIIA binds, and along with TFIIB helps to stabilize the TBP-DNA complex. TFIIB provides an important link to DNA polymerase II, and the TFIIB-TBP complex is next bound by another complex consisting of TFIIF and Pol II. TFIIF helps target Pol II to its promoters, both by interacting with TFIIB and by reducing the binding of the polymerase to nonspecific sites on the DNA. Finally, TFIIE and TFIIH bind to create the closed complex. TFIIH has multiple subunits and includes a DNA helicase activity that promotes the unwinding of DNA near the RNA start site (a process requiring the hydrolysis of ATP), thereby creating an open complex (Fig. 26–9a, step ❷). Counting all the subunits of the various essential factors (excluding TFIIA and some subunits of TFIID), this minimal active assembly has more than 30 polypeptides. Structural studies by Roger Kornberg and his collaborators have provided a more detailed look at the core structure of RNA polymerase II during elongation (Fig. 26–9c).

RNA Strand Initiation and Promoter Clearance TFIIH has an additional function during the initiation phase. A kinase activity in one of its subunits phosphorylates Pol II at many places in the CTD (Fig. 26–9a). Several other protein kinases, including CDK9 (cyclin-dependent kinase 9), which is part of the complex pTEFb (*p*ositive *t*ranscription *e*longation *f*actor *b*), also phosphorylate the CTD, primarily on the Ser residues of the CTD repeat sequence. This causes a conformational change in the overall complex, initiating transcription. Phosphorylation of the CTD is also important during the subsequent elongation phase, with the phosphorylation state of the CTD changing as transcription proceeds. The changes affect the interactions between the transcription complex and other proteins and enzymes, such that different sets of proteins are bound at initiation than at later stages. Some of these proteins are involved in processing the transcript (as described below).

During synthesis of the initial 60 to 70 nucleotides of RNA, first TFIIE and then TFIIH is released, and Pol II enters the elongation phase of transcription.

Elongation, Termination, and Release TFIIF remains associated with Pol II throughout elongation. During this stage, the activity of the polymerase is greatly enhanced by proteins called elongation factors (Table 26–2). The elongation factors, some bound to the phosphorylated

CTD, suppress pausing during transcription and also coordinate interactions between protein complexes involved in the posttranscriptional processing of mRNAs. Once the RNA transcript is completed, transcription is terminated. Pol II is dephosphorylated and recycled, ready to initiate another transcript (Fig. 26–9a, steps ❸ to ❺).

Regulation of RNA Polymerase II Activity Regulation of transcription at Pol II promoters is quite elaborate. It involves the interaction of a wide variety of other proteins with the preinitiation complex. Some of these regulatory proteins interact with transcription factors, others with Pol II itself. The regulation of transcription is described in more detail in Chapter 28.

Diverse Functions of TFIIH In eukaryotes, the repair of damaged DNA (see Table 25–5) is more efficient within genes that are actively being transcribed than for other damaged DNA, and the template strand is repaired somewhat more efficiently than the nontemplate strand. These remarkable observations are explained by the alternative roles of the TFIIH subunits. Not only does TFIIH participate in formation of the closed complex during assembly of a transcription complex (as described above), but some of its subunits are also essential components of the separate nucleotide-excision repair complex (see Fig. 25–25).



When Pol II transcription halts at the site of a DNA lesion, TFIIH can interact with the lesion and recruit the entire nucleotide-excision repair complex. Genetic loss of certain TFIIH subunits can produce human diseases. Some examples are xeroderma pigmentosum (see Box 25–1) and Cockayne syndrome, which is characterized by arrested growth, photosensitivity, and neurological disorders. ■

DNA-Dependent RNA Polymerase Undergoes Selective Inhibition

The elongation of RNA strands by RNA polymerase in both bacteria and eukaryotes is inhibited by the antibiotic **actinomycin D** (Fig. 26–10). The planar portion of this molecule inserts (intercalates) into the double-helical DNA between successive G≡C base pairs, deforming the DNA. This prevents movement of the polymerase along the template. Because actinomycin D inhibits RNA elongation in intact cells as well as in cell extracts, it is used to identify cell processes that depend on RNA synthesis. **Acridine** inhibits RNA synthesis in a similar fashion.

Rifampicin inhibits bacterial RNA synthesis by binding to the β subunit of bacterial RNA polymerases, preventing the promoter clearance step of transcription (Fig. 26–6). It is sometimes used as an antibiotic.

The mushroom *Amanita phalloides* has a very effective defense mechanism against predators. It produces **α -amanitin**, which disrupts mRNA formation in animal

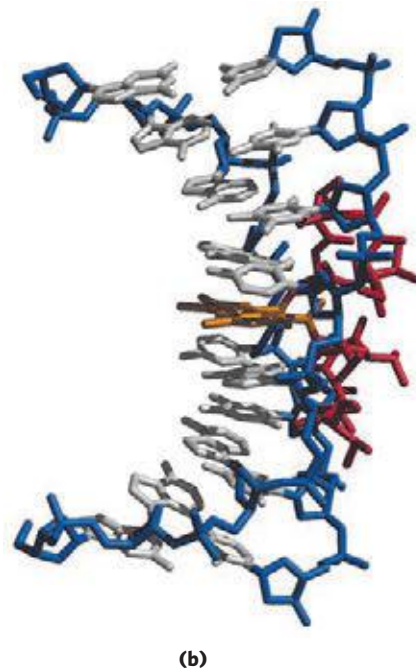
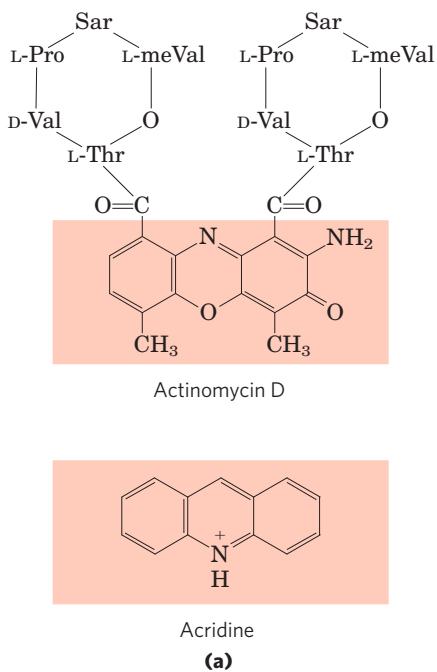


FIGURE 26-10 Actinomycin D and acridine, inhibitors of DNA transcription. (a) The shaded portion of actinomycin D is planar and intercalates between two successive G≡C base pairs in duplex DNA. The two cyclic peptide moieties of actinomycin D bind to the minor groove of the double helix. Sarcosine (Sar) is *N*-methylglycine; meVal is methylvaline. Acridine also acts by intercalation in DNA. (b) A complex of actinomycin D with DNA (PDB ID 1DSC). The DNA backbone is shown in blue, the bases are white, the intercalated part of actinomycin (shaded in (a)) is orange, and the remainder of the actinomycin is red. The DNA is bent as a result of the actinomycin binding.

cells by blocking Pol II and, at higher concentrations, Pol III. Neither Pol I nor bacterial RNA polymerase is sensitive to α -amanitin—nor is the RNA polymerase II of *A. phalloides* itself!

SUMMARY 26.1 DNA-Dependent Synthesis of RNA

- ▶ Transcription is catalyzed by DNA-dependent RNA polymerases, which use ribonucleoside 5'-triphosphates to synthesize RNA complementary to the template strand of duplex DNA. Transcription occurs in several phases: binding of RNA polymerase to a DNA site called a promoter, initiation of transcript synthesis, elongation, and termination.
- ▶ Bacterial RNA polymerase requires a special subunit to recognize the promoter. As the first committed step in transcription, binding of RNA polymerase to the promoter and initiation of transcription are closely regulated. Transcription stops at sequences called terminators.
- ▶ Eukaryotic cells have three types of RNA polymerases. Binding of RNA polymerase II to its promoters requires an array of proteins called transcription factors. Elongation factors participate in the elongation phase of transcription. The largest subunit of Pol II has a long carboxyl-terminal domain, which is phosphorylated during the initiation and elongation phases.

26.2 RNA Processing

Many of the RNA molecules in bacteria and virtually all RNA molecules in eukaryotes are processed to some degree after synthesis. Some of the most interesting

molecular events in RNA metabolism occur during this postsynthetic processing. Intriguingly, several of the enzymes that catalyze these reactions consist of RNA rather than protein. The discovery of these catalytic RNAs, or **ribozymes**, has brought a revolution in thinking about RNA function and about the origin of life.

A newly synthesized RNA molecule is called a **primary transcript**. Perhaps the most extensive processing of primary transcripts occurs in eukaryotic mRNAs and in tRNAs of both bacteria and eukaryotes. Special-function RNAs are also processed.

The primary transcript for a eukaryotic mRNA typically contains sequences encompassing one gene, although the sequences encoding the polypeptide may not be contiguous. Noncoding tracts that break up the coding region of the transcript are called introns, and the coding segments are called exons (see the discussion of introns and exons in DNA in Chapter 24). In a process called **RNA splicing**, the introns are removed from the primary transcript and the exons are joined to form a continuous sequence that specifies a functional polypeptide. Eukaryotic mRNAs are also modified at each end. A modified residue called a 5' cap is added at the 5' end. The 3' end is cleaved, and 80 to 250 A residues are added to create a poly(A) "tail." The sometimes elaborate protein complexes that carry out each of these three mRNA-processing reactions do not operate independently. They seem to be organized in association with each other and with the phosphorylated CTD of Pol II; each complex affects the function of the others. Proteins involved in mRNA transport to the cytoplasm are also associated with the mRNA in the nucleus, and the processing of the transcript is coupled to its transport. In effect, a eukaryotic mRNA, as it is synthesized,

is ensconced in an elaborate complex involving dozens of proteins. The composition of the complex changes as the primary transcript is processed, transported to the cytoplasm, and delivered to the ribosome for translation. The associated proteins modulate all aspects of the function and fate of the mRNA. These processes are outlined in **Figure 26–11** and described in more detail below.

The primary transcripts of bacterial and eukaryotic tRNAs are processed by the removal of sequences from each end (cleavage) and in a few cases by the removal of introns (splicing). Many bases and sugars in tRNAs are also modified; mature tRNAs are replete with unusual bases not found in other nucleic acids (see Fig. 26–22). Many of the special-function RNAs also undergo elaborate processing, often involving the removal of segments from one or both ends.

The ultimate fate of any RNA is its complete and regulated degradation. The rate of turnover of RNAs plays a critical role in determining their steady-state levels and the rate at which cells can shut down expression of a gene whose product is no longer needed. During the development of multicellular organisms, for example, certain proteins must be expressed at one stage only, and the mRNA encoding such a protein must be made and destroyed at the appropriate times.

Eukaryotic mRNAs Are Capped at the 5' End

Most eukaryotic mRNAs have a **5' cap**, a residue of 7-methylguanosine linked to the 5'-terminal residue of

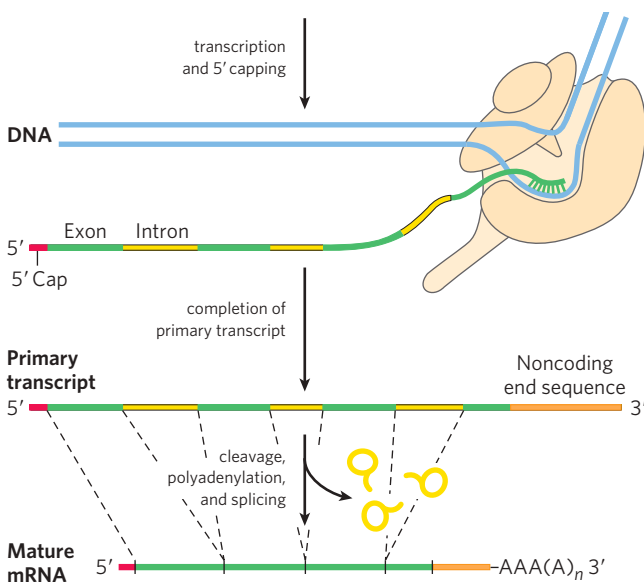


FIGURE 26–11 Formation of the primary transcript and its processing during maturation of mRNA in a eukaryotic cell. The 5' cap (red) is added before synthesis of the primary transcript is complete. A noncoding end sequence (intron) following the last exon is shown in orange. Splicing can occur either before or after the cleavage and polyadenylation steps. All the processes shown here take place in the nucleus.

the mRNA through an unusual 5',5'-triphosphate linkage (**Fig. 26–12**). The 5' cap helps protect mRNA from ribonucleases. It also binds to a specific cap-binding complex of proteins and participates in binding of the mRNA to the ribosome to initiate translation (Chapter 27).

The 5' cap is formed by condensation of a molecule of GTP with the triphosphate at the 5' end of the transcript. The guanine is subsequently methylated at N-7, and additional methyl groups are often added at the 2' hydroxyls of the first and second nucleotides adjacent to the cap (Fig. 26–12a). The methyl groups are derived from *S*-adenosylmethionine. All these reactions occur very early in transcription, after the first 20 to 30 nucleotides of the transcript have been added. All three of the capping enzymes, and through them the 5' end of the transcript itself, are associated with the RNA polymerase II CTD until the cap is synthesized. The capped 5' end is then released from the capping enzymes and bound by the cap-binding complex (Fig. 26–12c).

Both Introns and Exons Are Transcribed from DNA into RNA

In bacteria, a polypeptide chain is generally encoded by a DNA sequence that is colinear with the amino acid sequence, continuing along the DNA template without interruption until the information needed to specify the polypeptide is complete. However, the notion that *all* genes are continuous was disproved in 1977 when Phillip Sharp and Richard Roberts independently discovered that many genes for polypeptides in eukaryotes are interrupted by noncoding sequences (introns).

The vast majority of genes in vertebrates contain introns; among the few exceptions are those that encode histones. The occurrence of introns in other eukaryotes varies. Many genes in the yeast *Saccharomyces cerevisiae* lack introns, although in some other yeast species introns are more common. Introns are also found in a few bacterial and archaeal genes. Introns in DNA are transcribed along with the rest of the gene by RNA polymerases. The introns in the primary RNA transcript are then spliced, and the exons are joined to form a mature, functional RNA. In eukaryotic mRNAs, most exons are less than 1,000 nucleotides long, with many in the 100 to 200 nucleotide size range, encoding stretches of 30 to 60 amino acids within a longer polypeptide. Introns vary in size from 50 to 20,000 nucleotides. Genes of higher eukaryotes, including humans, typically have much more DNA devoted to introns than to exons. Many genes have introns; some genes have dozens of them.

RNA Catalyzes the Splicing of Introns

There are four classes of introns. The first two, the group I and group II introns, differ in the details of their splicing mechanisms but share one surprising characteristic: they are *self-splicing*—no protein enzymes are involved. Group I introns are found in some nuclear,

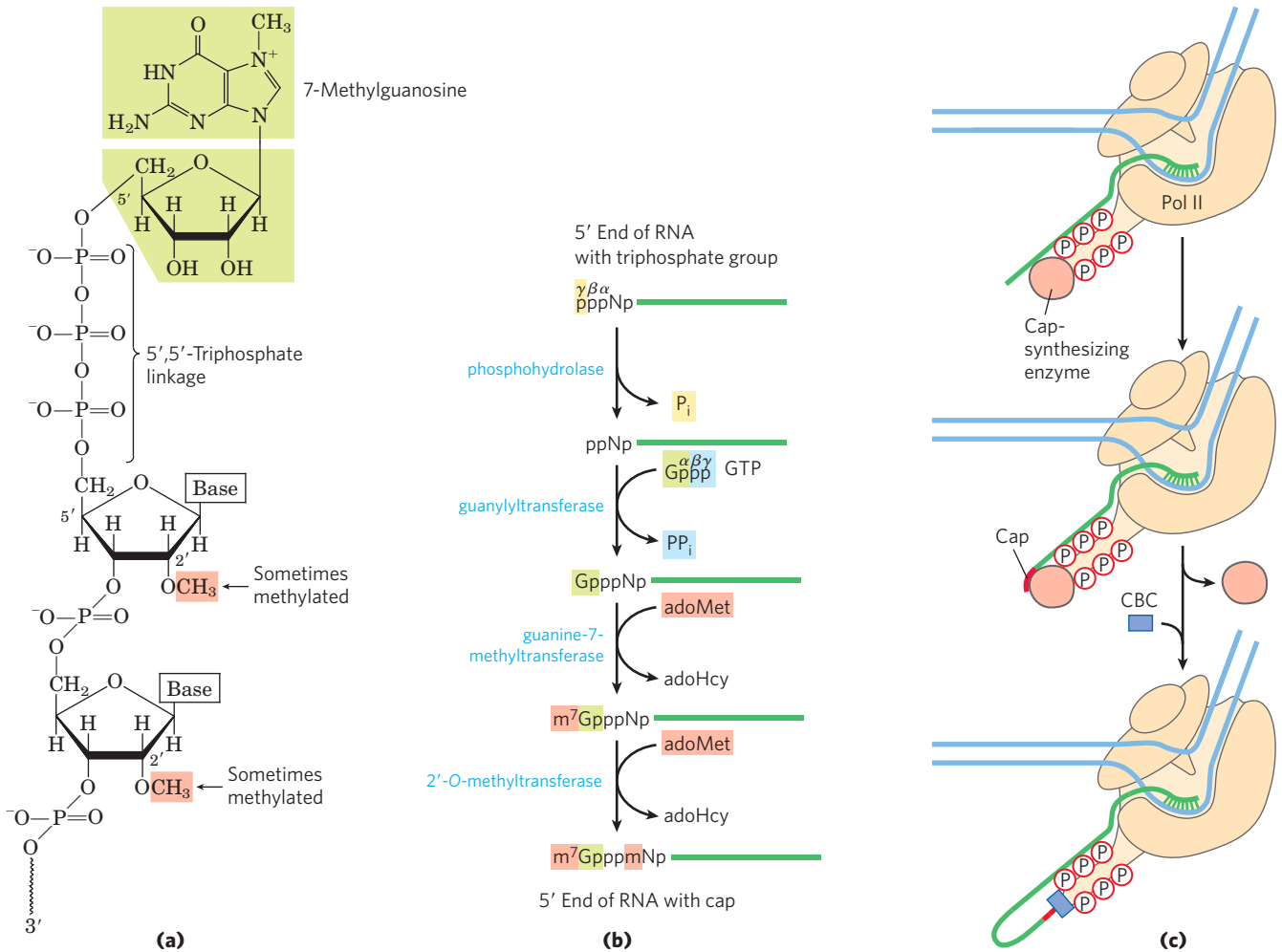


FIGURE 26-12 The 5' cap of mRNA. **(a)** 7-Methylguanosine (m⁷G) is joined to the 5' end of almost all eukaryotic mRNAs in an unusual 5',5'-triphosphate linkage. Methyl groups (light red) are often found at the 2' position of the first and second nucleotides. RNAs in yeast cells lack the 2'-methyl groups. The 2'-methyl group on the second nucleotide

is generally found only in RNAs from vertebrate cells. **(b)** Generation of the 5' cap involves four to five separate steps (adohcy is *S*-adenosylhomocysteine). **(c)** Synthesis of the cap is carried out by enzymes tethered to the CTD of Pol II. The cap remains tethered to the CTD through an association with the cap-binding complex (CBC).

mitochondrial, and chloroplast genes that code for rRNAs, mRNAs, and tRNAs. Group II introns are generally found in the primary transcripts of mitochondrial or chloroplast mRNAs in fungi, algae, and plants. Group I and group II introns are also found among the rare examples of introns in bacteria. Neither class requires a high-energy cofactor (such as ATP) for splicing. The splicing mechanisms in both groups involve two transesterification reaction steps (**Fig. 26-13**), in which a ribose 2'- or 3'-hydroxyl group makes a nucleophilic attack on a phosphorus and a new phosphodiester bond is formed at the expense of the old, maintaining the balance of energy. These reactions are very similar to the DNA breaking and rejoining reactions promoted by topoisomerases (see **Fig. 24-20**) and site-specific recombinases (see **Fig. 25-37**).

The group I splicing reaction requires a guanine nucleoside or nucleotide cofactor, but the cofactor is not

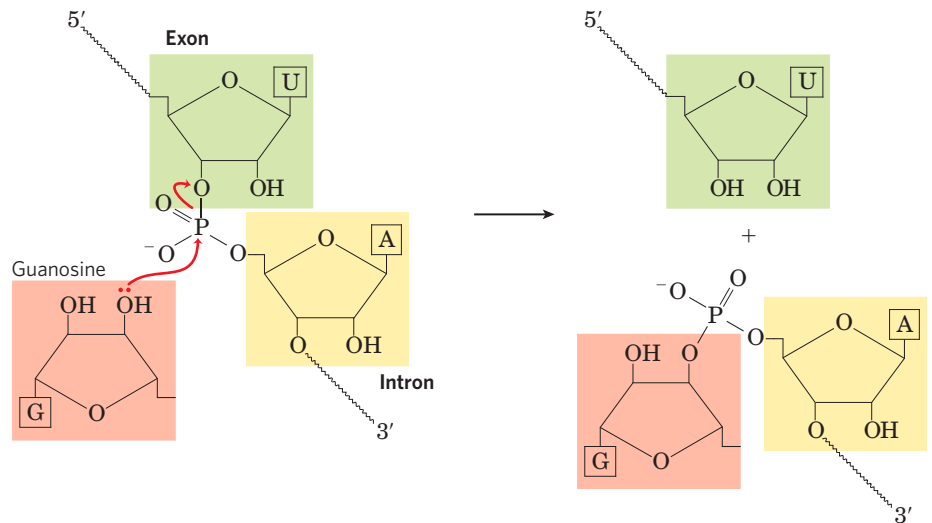
used as a source of energy; instead, the 3'-hydroxyl group of guanosine is used as a nucleophile in the first step of the splicing pathway. The guanosine 3'-hydroxyl group forms a normal 3',5'-phosphodiester bond with the 5' end of the intron (**Fig. 26-14**). The 3' hydroxyl of the exon that is displaced in this step then acts as a nucleophile in a similar reaction at the 3' end of the intron. The result is precise excision of the intron and ligation of the exons.

In group II introns the reaction pattern is similar except for the nucleophile in the first step, which in this case is the 2'-hydroxyl group of an A residue *within* the intron (**Fig. 26-15**). A branched lariat structure is formed as an intermediate.

Self-splicing of introns was first revealed in 1982 in studies of the splicing mechanism of the group I rRNA intron from the ciliated protozoan *Tetrahymena thermophila*, conducted by Thomas Cech and colleagues.

FIGURE 26-13 Transesterification reaction.

Shown here is the first step in the two-step splicing of group I introns. In this example, the 3' OH of a guanosine molecule acts as nucleophile, attacking the phosphodiester linkage between U and A residues at an exon-intron junction of an mRNA molecule (see Fig. 26-14).



Thomas Cech

These workers transcribed isolated *Tetrahymena* DNA (including the intron) in vitro using purified bacterial RNA polymerase. The resulting RNA spliced itself accurately without any protein enzymes from *Tetrahymena*. The discovery that RNAs could have catalytic functions was a milestone in our understanding of biological systems.

Most introns are *not* self-splicing, and these types are not designated with a group number. The third and largest class of introns includes those found in nuclear mRNA primary transcripts. These are called **spliceosomal introns**, because their removal occurs within and is catalyzed by a large protein complex called a **spliceosome**. Within the spliceosome, the introns undergo splicing by the same lariat-forming mechanism as the group II introns. The spliceosome is made up of specialized RNA-protein complexes, small nuclear ribonucleoproteins (snRNPs, often pronounced *snurps*). Each snRNP contains one of a class of eukaryotic RNAs, 100 to 200

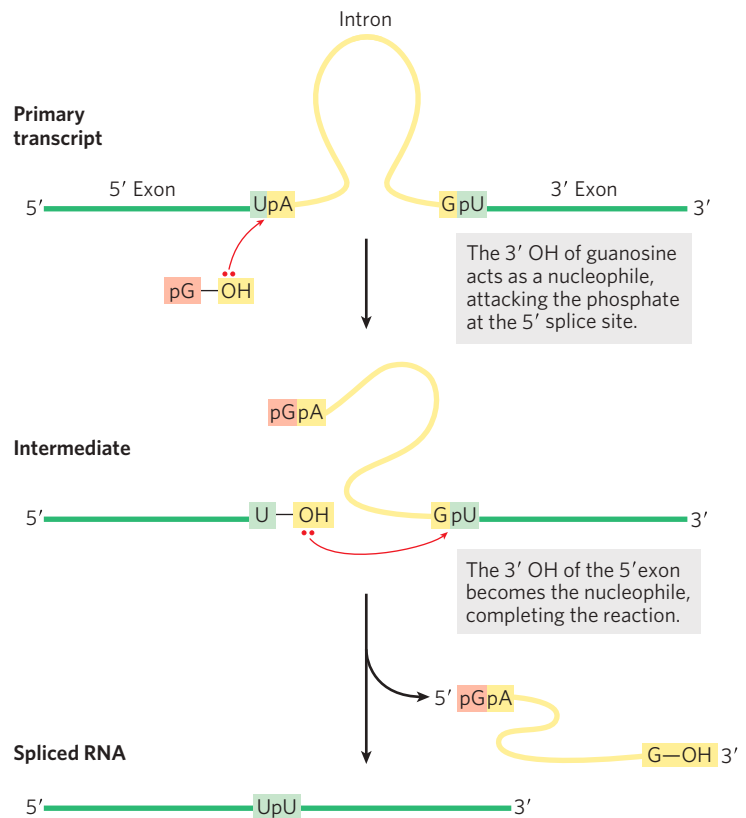


FIGURE 26-14 Splicing mechanism of group I introns. The nucleophile in the first step may be guanosine, GMP, GDP, or GTP. The spliced intron is eventually degraded.

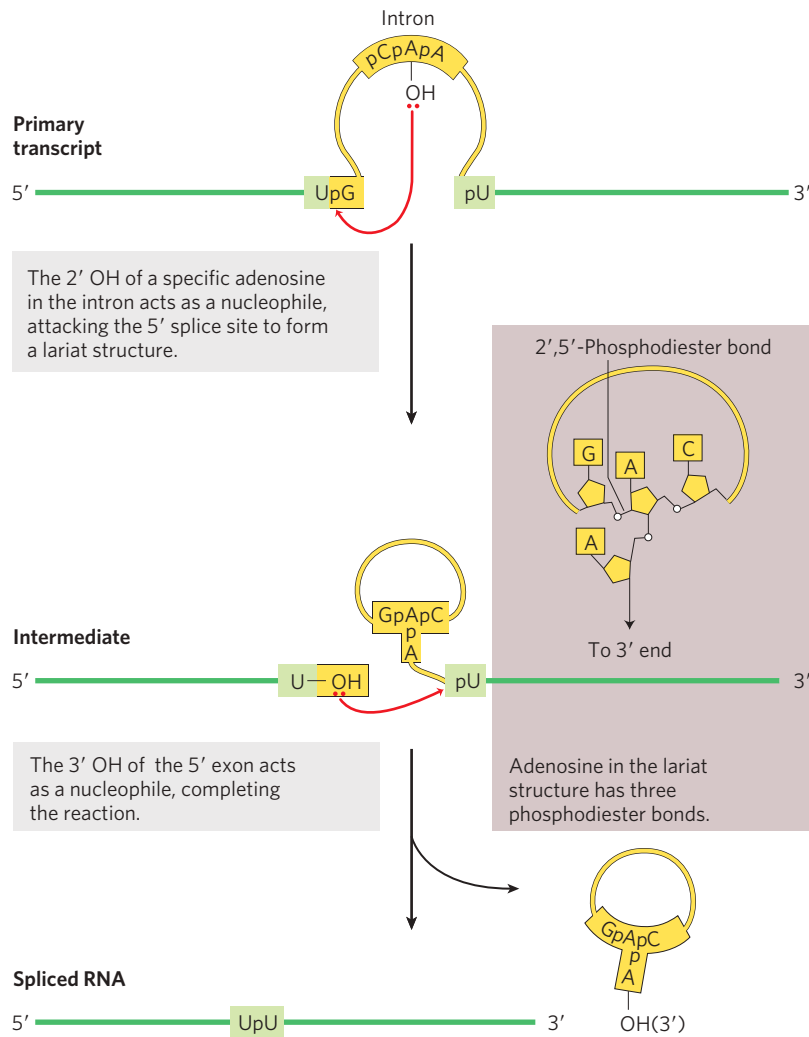


FIGURE 26-15 Splicing mechanism of group II introns. The chemistry is similar to that of group I intron splicing, except for the identity of the nucleophile in the first step and formation of a lariatlike intermediate, in which one branch is a 2',5'-phosphodiester bond.

nucleotides long, known as **small nuclear RNAs (snRNAs)**. Five snRNAs (U1, U2, U4, U5, and U6) involved in splicing reactions are generally found in abundance in eukaryotic nuclei. The RNAs and proteins in snRNPs are highly conserved in eukaryotes from yeasts to humans. **mRNA Splicing**

Spliceosomal introns generally have the dinucleotide sequence GU at the 5' end and AG at the 3' end, and these sequences mark the sites where splicing occurs. The U1 snRNA contains a sequence complementary to sequences near the 5' splice site of nuclear mRNA introns (**Fig. 26-16a**), and the U1 snRNP binds to this region in the primary transcript. Addition of the U2, U4, U5, and U6 snRNPs leads to formation of the spliceosome (**Fig. 26-16b**). The snRNPs together contribute five RNAs and about 50 proteins to the core spliceosome, a supramolecular assembly nearly as complex as the ribosome (described in Chapter 27). Perhaps 50 additional proteins are associated with the spliceosome at different stages in the splicing process, with some of these proteins having multiple functions: in splicing, mRNA transport to the cytoplasm, translation, and eventual mRNA degradation. ATP is required for assembly of the spliceo-

some, but the RNA cleavage-ligation reactions do not seem to require ATP. Some mRNA introns are spliced by a less common type of spliceosome, in which the U1 and U2 snRNPs are replaced by the U11 and U12 snRNPs. Whereas U1- and U2-containing spliceosomes remove introns with (5')GU and AG(3') terminal sequences, as shown in Figure 26-16, the U11- and U12-containing spliceosomes remove a rare class of introns that have (5')AU and AC(3') terminal sequences to mark the intronic splice sites. The spliceosomes used in nuclear RNA splicing may have evolved from more ancient group II introns, with the snRNPs replacing the catalytic domains of their self-splicing ancestors.

Some components of the splicing apparatus are tethered to the CTD of RNA polymerase II, indicating that splicing, like other RNA processing reactions, is tightly coordinated with transcription (**Fig. 26-16c**). As the first splice junction is synthesized, it is bound by a tethered spliceosome. The second splice junction is then captured by this complex as it passes, facilitating the juxtaposition of the intron ends and the subsequent splicing process. After splicing, the intron remains in the nucleus and is eventually degraded.

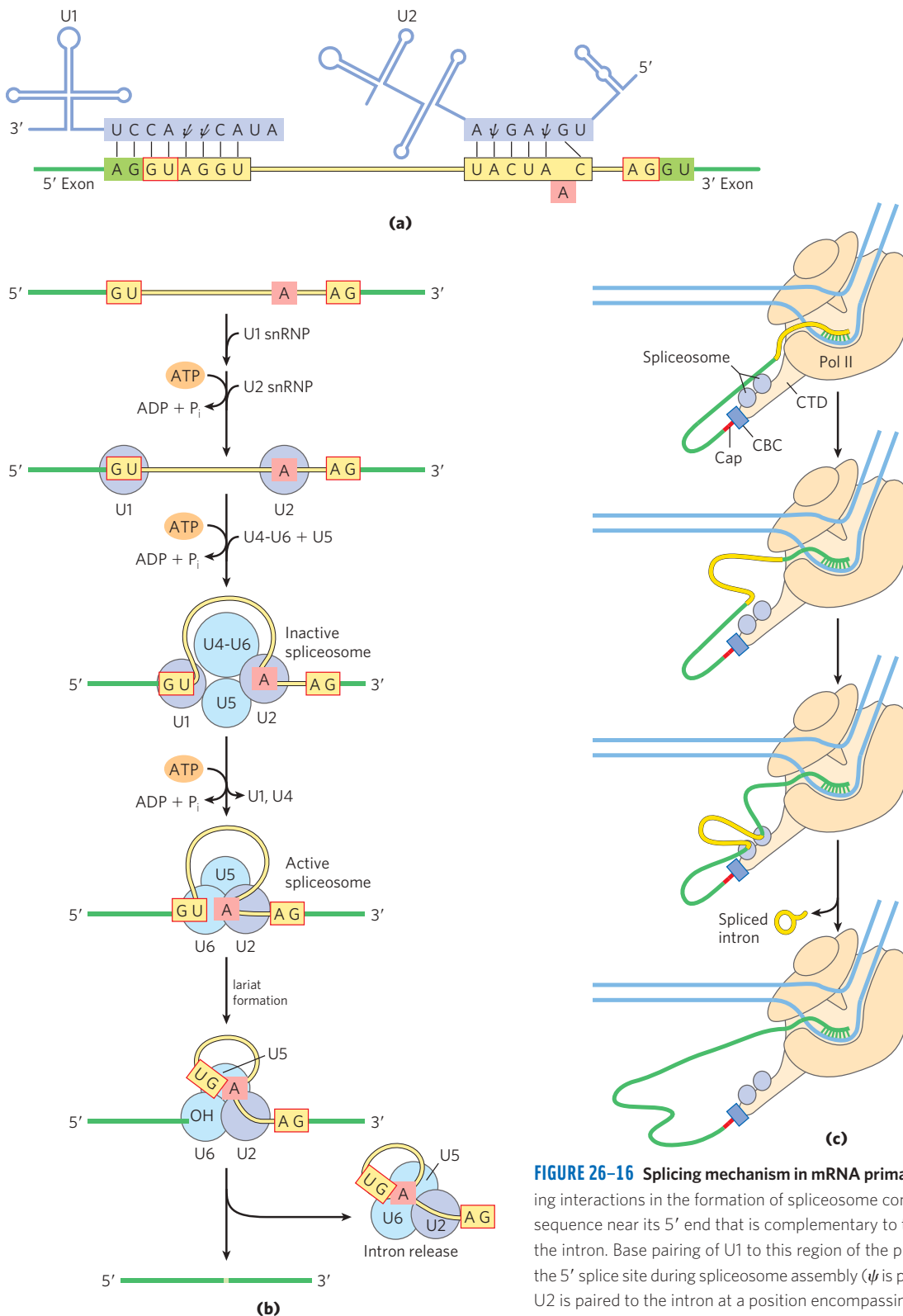


FIGURE 26-16 Splicing mechanism in mRNA primary transcripts. (a) RNA pairing interactions in the formation of spliceosome complexes. The U1 snRNA has a sequence near its 5' end that is complementary to the splice site at the 5' end of the intron. Base pairing of U1 to this region of the primary transcript helps define the 5' splice site during spliceosome assembly (ψ is pseudouridine; see Fig. 26-22). U2 is paired to the intron at a position encompassing the A residue (shaded light red) that becomes the nucleophile during the splicing reaction. Base pairing of U2 snRNA causes a bulge that displaces and helps to activate the adenylate, the 2' OH of which will form the lariat structure through a 2',5'-phosphodiester bond.

(b) Assembly of spliceosomes. The U1 and U2 snRNPs bind, then the remaining snRNPs (the U4-U6 complex and U5) bind to form an inactive spliceosome. Internal rearrangements convert this species to an active spliceosome in which U1 and U4 have been expelled and U6 is paired with both the 5' splice site and U2. This is followed by the catalytic steps, which parallel those of the splicing of group II introns (see Fig. 26-15).

(c) Coordination of splicing and transcription brings the two splice sites together. See the text for details. The spliceosome is much larger than indicated here.

The fourth class of introns, found in certain tRNAs, is distinguished from the group I and II introns in that the splicing reaction requires ATP and an endonuclease. The splicing endonuclease cleaves the phosphodiester bonds at both ends of the intron, and the two exons are joined by a mechanism similar to the DNA ligase reaction (see Fig. 25–16).

Although spliceosomal introns seem to be limited to eukaryotes, the other intron classes are not. Genes with group I and II introns have now been found in both bacteria and bacterial viruses. Bacteriophage T4, for example, has several protein-encoding genes with group I introns. Introns may be more common in archaea than in bacteria.

Eukaryotic mRNAs Have a Distinctive 3' End Structure

At their 3' end, most eukaryotic mRNAs have a string of 80 to 250 A residues, making up the **poly(A) tail**. This tail serves as a binding site for one or more specific proteins. The poly(A) tail and its associated proteins probably help protect mRNA from enzymatic destruction. Many bacterial mRNAs also acquire poly(A) tails, but these tails stimulate decay of mRNA rather than protecting it from degradation.

The poly(A) tail is added in a multistep process. The transcript is extended beyond the site where the poly(A) tail is to be added, then is cleaved at the poly(A) addition site by an endonuclease component of a large enzyme complex, again associated with the CTD of RNA polymerase II (Fig. 26–17). The mRNA site where cleavage occurs is marked by two sequence elements: the highly conserved sequence (5')AAUAAA(3'), 10 to 30 nucleotides on the 5' side (upstream) of the cleavage site, and a less well-defined sequence rich in G and U residues, 20 to 40 nucleotides downstream of the cleavage site. Cleavage generates the free 3'-hydroxyl group that defines the end of the mRNA, to which A residues are immediately added by **polyadenylate polymerase**, which catalyzes the reaction



where $n = 80$ to 250. This enzyme does not require a template but does require the cleaved mRNA as a primer.

The overall processing of a typical eukaryotic mRNA is summarized in Figure 26–18. In some cases the polypeptide-coding region of the mRNA is also modified by RNA “editing” (see Section 27.1 for details). This editing includes processes that add or delete bases in the coding regions of primary transcripts or that change the sequence (by, for example, enzymatic deamination of a C residue to create a U residue). A particularly dramatic example occurs in trypanosomes, which are parasitic protozoa: large regions of an mRNA are synthesized without any uridylate, and the U residues are inserted later by RNA editing.

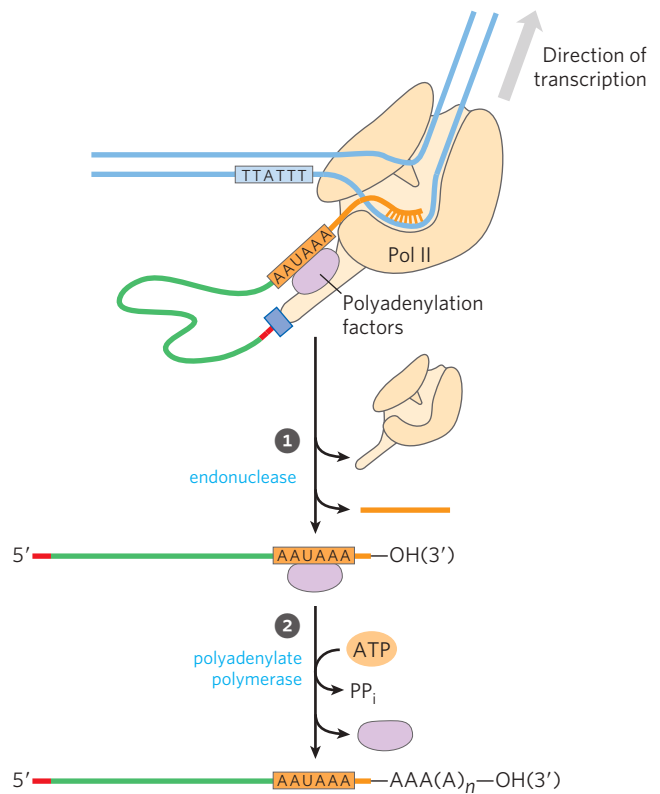


FIGURE 26–17 Addition of the poly(A) tail to the primary RNA transcript of eukaryotes. Pol II synthesizes RNA beyond the segment of the transcript containing the cleavage signal sequences, including the highly conserved upstream sequence (5')AAUAAA. This cleavage signal sequence is bound by an enzyme complex that includes an endonuclease, a polyadenylate polymerase, and several other multisubunit proteins involved in sequence recognition, stimulation of cleavage, and regulation of the length of the poly(A) tail, all of which are tethered to the CTD. ① The RNA is cleaved by the endonuclease at a point 10 to 30 nucleotides 3' to (downstream of) the sequence AAUAAA. ② The polyadenylate polymerase synthesizes a poly(A) tail 80 to 250 nucleotides long, beginning at the cleavage site.

A Gene Can Give Rise to Multiple Products by Differential RNA Processing

One of the paradoxes of modern genomics is that the apparent complexity of organisms does not correlate with the number of protein-encoding genes or even with the amount of genomic DNA. However, the traditional focus on protein-encoding genes ignores the complexity of an organism's transcriptome. As the functions of RNA are better appreciated, new genomic complexities become apparent.

Some eukaryotic mRNA transcripts produce only one mature mRNA and one corresponding polypeptide, but others can be processed in more than one way to produce *different* mRNAs and thus different polypeptides. The primary transcript contains molecular signals for all the alternative processing pathways, and the pathway favored in a given cell is determined by processing factors, RNA-binding proteins that promote one particular path.

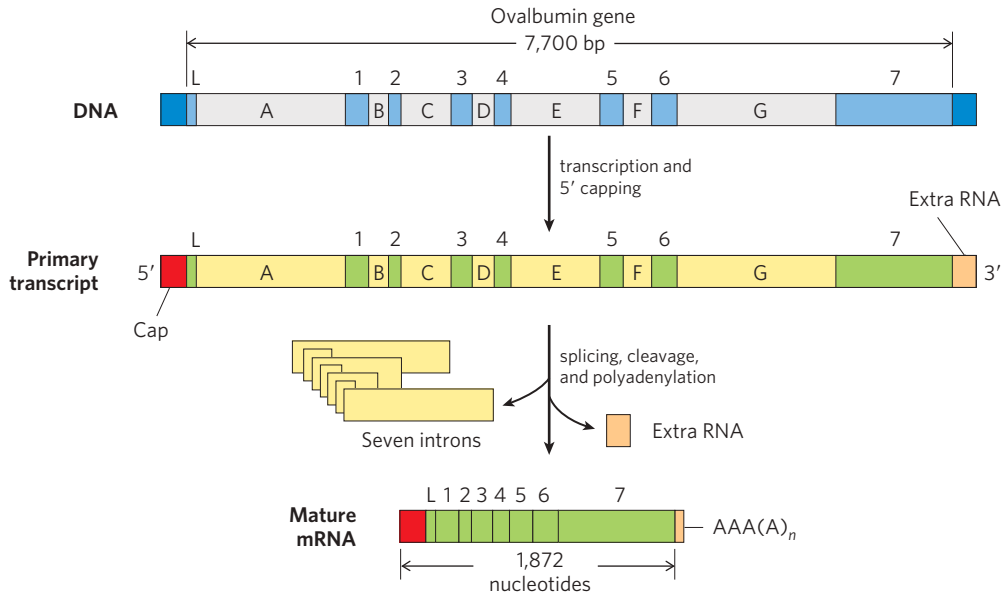


FIGURE 26-18 Overview of the processing of a eukaryotic mRNA. The ovalbumin gene, shown here, has introns A to G and exons 1 to 7 and L (L encodes a signal peptide sequence that targets the protein for export from the cell; see Fig. 27-38). About three-quarters of the RNA is

removed during processing. Pol II extends the primary transcript well beyond the cleavage and polyadenylation site (“extra RNA”) before terminating transcription. Termination signals for Pol II have not yet been defined.

Complex transcripts can have either more than one site for cleavage and polyadenylation or alternative splicing patterns, or both. If there are two or more sites for cleavage and polyadenylation, use of the one closest to the 5' end will remove more of the primary transcript sequence (**Fig. 26-19a**). This mechanism, called poly(A)

site choice, generates diversity in the variable domains of immunoglobulin heavy chains (see Fig. 25-42). Alternative splicing patterns (**Fig. 26-19b**) produce, from a common primary transcript, three different forms of the myosin heavy chain at different stages of fruit fly development. *Both* mechanisms come into play when a single

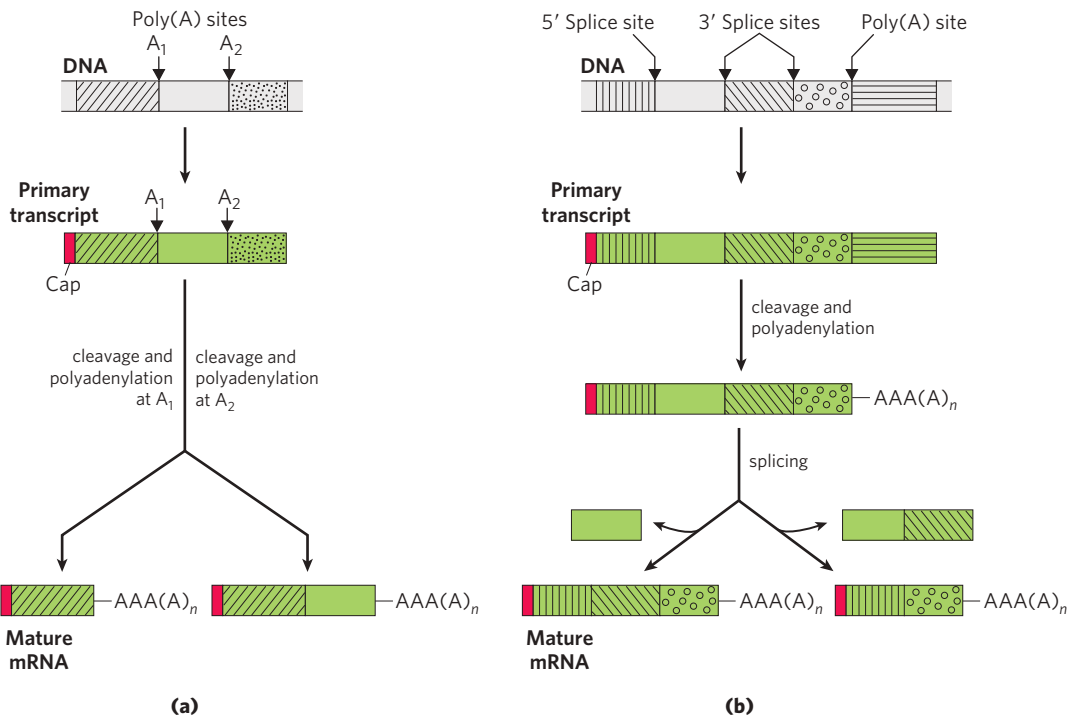


FIGURE 26-19 Two mechanisms for the alternative processing of complex transcripts in eukaryotes. (a) Alternative cleavage and polyadenylation patterns. Two poly(A) sites, A_1 and A_2 , are shown. (b) Alternative

splicing patterns. Two different 3' splice sites are shown. In both mechanisms, different mature mRNAs are produced from the same primary transcript.

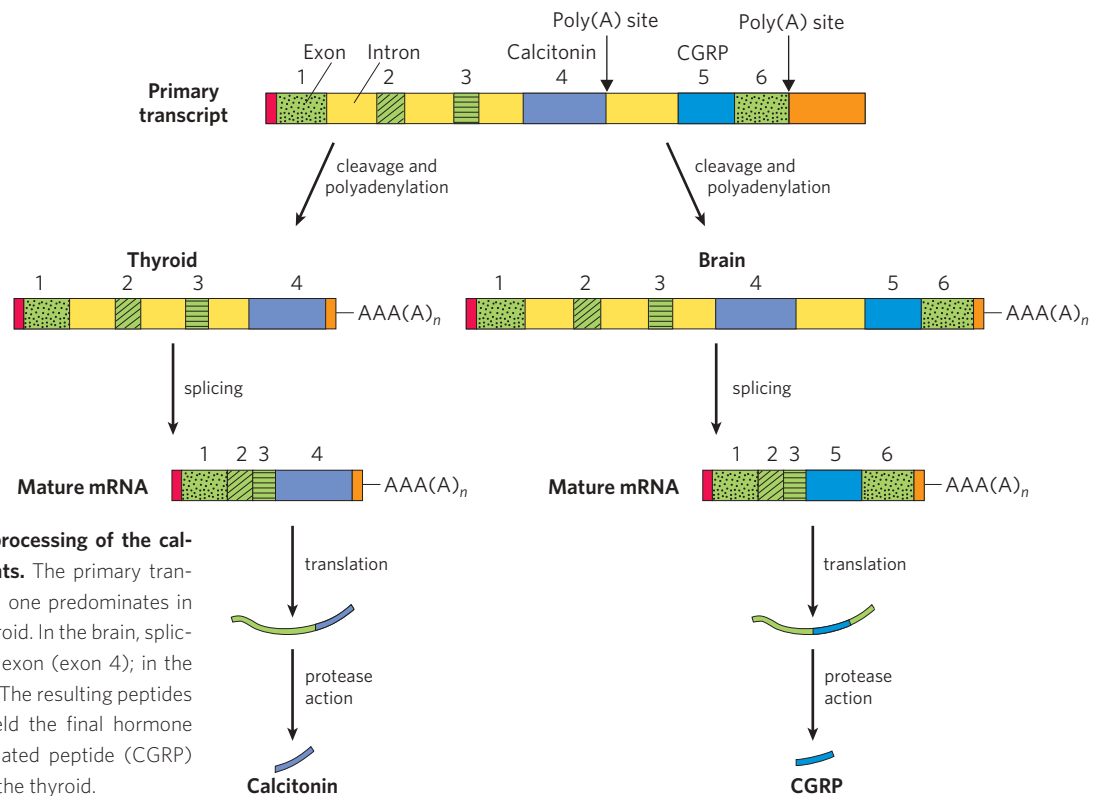


FIGURE 26-20 Alternative processing of the calcitonin gene transcript in rats. The primary transcript has two poly(A) sites; one predominates in the brain, the other in the thyroid. In the brain, splicing eliminates the calcitonin exon (exon 4); in the thyroid, this exon is retained. The resulting peptides are processed further to yield the final hormone products: calcitonin-gene-related peptide (CGRP) in the brain and calcitonin in the thyroid.

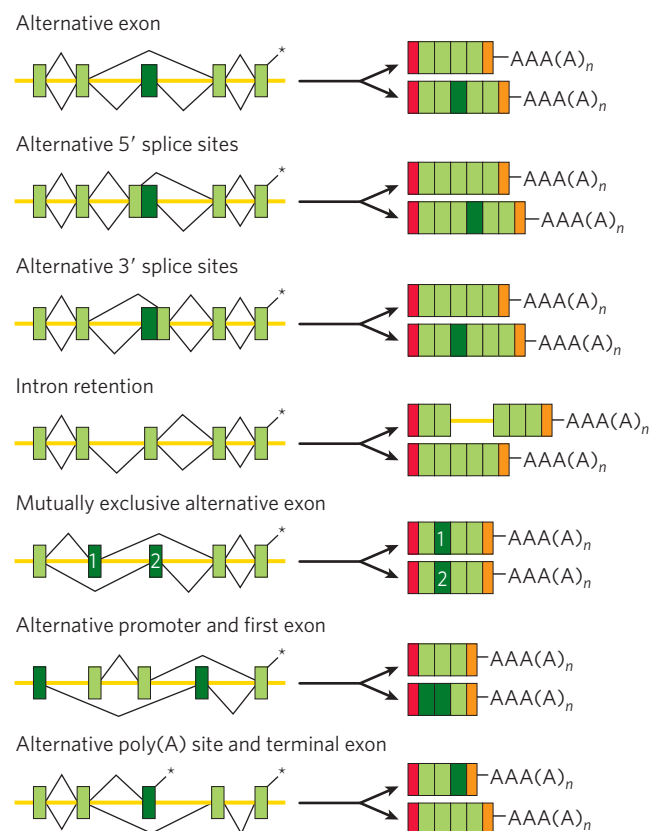
RNA transcript is processed differently to produce two different hormones: the calcium-regulating hormone calcitonin in rat thyroid and calcitonin-gene-related peptide (CGRP) in rat brain (Fig. 26-20). There are many additional patterns of alternative splicing (Fig. 26-21). Many, perhaps most, of the genes in mammalian genomes are subject to alternative splicing, substantially increasing the number of proteins encoded by the genes. The same processes play a much smaller role in lower eukaryotes, with only a few genes subject to alternative splicing in yeast.

Ribosomal RNAs and tRNAs Also Undergo Processing

Posttranscriptional processing is not limited to mRNA. Ribosomal RNAs of bacterial, archaeal, and eukaryotic cells are made from longer precursors called **preribosomal RNAs**, or pre-rRNAs. Transfer RNAs are similarly derived from longer precursors. These RNAs may also contain a variety of modified nucleosides; some examples are shown in Figure 26-22.

FIGURE 26-21 Summary of splicing patterns. Exons are shown in shades of green, and introns/untranslated regions as yellow lines. Positions where polyadenosine is to be added are marked with asterisks. Exons joined in a particular splicing scheme are linked with black lines. The alternative linkage patterns above and below the transcript lead to the top and bottom spliced mRNA products, respectively. In the products, red and orange boxes represent the 5' cap and 3' untranslated regions, respectively.

Ribosomal RNAs In bacteria, 16S, 23S, and 5S rRNAs (and some tRNAs, although most tRNAs are encoded elsewhere) arise from a single 30S RNA precursor of



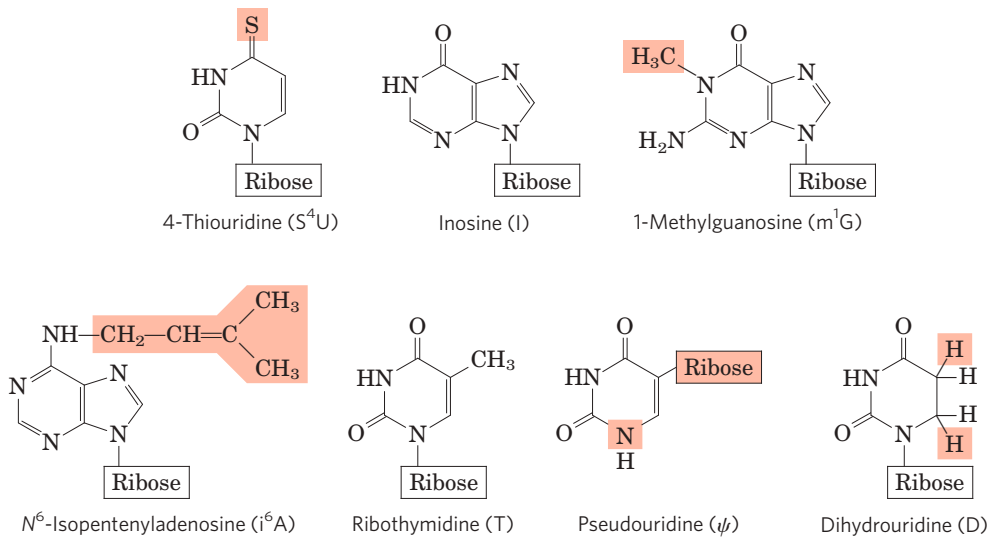


FIGURE 26-22 Some modified bases of rRNAs and tRNAs, produced in posttranscriptional reactions. The standard symbols are shown in parentheses. Note the unusual ribose attachment point in pseudouridine. This is just a small sampling of the 96 modified nucleosides known to occur

in different RNA species, with 81 different types known in tRNAs and 30 observed to date in rRNAs. A complete listing of these modified bases can be found in the RNA modification database (<http://rna-mdb.cas.albany.edu/RNAmods/>).

about 6,500 nucleotides. RNA at both ends of the 30S precursor and segments between the rRNAs are removed during processing (Fig. 26-23). The 16S and 23S rRNAs contain modified nucleosides. In *E. coli*, the 11 modifications in the 16S rRNA include a pseudouridine and 10 nucleosides methylated on the base or the 2'-hydroxyl group or both. The 23S rRNA has 10 pseudouridines, 1 dihydrouridine, and 12 methylated nucleosides. In bacteria, each modification is generally catalyzed by a distinct enzyme. Methylation reactions use *S*-adenosylmethionine as cofactor. No cofactor is required for pseudouridine formation.

The genome of *E. coli* encodes seven pre-rRNA molecules. All of these genes have essentially identical rRNA-coding regions, but they differ in the segments between these regions. The segment between the 16S and 23S rRNA genes generally encodes one or two tRNAs, with different tRNAs produced from different pre-rRNA transcripts. Coding sequences for tRNAs are also found on the 3' side of the 5S rRNA in some precursor transcripts.

The situation in eukaryotes is more complicated. A 45S pre-rRNA transcript is synthesized by RNA polymerase I and processed in the nucleolus to form the

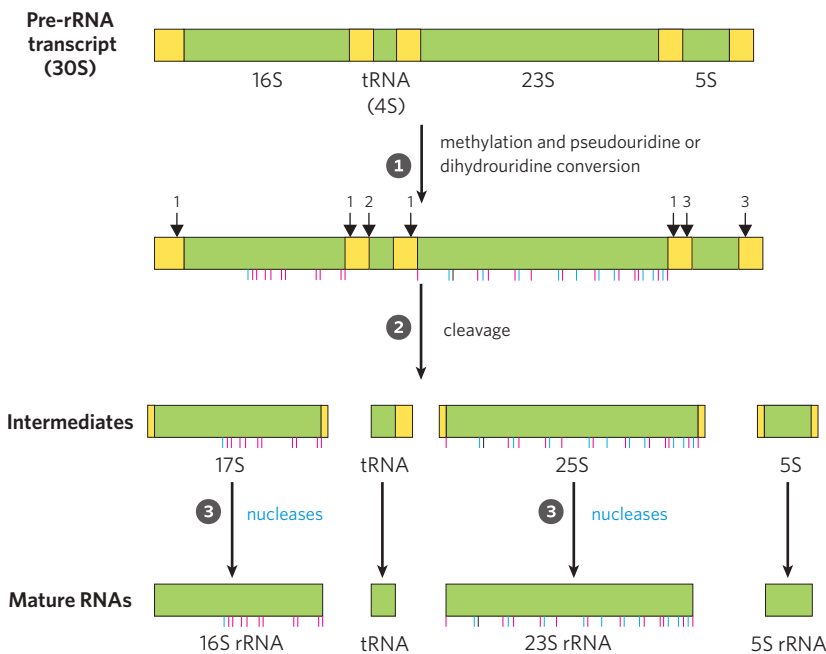


FIGURE 26-23 Processing of pre-rRNA transcripts in bacteria. **1** Before cleavage, the 30S RNA precursor is methylated at specific bases (red tick marks), and some uridine residues are converted to pseudouridine (blue tick marks) or dihydrouridine (black tick mark) residues. The methylation reactions are of multiple types, some occurring on bases and some on 2'-hydroxyl groups. **2** Cleavage liberates precursors of rRNAs and tRNA(s). Cleavage at the points labeled 1, 2, and 3 is carried out by the enzymes RNase III, RNase P, and RNase E, respectively. As discussed later in the text, RNase P is a ribozyme. **3** The final 16S, 23S, and 5S rRNA products result from the action of a variety of specific nucleases. The seven copies of the gene for pre-rRNA in the *E. coli* chromosome differ in the number, location, and identity of tRNAs included in the primary transcript. Some copies of the gene have additional tRNA gene segments between the 16S and 23S rRNA segments and at the far 3' end of the primary transcript.

18S, 28S, and 5.8S rRNAs characteristic of eukaryotic ribosomes (Fig. 26–24). As in bacteria, the processing includes cleavage reactions mediated by endo- or exoribonucleases and nucleoside modification reactions. Some pre-rRNAs also include introns that must be spliced. The entire process is initiated in the nucleolus, in large complexes that assemble on the rRNA precursor as it is synthesized by Pol I. There is a tight coupling between rRNA transcription, rRNA maturation, and

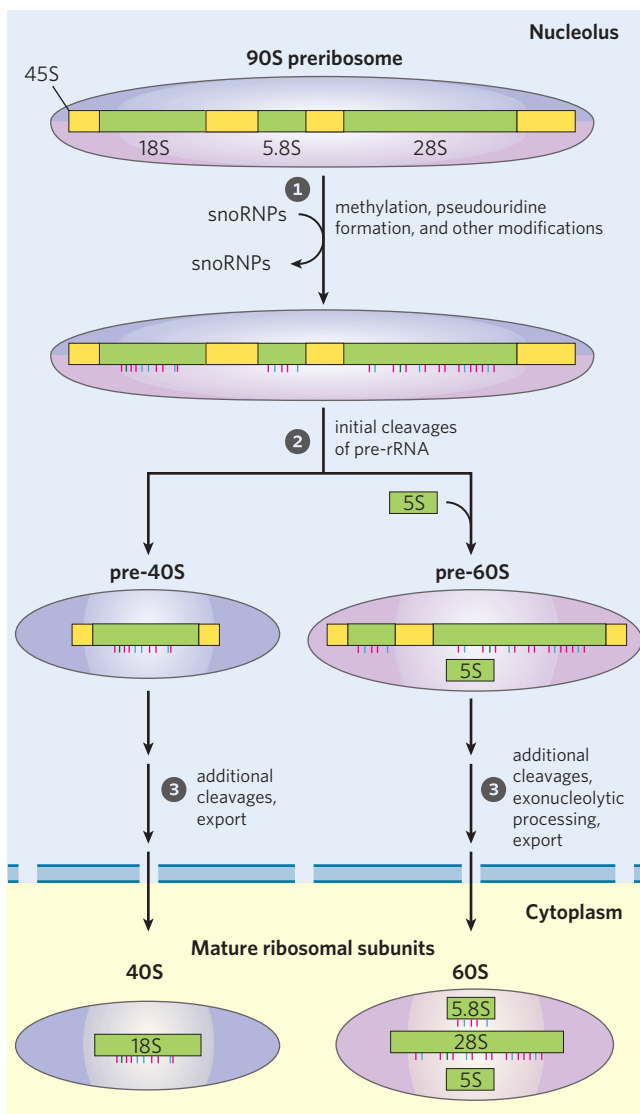


FIGURE 26–24 Processing of pre-rRNA transcripts in vertebrates. During transcription, the 45S primary transcript is incorporated into a nucleolar 90S preribosomal complex, in which rRNA processing and ribosome assembly are tightly coupled. **1** The 45S precursor is methylated at more than 100 of its 14,000 nucleotides, either on the bases or on the 2'-OH groups (red ticks), some uridines are converted to pseudouridine (blue ticks), and a few other modifications occur (green ticks are dihydrouridine). **2** and **3** A series of enzymatic cleavages of the 45S precursor produces the 18S, 5.8S, and 28S rRNAs, and the ribosomal subunits gradually take shape with the assembling ribosomal proteins. The cleavage reactions and all of the modifications require small nucleolar RNAs (snoRNAs) found in protein complexes (snoRNPs) in the nucleolus that are reminiscent of spliceosomes. The 5S rRNA is produced separately.

ribosome assembly in the nucleolus. Each complex includes the ribonucleases that cleave the rRNA precursor, the enzymes that modify particular bases, large numbers of **small nucleolar RNAs**, or **snoRNAs**, that guide nucleoside modification and some cleavage reactions, and ribosomal proteins. In yeast, the entire process involves the pre-rRNA, more than 170 nonribosomal proteins, snoRNAs for each nucleoside modification (about 70 in all, since some snoRNAs guide two types of modification), and the 78 ribosomal proteins. Humans have an even greater number of modified nucleosides, about 200, and a greater number of associated snoRNAs. The composition of the complexes may change as the ribosomes are assembled, and many of the intermediate complexes may rival the ribosome itself, and the snRNPs, in complexity. The 5S rRNA of most eukaryotes is made as a completely separate transcript by a different polymerase (Pol III).

The most common nucleoside modifications in eukaryotic rRNAs are, again, conversion of uridine to pseudouridine and adomet-dependent nucleoside methylation (often at 2'-hydroxyl groups). These reactions rely on snoRNA-protein complexes, or **snoRNPs**, each consisting of a snoRNA and four or five proteins, which include the enzyme that carries out the modification. There are two classes of snoRNPs, both defined by key conserved sequence elements referred to as lettered boxes. The box H/ACA snoRNPs are involved in pseudouridylation, and box C/D snoRNPs function in 2'-O-methylations. Unlike the situation in bacteria, the same enzyme may participate in modifications at many sites, guided by the snoRNAs.

The snoRNAs are 60 to 300 nucleotides long. Many are encoded within the introns of other genes and cotranscribed with those genes. Each snoRNA includes a 10 to 21 nucleotide sequence that is perfectly complementary to some site on an rRNA. The conserved sequence elements in the remainder of the snoRNA fold into structures that are bound by the snoRNP proteins (Fig. 26–25).

Transfer RNAs Most cells have 40 to 50 distinct tRNAs, and eukaryotic cells have multiple copies of many of the tRNA genes. Transfer RNAs are derived from longer RNA precursors by enzymatic removal of nucleotides from the 5' and 3' ends (Fig. 26–26). In eukaryotes, introns are present in a few tRNA transcripts and must be excised. Where two or more different tRNAs are contained in a single primary transcript, they are separated by enzymatic cleavage. The endonuclease RNase P, found in all organisms, removes RNA at the 5' end of tRNAs. This enzyme contains both protein and RNA. The RNA component is essential for activity, and in bacterial cells it can carry out its processing function with precision even without the protein component. RNase P is therefore another example of a catalytic RNA, as described in more detail below. The 3' end of tRNAs is processed by one or more nucleases, including the exonuclease RNase D.

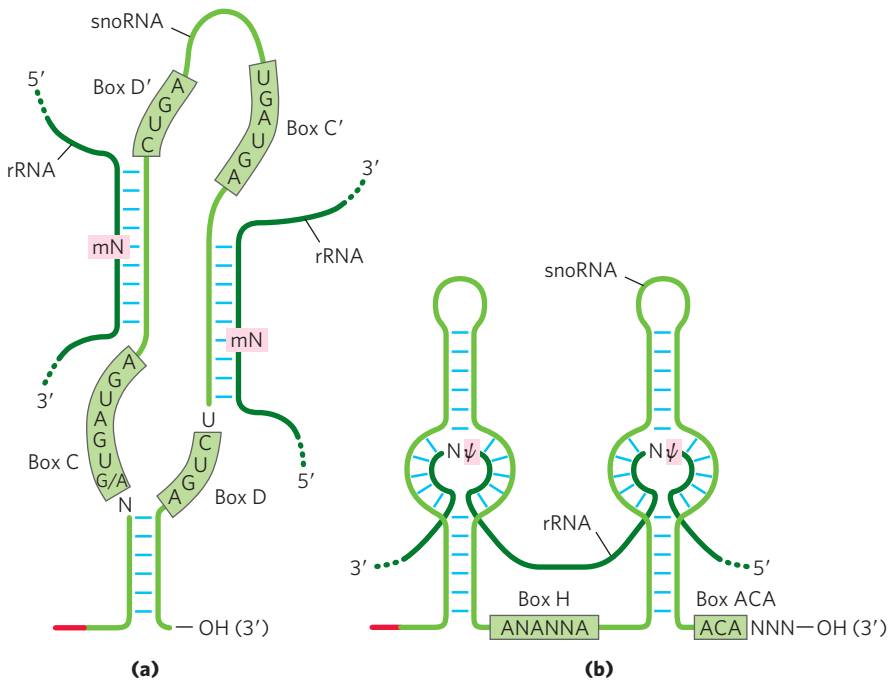


FIGURE 26-25 The function of snoRNAs in guiding rRNA modification. **(a)** RNA pairing with box C/D snoRNAs to guide methylation reactions. The methylation sites in the target rRNA (dark green) are in the regions paired with the C/D snoRNA. The highly conserved C and D (and C' and D') box sequences are binding sites for proteins that make up the larger snoRNP. **(b)** RNA pairing with box H/ACA snoRNAs to guide pseudouridylation. The pseudouridine conversion sites in the target rRNA (green segments) are again in the regions paired with the snoRNA, and the conserved H/ACA box sequences are protein-binding sites.

Transfer RNA precursors may undergo further posttranscriptional processing. The 3'-terminal trinucleotide CCA(3') to which an amino acid is attached during protein synthesis (Chapter 27) is absent from some bacterial and all eukaryotic tRNA precursors and is added during processing (Fig. 26-26). This addition is carried out by tRNA nucleotidyltransferase, an unusual enzyme that binds the three ribonucleoside triphosphate precursors in separate active sites and catalyzes formation of the phosphodiester bonds to produce the

CCA(3') sequence. The creation of this defined sequence of nucleotides is therefore not dependent on a DNA or RNA template—the template is the binding site of the enzyme.

The final type of tRNA processing is the modification of some bases by methylation, deamination, or reduction (Fig. 26-22). In the case of pseudouridine, the base (uracil) is removed and reattached to the sugar through C-5. Some of these modified bases occur at characteristic positions in all tRNAs (Fig. 26-26).

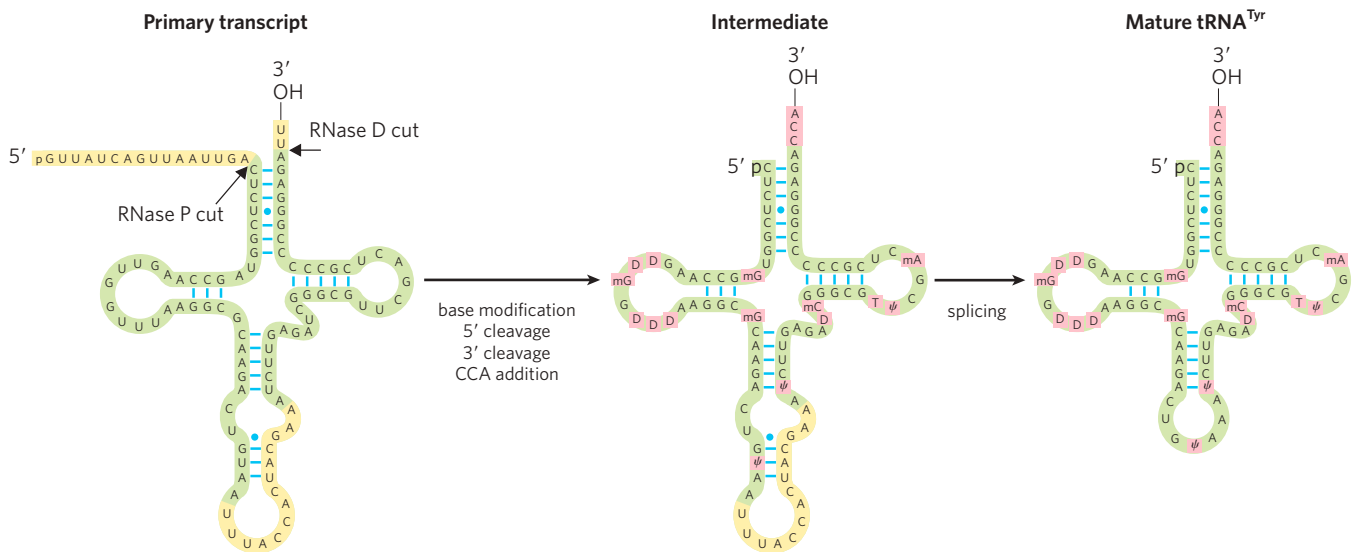


FIGURE 26-26 Processing of tRNAs in bacteria and eukaryotes. The yeast tRNA^{Tyr} (the tRNA specific for tyrosine binding; see Chapter 27) is used to illustrate the important steps. The nucleotide sequences shown in yellow are removed from the primary transcript. The ends are processed first, the 5' end before the 3' end. CCA is then added to the 3' end, a necessary step in processing eukaryotic tRNAs and those bacterial tRNAs

that lack this sequence in the primary transcript. While the ends are being processed, specific bases in the rest of the transcript are modified (see Fig. 26-22). For the eukaryotic tRNA shown here, the final step is splicing of the 14 nucleotide intron. Introns are found in some eukaryotic tRNAs but not in bacterial tRNAs.

Special-Function RNAs Undergo Several Types of Processing

The number of known classes of special-function RNAs is expanding rapidly, as is the variety of functions known to be associated with them. Many of these RNAs undergo processing.

The snRNAs and snoRNAs not only facilitate RNA processing reactions but are themselves synthesized as larger precursors and then processed. Many snoRNAs are encoded within the introns of other genes. As the introns are spliced from the pre-mRNA, the snoRNP proteins bind to the snoRNA sequences and ribonucleases remove the extra RNA at the 5' and 3' ends. The snRNAs destined for spliceosomes are synthesized as pre-snRNAs by RNA polymerase II, and ribonucleases remove the extra RNA at each end. Particular nucleosides in snRNAs are also subject to 11 types of modification, with 2'-*O*-methylation and conversion of uridine to pseudouridine predominating.

MicroRNAs (miRNAs) are a special class of RNAs involved in gene regulation. They are noncoding RNAs, about 22 nucleotides long, complementary in sequence to particular regions of mRNAs. They regulate mRNA function by cleaving the mRNA or suppressing its translation. The miRNAs are found in multicellular eukaryotes ranging from worms and fruit flies to plants and mammals. Up to 1% of the human genome may encode miRNAs, and miRNAs may target up to one-third of human mRNAs. Their function in gene regulation is described in Chapter 28.

The miRNAs are synthesized from much larger precursors, in several steps (Fig. 26–27). The primary transcripts for miRNAs (pri-miRNAs) vary greatly in size; some are encoded in the introns of other genes and are coexpressed with these host genes. Their roles in gene regulation also are detailed in Chapter 28.

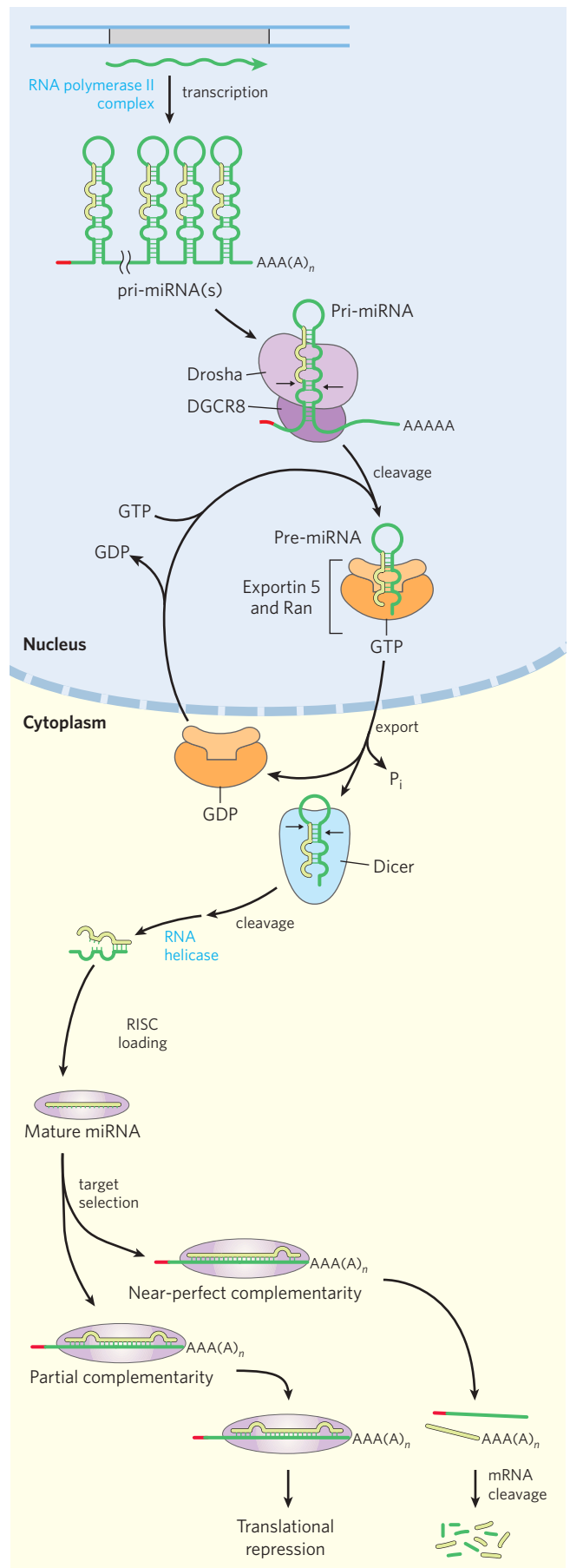


FIGURE 26–27 Synthesis and processing of miRNAs. The primary transcript of miRNAs is larger RNA of variable length, a pri-miRNA. Much of its processing is mediated by two endoribonucleases in the RNase III family, Drosha and Dicer. First, in the nucleus, the pri-miRNA is reduced to a 70 to 80 nucleotide precursor miRNA (pre-miRNA) by a protein complex including Drosha and another protein, DGCR8. The pre-miRNA is then exported to the cytoplasm in a complex with a protein called exportin-5 and the Ran GTPase (see Fig. 27–42). In the cytoplasm, Ran hydrolyzes the GTP, then exportin-5 protein and the pre-miRNA are released. The Ran-GDP and exportin-5 proteins are transported back into the nucleus. The pre-miRNA is acted on by Dicer to produce the nearly mature miRNA paired with a short RNA complement. The complement is removed by an RNA helicase, and the mature miRNA is incorporated into protein complexes, such as the RNA-induced silencing complex (RISC), which then bind a target mRNA. If the complementarity between miRNA and its target is nearly perfect, the target mRNA is cleaved. If the complementarity is only partial, the complex blocks translation of the target mRNA.

RNA Enzymes Are the Catalysts of Some Events in RNA Metabolism

The study of posttranscriptional processing of RNA molecules led to one of the most exciting discoveries in modern biochemistry—the existence of RNA enzymes. The best-characterized ribozymes are the self-splicing group I introns, RNase P, and the hammerhead ribozyme (discussed below). Most of the activities of these ribozymes are based on two fundamental reactions: transesterification (Fig. 26–13) and phosphodiester bond hydrolysis (cleavage). The substrate for ribozymes is often an RNA molecule, and it may even be part of the ribozyme itself. When its substrate is RNA, the RNA catalyst can make use of base-pairing interactions to align the substrate for the reaction.

Ribozymes vary greatly in size. A self-splicing group I intron may have more than 400 nucleotides. The hammerhead ribozyme consists of two RNA strands with only 41 nucleotides in all (Fig. 26–28). As with protein enzymes, the three-dimensional structure of ribozymes is important for function. Ribozymes are inactivated by heating above their melting temperature or by addition of denaturing agents or complementary oligonucleotides, which disrupt normal base-pairing patterns. Ribozymes can also be inactivated if essential nucleotides are changed. The secondary structure of a self-splicing group I intron from the 26S rRNA precursor of *Tetrahymena* is shown in detail in Figure 26–29.

Enzymatic Properties of Group I Introns Self-splicing group I introns share several properties with enzymes besides accelerating the reaction rate, including their kinetic behavior and their specificity. Binding of the guanosine cofactor (Fig. 26–13) to the *Tetrahymena* group I rRNA intron is saturable ($K_m \approx 30 \mu\text{M}$) and can be competitively inhibited by 3'-deoxyguanosine. The intron is very precise in its excision reaction, largely due to a segment called the **internal guide sequence** that can base-pair with exon sequences near the 5' splice site (Fig. 26–29). This pairing promotes the alignment of specific bonds to be cleaved and rejoined.

Because the intron itself is chemically altered during the splicing reaction—its ends are cleaved—it may seem to lack one key enzymatic property: the ability to catalyze multiple reactions. Closer inspection has shown that after excision, the 414 nucleotide intron from *Tetrahymena* rRNA can, in vitro, act as a true enzyme (but in vivo it is quickly degraded). A series of intramolecular cyclization and cleavage reactions in the excised intron leads to the loss of 19 nucleotides from its 5' end. The remaining 395 nucleotide, linear RNA—referred to as L-19 IVS (*intervening sequence*)—promotes nucleotidyl transfer reactions in which some oligonucleotides are lengthened at the expense of others (Fig. 26–30). The best substrates are oligonucleotides, such as a synthetic (C)₅ oligomer, that can base-pair with the same

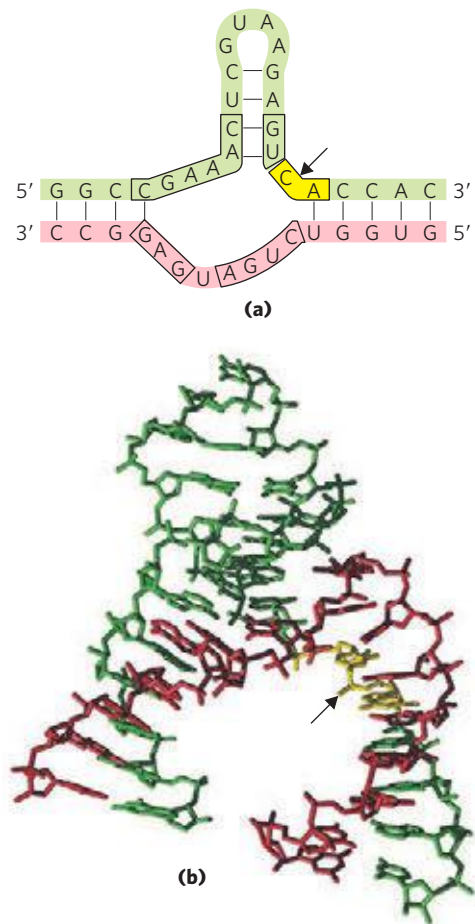



FIGURE 26-28 Hammerhead ribozyme. Certain viruslike elements, or virusoids, have small RNA genomes and usually require another virus to assist in their replication or packaging or both. Some virusoid RNAs include small segments that promote site-specific RNA cleavage reactions associated with replication. These segments are called hammerhead ribozymes, because their secondary structures are shaped like the head of a hammer. Hammerhead ribozymes have been defined and studied separately from the much larger viral RNAs. **(a)** The minimal sequences required for catalysis by the ribozyme. The boxed nucleotides are highly conserved and are required for catalytic function. The arrow indicates the site of self-cleavage. **(b)** Three-dimensional structure (PDB ID 1MME; see Fig. 8-25b for a space-filling view). The strands are colored as in (a). The hammerhead ribozyme is a metalloenzyme; Mg^{2+} ions are required for activity in vivo. The phosphodiester bond at the site of self-cleavage is indicated by an arrow.  **Hammerhead Ribozyme**

guanylate-rich internal guide sequence that held the 5' exon in place for self-splicing.

The enzymatic activity of the L-19 IVS ribozyme results from a cycle of transesterification reactions mechanistically similar to self-splicing. Each ribozyme molecule can process about 100 substrate molecules per hour and is not altered in the reaction; therefore the intron acts as a catalyst. It follows Michaelis-Menten kinetics, is specific for RNA oligonucleotide substrates, and can be competitively inhibited. The k_{cat}/K_m (specificity constant) is $10^3 \text{ M}^{-1}\text{s}^{-1}$, lower than that of many enzymes, but the ribozyme accelerates hydrolysis by a factor of 10^{10} relative

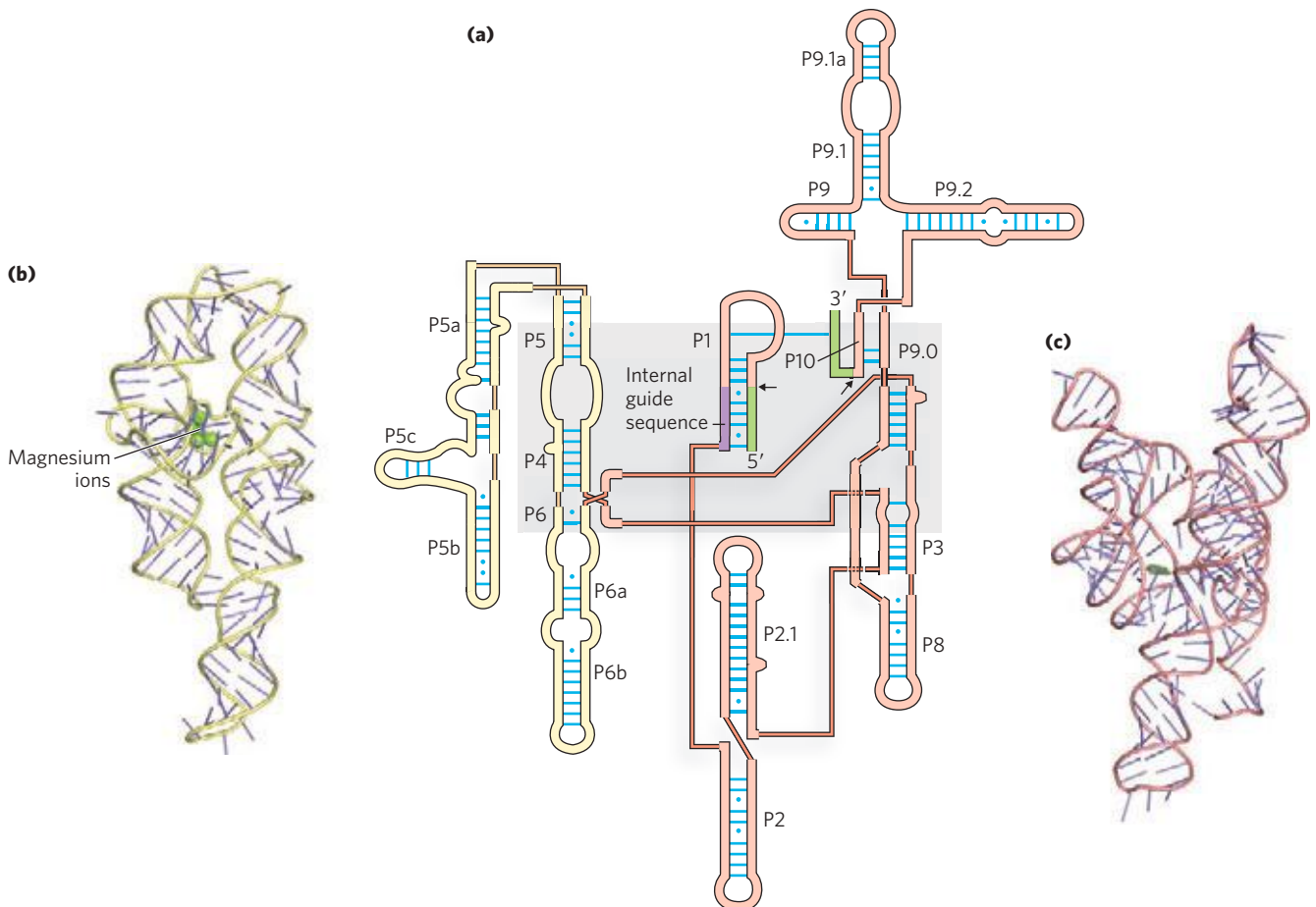


FIGURE 26-29 Secondary structure of the self-splicing rRNA intron of *Tetrahymena*. (a) Intron sequences are shaded yellow and light red, exon sequences green. Each thin, light red line represents a bond between neighboring nucleotides in a continuous sequence (a device necessitated by showing this complex molecule in two dimensions). Short blue lines represent normal base pairing; blue dots indicate G-U base pairs. All nucleotides are shown. The catalytic core of the self-splicing activity is shaded in gray. Some base-paired regions are labeled (P1, P3, P2.1, P5a, and

so forth) according to an established convention for this RNA molecule. The P1 region, which contains the internal guide sequence (purple), is the location of the 5' splice site (black arrow). Part of the internal guide sequence pairs with the end of the 3' exon, bringing the 5' and 3' splice sites (black arrows) into close proximity. (b) (PDB ID 1GID) The three-dimensional structure of the P4-P6 domain, shown in yellow in (a). (c) (PDB ID 1U6B) The three-dimensional structure of most of the remainder of the intron, shown in light red in (a).

to the uncatalyzed reaction. It makes use of substrate orientation, covalent catalysis, and metal-ion catalysis—strategies used by protein enzymes.

Characteristics of Other Ribozymes *E. coli* RNase P has both an RNA component (the M1 RNA, with 377 nucleotides) and a protein component (M_r 17,500). In 1983 Sidney Altman and Norman Pace and their coworkers discovered that under some conditions, the M1 RNA alone is capable of catalysis, cleaving tRNA precursors at the correct position. The protein component apparently serves to stabilize the RNA or facilitate its function in vivo. The RNase P ribozyme recognizes the three-dimensional shape of its pre-tRNA substrate, along with the CCA sequence, and thus can cleave the 5' leaders from diverse tRNAs (Fig. 26-26).

The known catalytic repertoire of ribozymes continues to expand. Some virusoids, small RNAs associated with plant RNA viruses, include a structure that promotes a self-cleavage reaction; the hammerhead ribozyme illustrated in Figure 26-28 is in this class, catalyzing the hydrolysis of an internal phosphodiester bond. The splicing reaction that occurs in a spliceosome seems to rely on a catalytic center formed by the U2, U5, and U6 snRNAs (Fig. 26-16). And perhaps most important, an RNA component of ribosomes catalyzes the synthesis of proteins (Chapter 27).

Exploring catalytic RNAs has provided new insights into catalytic function in general and has important implications for our understanding of the origin and evolution of life on this planet, a topic discussed in Section 26.3.

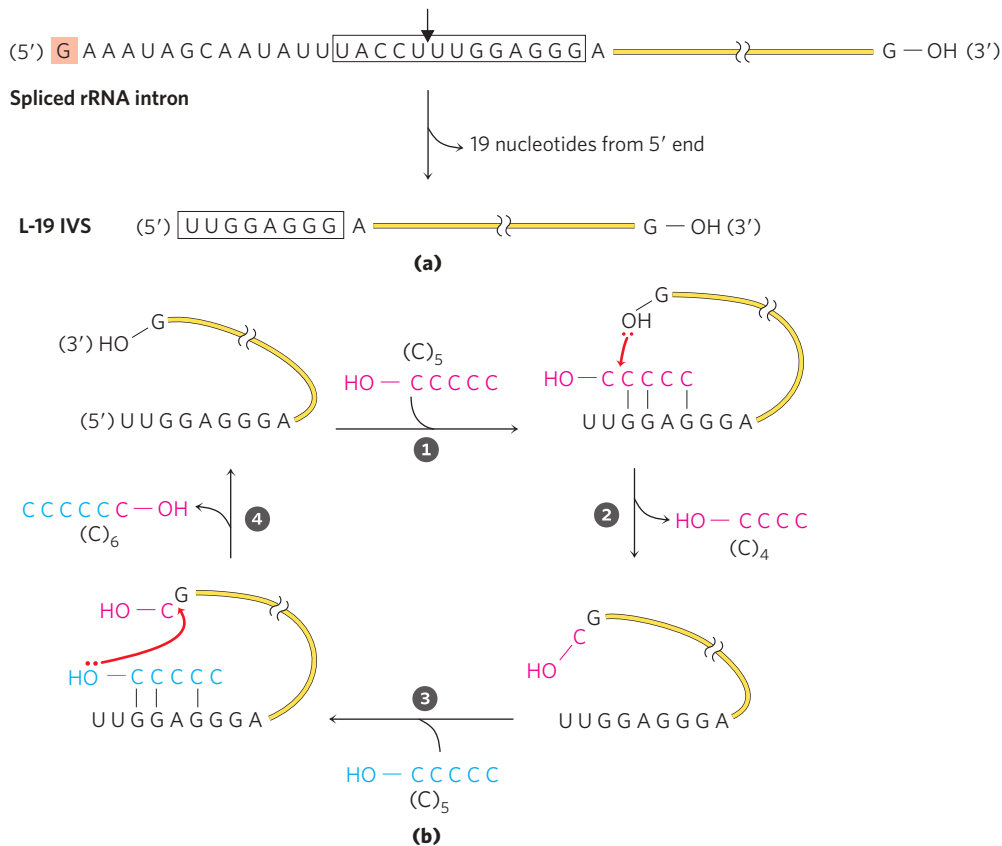


FIGURE 26-30 In vitro catalytic activity of L-19 IVS. (a) L-19 IVS is generated by the autocatalytic removal of 19 nucleotides from the 5' end of the spliced *Tetrahymena* intron. The cleavage site is indicated by the arrow in the internal guide sequence (boxed). The G residue (shaded light red) added in the first step of the splicing reaction (see Fig. 26-14) is part of the removed sequence. A portion of the internal guide sequence remains at the 5' end of L-19 IVS. (b) L-19 IVS lengthens some RNA oligonucleotides at the expense of others in a cycle of transesterification reactions

(steps 1 through 4). The 3' OH of the G residue at the 3' end of L-19 IVS plays a key role in this cycle (note that this is *not* the G residue added in the splicing reaction). (C)₅ is one of the ribozyme's better substrates because it can base-pair with the guide sequence remaining in the intron. Although this catalytic activity is probably irrelevant to the cell, it has important implications for current hypotheses on evolution, discussed at the end of this chapter.

Cellular mRNAs Are Degraded at Different Rates

The expression of genes is regulated at many levels. A crucial factor governing a gene's expression is the cellular concentration of its associated mRNA. The concentration of any molecule depends on two factors: its rate of synthesis and its rate of degradation. When synthesis and degradation of an mRNA are balanced, the concentration of the mRNA remains in a steady state. A change in either rate will lead to net accumulation or depletion of the mRNA. Degradative pathways ensure that mRNAs do not build up in the cell and direct the synthesis of unnecessary proteins.

The rates of degradation vary greatly for mRNAs from different eukaryotic genes. For a gene product that is needed only briefly, the half-life of its mRNA may be only minutes or even seconds. Gene products needed constantly by the cell may have mRNAs that are stable over many cell generations. The average half-life of the mRNAs of a vertebrate cell is about 3 hours, with the pool of each type of mRNA turning over about 10 times per

cell generation. The half-life of bacterial mRNAs is much shorter—only about 1.5 min—perhaps because of regulatory requirements.

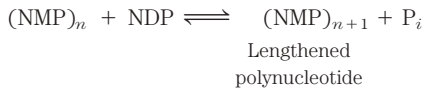
Messenger RNA is degraded by ribonucleases present in all cells. In *E. coli*, the process begins with one or several cuts by an endoribonuclease, followed by 3'→5' degradation by exoribonucleases. In lower eukaryotes, the major pathway involves first shortening the poly(A) tail, then decapping the 5' end and degrading the mRNA in the 5'→3' direction. A 3'→5' degradative pathway also exists and may be the major path in higher eukaryotes. All eukaryotes have a complex of up to 10 conserved 3'→5' exoribonucleases, called the **exosome**, which is involved in the processing of the 3' end of rRNAs, tRNAs, and some special-function RNAs (including snRNAs and snoRNAs), as well as the degradation of mRNAs.

A hairpin structure in bacterial mRNAs with a ρ -independent terminator (Fig. 26-7) confers stability against degradation. Similar hairpin structures can make some parts of a primary transcript more stable, leading to

nonuniform degradation of transcripts. In eukaryotic cells, both the 3' poly(A) tail and the 5' cap are important to the stability of many mRNAs. 🌊 **Life Cycle of an mRNA**

Polynucleotide Phosphorylase Makes Random RNA-like Polymers

In 1955, Marianne Grunberg-Manago and Severo Ochoa discovered the bacterial enzyme **polynucleotide phosphorylase**, which in vitro catalyzes the reaction



Polynucleotide phosphorylase was the first nucleic acid-synthesizing enzyme discovered (Arthur Kornberg's discovery of DNA polymerase followed soon thereafter). The reaction catalyzed by polynucleotide phosphorylase differs fundamentally from the polymerase activities discussed so far in that it is not template-dependent. The enzyme uses the 5'-diphosphates of ribonucleosides as substrates and cannot act on the homologous 5'-triphosphates or on deoxyribonucleoside 5'-diphosphates. The RNA polymer formed by polynucleotide phosphorylase contains the usual 3',5'-phosphodiester linkages, which can be hydrolyzed by ribonuclease. The reaction is readily reversible and can be pushed in the direction of breakdown of the polyribonucleotide by increasing the phosphate concentration. The probable function of this enzyme in the cell is the degradation of mRNAs to nucleoside diphosphates.



Marianne Grunberg-Manago



Severo Ochoa, 1905–1993

Because the polynucleotide phosphorylase reaction does not use a template, the polymer it forms does not have a specific base sequence. The reaction proceeds equally well with any or all of the four nucleoside diphosphates, and the base composition of the resulting polymer reflects nothing more than the relative concentrations of the 5'-diphosphate substrates in the medium.

Polynucleotide phosphorylase can be used in the laboratory to prepare RNA polymers with many different base sequences and frequencies. Synthetic RNA polymers of this sort were critical for deducing the genetic code for the amino acids (Chapter 27).

SUMMARY 26.2 RNA Processing

- ▶ Eukaryotic mRNAs are modified by addition of a 7-methylguanosine residue at the 5' end and by cleavage and polyadenylation at the 3' end to form a long poly(A) tail.
- ▶ Many primary mRNA transcripts contain introns (noncoding regions), which are removed by splicing. Excision of the group I introns found in some rRNAs requires a guanosine cofactor. Some group I and group II introns are capable of self-splicing; no protein enzymes are required. Nuclear mRNA precursors have a third (the largest) class of introns, which are spliced with the aid of RNA-protein complexes called snRNPs, assembled into spliceosomes. A fourth class of introns, found in some tRNAs, consists of the only introns known to be spliced by protein enzymes.
- ▶ The function of many eukaryotic mRNAs is regulated by complementary microRNAs (miRNAs). The miRNAs are themselves derived from larger precursors through a series of processing reactions.
- ▶ Ribosomal RNAs and transfer RNAs are derived from longer precursor RNAs, trimmed by nucleases. Some bases are modified enzymatically during the maturation process. Some nucleoside modifications are guided by snoRNAs, within protein complexes called snoRNPs.
- ▶ The self-splicing introns and the RNA component of RNase P (which cleaves the 5' end of tRNA precursors) are two examples of ribozymes. These biological catalysts have the properties of true enzymes. They generally promote hydrolytic cleavage and transesterification, using RNA as substrate. Combinations of these reactions can be promoted by the excised group I intron of *Tetrahymena* rRNA, resulting in a type of RNA polymerization reaction.
- ▶ Polynucleotide phosphorylase reversibly forms RNA-like polymers from ribonucleoside 5'-diphosphates, adding or removing ribonucleotides at the 3'-hydroxyl end of the polymer. The enzyme degrades RNA in vivo.

26.3 RNA-Dependent Synthesis of RNA and DNA

In our discussion of DNA and RNA synthesis up to this point, the role of the template strand has been reserved for DNA. However, some enzymes use an RNA template for nucleic acid synthesis. With the very important exception of viruses with an RNA genome, these enzymes play only a modest role in information pathways. RNA viruses are the source of most RNA-dependent polymerases characterized so far.

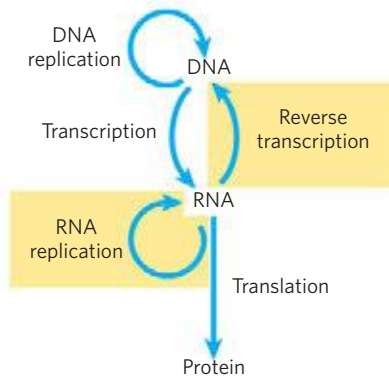


FIGURE 26-31 Extension of the central dogma to include RNA-dependent synthesis of RNA and DNA.

The existence of RNA replication requires an elaboration of the central dogma (**Fig. 26-31**; contrast this with the diagram on p. 977). The enzymes involved in RNA replication have profound implications for investigations into the nature of self-replicating molecules that may have existed in prebiotic times.

Reverse Transcriptase Produces DNA from Viral RNA

Certain RNA viruses that infect animal cells carry within the viral particle an RNA-dependent DNA polymerase called **reverse transcriptase**. On infection, the single-stranded RNA viral genome (~10,000 nucleotides) and the enzyme enter the host cell. The reverse transcriptase first catalyzes the synthesis of a DNA strand complementary to the viral RNA (**Fig. 26-32**), then degrades the RNA strand of the viral RNA-DNA hybrid and replaces it with DNA. The resulting duplex DNA often becomes incorporated into the genome of the

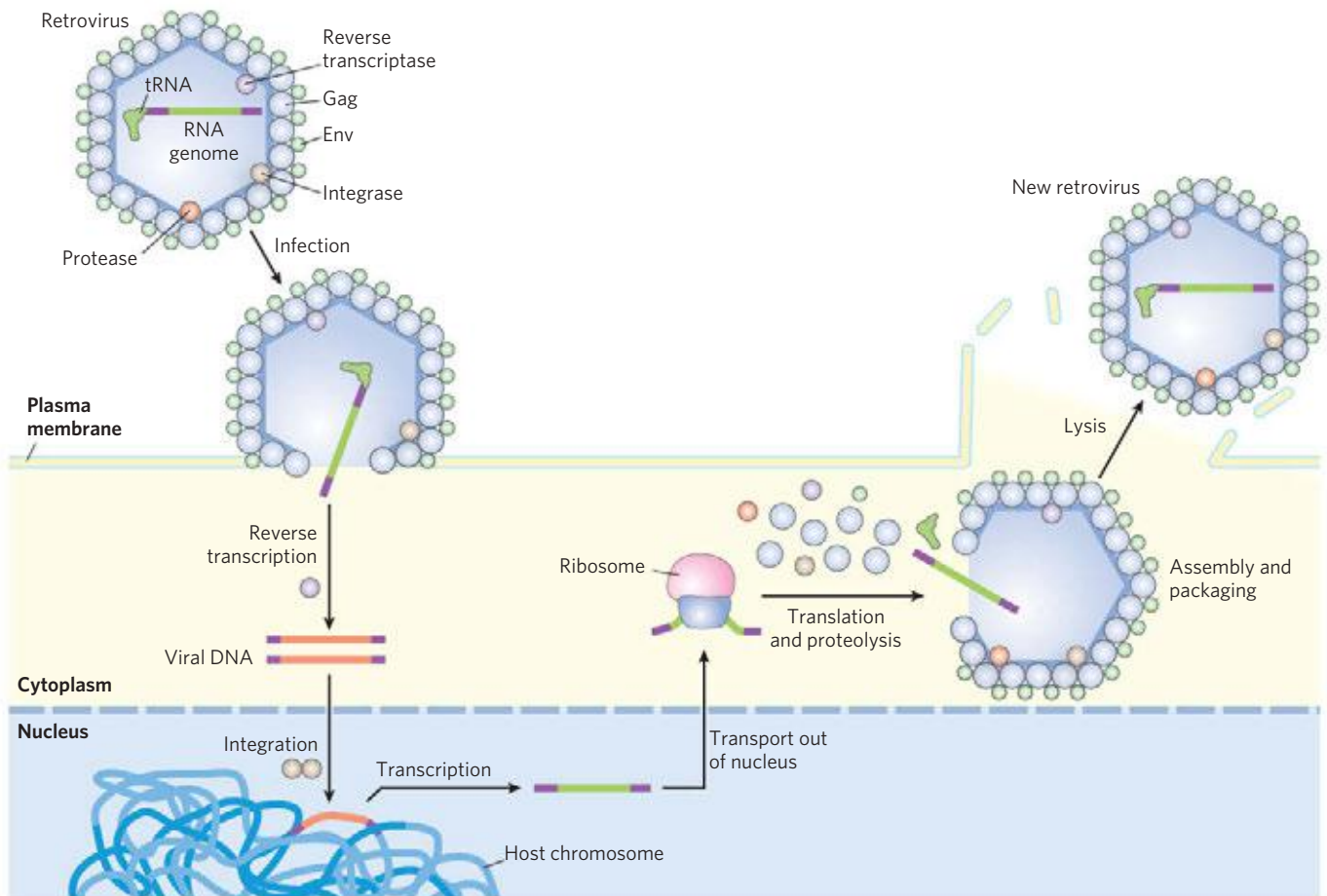


FIGURE 26-32 Retroviral infection of a mammalian cell and integration of the retrovirus into the host chromosome. Viral particles entering the host cell carry viral reverse transcriptase and a cellular tRNA (picked up from a former host cell) already base-paired to the viral RNA. The purple segments represent the long terminal repeats on the viral RNA. The tRNA facilitates immediate conversion of viral RNA to double-stranded DNA by the action of reverse transcriptase, as described in the text. Once converted to double-stranded DNA, the DNA enters the nucleus and is integrated into the host genome. The integration is catalyzed by a virally encoded integrase. Integration of viral DNA into host DNA is mechanis-

tically similar to the insertion of transposons in bacterial chromosomes (see Fig. 25-41). For example, a few base pairs of host DNA become duplicated at the site of integration, forming short repeats of 4 to 6 bp at each end of the inserted retroviral DNA (not shown). On transcription and translation of the integrated viral DNA, new viruses are formed and released by cell lysis (right). In the viruses, the viral RNA is enclosed by capsid proteins called Gag and outer envelope proteins called Env. Additional viral proteins (reverse transcriptase, integrase, and a viral protease needed for posttranslational processing of viral proteins) are packaged within the virus particle with the RNA.

viral genes can be activated and transcribed, and the gene products—viral proteins and the viral RNA genome itself—packaged as new viruses. The RNA viruses that contain reverse transcriptases are known as **retroviruses** (*retro* is the Latin prefix for “backward”).

The existence of reverse transcriptases in RNA viruses was predicted by Howard Temin in 1962, and the enzymes were ultimately detected by Temin and, independently, by David Baltimore in 1970. Their discovery aroused much attention as dogma-shaking proof that genetic information can flow “backward” from RNA to DNA.



Howard Temin, 1934-1994



David Baltimore

Retroviruses typically have three genes: *gag* (a name derived from the historical designation *group associated antigen*), *pol*, and *env* (Fig. 26-33). The transcript that contains *gag* and *pol* is translated into a long “polyprotein,” a single large polypeptide that is cleaved into six proteins with distinct functions. The proteins derived from the *gag* gene make up the interior core of the viral particle. The *pol* gene encodes the protease that cleaves the long polypeptide, an integrase that inserts the viral DNA into the host chromosomes, and reverse transcriptase. Many reverse transcriptases have two subunits, α and β . The *pol* gene specifies the β subunit (M_r 90,000), and the α subunit (M_r 65,000) is simply a proteolytic fragment of the β subunit. The *env* gene encodes the proteins of the viral envelope. At each end of the linear RNA genome are long terminal repeat (LTR) sequences of a few hundred nucleotides. Transcribed into the duplex DNA, these sequences facilitate integration of the viral chromosome into the host DNA and contain promoters for viral gene expression.

Reverse transcriptases catalyze three different reactions: (1) RNA-dependent DNA synthesis, (2) RNA degradation, and (3) DNA-dependent DNA synthesis. Like many DNA and RNA polymerases, reverse transcriptases contain Zn^{2+} . Each transcriptase is most active with the RNA of its own virus, but each can be used experimentally to make DNA complementary to a variety of RNAs. The DNA and RNA synthesis and RNA degradation activities use separate active sites on the protein. For DNA synthesis to begin, the reverse transcriptase requires a primer, a cellular tRNA obtained during an earlier infection and carried in the viral particle. This

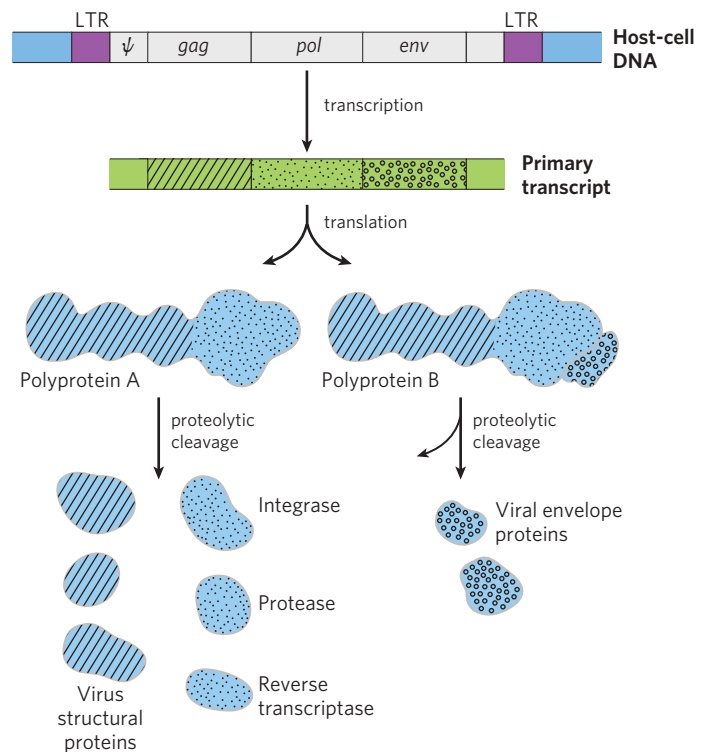


FIGURE 26-33 Structure and gene products of an integrated retroviral genome. The long terminal repeats (LTRs) have sequences needed for the regulation and initiation of transcription. The sequence denoted ψ is required for packaging of retroviral RNAs into mature viral particles. Transcription of the retroviral DNA produces a primary transcript encompassing the *gag*, *pol*, and *env* genes. Translation (Chapter 27) produces a polyprotein, a single long polypeptide derived from the *gag* and *pol* genes, which is cleaved into six distinct proteins. Splicing of the primary transcript yields an mRNA derived largely from the *env* gene, which is also translated into a polyprotein, then cleaved to generate viral envelope proteins.

tRNA is base-paired at its 3' end with a complementary sequence in the viral RNA. The new DNA strand is synthesized in the 5'→3' direction, as in all RNA and DNA polymerase reactions. Reverse transcriptases, like RNA polymerases, do not have 3'→5' proofreading exonucleases. They generally have error rates of about 1 per 20,000 nucleotides added. An error rate this high is extremely unusual in DNA replication and seems to be a feature of most enzymes that replicate the genomes of RNA viruses. A consequence is a higher mutation rate and faster rate of viral evolution, which is a factor in the frequent appearance of new strains of disease-causing retroviruses.

Reverse transcriptases have become important reagents in the study of DNA-RNA relationships and in DNA cloning techniques. They make possible the synthesis of DNA complementary to an mRNA template, and synthetic DNA prepared in this manner, called **complementary DNA (cDNA)**, can be used to clone cellular genes (see Fig. 9-14).



FIGURE 26-34 Rous sarcoma virus genome. The *src* gene encodes a tyrosine kinase, one of a class of enzymes that function in systems affecting cell division, cell-cell interactions, and intercellular communication (Chapter 12). The same gene is found in chicken DNA (the usual host

for this virus) and in the genomes of many other eukaryotes, including humans. When associated with the Rous sarcoma virus, this oncogene is often expressed at abnormally high levels, contributing to unregulated cell division and cancer.

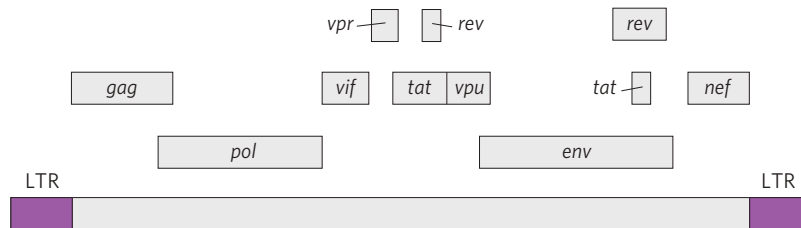


FIGURE 26-35 The genome of HIV, the virus that causes AIDS. In addition to the typical retroviral genes, HIV contains several small genes with a variety of functions (not identified here and not all known). Some

of these genes overlap. Alternative splicing mechanisms produce many different proteins from this small (9.7×10^3 nucleotides) genome.

Some Retroviruses Cause Cancer and AIDS



Retroviruses have featured prominently in recent advances in the molecular understanding of cancer. Most retroviruses do not kill their host cells but remain integrated in the cellular DNA, replicating when the cell divides. Some retroviruses, classified as RNA tumor viruses, contain an oncogene that can cause the cell to grow abnormally. The first retrovirus of this type to be studied was the Rous sarcoma virus (also called avian sarcoma virus; **Fig. 26-34**), named for F. Peyton Rous, who studied chicken tumors now known to be caused by this virus. Since the initial discovery of oncogenes by Harold Varmus and Michael Bishop, many dozens of such genes have been found in retroviruses.

The human immunodeficiency virus (HIV), which causes acquired immune deficiency syndrome (AIDS), is a retrovirus. Identified in 1983, HIV has an RNA genome with standard retroviral genes along with several other unusual genes (**Fig. 26-35**). Unlike many other retroviruses, HIV kills many of the cells it infects (principally T lymphocytes) rather than causing tumor formation. This gradually leads to suppression of the immune system in the host organism. The reverse transcriptase of HIV is even more error-prone than other known reverse transcriptases—10 times more so—resulting in high mutation rates in this virus. One or more errors are generally made every time the viral genome is replicated, so any two viral RNA molecules are likely to differ.

Many modern vaccines for viral infections consist of one or more coat proteins of the virus, produced by methods described in Chapter 9. These proteins are not infectious on their own but stimulate the immune system to recognize and resist subsequent viral invasions (Chapter 5). Because of the high error rate of the HIV reverse transcriptase, the *env* gene in this virus (along

with the rest of the genome) undergoes very rapid mutation, complicating the development of an effective vaccine. However, repeated cycles of cell invasion and replication are needed to propagate an HIV infection, so inhibition of viral enzymes offers the most effective therapy currently available. The HIV protease is targeted by a class of drugs called protease inhibitors (see **Fig. 6-24**). Reverse transcriptase is the target of some additional drugs widely used to treat HIV-infected individuals (**Box 26-2**). ■

Many Transposons, Retroviruses, and Introns May Have a Common Evolutionary Origin

Some well-characterized eukaryotic DNA transposons from sources as diverse as yeast and fruit flies have a structure very similar to that of retroviruses; these are sometimes called retrotransposons (**Fig. 26-36**). Retrotransposons encode an enzyme homologous to the retroviral reverse transcriptase, and their coding regions are flanked by LTR sequences. They transpose from one position to another in the cellular genome by means of an RNA intermediate, using reverse transcriptase to

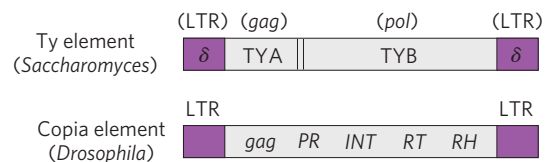


FIGURE 26-36 Eukaryotic transposons. The Ty element of the yeast *Saccharomyces* and the copia element of the fruit fly *Drosophila* serve as examples of eukaryotic retrotransposons, which often have a structure similar to retroviruses but lack the *env* gene. The δ sequences of the Ty element are functionally equivalent to retroviral LTRs. In the copia element, *INT* and *RT* are homologous to the integrase and reverse transcriptase segments, respectively, of the *pol* gene.

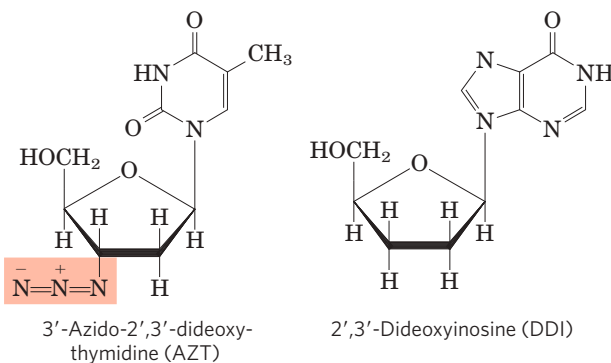
BOX 26-2 MEDICINE Fighting AIDS with Inhibitors of HIV Reverse Transcriptase

Research into the chemistry of template-dependent nucleic acid biosynthesis, combined with modern techniques of molecular biology, has elucidated the life cycle and structure of the human immunodeficiency virus, the retrovirus that causes AIDS. A few years after the isolation of HIV, this research resulted in the development of drugs capable of prolonging the lives of people infected by HIV.

The first drug to be approved for clinical use was AZT, a structural analog of deoxythymidine. AZT was first synthesized in 1964 by Jerome P. Horwitz. It failed as an anticancer drug (the purpose for which it was made), but in 1985 it was found to be a useful treatment for AIDS. AZT is taken up by T lymphocytes, immune system cells that are particularly vulnerable to

HIV infection, and converted to AZT triphosphate. (AZT triphosphate taken directly would be ineffective because it cannot cross the plasma membrane.) HIV's reverse transcriptase has a higher affinity for AZT triphosphate than for dTTP, and binding of AZT triphosphate to this enzyme competitively inhibits dTTP binding. When AZT is added to the 3' end of the growing DNA strand, lack of a 3' hydroxyl means that the DNA strand is terminated prematurely and viral DNA synthesis grinds to a halt.

AZT triphosphate is not as toxic to the T lymphocytes themselves because *cellular* DNA polymerases have a lower affinity for this compound than for dTTP. At concentrations of 1 to 5 μM , AZT affects HIV reverse transcription but not most cellular DNA replication. Unfortunately, AZT seems to be toxic to the bone marrow cells that are the progenitors of erythrocytes, and many individuals taking AZT develop anemia. AZT can increase the survival time of people with advanced AIDS by about a year, and it delays the onset of AIDS in those who are still in the early stages of HIV infection. Some other AIDS drugs, such as dideoxyinosine (DDI), have a similar mechanism of action. Newer drugs target and inactivate the HIV protease. Because of the high error rate of HIV reverse transcriptase and the resulting rapid evolution of HIV, the most effective treatments of HIV infection use a combination of drugs directed at both the protease and the reverse transcriptase.



make a DNA copy of the RNA, followed by integration of the DNA at a new site. Most transposons in eukaryotes use this mechanism for transposition, distinguishing them from bacterial transposons, which move as DNA directly from one chromosomal location to another (see Fig. 25-41).

Retrotransposons lack an *env* gene and so cannot form viral particles. They can be thought of as defective viruses, trapped in cells. Comparisons between retroviruses and eukaryotic transposons suggest that reverse transcriptase is an ancient enzyme that predates the evolution of multicellular organisms.

Interestingly, many group I and group II introns are also mobile genetic elements. In addition to their self-splicing activities, they encode DNA endonucleases that promote their movement. During genetic exchanges between cells of the same species, or when DNA is introduced into a cell by parasites or by other means, these endonucleases promote insertion of the intron into an identical site in another DNA copy of a homologous gene that does not contain the intron, in a process termed **homing** (Fig. 26-37). Whereas group I intron homing is DNA-based, group II intron homing occurs

through an RNA intermediate. The endonucleases of the group II introns have associated reverse transcriptase activity. The proteins can form complexes with the intron RNAs themselves, after the introns are spliced from the primary transcripts. Because the homing process involves insertion of the RNA intron into DNA and reverse transcription of the intron, the movement of these introns has been called retrohoming. Over time, every copy of a particular gene in a population may acquire the intron. Much more rarely, the intron may insert itself into a new location in an unrelated gene. If this event does not kill the host cell, it can lead to the evolution and distribution of an intron in a new location. The structures and mechanisms used by mobile introns support the idea that at least some introns originated as molecular parasites whose evolutionary past can be traced to retroviruses and transposons.

Telomerase Is a Specialized Reverse Transcriptase

Telomeres, the structures at the ends of linear eukaryotic chromosomes (see Fig. 24-8), generally consist of many tandem copies of a short oligonucleotide sequence.

FIGURE 26-37 Introns that move: homing and retrohoming. Certain introns include a gene (shown in red) for enzymes that promote homing (certain group I introns) or retrohoming (certain group II introns). **(a)** The gene in the spliced intron is bound by a ribosome and translated. Group I homing introns specify a site-specific endonuclease, called a homing endonuclease. Group II retrohoming introns specify a protein with both endonuclease and reverse transcriptase activities (not shown here).

(b) Homing. Allele *a* of a gene *X* containing a group I homing intron is present in a cell containing allele *b* of the same gene, which lacks the intron. The homing endonuclease produced by *a* cleaves *b* at the position corresponding to the intron in *a*, and double-strand break repair (recombination with allele *a*; see Fig. 25-35) then creates a new copy of the intron in *b*. **(c)** Retrohoming. Allele *a* of gene *Y* contains a retrohoming group II intron; allele *b* lacks the intron. The spliced intron inserts itself into the coding strand of *b* in a reaction that is the reverse of the splicing that excised the intron from the primary transcript (see Fig. 26-15), except that here the insertion is into DNA rather than RNA. The noncoding DNA strand of *b* is then cleaved by the intron-encoded endonuclease/reverse transcriptase. This same enzyme uses the inserted RNA as a template to synthesize a complementary DNA strand. The RNA is then degraded by cellular ribonucleases and replaced with DNA.

This sequence usually has the form T_xG_y in one strand and C_yA_x in the complementary strand, where *x* and *y* are typically in the range of 1 to 4 (p. 984). Telomeres vary in length from a few dozen base pairs in some ciliated protozoans to tens of thousands of base pairs in mammals. The TG strand is longer than its complement, leaving a region of single-stranded DNA of up to a few hundred nucleotides at the 3' end.

The ends of a linear chromosome are not readily replicated by cellular DNA polymerases. DNA replication requires a template and primer, and beyond the end of a linear DNA molecule no template is available for the pairing of an RNA primer. Without a special mechanism for replicating the ends, chromosomes would be shortened somewhat in each cell generation. The enzyme **telomerase**, discovered by Carol Greider and Elizabeth Blackburn, solves this problem by adding telomeres to chromosome ends.

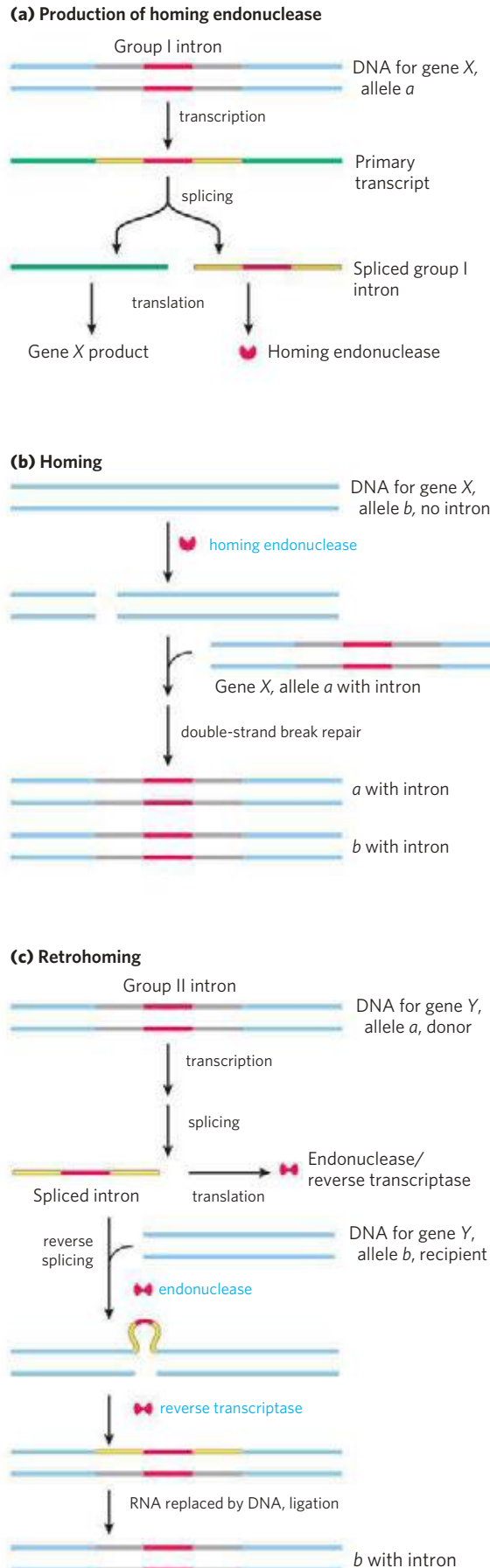


Carol Greider



Elizabeth Blackburn

The discovery and purification of this enzyme provided insight into a reaction mechanism that is remarkable and unprecedented. Telomerase, like some other



enzymes described in this chapter, contains both RNA and protein components. The RNA component is about 150 nucleotides long and contains about 1.5 copies of the appropriate C_yA_x telomere repeat. This region of the RNA acts as a template for synthesis of the T_xG_y strand of the telomere. Telomerase thereby acts as a cellular reverse transcriptase that provides the active site for RNA-dependent DNA synthesis. Unlike retroviral reverse transcriptases, telomerase copies only a small segment of RNA that it carries within itself. Telomere synthesis requires the 3' end of a chromosome as primer and proceeds in the usual 5'→3' direction. Having synthesized one copy of the repeat, the enzyme repositions to resume extension of the telomere (Fig. 26–38a).

After extension of the T_xG_y strand by telomerase, the complementary C_yA_x strand is synthesized by cellular DNA polymerases, starting with an RNA primer (see Fig. 25–12). The single-stranded region is protected by specific binding proteins in many lower eukaryotes, especially those species with telomeres of less than a few hundred base pairs. In higher eukaryotes (including mammals) with telomeres many thousands of base pairs long, the single-stranded end is sequestered in a specialized structure called a **T loop** (Fig. 26–38b). The single-stranded end is folded back and paired with its complement in the double-stranded portion of the telomere. The formation of a T loop involves invasion of the 3' end of the telomere's single strand into the duplex

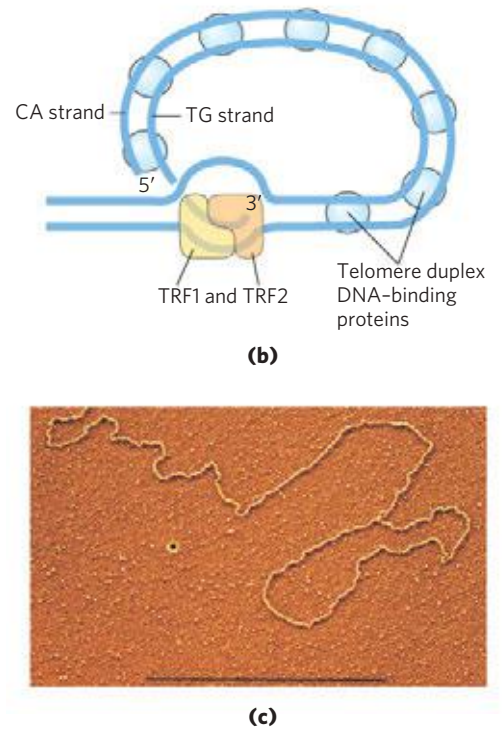
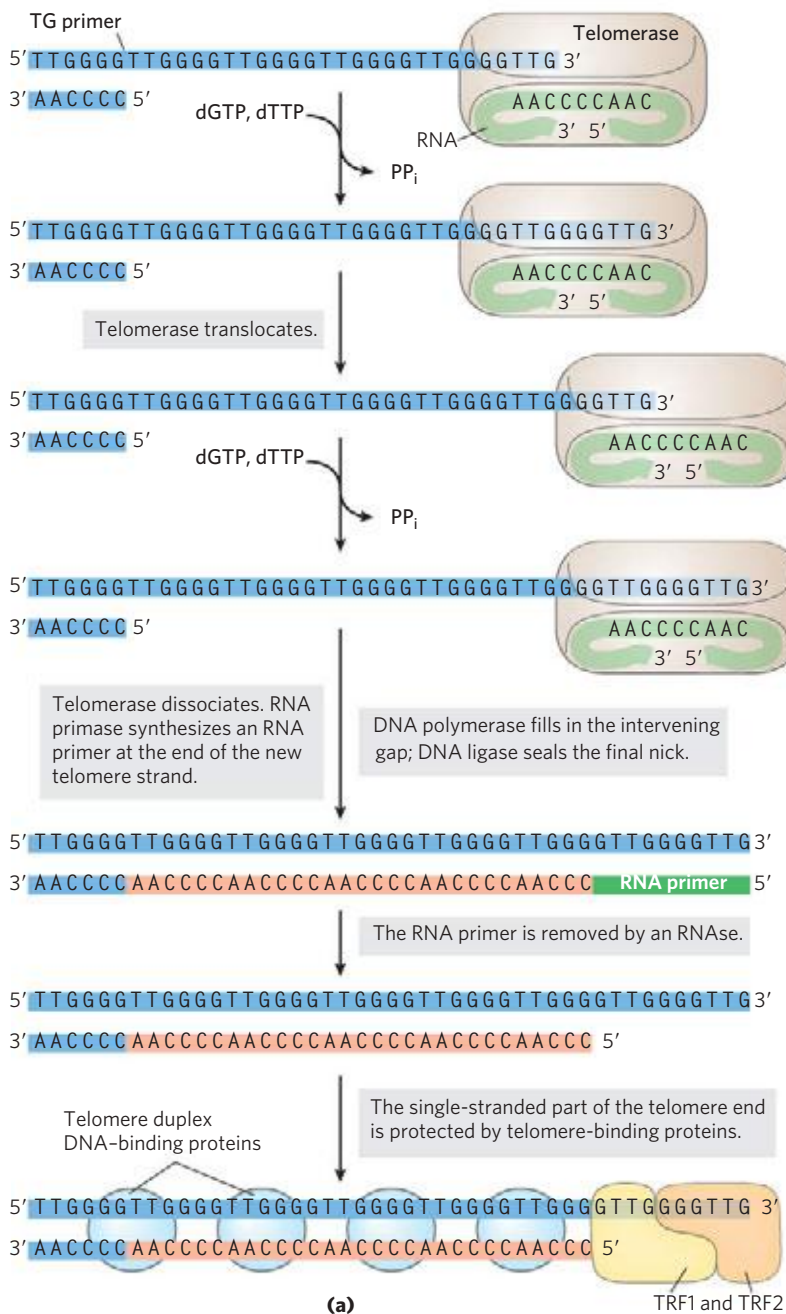


FIGURE 26–38 Telomere synthesis and structure. (a) The internal template RNA of telomerase binds to and base-pairs with the TG primer (T_xG_y) of DNA. Telomerase adds more T and G residues to the TG primer, then repositions the internal template RNA to allow the addition of more T and G residues that generate the TG strand of the telomere. The complementary strand is synthesized by cellular DNA polymerases after priming by an RNA primase. (b) Proposed structure of T loops in telomeres. The single-stranded tail synthesized by telomerase is folded back and paired with its complement in the duplex portion of the telomere. The telomere is bound by several telomere-binding proteins, including TRF1 and TRF2 (telomere repeat binding factors). (c) Electron micrograph of a T loop at the end of a chromosome isolated from a mouse hepatocyte. The bar at the bottom of the micrograph represents a length of 5,000 bp.

DNA, perhaps by a mechanism similar to the initiation of homologous genetic recombination (see Fig. 25–35). In mammals, the looped DNA is bound by two proteins, TRF1 and TRF2, with the latter protein involved in formation of the T loop. T loops protect the 3' ends of chromosomes, making them inaccessible to nucleases and the enzymes that repair double-strand breaks.

In protozoans (such as *Tetrahymena*), loss of telomerase activity results in a gradual shortening of telomeres with each cell division, ultimately leading to the death of the cell line. A similar link between telomere length and cell senescence (cessation of cell division) has been observed in humans. In germ-line cells, which contain telomerase activity, telomere lengths are maintained; in somatic cells, which lack telomerase, they are not. There is a linear, inverse relationship between the length of telomeres in cultured fibroblasts and the age of the individual from whom the fibroblasts were taken: telomeres in human somatic cells gradually shorten as an individual ages. If the telomerase reverse transcriptase is introduced into human somatic cells in vitro, telomerase activity is restored and the cellular life span increases markedly.

Is the gradual shortening of telomeres a key to the aging process? Is our natural life span determined by the length of the telomeres we are born with? Further research in this area should yield some fascinating insights.

Some Viral RNAs Are Replicated by RNA-Dependent RNA Polymerase

Some *E. coli* bacteriophages, including $\phi 2$, MS2, R17, and Q β , as well as some eukaryotic viruses (including influenza and Sindbis viruses, the latter associated with a form of encephalitis) have RNA genomes. The single-stranded RNA chromosomes of these viruses, which also function as mRNAs for the synthesis of viral proteins, are replicated in the host cell by an **RNA-dependent RNA polymerase (RNA replicase)**. All RNA viruses—with the exception of retroviruses—must encode a protein with RNA-dependent RNA polymerase activity, because the host cells do not possess this enzyme.

The RNA replicase of most RNA bacteriophages has a molecular weight of $\sim 210,000$ and consists of four subunits. One subunit (M_r 65,000) is the product of the replicase gene encoded by the viral RNA and has the active site for replication. The other three subunits are host proteins normally involved in host-cell protein synthesis: the *E. coli* elongation factors Tu (M_r 45,000) and Ts (M_r 34,000) (which ferry amino acyl-tRNAs to the ribosomes) and the protein S1 (an integral part of the 30S ribosomal subunit). These three host proteins may help the RNA replicase locate and bind to the 3' ends of the viral RNAs.

The RNA replicase isolated from Q β -infected *E. coli* cells catalyzes the formation of an RNA complementary to the viral RNA, in a reaction equivalent to that catalyzed by DNA-dependent RNA polymerases.

New RNA strand synthesis proceeds in the 5'→3' direction by a chemical mechanism identical to that used in all other nucleic acid synthetic reactions that require a template. RNA replicase requires RNA as its template and will not function with DNA. It lacks a separate proofreading endonuclease activity and has an error rate similar to that of RNA polymerase. Unlike the DNA and RNA polymerases, RNA replicases are specific for the RNA of their own virus; the RNAs of the host cell are generally not replicated. This explains how RNA viruses are preferentially replicated in the host cell, which contains many other types of RNA.

RNA Synthesis Offers Important Clues to Biochemical Evolution

The extraordinary complexity and order that distinguish living from inanimate systems are key manifestations of fundamental life processes. Maintaining the living state requires that *selected* chemical transformations occur very rapidly—especially those that use environmental energy sources and synthesize elaborate or specialized cellular macromolecules. Life depends on powerful and selective catalysts—enzymes—and on informational systems capable of both securely storing the blueprint for these enzymes and accurately reproducing the blueprint for generation after generation. Chromosomes encode the blueprint not for the cell but for the enzymes that construct and maintain the cell. The parallel demands for information and catalysis present a classic conundrum: what came first, the information needed to specify structure or the enzymes needed to maintain and transmit the information?

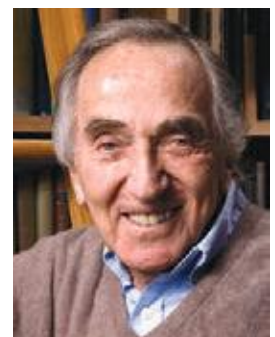
The unveiling of the structural and functional complexity of RNA led Carl Woese, Francis Crick, and Leslie Orgel to propose in the 1960s that this macromolecule might serve as both information carrier and catalyst. The discovery of catalytic RNAs took this proposal from conjecture to hypothesis and has led to widespread speculation that an “RNA world” might have been important in the transition from



Carl Woese



Francis Crick, 1916–2004



Leslie Orgel, 1927–2007

prebiotic chemistry to life (see Fig. 1–36). The parent of all life on this planet, in the sense that it could reproduce itself across the generations from the origin of life to the present, might have been a self-replicating RNA or a polymer with equivalent chemical characteristics.

How might a self-replicating polymer come to be? How might it maintain itself in an environment where the precursors for polymer synthesis are scarce? How could evolution progress from such a polymer to the modern DNA-protein world? These difficult questions can be addressed by careful experimentation, providing clues about how life on Earth began and evolved.

The probable origin of purine and pyrimidine bases is suggested by experiments designed to test hypotheses about prebiotic chemistry (pp. 33–34). Beginning with simple molecules thought to be present in the early atmosphere (CH_4 , NH_3 , H_2O , H_2), electrical discharges such as lightning generate, first, more reactive molecules such as HCN and aldehydes, then an array of amino acids and organic acids (see Fig. 1–34). When molecules such as HCN become abundant, purine and pyrimidine bases are synthesized in detectable amounts. Remarkably, a concentrated solution of ammonium cyanide, refluxed for a few days, generates adenine in yields of up to 0.5% (Fig. 26–39). Adenine may well have been the first and most abundant nucleotide constituent to appear on Earth. Intriguingly, most enzyme cofactors contain adenosine as part of their structure, although it plays no direct role in the cofactor function (see Fig. 8–38). This may suggest an evolutionary relationship. Based on the simple synthesis of adenine from cyanide, adenine may simply have been abundant and available.

The RNA world hypothesis requires a nucleotide polymer to reproduce itself. Can a ribozyme bring about its own synthesis in a template-directed manner? Researchers are getting closer to finding such a ribozyme or ribozyme system. For example, Gerald Joyce and colleagues, in 2009, reported on the first set of two ribozymes that could cross-catalyze each other's formation (Fig. 26–40). One ribozyme, E, catalyzes the joining of two oligonucleotides (A' and B') to create a second and complementary ribozyme called E' . E' could then catalyze the joining of two other oligonucleotides (A and B) to form another molecule of E. In this system, the formation of E and E' was templated, and the amounts grew exponentially as long as substrates were available and proteins were absent. The system evolved so that more efficient enzymes appeared in the population.

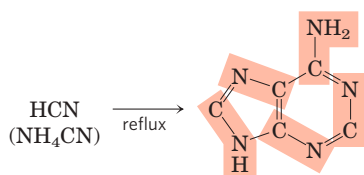


FIGURE 26–39 Possible prebiotic synthesis of adenine from ammonium cyanide. Adenine is derived from five molecules of cyanide, denoted by shading.

A more general RNA-polymerase-like ribozyme was described in 2011 by Philipp Holliger and colleagues. Indeed, the pace of discovery in this field is accelerating.

A self-replicating polymer would quickly use up available supplies of precursors provided by the relatively slow processes of prebiotic chemistry. Thus, from an early stage in evolution, metabolic pathways would be required to generate precursors efficiently, with the synthesis of precursors presumably catalyzed by ribozymes. The extant ribozymes found in nature have a limited repertoire of catalytic functions, and of the ribozymes that may once have existed, no trace is left. To explore the RNA world hypothesis more deeply, we need to know whether RNA has the potential to catalyze the many different reactions needed in a primitive system of metabolic pathways.

The search for RNAs with new catalytic functions has been aided by the development of a method that rapidly searches pools of random polymers of RNA and extracts those with particular activities: **SELEX** is nothing less than accelerated evolution in a test tube (Box 26–3). It has been used to generate RNA molecules that bind to amino acids, organic dyes, nucleotides, cyanocobalamin, and other molecules. Researchers have isolated ribozymes that catalyze ester and amide bond formation, $\text{S}_\text{N}2$ reactions, metallation of (addition of metal ions to) porphyrins, and carbon–carbon bond formation. The evolution of enzymatic cofactors with nucleotide “handles” that facilitate their binding to ribozymes might have further expanded the repertoire of chemical processes available to primitive metabolic systems.

As we shall see in the next chapter, some natural RNA molecules catalyze the formation of peptide bonds, offering an idea of how the RNA world might have been transformed by the greater catalytic potential of proteins. The synthesis of proteins would have been a major event in the evolution of the RNA world but would also have hastened its demise. The information-carrying role of RNA may have passed to DNA because DNA is chemically more stable. RNA replicase and reverse transcriptase may be modern versions of enzymes that once played important roles in making the transition to the modern DNA-based system.

The RNA world hypothesis is now supported by at least five lines of evidence. First, RNA catalysts exist. RNA thus clearly has the capacity to both store information and catalyze reactions. Second, the wide range of reactions for which RNA catalysts have been developed is sufficient to form the basis of a primordial metabolic system. Third, there are numerous extant RNAs that are likely vestiges of an RNA world, including ribozymes, retroviruses, RNA viruses, and retrotransposons. Fourth, an RNA catalyst is responsible for the synthesis of proteins (Chapter 27). Finally, RNA catalysts with a capacity for self-replication are coming to light in the laboratory. An active field of investigation—prebiotic chemistry (not described here)—is revealing

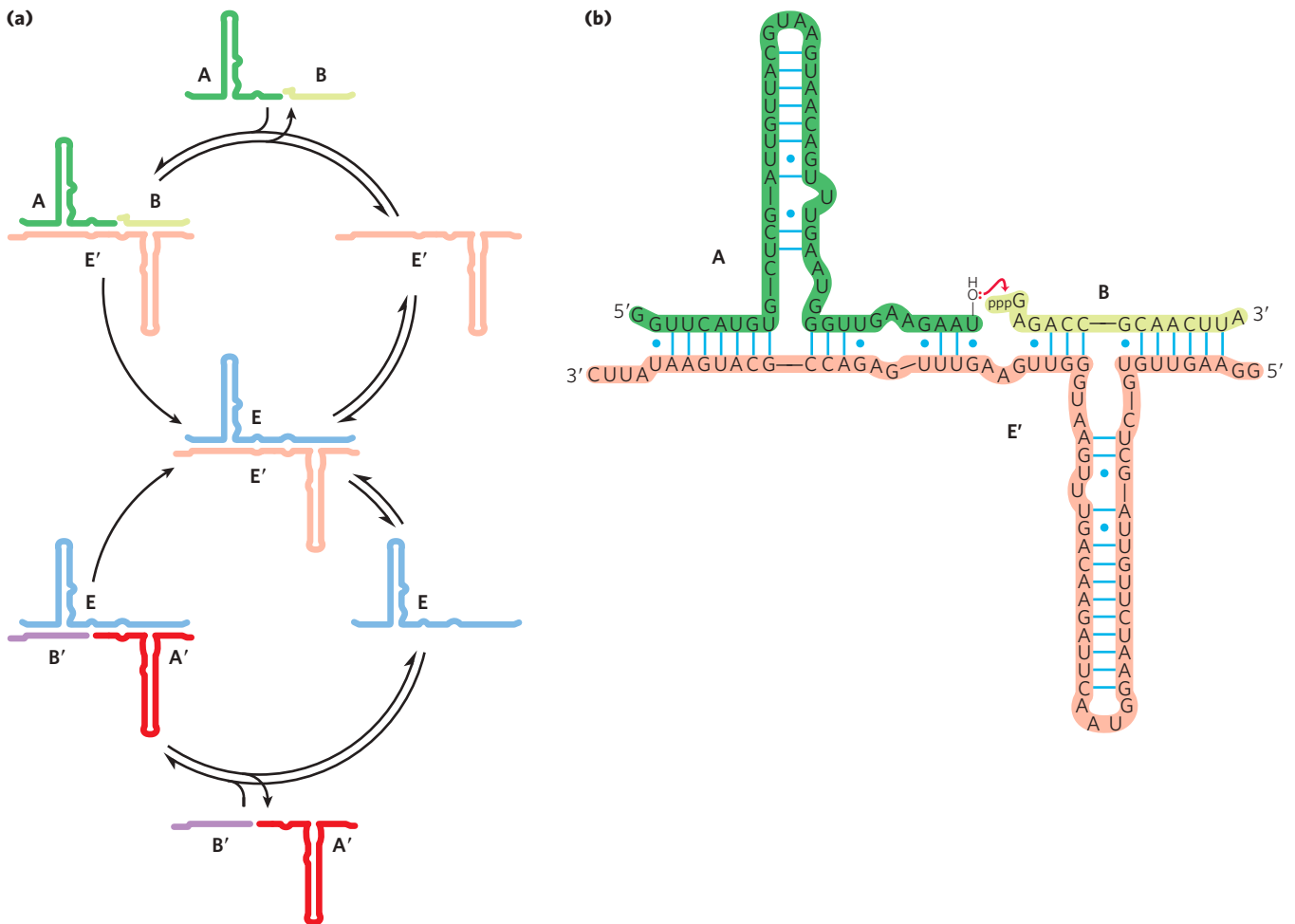


FIGURE 26-40 Self-sustained replication of an RNA enzyme. This system has many of the properties of a living system. The RNA molecules incorporate information and catalytic function, and the reactions produce an exponential increase in the product RNAs. When variants of the RNA substrates are introduced, the system undergoes natural selection such that the best replicators come to dominate the population. **(a)** The reaction scheme is outlined. Oligoribonucleotides A and B anneal to ribozyme E' and are ligated catalytically to form ribozyme E. The joining of oligo-

ribonucleotides A' and B' is similarly catalyzed by ribozyme E. The levels of E and E' grow exponentially, with a doubling time of about one hour at 42°C, as long as there is a supply of the precursors A, B, A', and B'. **(b)** The ligation reaction involves attack of the 3' OH of one oligoribonucleotide on the α -phosphate of the 5'-triphosphate of the other oligoribonucleotide. Pyrophosphate is released. Base pairing of the substrates with the ribozyme plays a key role in aligning the substrates for the reaction.

plausible pathways for the appearance of nucleotide precursors in the prebiotic soup.

Molecular parasites may also have originated in an RNA world. With the appearance of the first inefficient self-replicators, transposition could have been a potentially important alternative to replication as a strategy for successful reproduction and survival. Early parasitic RNAs would simply hop into a self-replicating molecule via catalyzed transesterification and then passively undergo replication. Natural selection would have driven transposition to become site-specific, targeting sequences that did not interfere with the catalytic activities of the host RNA. Replicators and RNA transposons could have existed in a primitive symbiotic relationship, each contributing to the evolution of the other. Modern introns, retroviruses, and transposons may all be vestiges of a "piggyback" strategy pursued by early parasitic RNAs.

These elements continue to make major contributions to the evolution of their hosts.

Although the RNA world remains a hypothesis, with many gaps yet to be explained, experimental evidence supports a growing list of its key elements. Further experimentation should increase our understanding. Important clues to the puzzle will be found in the workings of fundamental chemistry, in living cells, and perhaps on other planets. Meanwhile, the extant RNA universe continues to expand (Box 26-4).

SUMMARY 26.3 RNA-Dependent Synthesis of RNA and DNA

- ▶ RNA-dependent DNA polymerases, also called reverse transcriptases, were first discovered in retroviruses, which must convert their RNA

BOX 26-3 METHODS The SELEX Method for Generating RNA Polymers with New Functions

SELEX (systematic evolution of ligands by exponential enrichment) is used to generate **aptamers**, oligonucleotides selected to tightly bind a specific molecular target. The process is generally automated to allow rapid identification of one or more aptamers with the desired binding specificity.

Figure 1 illustrates how SELEX is used to select an RNA species that binds tightly to ATP. In step ①, a random mixture of RNA polymers is subjected to “unnatural selection” by passing it through a resin to which ATP is attached. The practical limit for the complexity of an RNA mixture in SELEX is about 10^{15} different sequences, which allows for the complete randomization of 25 nucleotides ($4^{25} = 10^{15}$). For longer RNAs, the RNA pool used to initiate the search does not include all possible sequences. ② RNA polymers that pass through the column are discarded; ③ those that bind to ATP are washed from the column with salt solution and collected. ④ The collected RNA polymers are amplified by reverse transcriptase to make many DNA complements to the selected RNAs; then an RNA polymerase makes many RNA complements of the resulting DNA molecules. ⑤ This new pool of RNA is subjected to the same selection procedure, and the cycle is repeated a dozen or more times. At the end, only a few aptamers—in this case, RNA sequences with considerable affinity for ATP—remain.

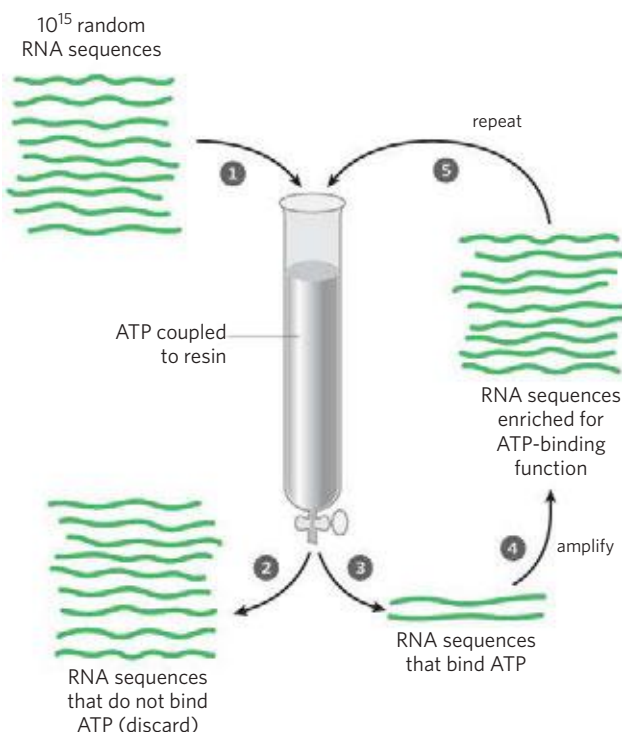


FIGURE 1 The SELEX procedure.

Critical sequence features of an RNA aptamer that binds ATP are shown in Figure 2; molecules with this general structure bind ATP (and other adenosine nucleotides) with $K_d < 50 \mu\text{M}$. Figure 3 presents the three-dimensional structure of a 36 nucleotide RNA aptamer (shown as a complex with AMP) generated by SELEX. This RNA has the backbone structure shown in Figure 2.

In addition to its use in exploring the potential functionality of RNA, SELEX has an important practical side in identifying short RNAs with pharmaceutical uses. Finding an aptamer that binds specifically to every potential therapeutic target may be impossible, but the capacity of SELEX to rapidly select and amplify a specific oligonucleotide sequence from a highly complex pool of sequences makes this a promising approach for the generation of new therapies. For example, one could select an RNA that binds tightly to a receptor protein prominent in the plasma membrane of cells in a particular cancerous tumor. Blocking the activity of the receptor, or targeting a toxin to the tumor cells by attaching it to the aptamer, would kill the cells. SELEX also has been used to select DNA aptamers that detect anthrax spores. Many other promising applications are under development. ■

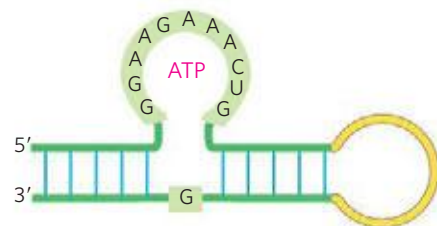


FIGURE 2 RNA aptamer that binds ATP. The shaded nucleotides are those required for the binding activity.

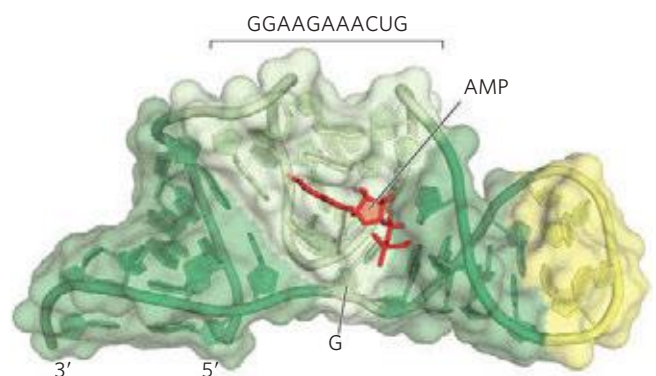


FIGURE 3 (Derived from PDB ID 1RAW) RNA aptamer bound to AMP. The bases of the conserved nucleotides (forming the binding pocket) are white; the bound AMP is red.

BOX 26-4 An Expanding RNA Universe Filled with TUF RNAs

Current estimates for the number of genes in the human genome, and in many other genomes, are mentioned in multiple places throughout this text. The estimates presume that scientists know a gene when they see it, based on our current understanding of DNA, RNA, and proteins. Is the presumption correct?

As noted in Chapter 9, less than 2% of the human genome seems to encode proteins. Even when introns

are factored in, one might expect only a tiny fraction of the genome to be transcribed into RNA, mostly mRNA to encode those proteins. The remainder of the genome has sometimes been referred to as junk DNA. The “junk” moniker simply reflects our ignorance, which is slowly giving way to the realization that most of the genome is fully functional.

In an effort to better map the boundaries of the human transcriptome, researchers have invented new tools to determine with higher accuracy which genomic sequences are transcribed into RNA. The answers are surprising. Much more of our genome is transcribed into RNA than anyone supposed. Much of this RNA seems not to encode proteins. Much of it lacks some of the structures (for example, the 3' poly(A) tail) that characterize mRNA. So what is this RNA doing?

Most of the methods for looking into this matter fall into two broad categories: cDNA cloning and microarrays. The creation of a cDNA library to study the genes transcribed in a particular eukaryotic genome is described in Chapter 9 (see Fig. 9-14). However, the classical methods for generating cDNA often lead to the cloning of only part of the sequence of a given transcript. Because reverse transcriptase may stall at regions of secondary structure in mRNA, or may simply dissociate, often 20% or less of the clones in a cDNA library are full-length DNAs. This makes it difficult to use the library to map transcription start sites (TSSs) and to study the part of a gene that encodes the amino-terminal sequence of a protein. One of the many approaches developed to overcome this problem is illustrated in Figure 1. Such refinements in technology have resulted in the creation of cDNA libraries in which more than 95% of the clones are full-length, providing an enriched source of information about cellular RNAs. However, cDNAs are generally created from RNA transcripts that have poly(A) tails. The use of microarrays, coupled to methods of cDNA preparation that do not rely on poly(A) tails (Fig. 2), has revealed that much of the RNA in eukaryotic cells lacks the common end structures.

A complete picture has not yet emerged, but some conclusions are already clear. If one excludes the repetitive sequences (transposons, for example) that can make up half of a mammalian genome, at least

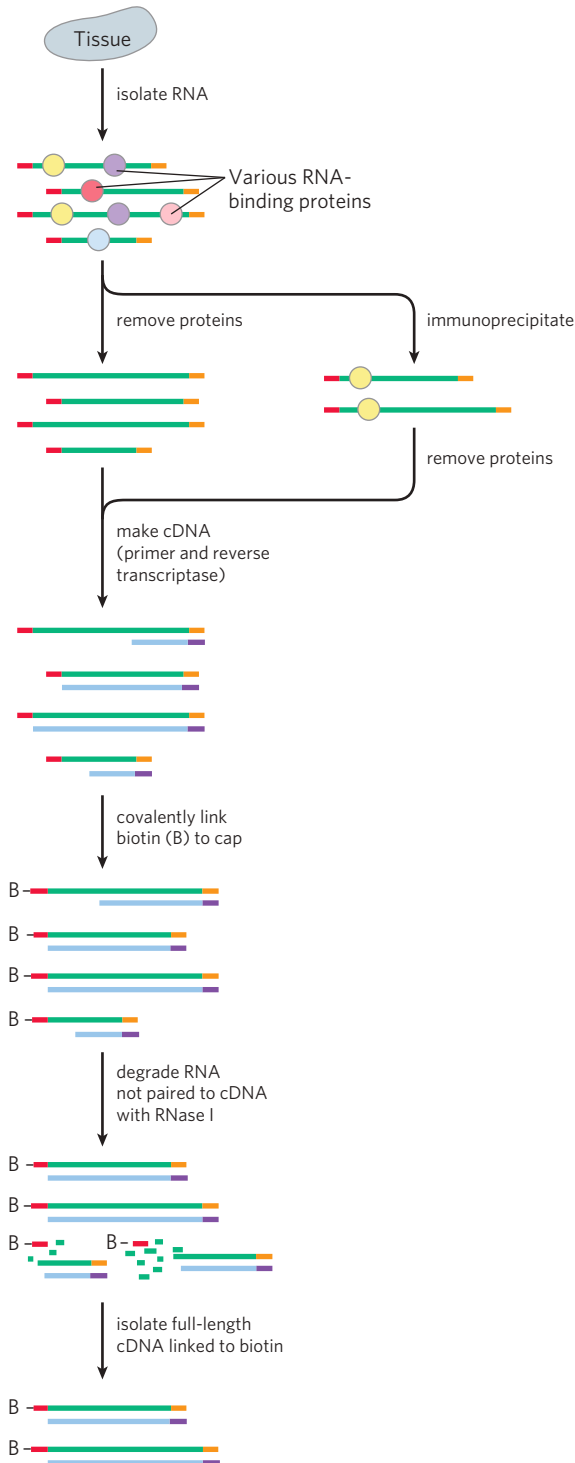


FIGURE 1 A strategy for cloning full-length cDNAs. A pool of mRNAs is isolated from a tissue sample. In some cases, mRNAs bound by a particular protein may be targeted by immunoprecipitating the protein and isolating the associated mRNAs. Biotin (B) is covalently attached to the 5' ends of the mRNAs, making use of unique features of the 5' cap. A poly(dT) primer (purple) is used to prime reverse transcription of the mRNAs. RNase I degrades RNA that is not part of a DNA-RNA hybrid and thus destroys the incomplete cDNA-RNA pairs. The full-length cDNA-RNA hybrids are collected with streptavidin beads (which bind biotin), converted to duplex DNA, and cloned.

40%—and perhaps the vast majority—of the remaining genomic DNA is transcribed into RNA. There seem to be more RNAs lacking poly(A) tails than RNAs with them. Much of this RNA is not transported to the cytoplasm but remains exclusively in the nucleus. Many segments of the genome are transcribed on both strands; one transcript is the complement of the other, a relationship referred to as antisense. Many of the antisense RNAs may be involved in regulation of the RNAs with which they pair. Many RNAs are produced in only one or a few tissues, and new transcripts are discovered every time a new tissue source is analyzed. Thus, the complete transcriptome has not yet been defined for any organism. Most important, many of the novel RNAs are transcribed from genomic segments, such as those illustrated in Figure 9–15, that share synteny in more than one organism. This evolutionary conservation strongly suggests that these RNAs have an important function.

Some of the novel RNAs are snoRNAs, snRNAs, or miRNAs, types of RNA recognized only in the past two decades. New TSS sequences are being discovered. New classes of RNA molecules are being defined. New patterns of alternative splicing are being elucidated. And all of these findings are challenging our definitions of a gene. In the mouse and human genomes, even the familiar protein-encoding mRNA transcripts may be much more numerous than initially

thought, and the number of known protein-encoding genes may soon increase. However, the function of most of the newly discovered transcripts is unknown, and they are simply called TUFs (transcripts of unknown function). This new RNA universe is a frontier that promises further insights into the workings of eukaryotic cells and perhaps a new glimpse of our origins in the RNA world of the distant past.

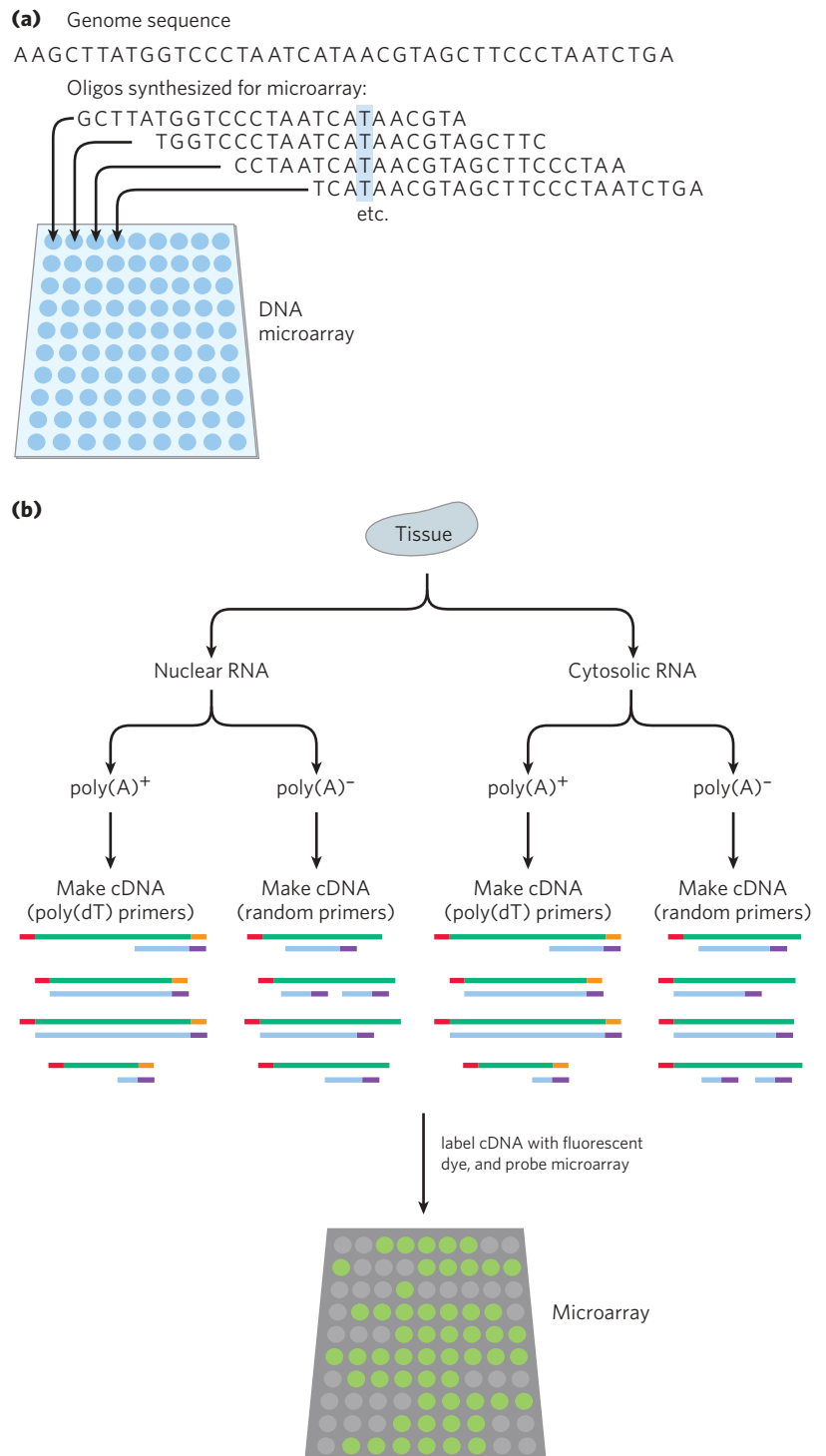


FIGURE 2 Defining the transcriptome with microarrays.

(a) Tiled microarrays are synthesized, representing the nonrepetitive parts of the genome. In a tiled array, the successive oligonucleotides in the individual spots overlap in sequence, so that each nucleotide (such as the T shown in blue) is represented multiple times.

(b) A tissue sample is fractionated to separate nuclear and cytoplasmic samples, and RNA is isolated from each. The RNA containing poly(A) tails is separated from RNA lacking the tails (by passing the RNA over a column with bound poly(dT)). RNA with a poly(A) tail is converted to cDNA, using methods described in Figure 1. RNA lacking a poly(A) tail is converted to cDNA by using primers with randomized sequences. The resulting DNA fragments do not correspond in length precisely to the RNAs from which they are derived, but the overall DNA pool includes most of the sequences present in the original RNAs. The cDNA samples are then labeled and used to probe the microarrays. The signals from the microarrays define the sequences that were transcribed into RNA.

genomes into double-stranded DNA as part of their life cycle. These enzymes transcribe the viral RNA into DNA, a process that can be used experimentally to form complementary DNA.

- ▶ Many eukaryotic transposons are related to retroviruses, and their mechanism of transposition includes an RNA intermediate.
- ▶ Telomerase, the enzyme that synthesizes the telomere ends of linear chromosomes, is a specialized reverse transcriptase that contains an internal RNA template.
- ▶ RNA-dependent RNA polymerases, such as the replicases of RNA bacteriophages, are template-specific for the viral RNA.
- ▶ The existence of catalytic RNAs and pathways for the interconversion of RNA and DNA has led to speculation that an important stage in evolution was the appearance of an RNA (or an equivalent polymer) that could catalyze its own replication. The biochemical potential of RNAs can be explored by SELEX, a method for rapidly selecting RNA sequences with particular binding or catalytic properties.

Key Terms

Terms in bold are defined in the glossary.

transcription 1057	poly(A) tail 1075
messenger RNA	polyadenylate
(mRNA) 1057	polymerase 1075
transfer RNA	small nucleolar RNA
(tRNA) 1057	(snoRNA) 1079
ribosomal RNA	snoRNP 1079
(rRNA) 1057	microRNA (miRNA) 1081
transcriptome 1057	internal guide
DNA-dependent RNA	sequence 1082
polymerase 1058	exosome 1084
template strand 1059	polynucleotide
nontemplate strand 1059	phosphorylase 1085
coding strand 1059	reverse
promoter 1060	transcriptase 1086
consensus sequence 1060	retrovirus 1087
footprinting 1062	complementary DNA
cAMP receptor protein	(cDNA) 1087
(CRP) 1063	homing 1089
repressor 1063	telomerase 1090
transcription factors 1066	T loop 1091
ribozymes 1069	RNA-dependent RNA
primary transcript 1069	polymerase (RNA
RNA splicing 1069	replicase) 1092
5' cap 1070	SELEX 1093
spliceosome 1072	aptamer 1095
small nuclear RNA	
(snRNA) 1073	

Further Reading

General

Cox, M.M., Doudna, J.A., & O'Donnell, M. (2012) *Molecular Biology: Principles and Practice*, W. H. Freeman and Company, New York.

Jacob, F. & Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356.

A classic article that introduced many important ideas.

DNA-Directed RNA Synthesis

Gruber, T.M. & Gross, C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57**, 441–466.

Herbert, K.M., Greenleaf, W.J., & Block, S.M. (2008) Single-molecule studies of RNA polymerase: motoring along. *Annu. Rev. Biochem.* **77**, 149–176.

Kuehner, J.N., Pearson, E.L., & Moore, C. (2011) Unravelling the means to an end: RNA polymerase II transcription termination. *Nat. Rev. Mol. Cell Biol.* **12**, 283–294.

Laine, J.P. & Egly, J.M. (2006) When transcription and repair meet: a complex system. *Trends Genet.* **22**, 430–436.

Mooney, R.A., Darst, S.A., & Landick, R. (2005) Sigma and RNA polymerase: an on-again, off-again relationship? *Mol. Cell* **20**, 335–345.

Nudler, E. (2009) RNA polymerase active center: the molecular engine of transcription. *Annu. Rev. Biochem.* **78**, 335–361.

Svejstrup, J.Q., Conaway, R.C., & Conaway, J.W. (2006) RNA polymerase II: a “Nobel” enzyme demystified. *Mol. Cell* **24**, 637–642.

Thomas, M.C. & Chiang, C.M. (2006) The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178.

Zhou, Q., Li, T., & Price, D.H. (2012) RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**, 119–143.

RNA Processing

Berezikov, E. (2011) Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* **12**, 846–860.

Buchan, J.R. & Parker, R. (2007) The two faces of miRNA. *Science* **318**, 1877–1878.

Butcher, S.E. & Brow, D.A. (2005) Towards understanding the catalytic core structure of the spliceosome. *Biochem. Soc. Trans.* **33**, 447–449.

Gogarten, J.P. & Hilario, E. (2006) Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol. Biol.* **6**, 94.

Hoskins, A.A., Gelles, J., & Moore, M.J. (2011) New insights into the spliceosome by single molecule fluorescence microscopy. *Curr. Opin. Chem. Biol.* **15**, 864–870.

Huang, Y.Q. & Steitz, J.A. (2005) SRprises along a messenger's journey. *Mol. Cell* **17**, 613–615.

Kaberdin, V.R. & Blasi, U. (2006) Translation initiation and the fate of bacterial mRNAs. *FEMS Microbiol. Rev.* **30**, 967–979.

Kalsotra, A. & Cooper, T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729.

Reiter, N.J., Chan, C.W., & Mondragon, A. (2011) Emerging structural themes in large RNA molecules. *Curr. Opin. Struct. Biol.* **21**, 319–326.

Rodriguez-Trelles, F., Tarrio, R., & Ayala, F.J. (2006) Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.* **40**, 47–76.

Schneider, D.A., Michel, A., Sikes, M.L., Vu, L., Dodd, J.A., Salgia, S., Osheim, Y.S., Beyer, A.L., & Nomura, M. (2007) Transcription elongation by RNA polymerase I is linked to efficient rRNA processing and ribosome assembly. *Mol. Cell* **26**, 217–229.

RNA-Directed RNA or DNA Synthesis

Blackburn, E.H., Greider, C.W., & Szostak, J.W. (2006) Telomeres and telomerase: the path from maize, *Tetrahymena* and yeast to human cancer and aging. *Nat. Med.* **12**, 1133–1138.

Boeke, J.D. & Devine, S.E. (1998) Yeast retrotransposons: finding a nice, quiet neighborhood. *Cell* **93**, 1087–1089.

Cordaux, R. & Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703.

Frankel, A.D. & Young, J.A.T. (1998) HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1–25.

Mason, M., Schuller, A., & Skordalakes, E. (2011) Telomerase structure function. *Curr Opin. Struct. Biol.* **21**, 92–100.

O'Sullivan, R.J. & Karlseder, J. (2010) Telomeres: protecting chromosomes against genome instability. *Nat. Rev. Mol. Cell Biol.* **11**, 171–181.

Temin, H.M. (1976) The DNA provirus hypothesis: the establishment and implications of RNA-directed DNA synthesis. *Science* **192**, 1075–1080.

Discussion of the original proposal for reverse transcription in retroviruses.

Ribozymes and Evolution

Carninci, P. (2007) Constructing the landscape of the mammalian genome. *J. Exp. Biol.* **210**, 1497–1506.

A good summary of the work showing that mammalian transcriptomes are much more extensive than previously thought.

Doudna, J.A. & Lorsch, J.R. (2005) Ribozyme catalysis: not different, just worse. *Nat. Struct. Mol. Biol.* **12**, 395–402.

Green, R. & Doudna, J.A. (2006) RNAs regulate biology. *ACS Chem. Biol.* **1**, 335–338.

Huttenhofer, A. & Vogel, J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res.* **34**, 635–646.

Joyce, G.F. (2002) The antiquity of RNA-based evolution. *Nature* **418**, 214–221.

Kazantsev, A.V. & Pace, N.R. (2006) Bacterial RNase P: a new view of an ancient enzyme. *Nat. Rev. Microbiol.* **4**, 729–740.

Lassila, J.K., Zalatan, J.G., & Herschlag, D. (2011) Biological phosphoryl-transfer reactions: understanding mechanism and catalysis. *Annu. Rev. Biochem.* **80**, 669–702.

Lincoln, T.A. & Joyce, G.F. (2009) Self-sustained replication of an RNA enzyme. *Science* **323**, 1229–1232.

Müller, U.F. (2006) Recreating an RNA world. *Cell. Mol. Life Sci.* **63**, 1278–1293.

Willingham, A.T. & Gingeras, T.R. (2006) TUF love for “junk” DNA. *Cell* **125**, 1215–1220.

Wilson, T.J. & Lilley, D.M. (2009) The evolution of ribozyme chemistry. *Science* **323**, 1436–1438.

Wochner, A., Attwater, J., Coulson, A., & Holliger, P. (2011) Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**, 209–212.

Yarus, M. (2002) Primordial genetics: phenotype of the ribocyte. *Annu. Rev. Genet.* **36**, 125–151.

Detailed speculations about what an RNA-based life form might have been like and a good summary of the research behind them.

Problems

1. RNA Polymerase (a) How long would it take for the *E. coli* RNA polymerase to synthesize the primary transcript for the *E. coli* genes encoding the enzymes for lactose metabolism (the 5,300 bp *lac* operon, considered in Chapter 28)? (b) How far along the DNA would the transcription “bubble” formed by RNA polymerase move in 10 seconds?

2. Error Correction by RNA Polymerases DNA polymerases are capable of editing and error correction, whereas the capacity for error correction in RNA polymerases seems to be quite limited. Given that a single base error in either replication or transcription can lead to an error in protein synthesis, suggest a possible biological explanation for this difference.

3. RNA Posttranscriptional Processing Predict the likely effects of a mutation in the sequence (5')AAUAAA in a eukaryotic mRNA transcript.

4. Coding versus Template Strands The RNA genome of phage Q β is the nontemplate strand, or coding strand, and when introduced into the cell, it functions as an mRNA. Suppose the RNA replicase of phage Q β synthesized primarily template-strand RNA and uniquely incorporated this, rather than nontemplate strands, into the viral particles. What would be the fate of the template strands when they entered a new cell? What enzyme would have to be included in the viral particles for successful invasion of a host cell?

5. Transcription The gene encoding the *E. coli* enzyme β -galactosidase begins with the sequence ATGACCATGAT-TACG. What is the sequence of the RNA transcript specified by this part of the gene?

6. The Chemistry of Nucleic Acid Biosynthesis Describe three properties common to the reactions catalyzed by DNA polymerase, RNA polymerase, reverse transcriptase, and RNA replicase. How is the enzyme polynucleotide phosphorylase similar to and different from these four enzymes?

7. RNA Splicing What is the minimum number of transesterification reactions needed to splice an intron from an mRNA transcript? Explain.

8. RNA Processing If the splicing of mRNA in a vertebrate cell is blocked, the rRNA modification reactions are also blocked. Suggest a reason for this.

9. RNA Genomes The RNA viruses have relatively small genomes. For example, the single-stranded RNAs of retroviruses have about 10,000 nucleotides and the Q β RNA is only 4,220 nucleotides long. Given the properties of reverse transcriptase and RNA replicase described in this chapter, can you suggest a reason for the small size of these viral genomes?

10. Screening RNAs by SELEX The practical limit for the number of different RNA sequences that can be screened in a SELEX experiment is 10^{15} . (a) Suppose you are working with oligonucleotides 32 nucleotides long. How many sequences exist in a randomized pool containing every sequence possible?

(b) What percentage of these can be screened in a SELEX experiment? (c) Suppose you wish to select an RNA molecule that catalyzes the hydrolysis of a particular ester. From what you know about catalysis, propose a SELEX strategy that might allow you to select the appropriate catalyst.

11. Slow Death The death cap mushroom, *Amanita phalloides*, contains several dangerous substances, including the lethal α -amanitin. This toxin blocks RNA elongation in consumers of the mushroom by binding to eukaryotic RNA polymerase II with very high affinity; it is deadly in concentrations as low as 10^{-8} M. The initial reaction to ingestion of the mushroom is gastrointestinal distress (caused by some of the other toxins). These symptoms disappear, but about 48 hours later, the mushroom-eater dies, usually from liver dysfunction. Speculate on why it takes this long for α -amanitin to kill.



12. Detection of Rifampicin-Resistant Strains of Tuberculosis Rifampicin is an important antibiotic used to treat tuberculosis and other mycobacterial diseases. Some strains of *Mycobacterium tuberculosis*, the causative agent of tuberculosis, are resistant to rifampicin. These strains become resistant through mutations that alter the *rpoB* gene, which encodes the β subunit of the RNA polymerase. Rifampicin cannot bind to the mutant RNA polymerase and so is unable to block the initiation of transcription. DNA sequences from a large number of rifampicin-resistant *M. tuberculosis* strains have been found to have mutations in a specific 69 bp region of *rpoB*. One well-characterized rifampicin-resistant strain has a single base pair alteration in *rpoB* that results in a His residue being replaced by an Asp residue in the β subunit.

(a) Based on your knowledge of protein chemistry, suggest a technique that would allow detection of the rifampicin-resistant strain containing this particular mutant protein.

(b) Based on your knowledge of nucleic acid chemistry, suggest a technique to identify the mutant form of *rpoB*.

Using the Web

13. The Ribonuclease Gene Human pancreatic ribonuclease has 128 amino acid residues.

(a) What is the minimum number of nucleotide pairs required to code for this protein?

(b) The mRNA expressed in human pancreatic cells was copied with reverse transcriptase to create a “library” of human DNA. The sequence of the mRNA coding for human pancreatic ribonuclease was determined by sequencing the complementary DNA (cDNA) from this library that included an open reading frame for the protein. Use the Entrez database system (www.ncbi.nlm.nih.gov/Entrez) to find the published sequence of this mRNA (search the CoreNucleotide

database for accession number D26129). What is the length of this mRNA?

(c) How can you account for the discrepancy between the size you calculated in (a) and the actual length of the mRNA?

Data Analysis Problem

14. A Case of RNA Editing The AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) receptor is an important component of the human nervous system. It is present in several forms, in different neurons, and some of this variety results from posttranscriptional modification. This problem explores research on the mechanism of this RNA editing.

An initial report by Sommer and coauthors (1991) looked at the sequence encoding a key Arg residue in the AMPA receptor. The sequence of the cDNA (see Fig. 9–14) for the AMPA receptor showed a CGG (Arg; see Fig. 27–7) codon for this amino acid. Surprisingly, the genomic DNA showed a CAG (Gln) codon at this position.

(a) Explain how this result is consistent with posttranscriptional modification of the AMPA receptor mRNA.

Rueter and colleagues (1995) explored this mechanism in detail. They first developed an assay to differentiate between edited and unedited transcripts, based on the Sanger method of DNA sequencing (see Fig. 8–33). They modified the technique to determine whether the base in question was an A (as in CAG) or not. They designed two DNA primers based on the genomic DNA sequence of this region of the AMPA gene. These primers, and the genomic DNA sequence of the nontemplate strand for the relevant region of the AMPA receptor gene, are shown at the bottom of the page; the A residue that is edited is in red.

To detect whether this A was present or had been edited to another base, Rueter and coworkers used the following procedure:

1. Prepared cDNA complementary to the mRNA, using primer 1, reverse transcriptase, dATP, dGTP, dCTP, and dTTP.
2. Removed the mRNA.
3. Annealed 32 P-labeled primer 2 to the cDNA and reacted this with DNA polymerase, dGTP, dCTP, dTTP, and ddATP (dideoxy ATP; see Fig. 8–33).
4. Denatured the resulting duplexes and separated them with polyacrylamide gel electrophoresis (see Fig. 3–18).
5. Detected the 32 P-labeled DNA species with autoradiography.

They found that edited mRNA produced a 22 nucleotide [32 P] DNA, whereas unedited mRNA produced a 19 nucleotide [32 P] DNA.

(b) Using the sequences below, explain how the edited and unedited mRNAs resulted in these different products.

(5') ...GTCTCTGGTTTTTCCTTGGGTGCCTTTATGCA GCAAGGATGCGATATTTTCGCCAAG...
 Primer 1: CGTTCCTACGCTATAAAGCGGTT C (5')
 Primer 2: (5') CCTTGGGTGCCTTTA

Using the same procedure, to measure the fraction of transcripts edited under different conditions, the researchers found that extracts of cultured epithelial cells (a common cell line called HeLa) could edit the mRNA at a high level. To determine the nature of the editing machinery, they pretreated an active HeLa cell extract as described in the table and measured its ability to edit AMPA mRNA. Proteinase K degrades only proteins; micrococcal nuclease, only DNA.

Sample	Pretreatment	% mRNA edited
1	None	18
2	Proteinase K	5
3	Heat to 65 °C	3
4	Heat to 85 °C	3
5	Micrococcal nuclease	17

(c) Use these data to argue that the editing machinery consists of protein. What is a key weakness in this argument?

To determine the exact nature of the edited base, Rueter and colleagues used the following procedure:

1. Produced mRNA, using [α - 32 P]ATP in the reaction mixture.
2. Edited the labeled mRNA by incubating with HeLa extract.
3. Hydrolyzed the edited mRNA to single nucleotide monophosphates with nuclease P1.
4. Separated the nucleotide monophosphates with thin-layer chromatography (TLC; see Fig. 10–25b).
5. Identified the resulting 32 P-labeled nucleotide monophosphates with autoradiography.

In unedited mRNA, they found only [32 P]AMP; in edited mRNA, they found mostly [32 P]AMP with some [32 P]IMP (inosine monophosphate; see Fig. 22–36).

(d) Why was it necessary to use [α - 32 P]ATP rather than [β - 32 P]ATP or [γ - 32 P]ATP in this experiment?

(e) Why was it necessary to use [α - 32 P]ATP rather than [α - 32 P]GTP, [α - 32 P]CTP, or [α - 32 P]UTP?

(f) How does the result exclude the possibility that the entire A nucleotide (sugar, base, and phosphate) was removed and replaced by an I nucleotide during the editing process?

The researchers next edited mRNA that was labeled with [2,8- 3 H]ATP and repeated the above procedure. The only 3 H-labeled mononucleotides produced were AMP and IMP.

(g) How does this result exclude removal of the A base (leaving the sugar-phosphate backbone intact) followed by replacement with an I base as a mechanism of editing? What, then, is the most likely mechanism of editing in this case?

(h) How does changing an A to an I residue in the mRNA explain the Gln to Arg change in protein sequence in the two forms of AMPA receptor protein? (Hint: See Fig. 27–8.)

References

Rueter, S.M., Burns, C.M., Coode, S.A., Mookherjee, P., & Emesont, R.B. (1995) Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. *Science* **267**, 1491–1494.

Sommer, B., Köhler, M., Sprengel, R., & Seeburg, P.H. (1991) RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11–19.

this page left intentionally blank

Protein Metabolism

27.1 The Genetic Code 1103

27.2 Protein Synthesis 1113

27.3 Protein Targeting and Degradation 1139

Proteins are the end products of most information pathways. A typical cell requires thousands of different proteins at any given moment. These must be synthesized in response to the cell's current needs, transported (targeted) to their appropriate cellular locations, and degraded when no longer needed. Many of the fundamental components and mechanisms utilized by the protein biosynthetic machinery are remarkably well conserved in all life-forms from bacteria to higher eukaryotes, indicating that they were present in the last universal common ancestor (LUCA) of all extant organisms.

An understanding of protein synthesis, the most complex biosynthetic process, has been one of the greatest challenges in biochemistry. Eukaryotic protein synthesis involves more than 70 different ribosomal proteins; 20 or more enzymes to activate the amino acid precursors; a dozen or more auxiliary enzymes and other protein factors for the initiation, elongation, and termination of polypeptides; perhaps 100 additional enzymes for the final processing of different proteins; and 40 or more kinds of transfer and ribosomal RNAs. Overall, almost 300 different macromolecules cooperate to synthesize polypeptides. Many of these macromolecules are organized into the complex three-dimensional structure of the ribosome.

To appreciate the central importance of protein synthesis, consider the cellular resources devoted to this process. Protein synthesis can account for up to 90% of the chemical energy used by a cell for all biosynthetic reactions. Every bacterial, archaeal, and eukaryotic cell contains from several to thousands of copies of many different proteins and RNAs. The 15,000 ribosomes, 100,000 molecules of protein synthesis-related protein factors and enzymes, and 200,000 tRNA molecules in a typical bacterial cell can account for more than 35% of the cell's dry weight.

Despite the great complexity of protein synthesis, proteins are made at exceedingly high rates. A polypeptide of 100 residues is synthesized in an *Escherichia coli* cell (at 37 °C) in about 5 seconds. Synthesis of the thousands of different proteins in a cell is tightly regulated, so that just enough copies are made to match the current metabolic circumstances. To maintain the appropriate mix and concentration of proteins, the targeting and degradative processes must keep pace with synthesis. Research is gradually uncovering the finely coordinated cellular choreography that guides each protein to its proper cellular location and selectively degrades it when it is no longer required.

The study of protein synthesis offers another important reward: a look at a world of RNA catalysts that may have existed before the dawn of life “as we know it.” Elucidation of the three-dimensional structures of ribosomes, beginning in the year 2000, has given us an increasingly detailed look at the mechanics of protein synthesis. It has also confirmed a hypothesis first put forward by Harry Noller two decades earlier: proteins are synthesized by a gigantic RNA enzyme!



Harry Noller

27.1 The Genetic Code

Three major advances set the stage for our present knowledge of protein biosynthesis. First, in the early 1950s, Paul Zamecnik and his colleagues designed a set of experiments to investigate where in the cell proteins are synthesized. They injected radioactive amino acids into rats and, at different time intervals after the injection, removed the liver, homogenized it, fractionated the homogenate by centrifugation, and examined the subcellular fractions for the presence of radioactive protein. When hours or days were allowed to elapse after injection of the labeled amino acids, *all* the subcellular



Paul Zamecnik, 1912-2009

fractions contained labeled proteins. However, when only minutes had elapsed, labeled protein appeared only in a fraction containing small ribonucleoprotein particles. These particles, visible in animal tissues by electron microscopy, were therefore identified as the site of protein synthesis from amino acids and later were named ribosomes (Fig. 27-1).

The second key advance was made by Mahlon Hoagland and Zamecnik when they found that amino acids were “activated” when incubated with ATP and the cytosolic fraction of liver cells. The amino acids became attached to a heat-stable soluble RNA of the type that had been discovered and characterized by Robert Holley and later called transfer RNA (tRNA), to form **aminoacyl-tRNAs**. The enzymes that catalyze this process are the **aminoacyl-tRNA synthetases**.

The third advance resulted from Francis Crick’s reasoning on how the genetic information encoded in the 4-letter language of nucleic acids could be translated into the 20-letter language of proteins. A small nucleic acid (perhaps RNA) could serve the role of an adaptor, one part of the adaptor molecule binding a specific amino acid and another part recognizing the nucleotide sequence encoding that amino acid in an mRNA (Fig. 27-2). This idea was soon verified. The tRNA adaptor “translates” the nucleotide sequence of an mRNA into the amino acid sequence of a polypeptide. The overall process of mRNA-guided protein synthesis is often referred to simply as **translation**.

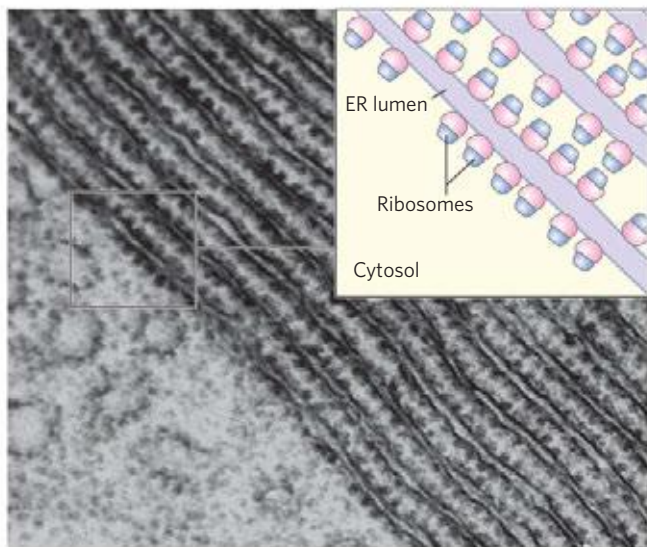


FIGURE 27-1 Ribosomes and endoplasmic reticulum. Electron micrograph and schematic drawing of a portion of a pancreatic cell, showing ribosomes attached to the outer (cytosolic) face of the endoplasmic reticulum (ER). The ribosomes are the numerous small dots bordering the parallel layers of membranes.

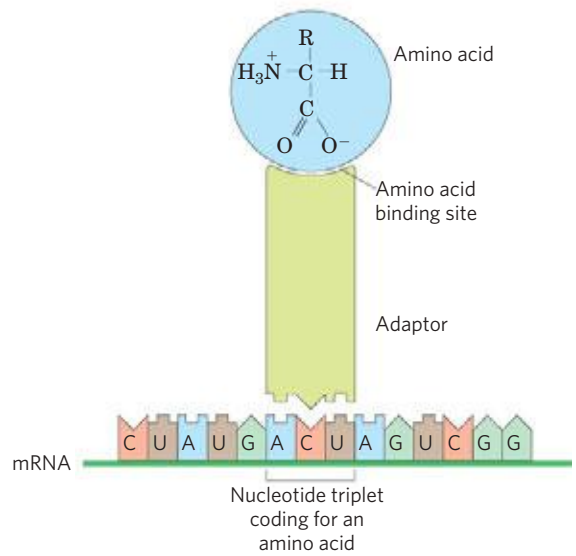


FIGURE 27-2 Crick’s adaptor hypothesis. Today we know that the amino acid is covalently bound at the 3’ end of a tRNA molecule and that a specific nucleotide triplet elsewhere in the tRNA interacts with a particular triplet codon in mRNA through hydrogen bonding of complementary bases.

These three developments soon led to recognition of the major stages of protein synthesis and ultimately to the elucidation of the genetic code that specifies each amino acid.

The Genetic Code Was Cracked Using Artificial mRNA Templates

By the 1960s it was apparent that at least three nucleotide residues of DNA are necessary to encode each amino acid. The four code letters of DNA (A, T, G, and C) in groups of two can yield only $4^2 = 16$ different combinations, insufficient to encode 20 amino acids. Groups of three, however, yield $4^3 = 64$ different combinations.

Several key properties of the genetic code were established in early genetic studies (Figs 27-3, 27-4).

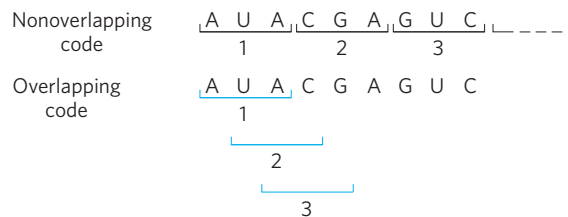


FIGURE 27-3 Overlapping versus nonoverlapping genetic codes. In a nonoverlapping code, codons (numbered consecutively) do not share nucleotides. In an overlapping code, some nucleotides in the mRNA are shared by different codons. In a triplet code with maximum overlap, many nucleotides, such as the third nucleotide from the left (A), are shared by three codons. Note that in an overlapping code, the triplet sequence of the first codon limits the possible sequences for the second codon. A nonoverlapping code provides much more flexibility in the triplet sequence of neighboring codons and therefore in the possible amino acid sequences designated by the code. The genetic code used in all living systems is now known to be nonoverlapping.

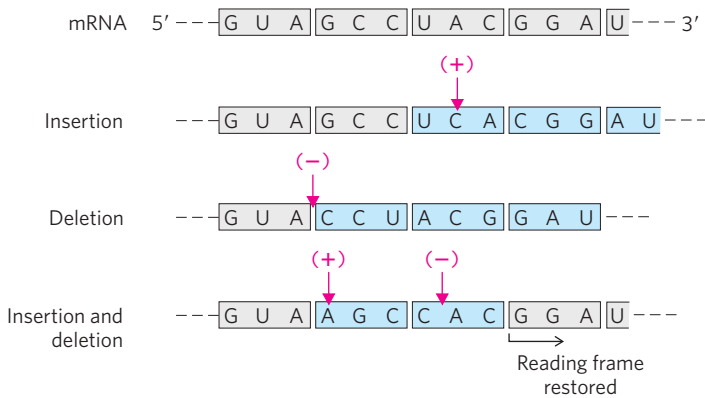


FIGURE 27-4 The triplet, nonoverlapping code. Evidence for the general nature of the genetic code came from many types of experiments, including genetic experiments on the effects of deletion and insertion mutations. Inserting or deleting one base pair (shown here in the mRNA transcript) alters the sequence of triplets in a nonoverlapping code; all amino acids coded by the mRNA following the change are affected. Combining insertion and deletion mutations affects some amino acids but can eventually restore the correct amino acid sequence. Adding or subtracting three nucleotides (not shown) leaves the remaining triplets intact, providing evidence that a codon has three, rather than four or five, nucleotides. The triplet codons shaded in gray are those transcribed from the original gene; codons shaded in blue are new codons resulting from the insertion or deletion mutations.

A **codon** is a triplet of nucleotides that codes for a specific amino acid. Translation occurs in such a way that these nucleotide triplets are read in a successive, nonoverlapping fashion. A specific first codon in the sequence establishes the **reading frame**, in which a new codon begins every three nucleotide residues. There is no punctuation between codons for successive amino acid residues. The amino acid sequence of a protein is defined by a linear sequence of contiguous triplets. In principle, any given single-stranded DNA or mRNA sequence has three possible reading frames. Each reading frame gives a different sequence of codons (Fig. 27-5), but only one is likely to encode a given protein. A key question remained: what were the three-letter code words for each amino acid?



Marshall Nirenberg

In 1961 Marshall Nirenberg and Heinrich Matthaei reported the first breakthrough. They incubated synthetic polyuridylylate, poly(U), with an *E. coli* extract, GTP, ATP, and a mixture of the 20 amino acids in 20 different tubes, each tube containing a different radioactively labeled amino acid. Because poly(U) mRNA is made up of many successive UUU triplets, it should promote the synthesis of a polypeptide containing only the amino acid encoded by the triplet UUU. A radioactive polypeptide was indeed formed in only one of the 20 tubes, the one containing radioactive phenylalanine. Nirenberg and Matthaei therefore concluded that the triplet codon UUU encodes phenylalanine. The same approach soon revealed that polycytidylylate, poly(C),

encodes a polypeptide containing only proline (polyproline), and polyadenylate, poly(A), encodes polylysine. Polyguanylylate did not generate any polypeptide in this experiment because it spontaneously forms tetraplexes (see Fig. 8-20d) that cannot be bound by ribosomes.

The synthetic polynucleotides used in such experiments were prepared with polynucleotide phosphorylase (p. 1085), which catalyzes the formation of RNA polymers starting from ADP, UDP, CDP, and GDP. This enzyme, discovered by Severo Ochoa, requires no template and makes polymers with a base composition that directly reflects the relative concentrations of the nucleoside 5'-diphosphate precursors in the medium. If polynucleotide phosphorylase is presented with UDP only, it makes only poly(U). If it is presented with a mixture of five parts ADP and one part CDP, it makes a polymer in which about five-sixths of the residues are adenylate and one-sixth are cytidylate. This random polymer is likely to have many triplets of the sequence AAA, smaller numbers of AAC, ACA, and CAA triplets, relatively few ACC, CCA, and CAC triplets, and very few CCC triplets (Table 27-1). Using a variety of artificial mRNAs made by polynucleotide phosphorylase from different starting mixtures of ADP, GDP, UDP, and CDP, the Nirenberg and Ochoa groups soon identified the base compositions of the triplets coding for almost all the amino acids. Although these experiments revealed the base composition of the coding triplets, they usually could not reveal the sequence of the bases.

KEY CONVENTION: Much of the following discussion deals with tRNAs. The amino acid specified by a tRNA is indicated by a superscript, such as tRNA^{Ala}, and the aminoacylated tRNA by a hyphenated name: alanyl-tRNA^{Ala} or Ala-tRNA^{Ala}. ■

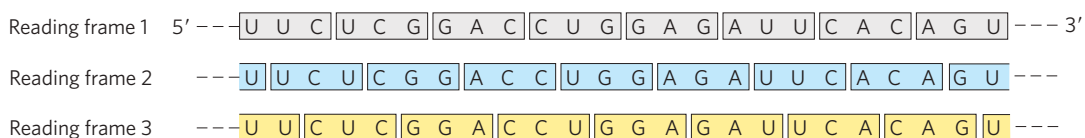


FIGURE 27-5 Reading frames in the genetic code. In a triplet, nonoverlapping code, all mRNAs have three potential reading frames, shaded here in different colors. The triplets, and hence the amino acids specified, are different in each reading frame.

TABLE 27-1 Incorporation of Amino Acids into Polypeptides in Response to Random Polymers of RNA

Amino acid	Observed frequency of incorporation (Lys = 100)	Tentative assignment for nucleotide composition of corresponding codon*	Expected frequency of incorporation based on assignment (Lys = 100)
Asparagine	24	A ₂ C	20
Glutamine	24	A ₂ C	20
Histidine	6	AC ₂	4
Lysine	100	AAA	100
Proline	7	AC ₂ , CCC	4.8
Threonine	26	A ₂ C, AC ₂	24

Note: Presented here is a summary of data from one of the early experiments designed to elucidate the genetic code. A synthetic RNA containing only A and C residues in a 5:1 ratio directed polypeptide synthesis, and both the identity and the quantity of incorporated amino acids were determined. Based on the relative abundance of A and C residues in the synthetic RNA, and assigning the codon AAA (the most likely codon) a frequency of 100, there should be three different codons of composition A₂C, each at a relative frequency of 20; three of composition AC₂, each at a relative frequency of 4.0; and CCC at a relative frequency of 0.8. The CCC assignment was based on information derived from prior studies with poly(C). Where two tentative codon assignments are made, both are proposed to code for the same amino acid.

*These designations of nucleotide composition contain no information on nucleotide sequence (except, of course, AAA and CCC).

In 1964 Nirenberg and Philip Leder achieved another experimental breakthrough. Isolated *E. coli* ribosomes would bind a specific aminoacyl-tRNA in the presence of the corresponding synthetic polynucleotide messenger. For example, ribosomes incubated with poly(U) and phenylalanyl-tRNA^{Phe} (Phe-tRNA^{Phe}) bind both RNAs, but if the ribosomes are incubated with poly(U) and some other aminoacyl-tRNA, the aminoacyl-tRNA is not bound, because it does not recognize the UUU triplets in poly(U) (Table 27-2). Even trinucleotides could promote specific binding of appropriate tRNAs, so these experiments could be carried out with chemically synthesized small oligonucleotides. With this technique researchers determined which aminoacyl-tRNA bound to 54 of the 64 possible triplet codons. For some codons, either no aminoacyl-tRNA or more than one would bind. Another method was needed to complete and confirm the entire genetic code.

TABLE 27-2 Trinucleotides That Induce Specific Binding of Aminoacyl-tRNAs to Ribosomes

Trinucleotide	Relative increase in ¹⁴ C-labeled aminoacyl-tRNA bound to ribosome*		
	Phe-tRNA ^{Phe}	Lys-tRNA ^{Lys}	Pro-tRNA ^{Pro}
UUU	4.6	0	0
AAA	0	7.7	0
CCC	0	0	3.1

Source: Modified from Nirenberg, M. & Leder, P. (1964) RNA code words and protein synthesis. *Science* 145, 1399.

*Each number represents the factor by which the amount of bound ¹⁴C increased when the indicated trinucleotide was present, relative to a control with no trinucleotide.



H. Gobind Khorana, 1922-2011

At about this time, a complementary approach was provided by H. Gobind Khorana, who developed chemical methods to synthesize polyribonucleotides with defined, repeating sequences of two to four bases. The polypeptides produced by these mRNAs had one or a few amino acids in repeating patterns. These patterns, when combined with information from the random polymers used by Nirenberg and colleagues, permitted unambiguous codon assignments. The copolymer (AC)_n, for example, has alternating ACA and CAC codons: ACACACACACACACA. The polypeptide synthesized on this messenger contained equal amounts of threonine and histidine. Given that a histidine codon has one A and two Cs (Table 27-1), CAC must code for histidine and ACA for threonine.

Consolidation of the results from many experiments permitted the assignment of 61 of the 64 possible codons. The other three were identified as termination codons, in part because they disrupted amino acid coding patterns when they occurred in a synthetic RNA polymer (Fig. 27-6). Meanings for all the triplet codons (tabulated in Fig. 27-7) were established by 1966 and have been verified in many different ways.

The cracking of the genetic code is regarded as one of the most important scientific discoveries of the twentieth century.

Codons are the key to the translation of genetic information, directing the synthesis of specific proteins. The reading frame is set when translation of an mRNA

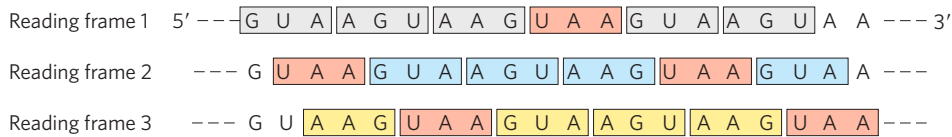


FIGURE 27-6 Effect of a termination codon in a repeating tetranucleotide. Termination codons (light red) are encountered every fourth codon in three different reading frames (shown in different colors). Dipeptides or tripeptides are synthesized, depending on where the ribosome initially binds.

		First letter of codon (5' end)			
		U	C	A	G
U	Second letter of codon	UUU Phe	UCU Ser	UAU Tyr	UGU Cys
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	
C	CUU Leu	CCU Pro	CAU His	CGU Arg	
	CUC Leu	CCC Pro	CAC His	CGC Arg	
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	
G	GUA Val	GCA Ala	GAA Glu	GGA Gly	
	GUG Val	GCG Ala	GAG Glu	GGG Gly	

FIGURE 27-7 “Dictionary” of amino acid code words in mRNAs. The codons are written in the 5'→3' direction. The third base of each codon (in bold type) plays a lesser role in specifying an amino acid than the first two. The three termination codons are shaded in light red, the initiation codon AUG in green. All the amino acids except methionine and tryptophan have more than one codon. In most cases, codons that specify the same amino acid differ only at the third base.

molecule begins and it is maintained as the synthetic machinery reads sequentially from one triplet to the next. If the initial reading frame is off by one or two bases, or if translation somehow skips a nucleotide in the mRNA, all the subsequent codons will be out of register; the result is usually a “missense” protein with a garbled amino acid sequence.

Several codons serve special functions (Fig. 27-7). The **initiation codon** AUG is the most common signal for the beginning of a polypeptide in all cells, in addition to coding for Met residues in internal positions of polypeptides. The **termination codons** (UAA, UAG, and UGA), also called stop codons or nonsense codons, normally signal the end of polypeptide synthesis and do not code for any known amino acids. Some deviations from these rules are discussed in Box 27-1.

As described in Section 27.2, initiation of protein synthesis in the cell is an elaborate process that relies on

initiation codons and other signals in the mRNA. In retrospect, the experiments of Nirenberg, Khorana, and others to identify codon function should not have worked in the absence of initiation codons. Serendipitously, experimental conditions caused the normal initiation requirements for protein synthesis to be relaxed. Diligence combined with chance to produce a breakthrough—a common occurrence in the history of biochemistry.

In a random sequence of nucleotides, 1 in every 20 codons in each reading frame is, on average, a termination codon. In general, a reading frame without a termination codon among 50 or more codons is referred to as an **open reading frame (ORF)**. Long open reading frames usually correspond to genes that encode proteins. In the analysis of sequence databases, sophisticated programs are used to search for open reading frames in order to find genes among the often huge background of nongenic DNA. An uninterrupted gene coding for a typical protein with a molecular weight of 60,000 would require an open reading frame with 500 or more codons.

A striking feature of the genetic code is that an amino acid may be specified by more than one codon, so the code is described as **degenerate**. This does *not* suggest that the code is flawed: although an amino acid may have two or more codons, each codon specifies only one amino acid. The degeneracy of the code is not uniform. Whereas methionine and tryptophan have single codons, for example, three amino acids (Arg, Leu, Ser) have six codons, five amino acids have four, isoleucine has three, and nine amino acids have two (Table 27-3).

TABLE 27-3 Degeneracy of the Genetic Code

Amino acid	Number of codons	Amino acid	Number of codons
Met	1	Tyr	2
Trp	1	Ile	3
Asn	2	Ala	4
Asp	2	Gly	4
Cys	2	Pro	4
Gln	2	Thr	4
Glu	2	Val	4
His	2	Arg	6
Lys	2	Leu	6
Phe	2	Ser	6

BOX 27-1 Exceptions That Prove the Rule: Natural Variations in the Genetic Code

In biochemistry, as in other disciplines, exceptions to general rules can be problematic for instructors and frustrating for students. At the same time, though, they teach us that life is complex and inspire us to search for more surprises. Understanding the exceptions can even reinforce the original rule in surprising ways.

One would expect little room for variation in the genetic code. Even a single amino acid substitution can have profoundly deleterious effects on the structure of a protein. Nevertheless, variations in the code do occur in some organisms, and they are both interesting and instructive. The types of variation and their rarity provide powerful evidence for a common evolutionary origin of all living things.

To alter the code, changes must occur in the gene(s) encoding one or more tRNAs, with the obvious target for alteration being the anticodon. Such a change would lead to the systematic insertion of an amino acid at a codon that, according to the standard code (see Fig. 27-7), does not specify that amino acid. The genetic code, in effect, is defined by two elements: (1) the anticodons on tRNAs (which determine where an amino acid is placed in a growing polypeptide) and (2) the specificity of the enzymes—the aminoacyl-tRNA synthetases—that charge the tRNAs, which determines the identity of the amino acid attached to a given tRNA.

Most sudden changes in the code would have catastrophic effects on cellular proteins, so code alterations are more likely to persist where relatively few proteins would be affected—such as in small genomes encoding only a few proteins. The biological consequences of a code change could also be limited by restricting changes to the three termination codons, which do not generally occur *within* genes (see Box 27-4 for exceptions to *this* rule). This pattern is in fact observed.

Of the very few variations in the genetic code that we know of, most occur in mitochondrial DNA (mtDNA), which encodes only 10 to 20 proteins. Mitochondria have their own tRNAs, so their code variations do not affect the much larger cellular genome. The most common changes in mitochondria involve termination codons. These changes affect termination in the products of only a subset of genes, and sometimes the effects are minor because the genes have multiple (redundant) termination codons.

Vertebrate mtDNAs have genes that encode 13 proteins, 2 rRNAs, and 22 tRNAs (see Fig. 19-40a). The

small number of codon reassignments, along with an unusual set of wobble rules (p. 1110), makes the 22 tRNAs sufficient to decode the protein genes, as opposed to the 32 tRNAs required for the standard code. In mitochondria, these changes can be viewed as a kind of genomic streamlining, as a smaller genome confers a replication advantage on the organelle. Four codon families (in which the amino acid is determined entirely by the first two nucleotides) are decoded by a single tRNA with a U residue in the first (or wobble) position in the anticodon. Either the U pairs somehow with any of the four possible bases in the third position of the codon or a “two out of three” mechanism is used—that is, no base pairing is needed at the third position. Other tRNAs recognize codons with either A or G in the third position, and yet others recognize U or C, so that virtually all the tRNAs recognize either two or four codons.

In the standard code, only two amino acids are specified by single codons: methionine and tryptophan (see Table 27-3). If all mitochondrial tRNAs recognize two codons, we would expect additional Met and Trp codons in mitochondria. And we find that the single most common code variation is the normal termination codon UGA specifying tryptophan. The tRNA^{Trp} recognizes and inserts a Trp residue at either UGA or the normal Trp codon, UGG. The second most common variation is conversion of AUA from an Ile codon to a Met codon; the normal Met codon is AUG, and a single tRNA recognizes both codons. The known coding variations in mitochondria are summarized in Table 1.

Turning to the much rarer changes in the codes for cellular (as distinct from mitochondrial) genomes, we find that the only known variation in a bacterium is again the use of UGA to encode Trp residues, occurring in the simplest free-living cell, *Mycoplasma capricolum*. Among eukaryotes, rare extramitochondrial coding changes occur in a few species of ciliated protists, in which both termination codons UAA and UAG can specify glutamine. There are also rare but interesting cases where stop codons have been adapted to encode amino acids that are not among the standard 20, as detailed in Box 27-3.

Changes in the code need not be absolute; a codon might not always encode the same amino acid. For example, in many bacteria—including *E. coli*—GUG (Val) is sometimes used as an initiation codon that specifies Met. This occurs only for those genes in

The genetic code is nearly universal. With the intriguing exception of a few minor variations in mitochondria, some bacteria, and some single-celled eukaryotes (Box 27-1), amino acid codons are identical in all species examined so far. Human beings, *E. coli*, tobacco plants, amphibians, and viruses share the same genetic code. Thus it would appear that all life-forms have a common evolutionary ancestor, whose genetic code has

been preserved throughout biological evolution. Even the variations reinforce this theme.

Wobble Allows Some tRNAs to Recognize More than One Codon

When several different codons specify one amino acid, the difference between them usually lies at the third

TABLE 1 Known Variant Codon Assignments in Mitochondria

	Codons*				
	UGA	AUA	AGA AGG	CUN	CGG
Normal code assignment	Stop	Ile	Arg	Leu	Arg
Animals					
Vertebrates	Trp	Met	Stop	+	+
<i>Drosophila</i>	Trp	Met	Ser	+	+
Yeasts					
<i>Saccharomyces cerevisiae</i>	Trp	Met	+	Thr	+
<i>Torulopsis glabrata</i>	Trp	Met	+	Thr	?
<i>Schizosaccharomyces pombe</i>	Trp	+	+	+	+
Filamentous fungi	Trp	+	+	+	+
Trypanosomes	Trp	+	+	+	+
Higher plants	+	+	+	+	Trp
<i>Chlamydomonas reinhardtii</i>	?	+	+	+	?

*N indicates any nucleotide; +, codon has the same meaning as in the normal code; ?, codon not observed in this mitochondrial genome.

which the GUG is properly located relative to particular mRNA sequences that affect the initiation of translation (as discussed in Section 27.2).

The most surprising alteration in the genetic code occurs in some fungal species of the genus *Candida*, as originally discovered for *Candida albicans*. *C. albicans* is an organism of high genomic complexity, yet its genetic code has undergone a dramatic change: the CUG codon, which normally encodes Leu, encodes Ser instead. The natural selection pressure for this change is completely unknown. Furthermore, the chemical nature of Ser and Leu are quite different. However, even this change can be understood based on the properties of a universal code. When several codons encode the same amino acid and utilize multiple tRNAs, not all of the codons are used with equal frequency. In a phenomenon called **codon bias**, some codons for a particular amino acid are used more frequently (sometimes much more frequently) than others. The tRNAs for the frequently used codons are often present at much higher concentrations than the tRNAs required for the rarely used codons. Code degeneracy leads to the presence of six codons for Leu. In bacteria, CUG is used often to encode Leu. However, in fungi in the genera that are very closely related to *Candida* but do not

have the coding change, the CUG codon is used quite rarely to encode Leu and is often entirely absent in highly expressed proteins. A change in the coding sense of CUG would thus have a much smaller effect on fungal cell metabolism than might be anticipated if all codons were used equally. The coding change may have occurred by a gradual loss of CUG codons in genes and the tRNA that recognizes CUG as a Leu codon, followed by a capture event—a mutation in the anticodon of a tRNA^{Ser} that allowed it to recognize CUG. Alternatively, there may have been an intermediate stage in which CUG was recognized as encoding both Leu and Ser, perhaps with contextual signals in the mRNAs that helped one tRNA or another to recognize specific CUG codons (see Box 27–3). Phylogenetic analysis indicates that the reassignment of CUG as a Ser codon occurred in *Candida* ancestors about 150 to 170 million years ago.

These variations tell us that the code is not quite as universal as once believed, but that its flexibility is severely constrained. The variations are obviously derivatives of the normal code, and no example of a completely different code has been found. The limited scope of code variants strengthens the principle that all life on this planet evolved on the basis of a single (slightly flexible) genetic code.

base position (at the 3' end). For example, alanine is coded by the triplets GCU, GCC, GCA, and GCG. The codons for most amino acids can be symbolized by XY_G^A or XY_C^U. The first two letters of each codon are the primary determinants of specificity, a feature that has some interesting consequences.

Transfer RNAs base-pair with mRNA codons at a three-base sequence on the tRNA called the **antico-**

don. The first base of the codon in mRNA (read in the 5'→3' direction) pairs with the third base of the anticodon (**Fig. 27–8a**). If the anticodon triplet of a tRNA recognized only one codon triplet through Watson-Crick base pairing at all three positions, cells would have a different tRNA for each amino acid codon. This is not the case, however, because the anticodons in some tRNAs include the nucleotide inosinate (designated I),

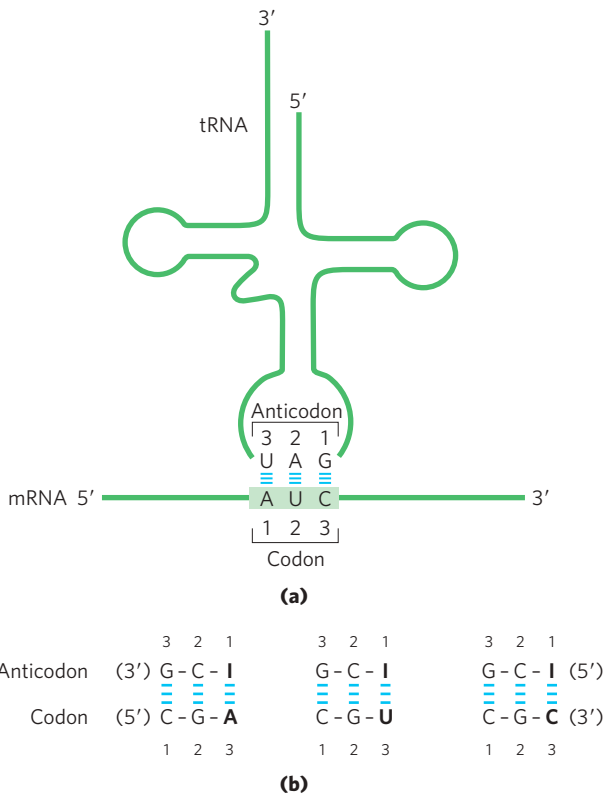


FIGURE 27-8 Pairing relationship of codon and anticodon. (a) Alignment of the two RNAs is antiparallel. The tRNA is shown in the traditional cloverleaf configuration. (b) Three different codon pairing relationships are possible when the tRNA anticodon contains inosinate.

which contains the uncommon base hypoxanthine (see Fig. 8-5b). Inosinate can form hydrogen bonds with three different nucleotides (U, C, and A; Fig. 27-8b), although these pairings are much weaker than the hydrogen bonds of Watson-Crick base pairs G≡C and A=U. In yeast, one tRNA^{Arg} has the anticodon (5')ICG, which recognizes three arginine codons: (5')CGA, (5')CGU, and (5')CGC. The first two bases are identical (CG) and form strong Watson-Crick base pairs with the corresponding bases of the anticodon, but the third base (A, U, or C) forms rather weak hydrogen bonds with the I residue at the first position of the anticodon.

Examination of these and other codon-anticodon pairings led Crick to conclude that the third base of most codons pairs rather loosely with the corresponding base of its anticodon; to use his picturesque word, the third base of such codons (and the first base of their corresponding anticodons) “wobbles.” Crick proposed a set of four relationships called the **wobble hypothesis**:

1. The first two bases of an mRNA codon always form strong Watson-Crick base pairs with the corresponding bases of the tRNA anticodon and confer most of the coding specificity.
2. The first base of the anticodon (reading in the 5'→3' direction; this pairs with the third base of

the codon) determines the number of codons recognized by the tRNA. When the first base of the anticodon is C or A, base pairing is specific and only one codon is recognized by that tRNA. When the first base is U or G, binding is less specific and two different codons may be read. When inosine (I) is the first (wobble) nucleotide of an anticodon, three different codons can be recognized—the maximum number for any tRNA. These relationships are summarized in Table 27-4.

3. When an amino acid is specified by several different codons, the codons that differ in either of the first two bases require different tRNAs.
4. A minimum of 32 tRNAs are required to translate all 61 codons (31 to encode the amino acids and 1 for initiation).

The wobble (or third) base of the codon contributes to specificity, but, because it pairs only loosely with its corresponding base in the anticodon, it permits rapid dissociation of the tRNA from its codon during protein synthesis. If all three bases of a codon engaged in strong Watson-Crick pairing with the three bases of the anticodon, tRNAs would dissociate too slowly and this would limit the rate of protein synthesis. Codon-anticodon interactions balance the requirements for accuracy and speed.

The Genetic Code Is Mutation-Resistant

The genetic code plays an interesting role in safeguarding the genomic integrity of every living organism. Evolution did not produce a code in which codon assignments

TABLE 27-4 How the Wobble Base of the Anticodon Determines the Number of Codons a tRNA Can Recognize

1. One codon recognized:	Anticodon (3') X-Y-C (5')	(3') X-Y-A (5')
	Codon (5') X'-Y'-G (3')	(5') X'-Y'-U (3')
2. Two codons recognized:	Anticodon (3') X-Y-U (5')	(3') X-Y-G (5')
	Codon (5') X'-Y'-A (3')	(5') X'-Y'-C (3')
3. Three codons recognized:	Anticodon (3') X-Y-I (5')	
	Codon (5') X'-Y'-A (3')	

Note: X and Y denote bases complementary to and capable of strong Watson-Crick base pairing with X' and Y', respectively. Wobble bases—in the 3' position of codons and 5' position of anticodons—are shaded in white.

appeared at random. Instead, the code is strikingly resistant to the deleterious effects of the most common kinds of mutations—**missense mutations**, in which a single new base pair replaces another. In the third, or wobble, position of the codon, single base substitutions produce a change in the encoded amino acid only about 25% of the time. Most such changes are thus **silent mutations**, where the nucleotide is different but the encoded amino acid remains the same.

Due to the types of spontaneous DNA damage that affect genomes (see Chapter 8), the most frequent missense mutation is a **transition mutation**, where a purine is replaced by a purine or a pyrimidine by a pyrimidine (for example, G \rightleftharpoons C changed to A \rightleftharpoons T). All three codon positions have evolved so that there is some resistance to transition mutations. A mutation in the first position of the codon will usually result in an amino acid coding change, but the change often involves an amino acid with similar chemical properties. This is especially true for the hydrophobic amino acids that dominate the first column of the code shown in Figure 27–7. Consider the Val codon GUU. A change to AUU would substitute Ile for Val. A change to CUU would replace Val with Leu. The resulting changes in the structure and/or function of the protein encoded by that gene would often (but not always) be small.

Computational studies have shown that genetic codes delineated at random are almost always less resistant to mutation than the existing code. The results indicate that the code underwent considerable streamlining prior to the appearance of LUCA, the ancestral cell.

The genetic code tells us how protein sequence information is stored in nucleic acids and provides some clues about how that information is translated into protein.

Translational Frameshifting and RNA Editing Affect How the Code Is Read

Once the reading frame has been set during protein synthesis, codons are translated without overlap or punctuation until the ribosomal complex encounters a termination codon. The other two possible reading frames usually contain no useful genetic information, but a few genes are structured so that ribosomes “hiccup” at a certain point in the translation of their mRNAs, changing the reading frame from that point on. This appears to be a mechanism either to allow two or more related but

distinct proteins to be produced from a single transcript or to regulate the synthesis of a protein.

One of the best-documented examples of **translational frameshifting** occurs during translation of the mRNA for the overlapping *gag* and *pol* genes of the Rous sarcoma virus (see Fig. 26–34). The reading frame for *pol* is offset to the left by one base pair (–1 reading frame) relative to the reading frame for *gag* (Fig. 27–9).

The product of the *pol* gene (reverse transcriptase) is translated as a larger polyprotein, on the same mRNA that is used for the *gag* protein alone (see Fig. 26–33). The polyprotein, or *gag-pol* protein, is then trimmed to the mature reverse transcriptase by proteolytic digestion. Production of the polyprotein requires a translational frameshift in the overlap region to allow the ribosome to bypass the UAG termination codon at the end of the *gag* gene (shaded light red in Fig. 27–9).

Frameshifts occur during about 5% of translations of this mRNA, and the *gag-pol* polyprotein (and ultimately reverse transcriptase) is synthesized at about one-twentieth the frequency of the *gag* protein, a level that suffices for efficient reproduction of the virus. In some retroviruses, another translational frameshift allows translation of an even larger polyprotein that includes the product of the *env* gene fused to the *gag* and *pol* gene products (see Fig. 26–33). A similar mechanism produces both the τ and γ subunits of *E. coli* DNA polymerase III from a single *dnaX* gene transcript (see Table 25–2).

Some mRNAs are edited before translation. **RNA editing** can involve the addition, deletion, or alteration of nucleotides in the RNA in a manner that affects the meaning of the transcript when it is translated. Addition or deletion of nucleotides has been most commonly observed in RNAs originating from the mitochondrial and chloroplast genomes of eukaryotes. The reactions require a special class of RNA molecules encoded by these same organelles, with sequences complementary to the edited mRNAs. These guide RNAs (gRNAs; Fig. 27–10) act as templates for the editing process.

The initial transcripts of the genes that encode cytochrome oxidase subunit II in some protist mitochondria provide an example of editing by insertion. These transcripts do not correspond precisely to the sequence needed at the carboxyl terminus of the protein product. A posttranscriptional editing process inserts four U residues that shift the translational reading

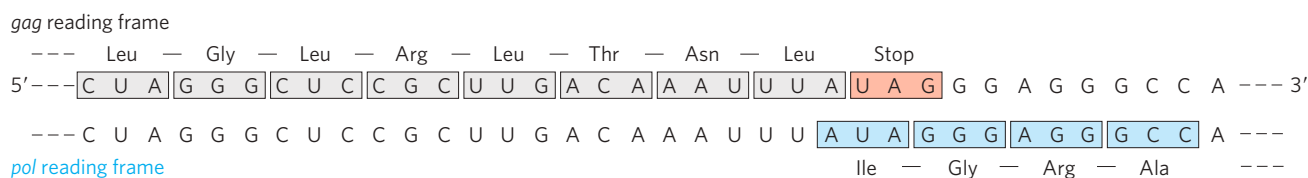


FIGURE 27–9 Translational frameshifting in a retroviral transcript. The *gag-pol* overlap region in Rous sarcoma virus RNA is shown.

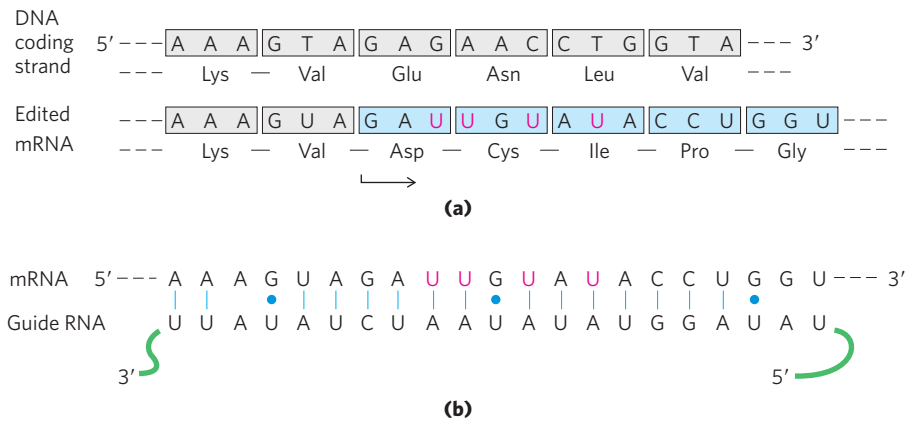


FIGURE 27-10 RNA editing of the transcript of the cytochrome oxidase subunit II gene from *Trypanosoma brucei* mitochondria. (a) Insertion of four U residues (red) produces a revised reading frame. (b) A special class of guide RNAs, complementary to the edited product, act as templates for the editing process. Note the presence of two G=U base pairs, signified by a blue dot to indicate non-Watson-Crick pairing.

frame of the transcript. Figure 27-10 shows the added U residues in the small part of the transcript that is affected by editing. Note that the base pairing between the initial transcript and the guide RNA involves a number of G=U base pairs (blue dots), which are common in RNA molecules.

RNA editing by alteration of nucleotides most commonly involves the enzymatic deamination of adenosine or cytidine residues, forming inosine or uridine, respectively (Fig. 27-11), although other base changes have been described. Inosine is interpreted as a G residue during translation. The adenosine deamination reactions are carried out by *adenosine deaminases* that act on RNA (**ADARs**). The cytidine deaminations are carried out by the *apoB* mRNA editing catalytic peptide

(**APOBEC**) family of enzymes, which includes the related *activation-induced deaminase (AID)* enzymes. Both groups of deaminase enzymes have a homologous zinc-coordinating catalytic domain.

A well-studied example of RNA editing by deamination occurs in the gene for the apolipoprotein B component of low-density lipoprotein in vertebrates. One form of apolipoprotein B, apoB-100 (M_r 513,000), is synthesized in the liver; a second form, apoB-48 (M_r 250,000), is synthesized in the intestine. Both are encoded by an mRNA produced from the gene for apoB-100. An APOBEC cytidine deaminase found only in the intestine binds to the mRNA at the codon for amino acid residue 2,153 (CAA = Gln) and converts the C to a U to create the termination codon UAA. The apoB-48 produced in

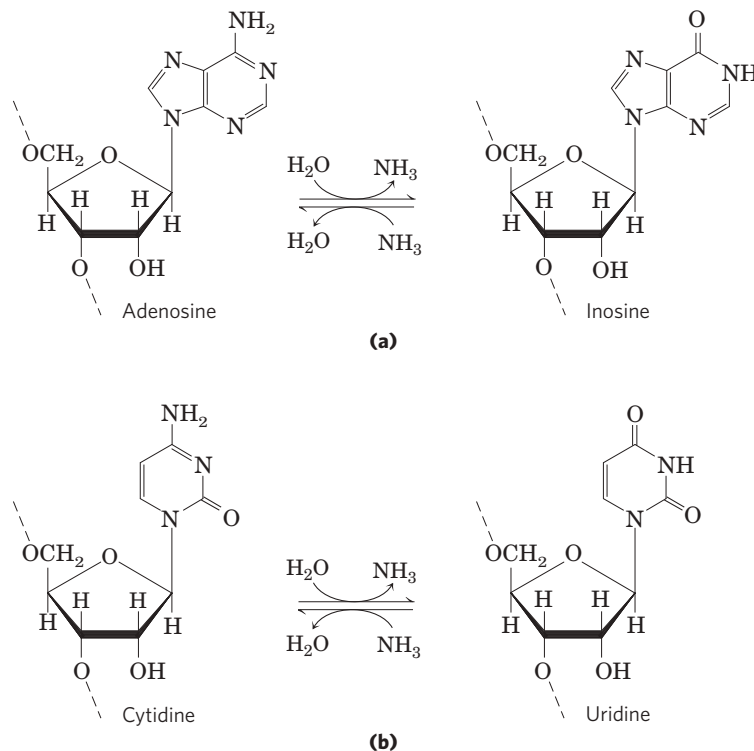


FIGURE 27-11 Deamination reactions that result in RNA editing. (a) The conversion of adenosine nucleotides to inosine nucleotides is catalyzed

by ADAR enzymes. (b) Cytidine-to-uridine conversions are catalyzed by the APOBEC family of enzymes.

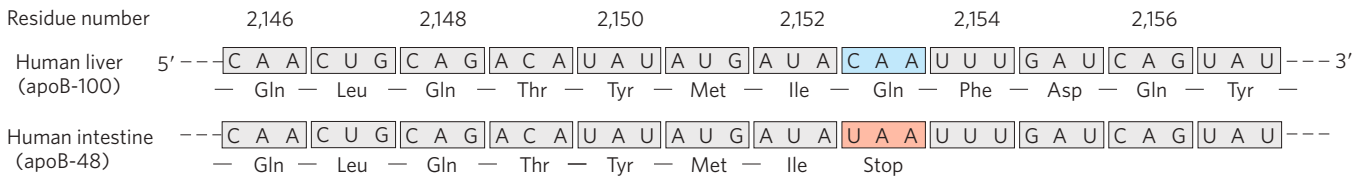


FIGURE 27-12 RNA editing of the transcript of the gene for the apoB-100 component of LDL. Deamination, which occurs only in the intestine,

converts a specific cytidine to uridine, changing a Gln codon to a stop codon and producing a truncated protein.

the intestine from this modified mRNA is simply an abbreviated form (corresponding to the amino-terminal half) of apoB-100 (Fig. 27-12). This reaction permits tissue-specific synthesis of two different proteins from one gene.

The ADAR-promoted A-to-I editing is particularly common in transcripts derived from the genes of primates. Perhaps 90% or more of the editing occurs in Alu elements, a subset of the eukaryotic transposons called short interspersed elements (SINEs), that are particularly common in mammalian genomes. There are over a million of the 300 bp Alu elements in human DNA, making up about 10% of the genome. These are concentrated near protein-encoding genes, often appearing in introns and untranslated regions at the 3' and 5' ends of transcripts. When it is first synthesized (prior to processing), the *average* human mRNA includes 10 to 20 Alu elements. The ADAR enzymes bind to and promote A-to-I editing only in duplex regions of RNA. The abundant Alu elements offer many opportunities for intramolecular base pairing within the transcripts, providing the duplex targets required by the ADARs. Some of the editing affects the coding sequences of genes. Defects in ADAR function have been associated with a variety of human neurological conditions, including amyotrophic lateral sclerosis (ALS), epilepsy, and major depression.

The genomes of all vertebrates are replete with SINEs, but many different types of SINEs are present in most of these organisms. The Alu elements predominate only in the primates. Careful screening of genes and transcripts indicates that A-to-I editing is 30 to 40 times more prevalent in humans than in mice, largely due to the presence of many Alu elements. Large-scale A-to-I editing and an increased level of alternative splicing (see Fig. 26-21) are two features that set primate genomes apart from those of other mammals. It is not yet clear whether these reactions are incidental or whether they played key roles in the evolution of primates and, ultimately, humans.

SUMMARY 27.1 The Genetic Code

- ▶ The particular amino acid sequence of a protein is constructed through the translation of information encoded in mRNA. This process is carried out by ribosomes.
- ▶ Amino acids are specified by mRNA codons consisting of nucleotide triplets. Translation

requires adaptor molecules, the tRNAs, that recognize codons and insert amino acids into their appropriate sequential positions in the polypeptide.

- ▶ The base sequences of the codons were deduced from experiments using synthetic mRNAs of known composition and sequence.
- ▶ The codon AUG signals initiation of translation. The triplets UAA, UAG, and UGA are signals for termination.
- ▶ The genetic code is degenerate: it has multiple codons for almost every amino acid.
- ▶ The standard genetic code is universal in all species, with some minor deviations in mitochondria and a few single-celled organisms. The deviations occur in a pattern that reinforces the concept of a universal code.
- ▶ The third position in each codon is much less specific than the first and second and is said to wobble.
- ▶ The genetic code is resistant to the effects of missense mutations.
- ▶ Translational frameshifting and RNA editing affect how the genetic code is read during translation.

27.2 Protein Synthesis

As we have seen for DNA and RNA (Chapters 25 and 26), the synthesis of polymeric biomolecules can be considered in terms of initiation, elongation, and termination stages. These fundamental processes are typically bracketed by two additional stages: activation of precursors before synthesis and postsynthetic processing of the completed polymer. Protein synthesis follows the same pattern. The activation of amino acids before their incorporation into polypeptides and the posttranslational processing of the completed polypeptide play particularly important roles in ensuring both the fidelity of synthesis and the proper function of the protein product. The process is outlined in Figure 27-13. The cellular components involved in the five stages of protein synthesis in *E. coli* and other bacteria are listed in Table 27-5; the requirements in eukaryotic cells are quite similar, although the components are in some cases more numerous. An initial overview of the stages of protein synthesis provides a useful outline for the discussion that follows.

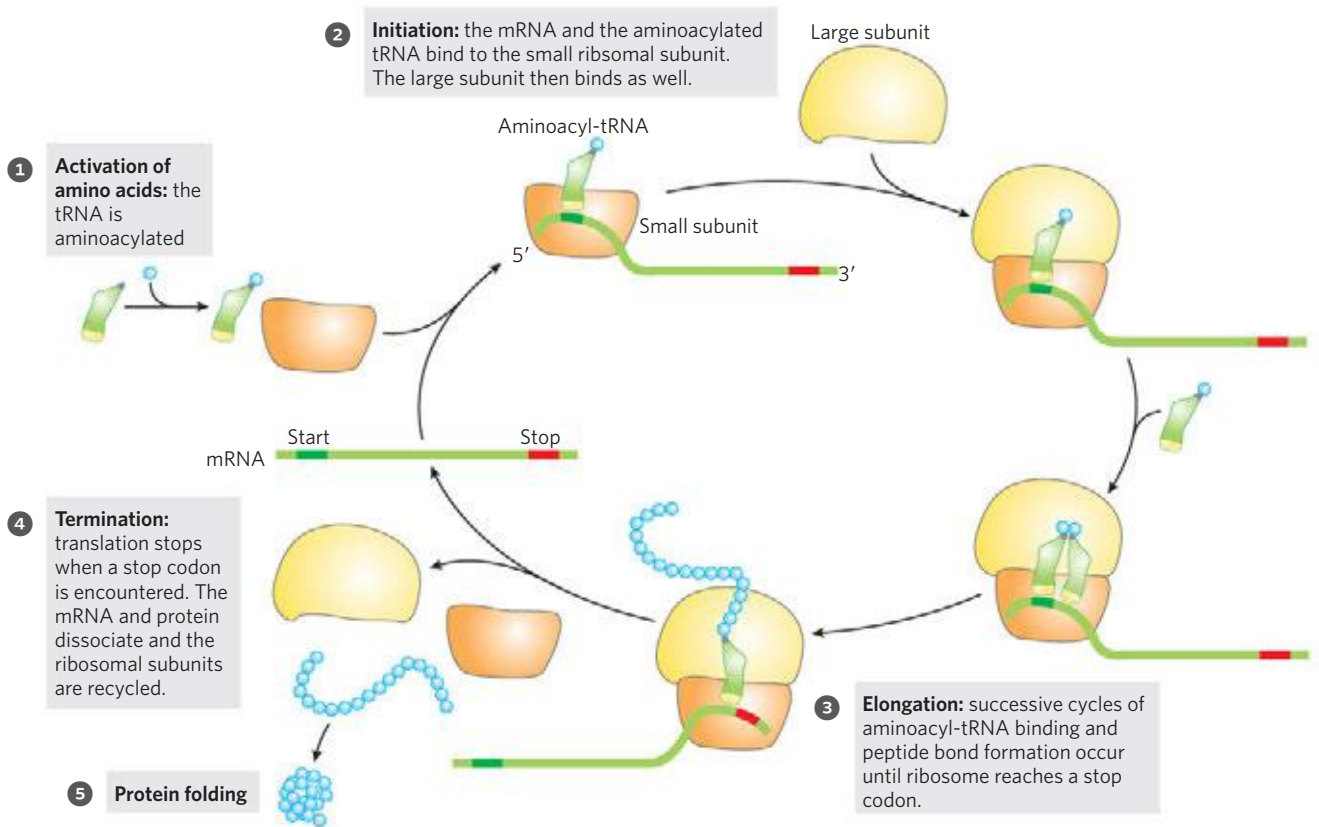


FIGURE 27-13 An overview of the five stages of protein synthesis. ① The tRNAs are aminoacylated. ② Translation initiation occurs when an mRNA and an aminoacylated tRNA are bound to the ribosome. ③ In elongation, the ribosome moves along the mRNA, matching tRNAs to each codon and catalyzing peptide bond formation. ④ Translation is

terminated at a stop codon, and the ribosomal subunits are released and recycled for another round of protein synthesis. ⑤ Following synthesis, the protein must fold into its active conformation and ribosome components are recycled.

Protein Biosynthesis Takes Place in Five Stages

Stage 1: Activation of Amino Acids For the synthesis of a polypeptide with a defined sequence, two fundamental chemical requirements must be met: (1) the carboxyl group of each amino acid must be activated to facilitate formation of a peptide bond, and (2) a link must be established between each new amino acid and the information in the mRNA that encodes it. Both these requirements are met by attaching the amino acid to a tRNA in the first stage of protein synthesis. Attaching the right amino acid to the right tRNA is critical. This reaction takes place in the cytosol, not on the ribosome. Each of the 20 amino acids is covalently attached to a specific tRNA at the expense of ATP energy, using Mg^{2+} -dependent activating enzymes known as aminoacyl-tRNA synthetases. When attached to their amino acid (aminoacylated) the tRNAs are said to be “charged.”

Stage 2: Initiation The mRNA bearing the code for the polypeptide to be synthesized binds to the smaller of two ribosomal subunits and to the initiating aminoacyl-tRNA. The large ribosomal subunit then binds to form an initiation complex. The initiating aminoacyl-tRNA

base-pairs with the mRNA codon AUG that signals the beginning of the polypeptide. This process, which requires GTP, is promoted by cytosolic proteins called initiation factors.

Stage 3: Elongation The nascent polypeptide is lengthened by covalent attachment of successive amino acid units, each carried to the ribosome and correctly positioned by its tRNA, which base-pairs to its corresponding codon in the mRNA. Elongation requires cytosolic proteins known as elongation factors. The binding of each incoming aminoacyl-tRNA and the movement of the ribosome along the mRNA are facilitated by the hydrolysis of GTP as each residue is added to the growing polypeptide.

Stage 4: Termination and Ribosome Recycling Completion of the polypeptide chain is signaled by a termination codon in the mRNA. The new polypeptide is released from the ribosome, aided by proteins called release factors, and the ribosome is recycled for another round of synthesis.

Stage 5: Folding and Posttranslational Processing In order to achieve its biologically active form, the new polypeptide must fold into its proper three-dimensional conformation.

TABLE 27-5 Components Required for the Five Major Stages of Protein Synthesis in *E. coli*

Stage	Essential components
1. Activation of amino acids	20 amino acids 20 aminoacyl-tRNA synthetases 32 or more tRNAs ATP Mg ²⁺
2. Initiation	mRNA <i>N</i> -Formylmethionyl-tRNA ^{fMet} Initiation codon in mRNA (AUG) 30S ribosomal subunit 50S ribosomal subunit Initiation factors (IF-1, IF-2, IF-3) GTP Mg ²⁺
3. Elongation	Functional 70S ribosome (initiation complex) Aminoacyl-tRNAs specified by codons Elongation factors (EF-Tu, EF-Ts, EF-G) GTP Mg ²⁺
4. Termination and ribosome recycling	Termination codon in mRNA Release factors (RF-1, RF-2, RF-3, RRF) EF-G IF-3
5. Folding and posttranslational processing	Chaperones and folding enzymes (PPI, PDI); specific enzymes, cofactors, and other components for removal of initiating residues and signal sequences, additional proteolytic processing, modification of terminal residues, and attachment of acetyl, phosphoryl, methyl, carboxyl, carbohydrate, or prosthetic groups

Before or after folding, the new polypeptide may undergo enzymatic processing, including removal of one or more amino acids (usually from the amino terminus); addition of acetyl, phosphoryl, methyl, carboxyl, or other groups to certain amino acid residues; proteolytic cleavage; and/or attachment of oligosaccharides or prosthetic groups.

Before looking at these five stages in detail, we must examine two key components in protein biosynthesis: the ribosome and tRNAs.

The Ribosome Is a Complex Supramolecular Machine

Each *E. coli* cell contains 15,000 or more ribosomes, which comprise nearly a quarter of the dry weight of the cell. Bacterial ribosomes contain about 65% rRNA and 35% protein; they have a diameter of about 18 nm and are composed of two unequal subunits with sedimentation coefficients of 30S and 50S and a combined sedimentation coefficient of 70S. Both subunits contain dozens of ribosomal proteins and at least one large rRNA (Table 27-6).

Following Zamecnik's discovery that ribosomes are the complexes responsible for protein synthesis, and following elucidation of the genetic code, the study of ribosomes accelerated. In the late 1960s Masayasu Nomura and colleagues demonstrated that both ribosomal subunits can be broken down into their RNA and protein components, then reconstituted *in vitro*. Under appropriate experimental conditions, the RNA and protein spontaneously reassemble to form 30S or 50S subunits nearly identical in structure and activity to native subunits. This breakthrough fueled decades of research into the function and structure of ribosomal RNAs and proteins. At the same time, increasingly sophisticated structural methods revealed more and more details about ribosome structure.



Masayasu Nomura,
1927–2011

TABLE 27-6 RNA and Protein Components of the *E. coli* Ribosome

Subunit	Number of different proteins	Total number of proteins	Protein designations	Number and type of rRNAs
30S	21	21	S1–S21	1 (16S rRNA)
50S	33	36	L1–L36*	2 (5S and 23S rRNAs)

*The L1 to L36 protein designations do not correspond to 36 different proteins. The protein originally designated L7 is in fact a modified form of L12, and L8 is a complex of three other proteins. Also, L26 proved to be the same protein as S20 (and not part of the 50S subunit). This gives 33 different proteins in the large subunit. There are four copies of the L7/L12 protein, with the three extra copies bringing the total protein count to 36.

The dawn of a new millennium brought with it the elucidation of the first high-resolution structures of bacterial ribosomal subunits by Thomas Steitz, Ada Yonath, Venki Ramakrishnan, Harry Noller, and others. This work yielded a wealth of surprises (Fig. 27-14a). First, a traditional focus on the protein components of ribosomes was shifted. The ribosomal subunits are huge RNA molecules. In the 50S subunit, the 5S and 23S rRNAs form the structural core. The proteins are secondary

elements in the complex, decorating the surface. Second and most important, there is no protein within 18 Å of the active site for peptide bond formation. The high-resolution structure thus confirms what Harry Noller had predicted much earlier: the ribosome is a ribozyme. In addition to the insight that the detailed structures of the ribosome and its subunits provide into the mechanism of protein synthesis (as elaborated below), they have stimulated a new look at the evolution of life (Box 27-2). The ribosomes of eukaryotic cells have also yielded to structural analysis (Fig. 27-14b).



Venkatraman Ramakrishnan

Thomas A. Steitz

Ada E. Yonath

The bacterial ribosome is complex, with a combined molecular weight of ~2.7 million. The two irregularly shaped ribosomal subunits fit together to form a cleft through which the mRNA passes as the ribosome moves along it during translation (Fig. 27-14a). The 57 proteins in bacterial ribosomes vary enormously in size and structure. Molecular weights range from

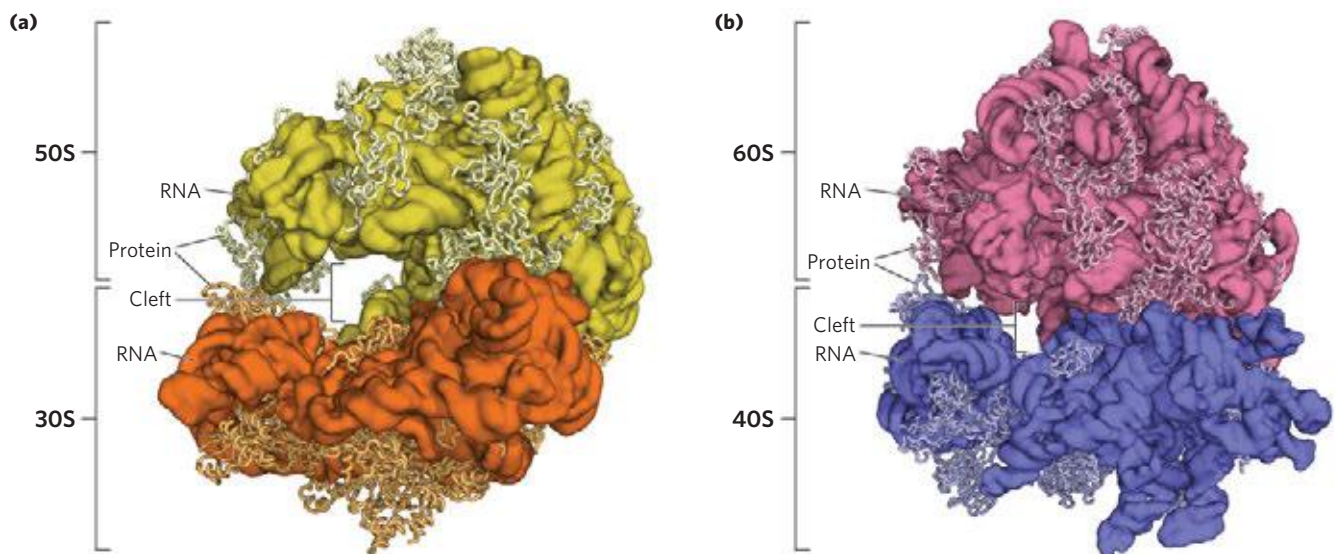


FIGURE 27-14 The structure of ribosomes. Our understanding of ribosome structure has been greatly enhanced by multiple high-resolution images of the ribosomes from bacteria and yeast. **(a)** The bacterial ribosome (derived from PDB ID 2OW8 and PDB ID 1VSA). The 50S and 30S

subunits come together to form the 70S ribosome. A cleft between them is where protein synthesis occurs. **(b)** The yeast ribosome has a similar structure with somewhat increased complexity (derived from PDB ID 3O58 and PDB ID 3O2Z).

BOX 27-2 From an RNA World to a Protein World

Extant ribozymes generally promote one of two types of reactions: hydrolytic cleavage of phosphodiester bonds or phosphoryl transfers (Chapter 26). In both cases, the substrates of the reactions are also RNA molecules. The ribosomal RNAs provide an important expansion of the catalytic range of known ribozymes. Coupled to the laboratory exploration of potential RNA catalytic function (see Box 26-3), the idea of an RNA world as a precursor to current life-forms becomes increasingly attractive.

A viable RNA world would require an RNA capable of self-replication, a primitive metabolism to generate the needed ribonucleotide precursors, and a cell boundary to aid in concentrating the precursors and sequestering them from the environment. The requirements for catalysis of reactions involving a growing range of metabolites and macromolecules could have led to larger and more complex RNA catalysts. The many negatively charged phosphoryl groups in the RNA backbone limit the stability of very large RNA molecules. In an RNA world, divalent cations or other positively charged groups could be incorporated into the structures to augment stability.

Certain peptides could stabilize large RNA molecules. For example, many ribosomal proteins in modern eukaryotic cells have long extensions, lacking a regular secondary structure, that snake into the rRNAs and help stabilize them (Fig. 1). Ribozyme-catalyzed synthesis of peptides could thus initially have evolved as part of a general solution to the structural maintenance of large RNA molecules. The synthesis of peptides may have helped stabilize large ribozymes, but this advance also marked the beginning of the end for the RNA world. Once peptide synthesis was possible, the greater catalytic potential of proteins would have set in motion an irreversible transition to a protein-dominated metabolic system.

Most enzymatic processes, then, were eventually surrendered to the proteins—but not all. In every organism, the critical task of synthesizing the proteins

remains, even now, a ribozyme-catalyzed process. There appears to be only one good arrangement (or just a very few) of nucleotide residues in a ribozyme active site that can catalyze peptide synthesis. The rRNA residues that seem to be involved in the peptidyl transferase activity of ribosomes are highly conserved in the large-subunit rRNAs of all species. Using in vitro evolution (SELEX; see Box 26-3), investigators have isolated artificial ribozymes that promote peptide synthesis. Intriguingly, most of them include the ribonucleotide octet (5')AUAACAGG(3'), a highly conserved sequence found at the peptidyl transferase active site in the ribosomes of all cells. There may be just one optimal solution to the overall chemical problem of ribozyme-catalyzed synthesis of proteins of defined sequence. Evolution found this solution once, and no life-form has notably improved on it.

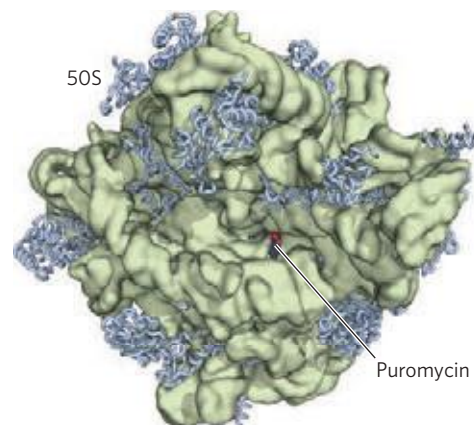


FIGURE 1 The 50S subunit of a bacterial ribosome (PDB ID 1Q7Y). The protein backbones are shown as blue wormlike structures; the rRNA components are transparent. The unstructured extensions of many of the ribosomal proteins snake into the rRNA structures, helping to stabilize them. Bound puromycin, in red, marks the rRNA active site for peptidyl transferase.

about 6,000 to 75,000. Most of the proteins have globular domains arranged on the ribosome surface. Some also have snakelike extensions that protrude into the rRNA core of the ribosome, stabilizing its structure. The functions of some of these proteins have not yet been elucidated in detail, although a structural role seems evident for many of them.

The sequences of the rRNAs of many organisms are now known. Each of the three single-stranded rRNAs of *E. coli* has a specific three-dimensional conformation featuring extensive intrachain base pairing. The folding patterns of the rRNAs are highly conserved in all organisms, particularly the regions implicated in key func-

tions (**Fig. 27-15**). The predicted secondary structure of the rRNAs has largely been confirmed by structural analysis but fails to convey the extensive network of tertiary interactions apparent in the complete structure.

The ribosomes of eukaryotic cells (other than mitochondrial and chloroplast ribosomes) are larger and more complex than bacterial ribosomes (**Fig. 27-16**; compare Fig. 27-14b), with a diameter of about 23 nm and a sedimentation coefficient of about 80S. They also have two subunits, which vary in size among species but on average are 60S and 40S. Altogether, eukaryotic ribosomes contain more than 80 different proteins. The ribosomes of mitochondria and chloroplasts are somewhat smaller

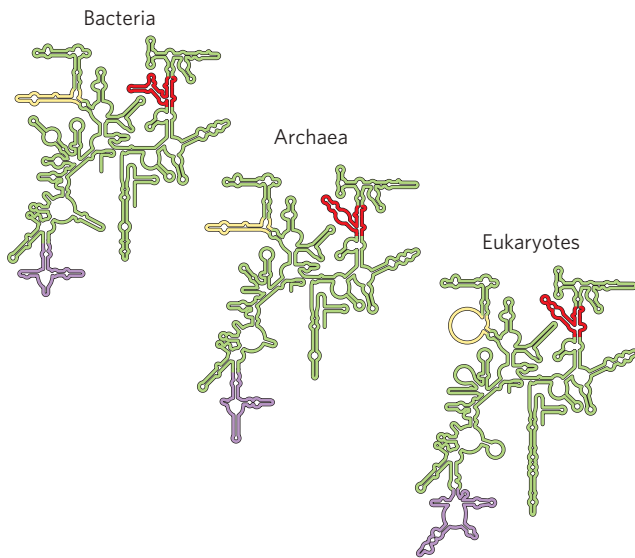


FIGURE 27-15 Conservation of secondary structure in the small subunit rRNAs from the three domains of life. The red, yellow, and purple indicate areas where the structures of the rRNAs from bacteria, archaea, and eukaryotes have diverged. Conserved regions are shown in green.

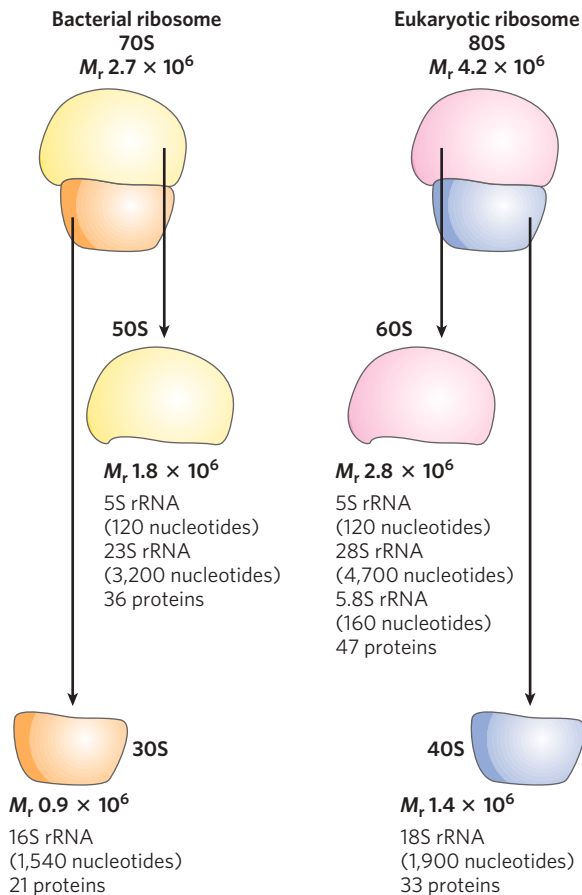


FIGURE 27-16 Summary of the composition and mass of ribosomes in bacteria and eukaryotes. Ribosomal subunits are identified by their S (Svedberg unit) values, sedimentation coefficients that refer to their rate of sedimentation in a centrifuge. The S values are not additive when subunits are combined, because S values are approximately proportional to the $2/3$ power of molecular weight and are also slightly affected by shape.

and simpler than bacterial ribosomes. Nevertheless, ribosomal structure and function are strikingly similar in all organisms and organelles.

Transfer RNAs Have Characteristic Structural Features

To understand how tRNAs can serve as adaptors in translating the language of nucleic acids into the language of proteins, we must first examine their structure in more detail. Transfer RNAs are relatively small and consist of a single strand of RNA folded into a precise three-dimensional structure (see Fig. 8–25a). The tRNAs in bacteria and in the cytosol of eukaryotes have between 73 and 93 nucleotide residues, corresponding to molecular weights of 24,000 to 31,000. Mitochondria and chloroplasts contain distinctive, somewhat smaller tRNAs. Cells have at least one kind of tRNA for each amino acid; at least 32 tRNAs are required to recognize all the amino acid codons (some recognize more than one codon), but some cells use more than 32.

Yeast alanine tRNA (tRNA^{Ala}) was the first nucleic acid to be completely sequenced, by Robert Holley in 1965. It contains 76 nucleotide residues, 10 of which have modified bases. Comparisons of tRNAs from various species have revealed many common structural features (Fig. 27-17). Eight or more of the nucleotide residues have modified bases and sugars, many of which are methylated derivatives of the principal bases. Most tRNAs have a guanylate (pG) residue at the 5' end, and all have the trinucleotide sequence CCA(3') at the 3' end. When drawn in two dimensions, the hydrogen-bonding pattern of all tRNAs forms a cloverleaf structure with four arms; the longer tRNAs have a short fifth arm, or extra arm. In three dimensions, a tRNA has the form of a twisted L (Fig. 27-18).



Robert W. Holley,
1922–1993

Two of the arms of a tRNA are critical for its adaptor function. The **amino acid arm** can carry a specific amino acid esterified by its carboxyl group to the 2'- or 3'-hydroxyl group of the A residue at the 3' end of the tRNA. The **anticodon arm** contains the anticodon. The other major arms are the **D arm**, which contains the unusual nucleotide dihydrouridine (D), and the **T ψ C arm**, which contains ribothymidine (T), not usually present in RNAs, and pseudouridine (ψ), which has an unusual carbon–carbon bond between the base and ribose (see Fig. 26–22). The D and T ψ C arms contribute important interactions for the overall folding of tRNA molecules, and the T ψ C arm interacts with the large-subunit rRNA.

Having looked at the structures of ribosomes and tRNAs, we now consider in detail the five stages of protein synthesis.

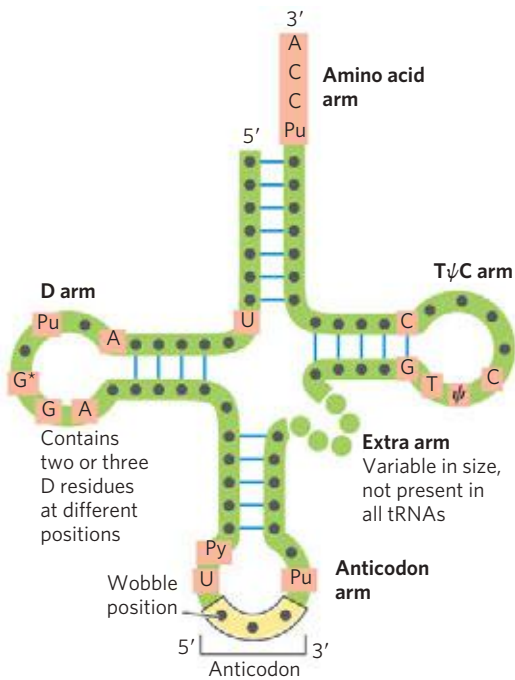
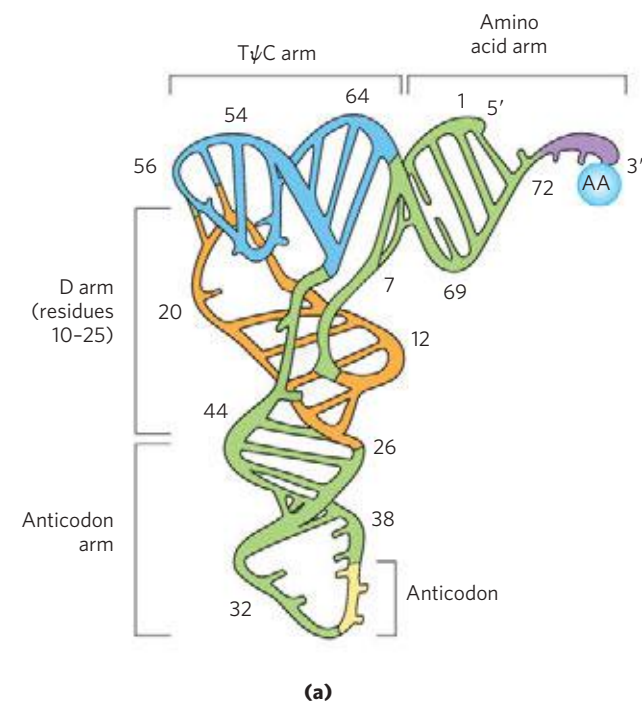


FIGURE 27-17 General cloverleaf secondary structure of tRNAs. The large dots on the backbone represent nucleotide residues; the blue lines represent base pairs. Characteristic and/or invariant residues common to all tRNAs are shaded in light red. Transfer RNAs vary in length from 73 to 93 nucleotides. Extra nucleotides occur in the extra arm or in the D arm. At the end of the anticodon arm is the anticodon loop, which always contains seven unpaired nucleotides. The D arm contains two or three D (5,6-dihydrouridine) residues, depending on the tRNA. In some tRNAs, the D arm has only three hydrogen-bonded base pairs. Symbols are: Pu, purine nucleotide; Py, pyrimidine nucleotide; ψ , pseudouridylate; G*, either guanylate or 2'-O-methylguanylate.



Stage 1: Aminoacyl-tRNA Synthetases Attach the Correct Amino Acids to Their tRNAs

During the first stage of protein synthesis, taking place in the cytosol, aminoacyl-tRNA synthetases esterify the 20 amino acids to their corresponding tRNAs. Each enzyme is specific for one amino acid and one or more corresponding tRNAs. Most organisms have one aminoacyl-tRNA synthetase for each amino acid. For amino acids with two or more corresponding tRNAs, the same enzyme usually aminoacylates all of them.

The structures of all the aminoacyl-tRNA synthetases of *E. coli* have been determined. Researchers have divided them into two classes (Table 27-7) based on substantial differences in primary and tertiary structure and in reaction mechanism (Fig. 27-19); these two classes are the same in all organisms. There is no

TABLE 27-7 The Two Classes of Aminoacyl-tRNA Synthetases

Class I		Class II	
Arg	Leu	Ala	Lys
Cys	Met	Asn	Phe
Gln	Trp	Asp	Pro
Glu	Tyr	Gly	Ser
Ile	Val	His	Thr

Note: Here, Arg represents arginyl-tRNA synthetase, and so forth. The classification applies to all organisms for which tRNA synthetases have been analyzed and is based on protein structural distinctions and on the mechanistic distinction outlined in Figure 27-19.

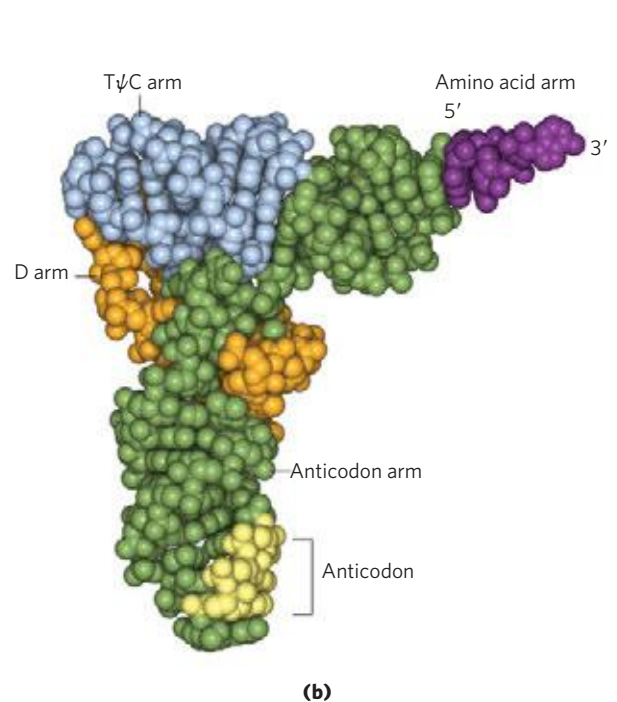
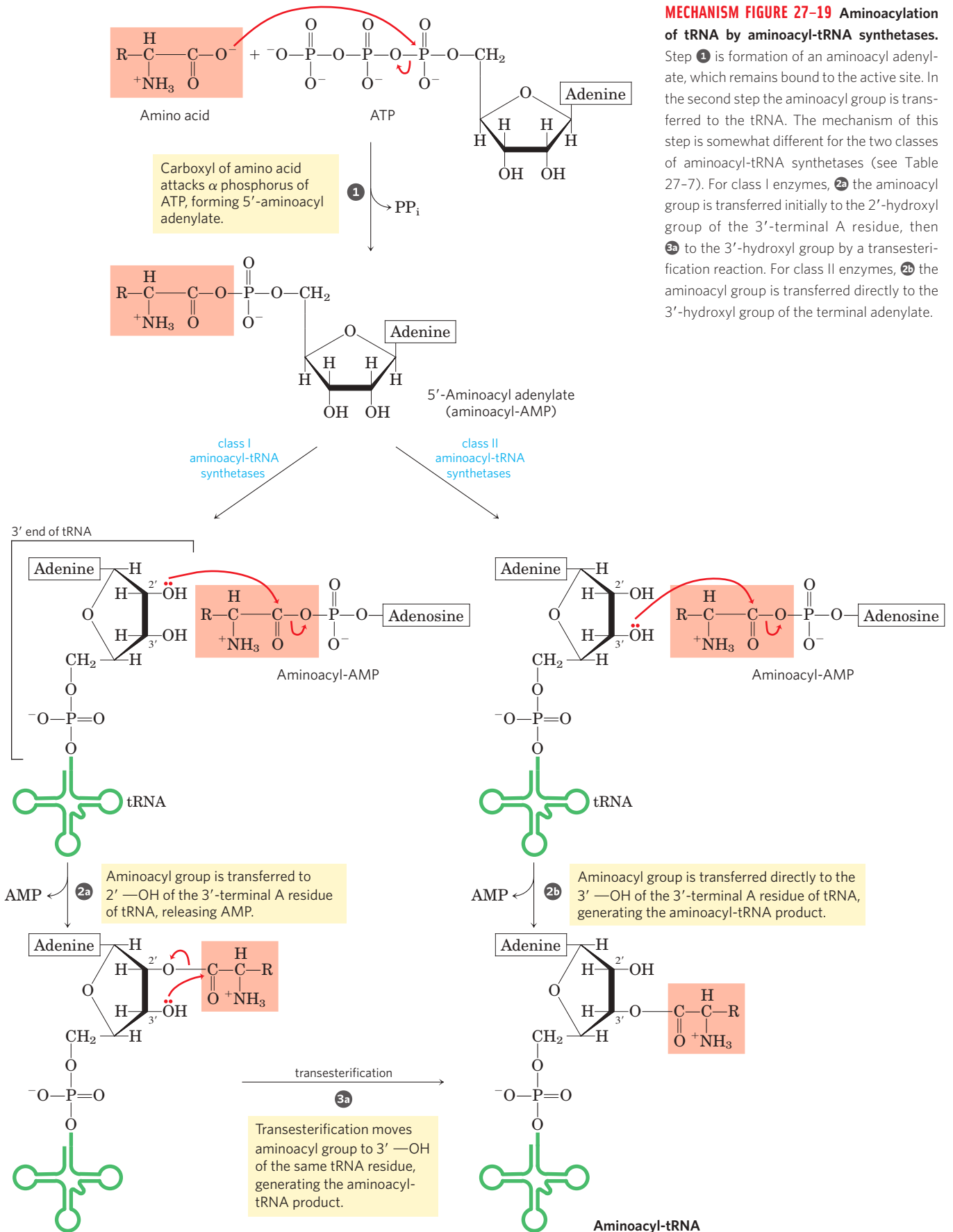


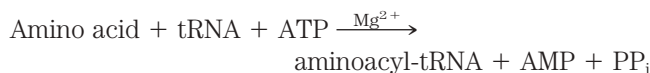
FIGURE 27-18 Three-dimensional structure of yeast tRNA^{Phe} deduced from x-ray diffraction analysis. The shape resembles a twisted L. (a) Schematic diagram with the various arms identified in Figure 27-17

shaded in different colors. (b) A space-filling model, with the same color coding (PDB ID 4TRA). The CCA sequence at the 3' end (purple) is the attachment point for the amino acid.

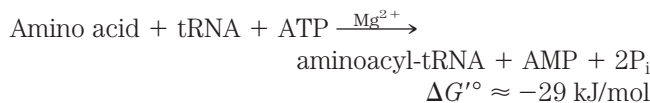


evidence that the two classes share a common ancestor, and the biological, chemical, or evolutionary reasons for two enzyme classes for essentially identical processes remain obscure.

The reaction catalyzed by an aminoacyl-tRNA synthetase is



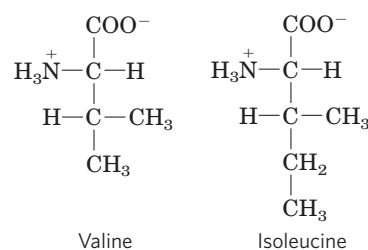
This reaction occurs in two steps in the enzyme's active site. In step ① (Fig. 27-19) an enzyme-bound intermediate, aminoacyl adenylate (aminoacyl-AMP), is formed. In the second step the aminoacyl group is transferred from enzyme-bound aminoacyl-AMP to its corresponding specific tRNA. The course of this second step depends on the class to which the enzyme belongs, as shown by pathways 2a and 2b in Figure 27-19. The resulting ester linkage between the amino acid and the tRNA (Fig. 27-20) has a highly negative standard free energy of hydrolysis ($\Delta G'^{\circ} = -29 \text{ kJ/mol}$). The pyrophosphate formed in the activation reaction undergoes hydrolysis to phosphate by inorganic pyrophosphatase. Thus *two* high-energy phosphate bonds are ultimately expended for each amino acid molecule activated, rendering the overall reaction for amino acid activation essentially irreversible:



Proofreading by Aminoacyl-tRNA Synthetases The aminoacylation of tRNA accomplishes two ends: (1) it activates an amino acid for peptide bond formation and (2) it

ensures appropriate placement of the amino acid in a growing polypeptide. The identity of the amino acid attached to a tRNA is not checked on the ribosome, so attachment of the correct amino acid to the tRNA is essential to the fidelity of protein synthesis.

As you will recall from Chapter 6, enzyme specificity is limited by the binding energy available from enzyme-substrate interactions. Discrimination between two similar amino acid substrates has been studied in detail in the case of Ile-tRNA synthetase, which distinguishes between valine and isoleucine, amino acids that differ by only a single methylene group ($-\text{CH}_2-$):



Ile-tRNA synthetase favors activation of isoleucine (to form Ile-AMP) over valine by a factor of 200—as we would expect, given the amount by which a methylene group (in Ile) could enhance substrate binding. Yet valine is erroneously incorporated into proteins in positions normally occupied by an Ile residue at a frequency of only about 1 in 3,000. How is this greater than 10-fold increase in accuracy brought about? Ile-tRNA synthetase, like some other aminoacyl-tRNA synthetases, has a proofreading function.

Recall a general principle from the discussion of proofreading by DNA polymerases (see Fig. 25-7): if available binding interactions do not provide sufficient discrimination between two substrates, the necessary specificity can be achieved by substrate-specific binding in *two successive* steps. The effect of forcing the system through two successive filters is multiplicative. In the case of Ile-tRNA synthetase, the first filter is the initial binding of the amino acid to the enzyme and its activation to aminoacyl-AMP. The second is the binding of any *incorrect* aminoacyl-AMP products to a separate active site on the enzyme; a substrate that binds in this second active site is hydrolyzed. The R group of valine is slightly smaller than that of isoleucine, so Val-AMP fits the hydrolytic (proofreading) site of the Ile-tRNA synthetase but Ile-AMP does not. Thus Val-AMP is hydrolyzed to valine and AMP in the proofreading active site, and tRNA bound to the synthetase does not become aminoacylated to the wrong amino acid.

In addition to proofreading after formation of the aminoacyl-AMP intermediate, most aminoacyl-tRNA synthetases can hydrolyze the ester linkage between amino acids and tRNAs in the aminoacyl-tRNAs. This hydrolysis is greatly accelerated for incorrectly charged tRNAs, providing yet a third filter to enhance the fidelity of the

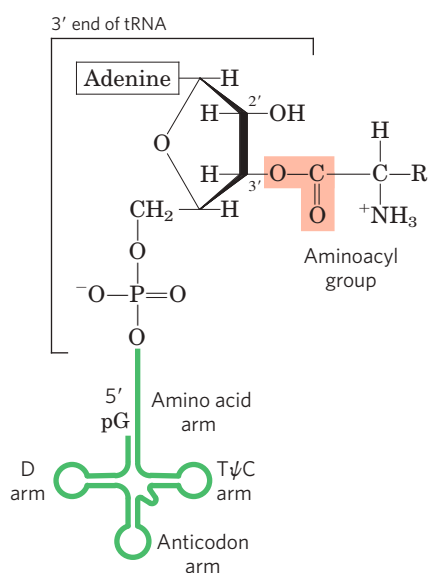


FIGURE 27-20 General structure of aminoacyl-tRNAs. The aminoacyl group is esterified to the 3' position of the terminal A residue. The ester linkage that both activates the amino acid and joins it to the tRNA is shaded light red.

overall process. The few aminoacyl-tRNA synthetases that activate amino acids with no close structural relatives (Cys-tRNA synthetase, for example) demonstrate little or no proofreading activity; in these cases, the active site for aminoacylation can sufficiently discriminate between the proper substrate and any incorrect amino acid.

The overall error rate of protein synthesis (~ 1 mistake per 10^4 amino acids incorporated) is not nearly as low as that of DNA replication. Because flaws in a protein are eliminated when the protein is degraded and are not passed on to future generations, they have less biological significance. The degree of fidelity in protein synthesis is sufficient to ensure that most proteins contain no mistakes and that the large amount of energy required to synthesize a protein is rarely wasted. One defective protein molecule is usually unimportant when many correct copies of the same protein are present.

Interaction between an Aminoacyl-tRNA Synthetase and a tRNA: A "Second Genetic Code" An individual aminoacyl-tRNA synthetase must be specific not only for a single amino acid but for certain tRNAs as well. Discriminating among dozens of tRNAs is just as important for the overall fidelity of protein biosynthesis as is distinguishing among amino acids. The interaction between aminoacyl-tRNA synthetases and tRNAs has been referred to

as the "second genetic code," reflecting its critical role in maintaining the accuracy of protein synthesis. The "coding" rules appear to be more complex than those in the "first" code.

Figure 27-21 summarizes what we know about the nucleotides involved in recognition by some aminoacyl-tRNA synthetases. Some nucleotides are conserved in all tRNAs and therefore cannot be used for discrimination. By observing changes in nucleotides that alter substrate specificity, researchers have identified nucleotide positions that are involved in discrimination by the aminoacyl-tRNA synthetases. These nucleotide positions seem to be concentrated in the amino acid arm and the anticodon arm, including the nucleotides of the anticodon itself, but are also located in other parts of the tRNA molecule. Determination of the crystal structures of aminoacyl-tRNA synthetases complexed with their cognate tRNAs and ATP has added a great deal to our understanding of these interactions (**Fig. 27-22**).

Ten or more specific nucleotides may be involved in recognition of a tRNA by its specific aminoacyl-tRNA synthetase. But in a few cases the recognition mechanism is quite simple. Across a range of organisms from bacteria to humans, the primary determinant of tRNA recognition by the Ala-tRNA synthetases is a single G=U base pair in the

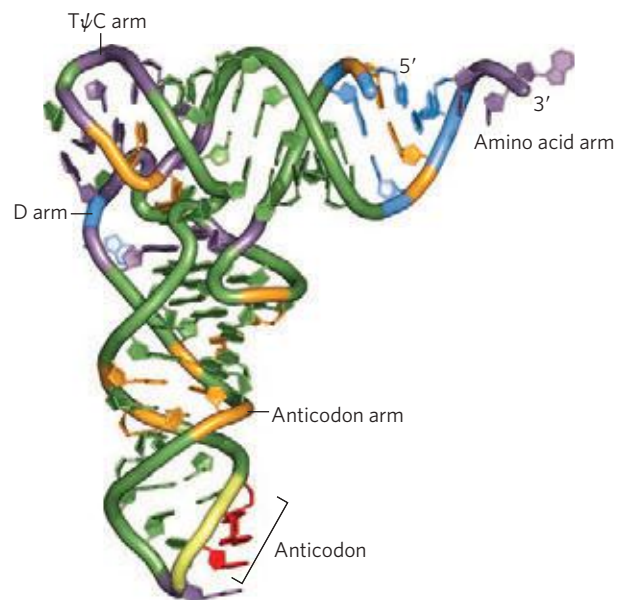
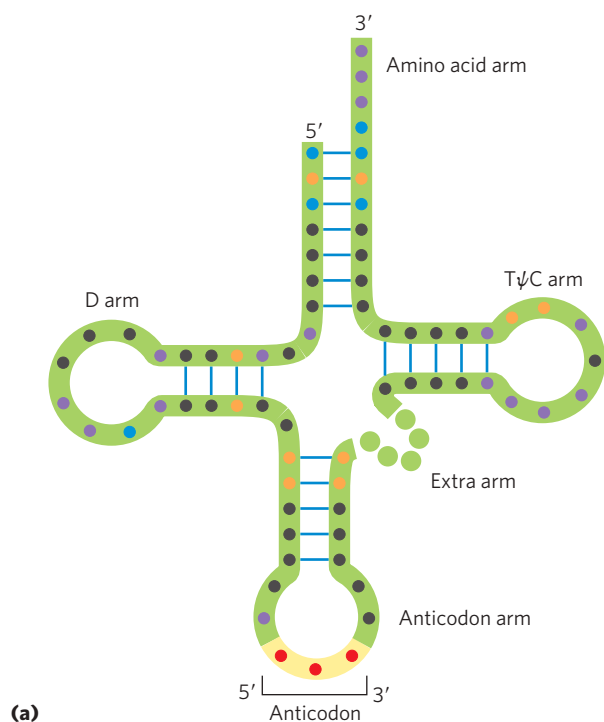


FIGURE 27-21 Nucleotide positions in tRNAs that are recognized by aminoacyl-tRNA synthetases. **(a)** Some positions (purple dots) are the same in all tRNAs and therefore cannot be used to discriminate one from another. Other positions are known recognition points for one (orange) or more (blue) aminoacyl-tRNA synthetases. Structural features other

than sequence are important for recognition by some of the synthetases. **(b)** (PDB ID 1EHZ) The same structural features are shown in three dimensions, with the orange and blue residues again representing positions recognized by one or more aminoacyl-tRNA synthetases, respectively.

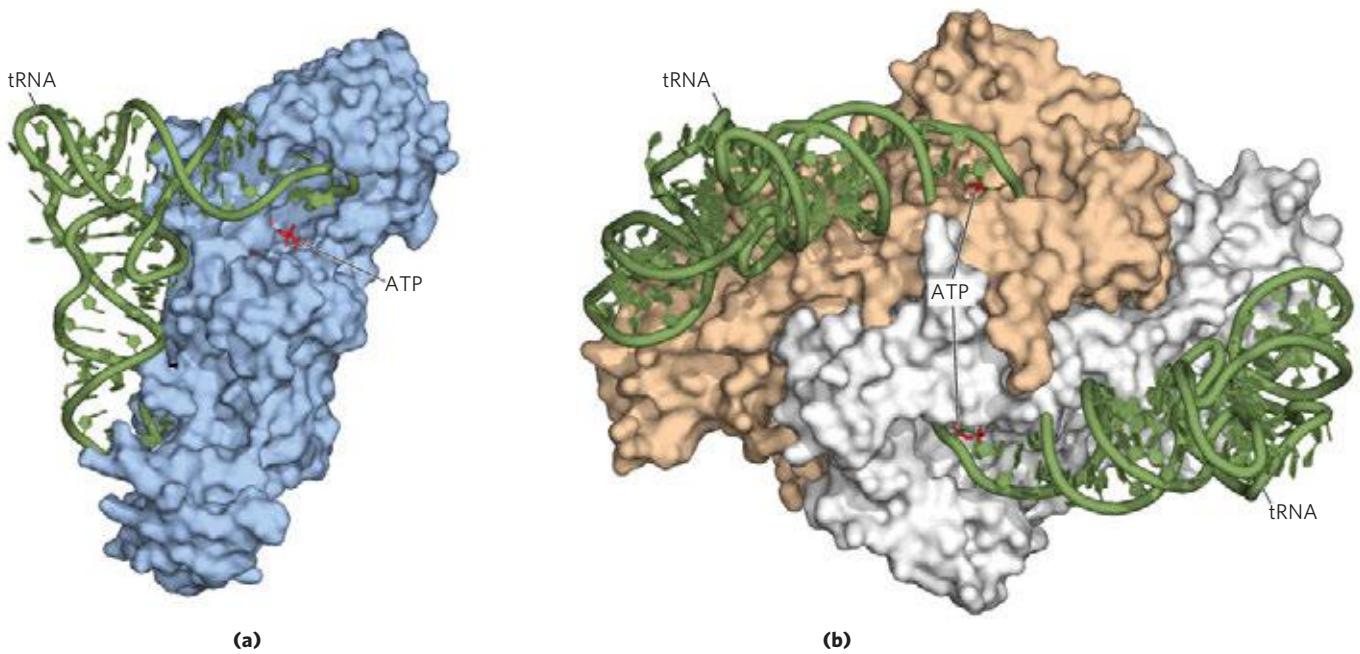


FIGURE 27-22 Aminoacyl-tRNA synthetases. Both synthetases are complexed with their cognate tRNAs (green). Bound ATP (red) pinpoints the active site near the end of the aminoacyl arm. **(a)** Gln-tRNA synthetase

from *E. coli*, a typical monomeric class I synthetase (PDB ID 1QRT). **(b)** Asp-tRNA synthetase from yeast, a typical dimeric class II synthetase (PDB ID 1ASZ).

amino acid arm of tRNA^{Ala} (**Fig. 27-23a**). A short synthetic RNA with as few as 7 bp arranged in a simple hairpin minihelix is efficiently aminoacylated by the Ala-tRNA synthetase, as long as the RNA contains the critical G=U (Fig. 27-23b). This relatively simple alanine system may be an evolutionary relic of a period when RNA oligonucleotides, ancestors to tRNA, were aminoacylated in a primitive system for protein synthesis.

The interaction of aminoacyl-tRNA synthetases and their cognate tRNAs is critical to accurate reading of the genetic code. Any expansion of the code to include new amino acids would necessarily require a new aminoacyl-tRNA synthetase:tRNA pair. A limited expansion of the genetic code has been observed in nature; a more extensive expansion has been accomplished in the laboratory (Box 27-3).

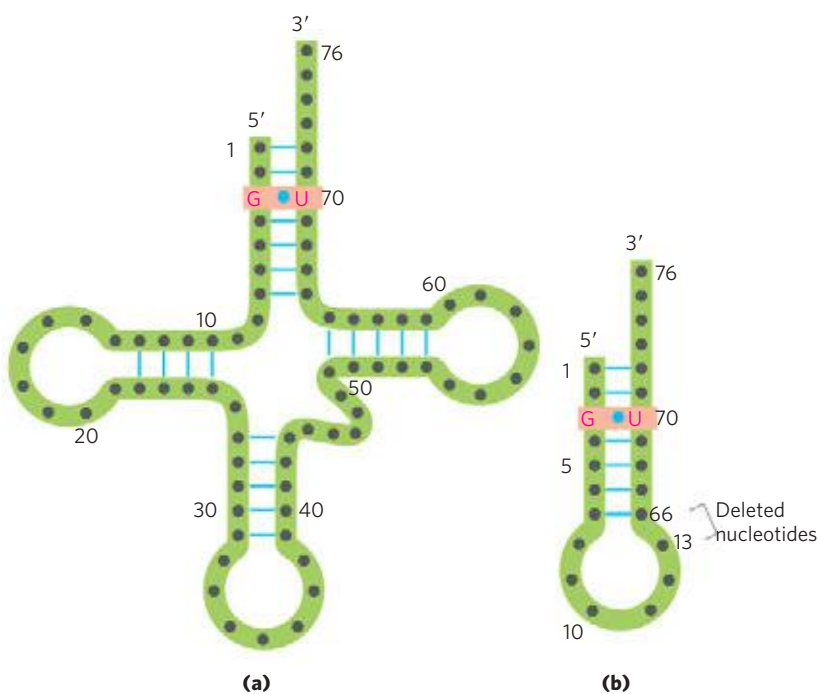


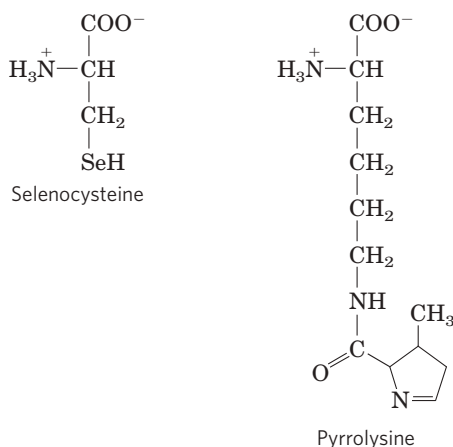
FIGURE 27-23 Structural elements of tRNA^{Ala} that are required for recognition by Ala-tRNA synthetase. **(a)** The tRNA^{Ala} structural elements recognized by the Ala-tRNA synthetase are unusually simple. A single G=U base pair (light red) is the only element needed for specific binding and aminoacylation. **(b)** A short synthetic RNA minihelix, with the critical G=U base pair but lacking most of the remaining tRNA structure. This is aminoacylated specifically with alanine almost as efficiently as the complete tRNA^{Ala}.

BOX 27-3 Natural and Unnatural Expansion of the Genetic Code

As we have seen, the 20 amino acids commonly found in proteins offer only limited chemical functionality. Living systems generally overcome these limitations by using enzymatic cofactors or by modifying particular amino acids after they have been incorporated into proteins. In principle, expansion of the genetic code to introduce new amino acids into proteins offers another route to new functionality, but it is a very difficult route to exploit. Such a change might just as easily result in the inactivation of thousands of cellular proteins.

Expanding the genetic code to include a new amino acid requires several cellular changes. A new aminoacyl-tRNA synthetase must generally be present, along with a cognate tRNA. Both of these components must be highly specific, interacting only with each other and the new amino acid. Significant concentrations of the new amino acid must be present in the cell, which may entail the evolution of new metabolic pathways. As outlined in Box 27-1, the anticodon on the tRNA would most likely pair with a codon that normally specifies termination. Making all of this work in a living cell seems unlikely, but it has happened both in nature and in the laboratory.

There are actually 22 rather than 20 amino acids specified by the known genetic code. The two extra ones are selenocysteine and pyrrolysine, each found in only very few proteins but both offering a glimpse into the intricacies of code evolution.



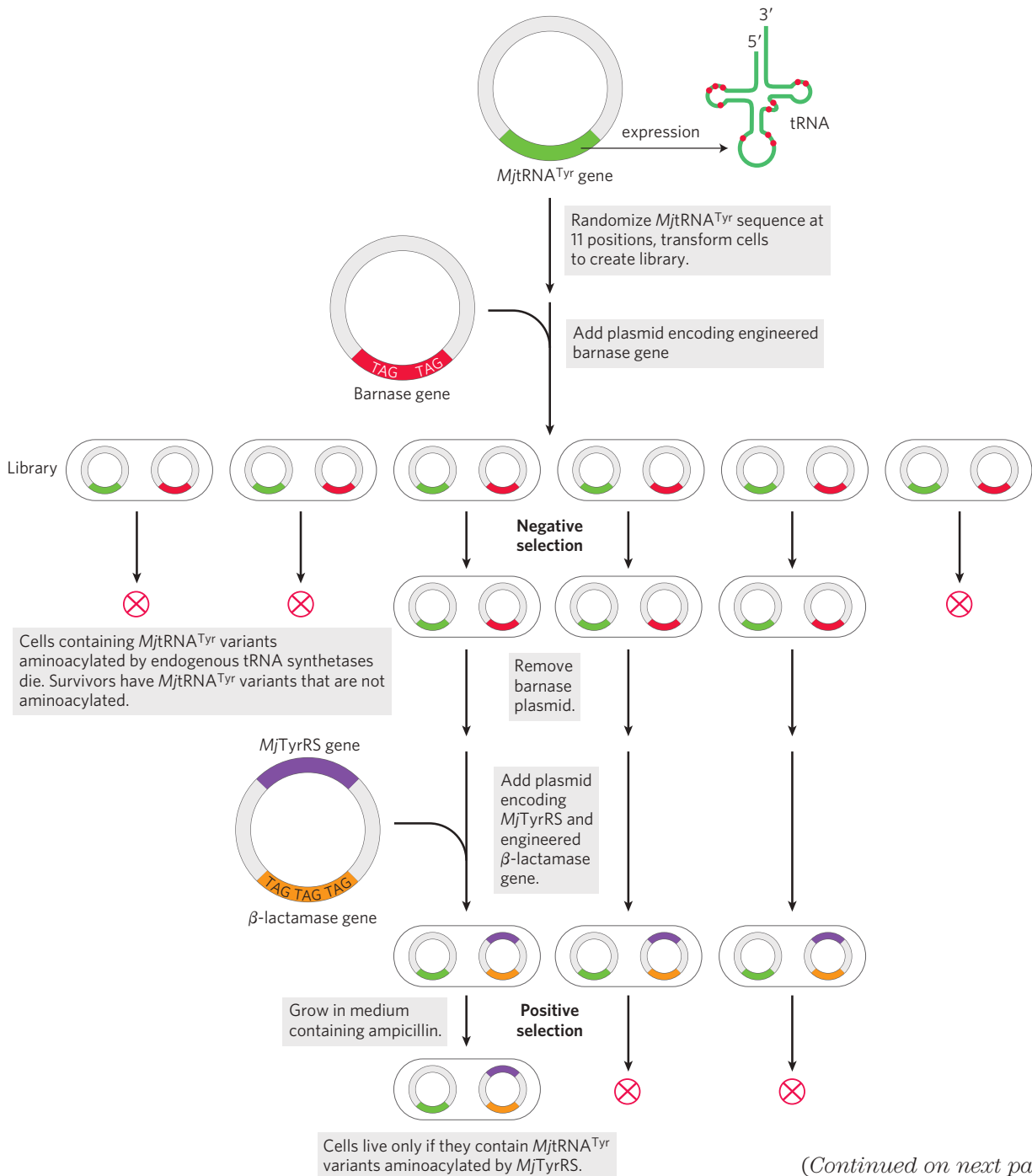
A few proteins in all cells (such as formate dehydrogenase in bacteria and glutathione peroxidase in mammals) require selenocysteine for their activity. In *E. coli* selenocysteine is introduced into the enzyme formate dehydrogenase during translation, in response to an in-frame UGA codon. A special type of Ser-tRNA, present at lower levels than other Ser-tRNAs, recognizes UGA and no other codons. This tRNA is charged with serine by the normal serine aminoacyl-tRNA

synthetase, and the serine is enzymatically converted to selenocysteine by a separate enzyme before its use at the ribosome. The charged tRNA does not recognize just any UGA codon; some contextual signal in the mRNA, still to be identified, ensures that this tRNA recognizes only the few UGA codons, within certain genes, that specify selenocysteine. In effect, UGA doubles as a codon for both termination and (very occasionally) selenocysteine. This particular code expansion has a dedicated tRNA, as described above, but it lacks a dedicated cognate aminoacyl-tRNA synthetase. The process works for selenocysteine, but one might consider it an intermediate step in the evolution of a complete new codon definition.

Pyrrolysine is found in a group of anaerobic archaea called methanogens (see Box 22-1). These organisms produce methane as a required part of their metabolism, and the Methanosarcinaceae group can use methylamines as substrates for methanogenesis. Producing methane from monomethylamine requires the enzyme monomethylamine methyltransferase. The gene encoding this enzyme has an in-frame UAG termination codon. The structure of the methyltransferase was elucidated in 2002, revealing the presence of the novel amino acid pyrrolysine at the position specified by the UAG codon. Subsequent experiments demonstrated that—unlike selenocysteine—pyrrolysine was attached directly to a dedicated tRNA by a cognate pyrrolysyl-tRNA synthetase. These cells produce pyrrolysine via a metabolic pathway that remains to be elucidated. The overall system has all the hallmarks of an established codon assignment, although it only works for UAG codons in this particular gene. As in the case of selenocysteine, there are probably contextual signals that direct this tRNA to the correct UAG codon.

Can scientists match this evolutionary feat? Modification of proteins with various functional groups can provide important insights into the activity and/or structure of the proteins. However, protein modification is often quite laborious. For example, an investigator who wishes to attach a new group to a particular Cys residue will have to somehow block the other Cys residues that may be present on the same protein. If one could instead adapt the genetic code to enable a cell to insert a modified amino acid at a particular location in a protein, the process could be rendered much more convenient. Peter Schultz and coworkers have done just that.

To develop a new codon assignment, one again needs a new aminoacyl-tRNA synthetase and a novel cognate tRNA, both adapted to work only with a particular new amino acid. Efforts to create such an “unnatural” code expansion initially focused on *E. coli*. The codon UAG was chosen as the best target for



(Continued on next page)

FIGURE 1 Selecting *MjtRNA^{Tyr}* variants that function only with the tyrosyl-tRNA synthetase *MjTyrRS*. The sequence of the gene encoding *MjtRNA^{Tyr}*, on a plasmid, is randomized at 11 positions that do not interact with *MjTyrRS* (red dots). The mutagenized plasmids are introduced into *E. coli* cells to create a library of millions of *MjtRNA^{Tyr}* variants, represented by the six cells shown here. The toxic barnase gene, engineered to include the sequence TAG so that its transcript includes UAG stop codons, is provided on a separate plasmid, providing a negative selection. If this gene is expressed, the cells die. It can only be expressed if the *MjtRNA^{Tyr}* variant expressed by that particular cell is

aminoacylated by endogenous (*E. coli*) aminoacyl-tRNA synthetases, inserting an amino acid instead of stopping translation. Another gene, encoding β -lactamase, and also engineered with TAG sequences to produce UAG stop codons, is provided on yet another plasmid that also expresses the gene encoding *MjTyrRS*. This serves as a means of positive selection for the remaining *MjtRNA^{Tyr}* variants. Those variants that are aminoacylated by *MjTyrRS* allow expression of the β -lactamase gene, which allows cells to grow on ampicillin. Multiple rounds of negative and positive selection yield the best *MjtRNA^{Tyr}* variants that are aminoacylated uniquely by *MjTyrRS* and used efficiently in translation.

BOX 27-3 Natural and Unnatural Expansion of the Genetic Code (Continued)

encoding a new amino acid. UAG is the least used of the three termination codons, and strains with tRNAs selected to recognize UAG (see Box 27-4) do not exhibit growth defects. To create the new tRNA and tRNA synthetase, the genes for a tyrosyl-tRNA and its cognate tyrosyl-tRNA synthetase were taken from the archaeon *Methanococcus jannaschii* ($MjtRNA^{Tyr}$ and $MjTyrRS$). $MjTyrRS$ does not bind to the anticodon loop of $MjtRNA^{Tyr}$, allowing the anticodon loop to be modified to CUA (complementary to UAG) without affecting the interaction. Because the archaeal and bacterial systems are orthologous, the modified archaeal components could be transferred to *E. coli* cells without disrupting the intrinsic translation system of the cells.

First, the gene encoding $MjtRNA^{Tyr}$ had to be modified to generate an ideal product tRNA—one that was not recognized by any aminoacyl-tRNA synthetases endogenous to *E. coli*, but was aminoacylated by $MjTyrRS$. Finding such a variant could be accomplished via a series of negative and positive selection cycles designed to efficiently sift through variants of the tRNA gene (Fig. 1). Parts of the $MjtRNA^{Tyr}$ sequence were randomized, allowing creation of a library of cells that each expressed a different version of the tRNA. A gene encoding barnase (a ribonuclease toxic to *E. coli*) was engineered so that its mRNA transcript contained several UAG codons, and this gene was also introduced into the cells on a plasmid. If the $MjtRNA^{Tyr}$ variant expressed in a particular cell in the library was aminoacylated by an endogenous tRNA synthetase, it would express the barnase gene and that cell would die (a negative selection). Surviving cells would contain tRNA variants that were not aminoacylated by endogenous tRNA synthetases, but could potentially be aminoacylated by $MjTyrRS$. A positive selection (Fig. 1) was then set up by engineer-

ing the β -lactamase gene (which confers resistance to the antibiotic ampicillin) so that its transcript contained several UAG codons and introducing this gene into the cells along with the gene encoding $MjTyrRS$. Those $MjtRNA^{Tyr}$ variants that could be aminoacylated by $MjTyrRS$ allowed growth on ampicillin only when $MjTyrRS$ was also expressed in the cell. Several rounds of this negative and positive selection scheme identified a new $MjtRNA^{Tyr}$ variant that was not affected by endogenous enzymes, was aminoacylated by $MjTyrRS$, and functioned well in translation.

Second, the $MjTyrRS$ had to be modified to recognize the new amino acid. The gene encoding $MjTyrRS$ was now mutagenized to create a large library of variants. Variants that would aminoacylate the new $MjtRNA^{Tyr}$ variant with endogenous amino acids were eliminated using the barnase gene selection. A second positive selection (similar to the ampicillin selection above) was carried out so that cells would survive only if the $MjtRNA^{Tyr}$ variant was aminoacylated only in the presence of the unnatural amino acid. Several rounds of negative and positive selection generated a cognate tRNA synthetase-tRNA pair that recognized only the unnatural amino acid. These components were then renamed to reflect the unnatural amino acid used in the selection.

Using this approach, many *E. coli* strains have been constructed, each capable of incorporating one particular unnatural amino acid into a protein in response to a UAG codon. The same approach has been used to artificially expand the genetic code of yeast and even mammalian cells. Over 30 different amino acids (Fig. 2) can be introduced site-specifically and efficiently into cloned proteins in this way. The result is an increasingly useful and flexible tool kit with which to advance the study of protein structure and function.

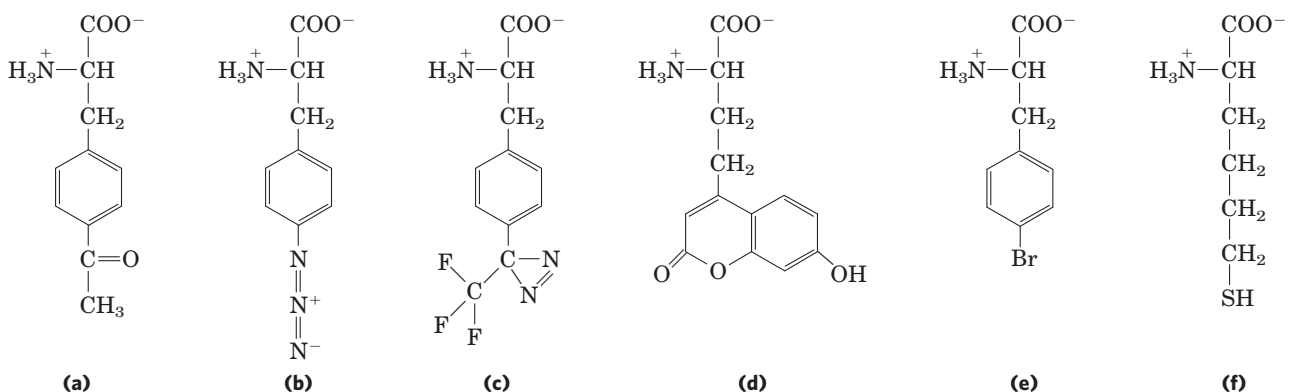


FIGURE 2 A sampling of unnatural amino acids that have been added to the genetic code. These unnatural amino acids add uniquely reactive chemical groups such as **(a)** a ketone, **(b)** an azide, **(c)** a photo-crosslinker (a functional group designed to form a covalent bond with

a nearby group when activated by light), **(d)** a highly fluorescent amino acid, **(e)** an amino acid with a heavy atom (Br) for use in crystallography, and **(f)** a long-chain cysteine analog that can form extended disulfide bonds.

Stage 2: A Specific Amino Acid Initiates Protein Synthesis

Protein synthesis begins at the amino-terminal end and proceeds by the stepwise addition of amino acids to the carboxyl-terminal end of the growing polypeptide, as determined by Howard Dintzis in 1961 (Fig. 27-24). The AUG initiation codon thus specifies an *amino-terminal* methionine residue. Although methionine has only one codon, (5')AUG, all organisms have two tRNAs for methionine. One is used exclusively when (5')AUG is the initiation codon for protein synthesis. The other is used to code for a Met residue in an internal position in a polypeptide.

The distinction between an initiating (5')AUG and an internal one is straightforward. In bacteria, the two types of tRNA specific for methionine are designated tRNA^{Met} and tRNA^{fMet}. The amino acid incorporated in response to the (5')AUG initiation codon is *N*-formylmethionine (fMet). It arrives at the ribosome as *N*-formylmethionyl-tRNA^{fMet} (fMet-tRNA^{fMet}), which is formed in two successive reactions. First, methionine is attached to tRNA^{fMet} by the Met-tRNA synthetase (which in *E. coli* aminoacylates both tRNA^{fMet} and tRNA^{Met}):

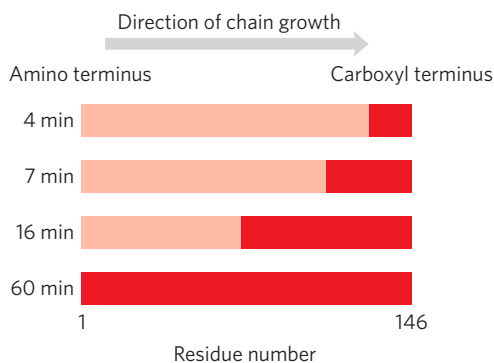
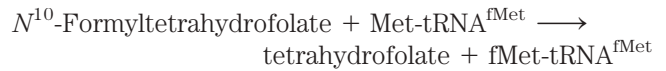
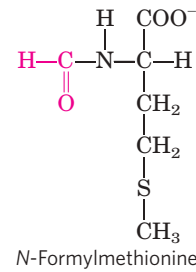


FIGURE 27-24 Proof that polypeptides grow by addition of amino acid residues to the carboxyl end: the Dintzis experiment. Reticulocytes (immature erythrocytes) actively synthesizing hemoglobin were incubated with radioactive leucine (selected because it occurs frequently in both the α - and β -globin chains). Samples of completed α chains were isolated from the reticulocytes at various times afterward, and the distribution of radioactivity was determined. The dark red zones show the portions of completed α -globin chains containing radioactive Leu residues. At 4 min, only a few residues at the carboxyl end of α -globin were labeled, because the only complete globin chains with incorporated label after 4 min were those that had nearly completed synthesis at the time the label was added. With longer incubation times, successively longer segments of the polypeptide contained labeled residues, always in a block at the carboxyl end of the chain. The unlabeled end of the polypeptide (the amino terminus) was thus defined as the initiating end, which means that polypeptides grow by successive addition of amino acids to the carboxyl end.

Next, a transformylase transfers a formyl group from N^{10} -formyltetrahydrofolate to the amino group of the Met residue:



The transformylase is more selective than the Met-tRNA synthetase; it is specific for Met residues attached to tRNA^{fMet}, presumably recognizing some unique structural feature of that tRNA. By contrast, Met-tRNA^{Met} inserts methionine in interior positions in polypeptides.



Addition of the *N*-formyl group to the amino group of methionine by the transformylase prevents fMet from entering interior positions in a polypeptide while also allowing fMet-tRNA^{fMet} to be bound at a specific ribosomal initiation site that accepts neither Met-tRNA^{Met} nor any other aminoacyl-tRNA.

In eukaryotic cells, all polypeptides synthesized by cytosolic ribosomes begin with a Met residue (rather than fMet), but, again, the cell uses a specialized initiating tRNA that is distinct from the tRNA^{Met} used at (5')AUG codons at interior positions in the mRNA. Polypeptides synthesized by mitochondrial and chloroplast ribosomes, however, begin with *N*-formylmethionine. This strongly supports the view that mitochondria and chloroplasts originated from bacterial ancestors that were symbiotically incorporated into precursor eukaryotic cells at an early stage of evolution (see Fig. 1-38).

How can the single (5')AUG codon determine whether a starting *N*-formylmethionine (or methionine, in eukaryotes) or an interior Met residue is ultimately inserted? The details of the initiation process provide the answer.

The Three Steps of Initiation The **initiation** of polypeptide synthesis in bacteria requires (1) the 30S ribosomal subunit, (2) the mRNA coding for the polypeptide to be made, (3) the initiating fMet-tRNA^{fMet}, (4) a set of three proteins called initiation factors (IF-1, IF-2, and IF-3), (5) GTP, (6) the 50S ribosomal subunit, and (7) Mg²⁺. Formation of the initiation complex takes place in three steps (Fig. 27-25).

In step 1 the 30S ribosomal subunit binds two initiation factors, IF-1 and IF-3. Factor IF-3 prevents the 30S and 50S subunits from combining prematurely. The mRNA then binds to the 30S subunit. The initiating (5')AUG is guided to its correct position by the **Shine-Dalgarno sequence** (named for Australian researchers John Shine and Lynn Dalgarno, who identified it) in the

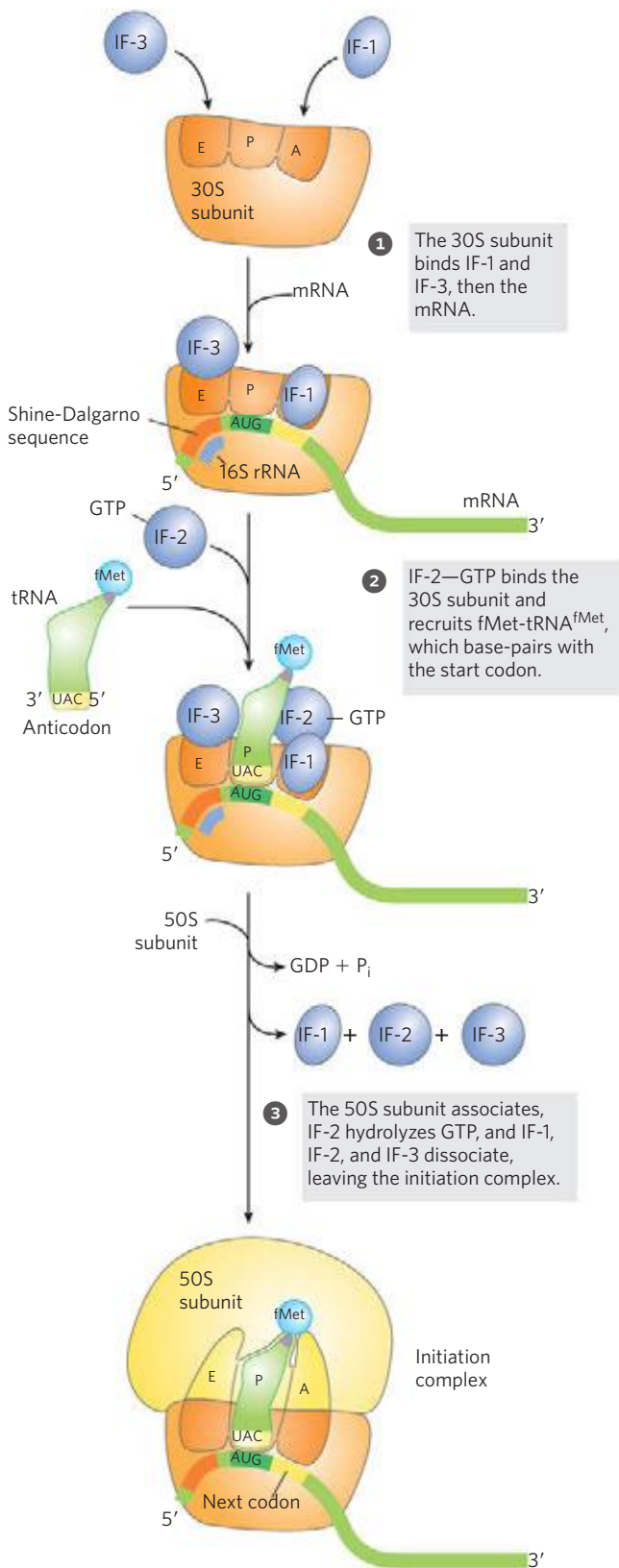


FIGURE 27-25 Formation of the initiation complex in bacteria. The complex forms in three steps (described in the text) at the expense of the hydrolysis of GTP to GDP and P_i. IF-1, IF-2, and IF-3 are initiation factors. E designates the exit site, P the peptidyl site, and A the aminoacyl site. Here the anticodon of the tRNA is oriented 3' to 5', left to right, as in Figure 27-8 but opposite to the orientation in Figures 27-21 and 27-23.

mRNA. This consensus sequence is an initiation signal of four to nine purine residues, 8 to 13 bp to the 5' side of the initiation codon (**Fig. 27-26a**). The sequence base-pairs with a complementary pyrimidine-rich sequence near the 3' end of the 16S rRNA of the 30S ribosomal subunit (Fig. 27-26b). This mRNA-rRNA interaction positions the initiating (5')AUG sequence of the mRNA in the precise position on the 30S subunit where it is required for initiation of translation. The particular (5')AUG where fMet-tRNA^{fMet} is to be bound is distinguished from other methionine codons by its proximity to the Shine-Dalgarno sequence in the mRNA.

Bacterial ribosomes have three sites that bind tRNAs, the **aminoacyl (A) site**, the **peptidyl (P) site**, and the **exit (E) site**. The A and P sites bind to aminoacyl-tRNAs, whereas the E site binds only to uncharged tRNAs that have completed their task on the ribosome. Both the 30S and the 50S subunits contribute to the characteristics of the A and P sites, whereas the E site is largely confined to the 50S subunit. The initiating (5')AUG is positioned at the P site, the only site to which fMet-tRNA^{fMet} can bind (Fig. 27-25). The fMet-tRNA^{fMet} is the only aminoacyl-tRNA that binds first to the P site; during the subsequent elongation stage, all other incoming aminoacyl-tRNAs (including the Met-tRNA^{Met} that binds to interior AUG codons) bind first to the A site and only subsequently to the P and E sites. The E site is the site from which the "uncharged" tRNAs leave during elongation. Factor IF-1 binds at the A site and prevents tRNA binding at this site during initiation.

In step **2** of the initiation process (Fig. 27-25), the complex consisting of the 30S ribosomal subunit, IF-3, and mRNA is joined by both GTP-bound IF-2 and the initiating fMet-tRNA^{fMet}. The anticodon of this tRNA now pairs correctly with the mRNA's initiation codon.

In step **3** this large complex combines with the 50S ribosomal subunit; simultaneously, the GTP bound to IF-2 is hydrolyzed to GDP and P_i, which are released from the complex. All three initiation factors depart from the ribosome at this point.

Completion of the steps in Figure 27-25 produces a functional 70S ribosome called the **initiation complex**, containing the mRNA and the initiating fMet-tRNA^{fMet}. The correct binding of the fMet-tRNA^{fMet} to the P site in the complete 70S initiation complex is assured by at least three points of recognition and attachment: the codon-anticodon interaction involving the initiation AUG fixed in the P site, interaction between the Shine-Dalgarno sequence in the mRNA and the 16S rRNA, and binding interactions between the ribosomal P site and the fMet-tRNA^{fMet}. The initiation complex is now ready for elongation.

Initiation in Eukaryotic Cells Translation is generally similar in eukaryotic and bacterial cells; most of the significant differences are in the number of components and in

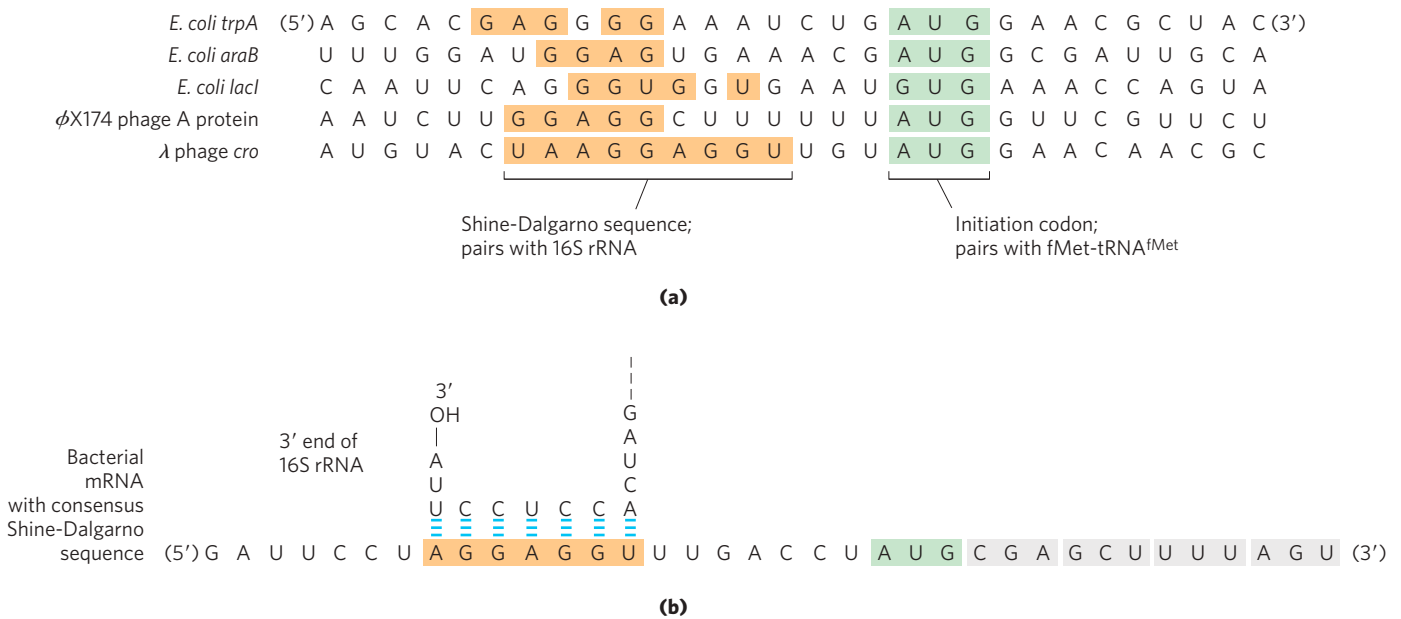


FIGURE 27-26 Messenger RNA sequences that serve as signals for initiation of protein synthesis in bacteria. **(a)** Alignment of the initiating AUG (shaded in green) at its correct location on the 30S ribosomal subunit depends in part on upstream Shine-Dalgarno sequences (light red). Portions of the mRNA transcripts of five bacterial genes are shown. Note

the unusual example of the *E. coli* LacI protein, which initiates with a GUG (Val) codon (see Box 27-1). In *E. coli*, AUG is the start codon in approximately 91% of the genes, with GUG (7%) and UUG (2%) assuming this role more rarely. **(b)** The Shine-Dalgarno sequence of the mRNA pairs with a sequence near the 3' end of the 16S rRNA.

mechanistic details. The initiation process in eukaryotes is outlined in **Figure 27-27**. Eukaryotic mRNAs are bound to the ribosome as a complex with a number of specific binding proteins. Eukaryotic cells have at least 12 initiation factors. Initiation factors eIF1A and eIF3 are the functional homologs of the bacterial IF-1 and IF-3, binding to the 40S subunit in step ① and blocking tRNA binding to the A site and premature joining of the large and small ribosomal subunits, respectively. The factor eIF1 binds to the E site. The charged initiator tRNA is bound by the initiation factor eIF2, which also has bound GTP. In step ② this ternary complex binds to the 40S ribosomal subunit, along with two other proteins involved in later steps, eIF5 (not shown in Fig. 27-27) and eIF5B. This creates a 43S preinitiation complex. The mRNA binds to the eIF4F complex, which, in step ③, mediates its association with the 43S preinitiation complex. The eIF4F complex is made up of eIF4E (binding to the 5' cap), eIF4A (an ATPase and RNA helicase), and eIF4G (a linker protein). The eIF4G protein binds to eIF3 and eIF4E to provide the first link between the 43S preinitiation complex and the mRNA. The eIF4G also binds to the poly(A) binding protein (PABP) at the 3' end of the mRNA, circularizing the mRNA (**Fig. 27-28**) and facilitating the translational regulation of gene expression, as described in Chapter 28.

The addition of the mRNA and its associated factors creates a 48S complex. This complex scans the bound mRNA, starting at the 5' cap, until an AUG codon is encountered. The scanning process (step ④ in Fig. 27-27)

may be facilitated by the RNA helicase of eIF4A and another bound factor (eIF4B, not shown in Fig. 27-27) whose precise molecular activity is not understood.

Once the initiating AUG site is encountered, the 60S ribosomal subunit associates with the complex in step ⑤, accompanied by the release of many of the initiation factors. This requires the activity of eIF5 and eIF5B. The eIF5 protein promotes the GTPase activity of eIF2, producing an eIF2-GDP complex with reduced affinity for the initiator tRNA. The eIF5B protein is homologous to the bacterial IF-2. It hydrolyzes its bound GTP and triggers dissociation of eIF2-GDP and other initiation factors, followed closely by the association of the 60S subunit. This completes formation of the initiation complex.

The roles of the various bacterial and eukaryotic initiation factors in the overall process are summarized in Table 27-8. The mechanism by which these proteins act is an important area of investigation.

Stage 3: Peptide Bonds Are Formed in the Elongation Stage

The third stage of protein synthesis is **elongation**. Again, we begin with bacterial cells. Elongation requires (1) the initiation complex described above, (2) aminoacyl-tRNAs, (3) a set of three soluble cytosolic proteins called **elongation factors** (EF-Tu, EF-Ts, and EF-G in bacteria), and (4) GTP. Cells use three steps to add

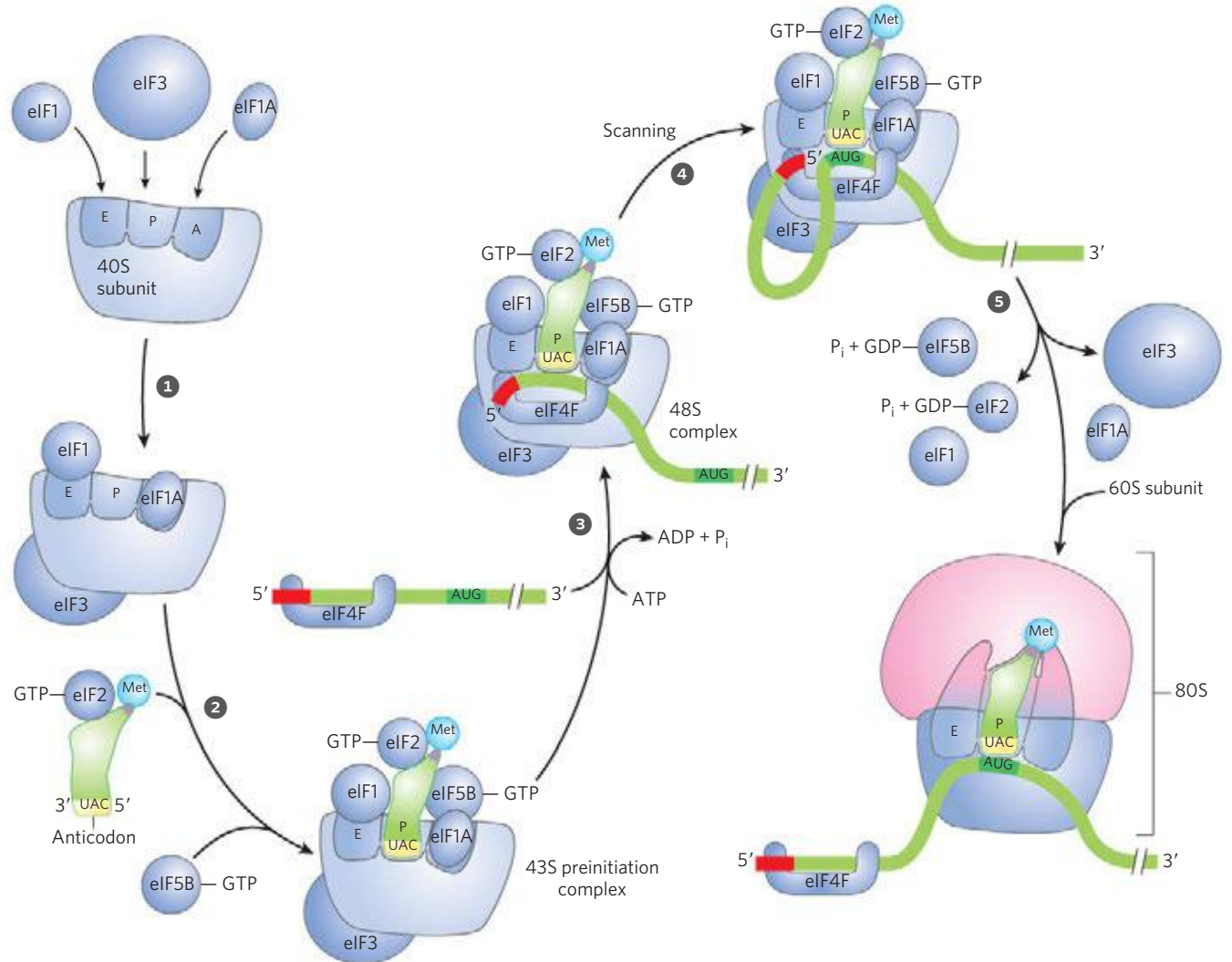


FIGURE 27-27 Initiation of protein synthesis in eukaryotes. The five steps are described in the text. Eukaryotic initiation factors mediate the association of first the charged initiator tRNA to form a 43S complex and

then the mRNA (with the 5' cap shown in red) to form a 48S complex. The final initiation complex is formed as the 60S subunit associates, coupled with the release of most of the initiation factors.

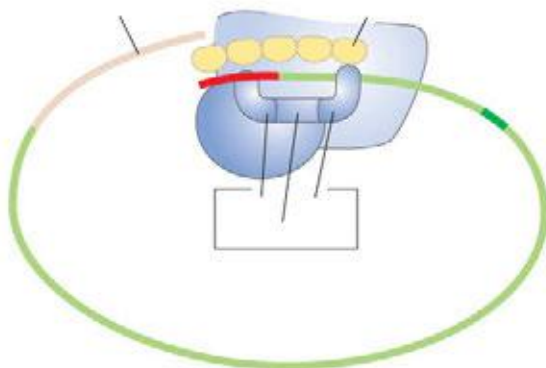


FIGURE 27-28 Circularization of the mRNA in the eukaryotic initiation complex. The 3' and 5' ends of eukaryotic mRNAs are linked by the eIF4F complex of proteins. The eIF4E subunit binds to the 5' cap, and the eIF4G protein binds to the poly(A) binding protein (PABP) at the 3' end of the mRNA. The eIF4G protein also binds to eIF3, linking the circularized mRNA to the 40S subunit of the ribosome.

each amino acid residue, and the steps are repeated as many times as there are residues to be added.

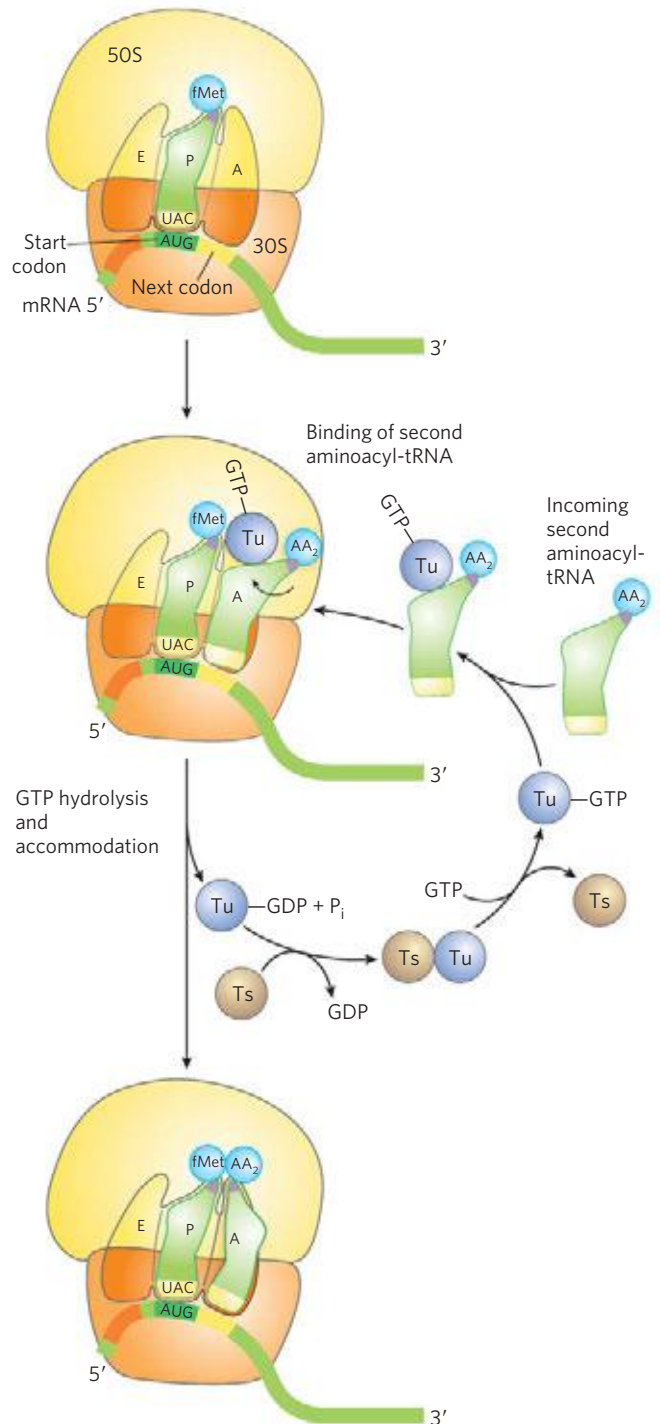
Elongation Step 1: Binding of an Incoming Aminoacyl-tRNA In the first step of the elongation cycle (**Fig. 27-29**), the appropriate incoming aminoacyl-tRNA binds to a complex of GTP-bound EF-Tu. The resulting aminoacyl-tRNA-EF-Tu-GTP complex binds to the A site of the 70S initiation complex. The GTP is hydrolyzed and an EF-Tu-GDP complex is released from the 70S ribosome. The EF-Tu-GTP complex is regenerated in a process involving EF-Ts and GTP.

Elongation Step 2: Peptide Bond Formation A peptide bond is now formed between the two amino acids bound by their tRNAs to the A and P sites on the ribosome. This occurs by the transfer of the initiating *N*-formylmethionyl group from its tRNA to the amino group of the second

TABLE 27-8 Protein Factors Required for Initiation of Translation in Bacterial and Eukaryotic Cells

Factor	Function
Bacterial	
IF-1	Prevents premature binding of tRNAs to A site
IF-2	Facilitates binding of fMet-tRNA ^{fMet} to 30S ribosomal subunit
IF-3	Binds to 30S subunit; prevents premature association of 50S subunit; enhances specificity of P site for fMet-tRNA ^{fMet}
Eukaryotic	
eIF1	Binds to the E site of the 40S subunit; facilitates interaction between eIF2-tRNA-GTP ternary complex and the 40S subunit
eIF1A	Homolog of bacterial IF-1; prevents premature binding of tRNAs to A site
eIF2	GTPase; facilitates binding of initiating Met-tRNA ^{Met} to 40S ribosomal subunit
eIF2B*, eIF3	First factors to bind 40S subunit; facilitate subsequent steps
eIF4F	Complex consisting of eIF4E, eIF4A, and eIF4G
eIF4A	RNA helicase activity; removes secondary structure in the mRNA to permit binding to 40S subunit; part of the eIF4F complex
eIF4B	Binds to mRNA; facilitates scanning of mRNA to locate the first AUG
eIF4E	Binds to the 5' cap of mRNA; part of the eIF4F complex
eIF4G	Binds to eIF4E and to poly(A) binding protein (PABP); part of the eIF4F complex
eIF5*	Promotes dissociation of several other initiation factors from 40S subunit as a prelude to association of 60S subunit to form 80S initiation complex
eIF5b	GTPase homologous to bacterial IF-2; promotes dissociation of initiation factors prior to final ribosome assembly

*Not shown in Figure 27-27.

**FIGURE 27-29** First elongation step in bacteria: binding of the second aminoacyl-tRNA. The second aminoacyl-tRNA (AA₂) enters the A site of the ribosome bound to GTP-bound EF-Tu (shown here as Tu). Binding of the second aminoacyl-tRNA to the A site is accompanied by hydrolysis of the GTP to GDP and P_i and release of the EF-Tu-GDP complex from the ribosome. The bound GDP is released when the EF-Tu-GDP complex binds to EF-Ts, and EF-Ts is subsequently released when another molecule of GTP binds to EF-Tu. This recycles EF-Tu and makes it available to repeat the cycle. Accommodation involves a change in the second tRNA conformation that pulls its aminoacyl end into the peptidyl transferase site.

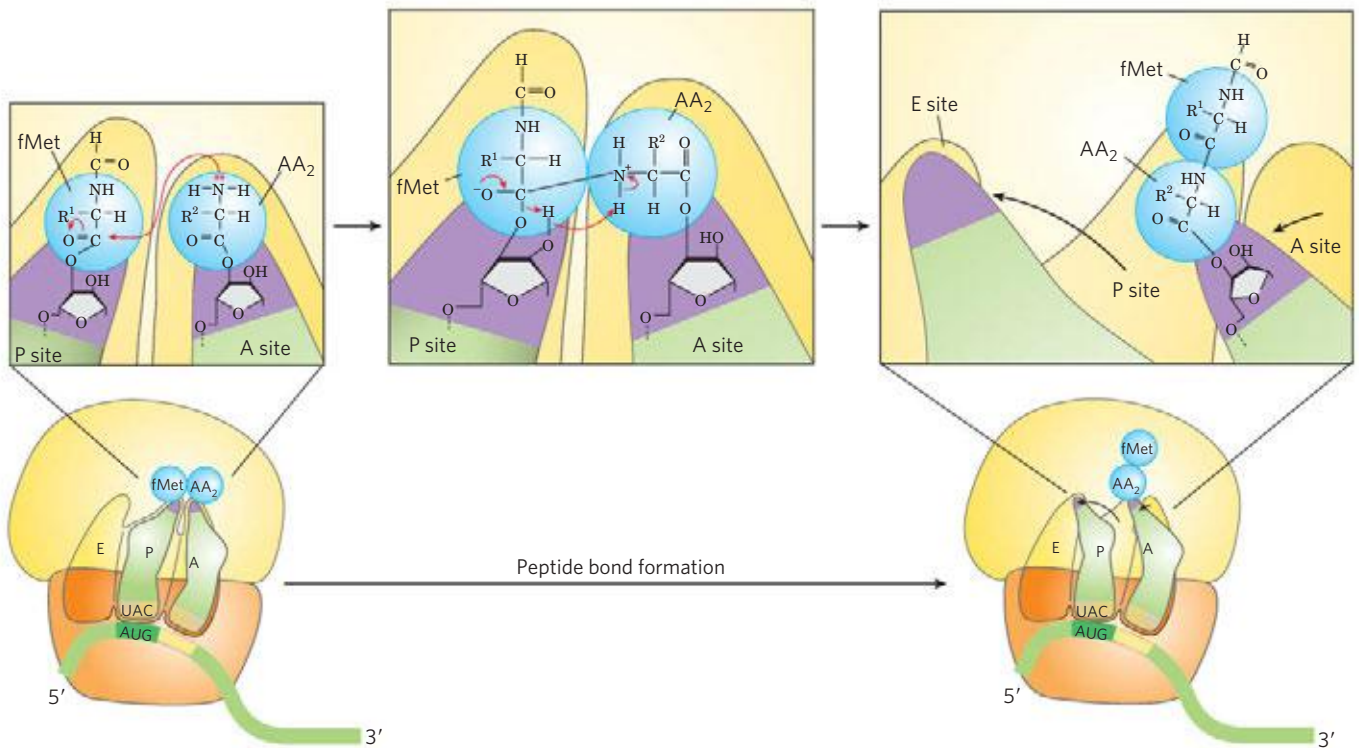


FIGURE 27-30 Second elongation step in bacteria: formation of the first peptide bond. The peptidyl transferase catalyzing this reaction is the 23S rRNA ribozyme. The *N*-formylmethionyl group is transferred to the amino group of the second aminoacyl-tRNA in the A site, forming a dipeptidyl-tRNA. At this stage, both tRNAs bound to the ribosome shift position in

amino acid, now in the A site (**Fig. 27-30**). The α -amino group of the amino acid in the A site acts as a nucleophile, displacing the tRNA in the P site to form the peptide bond. This reaction produces a dipeptidyl-tRNA in the A site, and the now “uncharged” (deacylated) tRNA^{fMet} remains bound to the P site. The tRNAs then shift to a hybrid binding state, with elements of each spanning two different sites on the ribosome, as shown in Figure 27-30.

The enzymatic activity that catalyzes peptide bond formation has historically been referred to as **peptidyl transferase** and was widely assumed to be intrinsic to one or more of the proteins in the large ribosomal subunit. We now know that this reaction is catalyzed by the 23S rRNA, adding to the known catalytic repertoire of ribozymes. This discovery has interesting implications for the evolution of life (see Box 27-2).

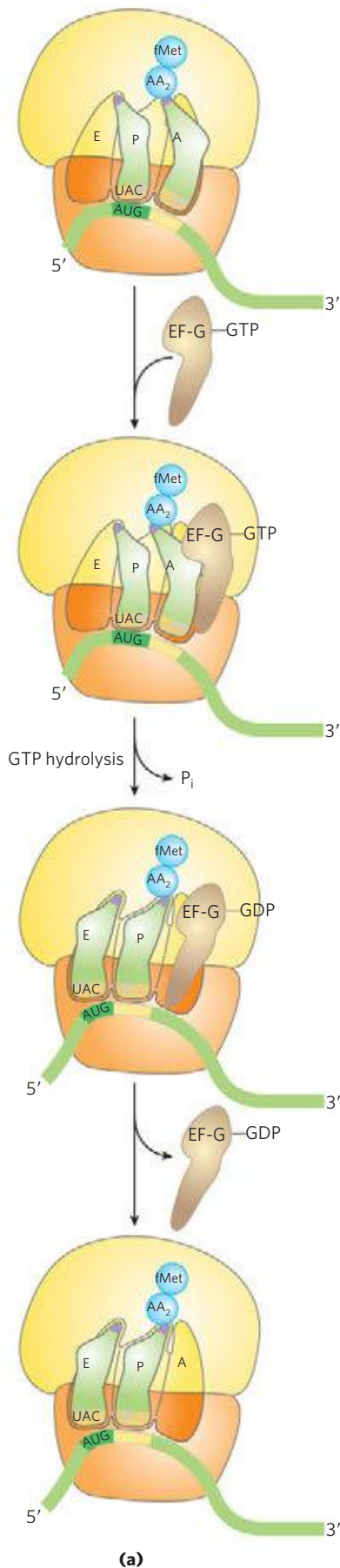
Elongation Step 3: Translocation In the final step of the elongation cycle, **translocation**, the ribosome moves one codon toward the 3' end of the mRNA (**Fig. 27-31a**). This movement shifts the anticodon of the dipeptidyl-tRNA, which is still attached to the second codon of the mRNA, from the A site to the P site, and shifts the deacylated tRNA from the P site to the E site, from where the tRNA is released into the cytosol. The third codon of the mRNA now lies in the A site and the

the 50S subunit to take up a hybrid binding state. The uncharged tRNA shifts so that its 3' and 5' ends are in the E site. Similarly, the 3' and 5' ends of the peptidyl tRNA shift to the P site. The anticodons remain in the P and A sites. Note the involvement of the 2'-hydroxyl group of the 3'-terminal adenosine as a general acid-base catalyst in this reaction.

second codon in the P site. Movement of the ribosome along the mRNA requires EF-G (also known as translocase) and the energy provided by hydrolysis of another molecule of GTP. A change in the three-dimensional conformation of the entire ribosome results in its movement along the mRNA. Because the structure of EF-G mimics the structure of the EF-Tu-tRNA complex (**Fig. 27-31b**), EF-G can bind the A site and presumably displace the peptidyl-tRNA.

After translocation, the ribosome, with its attached dipeptidyl-tRNA and mRNA, is ready for the next elongation cycle and attachment of a third amino acid residue. This process occurs in the same way as addition of the second residue (as shown in Figs 27-29, 27-30, and 27-31). For each amino acid residue correctly added to the growing polypeptide, two GTPs are hydrolyzed to GDP and P_i as the ribosome moves from codon to codon along the mRNA toward the 3' end.

The polypeptide remains attached to the tRNA of the most recent amino acid to be inserted. This association maintains the functional connection between the information in the mRNA and its decoded polypeptide output. At the same time, the ester linkage between this tRNA and the carboxyl terminus of the growing polypeptide activates the terminal carboxyl group for nucleophilic attack by the incoming amino acid to form a new peptide bond (**Fig. 27-30**). As the existing ester linkage between

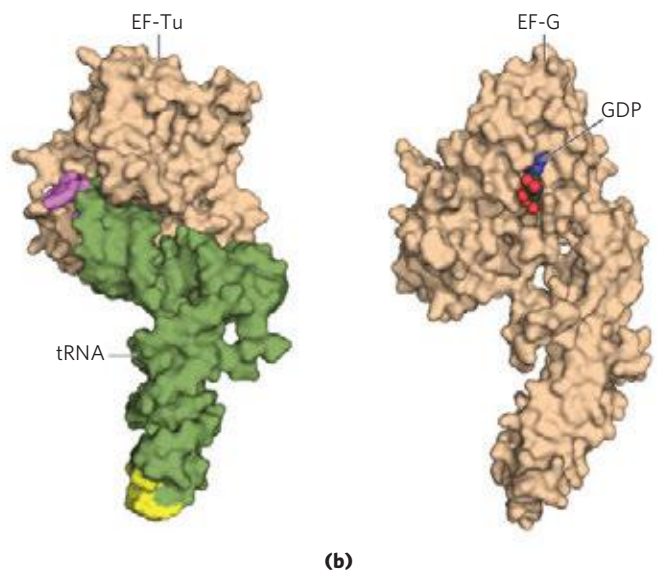


the polypeptide and tRNA is broken during peptide bond formation, the linkage between the polypeptide and the information in the mRNA persists, because each newly added amino acid is still attached to its tRNA.

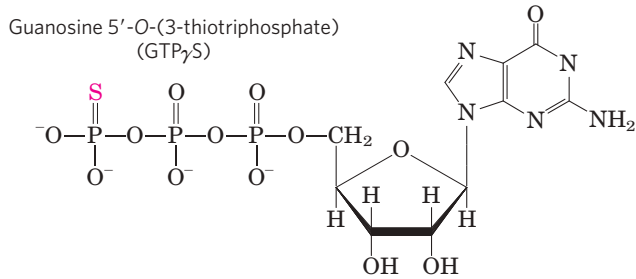
The elongation cycle in eukaryotes is quite similar to that in bacteria. Three eukaryotic elongation factors (eEF1 α , eEF1 $\beta\gamma$, and eEF2) have functions analogous to those of the bacterial elongation factors (EF-Tu, EF-Ts, and EF-G, respectively). Eukaryotic ribosomes do not have an E site; uncharged tRNAs are expelled directly from the P site.

Proofreading on the Ribosome The GTPase activity of EF-Tu during the first step of elongation in bacterial cells (Fig. 27–29) makes an important contribution to the rate and fidelity of the overall biosynthetic process. Both the EF-Tu–GTP and EF-Tu–GDP complexes exist for a few milliseconds before they dissociate. These two intervals provide opportunities for the codon-anticodon interactions to be proofread. Incorrect aminoacyl-tRNAs normally dissociate from the A site during one of these periods. If the GTP analog guanosine 5'-O-(3-thiotriphosphate) (GTP γ S) is used in place of GTP, hydrolysis is

FIGURE 27–31 Third elongation step in bacteria: translocation. (a) The ribosome moves one codon toward the 3' end of the mRNA, using energy provided by hydrolysis of GTP bound to EF-G (translocase). The dipeptidyl-tRNA is now entirely in the P site, leaving the A site open for the incoming (third) aminoacyl-tRNA. The uncharged tRNA later dissociates from the E site, and the elongation cycle begins again. (b) The structure of EF-G mimics the structure of EF-Tu complexed with tRNA. Shown here are (left) EF-Tu complexed with tRNA (PDB ID 1B23) and (right) EF-G complexed with GDP (PDB ID 1DAR). The carboxyl-terminal part of EF-G mimics the structure of the anticodon loop of tRNA in both shape and charge distribution.



slowed, improving the fidelity (by increasing the proofreading intervals) but reducing the rate of protein synthesis.



The process of protein synthesis (including the characteristics of codon-anticodon pairing already described) has clearly been optimized through evolution to balance the requirements for speed and fidelity. Improved fidelity might diminish speed, whereas increases in speed would probably compromise fidelity. And, recall that the proofreading mechanism on the ribosome establishes only that the proper codon-anticodon pairing has taken place, not that the correct amino acid is attached to the tRNA. If a tRNA is successfully aminoacylated with the wrong amino acid (as can be done experimentally), this incorrect amino acid is efficiently incorporated into a protein in response to whatever codon is normally recognized by the tRNA.

Stage 4: Termination of Polypeptide Synthesis Requires a Special Signal

Elongation continues until the ribosome adds the last amino acid coded by the mRNA. **Termination**, the fourth stage of polypeptide synthesis, is signaled by the presence of one of three termination codons in the mRNA (UAA, UAG, UGA), immediately following the final coded amino acid. Mutations in a tRNA anticodon that allow an amino acid to be inserted at a termination codon are generally deleterious to the cell (Box 27–4). In bacteria, once a termination codon occupies the ribosomal A site, three **termination factors**, or **release factors**—the proteins RF-1, RF-2, and RF-3—contribute to (1) hydrolysis of the terminal peptidyl-tRNA bond, (2) release of the free polypeptide and the last tRNA, now uncharged, from the P site, and (3) dissociation of the 70S ribosome into its 30S and 50S subunits, ready to start a new cycle of polypeptide synthesis (Fig. 27–32). RF-1 recognizes the termination codons UAG and UAA, and RF-2 recognizes UGA and UAA. Either RF-1 or RF-2 (depending on which codon is present) binds at a termination codon and induces peptidyl transferase to transfer the growing polypeptide to a water molecule rather than to another amino acid. The release factors have domains thought to mimic the structure of tRNA, as shown for the elongation factor EF-G in Figure 27–31b. The specific function of

BOX 27–4 Induced Variation in the Genetic Code: Nonsense Suppression

When a mutation produces a termination codon in the interior of a gene, translation is prematurely halted and the incomplete polypeptide is usually inactive. These are called nonsense mutations. The gene can be restored to normal function if a second mutation either (1) converts the misplaced termination codon to a codon specifying an amino acid or (2) suppresses the effects of the termination codon. Such restorative mutations are called **nonsense suppressors**; they generally involve mutations in tRNA genes to produce altered (suppressor) tRNAs that can recognize the termination codon and insert an amino acid at that position. Most known suppressor tRNAs have single base substitutions in their anticodons.

Suppressor tRNAs constitute an experimentally induced variation in the genetic code to allow the reading of what are usually termination codons, much like the naturally occurring code variations described in Box 27–1. Nonsense suppression does not completely disrupt normal information transfer in a cell, because the cell usually has several copies of each tRNA gene; some of these duplicate genes are weakly expressed and account for only a minor part of the cellular pool of a particular tRNA. Suppressor mutations usually involve a “minor” tRNA, leaving the major tRNA to read its codon normally.

For example, *E. coli* has three identical genes for tRNA^{Tyr}, each producing a tRNA with the anticodon (5')GUA. One of these genes is expressed at relatively high levels and thus its product represents the major tRNA^{Tyr} species; the other two genes are transcribed in only small amounts. A change in the anticodon of the tRNA product of one of these duplicate tRNA^{Tyr} genes, from (5')GUA to (5')CUA, produces a minor tRNA^{Tyr} species that will insert tyrosine at UAG stop codons. This insertion of tyrosine at UAG is carried out inefficiently, but it can produce enough full-length protein from a gene with a nonsense mutation to allow the cell to survive. The major tRNA^{Tyr} continues to translate the genetic code normally for the majority of proteins.

The mutation that leads to creation of a suppressor tRNA does not always occur in the anticodon. The suppression of UGA nonsense codons generally involves the tRNA^{Trp} that normally recognizes UGG. The alteration that allows it to read UGA (and insert Trp residues at these positions) is a G to A change at position 24 (in an arm of the tRNA somewhat removed from the anticodon); this tRNA can now recognize *both* UGG and UGA. A similar change is found in tRNAs involved in the most common naturally occurring variation in the genetic code (UGA = Trp; see Box 27–1).

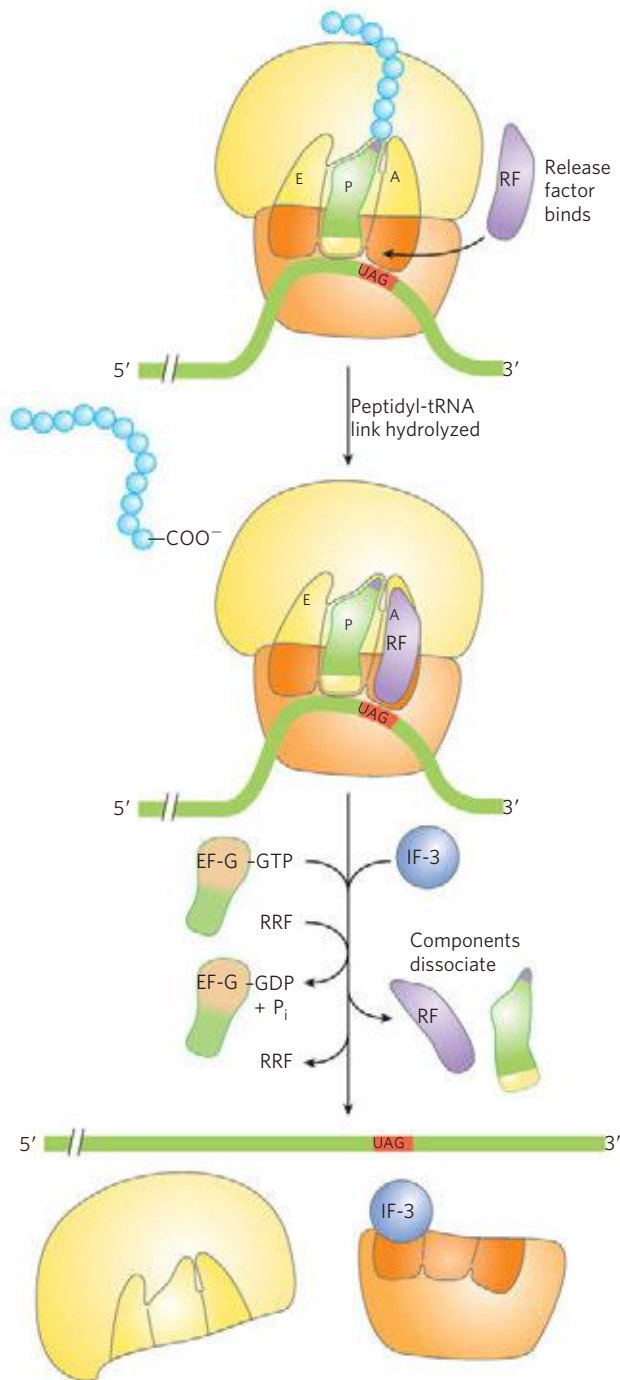


FIGURE 27-32 Termination of protein synthesis in bacteria. Termination occurs in response to a termination codon in the A site. First, a release factor, RF (RF-1 or RF-2, depending on which termination codon is present), binds to the A site. This leads to hydrolysis of the ester linkage between the nascent polypeptide and the tRNA in the P site and release of the completed polypeptide. Finally, the mRNA, deacylated tRNA, and release factor leave the ribosome, which dissociates into its 30S and 50S subunits, aided by ribosome recycling factor (RRF), IF-3, and energy provided by EF-G-mediated GTP hydrolysis. The 30S subunit complex with IF-3 is ready to begin another cycle of translation (see Fig. 27-25).

RF-3 has not been firmly established, although it is thought to release the ribosomal subunit. In eukaryotes, a single release factor, eRF, recognizes all three termination codons.

Ribosome recycling leads to dissociation of the translation components. The release factors dissociate from the posttermination complex (with an uncharged tRNA in the P site) and are replaced by EF-G and a protein called ribosome recycling factor (RRF; M_r 20,300). Hydrolysis of GTP by EF-G leads to dissociation of the 50S subunit from the 30S-tRNA-mRNA complex. EF-G and RRF are replaced by IF-3, which promotes the dissociation of the tRNA. The mRNA is then released. The complex of IF-3 and the 30S subunit is then ready to initiate another round of protein synthesis (Fig. 27-25).

Energy Cost of Fidelity in Protein Synthesis Synthesis of a protein true to the information specified in its mRNA requires energy. Formation of each aminoacyl-tRNA uses two high-energy phosphate groups. An additional ATP is consumed each time an incorrectly activated amino acid is hydrolyzed by the deacylation activity of an aminoacyl-tRNA synthetase as part of its proofreading activity. A GTP is cleaved to GDP and P_i during the first elongation step, and another during the translocation step. Thus, on average, the energy derived from the hydrolysis of more than four NTPs to NDPs is required for the formation of each peptide bond of a polypeptide.

This represents an exceedingly large thermodynamic “push” in the direction of synthesis: at least $4 \times 30.5 \text{ kJ/mol} = 122 \text{ kJ/mol}$ of phosphodiester bond energy to generate a peptide bond, which has a standard free energy of hydrolysis of only about -21 kJ/mol . The net free-energy change during peptide bond synthesis is thus -101 kJ/mol . Proteins are information-containing polymers. The biochemical goal is not simply the formation of a peptide bond but the formation of a peptide bond between two *specified* amino acids. Each of the high-energy phosphate compounds expended in this process plays a critical role in maintaining proper alignment between each new codon in the mRNA and its associated amino acid at the growing end of the polypeptide. This energy permits very high fidelity in the biological translation of the genetic message of mRNA into the amino acid sequence of proteins.

Rapid Translation of a Single Message by Polysomes Large clusters of 10 to 100 ribosomes that are very active in protein synthesis can be isolated from both eukaryotic and bacterial cells. Electron micrographs show a fiber between adjacent ribosomes in the cluster, which is called a **polysome** (Fig. 27-33a). The connecting strand is a single molecule of mRNA that is being translated simultaneously by many closely spaced ribosomes, allowing the highly efficient use of the mRNA.

In bacteria, transcription and translation are tightly coupled. Messenger RNAs are synthesized and translated in the same $5' \rightarrow 3'$ direction. Ribosomes begin translating the $5'$ end of the mRNA before transcription is complete (Fig. 27-33b). The situation is quite different in eukaryotic cells, where newly transcribed mRNAs must leave the nucleus before they can be translated.

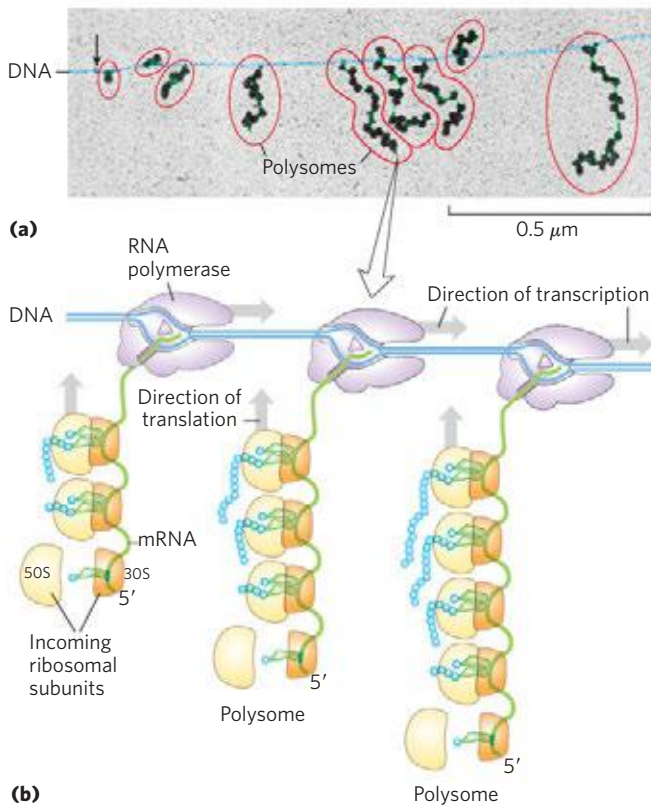


FIGURE 27-33 Coupling of transcription and translation in bacteria.

(a) Electron micrograph of polysomes forming during the transcription of a segment of DNA from *E. coli*. Each mRNA is being translated by many ribosomes simultaneously. The nascent polypeptide chains forming on the ribosomes are difficult to see under the spreading conditions used to prepare the samples shown in these electron micrographs. The arrow marks the approximate beginning of the gene that is being transcribed. (b) Each mRNA is translated by ribosomes while it is still being transcribed from DNA by RNA polymerase. This is possible because the mRNA in bacteria does not have to be transported from a nucleus to the cytoplasm before encountering ribosomes. In this schematic diagram the ribosomes are depicted as smaller than the RNA polymerase. In reality the ribosomes ($M_r, 2.7 \times 10^6$) are an order of magnitude larger than the RNA polymerase ($M_r, 3.9 \times 10^5$).

Bacterial mRNAs generally exist for just a few minutes (p. 1084) before they are degraded by nucleases. In order to maintain high rates of protein synthesis, the mRNA for a given protein or set of proteins must be made continuously and translated with maximum efficiency. The short lifetime of mRNAs in bacteria allows a rapid cessation of synthesis when the protein is no longer needed.

Stage 5: Newly Synthesized Polypeptide Chains Undergo Folding and Processing

In the final stage of protein synthesis, the nascent polypeptide chain is folded and processed into its biologically active form. During or after its synthesis, the polypeptide progressively assumes its native conformation, with the formation of appropriate hydrogen bonds and van der Waals, ionic, and hydrophobic interactions. In this way the linear, or one-dimensional, genetic message in the

mRNA is converted into the three-dimensional structure of the protein. Some newly made proteins, bacterial, archaeal, and eukaryotic, do not attain their final biologically active conformation until they have been altered by one or more processing reactions called **posttranslational modifications**.

Amino-Terminal and Carboxyl-Terminal Modifications The first residue inserted in all polypeptides is *N*-formylmethionine (in bacteria) or methionine (in eukaryotes). However, the formyl group, the amino-terminal Met residue, and often additional amino-terminal (and, in some cases, carboxyl-terminal) residues may be removed enzymatically in formation of the final functional protein. In as many as 50% of eukaryotic proteins, the amino group of the amino-terminal residue is *N*-acetylated after translation. Carboxyl-terminal residues are also sometimes modified.

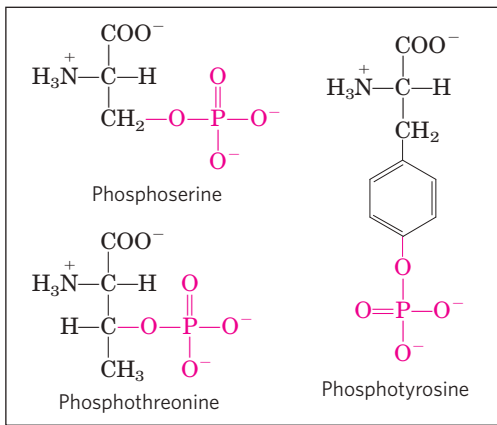
Loss of Signal Sequences As we shall see in Section 27.3, the 15 to 30 residues at the amino-terminal end of some proteins play a role in directing the protein to its ultimate destination in the cell. Such **signal sequences** are eventually removed by specific peptidases.

Modification of Individual Amino Acids The hydroxyl groups of certain Ser, Thr, and Tyr residues of some proteins are enzymatically phosphorylated by ATP (Fig. 27-34a); the phosphate groups add negative charges to these polypeptides. The functional significance of this modification varies from one protein to the next. For example, the milk protein casein has many phosphoserine groups that bind Ca^{2+} . Calcium, phosphate, and amino acids are all valuable to suckling young, so casein efficiently provides three essential nutrients. And as we have seen in numerous instances, phosphorylation-dephosphorylation cycles regulate the activity of many enzymes and regulatory proteins.

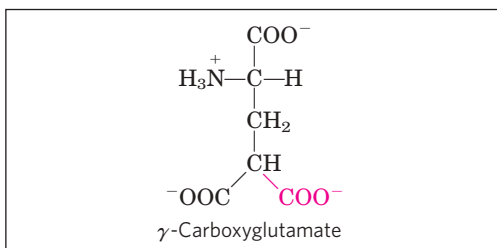
Extra carboxyl groups may be added to Glu residues of some proteins. For example, the blood-clotting protein prothrombin contains a number of γ -carboxyglutamate residues (Fig. 27-34b) in its amino-terminal region, introduced by an enzyme that requires vitamin K. These carboxyl groups bind Ca^{2+} , which is required to initiate the clotting mechanism.

Monomethyl- and dimethyllysine residues (Fig. 27-34c) occur in some muscle proteins and in cytochrome *c*. The calmodulin of most species contains one trimethyllysine residue at a specific position. In other proteins, the carboxyl groups of some Glu residues undergo methylation, removing their negative charge.

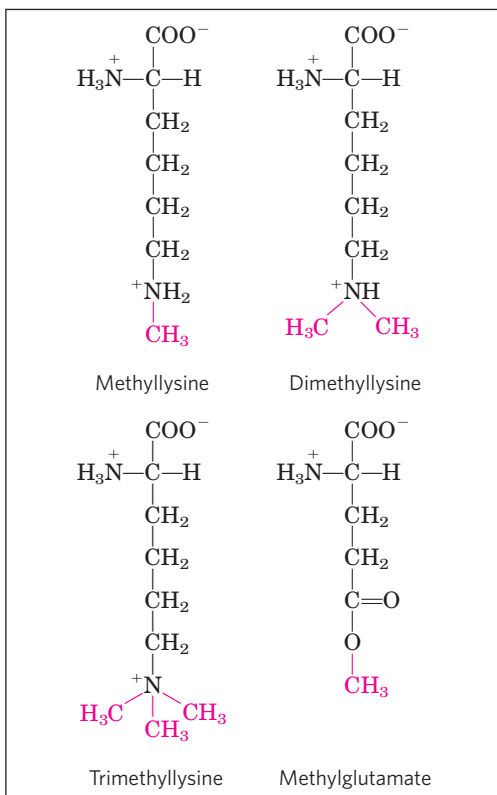
Attachment of Carbohydrate Side Chains The carbohydrate side chains of glycoproteins are attached covalently during or after synthesis of the polypeptide. In some glycoproteins, the carbohydrate side chain is attached enzymatically to Asn residues (*N*-linked oligosaccharides), in others to Ser or Thr residues (*O*-linked oligosaccharides) (see Fig. 7-30). Many proteins that function extracellularly, as well as the lubricating proteoglycans that coat mucous



(a)



(b)



(c)

FIGURE 27-34 Some modified amino acid residues. (a) Phosphorylated amino acids. (b) A carboxylated amino acid. (c) Some methylated amino acids.

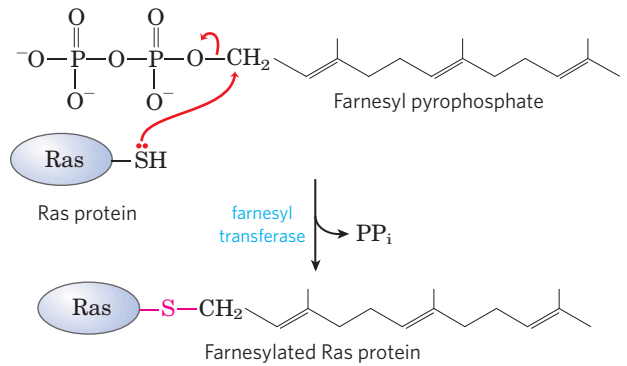


FIGURE 27-35 Farnesylation of a Cys residue. The thioether linkage is shown in red. The Ras protein is the product of the *ras* oncogene.

membranes, contain oligosaccharide side chains (see Fig. 7-28).

Addition of Isoprenyl Groups A number of eukaryotic proteins are modified by the addition of groups derived from isoprene (isoprenyl groups). A thioether bond is formed between the isoprenyl group and a Cys residue of the protein (see Fig. 11-15). The isoprenyl groups are derived from pyrophosphorylated intermediates of the cholesterol biosynthetic pathway (see Fig. 21-35), such as farnesyl pyrophosphate (Fig. 27-35). Proteins modified in this way include the Ras proteins (small G proteins), which are products of the *ras* oncogenes and proto-oncogenes, and the trimeric G proteins (both discussed in Chapter 12), as well as lamins, proteins found in the nuclear matrix. The isoprenyl group helps to anchor the protein in a membrane. The transforming (carcinogenic) activity of the *ras* oncogene is lost when isoprenylation of the Ras protein is blocked, a finding that has stimulated interest in identifying inhibitors of this posttranslational modification pathway for use in cancer chemotherapy.

Addition of Prosthetic Groups Many proteins require for their activity covalently bound prosthetic groups. Two examples are the biotin molecule of acetyl-CoA carboxylase and the heme group of hemoglobin or cytochrome *c*.

Proteolytic Processing Many proteins are initially synthesized as large, inactive precursor polypeptides that are proteolytically trimmed to form their smaller, active forms. Examples include proinsulin, some viral proteins, and proteases such as chymotrypsinogen and trypsinogen (see Fig. 6-38).

Formation of Disulfide Cross-Links After folding into their native conformations, some proteins form intrachain or interchain disulfide bridges between Cys residues. In eukaryotes, disulfide bonds are common in proteins to be exported from cells. The cross-links formed in this way help to protect the native conformation of the protein molecule from denaturation in the extracellular environment, which can differ greatly from intracellular conditions and is generally oxidizing.

Protein Synthesis Is Inhibited by Many Antibiotics and Toxins

Protein synthesis is a central function in cellular physiology and is the primary target of many naturally occurring antibiotics and toxins. Except as noted, these antibiotics inhibit protein synthesis in bacteria. The differences between bacterial and eukaryotic protein synthesis, though in some cases subtle, are sufficient that most of the compounds discussed below are relatively harmless to eukaryotic cells. Natural selection has favored the evolution of compounds that exploit minor differences in order to affect bacterial systems selectively, such that these biochemical weapons are synthesized by some microorganisms and are extremely toxic to others. Because nearly every step in protein synthesis can be specifically inhibited by one antibiotic or another, antibiotics have become valuable tools in the study of protein biosynthesis.

Puromycin, made by the mold *Streptomyces alboniger*, is one of the best-understood inhibitory antibiotics. Its structure is very similar to the 3' end of an aminoacyl-tRNA, enabling it to bind to the ribosomal A site and participate in peptide bond formation, producing peptidylpuromycin (Fig. 27-36). However, because puromycin resembles only the 3' end of the tRNA, it does not engage in translocation and dissociates from the ribosome shortly after it is linked to the carboxyl terminus of the peptide. This prematurely terminates polypeptide synthesis.

Tetracyclines inhibit protein synthesis in bacteria by blocking the A site on the ribosome, preventing the binding of aminoacyl-tRNAs. **Chloramphenicol** inhibits protein synthesis by bacterial (and mitochondrial and chloroplast) ribosomes by blocking peptidyl transfer; it does not affect cytosolic protein synthesis in eukaryotes. Conversely, **cycloheximide** blocks the peptidyl transferase of 80S eukaryotic ribosomes but not that of 70S bacterial (and mitochondrial and chloroplast) ribosomes. **Streptomycin**, a basic trisaccharide, causes misreading of the genetic code (in bacteria) at relatively low concentrations and inhibits initiation at higher concentrations.

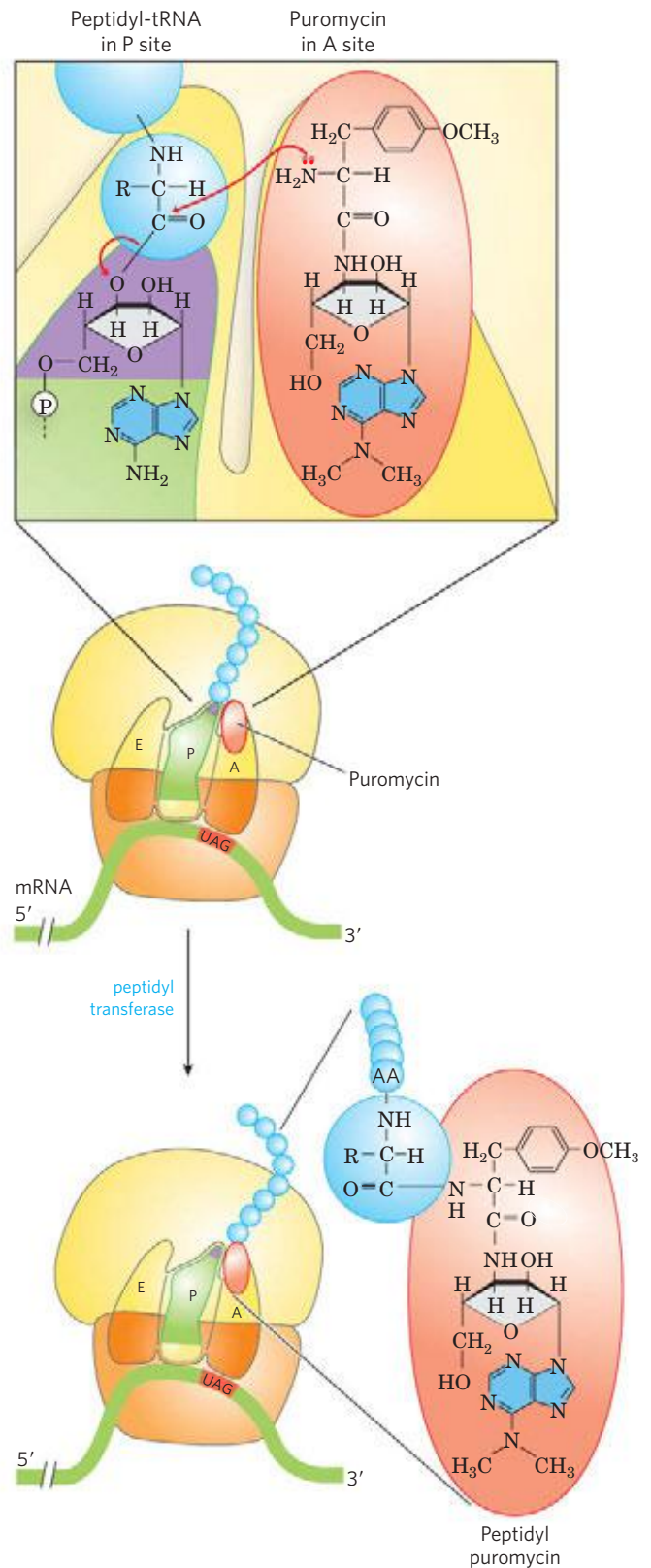
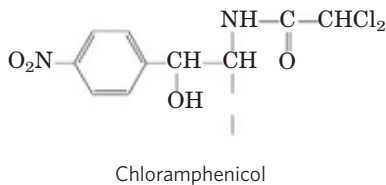
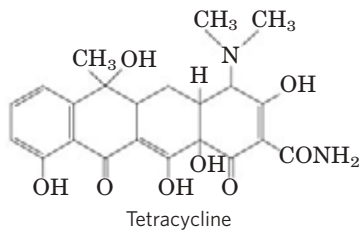
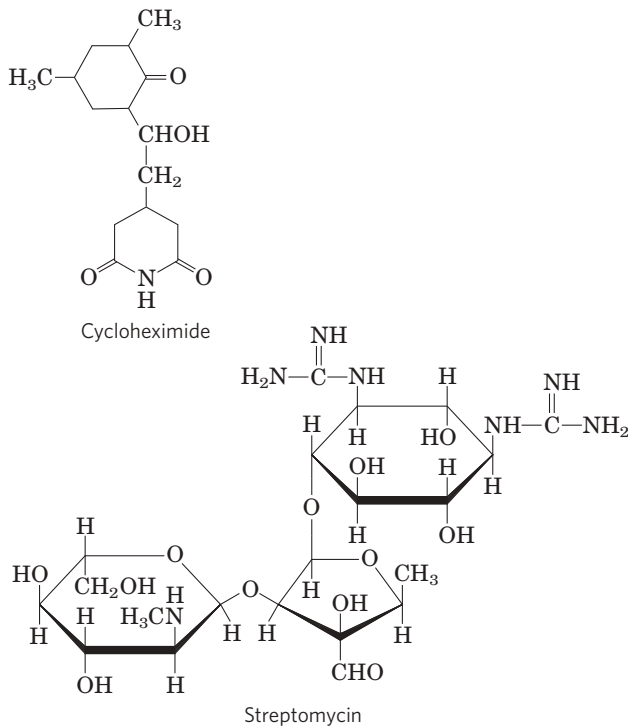


FIGURE 27-36 Disruption of peptide bond formation by puromycin. The antibiotic puromycin resembles the aminoacyl end of a charged tRNA, and it can bind to the ribosomal A site and participate in peptide bond formation. The product of this reaction, peptidyl puromycin, is not translocated to the P site. Instead, it dissociates from the ribosome, causing premature chain termination.



Several other inhibitors of protein synthesis are notable because of their toxicity to humans and other mammals. **Diphtheria toxin** (M_r 58,330) catalyzes the ADP-ribosylation of a diphthamide (a modified histidine) residue of eukaryotic elongation factor eEF2, thereby inactivating it. **Ricin** (M_r 29,895), an extremely toxic protein of the castor bean, inactivates the 60S subunit of eukaryotic ribosomes by depurinating a specific adenosine in 23S rRNA. Ricin was used in the infamous 1978 murder of BBC journalist and Bulgarian dissident Georgi Markov, presumably by the Bulgarian secret police. Using a syringe hidden at the end of an umbrella, a member of the secret police injected Markov in the leg with a ricin-infused pellet. He died 4 days later.

SUMMARY 27.2 Protein Synthesis

- ▶ Protein synthesis occurs on the ribosomes, which consist of protein and rRNA. Bacteria have 70S ribosomes, with a large (50S) and a small (30S) subunit. Eukaryotic ribosomes are significantly larger (80S) and contain more proteins.
- ▶ Transfer RNAs have 73 to 93 nucleotide residues, some of which have modified bases. Each tRNA has an amino acid arm with the terminal sequence CCA(3') to which an amino acid is esterified, an anticodon arm, a T ψ C arm, and a D arm; some tRNAs have a fifth arm. The anticodon is responsible for the specificity of interaction between the aminoacyl-tRNA and the complementary mRNA codon.
- ▶ The growth of polypeptides on ribosomes begins with the amino-terminal amino acid and proceeds by successive additions of new residues to the carboxyl-terminal end.

- ▶ Protein synthesis occurs in five stages.

1. Amino acids are activated by specific aminoacyl-tRNA synthetases in the cytosol. These enzymes catalyze the formation of aminoacyl-tRNAs, with simultaneous cleavage of ATP to AMP and PP_i. The fidelity of protein synthesis depends on the accuracy of this reaction, and some of these enzymes carry out proofreading steps at separate active sites.

2. In bacteria, the initiating aminoacyl-tRNA in all proteins is *N*-formylmethionyl-tRNA^{fMet}. Initiation of protein synthesis involves formation of a complex between the 30S ribosomal subunit, mRNA, GTP, fMet-tRNA^{fMet}, three initiation factors, and the 50S subunit; GTP is hydrolyzed to GDP and P_i.

3. In the elongation steps, GTP and elongation factors are required for binding the incoming aminoacyl-tRNA to the A site on the ribosome. In the first peptidyl transfer reaction, the fMet residue is transferred to the amino group of the incoming aminoacyl-tRNA. Movement of the ribosome along the mRNA then translocates the dipeptidyl-tRNA from the A site to the P site, a process requiring hydrolysis of GTP. Deacylated tRNAs dissociate from the ribosomal E site.

4. After many such elongation cycles, synthesis of the polypeptide is terminated with the aid of release factors. At least four high-energy phosphate equivalents (from ATP and GTP) are required to generate each peptide bond, an energy investment required to guarantee fidelity of translation.

5. Polypeptides fold into their active, three-dimensional forms. Many proteins are further processed by posttranslational modification reactions.

- ▶ Many well-studied antibiotics and toxins inhibit some aspect of protein synthesis.

27.3 Protein Targeting and Degradation

The eukaryotic cell is made up of many structures, compartments, and organelles, each with specific functions that require distinct sets of proteins and enzymes. These proteins (with the exception of those produced in mitochondria and plastids) are synthesized on ribosomes in the cytosol, so how are they directed to their final cellular destinations?

We are now beginning to understand this complex and fascinating process. Proteins destined for secretion, integration in the plasma membrane, or inclusion in lysosomes generally share the first few steps of a pathway that begins in the endoplasmic reticulum. Proteins destined for mitochondria, chloroplasts, or the nucleus use three separate mechanisms. And proteins destined for the cytosol simply remain where they are synthesized.

The most important element in many of these targeting pathways is a short sequence of amino acids called a **signal sequence**, whose function was first postulated by Günter Blobel and colleagues in 1970. The signal sequence directs a protein to its appropriate location in the cell and, for many proteins, is removed during transport or after the protein has reached its final destination. In proteins slated for transport into mitochondria, chloroplasts, or the ER, the signal sequence is at the amino terminus of a newly synthesized polypeptide. In many cases, the targeting capacity of particular signal sequences has been confirmed by fusing the signal sequence from one protein to a second protein and showing that the signal directs the second protein to the location where the first protein is normally found. The selective degradation of proteins no longer needed by the cell also relies largely on a set of molecular signals embedded in each protein's structure.

In this concluding section we examine protein targeting and degradation, emphasizing the underlying signals and molecular regulation that are so crucial to cellular metabolism. Except where noted, the focus is now on eukaryotic cells.

Posttranslational Modification of Many Eukaryotic Proteins Begins in the Endoplasmic Reticulum

Perhaps the best-characterized targeting system begins in the ER. Most lysosomal, membrane, or secreted proteins have an amino-terminal signal sequence (**Fig. 27-37**) that marks them for translocation into the lumen of the ER; hundreds of such signal sequences have been determined. The carboxyl terminus of the signal sequence is defined by a cleavage site, where protease action removes the sequence after the protein is imported into the ER. Signal sequences vary in length from 13 to 36 amino acid residues, but all have the

following features: (1) about 10 to 15 hydrophobic amino acid residues; (2) one or more positively charged residues, usually near the amino terminus, preceding the hydrophobic sequence; and (3) a short sequence at the carboxyl terminus (near the cleavage site) that is relatively polar, typically having amino acid residues with short side chains (especially Ala) at the positions closest to the cleavage site.

As originally demonstrated by George Palade, proteins with these signal sequences are synthesized on ribosomes attached to the ER. The signal sequence itself helps to direct the ribosome to the ER, as illustrated by steps 1 through 8 in **Figure 27-38**. 1 The targeting pathway begins with initiation of protein synthesis on free ribosomes. 2 The signal sequence appears early in the synthetic process, because it is at the amino terminus, which as we have seen is synthesized first. 3 As it emerges from the ribosome, the signal sequence—and the ribosome itself—is bound by the large **signal recognition particle (SRP)**; SRP then binds GTP and halts elongation of the polypeptide when it is about 70 amino acids long and the signal sequence has completely emerged from the ribosome. 4 The GTP-bound SRP now directs the ribosome (still bound to the mRNA) and the incomplete polypeptide to GTP-bound SRP receptors in the cytosolic face of the ER; the nascent polypeptide is delivered to a **peptide translocation complex** in the ER, which may interact directly with the ribosome. 5 SRP dissociates from the ribosome, accompanied by hydrolysis of GTP in both SRP and the SRP receptor. 6 Elongation of the polypeptide now resumes, with the ATP-driven translocation complex feeding the growing polypeptide into the ER lumen until the complete protein has been synthesized. 7 The signal sequence is removed by a signal peptidase within the ER lumen; 8 the ribosome dissociates and 9 is recycled.



Günter Blobel



George Palade

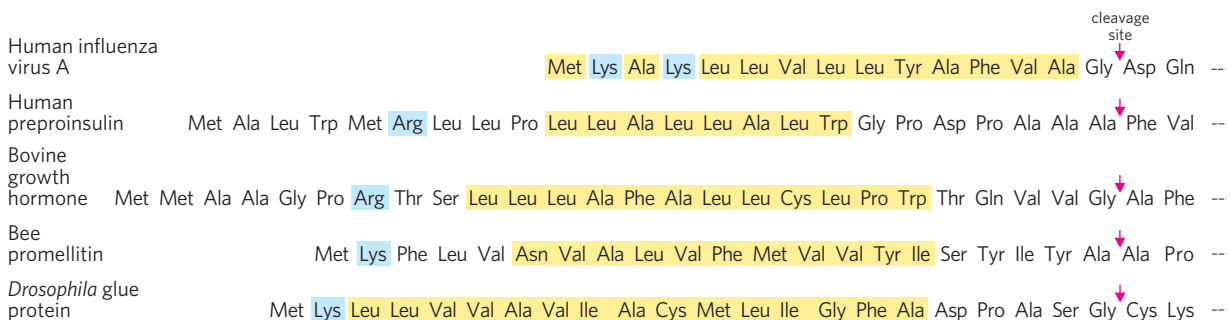


FIGURE 27-37 Amino-terminal signal sequences of some eukaryotic proteins that direct their translocation into the ER. The hydrophobic core (yellow) is preceded by one or more basic residues (blue). Note the polar

and short-side-chain residues immediately preceding (to the left of, as shown here) the cleavage sites (indicated by red arrows).

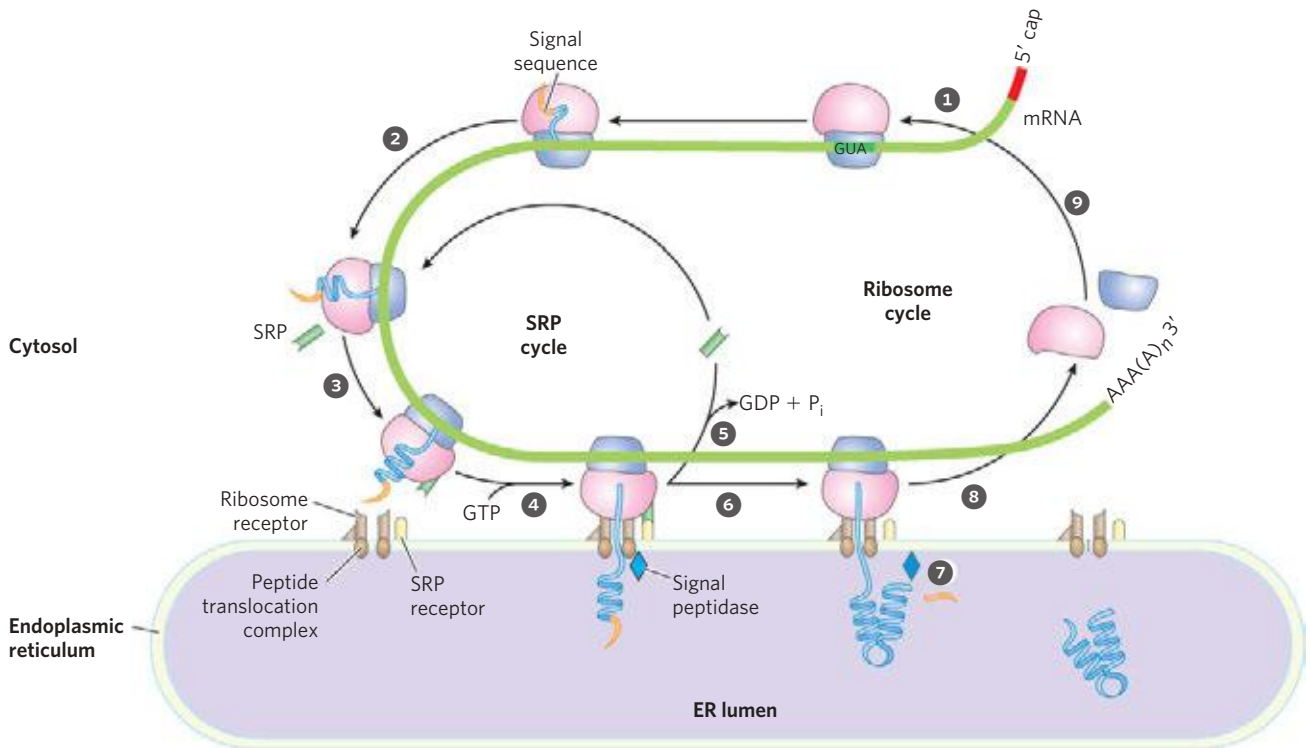


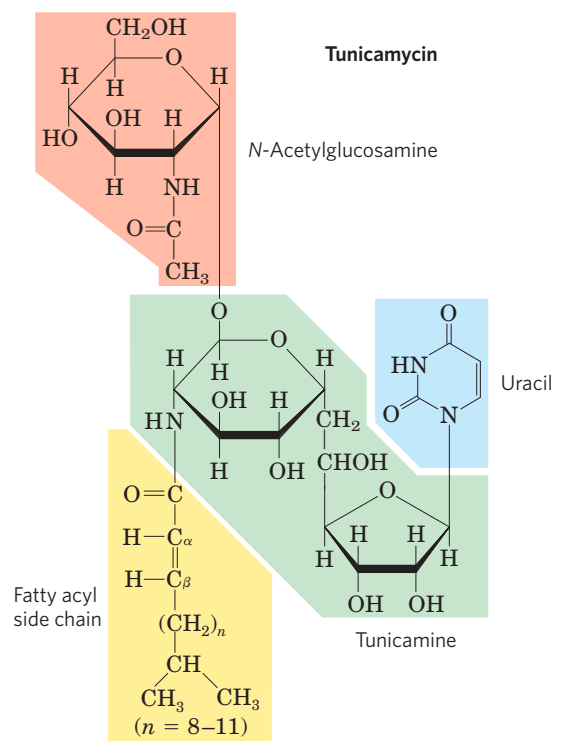
FIGURE 27-38 Directing eukaryotic proteins with the appropriate signals to the endoplasmic reticulum. This process involves the SRP cycle and translocation and cleavage of the nascent polypeptide. The steps are described in the text. SRP is a rod-shaped complex containing a 300 nucleotide RNA (7SL-RNA) and six different proteins (combined M_r 325,000). One protein subunit of SRP binds directly to the signal se-

quence, inhibiting elongation by sterically blocking the entry of aminoacyl-tRNAs and inhibiting peptidyl transferase. Another protein subunit binds and hydrolyzes GTP. The SRP receptor is a heterodimer of α (M_r 69,000) and β (M_r 30,000) subunits, both of which bind and hydrolyze multiple GTP molecules during this process.

Glycosylation Plays a Key Role in Protein Targeting

In the ER lumen, newly synthesized proteins are further modified in several ways. Following the removal of signal sequences, polypeptides are folded, disulfide bonds formed, and many proteins glycosylated to form glycoproteins. In many glycoproteins the linkage to their oligosaccharides is through Asn residues. These *N*-linked oligosaccharides are diverse (Chapter 7), but the pathways by which they form have a common first step. A 14 residue core oligosaccharide is built up in a stepwise fashion, then transferred from a dolichol phosphate donor molecule to certain Asn residues in the protein (Fig. 27-39). The transferase is on the luminal face of the ER and thus cannot catalyze glycosylation of cytosolic proteins. After transfer, the core oligosaccharide is trimmed and elaborated in different ways on different proteins, but all *N*-linked oligosaccharides retain a pentasaccharide core derived from the original 14 residue oligosaccharide. Several antibiotics act by interfering with one or more steps in this process and have aided in elucidating the steps of protein glycosylation. The best characterized is **tunicamycin**, which mimics the structure of UDP-*N*-acetylglucosamine and blocks the first step of the process (Fig. 27-39, step 1). A few proteins are *O*-glycosylated in the ER, but most *O*-glycosylation

occurs in the Golgi complex or in the cytosol (for proteins that do not enter the ER).



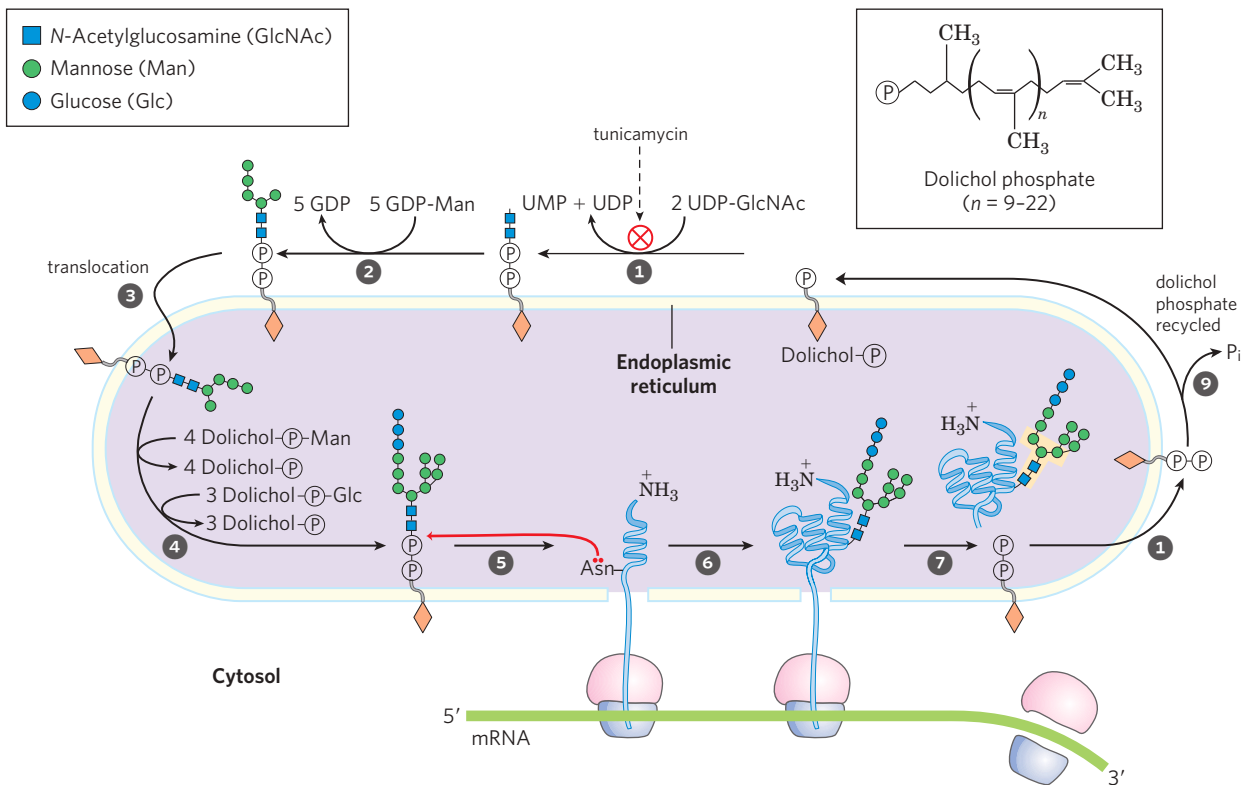


FIGURE 27-39 Synthesis of the core oligosaccharide of glycoproteins.

The core oligosaccharide is built up by the successive addition of monosaccharide units. **1, 2** The first steps occur on the cytosolic face of the ER. **3** Translocation moves the incomplete oligosaccharide across the membrane (mechanism not shown), and **4** completion of the core oligosaccharide occurs within the lumen of the ER. The precursors that contribute additional mannose and glucose residues to the growing oligosaccharide in the lumen are dolichol phosphate derivatives. In the first step in the construction of the *N*-linked oligosaccharide moiety of a glycoprotein, **5**

6 the core oligosaccharide is transferred from dolichol phosphate to an Asn residue of the protein within the ER lumen. The core oligosaccharide is then further modified in the ER and the Golgi complex in pathways that differ for different proteins. The five sugar residues shown surrounded by a beige screen (after step **7**) are retained in the final structure of all *N*-linked oligosaccharides. **8** The released dolichol pyrophosphate is again translocated so that the pyrophosphate is on the cytosolic face of the ER, then **9** a phosphate is hydrolytically removed to regenerate dolichol phosphate.

Suitably modified proteins can now be moved to a variety of intracellular destinations. Proteins travel from the ER to the Golgi complex in transport vesicles (**Fig. 27-40**). In the Golgi complex, oligosaccharides are *O*-linked to some proteins, and *N*-linked oligosaccharides are further modified. By mechanisms not yet fully understood, the Golgi complex also sorts proteins and sends them to their final destinations. The processes that segregate proteins targeted for secretion from those targeted for the plasma membrane or lysosomes must distinguish among these proteins on the basis of structural features other than signal sequences, which were removed in the ER lumen.

This sorting process is best understood in the case of hydrolases destined for transport to lysosomes. On arrival of a hydrolase (a glycoprotein) in the Golgi complex, an as yet undetermined feature (sometimes called a signal patch) of the three-dimensional structure of hydrolase is recognized by a phosphotransferase, which phosphorylates certain mannose residues in oligosaccharide (**Fig. 27-41**). The presence of one or more mannose 6-phosphate residues in its *N*-linked

oligosaccharide is the structural signal that targets a protein to lysosomes. A receptor protein in the membrane of the Golgi complex recognizes the mannose 6-phosphate signal and binds the hydrolase so marked. Vesicles containing these receptor-hydrolase complexes bud from the trans side of the Golgi complex and make their way to sorting vesicles. Here, the receptor-hydrolase complex dissociates in a process facilitated by the lower pH in the vesicle and by phosphatase-catalyzed removal of phosphate groups from the mannose 6-phosphate residues. The receptor is then recycled to the Golgi complex, and vesicles containing the hydrolases bud from the sorting vesicles and move to the lysosomes. In cells treated with tunicamycin (**Fig. 27-39**, step **1**), hydrolases that should be targeted to lysosomes are instead secreted, confirming that the *N*-linked oligosaccharide plays a key role in targeting these enzymes to lysosomes.

The pathways that target proteins to mitochondria and chloroplasts also rely on amino-terminal signal sequences. Although mitochondria and chloroplasts contain DNA, most of their proteins are encoded by

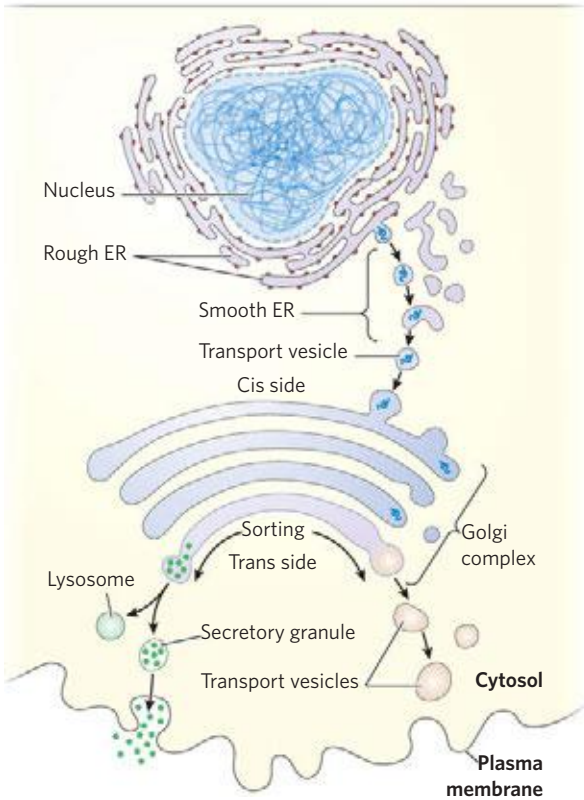


FIGURE 27-40 Pathway taken by proteins destined for lysosomes, the plasma membrane, or secretion. Proteins are moved from the ER to the cis side of the Golgi complex in transport vesicles. Sorting occurs primarily in the trans side of the Golgi complex.

nuclear DNA and must be targeted to the appropriate organelle. Unlike other targeting pathways, however, the mitochondrial and chloroplast pathways begin only *after* a precursor protein has been completely synthesized and released from the ribosome. Precursor proteins destined for mitochondria or chloroplasts are bound by cytosolic chaperone proteins and delivered to receptors on the exterior surface of the target organelle. Specialized translocation mechanisms then transport the protein to its final destination in the organelle, after which the signal sequence is removed.

Signal Sequences for Nuclear Transport Are Not Cleaved

Molecular communication between the nucleus and the cytosol requires the movement of macromolecules through nuclear pores. RNA molecules synthesized in the nucleus are exported to the cytosol. Ribosomal proteins synthesized on cytosolic ribosomes are imported

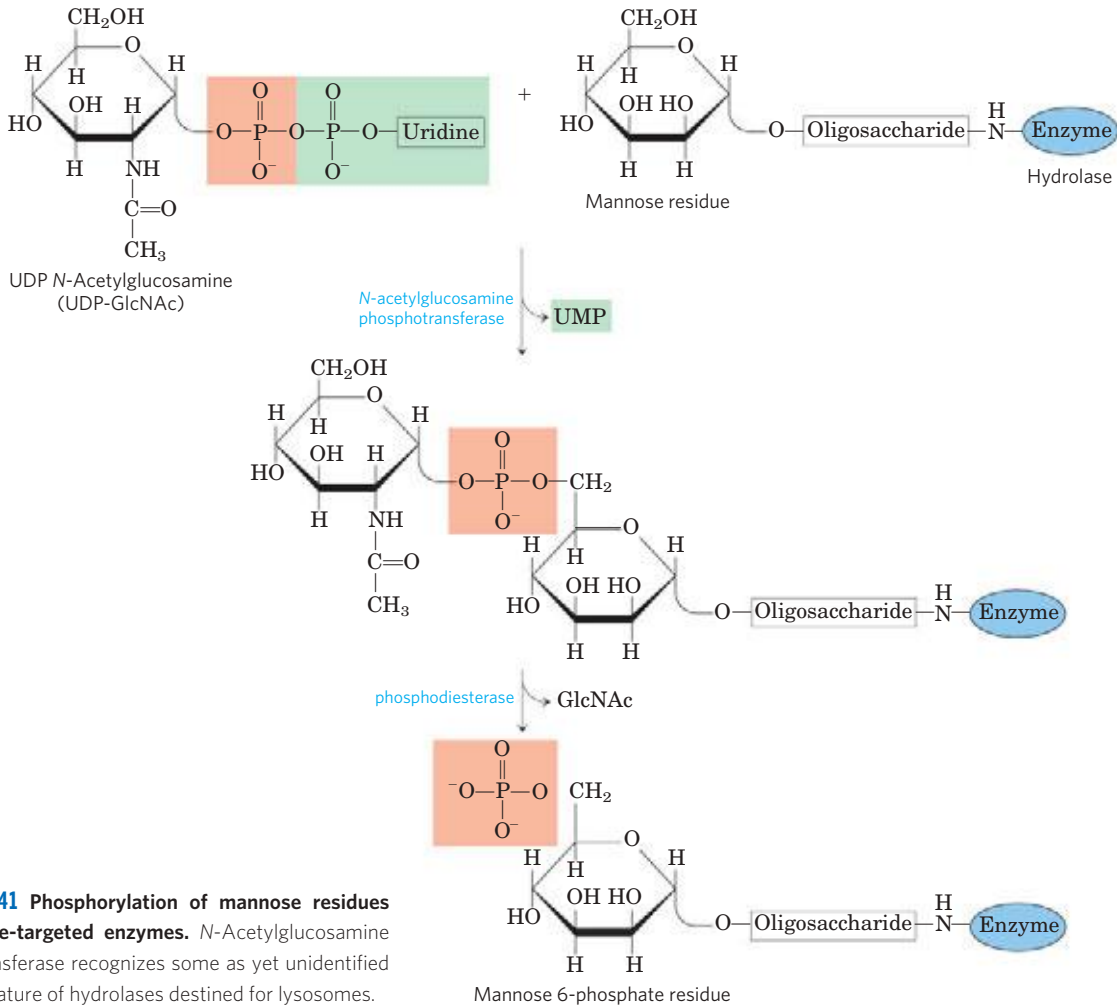


FIGURE 27-41 Phosphorylation of mannose residues on lysosome-targeted enzymes. N-Acetylglucosamine phosphotransferase recognizes some as yet unidentified structural feature of hydrolases destined for lysosomes.

into the nucleus and assembled into 60S and 40S ribosomal subunits in the nucleolus; completed subunits are then exported back to the cytosol. A variety of nuclear proteins (RNA and DNA polymerases, histones, topoisomerases, proteins that regulate gene expression, and so forth) are synthesized in the cytosol and imported into the nucleus. This traffic is modulated by a complex system of molecular signals and transport proteins that is gradually being elucidated.

In most multicellular eukaryotes, the nuclear envelope breaks down at each cell division, and once division is completed and the nuclear envelope reestablished, the dispersed nuclear proteins must be reimported. To allow this repeated nuclear importation, the signal sequence that targets a protein to the nucleus—the **nuclear localization sequence (NLS)**—is not removed after the protein arrives at its destination. An NLS, unlike other signal sequences, may be located almost anywhere along the primary sequence of the protein.

NLSs can vary considerably, but many consist of four to eight amino acid residues and include several consecutive basic (Arg or Lys) residues.

Nuclear importation is mediated by a number of proteins that cycle between the cytosol and the nucleus (**Fig. 27-42**), including importin α and β and a small GTPase known as Ran (*Ras*-related nuclear protein). A heterodimer of importin α and β functions as a soluble receptor for proteins targeted to the nucleus, with the α subunit binding NLS-bearing proteins in the cytosol. The complex of the NLS-bearing protein and the importin docks at a nuclear pore and is translocated through the pore by an energy-dependent mechanism. In the nucleus, the importin β is bound by Ran GTPase, releasing importin β from the imported protein. Importin β is bound by Ran and by CAS (*cellular apoptosis susceptibility protein*) and separated from the NLS-bearing protein. Importin α and β , in their complexes with Ran and CAS, are then exported from the nucleus. Ran

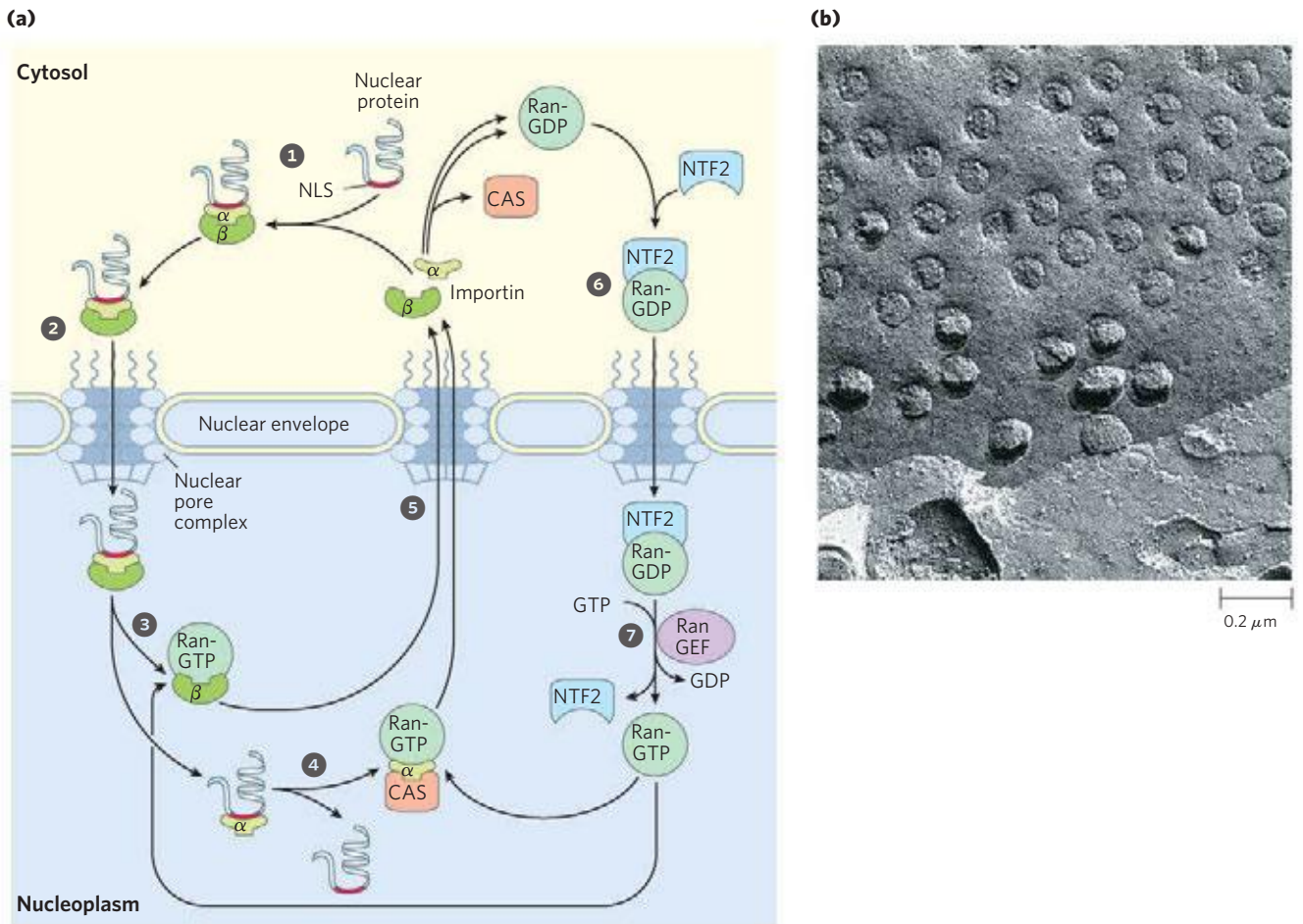


FIGURE 27-42 Targeting of nuclear proteins. (a) 1 A protein with an appropriate nuclear localization signal (NLS) is bound by a complex of importin α and β . 2 The resulting complex binds to a nuclear pore and translocates. 3 Inside the nucleus, dissociation of importin β is promoted by the binding of Ran-GTP. 4 Importin α binds to Ran-GTP and CAS (cellular apoptosis susceptibility protein), releasing the nuclear protein. 5 Importin α and β and CAS are transported out of the nucleus and recycled. They are released in the cytosol when Ran hydrolyzes its bound GTP.

6 Ran-GDP is bound by NTF2, and transported back into the nucleus. 7 RanGEF promotes the exchange of GDP for GTP in the nucleus, and Ran-GTP is ready to process another NLS-bearing protein-importin complex. (b) Scanning electron micrograph of the surface of the nuclear envelope, showing numerous nuclear pores. The nuclear pore complex is one of the largest molecular aggregates in the cell ($M_r \sim 5 \times 10^7$). It is made up of multiple copies of over 30 different proteins.

Inner membrane proteins

Phage fd, major coat protein

Met Lys Lys Ser Leu Val Leu Lys Ala Ser Val Ala Val Ala Thr Leu Val Pro Met Leu Ser Phe Ala Ala Glu --

Phage fd, minor coat protein

Met Lys Lys Leu Leu Phe Ala Ile Pro Leu Val Val Pro Phe Tyr Ser His Ser Ala Glu --

Periplasmic proteins

Alkaline phosphatase

Met Lys Gln Ser Thr Ile Ala Leu Ala Leu Leu Pro Leu Leu Phe Thr Pro Val Thr Lys Ala Arg Thr --

Leucine-specific binding protein

Met Lys Ala Asn Ala Lys Thr Ile Ile Ala Gly Met Ile Ala Leu Ala Ile Ser His Thr Ala Met Ala Asp Asp --

 β -Lactamase of pBR322

Met Ser Ile Gln His Phe Arg Val Ala Leu Ile Pro Phe Phe Ala Ala Phe Cys Leu Pro Val Phe Ala His Pro --

Outer membrane proteins

Lipoprotein

Met Lys Ala Thr Lys Leu Val Leu Gly Ala Val Ile Leu Gly Ser Thr Leu Leu Ala Gly Cys Ser --

LamB

Leu Arg Lys Leu Pro Leu Ala Val Ala Val Ala Ala Gly Val Met Ser Ala Gln Ala Met Ala Val Asp --

OmpA

Met Met Ile Thr Met Lys Lys Thr Ala Ile Ala Ile Ala Val Ala Leu Ala Gly Phe Ala Thr Val Ala Gln Ala Ala Pro --

FIGURE 27-43 Signal sequences that target proteins to different locations in bacteria. Basic amino acids (blue) near the amino terminus and hydrophobic core amino acids (yellow) are highlighted. The cleavage sites marking the ends of the signal sequences are indicated by red arrows.

Note that the inner bacterial cell membrane (see Fig. 1-6) is where phage fd coat proteins and DNA are assembled into phage particles. OmpA is outer membrane protein A; LamB is a cell surface receptor protein for λ phage.

hydrolyzes GTP in the cytosol to release the importins, which are then free to begin another importation cycle. Ran itself is also cycled back into the nucleus by the binding of Ran-GDP to nuclear transport factor 2 (NTF2). Inside the nucleus, the GDP bound to Ran is replaced with GTP through the action of Ran guanine nucleotide-exchange factor (RanGEF; see Box 12-2).

Bacteria Also Use Signal Sequences for Protein Targeting

Bacteria can target proteins to their inner or outer membranes, to the periplasmic space between these membranes, or to the extracellular medium. They use signal sequences at the amino terminus of the proteins

(Fig. 27-43), much like those on eukaryotic proteins targeted to the ER, mitochondria, and chloroplasts.

Most proteins exported from *E. coli* make use of the pathway shown in Figure 27-44. Following translation, a protein to be exported may fold only slowly, the amino-terminal signal sequence impeding the folding. The soluble chaperone protein SecB binds to the protein's signal sequence or other features of its incompletely folded structure. The bound protein is then delivered to SecA, a protein associated with the translocation complex (SecYEG) in the bacterial cell membrane. SecB is released, and SecA inserts itself into the membrane, forcing about 20 amino acid residues of the protein to be exported through the translocation complex. Hydrolysis of an ATP by SecA provides the energy for a conformational change that causes SecA to withdraw from the membrane, releasing the polypeptide. SecA binds another ATP, and the next stretch of 20 amino acid residues is pushed across the membrane through the translocation complex. Steps 4 and 5 are repeated until the entire protein has passed through and is released to the periplasm. The electrochemical potential across the membrane (denoted by + and -) also provides some of the driving force required for protein translocation.

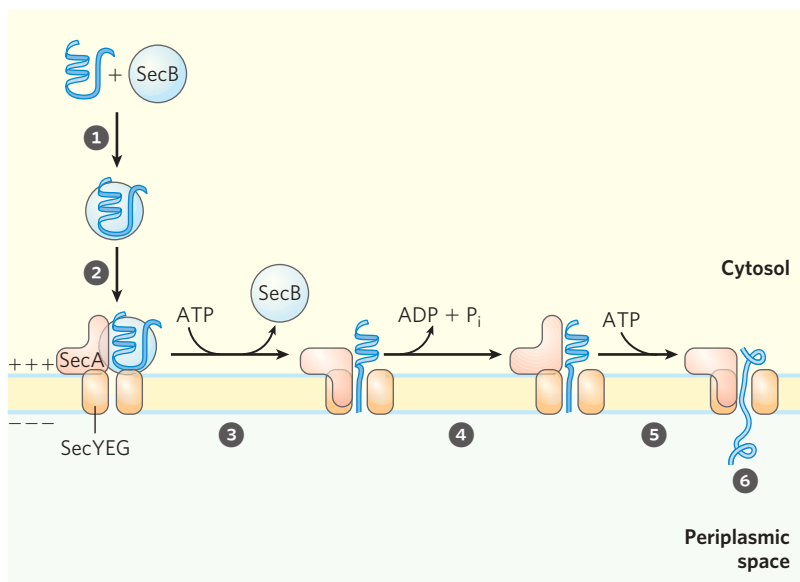


FIGURE 27-44 Model for protein export in bacteria. 1 A newly translated polypeptide binds to the cytosolic chaperone protein SecB, which 2 delivers it to SecA, a protein associated with the translocation complex (SecYEG) in the bacterial cell membrane. 3 SecB is released, and SecA inserts itself into the membrane, forcing about 20 amino acid residues of the protein to be exported through the translocation complex. 4 Hydrolysis of an ATP by SecA provides the energy for a conformational change that causes SecA to withdraw from the membrane, releasing the polypeptide. 5 SecA binds another ATP, and the next stretch of 20 amino acid residues is pushed across the membrane through the translocation complex. Steps 4 and 5 are repeated until 6 the entire protein has passed through and is released to the periplasm. The electrochemical potential across the membrane (denoted by + and -) also provides some of the driving force required for protein translocation.

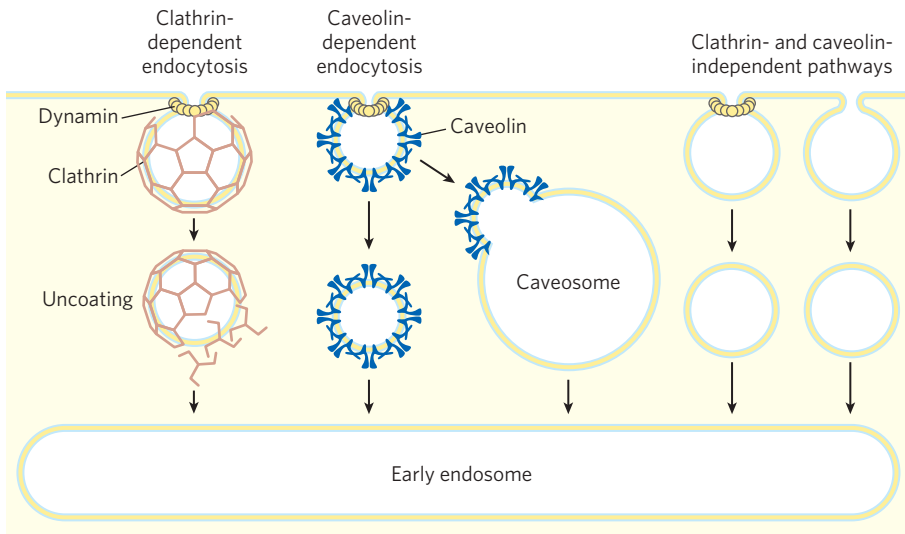


FIGURE 27-45 Summary of endocytosis pathways in eukaryotic cells. Pathways dependent on clathrin or caveolin make use of the GTPase dynamin to pinch vesicles from the plasma membrane. Some pathways do not use clathrin or caveolin; some of these make use of dynamin and some do not.

in lengths of about 20 amino acid residues. Each step is facilitated by the hydrolysis of ATP, catalyzed by SecA.

An exported protein is thus pushed through the membrane by a SecA protein located on the cytoplasmic surface, not pulled through the membrane by a protein on the periplasmic surface. This difference may simply reflect the need for the translocating ATPase to be where the ATP is. The transmembrane electrochemical potential can also provide energy for translocation of the protein, by an as yet unknown mechanism.

Although most exported bacterial proteins use this pathway, some follow an alternative pathway that uses signal recognition and receptor proteins homologous to components of the eukaryotic SRP and SRP receptor (Fig. 27-38).

Cells Import Proteins by Receptor-Mediated Endocytosis

Some proteins are imported into eukaryotic cells from the surrounding medium; examples include low-density lipoprotein (LDL), the iron-carrying protein transferrin, peptide hormones, and circulating proteins destined for degradation. There are several importation pathways (Fig. 27-45). In one path, proteins bind to receptors in invaginations of the membrane called **coated pits**, which concentrate endocytic receptors in preference to other cell-surface proteins. The pits are coated on their cytosolic side with a lattice of the protein **clathrin**, which forms closed polyhedral structures (Fig. 27-46). The clathrin lattice grows as more receptors are occupied by target proteins. Eventually, a complete membrane-bounded endocytic vesicle is pinched off the plasma membrane with the aid of the large GTPase **dynamin**, and enters the cytoplasm. The clathrin is quickly removed by uncoating enzymes, and the vesicle fuses with an endosome. ATPase activity in the endosomal membranes reduces the pH therein, facilitating dissociation of recep-

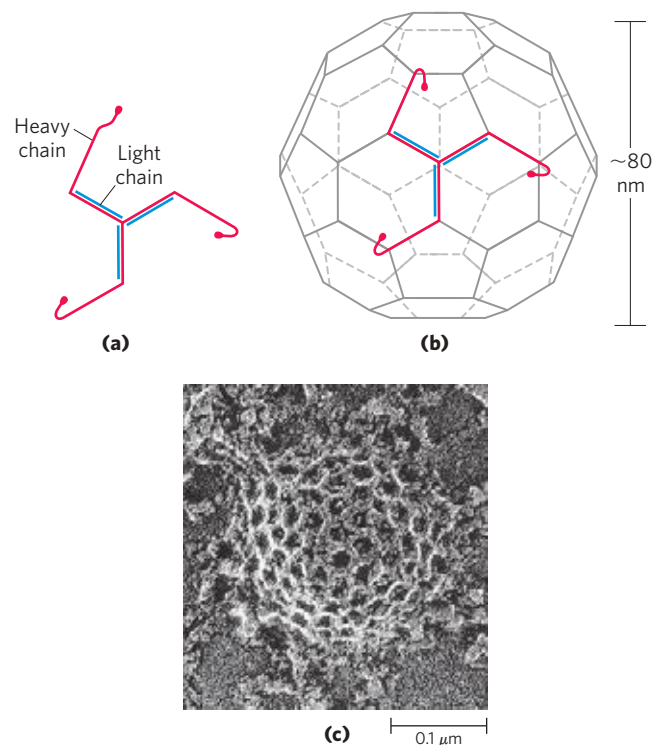


FIGURE 27-46 Clathrin. (a) Three light (L) chains (M_r 35,000) and three heavy (H) chains (M_r 180,000) of the (HL)₃ clathrin unit, organized as a three-legged structure called a triskelion. (b) Triskelions tend to assemble into polyhedral lattices. (c) Electron micrograph of a coated pit on the cytosolic face of the plasma membrane of a fibroblast.

tors from their target proteins. In a related pathway, caveolin causes invagination of patches of membrane containing lipid rafts associated with certain types of receptors (see Fig. 11-22). These endocytic vesicles then fuse with caveolin-containing internal structures, caveosomes, where the internalized molecules are sorted and redirected to other parts of the cell and the caveolins are prepared for recycling to the membrane surface. There

are also clathrin- and caveolin-independent pathways; some make use of dynamin and others do not.

The imported proteins and receptors then go their separate ways, their fates varying with the cell and protein type. Transferrin and its receptor are eventually recycled. Some hormones, growth factors, and immune complexes, after eliciting the appropriate cellular response, are degraded along with their receptors. LDL is degraded after the associated cholesterol has been delivered to its destination, but the LDL receptor is recycled (see Fig. 21–41).

Receptor-mediated endocytosis is exploited by some toxins and viruses to gain entry to cells. Influenza virus, diphtheria toxin, and cholera toxin all enter cells in this way.

Protein Degradation Is Mediated by Specialized Systems in All Cells

Protein degradation prevents the buildup of abnormal or unwanted proteins and permits the recycling of amino acids. The half-lives of eukaryotic proteins vary from 30 seconds to many days. Most proteins turn over rapidly relative to the lifetime of a cell, although a few (such as hemoglobin) can last for the life of the cell (about 110 days for an erythrocyte). Rapidly degraded proteins include those that are defective because of incorrectly inserted amino acids or because of damage accumulated during normal functioning. And enzymes that act at key regulatory points in metabolic pathways often turn over rapidly.

Defective proteins and those with characteristically short half-lives are generally degraded in both bacterial and eukaryotic cells by selective ATP-dependent cytosolic systems. A second system in vertebrates, operating in lysosomes, recycles the amino acids of membrane proteins, extracellular proteins, and proteins with characteristically long half-lives.

In *E. coli*, many proteins are degraded by an ATP-dependent protease called Lon (the name refers to the “long form” of proteins, observed only when this protease is absent). The protease is activated in the presence of defective proteins or those slated for rapid turnover; two ATP molecules are hydrolyzed for every peptide bond cleaved. The precise role of this ATP hydrolysis is not yet clear. Once a protein has been reduced to small inactive peptides, other ATP-independent proteases complete the degradation process.

The ATP-dependent pathway in eukaryotic cells is quite different, involving the protein **ubiquitin**, which, as its name suggests, occurs throughout the eukaryotic kingdoms. One of the most highly conserved proteins known, ubiquitin (76 amino acid residues) is essentially identical in organisms as different as yeasts and humans. Ubiquitin is covalently linked to proteins slated for destruction via an ATP-dependent pathway involving three separate types of enzymes, called E1 activating enzymes, E2 conjugating enzymes, and E3 ligases (**Fig. 27–47**).

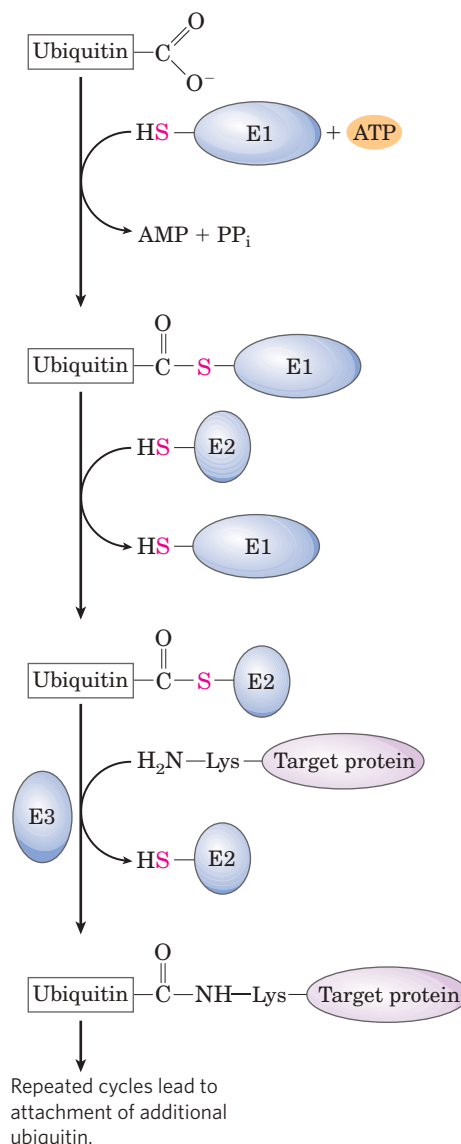


FIGURE 27–47 Three-step pathway by which ubiquitin is attached to a protein. Two different enzyme-ubiquitin intermediates are involved. The free carboxyl group of ubiquitin’s carboxyl-terminal Gly residue first becomes linked to an E1-class activating enzyme via a thioester. The ubiquitin is then transferred to an E2 conjugating enzyme. An E3 ligase ultimately catalyzes the transfer of the ubiquitin from E2 to the target, linking ubiquitin through an amide (isopeptide) bond to an ϵ -amino group of a Lys residue of the target protein. Additional cycles produce polyubiquitin, a covalent polymer of ubiquitin subunits that targets the attached protein for destruction in eukaryotes. Multiple pathways of this sort, with different protein targets, are present in most eukaryotic cells.

Ubiquitinated proteins are degraded by a large complex known as the **26S proteasome** (M_r 2.5×10^6) (**Fig. 27–48**). The eukaryotic proteasome consists of two copies each of at least 32 different subunits, most of which are highly conserved from yeasts to humans. The proteasome contains two main types of subcomplexes, a barrel-like core particle and regulatory particles on either end of the barrel. The 20S core particle consists of four rings; the outer rings are formed from seven α

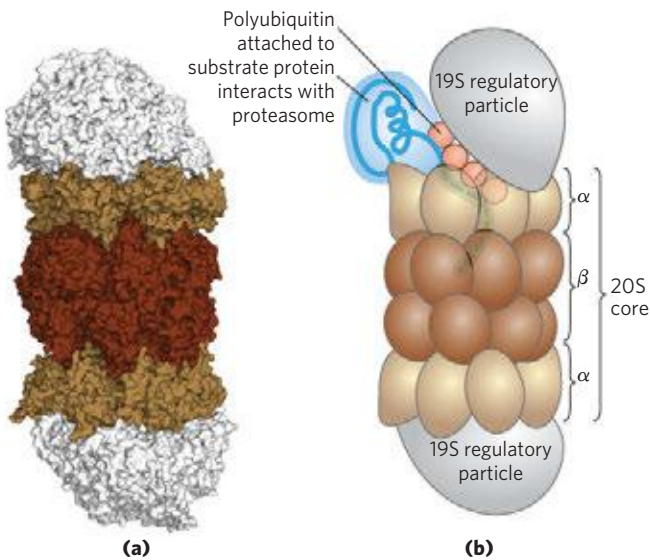


FIGURE 27-48 Three-dimensional structure of the eukaryotic proteasome. The 26S proteasome is highly conserved in all eukaryotes. The two subassemblies are the 20S core particle and the 19S regulatory particle, or cap. **(a)** (PDB ID 3L5Q) The core particle consists of four rings arranged to form a barrel-like structure. Each of the inner rings has seven different β subunits (dark brown), three of which have protease activities. The outer rings each have seven different α subunits (light brown). A regulatory particle forms a cap on each end of the core particle (gray). **(b)** The regulatory particle binds ubiquitinated proteins, unfolds them, and translocates them into the core particle, where they are degraded to peptides of 3 to 25 amino acid residues.

subunits, and the inner rings from seven β subunits. Three of the seven subunits in each β ring have protease activities, each with different substrate specificities. The stacked rings of the core particle form the barrel-like structure within which target proteins are degraded. The 19S regulatory particle on each end of the core particle contains approximately 18 subunits, including some that recognize and bind to ubiquitinated proteins. Six of the subunits are AAA+ ATPases (see Chapter 25) that probably function in unfolding the ubiquitinated proteins and translocating the unfolded polypeptide into the core particle for degradation. The 19S particle also deubiquitinates the proteins as they are degraded in the proteasome. Most cells have additional regulatory complexes that can replace the 19S particle. These alternative regulators do not hydrolyze ATP and do not bind to ubiquitin, but they are important for the degradation of particular cellular proteins. The 26S proteasome can be effectively “accessorized,” with regulatory complexes changing with changing cellular conditions.

Although we do not yet understand all the signals that trigger ubiquitination, one simple signal has been found. For many proteins, the identity of the first residue that remains after removal of the amino-terminal Met residue, and any other posttranslational proteolytic processing of the amino-terminal end, has a profound

influence on half-life (Table 27-9). These amino-terminal signals have been conserved over billions of years of evolution and are the same in bacterial protein degradation systems and in the human ubiquitination pathway. More complex signals, such as the destruction box discussed in Chapter 12 (see Fig. 12-47), are also being identified.

Ubiquitin-dependent proteolysis is as important for the regulation of cellular processes as for the elimination of defective proteins. Many proteins required at only one stage of the eukaryotic cell cycle are rapidly degraded by the ubiquitin-dependent pathway after completing their function. Ubiquitin-dependent destruction of cyclin is critical to cell-cycle regulation (see Fig. 12-47). The E1, E2, and E3 components of the ubiquitination pathway (Fig. 27-47) are in fact large families of proteins. Different E1, E2, and E3 enzymes exhibit different specificities for target proteins and thus regulate different cellular processes. Some of these enzymes are highly localized in certain cellular compartments, reflecting a specialized function.



Not surprisingly, defects in the ubiquitination pathway have been implicated in a wide range of disease states. An inability to degrade certain proteins that activate cell division (the products of oncogenes) can lead to tumor formation, whereas a too-rapid degradation of proteins that act as tumor suppressors can have the same effect. The ineffective or overly rapid degradation of cellular proteins also appears to play a role in a range of other conditions: renal diseases, asthma, neurodegenerative disorders such as Alzheimer and Parkinson diseases (associated with the formation of characteristic proteinaceous structures in neurons), cystic fibrosis (caused in some cases by a too-rapid degradation of a chloride ion channel, with resultant loss of function; see Box 11-2), Liddle syndrome (in which a

TABLE 27-9 Relationship between Protein Half-Life and Amino-Terminal Amino Acid Residue

Amino-terminal residue	Half-life*
Stabilizing	
Ala, Gly, Met, Ser, Thr, Val	>20 h
Destabilizing	
Gln, Ile	~30 min
Glu, Tyr	~10 min
Pro	~7 min
Asp, Leu, Lys, Phe	~3 min
Arg	~2 min

Source: Modified from Bachmair, A., Finley, D., & Varshavsky, A. (1986) In vivo half-life of a protein is a function of its amino-terminal residue. *Science* 234, 179-186.

*Half-lives were measured in yeast for the β -galactosidase protein modified so that in each experiment it had a different amino-terminal residue. Half-lives may vary for different proteins and in different organisms, but this general pattern appears to hold for all organisms.

sodium channel in the kidney is not degraded, leading to excessive Na⁺ absorption and early-onset hypertension—and many other disorders. Drugs designed to inhibit proteasome function are being developed as potential treatments for some of these conditions. In a changing metabolic environment, protein degradation is as important to a cell's survival as is protein synthesis, and much remains to be learned about these interesting pathways. ■

SUMMARY 27.3 Protein Targeting and Degradation

- ▶ After synthesis, many proteins are directed to particular locations in the cell. One targeting mechanism involves a peptide signal sequence, generally found at the amino terminus of a newly synthesized protein.
- ▶ In eukaryotic cells, one class of signal sequences is recognized by the signal recognition particle (SRP), which binds the signal sequence as soon as it appears on the ribosome and transfers the entire ribosome and incomplete polypeptide to the ER. Polypeptides with these signal sequences are moved into the ER lumen as they are synthesized; once in the lumen they may be modified and moved to the Golgi complex, then sorted and sent to lysosomes, the plasma membrane, or transport vesicles.
- ▶ Proteins targeted to mitochondria and chloroplasts in eukaryotic cells, and those destined for export in bacteria, also make use of an amino-terminal signal sequence.
- ▶ Proteins targeted to the nucleus have an internal signal sequence that, unlike other signal sequences, is not cleaved once the protein is successfully targeted.
- ▶ Some eukaryotic cells import proteins by receptor-mediated endocytosis.
- ▶ All cells eventually degrade proteins, using specialized proteolytic systems. Defective proteins and those slated for rapid turnover are generally degraded by an ATP-dependent system. In eukaryotic cells, the proteins are first tagged by linkage to ubiquitin, a highly conserved protein. Ubiquitin-dependent proteolysis is carried out by proteasomes, also highly conserved, and is critical to the regulation of many cellular processes.

Key Terms

Terms in bold are defined in the glossary.

aminoacyl-tRNA 1104	codon 1105
aminoacyl-tRNA synthetases 1104	reading frame 1105
translation 1104	initiation codon 1107
	termination codons 1107

open reading frame (ORF) 1107	polysome 1135
anticodon 1109	posttranslational modification 1136
wobble 1110	puromycin 1138
translational frameshifting 1111	tetracycline 1138
RNA editing 1111	chloramphenicol 1138
initiation 1127	cycloheximide 1138
Shine-Dalgarno sequence 1127	streptomycin 1138
aminoacyl (A) site 1128	diphtheria toxin 1139
peptidyl (P) site 1128	ricin 1139
exit (E) site 1128	signal sequence 1140
initiation complex 1128	signal recognition particle (SRP) 1140
elongation 1129	peptide translocation complex 1140
elongation factors 1129	tunicamycin 1141
peptidyl transferase 1132	nuclear localization sequence (NLS) 1144
translocation 1132	coated pits 1146
termination 1134	clathrin 1146
release factors 1134	dynamin 1146
nonsense suppressor 1134	ubiquitin 1147
	proteasome 1147

Further Reading

Genetic Code

Ambrogelly, A., Palioura, S., & Söll, D. (2007) Natural expansion of the genetic code. *Nat. Chem. Biol.* **3**, 29–35.

Blanc, V. & Davidson, N.O. (2003) C-to-U RNA editing: mechanisms leading to genetic diversity. *J. Biol. Chem.* **278**, 1395–1398.

Crick, F.H.C. (1966) The genetic code: III. *Sci. Am.* **215** (October), 55–62.

An insightful overview of the genetic code at a time when the code words had just been worked out.

Farajollahi, S. & Maas, S. (2010) Molecular diversity through RNA editing: a balancing act. *Trends Genet.* **26**, 221–230.

Hohn, M.J., Park, H.S., O'Donoghue, P., Schnitzbauer, M., & Söll, D. (2006) Emergence of the universal genetic code imprinted in an RNA record. *Proc. Natl. Acad. Sci. USA* **103**, 18,095–18,100.

Levanon, K., Eisenberg E., Rechavi G., & Levanon, E.Y. (2005) Letter from the editor: adenosine-to-inosine RNA editing in Alu repeats in the human genome. *EMBO Rep.* **6**, 831–835.

Liu, M. & Schatz, D.G. (2009) Balancing AID and DNA repair during somatic hypermutation. *Trends Immunol.* **30**, 173–181.

Lobanov, A.V., Turanov, A.A., Hatfield, D.L., & Gladyshev, V.N. (2010) Dual functions of codons in the genetic code. *Crit. Rev. Biochem. Mol. Biol.* **45**, 257–265.

Neeman, Y., Dahary, D., & Nishikura, K. (2006) Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.* **7**, 919–931.

Nirenberg, M. (2004) Historical review: deciphering the genetic code—a personal account. *Trends Biochem. Sci.* **29**, 46–54.

Schimmel, P. & Beebe, K. (2004) Molecular biology—genetic code seizes pyrrolysine. *Nature* **431**, 257–258.

Vetsigian, K., Woese, C., & Goldenfeld, N. (2006) Collective evolution and the genetic code. *Proc. Natl. Acad. Sci. USA* **103**, 10,696–10,701.

Yanofsky, C. (2007) Establishing the triplet nature of the genetic code. *Cell* **128**, 815–818.

Yarus, M., Caporaso, J.G., & Knight, R. (2005) Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.* **74**, 179–198.

Protein Synthesis

Ban, N., Nissen, P., Hansen, J., Moore, P.B., & Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920.

The first high-resolution structure of a major ribosomal subunit.

Ben-Shem, A., de Loubresse, N.G., Melnikov, S., Jenner, L., Yusupova, G., & Yusupov, M. (2011) The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529.

Chapeville, F., Lipmann, F., von Ehrenstein, G., Weisblum, B., Ray, W.J., Jr., & Benzer, S. (1962) On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. USA* **48**, 1086–1092.

Classic experiments providing proof for Crick's adaptor hypothesis and showing that amino acids are not checked after they are linked to tRNAs.

Decatur, W.A. & Fournier, M.J. (2002) rRNA modifications and ribosome function. *Trends Biochem. Sci.* **27**, 344–351.

Dintzis, H.M. (1961) Assembly of the peptide chains of hemoglobin. *Proc. Natl. Acad. Sci. USA* **47**, 247–261.

A classic experiment establishing that proteins are assembled beginning at the amino terminus.

Dunkle, J.A. & Cate, J.H.D. (2010) Ribosome structure and dynamics during translocation and termination. *Annu. Rev. Biophys.* **39**, 227–244.

Gray, N.K. & Wickens, M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.* **14**, 399–458.

Ibba, M. & Söll, D. (2000) Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69**, 617–650.

Korostelev, A. & Noller, H.F. (2007) The ribosome in focus: new structures bring new insights. *Trends Biochem. Sci.* **32**, 434–441.

Korostelev, A., Trakhanov, S., Laurberg, M., & Noller, H.F. (2006) Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**, 1065–1077.

Liu, C.C. & Schultz, P.G. (2010) Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79**, 413–444.

Poehlsgaard, J. & Douthwaite, S. (2005) The bacterial ribosome as a target for antibiotics. *Nat. Rev. Microbiol.* **3**, 870–881.

Rodnina, M.V. & Wintermeyer, W. (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu. Rev. Biochem.* **70**, 415–435.

Woese, C.R., Olsen, G.J., Ibba, M., & Söll, D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236.

Protein Targeting and Degradation

Bedford, L., Lowe, J., Dick, L.R., Mayer, R.J., & Brownell, J.E. (2011) Ubiquitin-like protein conjugation and the ubiquitin-proteasome system as drug targets. *Nat. Rev. Drug Discov.* **10**, 29–46.

DeMartino, G.N. & Gillette, T.G. (2007) Proteasomes: machines for all reasons. *Cell* **129**, 659–662.

Ferguson, S.M. & De Camilli, P. (2012) Dynamin, a membrane-remodelling GTPase. *Nat. Rev. Mol. Cell Biol.* **13**, 75–88.

Hartmann-Petersen, R., Seeger, M., & Gordon C. (2003) Transferring substrates to the 26S proteasome. *Trends Biochem. Sci.* **28**, 26–31.

Komander, D. & Rape, M. (2012) The ubiquitin code. *Annu. Rev. Biochem.* **81**, 203–229.

Liu, C.W., Li, X.H., Thompson, D., Wooding, K., Chang, T., Tang, Z., Yu, H., Thomas, P.J., & DeMartino, G.N. (2006) ATP binding and ATP hydrolysis play distinct roles in the function of 26S proteasome. *Mol. Cell* **24**, 39–50.

Luzio, J.P., Pryor, P.R., & Bright, N.A. (2007) Lysosomes: fusion and function. *Nat. Rev. Mol. Cell Biol.* **8**, 622–632.

Mayor, S. & Pagano, R.E. (2007) Pathways of clathrin-independent endocytosis. *Nat. Rev. Mol. Cell Biol.* **8**, 603–612.

McMahon, H.T. & Boucrot, E. (2011) Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nat. Rev. Mol. Cell Biol.* **12**, 517–533.

Pickart, C.M. & Cohen, R.E. (2004) Proteasomes and their kin: proteases in the machine age. *Nat. Rev. Mol. Cell Biol.* **5**, 177–187.

Royle, S.J. (2006) The cellular functions of clathrin. *Cell. Mol. Life Sci.* **63**, 1823–1832.

Schekman, R. (2007) How sterols regulate protein sorting and traffic. *Proc. Natl. Acad. Sci. USA* **104**, 6496–6497.

Stewart, M. (2007) Molecular mechanism of the nuclear protein import cycle. *Nat. Rev. Mol. Cell Biol.* **8**, 195–208.

Strambio-De-Castillia, C., Niepel, M., & Rout, M.P. (2010) The nuclear pore complex: bridging nuclear transport and gene regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 490–501.

Tasaki, T., Sriram, S.M., Park, K.S., & Kwon, Y.T. (2012) The N-end rule pathway. *Annu. Rev. Biochem.* **81**, 261–289.

Problems

1. Messenger RNA Translation Predict the amino acid sequences of peptides formed by ribosomes in response to the following mRNA sequences, assuming that the reading frame begins with the first three bases in each sequence.

- GGUCAGUCGCUCCUGAAU
- UUGGAUGCGCCAAUAAUUUGCU
- CAUGAUGCCUGUUGCUAC
- AUGGACGAA

2. How Many Different mRNA Sequences Can Specify One Amino Acid Sequence? Write all the possible mRNA sequences that can code for the simple tripeptide segment Leu–Met–Tyr. Your answer will give you some idea about the number of possible mRNAs that can code for one polypeptide.

3. Can the Base Sequence of an mRNA Be Predicted from the Amino Acid Sequence of Its Polypeptide Product? A given sequence of bases in an mRNA will code for one and only one sequence of amino acids in a polypeptide, if the reading frame is specified. From a given sequence of amino acid residues in a protein such as cytochrome *c*, can we predict the base sequence of the unique mRNA that coded it? Give reasons for your answer.

4. Coding of a Polypeptide by Duplex DNA The template strand of a segment of double-helical DNA contains the sequence

(5')CTTAACACCCCTGACTTTCGCGCCGTCG(3')

(a) What is the base sequence of the mRNA that can be transcribed from this strand?

(b) What amino acid sequence could be coded by the mRNA in (a), starting from the 5' end?

(c) If the complementary (nontemplate) strand of this DNA were transcribed and translated, would the resulting amino acid sequence be the same as in (b)? Explain the biological significance of your answer.

5. Methionine Has Only One Codon Methionine is one of two amino acids with only one codon. How does the single codon for methionine specify both the initiating residue and interior Met residues of polypeptides synthesized by *E. coli*?

6. Synthetic mRNAs The genetic code was elucidated with polyribonucleotides synthesized either enzymatically or chemically in the laboratory. Given what we now know about the genetic code, how would you make a polyribonucleotide that could serve as an mRNA coding predominantly for many Phe residues and a small number of Leu and Ser residues? What other amino acid(s) would be coded for by this polyribonucleotide, but in smaller amounts?

7. Energy Cost of Protein Biosynthesis Determine the minimum energy cost, in terms of ATP equivalents expended, required for the biosynthesis of the β -globin chain of hemoglobin (146 residues), starting from a pool including all necessary amino acids, ATP, and GTP. Compare your answer with the direct energy cost of the biosynthesis of a linear glycogen chain of 146 glucose residues in (α 1 \rightarrow 4) linkage, starting from a pool including glucose, UTP, and ATP (Chapter 15). From your data, what is the *extra* energy cost of making a protein, in which all the residues are ordered in a specific sequence, compared with the cost of making a polysaccharide containing the same number of residues but lacking the informational content of the protein?

In addition to the direct energy cost for the synthesis of a protein, there are indirect energy costs—those required for the cell to make the necessary enzymes for protein synthesis. Compare the magnitude of the indirect costs to a eukaryotic cell of the biosynthesis of linear (α 1 \rightarrow 4) glycogen chains and the biosynthesis of polypeptides, in terms of the enzymatic machinery involved.

8. Predicting Anticodons from Codons Most amino acids have more than one codon and attach to more than one tRNA, each with a different anticodon. Write all possible anticodons for the four codons of glycine: (5')GGU, GGC, GGA, and GGG.

(a) From your answer, which of the positions in the anticodons are primary determinants of their codon specificity in the case of glycine?

(b) Which of these anticodon-codon pairings has/have a wobbly base pair?

(c) In which of the anticodon-codon pairings do all three positions exhibit strong Watson-Crick hydrogen bonding?

9. Effect of Single-Base Changes on Amino Acid Sequence Much important confirmatory evidence on the genetic code has come from assessing changes in the amino acid sequence of mutant proteins after a single base has been changed in the gene that encodes the protein. Which of the following amino acid replacements would be consistent with the genetic code if the replacements were caused by a single base change? Which cannot be the result of a single-base mutation? Why?

(a) Phe \rightarrow Leu

(b) Lys \rightarrow Ala

(c) Ala \rightarrow Thr

(d) Phe \rightarrow Lys

(e) Ile \rightarrow Leu

(f) His \rightarrow Glu

(g) Pro \rightarrow Ser

10. Resistance of the Genetic Code to Mutation The following RNA sequence represents the beginning of an open reading frame. What changes (if any) can occur at each position without generating a change in the encoded amino acid residue?

(5')AUGAUUUGCUAUCUUGGACU

11. Basis of the Sickle-Cell Mutation Sickle-cell hemoglobin has a Val residue at position 6 of the β -globin chain instead of the Glu residue found in normal hemoglobin A. Can you predict what change took place in the DNA codon for glutamate to account for replacement of the Glu residue by Val?

12. Proofreading by Aminoacyl-tRNA Synthetases The isoleucyl-tRNA synthetase has a proofreading function that ensures the fidelity of the aminoacylation reaction, but the histidyl-tRNA synthetase lacks such a proofreading function. Explain.

13. Importance of the “Second Genetic Code” Some aminoacyl-tRNA synthetases do not recognize and bind the anticodon of their cognate tRNAs but instead use other structural features of the tRNAs to impart binding specificity. The tRNAs for alanine apparently fall into this category.

(a) What features of tRNA^{Ala} are recognized by Ala-tRNA synthetase?

(b) Describe the consequences of a C \rightarrow G mutation in the third position of the anticodon of tRNA^{Ala}.

(c) What other kinds of mutations might have similar effects?

(d) Mutations of these types are never found in natural populations of organisms. Why? (Hint: Consider what might happen both to individual proteins and to the organism as a whole.)

14. The Role of Translation Factors A researcher isolates mutant variants of the bacterial translation factors IF-2, EF-Tu, and EF-G. In each case, the mutation allows proper folding of the protein and the binding of GTP but does not allow GTP hydrolysis. At what stage would translation be blocked by each mutant protein?

15. Maintaining the Fidelity of Protein Synthesis The chemical mechanisms used to avoid errors in protein synthesis are different from those used during DNA replication. DNA polymerases use a 3' \rightarrow 5' exonuclease proofreading activity to remove mispaired nucleotides incorrectly inserted into a growing DNA strand. There is no analogous proofreading function on ribosomes and, in fact, the identity of an amino acid attached to an incoming tRNA and added to the growing polypeptide is never checked. A proofreading step that hydrolyzed the previously formed peptide bond after an incorrect amino acid had been inserted into a growing polypeptide (analogous to the proofreading step of DNA polymerases) would be impractical. Why? (Hint: Consider how the link between the growing polypeptide and the mRNA is maintained during elongation; see Figs 27–29 and 27–30.)

16. Predicting the Cellular Location of a Protein The gene for a eukaryotic polypeptide 300 amino acid residues long is altered so that a signal sequence recognized by SRP occurs at the polypeptide's amino terminus and a nuclear localization signal (NLS) occurs internally, beginning at residue 150. Where is the protein likely to be found in the cell?

17. Requirements for Protein Translocation across a Membrane The secreted bacterial protein OmpA has a precursor, ProOmpA, which has the amino-terminal signal sequence required for secretion. If purified ProOmpA is denatured with 8 M urea and the urea is then removed (such as by running the protein solution rapidly through a gel filtration column), the protein can be translocated across isolated bacterial inner membranes *in vitro*. However, translocation becomes impossible if ProOmpA is first allowed to incubate for a few hours in the absence of urea. Furthermore, the capacity for translocation is maintained for an extended period if ProOmpA is first incubated in the presence of another bacterial protein called trigger factor. Describe the probable function of this factor.

18. Protein-Coding Capacity of a Viral DNA The 5,386 bp genome of bacteriophage ϕ X174 includes genes for 10 proteins, designated A to K, with sizes given in the table below. How much DNA would be required to encode these 10 proteins? How can you reconcile the size of the ϕ X174 genome with its protein-coding capacity?

Protein	Number of amino acid residues	Protein	Number of amino acid residues
A	455	F	427
B	120	G	175
C	86	H	328
D	152	J	38
E	91	K	56

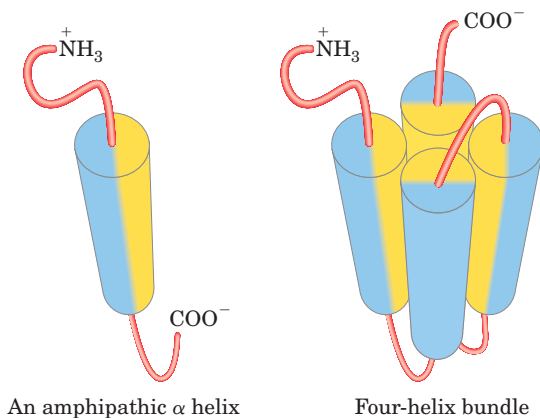
Data Analysis Problem

19. Designing Proteins by Using Randomly Generated Genes Studies of the amino acid sequence and corresponding three-dimensional structure of wild-type or mutant proteins have led to significant insights into the principles that govern protein folding. An important test of this understanding would be to *design* a protein based on these principles and see whether it folds as expected.

Kamtekar and colleagues (1993) used aspects of the genetic code to generate random protein sequences with defined patterns of hydrophilic and hydrophobic residues. Their clever approach combined knowledge about protein structure, amino acid properties, and the genetic code to explore the factors that influence protein structure.

They set out to generate a set of proteins with the simple four-helix bundle structure shown at the end of the paragraph, with α helices (shown as cylinders) connected by segments of random coil (light red). Each α helix is amphipathic—the

R groups on one side of the helix are exclusively hydrophobic (yellow) and those on the other side are exclusively hydrophilic (blue). A protein consisting of four of these helices separated by short segments of random coil would be expected to fold so that the hydrophilic sides of the helices face the solvent.



(a) What forces or interactions hold the four α helices together in this bundled structure?

Figure 4–4a shows a segment of α helix consisting of 10 amino acid residues. With the gray central rod as a divider, four of the R groups (purple spheres) extend from the left side of the helix and six extend from the right.

(b) Number the R groups in Figure 4–4a, from top (amino terminus; 1) to bottom (carboxyl terminus; 10). Which R groups extend from the left side and which from the right?

(c) Suppose you wanted to design this 10 amino acid segment to be an amphipathic helix, with the left side hydrophilic and the right side hydrophobic. Give a sequence of 10 amino acids that could potentially fold into such a structure. There are many possible correct answers here.

(d) Give one possible double-stranded DNA sequence that could encode the amino acid sequence you chose for (c). (It is an internal portion of a protein, so you do not need to include start or stop codons.)

Rather than designing proteins with specific sequences, Kamtekar and colleagues designed proteins with partially random sequences, with hydrophilic and hydrophobic amino acid residues placed in a controlled pattern. They did this by taking advantage of some interesting features of the genetic code to construct a library of synthetic DNA molecules with partially random sequences arranged in a particular pattern.

To design a DNA sequence that would encode random hydrophobic amino acid sequences, the researchers began with the degenerate codon NTN, where N can be A, G, C, or T. They filled each N position by including an equimolar mixture of A, G, C, and T in the DNA synthesis reaction to generate a mixture of DNA molecules with different nucleotides at that position (see Fig. 8–35). Similarly, to encode random polar amino acid sequences, they began with the degenerate codon NAN and used an equimolar mixture of A, G, and C (but in this case, no T) to fill the N positions.

(e) Which amino acids can be encoded by the NTN triplet? Are all amino acids in this set hydrophobic? Does the set include *all* the hydrophobic amino acids?

(f) Which amino acids can be encoded by the NAN triplet? Are all of these polar? Does the set include *all* the polar amino acids?

(g) In creating the NAN codons, why was it necessary to leave T out of the reaction mixture?

Kamtekar and coworkers cloned this library of random DNA sequences into plasmids, selected 48 that produced the correct patterning of hydrophilic and hydrophobic amino acids, and expressed these in *E. coli*. The next challenge was to determine whether the proteins folded as expected. It would be very time-consuming to express each protein, crystallize it, and determine its complete three-dimensional structure. Instead, the investigators used the *E. coli* protein-processing machinery to screen out sequences that led to highly defective proteins. In this initial screening, they kept only those

clones that resulted in a band of protein with the expected molecular weight on SDS polyacrylamide gel electrophoresis (see Fig. 3–18).

(h) Why would a grossly misfolded protein fail to produce a band of the expected molecular weight on electrophoresis?

Several proteins passed this initial test, and further exploration showed that they had the expected four-helix structure.

(i) Why didn't all of the random-sequence proteins that passed the initial screening test produce four-helix structures?

Reference

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., & Hecht, M.H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680–1685.

this page left intentionally blank

Regulation of Gene Expression

28.1 Principles of Gene Regulation 1156

28.2 Regulation of Gene Expression in Bacteria 1165

28.3 Regulation of Gene Expression in Eukaryotes 1175

Of the 4,000 or so genes in the typical bacterial genome, or the perhaps 25,000 genes in the human genome, only a fraction are expressed in a cell at any given time. Some gene products are present in very large amounts: the elongation factors required for protein synthesis, for example, are among the most abundant proteins in bacteria, and ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco) of plants and photosynthetic bacteria is, as far as we know, the most abundant enzyme in the biosphere. Other gene products occur in much smaller amounts; for instance, a cell may contain only a few molecules of the enzymes that repair rare DNA lesions. Requirements for some gene products change over time. The need for enzymes in certain metabolic pathways may wax and wane as food sources change or are depleted. During development of a multicellular organism, some proteins that influence cellular differentiation are present for just a brief time in only a few cells. Specialization of cellular function can dramatically affect the need for various gene products; an example is the uniquely high concentration of a single protein—hemoglobin—in erythrocytes. Given the high cost of protein synthesis, regulation of gene expression is essential to making optimal use of available energy.

The cellular concentration of a protein is determined by a delicate balance of at least seven processes, each having several potential points of regulation:

1. Synthesis of the primary RNA transcript (transcription)
2. Posttranscriptional modification of mRNA
3. Messenger RNA degradation
4. Protein synthesis (translation)
5. Posttranslational modification of proteins
6. Protein targeting and transport
7. Protein degradation

These processes are summarized in **Figure 28–1**. We have examined several of these mechanisms in previous chapters. Posttranscriptional modification of mRNA, by processes such as alternative splicing patterns (see Fig. 26–21) or RNA editing (see Figs 27–10, 27–12), can affect which proteins are produced from an mRNA transcript and in what amounts. A variety of nucleotide sequences in an mRNA can affect the rate of its degradation (p. 1084). Many factors affect the rate at which an mRNA is translated into a protein, as well as the post-translational modification, targeting, and eventual degradation of that protein (Chapter 27).

Of the regulatory processes illustrated in Figure 28–1, those operating at the level of transcription initiation are particularly well-documented. These processes are a major focus of this chapter, although other mechanisms are also considered. Researchers continue to discover complex and sometimes surprising regulatory mechanisms, leading to an increasing appreciation of the importance of posttranscriptional and translational regulation, especially in eukaryotes. For many genes, the regulatory processes are elaborate and redundant and can involve a considerable investment of chemical energy.

Control of transcription initiation permits the synchronized regulation of multiple genes encoding products with interdependent activities. For example, when their DNA is heavily damaged, bacterial cells require a coordinated increase in the levels of the many DNA repair enzymes. And perhaps the most sophisticated form of coordination occurs in the complex regulatory circuits that guide the development of multicellular eukaryotes, which can involve many types of regulatory mechanisms.

We begin by examining the interactions between proteins and DNA that are the key to transcriptional regulation. We next discuss the specific proteins that influence the expression of specific genes, first in bacterial

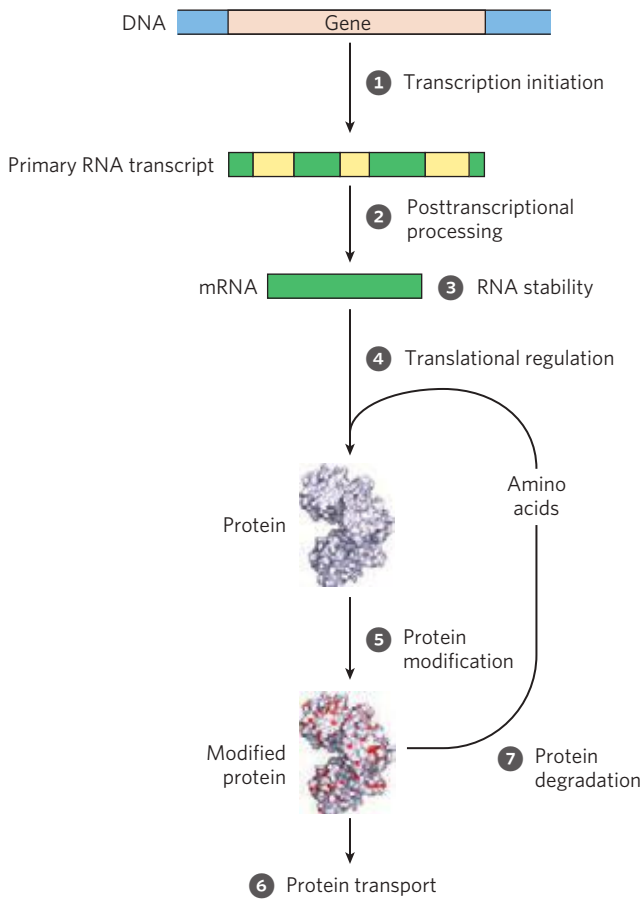


FIGURE 28-1 Seven processes that affect the steady-state concentration of a protein. Each process has several potential points of regulation.

and then in eukaryotic cells. Information about post-transcriptional and translational regulation is included in the discussion, where relevant, to provide a more complete overview of the rich complexity of regulatory mechanisms.

28.1 Principles of Gene Regulation

Genes for products that are required at all times, such as those for the enzymes of central metabolic pathways, are expressed at a more or less constant level in virtually every cell of a species or organism. Such genes are often referred to as **housekeeping genes**. Unvarying expression of a gene is called **constitutive gene expression**.

For other gene products, cellular levels rise and fall in response to molecular signals; this is **regulated gene expression**. Gene products that increase in concentration under particular molecular circumstances are referred to

as **inducible**; the process of increasing their expression is **induction**. The expression of many of the genes encoding DNA repair enzymes, for example, is induced by a system of regulatory proteins that responds to high levels of DNA damage. Conversely, gene products that decrease in concentration in response to a molecular signal are referred to as **repressible**, and the process is called **repression**. For example, in bacteria, ample supplies of tryptophan lead to repression of the genes for the enzymes that catalyze tryptophan biosynthesis.

Transcription is mediated and regulated by protein-DNA interactions, especially those involving the protein components of RNA polymerase (Chapter 26). We first consider how the activity of RNA polymerase is regulated, and proceed to a general description of the proteins participating in this regulation. We then examine the molecular basis for the recognition of specific DNA sequences by DNA-binding proteins.

RNA Polymerase Binds to DNA at Promoters

RNA polymerases bind to DNA and initiate transcription at promoters (see Fig. 26-5), sites generally found near points at which RNA synthesis begins on the DNA template. The regulation of transcription initiation often entails changes in how RNA polymerase interacts with a promoter.

The nucleotide sequences of promoters vary considerably, affecting the binding affinity of RNA polymerases and thus the frequency of transcription initiation. Some *Escherichia coli* genes are transcribed once per second, others less than once per cell generation. Much of this variation is due to differences in promoter sequence. In the absence of regulatory proteins, differences in promoter sequence may affect the frequency of transcription initiation by a factor of 1,000 or more. Most *E. coli* promoters have a sequence close to a consensus (Fig. 28-2). Mutations that result in a shift away from the consensus sequence usually decrease promoter function; conversely, mutations toward consensus usually enhance promoter function.

KEY CONVENTION: By convention, DNA sequences are shown as they exist in the nontemplate strand, with the 5' terminus on the left. Nucleotides are numbered from the transcription start site, with positive numbers to the right (in the direction of transcription) and negative numbers to the left. N indicates any nucleotide. ■

Although housekeeping genes are expressed constitutively, the cellular concentrations of the proteins they

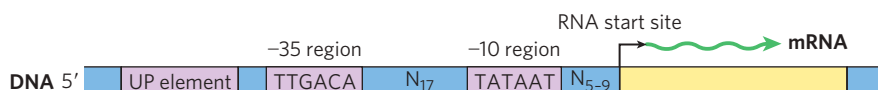


FIGURE 28-2 Consensus sequence for many *E. coli* promoters. Most base substitutions in the -10 and -35 regions have a negative effect on promoter function. Some promoters also include the UP (upstream promoter) element (see Fig. 26-5).

encode vary widely. For these genes, the RNA polymerase–promoter interaction strongly influences the rate of transcription initiation; differences in promoter sequence allow the cell to synthesize the appropriate level of each housekeeping gene product.

The basal rate of transcription initiation at the promoters of nonhousekeeping genes is also determined by the promoter sequence, but expression of these genes is further modulated by regulatory proteins. Many of these proteins work by enhancing or interfering with the interaction between RNA polymerase and the promoter.

The sequences of eukaryotic promoters are more variable than their bacterial counterparts (see Fig. 26–8). The three eukaryotic RNA polymerases usually require an array of general transcription factors in order to bind to a promoter. Yet, as with bacterial gene expression, the basal level of transcription is determined by the effect of promoter sequences on the function of RNA polymerase and its associated transcription factors.

Transcription Initiation Is Regulated by Proteins That Bind to or Near Promoters

At least three types of proteins regulate transcription initiation by RNA polymerase: **specificity factors** alter the specificity of RNA polymerase for a given promoter or set of promoters, **repressors** impede access of RNA polymerase to the promoter, and **activators** enhance the RNA polymerase–promoter interaction.

We introduced bacterial specificity factors in Chapter 26 (see Fig. 26–5), although we did not refer to them by that name. The σ subunit of the *E. coli* RNA polymerase holoenzyme is a specificity factor that mediates promoter recognition and binding. Most *E. coli* promoters are recognized by a single σ subunit (M_r 70,000), σ^{70} . Under some conditions, some of the σ^{70} subunits are replaced by one of six other specificity factors. One notable case arises when the bacteria are subjected to heat stress, leading to the replacement of σ^{70} by σ^{32} (M_r 32,000). When bound to σ^{32} , RNA polymerase is directed to a specialized set of promoters with a different consensus sequence (Fig. 28–3). These promoters control the expression of a set of genes that encode proteins, including some protein chaperones (p. 146), that are part of a stress-induced system called the heat shock response. Thus, through changes in the binding affinity of the polymerase that direct it to different

promoters, a set of genes involved in related processes is coordinately regulated. In eukaryotic cells, some of the general transcription factors, in particular the TATA-binding protein (TBP; see Fig. 26–9), may be considered specificity factors.

Repressors bind to specific sites on the DNA. In bacterial cells, such binding sites, called **operators**, are generally near a promoter. RNA polymerase binding, or its movement along the DNA after binding, is blocked when the repressor is present. Regulation by means of a repressor protein that blocks transcription is referred to as **negative regulation**. Repressor binding to DNA is regulated by a molecular signal (or **effector**), usually a small molecule or a protein, that binds to the repressor and causes a conformational change. The interaction between repressor and signal molecule either increases or decreases transcription. In some cases, the conformational change results in dissociation of a DNA-bound repressor from the operator (Fig. 28–4a). Transcription initiation can then proceed unhindered. In other cases, interaction between an inactive repressor and the signal molecule causes the repressor to bind to the operator (Fig. 28–4b). In eukaryotic cells, the binding site for a repressor may be some distance from the promoter. Binding of these repressors to their binding sites has the same effect as in bacterial cells: inhibiting the assembly or activity of a transcription complex at the promoter.

Activators provide a molecular counterpoint to repressors; they bind to DNA and *enhance* the activity of RNA polymerase at a promoter; this is **positive regulation**. In bacteria, activator-binding sites are often adjacent to promoters that are bound weakly or not at all by RNA polymerase alone, such that little transcription occurs in the absence of the activator. Many eukaryotic activators bind to DNA sites, called enhancers, that are quite distant from the promoter. These activators affect the rate of transcription at a promoter that may be located thousands of base pairs away. Some activators are usually bound to DNA, enhancing transcription until dissociation of the activator is triggered by the binding of a signal molecule (Fig. 28–4c). In other cases the activator binds to DNA only after interaction with a signal molecule (Fig. 28–4d). Signal molecules can therefore increase or decrease transcription, depending on how they affect the activator. Positive regulation is particularly common in eukaryotes, as we shall see.



FIGURE 28–3 Consensus sequence for promoters that regulate expression of the *E. coli* heat shock genes. This system responds to temperature increases as well as some other environmental stresses, resulting in the induction of a set of proteins. Binding of RNA polymerase to heat shock promoters is mediated by a specialized σ subunit of the polymerase, σ^{32} , which replaces σ^{70} in the RNA polymerase initiation complex.

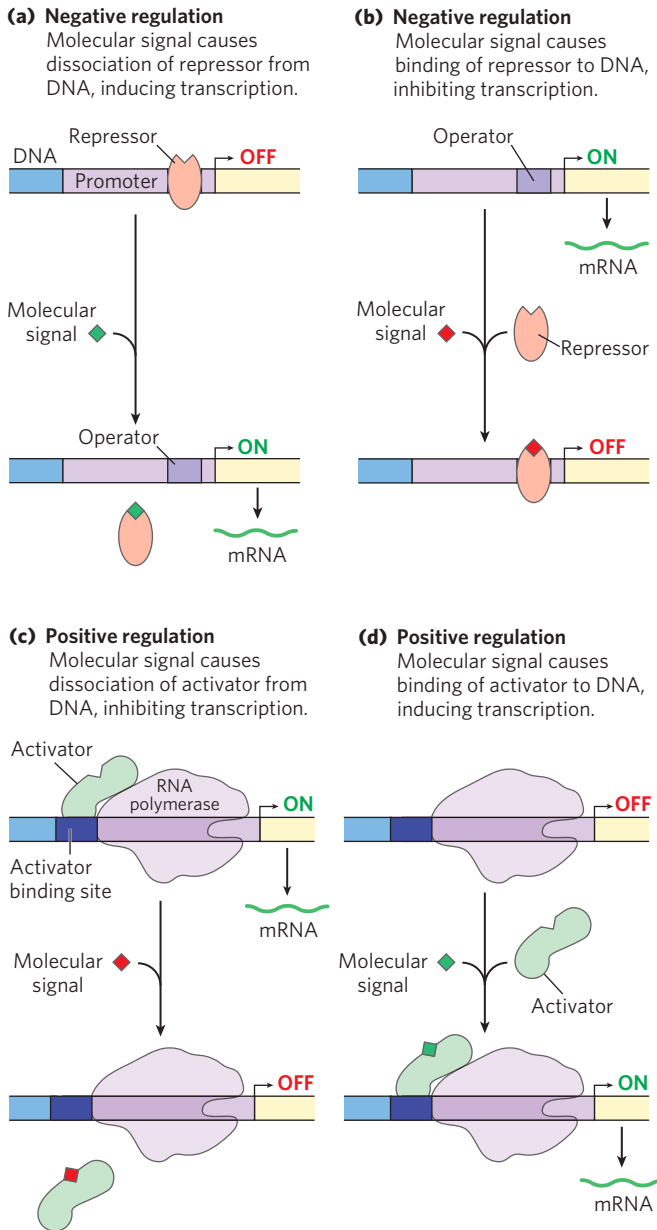


FIGURE 28-4 Common patterns of regulation of transcription initiation.

Two types of negative regulation are illustrated. **(a)** Repressor binds to the operator in the absence of the molecular signal; the external signal causes dissociation of the repressor to permit transcription. **(b)** Repressor binds in the presence of the signal; the repressor dissociates and transcription ensues when the signal is removed. Positive regulation is mediated by gene activators. Again, two types are shown. **(c)** Activator binds in the absence of the molecular signal and transcription proceeds; when the signal is added, the activator dissociates and transcription is inhibited. **(d)** Activator binds in the presence of the signal; it dissociates only when the signal is removed. Note that “positive” and “negative” regulation refer to the type of regulatory protein involved: the bound protein either facilitates or inhibits transcription. In either case, addition of the molecular signal may increase or decrease transcription, depending on its effect on the regulatory protein.

In eukaryotes, the distance between the binding sites of activators or repressors and promoters is bridged by looping out the DNA in between (**Fig. 28-5**). The looping is facilitated in some cases by proteins called **archi-**

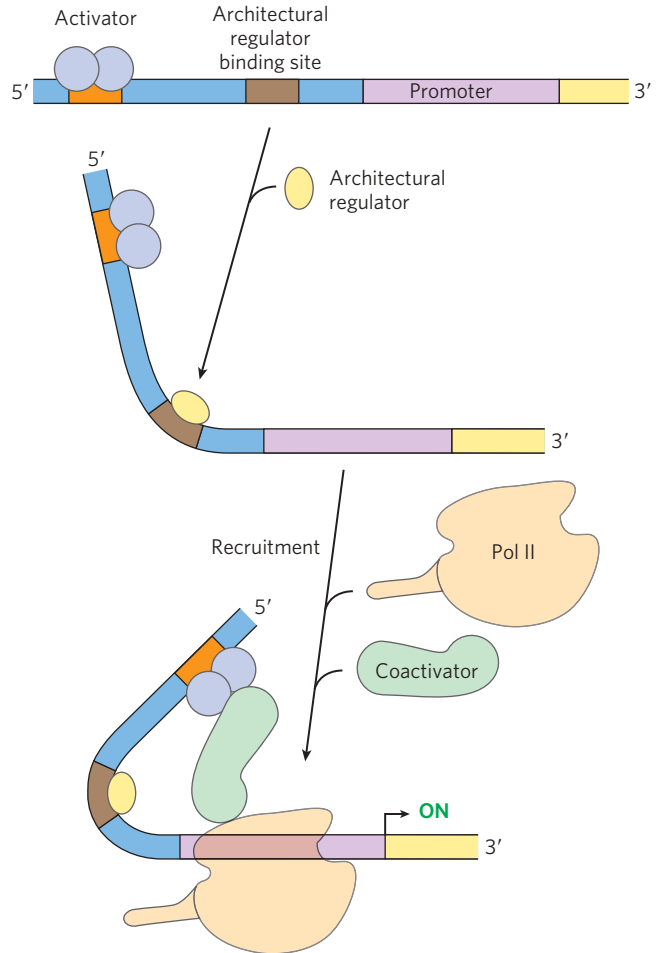


FIGURE 28-5 The interaction between activators/repressors and RNA polymerase in eukaryotes. Eukaryotic activators and repressors often bind sites thousands of base pairs from the promoters they regulate. DNA looping, often facilitated by architectural regulators, brings the sites together. The interaction between activators and RNA polymerase is often mediated by coactivators, as shown. Repression is sometimes mediated by repressors (described later) that bind to activators, thereby preventing the activating interaction with RNA polymerase.

tectural regulators that bind to intervening sites and facilitate the looping of the DNA. Most of the eukaryotic systems involve protein activators. The actual interaction between the activators and the RNA polymerase at the promoter is often mediated by intermediary proteins called coactivators. In some instances, protein repressors may take the place of coactivators, binding to the activators and preventing the activating interaction.

Many Bacterial Genes Are Clustered and Regulated in Operons

Bacteria have a simple general mechanism for coordinating the regulation of genes encoding products that participate in a set of related processes: these genes are clustered on the chromosome and are transcribed together. Many bacterial mRNAs are polycistronic—multiple genes on a single transcript—and the single promoter that initiates transcription of the cluster is the

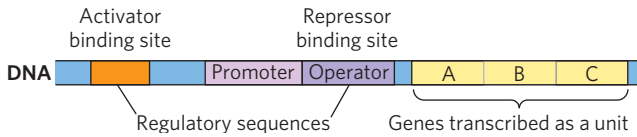


FIGURE 28-6 Representative bacterial operon. Genes A, B, and C are transcribed on one polycistronic mRNA. Typical regulatory sequences include binding sites for proteins that either activate or repress transcription from the promoter.

site of regulation for expression of all the genes in the cluster. The gene cluster and promoter, plus additional sequences that function together in regulation, are called an **operon (Fig. 28-6)**. Operons that include two to six genes transcribed as a unit are common; some operons contain 20 or more genes.

Many of the principles of bacterial gene expression were first defined by studies of lactose metabolism in *E. coli*, which can use lactose as its sole carbon source. In 1960, François Jacob and Jacques Monod published a short paper in the *Proceedings* of the French Academy of Sciences that described how two adjacent genes involved in lactose metabolism were coordinately regulated by a genetic element located at one end of the gene cluster. The genes were those for β -galactosidase, which cleaves lactose to galactose and glucose, and for galactoside permease (lactose permease, p. 416), which transports lactose into the cell (Fig. 28-7). The terms “operon” and “operator” were first introduced in this paper. With the operon model, gene regulation could, for the first time, be considered in molecular terms.



François Jacob



Jacques Monod, 1910–1976

The *lac* Operon Is Subject to Negative Regulation

The lactose (*lac*) operon (Fig. 28-8a) includes the genes for β -galactosidase (*Z*), galactoside permease (*Y*), and thiogalactoside transacetylase (*A*). The last of these enzymes seems to modify toxic galactosides to facilitate their removal from the cell. Each of the three genes is preceded by a ribosome-binding site (not shown in Fig. 28-8) that independently directs the translation of that gene (Chapter 27). Regulation of the *lac* operon by the *lac* repressor protein (Lac) follows the pattern outlined in Figure 28-4a.

The study of *lac* operon mutants has revealed some details of the workings of the operon’s regulatory system. In the absence of lactose, the *lac* operon genes are repressed. Mutations in the operator or in another gene, the *I* gene, result in constitutive synthesis of the gene products. When the *I* gene is defective, repression can be restored by introducing a functional *I* gene into the cell on another DNA molecule, demonstrating that the *I* gene encodes a diffusible molecule that causes gene repression. This molecule proved to be a protein, now called the Lac repressor, a tetramer of identical monomers. The operator to which it binds most tightly (O_1) abuts the transcription start site (Fig. 28-8a). The *I* gene is transcribed from its own promoter (P_I) independent of the *lac* operon genes. The *lac* operon has two secondary binding sites for the Lac repressor. One (O_2) is centered near position +410, within the gene encoding β -galactosidase (*Z*); the other (O_3) is near position -90, within the *I* gene. To repress the operon, the Lac repressor seems to bind to both the main operator and one of the two secondary sites, with the intervening

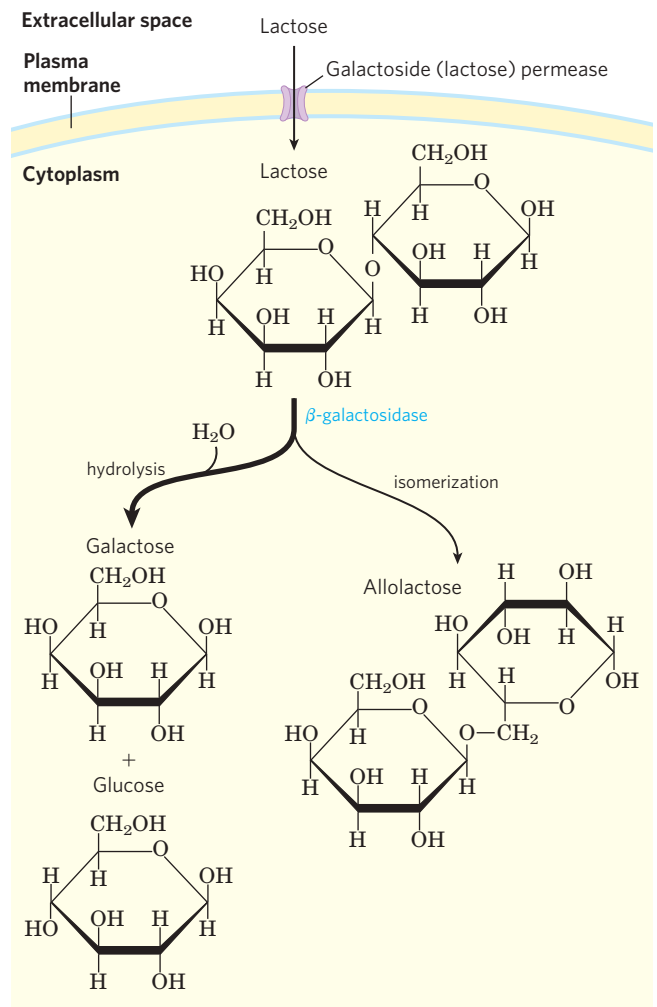


FIGURE 28-7 Lactose metabolism in *E. coli*. Uptake and metabolism of lactose require the activities of galactoside (lactose) permease and β -galactosidase. Conversion of lactose to allolactose by transglycosylation is a minor reaction also catalyzed by β -galactosidase.

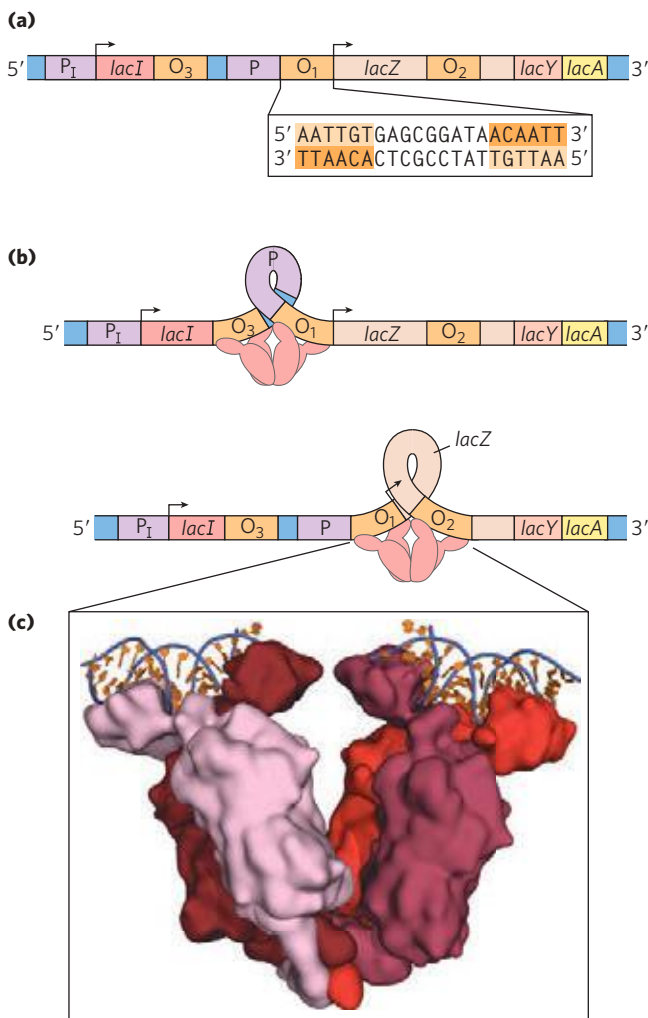


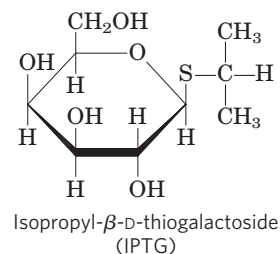
FIGURE 28-8 The *lac* operon. **(a)** The *lac* operon. The *lacI* gene encodes the Lac repressor. The *lac Z*, *Y*, and *A* genes encode β -galactosidase, galactoside permease, and thiogalactoside transacetylase, respectively. P is the promoter for the *lac* genes, and P_I is the promoter for the *I* gene. O₁ is the main operator for the *lac* operon; O₂ and O₃ are secondary operator sites of lesser affinity for the Lac repressor. The inverted repeat to which the Lac repressor binds in O₁ is shown in the inset. **(b)** The Lac repressor binds to the main operator and O₂ or O₃, apparently forming a loop in the DNA. **(c)** (PDB ID 2PE5) Lac repressor (shades of red) is shown bound to short, discontinuous segments of DNA (blue and orange).

DNA looped out (Fig. 28-8b, c). Either binding arrangement blocks transcription initiation.

Despite this elaborate binding complex, repression is not absolute. Binding of the Lac repressor reduces the rate of transcription initiation by a factor of 10^3 . If the O₂ and O₃ sites are eliminated by deletion or mutation, the binding of repressor to O₁ alone reduces transcription by a factor of about 10^2 . Even in the repressed state, each cell has a few molecules of β -galactosidase and galactoside permease, presumably synthesized on the rare occasions when the repressor transiently dissociates from the operators. This basal level of transcription is essential to operon regulation.

When cells are provided with lactose, the *lac* operon is induced. An inducer (signal) molecule binds to a specific site on the Lac repressor, causing a conformational change that results in dissociation of the repressor from the operator. The inducer in the *lac* operon system is not lactose itself but allolactose, an isomer of lactose (Fig. 28-7). After entry into the *E. coli* cell (via the few existing molecules of lactose permease), lactose is converted to allolactose by one of the few existing β -galactosidase molecules. Release of the operator by Lac repressor, triggered as the repressor binds to allolactose, allows expression of the *lac* operon genes and leads to a 10^3 -fold increase in the concentration of β -galactosidase.

Several β -galactosides structurally related to allolactose are inducers of the *lac* operon but are not substrates for β -galactosidase; others are substrates but not inducers. One particularly effective and nonmetabolizable inducer of the *lac* operon that is often used experimentally is isopropylthiogalactoside (IPTG).



An inducer that cannot be metabolized allows researchers to explore the physiological function of lactose as a carbon source for growth, separate from its function in the regulation of gene expression.

In addition to the multitude of operons now known in bacteria, a few polycistronic operons have been found in the cells of lower eukaryotes. In the cells of higher eukaryotes, however, almost all protein-encoding genes are transcribed separately.

The mechanisms by which operons are regulated can vary significantly from the simple model presented in Figure 28-8. Even the *lac* operon is more complex than indicated here, with an activator also contributing to the overall scheme, as we shall see in Section 28.2. Before any further discussion of the layers of regulation of gene expression, however, we examine the critical molecular interactions between DNA-binding proteins (such as repressors and activators) and the DNA sequences to which they bind.

Regulatory Proteins Have Discrete DNA-Binding Domains

Regulatory proteins generally bind to specific DNA sequences. Their affinity for these target sequences is roughly 10^4 to 10^6 times higher than their affinity for any other DNA sequence. Most regulatory proteins have discrete DNA-binding domains containing substructures

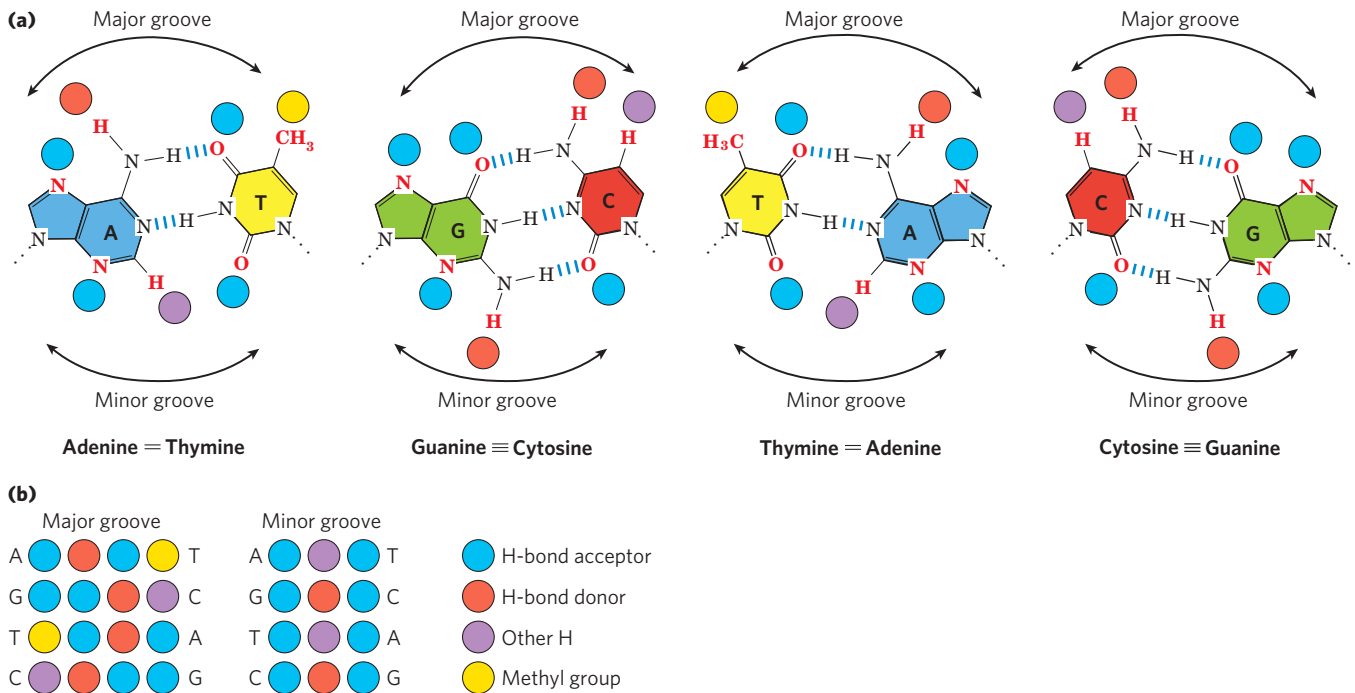


FIGURE 28-9 Groups in DNA available for protein binding. **(a)** Shown here are functional groups on all four base pairs that are displayed in the major and minor grooves of DNA (see Fig. 8-13). Hydrogen-bond acceptor and donor atoms are marked by blue and red disks, respectively. Other hydrogen atoms are marked with purple disks, and methyl groups with

yellow disks. **(b)** Recognition patterns for each base pair, from left to right, are summarized at bottom. The much greater variation in the patterns for the major groove relative to the minor groove.

that interact closely and specifically with the DNA. These binding domains usually include one or more of a relatively small group of recognizable and characteristic structural motifs.

To bind specifically to DNA sequences, regulatory proteins must recognize surface features on the DNA. Most of the chemical groups that differ among the four bases and thus permit discrimination between base pairs are hydrogen-bond donor and acceptor groups exposed in the major groove of DNA (**Fig. 28-9**), and most of the protein-DNA contacts that impart specificity are hydrogen bonds. A notable exception is the nonpolar surface near C-5 of pyrimidines, where thymine is readily distinguished from cytosine by its protruding methyl group. Protein-DNA contacts are also possible in the minor groove of the DNA, but the hydrogen-bonding patterns here generally do not allow ready discrimination between base pairs.

Within regulatory proteins, the amino acid side chains most often hydrogen-bonding to bases in the DNA are those of Asn, Gln, Glu, Lys, and Arg residues. Is there a simple recognition code in which a particular amino acid always pairs with a particular base? The two hydrogen bonds that can form between Gln or Asn and the N^6 and N^7 positions of adenine cannot form with any other base. And an Arg residue can form two hydrogen bonds with N^7 and O^6 of guanine (**Fig. 28-10**). Examination of the structure of many DNA-binding proteins, however, has shown that a protein can recognize

each base pair in more than one way, leading to the conclusion that there is no simple amino acid-base code. For some proteins, the Gln-adenine interaction can specify A=T base pairs, but in others a van der Waals pocket for the methyl group of thymine can recognize A=T base pairs. Researchers cannot yet examine the structure of a DNA-binding protein and infer the DNA sequence to which it binds.

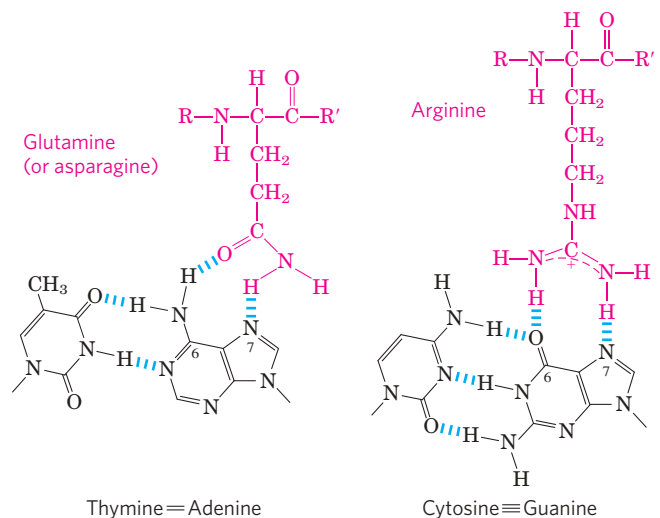


FIGURE 28-10 Specific amino acid residue-base pair interactions. The two examples shown have been observed in DNA-protein binding.

To interact with bases in the major groove of DNA, a protein requires a relatively small substructure that can stably protrude from the protein surface. The DNA-binding domains of regulatory proteins tend to be small (60 to 90 amino acid residues), and the structural motifs within these domains that are actually in contact with the DNA are smaller still. Many small proteins are unstable because of their limited capacity to form layers of structure to bury hydrophobic groups (p. 116). The DNA-binding motifs provide either a very compact stable structure or a way of allowing a segment of protein to protrude from the protein surface.

The DNA-binding sites for regulatory proteins are often inverted repeats of a short DNA sequence (a palindrome) at which multiple (usually two) subunits of a regulatory protein bind cooperatively. The Lac repressor is unusual in that it functions as a tetramer, with two dimers tethered together at the end distant from the DNA-binding sites (Fig. 28–8b). An *E. coli* cell usually contains about 20 tetramers of the Lac repressor. Each of the tethered dimers separately binds to a palindromic operator sequence, in contact with 17 bp of a 22 bp region in the *lac* operon. And each of the tethered dimers can independently bind to an operator sequence, with one generally binding to O_1 and the other to O_2 or O_3 (as in Fig. 28–8b). The symmetry of the O_1 operator sequence corresponds to the twofold axis of symmetry of two paired Lac repressor subunits. The tetrameric Lac repressor binds to its operator sequences *in vivo* with an estimated dissociation constant of about 10^{-10} M. The repressor discriminates between the operators and other sequences by a factor of about 10^6 , so binding to these few base pairs among the 4.6 million or so of the *E. coli* chromosome is highly specific.

Several DNA-binding motifs have been described, but here we focus on two that play prominent roles in the binding of DNA by regulatory proteins: the **helix-turn-helix** and the **zinc finger**. We also consider a type of DNA-binding domain—the homeodomain—found in some eukaryotic proteins.

Helix-Turn-Helix This DNA-binding motif is crucial to the interaction of many bacterial regulatory proteins with DNA, and similar motifs occur in some eukaryotic regulatory proteins. The helix-turn-helix motif comprises about 20 amino acids in two short α -helical segments, each seven to nine amino acid residues long, separated by a β turn (Fig. 28–11). This structure generally is not stable by itself; it is simply the reactive portion of a somewhat larger DNA-binding domain. One of the two α -helical segments is called the recognition helix because it usually contains many of the amino acids that interact with the DNA in a sequence-specific way. This α helix is stacked on other segments of the protein structure so that it protrudes from the protein surface. When bound to DNA, the recognition helix is positioned

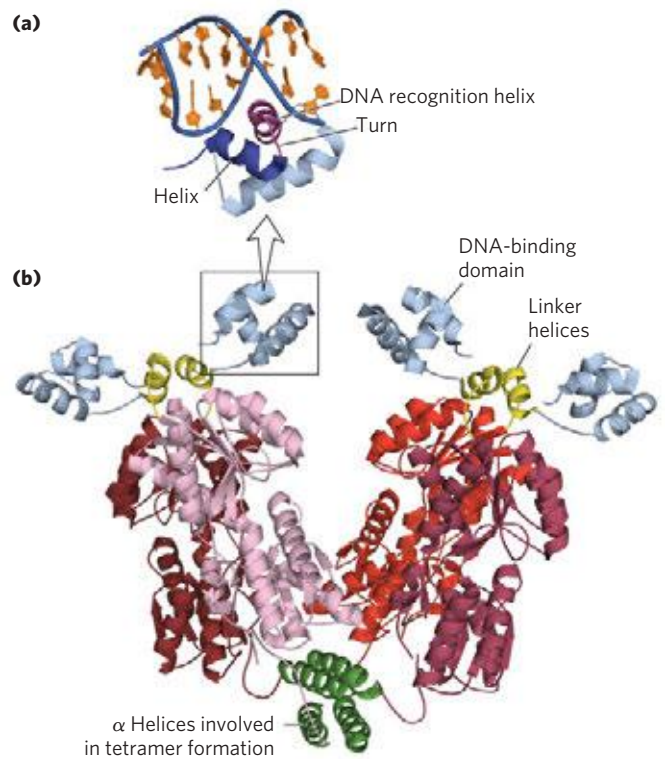


FIGURE 28–11 Helix-turn-helix. (PDB ID 2PE5) **(a)** DNA-binding domain of the Lac repressor bound to DNA (blue and orange). The helix-turn-helix motif is shown in dark blue and purple; the DNA recognition helix is purple. **(b)** Entire Lac repressor. The DNA-binding domains are light blue, and the α helices involved in tetramer formation are green. The remainder of the protein (shades of red) has the binding sites for allolactose. The allolactose-binding domains are linked to the DNA-binding domains through linker helices (yellow).

in or nearly in the major groove. The Lac repressor has this DNA-binding motif (Fig. 28–11).

Zinc Finger In a zinc finger, about 30 amino acid residues form an elongated loop held together at the base by a single Zn^{2+} ion, which is coordinated to four of the residues (four Cys, or two Cys and two His). The zinc does not itself interact with DNA; rather, the coordination of zinc with the amino acid residues stabilizes this small structural motif. Several hydrophobic side chains in the core of the structure also lend stability. **Figure 28–12** shows the interaction between DNA and three zinc fingers of a single polypeptide from the mouse regulatory protein Zif268.

Many eukaryotic DNA-binding proteins contain zinc fingers. The interaction of a single zinc finger with DNA is typically weak, and many DNA-binding proteins, like Zif268, have multiple zinc fingers that substantially enhance binding by interacting simultaneously with the DNA. One DNA-binding protein of the frog *Xenopus* has 37 zinc fingers. There are few known examples of the zinc finger motif in bacterial proteins.

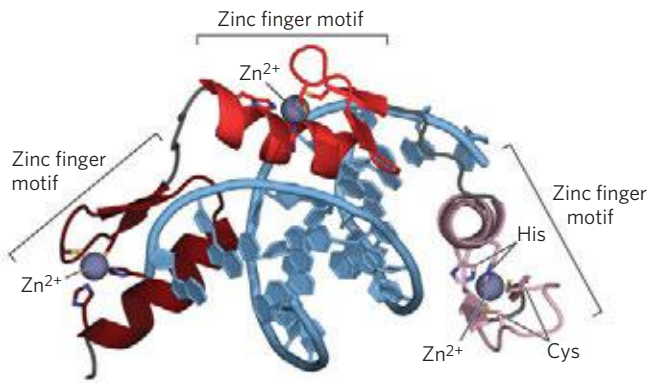


FIGURE 28-12 Zinc fingers. (PDB ID 1ZAA) Three zinc fingers (shades of red) of the regulatory protein Zif268, complexed with DNA (blue). Each Zn^{2+} coordinates with two His and two Cys residues.

The precise manner in which proteins with zinc fingers bind to DNA differs from one protein to the next. Some zinc fingers contain the amino acid residues that are important in sequence discrimination, whereas others seem to bind DNA nonspecifically (the amino acids required for specificity are located elsewhere in the protein). Zinc fingers can also function as RNA-binding motifs—for example, in certain proteins that bind eukaryotic mRNAs and act as translational repressors. We discuss this role later (Section 28.3).

Homeodomain Another type of DNA-binding domain has been identified in some proteins that function as transcriptional regulators, especially during eukaryotic development. This domain of 60 amino acids—called the **homeodomain** because it was discovered in homeotic genes (genes that regulate the development of body patterns)—is highly conserved and has now been identified in proteins from a wide variety of organisms, including humans (Fig. 28-13). The DNA-binding segment of the domain is related to the helix-turn-helix motif. The DNA sequence that encodes this domain is known as the **homeobox**.

Regulatory Proteins Also Have Protein-Protein Interaction Domains

Regulatory proteins contain domains not only for DNA binding but also for protein-protein interactions—with RNA polymerase, other regulatory proteins, or other subunits of the same regulatory protein. Examples include many eukaryotic transcription factors that function as gene activators, which often bind as dimers to the DNA, through DNA-binding domains that contain zinc fingers. Some structural domains are devoted to the interactions required for dimer formation, which is generally a prerequisite for DNA binding. Like DNA-binding motifs, the structural motifs that mediate protein-protein interactions tend to fall within one of a few common categories. Two important examples

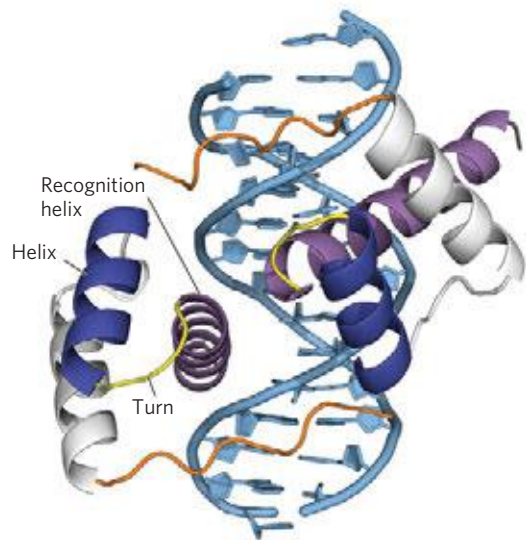


FIGURE 28-13 Homeodomains. (PDB ID 1FJL) Shown here are two homeodomains bound to DNA. In each homeodomain, one of the α helices (purple), layered on two others (dark blue and gray), can be seen protruding into the major groove. This is only a small part of a larger regulatory protein from a class called Pax, active in the regulation of development in fruit flies (see Section 28.3).

are the **leucine zipper** and the **basic helix-loop-helix**. Structural motifs such as these are the basis for classifying some regulatory proteins into structural families.

Leucine Zipper This motif is an amphipathic α helix with a series of hydrophobic amino acid residues concentrated on one side (Fig. 28-14), with the hydrophobic surface forming the area of contact between the two polypeptides of a dimer. A striking feature of these α helices is the occurrence of Leu residues at every seventh position, forming a straight line along the hydrophobic surface. Although researchers initially thought the Leu residues interdigitated (hence the name “zipper”), we now know that they line up side by side as the interacting α helices coil around each other (forming a coiled coil; Fig. 28-14b). Regulatory proteins with leucine zippers often have a separate DNA-binding domain with a high concentration of basic (Lys or Arg) residues that can interact with the negatively charged phosphates of the DNA backbone. Leucine zippers have been found in many eukaryotic and a few bacterial proteins.

Basic Helix-Loop-Helix Another common structural motif occurs in some eukaryotic regulatory proteins implicated in the control of gene expression during the development of multicellular organisms. These proteins share a conserved region of about 50 amino acid residues important in both DNA binding and protein dimerization. This region can form two short amphipathic α helices linked by a loop of variable length, the helix-loop-helix (distinct

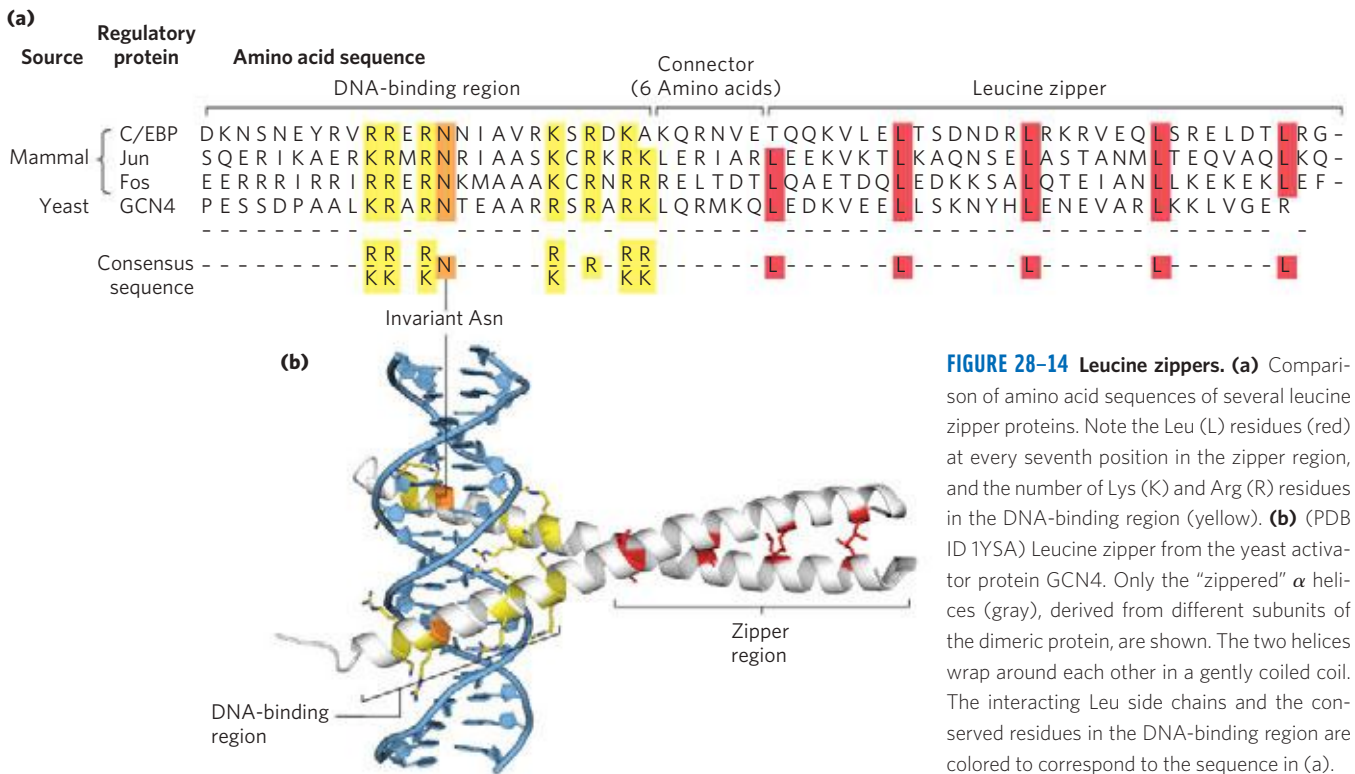


FIGURE 28-14 Leucine zippers. (a) Comparison of amino acid sequences of several leucine zipper proteins. Note the Leu (L) residues (red) at every seventh position in the zipper region, and the number of Lys (K) and Arg (R) residues in the DNA-binding region (yellow). (b) (PDB ID 1YSA) Leucine zipper from the yeast activator protein GCN4. Only the “zippered” α helices (gray), derived from different subunits of the dimeric protein, are shown. The two helices wrap around each other in a gently coiled coil. The interacting Leu side chains and the conserved residues in the DNA-binding region are colored to correspond to the sequence in (a).

from the helix-turn-helix motif associated with DNA binding). The helix-loop-helix motifs of two polypeptides interact to form dimers (Fig. 28-15). In these proteins, DNA binding is mediated by an adjacent short amino acid sequence rich in basic residues, similar to the separate DNA-binding region in proteins containing leucine zippers.

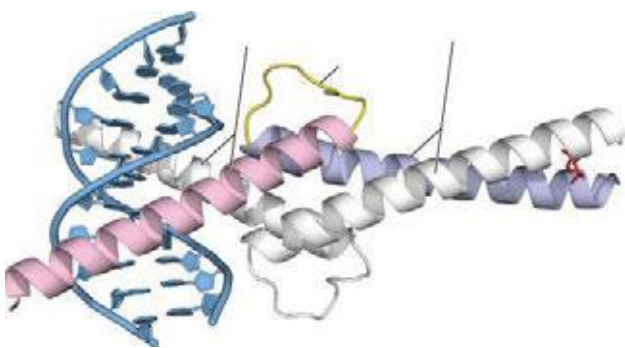


FIGURE 28-15 Helix-loop-helix. (PDB ID 1HLO) The human transcription factor Max, bound to its DNA target site. The protein is dimeric; one subunit is colored. The recognition helix (pink) is linked via the loop to the dimer-forming helix (light blue), which merges with the carboxyl-terminal end of the subunit. Interaction of the carboxyl-terminal helices of the two subunits describes a coiled coil very similar to that of a leucine zipper (see Fig. 28-14b), but with only one pair of interacting Leu residues (red side chains at the right) in this example. The overall structure is sometimes called a helix-loop-helix/leucine zipper motif.

Protein-Protein Interactions in Eukaryotic Regulatory Proteins In eukaryotes most genes are regulated by activators, and most genes are monocistronic. If a different activator were required for each gene, the number of activators (and genes encoding them) would need to be equivalent to the number of regulated genes. However, in yeast about 300 transcription factors (many of them activators) are responsible for the regulation of many thousands of yeast genes. Many of the transcription factors regulate the induction of multiple genes, but most genes are subject to regulation by multiple transcription factors. Appropriate regulation of different genes is accomplished by utilizing different combinations of a limited repertoire of transcription factors at each gene, a phenomenon referred to as **combinatorial control**.

Combinatorial control is accomplished in part by mixing and matching the variants within a regulatory protein family to form a series of different active protein dimers. Several families of eukaryotic transcription factors have been defined based on close structural similarities. Within each family, dimers can sometimes form between two identical proteins (a homodimer) or between two different members of the family (a heterodimer). A hypothetical family of four different leucine-zipper proteins could thus form up to 10 different dimeric species. In many cases, the different combinations have distinct regulatory and functional properties and function in the regulation of different genes. As we shall see, multiple regulatory proteins of this kind function in the regulation

of most eukaryotic genes, contributing further to combinatorial control.

In addition to having structural domains devoted to DNA binding and dimerization that direct a particular protein dimer to a particular gene, many regulatory proteins have domains that interact with RNA polymerase, with unrelated regulatory proteins, or with both. At least three types of additional domains for protein-protein interaction have been characterized (primarily in eukaryotes): glutamine-rich, proline-rich, and acidic domains, the names reflecting the amino acid residues that are especially abundant.

Protein-DNA binding interactions are the basis of the intricate regulatory circuits fundamental to gene function. We now turn to a closer examination of these gene regulatory schemes, first in bacterial, then in eukaryotic systems.

SUMMARY 28.1 Principles of Gene Regulation

- ▶ The expression of genes is regulated by processes that affect the rates at which gene products are synthesized and degraded. Much of this regulation occurs at the level of transcription initiation, mediated by regulatory proteins that either repress transcription (negative regulation) or activate transcription (positive regulation) at specific promoters.
- ▶ In bacteria, genes that encode products with interdependent functions are often clustered in an operon, a single transcriptional unit. Transcription of the genes is generally blocked by binding of a specific repressor protein at a DNA site called an operator. Dissociation of the repressor from the operator is mediated by a specific small molecule, an inducer. These principles were first elucidated in studies of the lactose (*lac*) operon. The Lac repressor dissociates from the *lac* operator when the repressor binds to its inducer, allolactose.
- ▶ Regulatory proteins are DNA-binding proteins that recognize specific DNA sequences; most have distinct DNA-binding domains. Within these domains, common structural motifs that bind DNA are the helix-turn-helix, zinc finger, and homeodomain.
- ▶ Regulatory proteins also contain domains for protein-protein interactions, including the leucine zipper and helix-loop-helix, which are involved in dimerization, and other motifs involved in activation of transcription. Mixing and matching of protein family variants in dimeric transcription factors provides for more efficient and responsive regulation through combinatorial control.

28.2 Regulation of Gene Expression in Bacteria

As in many other areas of biochemical investigation, the study of the regulation of gene expression advanced earlier and faster in bacteria than in other experimental organisms. The examples of bacterial gene regulation presented here are chosen from among scores of well-studied systems, partly for their historical significance, but primarily because they provide a good overview of the range of regulatory mechanisms in bacteria. Many of the principles of bacterial gene regulation are also relevant to understanding gene expression in eukaryotic cells.

We begin by examining the lactose and tryptophan operons; each system has regulatory proteins, but the overall mechanisms of regulation are very different. This is followed by a short discussion of the SOS response in *E. coli*, illustrating how genes scattered throughout the genome can be coordinately regulated. We then describe two bacterial systems of quite different types, illustrating the diversity of gene regulatory mechanisms: regulation of ribosomal protein synthesis at the level of translation, with many of the regulatory proteins binding to RNA (rather than DNA), and regulation of the process of “phase variation” in *Salmonella*, which results from genetic recombination. Finally, we examine some additional examples of posttranscriptional regulation in which the RNA modulates its own function.

The *lac* Operon Undergoes Positive Regulation

The operator-repressor-inducer interactions described earlier for the *lac* operon (Fig. 28–8) provide an intuitively satisfying model for an on/off switch in the regulation of gene expression. In truth, operon regulation is rarely so simple. A bacterium’s environment is too complex for its genes to be controlled by one signal. Other factors besides lactose affect the expression of the *lac* genes, such as the availability of glucose. Glucose, metabolized directly by glycolysis, is the preferred energy source in *E. coli*. Other sugars can serve as the main or sole nutrient, but extra enzymatic steps are required to prepare them for entry into glycolysis, necessitating the synthesis of additional enzymes. Clearly, expressing the genes for proteins that metabolize sugars such as lactose or arabinose is wasteful when glucose is abundant.

What happens to the expression of the *lac* operon when both glucose and lactose are present? A regulatory mechanism known as **catabolite repression** restricts expression of the genes required for catabolism of lactose, arabinose, and other sugars in the presence of glucose, even when these secondary sugars are also present. The effect of glucose is mediated by cAMP, as a coactivator, and an activator protein known as **cAMP receptor protein**, or **CRP** (the protein is sometimes

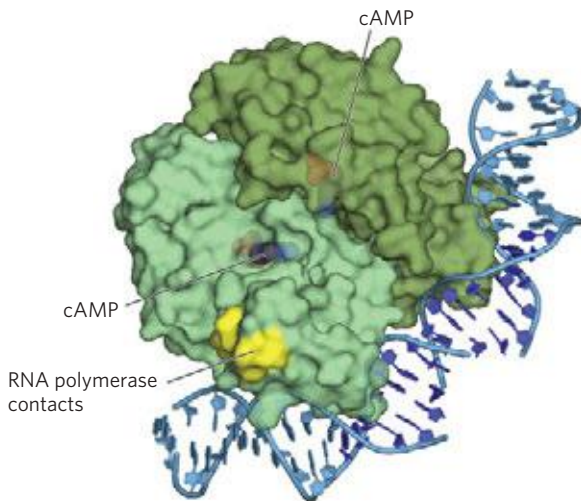


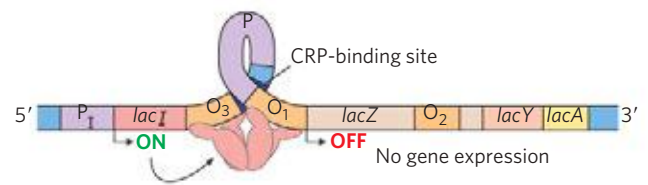
FIGURE 28-16 CRP homodimer with bound cAMP. (PDB ID 1RUN) Note the bending of the DNA around the protein. The region that interacts with RNA polymerase is indicated.

called CAP, for catabolite gene *activator protein*). CRP is a homodimer (subunit M_r 22,000) with binding sites for DNA and cAMP. Binding is mediated by a helix-turn-helix motif in the protein's DNA-binding domain (**Fig. 28-16**). When glucose is absent, CRP-cAMP binds to a site near the *lac* promoter (**Fig. 28-17**) and stimulates RNA transcription 50-fold. CRP-cAMP is therefore a positive regulatory element responsive to glucose levels, whereas the Lac repressor is a negative regulatory element responsive to lactose. The two act in concert. CRP-cAMP has little effect on the *lac* operon when the Lac repressor is blocking transcription, and dissociation of the repressor from the *lac* operator has little effect on transcription of the *lac* operon unless CRP-cAMP is present to facilitate transcription; when CRP is not bound, the wild-type *lac* promoter is a relatively weak promoter (Fig. 28-17a, c). The open complex of RNA polymerase and the promoter (see Fig. 26-6) does not form readily unless CRP-cAMP is present. CRP interacts directly with RNA polymerase (at the region shown in Fig. 28-16) through the polymerase's α subunit.

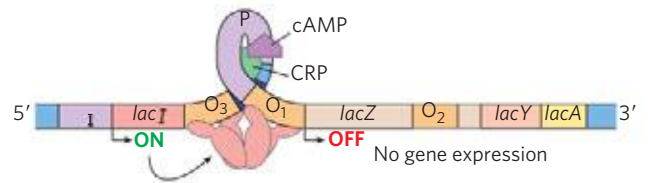
The effect of glucose on CRP is mediated by the cAMP interaction (Fig. 28-17). CRP binds to DNA most avidly when cAMP concentrations are high. In the presence of glucose, the synthesis of cAMP is inhibited and efflux of cAMP from the cell is stimulated. As [cAMP] declines, CRP binding to DNA declines, thereby decreasing the expression of the *lac* operon. Strong induction of the *lac* operon therefore requires both lactose (to inactivate the *lac* repressor) and a lowered concentration of glucose (to trigger an increase in [cAMP] and increased binding of cAMP to CRP).

CRP and cAMP are involved in the coordinated regulation of many operons, primarily those that encode enzymes for the metabolism of secondary sugars such as lactose and arabinose. A network of operons with a

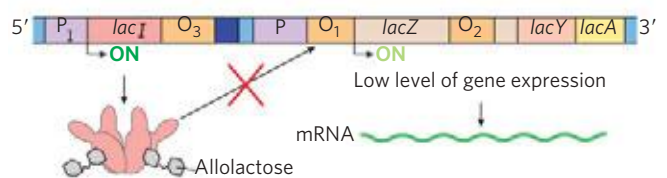
(a) Glucose high, cAMP low, lactose absent



(b) Glucose low, cAMP high, lactose absent



(c) Glucose high, cAMP low, lactose present



(d) Glucose low, cAMP high, lactose present

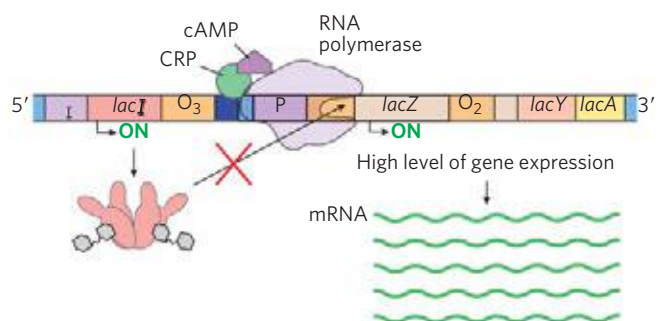


FIGURE 28-17 Positive regulation of the *lac* operon by CRP. The binding site for CRP-cAMP is near the promoter. The combined effects of glucose and lactose availability on *lac* operon expression are shown. When lactose is absent, the repressor binds to the operator and prevents transcription of the *lac* genes. It does not matter whether glucose is **(a)** present or **(b)** absent. **(c)** If lactose is present, the repressor dissociates from the operator. However, if glucose is also available, low cAMP levels prevent CRP-cAMP formation and DNA binding. RNA polymerase may occasionally bind and initiate transcription, resulting in a very low level of *lac* gene transcription. **(d)** When lactose is present and glucose levels are low, cAMP levels rise. The CRP-cAMP complex forms and facilitates robust binding of RNA polymerase to the *lac* promoter and high levels of transcription.

common regulator is called a **regulon**. This arrangement, which allows for coordinated shifts in cellular functions that can require the action of hundreds of genes, is a major theme in the regulated expression of dispersed networks of genes in eukaryotes. Other bacterial regulons include the heat shock gene system that

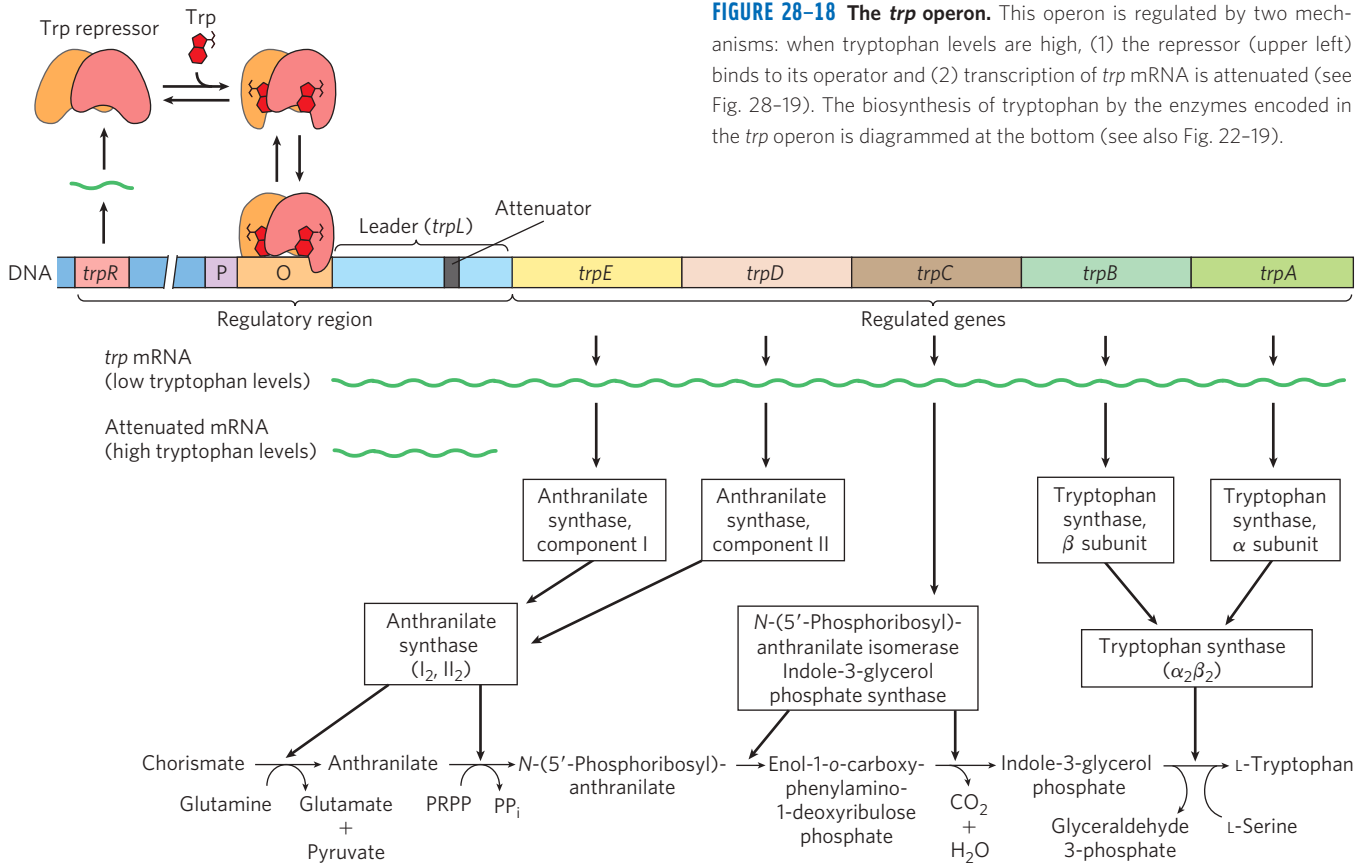


FIGURE 28-18 The *trp* operon. This operon is regulated by two mechanisms: when tryptophan levels are high, (1) the repressor (upper left) binds to its operator and (2) transcription of *trp* mRNA is attenuated (see Fig. 28-19). The biosynthesis of tryptophan by the enzymes encoded in the *trp* operon is diagrammed at the bottom (see also Fig. 22-19).

responds to changes in temperature (p. 1061) and the genes induced in *E. coli* as part of the SOS response to DNA damage, described later.

Many Genes for Amino Acid Biosynthetic Enzymes Are Regulated by Transcription Attenuation

The 20 common amino acids are required in large amounts for protein synthesis, and *E. coli* can synthesize all of them. The genes for the enzymes needed to synthesize a given amino acid are generally clustered in an operon and are expressed whenever existing supplies of that amino acid are inadequate for cellular requirements. When the amino acid is abundant, the biosynthetic enzymes are not needed and the operon is repressed.

The *E. coli* tryptophan (*trp*) operon (**Fig. 28-18**) includes five genes for the enzymes required to convert chorismate to tryptophan. Note that two of the enzymes catalyze more than one step in the pathway. The mRNA from the *trp* operon has a half-life of only about 3 min, allowing the cell to respond rapidly to changing needs for this amino acid. The Trp repressor is a homodimer. When tryptophan is abundant, it binds to the Trp repressor, causing a conformational change that permits the repressor to bind to the *trp* operator and inhibit expression of the *trp* operon. The *trp* operator site overlaps the promoter, so binding of the repressor blocks binding of RNA polymerase.

Once again, this simple on/off circuit mediated by a repressor is not the entire regulatory story. Different cellular concentrations of tryptophan can vary the rate of synthesis of the biosynthetic enzymes over a 700-fold range. Once repression is lifted and transcription begins, the rate of transcription is fine-tuned to cellular tryptophan requirements by a second regulatory process, called **transcription attenuation**, in which transcription is initiated normally but is abruptly halted *before* the operon genes are transcribed. The frequency with which transcription is attenuated is regulated by the availability of tryptophan and relies on the very close coupling of transcription and translation in bacteria.

The *trp* operon attenuation mechanism uses signals encoded in four sequences within a 162 nucleotide **leader** region at the 5' end of the mRNA, preceding the initiation codon of the first gene (**Fig. 28-19a**). Within the leader lies a region known as the **attenuator**, made up of sequences 3 and 4. These sequences base-pair to form a G≡C-rich stem-and-loop structure closely followed by a series of U residues. The attenuator structure acts as a transcription terminator (**Fig. 28-19b**). Sequence 2 is an alternative complement for sequence 3 (**Fig. 28-19c**). If sequences 2 and 3 base-pair, the attenuator structure cannot form and transcription continues into the *trp* biosynthetic genes; the loop formed by the pairing of sequences 2 and 3 does not obstruct transcription.

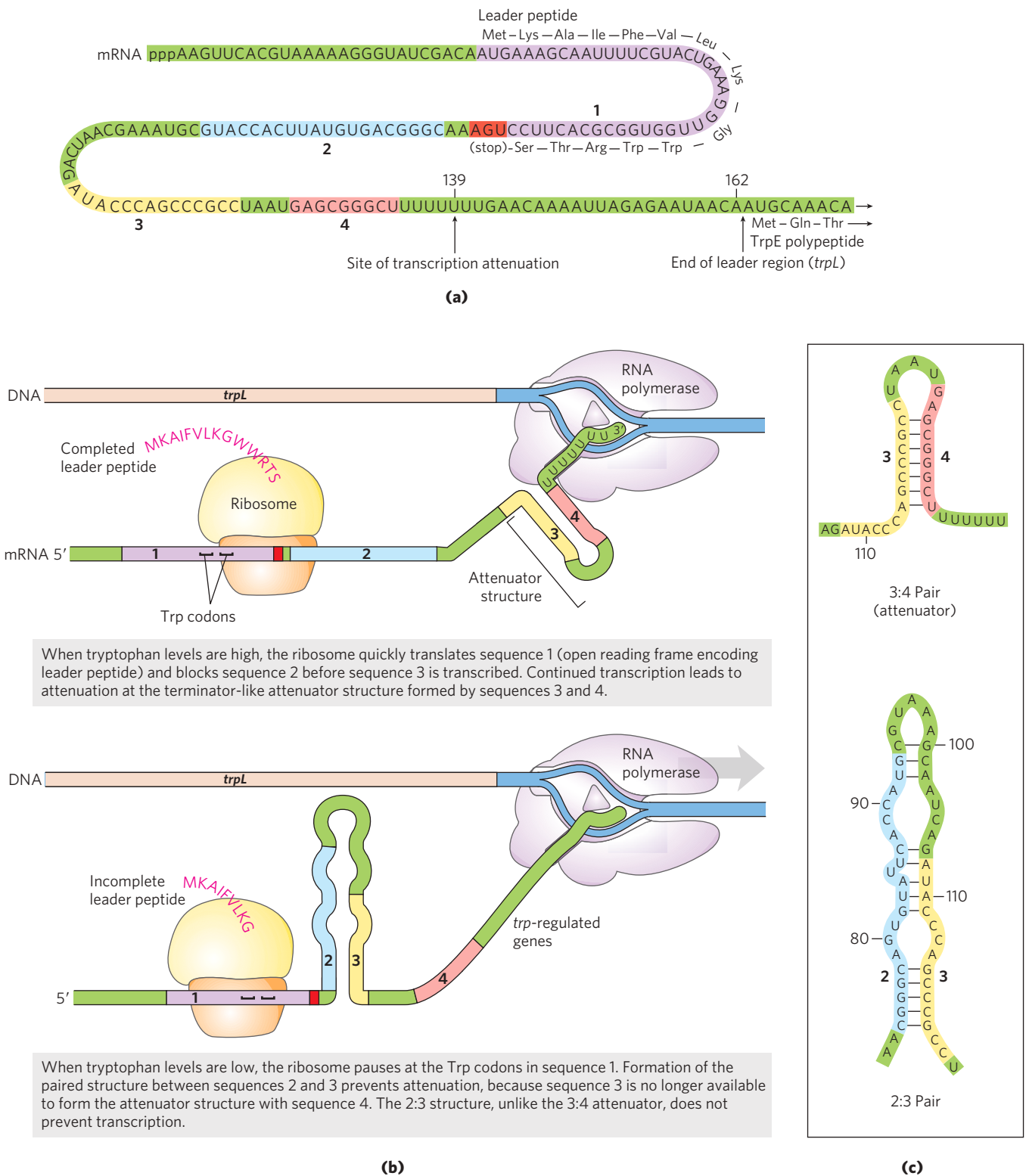


FIGURE 28-19 Transcriptional attenuation in the *trp* operon. Transcription is initiated at the beginning of the 162 nucleotide mRNA leader encoded by a DNA region called *trpL* (see Fig. 28-18). A regulatory mechanism determines whether transcription is attenuated at the end of the leader or continues into the structural genes. **(a)** The *trp* mRNA leader (*trpL*). The attenuation mechanism in the *trp* operon involves sequences 1 to 4 (highlighted). **(b)** Sequence 1 encodes a small peptide, the leader peptide, containing two Trp residues (W); it is translated immediately after transcription begins. Sequences 2 and 3 are complementary, as are sequences

3 and 4. The attenuator structure forms by the pairing of sequences 3 and 4 (top). Its structure and function are similar to those of a transcription terminator (see Fig. 26-7a). Pairing of sequences 2 and 3 (bottom) prevents the attenuator structure from forming. Note that the leader peptide has no other cellular function. Translation of its open reading frame has a purely regulatory role that determines which complementary sequences (2 and 3 or 3 and 4) are paired. **(c)** Base-pairing schemes for the complementary regions of the *trp* mRNA leader.

Regulatory sequence 1 is crucial for a tryptophan-sensitive mechanism that determines whether sequence 3 pairs with sequence 2 (allowing transcription to continue) or with sequence 4 (attenuating transcription). Formation of the attenuator stem-and-loop structure depends on events that occur during *translation* of regulatory sequence 1, which encodes a leader peptide (so called because it is encoded by the leader region of the mRNA) of 14 amino acids, two of which are Trp residues. The leader peptide has no other known cellular function; its synthesis is simply an operon regulatory device. This peptide is translated immediately after it is transcribed, by a ribosome that follows closely behind RNA polymerase as transcription proceeds.

When tryptophan concentrations are high, concentrations of charged tryptophan tRNA (Trp-tRNA^{Trp}) are also high. This allows translation to proceed rapidly past the two Trp codons of sequence 1 and into sequence 2, before sequence 3 is synthesized by RNA polymerase. In this situation, sequence 2 is covered by the ribosome and unavailable for pairing to sequence 3 when sequence 3 is synthesized; the attenuator structure (sequences 3 and 4) forms and transcription halts (Fig. 28–19b, top). When tryptophan concentrations are low, however, the ribosome stalls at the two Trp codons in sequence 1 because charged tRNA^{Trp} is less available. Sequence 2 remains free while sequence 3 is synthesized, allowing these two sequences to base-pair and permitting transcription to proceed (Fig. 28–19b, bottom). In this way, the proportion of transcripts that are attenuated declines as tryptophan concentration declines.

Many other amino acid biosynthetic operons use a similar attenuation strategy to fine-tune biosynthetic enzymes to meet the prevailing cellular requirements. The 15 amino acid leader peptide produced by the *phe* operon contains seven Phe residues. The *leu* operon leader peptide has four contiguous Leu residues. The leader peptide for the *his* operon contains seven contiguous His residues. In fact, in the *his* operon and a number of others, attenuation is sufficiently sensitive to be the *only* regulatory mechanism.

Induction of the SOS Response Requires Destruction of Repressor Proteins

Extensive DNA damage in the bacterial chromosome triggers the induction of many distantly located genes. This response, called the SOS response (p. 1035), provides another good example of coordinated gene regulation. Many of the induced genes are involved in DNA repair (see Table 25–6). The key regulatory proteins are the RecA protein and the LexA repressor.

The LexA repressor (M_r 22,700) inhibits transcription of all the SOS genes (Fig. 28–20), and induction of the SOS response requires removal of LexA. This is not a simple dissociation from DNA in response to binding of a small molecule, as in the regulation of the *lac* operon described above. Instead, the LexA repressor is inactivated when it catalyzes its own cleavage at a specific Ala–Gly peptide bond, producing two roughly equal protein fragments. At physiological pH, this autocleavage reaction requires the RecA protein. RecA is not a

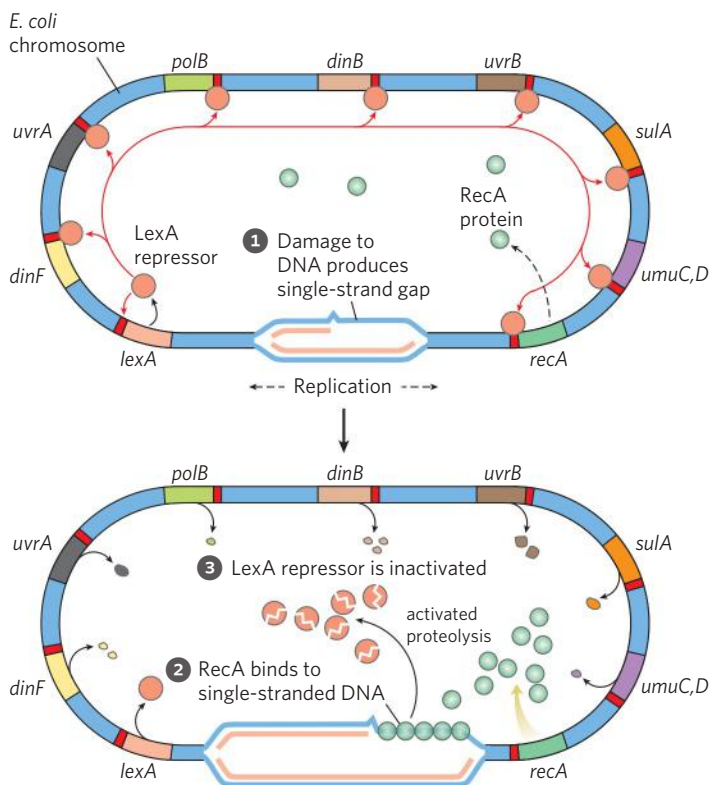


FIGURE 28–20 SOS response in *E. coli*. See Table 25–6 for the functions of many of these proteins. The LexA protein is the repressor in this system, which has an operator site (red) near each gene. Because the *recA* gene is not entirely repressed by the LexA repressor, the normal cell contains about 1,000 RecA monomers. ① When DNA is extensively damaged (such as by UV light), DNA replication is halted and the number of single-strand gaps in the DNA increases. ② RecA protein binds to this damaged, single-stranded DNA, activating the protein's coprotease activity. ③ While bound to DNA, the RecA protein facilitates cleavage and inactivation of the LexA repressor. When the repressor is inactivated, the SOS genes, including *recA*, are induced; RecA levels increase 50- to 100-fold.

protease in the classical sense, but its interaction with LexA facilitates the repressor's self-cleavage reaction. This function of RecA is sometimes called a co-protease activity.

The RecA protein provides the functional link between the biological signal (DNA damage) and induction of the SOS genes. Heavy DNA damage leads to numerous single-strand gaps in the DNA, and only RecA that is bound to single-stranded DNA can facilitate cleavage of the LexA repressor (Fig. 28–20, bottom). Binding of RecA at the gaps eventually activates its co-protease activity, leading to cleavage of the LexA repressor and SOS induction.

During induction of the SOS response in a severely damaged cell, RecA also cleaves and thus inactivates the repressors that otherwise allow propagation of certain viruses in a dormant lysogenic state within the bacterial host. This provides a remarkable illustration of evolutionary adaptation. These repressors, like LexA, also undergo self-cleavage at a specific Ala–Gly peptide bond, so induction of the SOS response permits replication of the virus and lysis of the cell, releasing new viral particles. Thus the bacteriophage can make a hasty exit from a compromised bacterial host cell.

Synthesis of Ribosomal Proteins Is Coordinated with rRNA Synthesis

In bacteria, an increased cellular demand for protein synthesis is met by increasing the number of ribosomes rather than altering the activity of individual ribosomes. In general, the number of ribosomes increases as the cellular growth rate increases. At high growth rates, ribosomes make up approximately 45% of the cell's dry weight. The proportion of cellular resources devoted to making ribosomes is so large, and the function of ribosomes so important, that cells must coordinate the synthesis of the ribosomal components: the ribosomal proteins (r-proteins) and RNAs (rRNAs). This regulation is distinct from the mechanisms described so far, because it occurs largely at the level of *translation*.

The 52 genes that encode the r-proteins occur in at least 20 operons, each with 1 to 11 genes. Some of these operons also contain the genes for the subunits of DNA primase (see Fig. 25–12), RNA polymerase (see Fig. 26–4), and protein synthesis elongation factors (see Fig. 27–29)—revealing the close coupling of replication, transcription, and protein synthesis during cell growth.

The r-protein operons are regulated primarily through a translational feedback mechanism. One r-protein encoded by each operon also functions as a **translational repressor**, which binds to the mRNA transcribed from that operon and blocks translation of all the genes the messenger encodes (Fig. 28–21). In general, the r-protein that plays the role of repressor also binds directly to an rRNA. Each translational repressor r-protein binds with higher affinity to the appropriate rRNA than to its mRNA, so the mRNA is bound and

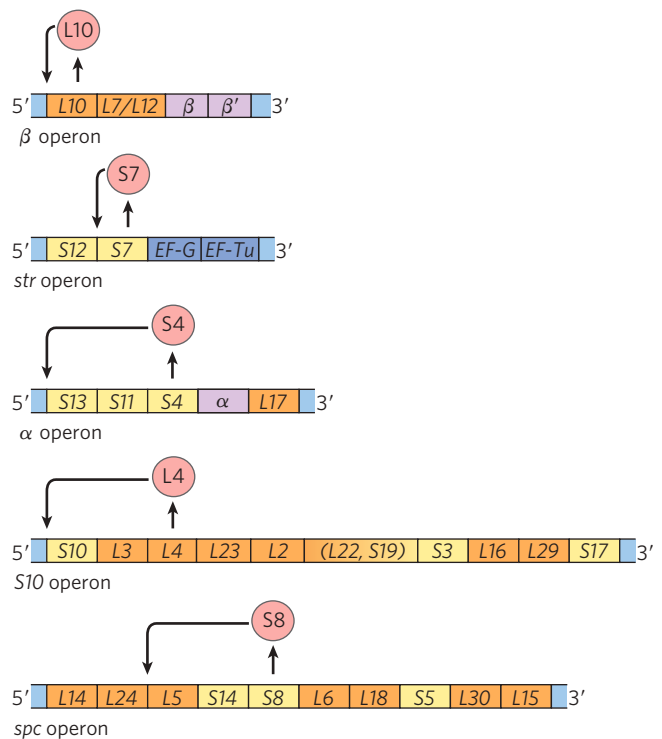


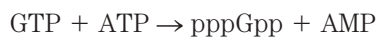
FIGURE 28–21 Translational feedback in some ribosomal protein operons.

The r-proteins that act as translational repressors are shaded light red. Each translational repressor blocks the translation of all genes in that operon by binding to the indicated site on the mRNA. Genes that encode subunits of RNA polymerase are shown in purple; genes that encode elongation factors are blue. The r-proteins of the large (50S) ribosomal subunit are designated L1 to L34; those of the small (30S) subunit, S1 to S21.

translation repressed only when the level of the r-protein exceeds that of the rRNA. This ensures that translation of the mRNAs encoding r-proteins is repressed only when synthesis of these r-proteins exceeds that needed to make functional ribosomes. In this way, the rate of r-protein synthesis is kept in balance with rRNA availability.

The mRNA-binding site for the translational repressor is near the translational start site of one of the genes in the operon, usually the first gene (Fig. 28–21). In other operons this would affect only that one gene, because in bacterial polycistronic mRNAs most genes have independent translation signals. In the r-protein operons, however, the translation of one gene depends on the translation of all the others. The mechanism of this translational coupling is not yet understood in detail. However, in some cases the translation of multiple genes seems to be blocked by folding of the mRNA into an elaborate three-dimensional structure that is stabilized both by internal base-pairing (as in Fig. 8–24) and by binding of the translational repressor protein. When the translational repressor is absent, ribosome binding and translation of one or more of the genes disrupts the folded structure of the mRNA and allows all the genes to be translated.

Because the synthesis of r-proteins is coordinated with the available rRNA, the regulation of ribosome production reflects the regulation of rRNA synthesis. In *E. coli*, rRNA synthesis from the seven rRNA operons responds to cellular growth rate and to changes in the availability of crucial nutrients, particularly amino acids. The regulation coordinated with amino acid concentrations is known as the **stringent response** (Fig. 28–22). When amino acid concentrations are low, rRNA synthesis is halted. Amino acid starvation leads to the binding of uncharged tRNAs to the ribosomal A site; this triggers a sequence of events that begins with the binding of an enzyme called **stringent factor** (RelA protein) to the ribosome. When bound to the ribosome, stringent factor catalyzes formation of the unusual nucleotide guanosine tetraphosphate (ppGpp; see Fig. 8–39); it adds pyrophosphate to the 3' position of GTP, in the reaction



then a phosphohydrolase cleaves off one phosphate to form ppGpp. The abrupt rise in ppGpp level in response to amino acid starvation results in a great reduction in rRNA synthesis, mediated at least in part by the binding of ppGpp to RNA polymerase.

The nucleotide ppGpp, along with cAMP, belongs to a class of modified nucleotides that act as cellular second

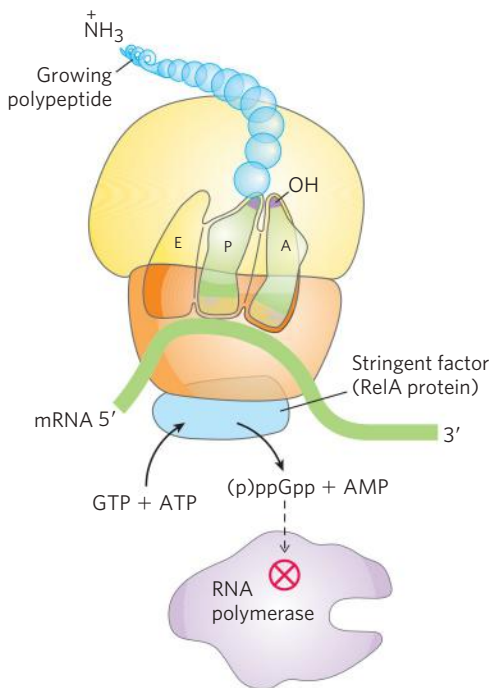


FIGURE 28–22 Stringent response in *E. coli*. This response to amino acid starvation is triggered by binding of an uncharged tRNA in the ribosomal A site. A protein called stringent factor binds to the ribosome and catalyzes the synthesis of pppGpp, which is converted by a phosphohydrolase to ppGpp. The signal ppGpp reduces transcription of some genes and increases that of others, in part by binding to the β subunit of RNA polymerase and altering the enzyme's promoter specificity. Synthesis of rRNA is reduced when ppGpp levels increase.

messengers (p. 308). In *E. coli*, these two nucleotides serve as starvation signals; they cause large changes in cellular metabolism by increasing or decreasing the transcription of hundreds of genes. In eukaryotic cells, similar nucleotide second messengers also have multiple regulatory functions. The coordination of cellular metabolism with cell growth is highly complex, and further regulatory mechanisms undoubtedly remain to be discovered.

The Function of Some mRNAs Is Regulated by Small RNAs in Cis or in Trans

As described throughout this chapter, proteins play an important and well-documented role in regulating gene expression. But RNA also has a crucial role—one that is becoming better recognized as more examples of regulatory RNAs are discovered. Once an mRNA is synthesized, its functions can be controlled by RNA-binding proteins, as seen for the r-protein operons just described, or by an RNA. A separate RNA molecule may bind to the mRNA “in trans” and affect its activity. Alternatively, a portion of the mRNA itself may regulate its own function. When part of a molecule affects the function of another part of the same molecule, it is said to act “in cis.”

A well-characterized example of RNA regulation in trans is seen in the regulation of the mRNA of the gene *rpoS* (RNA polymerase sigma factor), which encodes σ^S , one of the seven *E. coli* sigma factors (see Table 26–1). The cell uses this specificity factor in certain stress situations, such as when it enters the stationary phase (a state of no growth, necessitated by lack of nutrients) and σ^S is needed to transcribe large numbers of stress response genes. The σ^S mRNA is present at low levels under most conditions but is not translated, because a large hairpin structure upstream of the coding region inhibits ribosome binding (Fig. 28–23). Under certain stress conditions, one or both of two small special-function RNAs, DsrA (downstream region A) and RprA (*Rpos* regulator RNA A), are induced. Both can pair with one strand of the hairpin in the σ^S mRNA, disrupting the hairpin and thus allowing translation of *rpoS*. Another small RNA, OxyS (oxidative stress gene S), is induced under conditions of oxidative stress and inhibits the translation of *rpoS*, probably by pairing with and blocking the ribosome-binding site on the mRNA. OxyS is expressed as part of a system that responds to a different type of stress (oxidative damage) than does *rpoS*, and its task is to prevent expression of unneeded repair pathways. DsrA, RprA, and OxyS are all relatively small bacterial RNA molecules (less than 300 nucleotides), designated sRNAs (*s* for small; there are of course other “small” RNAs with other designations in eukaryotes). All require for their function a protein called Hfq, an RNA chaperone that facilitates RNA-RNA pairing. The known bacterial genes regulated in this way are few in number, just a few dozen in a typical bacterial species. However, these examples provide

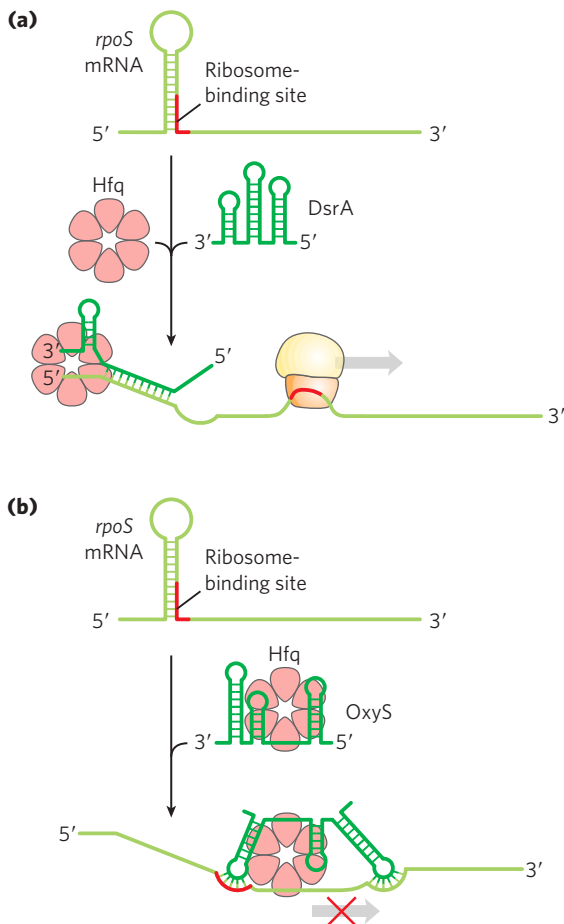


FIGURE 28-23 Regulation of bacterial mRNA function in trans by sRNAs.

Several sRNAs (small RNAs)—DsrA, RprA, and OxyS—are involved in regulation of the *rpoS* gene. All require the protein Hfq, an RNA chaperone that facilitates RNA-RNA pairing. Hfq has a toroid structure, with a pore in the center. **(a)** DsrA promotes translation by pairing with one strand of a stem-loop structure that otherwise blocks the ribosome-binding site. RprA (not shown) acts in a similar way. **(b)** OxyS blocks translation by pairing with the ribosome-binding site.

good model systems for understanding patterns present in the more complex and numerous examples of RNA-mediated regulation in eukaryotes.

Regulation in cis involves a class of RNA structures known as **riboswitches**. As described in Box 26-3, aptamers are RNA molecules, generated in vitro, that are capable of specific binding to a particular ligand. As one might expect, such ligand-binding RNA domains are also present in nature—in riboswitches—in a significant number of bacterial mRNAs (and even in some eukaryotic mRNAs). These natural aptamers are structured domains found in untranslated regions at the 5' ends of certain bacterial mRNAs. Binding of an mRNA's riboswitch to its appropriate ligand results in a conformational change in the mRNA, and transcription is inhibited by stabilization of a premature transcription termination structure, or translation is inhibited (in cis) by occlusion of the ribosome-binding site (**Fig. 28-24**). In most cases, the riboswitch acts in a kind of feedback

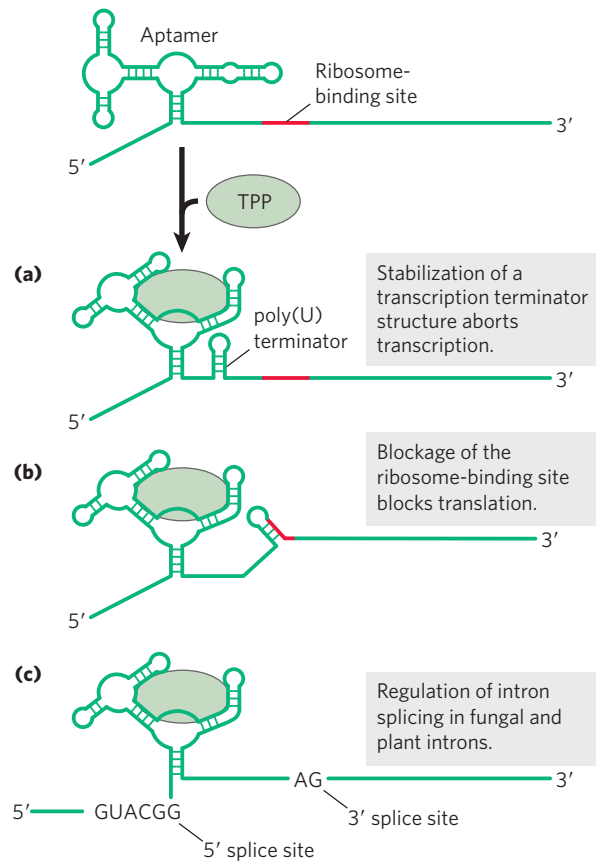



FIGURE 28-24 Regulation of bacterial mRNA function in cis by riboswitches. The known modes of action are illustrated by several different riboswitches based on a widespread natural aptamer that binds thiamine pyrophosphate. TPP binding to the aptamer leads to a conformational change that produces the varied results illustrated in parts **(a)**, **(b)**, and **(c)** in the different systems in which the aptamer is utilized.

loop. Most genes regulated in this way are involved in the synthesis or transport of the ligand that is bound by the riboswitch; thus, when the ligand is present in high concentrations, the riboswitch inhibits expression of the genes needed to replenish this ligand.

Each riboswitch binds only one ligand. Distinct riboswitches have been detected that respond to more than a dozen different ligands, including thiamine pyrophosphate (TPP, vitamin B₁), cobalamin (vitamin B₁₂), flavin mononucleotide, lysine, *S*-adenosylmethionine (adoMet), purines, *N*-acetylglucosamine 6-phosphate, and glycine. It is likely that many more remain to be discovered. The riboswitch that responds to TPP seems to be the most widespread; it is found in many bacteria, fungi, and some plants. The bacterial TPP riboswitch inhibits translation in some species and induces premature transcription termination in others (**Fig. 28-24**). The eukaryotic TPP riboswitch is found in the introns of certain genes and modulates the alternative splicing of those genes (see **Fig. 26-21**). It is not yet clear how common riboswitches are. However, estimates suggest that more than 4% of the genes of *Bacillus subtilis* are regulated by riboswitches.

 As riboswitches become better understood, researchers are finding medical applications. For example, most of the riboswitches described to date, including the one that responds to adoMet, have been found only in bacteria. A drug that bound to and activated the adoMet riboswitch would shut down the genes encoding the enzymes that synthesize and transport adoMet, effectively starving the bacterial cells of this essential cofactor. Drugs of this type are being sought for use as a new class of antibiotics. ■

The pace of discovery of functional RNAs shows no signs of abatement and continues to enrich the hypothesis that RNA played a special role in the evolution of life (Chapter 26). The sRNAs and riboswitches, like ribozymes and ribosomes, may be vestiges of an RNA world obscured by time but persisting as a rich array of biological devices still functioning in the extant biosphere. The laboratory selection of aptamers and ribozymes with novel ligand-binding and enzymatic functions (see Box 26–3) tells us that the RNA-based activities necessary for a viable RNA world are possible. Discovery of many of the same RNA functions in living organisms tells us that key components for RNA-based metabolism do exist. For example, the natural aptamers of riboswitches may be derived from RNAs that, billions of years ago, bound to cofactors needed to promote the enzymatic processes required for metabolism in the RNA world.

Some Genes Are Regulated by Genetic Recombination

We turn now to another mode of bacterial gene regulation, at the level of DNA rearrangement—recombination. *Salmonella typhimurium*, which inhabits the mammalian intestine, moves by rotating the flagella on its cell surface (Fig. 28–25). The many copies of the protein



FIGURE 28–25 *Salmonella typhimurium*, with flagella evident.

flagellin (M_r 53,000) that make up the flagella are prominent targets of mammalian immune systems. But *Salmonella* cells have a mechanism that evades the immune response: they switch between two distinct flagellin proteins (FljB and FliC) roughly once every 1,000 generations, using a process called **phase variation**.

The switch is accomplished by periodic inversion of a segment of DNA containing the promoter for a flagellin gene. The inversion is a site-specific recombination reaction (see Fig. 25–37) mediated by the Hin recombinase at specific 14 bp sequences (*hix* sequences) at either end of the DNA segment. When the DNA segment is in one orientation, the gene for FljB flagellin and the gene encoding a repressor (FljA) are expressed (Fig. 28–26a); the repressor shuts down expression of the gene for FliC flagellin. When the DNA segment is inverted (Fig. 28–26b), the *fljA* and *fljB* genes are no longer transcribed, and the *fliC* gene is induced as the repressor becomes depleted. The Hin recombinase, encoded

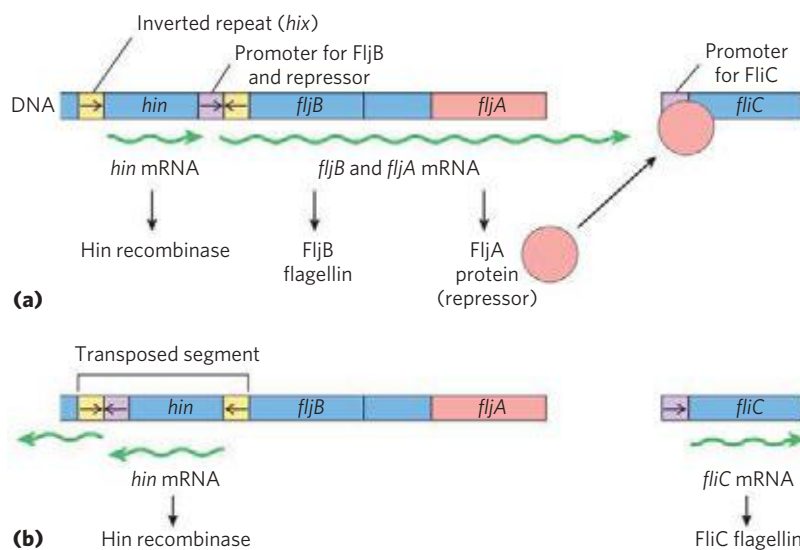


FIGURE 28–26 Regulation of flagellin genes in *Salmonella*: phase variation. The products of genes *fliC* and *fljB* are different flagellins. The *hin* gene encodes the recombinase that catalyzes inversion of the DNA segment containing the *fljB* promoter and the *hin* gene. The recombination sites (inverted repeats) are called *hix* (yellow). (a) In one orientation, *fljB* is expressed along

with a repressor protein (product of the *fljA* gene) that represses transcription of the *fliC* gene. (b) In the opposite orientation, only the *fliC* gene is expressed; the *fljA* and *fljB* genes cannot be transcribed. The interconversion between these two states, known as phase variation, also requires two other nonspecific DNA-binding proteins (not shown), HU and FIS.

TABLE 28–1 Examples of Gene Regulation by Recombination

System	Recombinase/ recombination site	Type of recombination	Function
Phase variation (<i>Salmonella</i>)	Hin/ <i>hix</i>	Site-specific	Alternative expression of two flagellin genes allows evasion of host immune response.
Host range (bacteriophage μ)	Gin/ <i>gix</i>	Site-specific	Alternative expression of two sets of tail fiber genes affects host range.
Mating-type switch (yeast)	HO endonuclease, RAD52 protein, other proteins/ <i>MAT</i>	Nonreciprocal gene conversion*	Alternative expression of two mating types of yeast, α and α , creates cells of different mating types that can mate and undergo meiosis.
Antigenic variation (trypanosomes) [†]	Varies	Nonreciprocal gene conversion*	Successive expression of different genes encoding the variable surface glycoproteins (VSGs) allows evasion of host immune response.

*In nonreciprocal gene conversion (a class of recombination events not discussed in Chapter 25), genetic information is moved from one part of the genome (where it is silent) to another (where it is expressed). The reaction is similar to replicative transposition (see Fig. 25–41).

[†]Trypanosomes cause African sleeping sickness and other diseases (see Box 6–3). The outer surface of a trypanosome is made up of multiple copies of a single VSG, the major surface antigen. A cell can change surface antigens to more than 100 different forms, precluding an effective defense by the host immune system.

by the *hin* gene in the DNA segment that undergoes inversion, is expressed when the DNA segment is in either orientation, so the cell can always switch from one state to the other.

This type of regulatory mechanism has the advantage of being absolute: gene expression is impossible when the gene is physically separated from its promoter (note the position of the *fljB* promoter in Fig. 28–26b). An absolute on/off switch may be important in this system (even though it affects only one of the two flagellin genes) because a flagellum with just one copy of the wrong flagellin might be vulnerable to host antibodies against that protein. The *Salmonella* system is by no means unique. Similar regulatory systems occur in some other bacteria and in some bacteriophages, and recombination systems with similar functions have been found in eukaryotes (Table 28–1). Gene regulation by DNA rearrangements that move genes and/or promoters is particularly common in pathogens that benefit by changing their host range or by changing their surface proteins, thereby staying ahead of host immune systems.

SUMMARY 28.2 Regulation of Gene Expression in Bacteria

▶ In addition to repression by the Lac repressor, the *E. coli lac* operon undergoes positive regulation by the cAMP receptor protein (CRP). When [glucose] is low, [cAMP] is high and CRP-cAMP binds to a specific site on the DNA, stimulating transcription of the *lac* operon and production of lactose-metabolizing enzymes. The presence of

glucose depresses [cAMP], decreasing expression of *lac* and other genes involved in metabolism of secondary sugars. A group of coordinately regulated operons is referred to as a regulon.

- ▶ Operons that produce the enzymes of amino acid synthesis have a regulatory circuit called attenuation, which uses a transcription termination site (the attenuator) in the mRNA. Formation of the attenuator is modulated by a mechanism that couples transcription and translation while responding to small changes in amino acid concentration.
- ▶ In the SOS system, multiple unlinked genes repressed by a single repressor are induced simultaneously when DNA damage triggers RecA protein-facilitated autocatalytic proteolysis of the repressor.
- ▶ In the synthesis of ribosomal proteins, one protein in each r-protein operon acts as a translational repressor. The mRNA is bound by the repressor, and translation is blocked only when the r-protein is present in excess of available rRNA.
- ▶ Posttranscriptional regulation of some mRNAs is mediated by sRNAs that act in trans or by riboswitches, part of the mRNA structure itself, that act in cis.
- ▶ Some genes are regulated by genetic recombination processes that move promoters relative to the genes being regulated. Regulation can also take place at the level of translation.

28.3 Regulation of Gene Expression in Eukaryotes

Initiation of transcription is a crucial regulation point for gene expression in all organisms. Although eukaryotes and bacteria use some of the same regulatory mechanisms, the regulation of transcription in the two systems is fundamentally different.

We can define a transcriptional ground state as the inherent activity of promoters and transcriptional machinery *in vivo* in the absence of regulatory sequences. In bacteria, RNA polymerase generally has access to every promoter and can bind and initiate transcription at some level of efficiency in the absence of activators or repressors; the transcriptional ground state is therefore nonrestrictive. In eukaryotes, however, strong promoters are generally inactive *in vivo* in the absence of regulatory proteins; that is, the transcriptional ground state is restrictive. This fundamental difference gives rise to at least four important features that distinguish the regulation of gene expression in eukaryotes from that in bacteria.

First, access to eukaryotic promoters is restricted by the structure of chromatin, and activation of transcription is associated with many changes in chromatin structure in the transcribed region. Second, although eukaryotic cells have both positive and negative regulatory mechanisms, positive mechanisms predominate in all systems characterized so far. Thus, given that the transcriptional ground state is restrictive, virtually every eukaryotic gene requires activation in order to be transcribed. Third, eukaryotic cells have larger, more complex multimeric regulatory proteins than do bacteria. Finally, transcription in the eukaryotic nucleus is separated from translation in the cytoplasm in both space and time.

The complexity of regulatory circuits in eukaryotic cells is extraordinary, as the following discussion shows. We conclude the section with an illustrated description of one of the most elaborate circuits: the regulatory cascade that controls development in fruit flies.

Transcriptionally Active Chromatin Is Structurally Distinct from Inactive Chromatin

The effects of chromosome structure on gene regulation in eukaryotes have no clear parallel in bacteria. In the eukaryotic cell cycle, interphase chromosomes appear, at first viewing, to be dispersed and amorphous (see Fig. 24–24). Nevertheless, several forms of chromatin can be found along these chromosomes. About 10% of the chromatin in a typical eukaryotic cell is in a more condensed form than the rest of the chromatin. This form, **heterochromatin**, is transcriptionally inactive. Heterochromatin is generally associated with particular chromosome structures—the centromeres, for example. The remaining, less condensed chromatin is called **euchromatin**.

Transcription of a eukaryotic gene is strongly repressed when its DNA is condensed within heterochromatin. Some, but not all, of the euchromatin is

transcriptionally active. Transcriptionally active chromosomal regions are distinguished from heterochromatin in at least three ways: the positioning of nucleosomes, the presence of histone variants, and the covalent modification of nucleosomes. These transcription-associated structural changes in chromatin are collectively called **chromatin remodeling**. The remodeling involves enzymes that promote these changes (Table 28–2).

Five known families of enzyme complexes actively reposition or displace nucleosomes, hydrolyzing ATP in the process. Three of these are particularly important in transcriptional activation (Table 28–2; see the table footnote for an explanation of the abbreviated names of the enzyme complexes described here). **SWI/SNF**, found in all eukaryotic cells, contains at least six core polypeptides that together remodel chromatin so that nucleosomes become more irregularly spaced. They also stimulate transcription factor binding. The complex includes a component called a bromodomain near the carboxyl terminus of the active ATPase subunit, which interacts with acetylated histone tails. SWI/SNF is not required for the transcription of every gene. **NURF**, a member of the ISW1 family, remodels chromatin in ways that complement and overlap the activity of SWI/SNF. These two enzyme complexes are crucial in preparing a region of chromatin for active transcription.

The third important protein family has a somewhat different role. Transcriptionally active chromatin tends to be deficient in histone H1, which binds to the linker DNA between nucleosome particles. These regions of chromatin are also enriched in the histone variants H3.3 and H2AZ (see Box 24–2). Alterations in histone content are again mediated by specialized enzymes and protein complexes. H2AZ deposition involves members of this third family of ATP-dependent remodeling enzymes, called **SWR1**.

The covalent modification of histones is altered dramatically within transcriptionally active chromatin. The core histones of nucleosome particles (H2A, H2B, H3, H4; see Fig. 24–26) are modified by methylation of Lys or Arg residues, phosphorylation of Ser or Thr residues, acetylation (see below), ubiquitination (see Fig. 27–47), or sumoylation. Each of the core histones has two distinct structural domains. A central domain is involved in histone-histone interaction and the wrapping of DNA around the nucleosome. A second, lysine-rich amino-terminal domain is generally positioned near the exterior of the assembled nucleosome particle; the covalent modifications occur at specific residues concentrated in this amino-terminal domain. The patterns of modification have led some researchers to propose the existence of a histone code, in which modification patterns are recognized by enzymes that alter the structure of chromatin. Indeed, some of the modifications are essential for interactions with proteins that play key roles in transcription.

The acetylation and methylation of histones figure prominently in the processes that activate chromatin

TABLE 28–2 Some Enzyme Complexes Catalyzing Chromatin Structural Changes Associated with Transcription

Enzyme complex*	Oligomeric structure (number of polypeptides)	Source	Activities
Histone modification			
GCN5-ADA2-ADA3	3	Yeast	GCN5 has type A HAT activity
SAGA/PCAF	>20	Eukaryotes	Includes GCN5-ADA2-ADA3; acetylates residues in H3 and H2B
NuA4	At least 12	Eukaryotes	Esa1 component has HAT activity; acetylates H4, H2A, and H2AZ
Histone movement/replacement enzymes that require ATP			
SWI/SNF	≥6; total M_r 2×10^6	Eukaryotes	Nucleosome remodeling; transcriptional activation
ISWI family	Varies	Eukaryotes	Nucleosome remodeling; transcriptional repression; transcriptional activation in some cases (NURF)
SWR1 family	~12	Eukaryotes	H2AZ deposition
Histone chaperones that do not require ATP			
HIRA	1	Eukaryotes	Deposition of H3.3 during transcription

*The abbreviations for eukaryotic genes and proteins are often more confusing or obscure than those used for bacteria. The complex of GCN5 (general control nonderepressible) and ADA (alteration/deficiency activation) proteins was discovered during investigation of the regulation of nitrogen metabolism genes in yeast. These proteins can be part of the larger SAGA complex (SPF, ADA2,3, GCN5, acetyltransferase) in yeasts. The equivalent of SAGA in humans is PCAF (p300/CBP-associated factor). NuA4 is nucleosome acetyltransferase of H4; Esa1 is essential SAS2-related acetyltransferase. SWI (switching) was discovered as a protein required for expression of certain genes involved in mating-type switching in yeast, and SNF (sucrose nonfermenting) as a factor for expression of the yeast gene for sucrose. Subsequent studies revealed multiple SWI and SNF proteins that acted in a complex. The SWI/SNF complex has a role in the expression of a wide range of genes and has been found in many eukaryotes, including humans. ISWI is imitation SWI; NURF, nuclear remodeling factor; SWR1, *Swi2/Snf2*-related ATPase 1; and HIRA, histone regulator A.

for transcription. During transcription, histone H3 is methylated (by specific histone methylases) at Lys⁴ in nucleosomes near the 5' end of the coding region and at Lys³⁶ throughout the coding region. These methylations facilitate the binding of **histone acetyltransferases (HATs)**, enzymes that acetylate particular Lys residues. Cytosolic (type B) HATs acetylate newly synthesized histones before the histones are imported into the nucleus. The subsequent assembly of the histones into chromatin after replication is facilitated by histone chaperones: CAF1 for H3 and H4 (see Box 24–2), and NAP1 for H2A and H2B.

Where chromatin is being activated for transcription, the nucleosomal histones are further acetylated by nuclear (type A) HATs. The acetylation of multiple Lys residues in the amino-terminal domains of histones H3 and H4 can reduce the affinity of the entire nucleosome for DNA. Acetylation of particular Lys residues is critical for the interaction of nucleosomes with other proteins. When transcription of a gene is no longer required, the extent of acetylation of nucleosomes in that vicinity is reduced by the activity of **histone deacetylases (HDACs)**, as part of a general gene-silencing process that restores the chromatin to a transcriptionally inactive

state. In addition to the removal of certain acetyl groups, new covalent modification of histones marks chromatin as transcriptionally inactive. For example, Lys⁹ of histone H3 is often methylated in heterochromatin.

The net effect of chromatin remodeling in the context of transcription is to make a segment of the chromosome more accessible and to “label” (chemically modify) it so as to facilitate the binding and activity of transcription factors that regulate expression of the gene or genes in that region.

Most Eukaryotic Promoters Are Positively Regulated

As already noted, eukaryotic RNA polymerases have little or no intrinsic affinity for their promoters; initiation of transcription is almost always dependent on the action of multiple activator proteins. One important reason for the apparent predominance of positive regulation seems obvious: the storage of DNA within chromatin effectively renders most promoters inaccessible, so genes are silent in the absence of other regulation. The structure of chromatin affects access to some promoters more than others, but repressors that bind to DNA so as to preclude access of RNA polymerase

(negative regulation) would often be simply redundant. Other factors must be at play in the use of positive regulation, and speculation generally centers around two: the large size of eukaryotic genomes and the greater efficiency of positive regulation.

First, nonspecific DNA binding of regulatory proteins becomes a more important problem in the much larger genomes of higher eukaryotes. And the chance that a single specific binding sequence will occur randomly at an inappropriate site also increases with genome size. Combinatorial control thus becomes important in a large genome (Fig. 28–27). Specificity for transcriptional activation can be improved if each of several positive-regulatory proteins must bind specific DNA sequences in order to activate a gene. The average number of regulatory sites for a gene in a multicellular organism is six, and genes that are regulated by a dozen such sites are common. The requirement for binding of several positive-regulatory proteins to specific DNA sequences vastly reduces the probability of the random occurrence of a functional juxtaposition of all the necessary binding sites. In addition, the actual number of regulatory proteins that must be encoded by a genome to regulate all of its genes can be reduced (Fig. 28–27). Thus, a new regulator is not needed for every gene, although regulation is complex enough in higher eukaryotes that regulator proteins may represent 5% to 10% of

all protein-encoding genes. In principle, a similar combinatorial strategy could be used by multiple negative-regulatory elements, but this brings us to the second reason for the use of positive regulation: it is simply more efficient. If the ~25,000 genes in the human genome were negatively regulated, each cell would have to synthesize, at all times, all of the different repressors in concentrations sufficient to permit specific binding to each “unwanted” gene. In positive regulation, most of the genes are usually inactive (that is, RNA polymerases do not bind to the promoters) and the cell synthesizes only the activator proteins needed to promote transcription of the subset of genes required in the cell at that time. These arguments notwithstanding, there are examples of negative regulation in eukaryotes, from yeasts to humans, as we shall see.

DNA-Binding Activators and Coactivators Facilitate Assembly of the General Transcription Factors

To continue our exploration of the regulation of gene expression in eukaryotes, we return to the interactions between promoters and RNA polymerase II (Pol II), the enzyme responsible for the synthesis of eukaryotic mRNAs. Although many (but not all) Pol II promoters include the TATA box and Inr (initiator) sequences, with their standard spacing (see Fig. 26–8), they vary

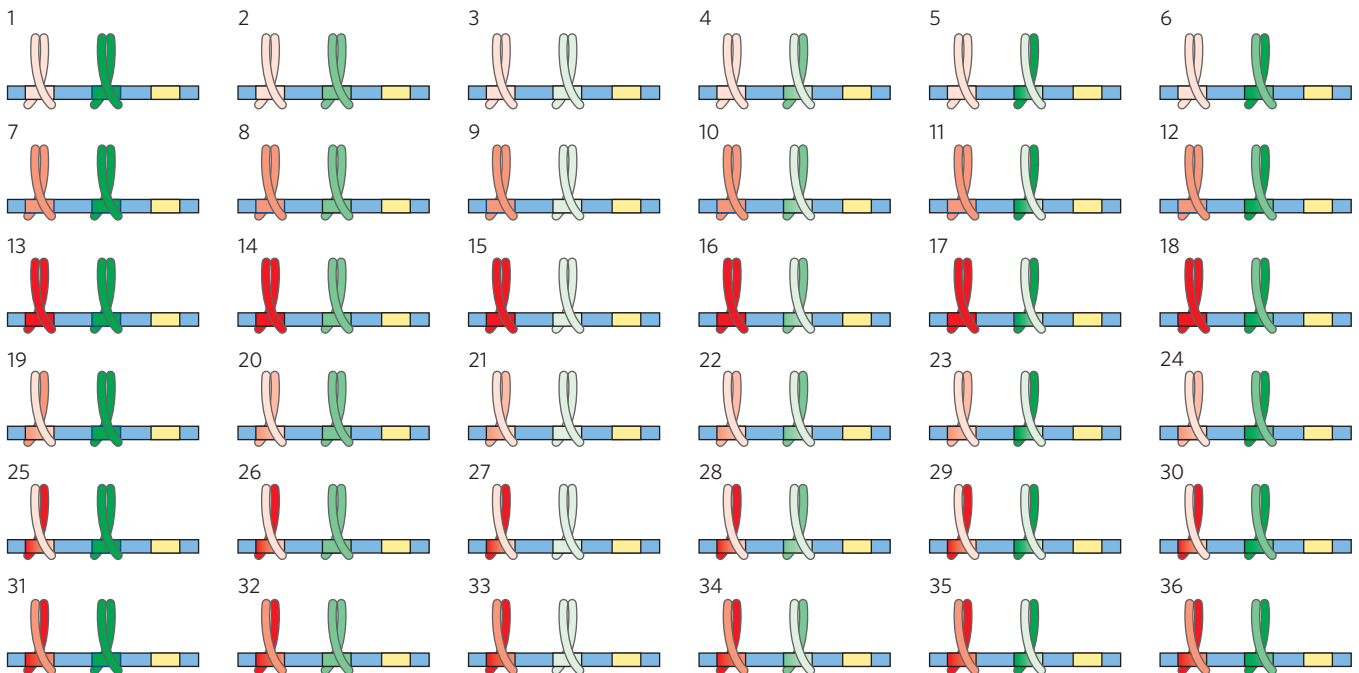


FIGURE 28–27 The advantages of combinatorial control. Combinatorial control allows specific regulation of many genes using a limited repertoire of regulatory proteins. Consider the possibilities inherent in regulation by two different families of leucine zipper proteins (red and green). If each regulatory gene family has three members (shown here as dark, medium, and light shades, each binding to a different DNA sequence) that can freely form either homo- or heterodimers, there are six possible dimeric

species in each family, each of which would recognize a different bipartite regulatory DNA sequence. If a gene had a regulatory site for each protein family, 36 different regulatory combinations would be possible, using just the six proteins from these two families. With six or more sites used in the regulation of a typical eukaryotic gene, the number of possible variants is much greater than this example suggests.

greatly in both the number and the location of additional sequences required for the regulation of transcription.

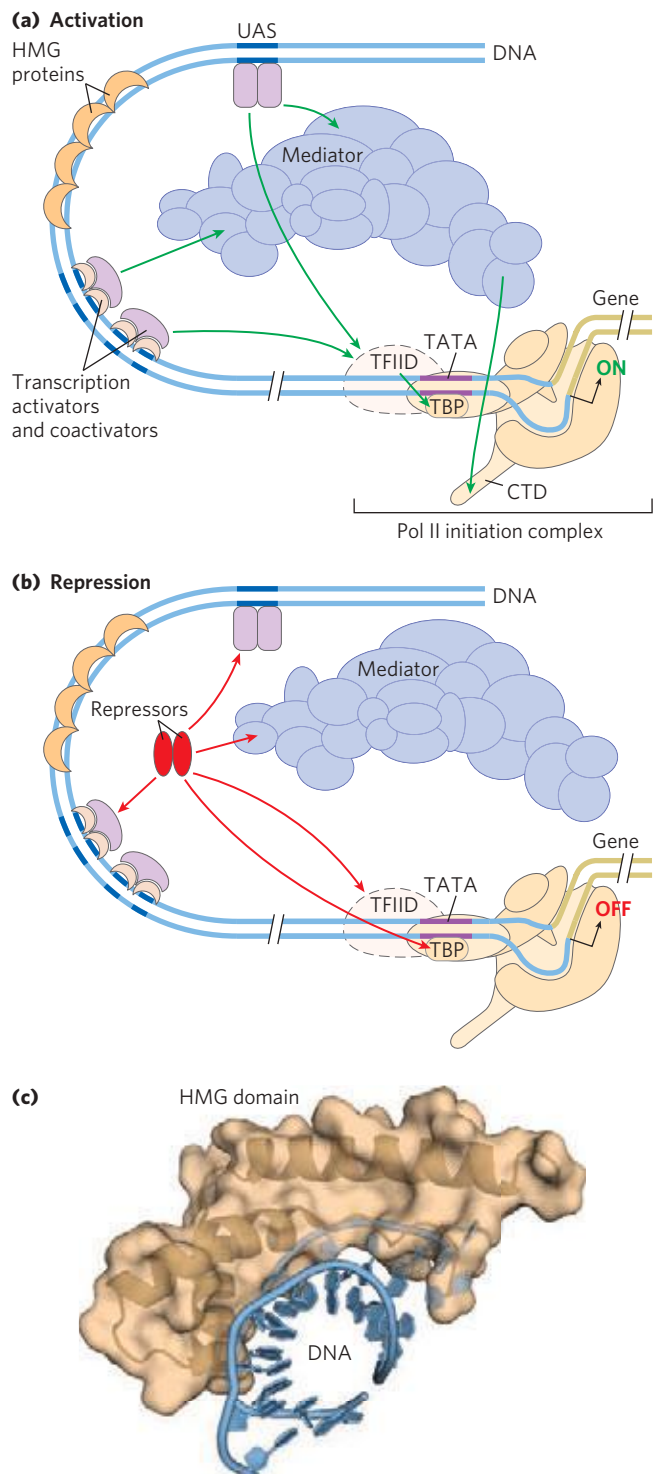
The additional regulatory sequences, generally bound by transcription activators, are usually called **enhancers** in higher eukaryotes and **upstream activator sequences (UASs)** in yeast. A typical enhancer may be found hundreds or even thousands of base pairs upstream from the transcription start site, or may even be downstream, within the gene itself. When bound by the appropriate regulatory proteins, an enhancer increases transcription at nearby promoters regardless of its orientation in the DNA. The UASs of yeast function in a similar way, although generally they must be positioned upstream and within a few hundred base pairs of the transcription start site.

Successful binding of active RNA polymerase II holoenzyme at one of its promoters usually requires the combined action of proteins of five types that have now been described: (1) **transcription activators**, which bind to enhancers or UASs and facilitate transcription; (2) **architectural regulators** that facilitate DNA looping; (3) **chromatin modification and remodeling proteins**, described above; (4) **coactivators**; and (5) **basal transcription factors** (see Fig. 26–9, Table 26–2), required at most Pol II promoters (Fig. 28–28). The coactivators act indirectly—not by binding to the DNA—and are required for essential communication between the activators and the complex composed of Pol II and the basal (or general) transcription factors. Furthermore, a variety of repressor proteins can interfere with communication between the RNA polymerase and the activators, resulting in repression of transcription (Fig. 28–28b). Here we focus on the protein complexes

FIGURE 28–28 Eukaryotic promoters and regulatory proteins. RNA polymerase II and its associated basal (general) transcription factors form a pre-initiation complex at the TATA box and Inr site of the cognate promoters, a process facilitated by transcription activators, acting through coactivators (Mediator, TFIIID, or both). **(a)** A composite promoter with typical sequence elements and protein complexes found in both yeast and higher eukaryotes. The carboxyl-terminal domain (CTD) of Pol II (see Fig. 26–9) is an important point of interaction with Mediator and other protein complexes. Histone modification enzymes (not shown) catalyze methylation and acetylation; remodeling enzymes alter the content and placement of nucleosomes. The transcription activators have distinct DNA-binding domains and activation domains. Green arrows indicate common modes of interaction often required for the activation of transcription, as discussed in the text. The HMG proteins are a common type of architectural regulator (see Fig. 28–5), facilitating the looping of the DNA required to bring together system components bound at distant binding sites. **(b)** Eukaryotic transcriptional repressors function by a range of mechanisms. Some bind directly to DNA, displacing a protein complex required for activation (not shown); many others interact with various parts of the transcription or activation protein complexes to prevent activation. Possible points of interaction are indicated with red arrows. **(c)** The structure of an HMG protein complex with DNA shows how HMG proteins facilitate DNA looping. The binding is relatively nonspecific, although DNA sequence preferences have been identified for many HMG proteins. Shown is the HMG domain of the protein HMG-D of *Drosophila*, bound to DNA (PDB ID 1QRV).

shown in Figure 28–28 and on how they interact to activate transcription.

Transcription Activators The requirements for activators vary greatly from one promoter to another. A few activators are known to facilitate transcription at hundreds of promoters, whereas others are specific for a few promoters. Many activators are sensitive to the binding of signal molecules, providing the capacity to activate or deactivate transcription in response to a changing cellular environment. Some enhancers bound by activators are quite distant from the promoter's TATA box. Multiple



enhancers (often six or more) are bound by a similar number of activators for a typical gene, providing combinatorial control and response to multiple signals.

Architectural Regulators How do the activators function at a distance? The answer in most cases seems to be that, as indicated earlier, the intervening DNA is looped so that the various protein complexes can interact directly. The looping is promoted by architectural regulators that are abundant in chromatin and bind to DNA with limited specificity. Most prominently, the **high mobility group (HMG)** proteins (Fig. 28–28c; “high mobility” refers to their electrophoretic mobility in polyacrylamide gels) play an important structural role in chromatin remodeling and transcriptional activation.

Coactivator Protein Complexes Most transcription requires the presence of additional protein complexes. Some major regulatory protein complexes that interact with Pol II have been defined both genetically and biochemically. These coactivator complexes act as intermediaries between the transcription activators and the Pol II complex.

A major eukaryotic coactivator, consisting of 20 to 30 or more polypeptides in a protein complex, is called **Mediator** (Fig. 28–28). Many of the 20 core polypeptides are highly conserved from fungi to humans. An additional complex of four subunits can interact with Mediator and inhibit transcription initiation. Mediator binds tightly to the carboxyl-terminal domain (CTD) of the largest subunit of Pol II. The Mediator complex is required for both basal and regulated transcription at many promoters used by Pol II, and it also stimulates phosphorylation of the CTD by TFIIH (a basal transcription factor). Transcription activators interact with one or more components of the Mediator complex, with the precise interaction sites differing from one activator to another. Coactivator complexes function at or near the promoter’s TATA box.

Additional coactivators, functioning with one or a few genes, have also been described. Some of these operate in conjunction with Mediator, and some may act in systems that do not employ Mediator.

TATA-Binding Protein The first component to bind in the assembly of a **preinitiation complex (PIC)** at the TATA box of a typical Pol II promoter is the **TATA-binding protein (TBP)**. TBP is often, but not always, delivered as part of a larger complex (~15 subunits) called TFIID. The complete complex also includes the basal transcription factors TFIIB, TFIIIE, TFIIIF, TFIIH; Pol II; and perhaps TFIIA. This minimal PIC, however, is often insufficient for the initiation of transcription and generally does not form at all if the promoter is obscured

within chromatin. Positive regulation, leading to transcription, is imposed by the activators and coactivators.

Choreography of Transcriptional Activation We can now begin to piece together the sequence of transcriptional activation events at a typical Pol II promoter (Fig. 28–29).

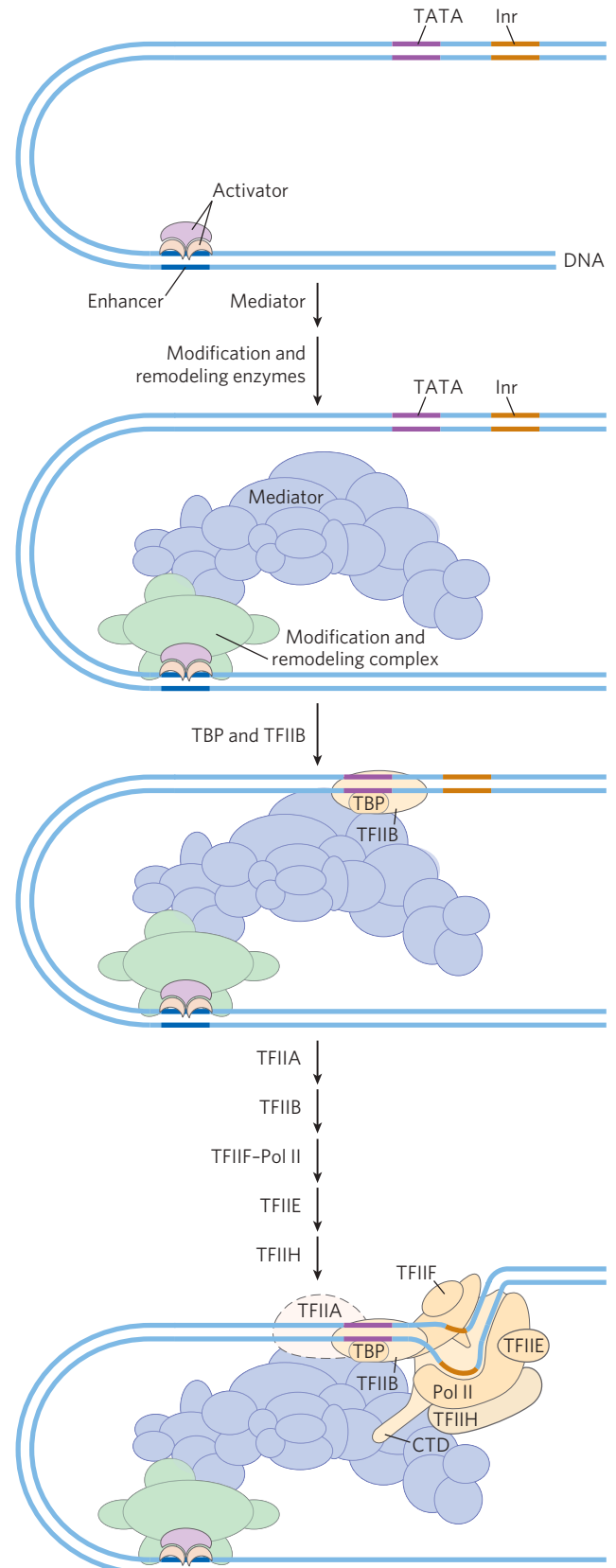


FIGURE 28–29 The components of transcriptional activation. Activators bind the DNA first. The activators recruit the histone modification/nucleosome remodeling complexes and a coactivator such as Mediator. Mediator facilitates the binding of TBP (or TFIID) and TFIIB, and the other basal transcription factors and Pol II then bind. Phosphorylation of the carboxyl-terminal domain (CTD) of Pol II leads to transcription initiation (not shown).

The exact order of binding of some components may vary, but the model in Figure 28–29 illustrates the principles of activation as well as one common path. Many transcription activators have significant affinity for their binding sites even when the sites are within condensed chromatin. The binding of activators is often the event that triggers subsequent activation of the promoter. Binding of one activator may enable the binding of others, gradually displacing some nucleosomes.

Crucial remodeling of the chromatin then takes place in stages, facilitated by interactions between activators and HATs or enzyme complexes such as SWI/SNF (or both). In this way, a bound activator can draw in other components necessary for further chromatin remodeling to permit transcription of specific genes. The bound activators interact with the large Mediator complex. Mediator, in turn, provides an assembly surface for the binding of first TBP (or TFIID), then TFIIB, and then other components of the PIC, including RNA polymerase II. Mediator stabilizes the binding of Pol II and its associated transcription factors and greatly facilitates formation of the PIC. Complexity in these regulatory circuits is the rule rather than the exception, with multiple DNA-bound activators promoting transcription.

The script can change from one promoter to another. For example, many promoters have a different set of recognition sequences and may not have a TATA box, and in multicellular eukaryotes the subunit composition of factors such as TFIID can vary from one tissue to another. However, most promoters seem to require a precisely ordered assembly of components to initiate transcription. The assembly process is not always fast. At some genes it may take minutes; at certain genes of higher eukaryotes the process can take days.

Reversible Transcriptional Activation Although rarer, some eukaryotic regulatory proteins that bind to Pol II promoters or that interact with transcriptional activators can act as repressors, inhibiting the formation of active PICs (Fig. 28–29). Some activators can adopt different conformations, enabling them to serve as transcription activators or as repressors. For example, some steroid hormone receptors (described later) function in the nucleus as transcription activators, stimulating the transcription of certain genes when a particular steroid hormone signal is present. When the hormone is absent, the receptor proteins revert to a repressor conformation, *preventing* the formation of PICs. In some cases this repression involves interaction with histone deacetylases and other proteins that help restore the surrounding chromatin to its transcriptionally inactive state. Mediator, when it includes the inhibitory subunits, may also block transcription initiation. This may be a regulatory mechanism to ensure ordered assembly of the PIC (by delaying transcriptional activation until all required factors are present), or it may be a mechanism that helps deactivate promoters when transcription is no longer required.

The Genes of Galactose Metabolism in Yeast Are Subject to Both Positive and Negative Regulation

Some of the general principles described above can be illustrated by one well-studied eukaryotic regulatory circuit (**Fig. 28–30**). The enzymes required for the importation and metabolism of galactose in yeast are

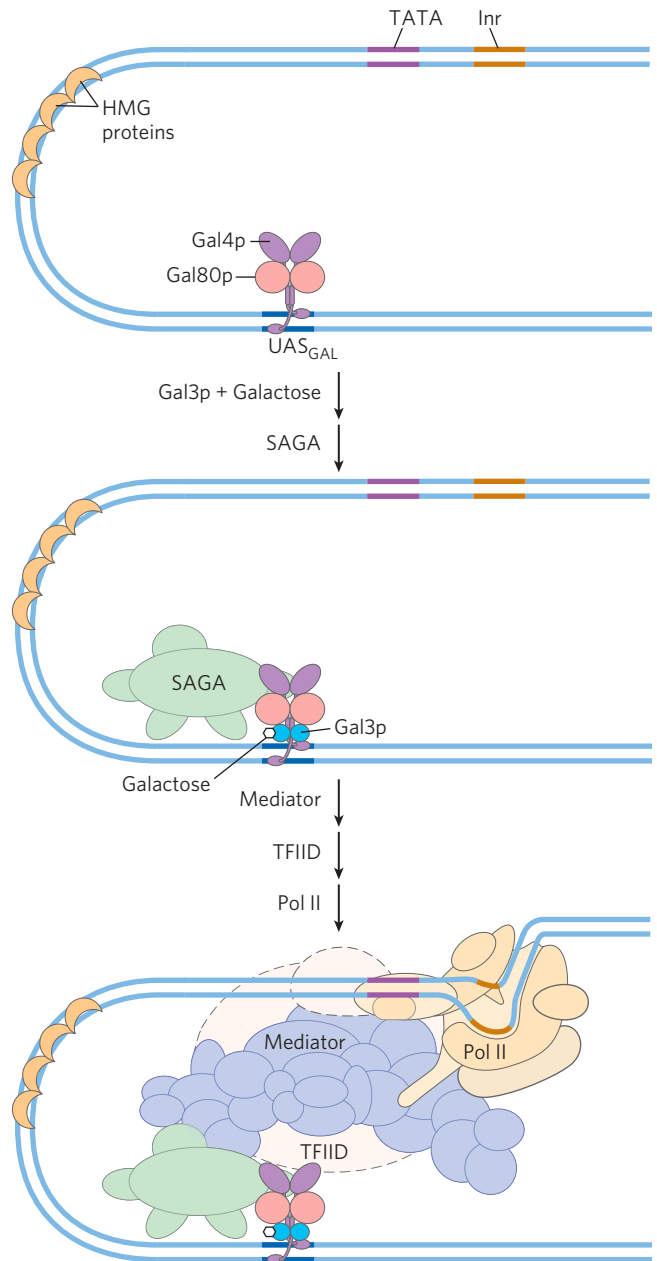


FIGURE 28–30 Regulation of transcription of *GAL* genes in yeast. Galactose imported into the yeast cell is converted to galactose 6-phosphate by a pathway involving six enzymes, whose genes are scattered over three chromosomes (see Table 28-3). Transcription of these genes is regulated by the combined actions of the proteins Gal4p, Gal80p, and Gal3p, with Gal4p playing the central role of transcription activator. The Gal4p-Gal80p complex is inactive. Binding of galactose to Gal3p leads to interaction of Gal3p with the Gal80p-Gal4p complex and activates Gal4p. The Gal4p subsequently recruits SAGA, Mediator, and TFIID to the galactose promoters, leading to DNA polymerase II recruitment and initiation of transcription.

TABLE 28-3 Genes of Galactose Metabolism in Yeast

Gene	Protein function	Chromosomal location	Protein size (number of residues)	Relative protein expression in different carbon sources		
				Glucose	Glycerol	Galactose
Regulated genes						
<i>GAL1</i>	Galactokinase	II	528	–	–	+++
<i>GAL2</i>	Galactose permease	XII	574	–	–	+++
<i>PGM2</i>	Phosphoglucomutase	XIII	569	+	+	++
<i>GAL7</i>	Galactose 1-phosphate uridylyltransferase	II	365	–	–	+++
<i>GAL10</i>	UDP-glucose 4-epimerase	II	699	–	–	+++
<i>MEL1</i>	α -Galactosidase	II	453	–	+	++
Regulatory genes						
<i>GAL3</i>	Inducer	IV	520	–	+	++
<i>GAL4</i>	Transcriptional activator	XVI	881	+/-	+	+
<i>GAL80</i>	Transcriptional inhibitor	XIII	435	+	+	++

Source: Adapted from Reece, R. & Platt, A. (1997) Signaling activation and repression of RNA polymerase II transcription in yeast. *Bioessays* 19, 1001–1010.

encoded by genes scattered over several chromosomes (Table 28-3). Each of the *GAL* genes is transcribed separately, and yeast cells have no operons like those in bacteria. However, all the *GAL* genes have similar promoters and are regulated coordinately by a common set of proteins. The promoters for the *GAL* genes consist of the TATA box and Inr sequences, as well as an upstream activator sequence (UAS_G) recognized by the transcription activator Gal4 protein (Gal4p). Regulation of gene expression by galactose entails an interplay between Gal4p and two other proteins, Gal80p and Gal3p (Fig. 28-30). Gal80p forms a complex with Gal4p, preventing Gal4p from functioning as an activator of the *GAL* promoters. When galactose is present, it binds Gal3p, which then interacts with the Gal80p-Gal4p complex, allowing Gal4p to function as an activator at the various *GAL* promoters. As the various galactose genes are induced and their products build up, Gal3p may be replaced with Gal1p (a galactokinase needed for galactose metabolism that also functions in regulation) for sustained activation of the regulatory circuit.

Other protein complexes also have a role in activating transcription of the *GAL* genes. These include the SAGA complex for histone acetylation and chromatin remodeling, the SWI/SNF complex for chromatin remodeling, and the Mediator complex. The Gal4 protein is responsible for the recruitment of these additional factors needed for transcriptional activation. SAGA may be the first and primary recruitment target for Gal4p.

Glucose is the preferred carbon source for yeast, as it is for bacteria. When glucose is present, most of the *GAL* genes are repressed—whether galactose is present or not. The *GAL* regulatory system described above is

effectively overridden by a complex catabolite repression system that includes several proteins (not depicted in Fig. 28-30).

Transcription Activators Have a Modular Structure

Transcription activators typically have a distinct structural domain for specific DNA binding and one or more additional domains for transcriptional activation or for interaction with other regulatory proteins. Interaction of two regulatory proteins is often mediated by domains containing leucine zippers (Fig. 28-14) or helix-loop-helix motifs (Fig. 28-15). We consider here three distinct types of structural domains used in activation by transcription activators (**Fig. 28-31a**): Gal4p, Sp1, and CTF1.

Gal4p contains a zinc finger-like structure in its DNA-binding domain, near the amino terminus; this domain has six Cys residues that coordinate two Zn²⁺. The protein functions as a homodimer (with dimerization mediated by interactions between two coiled coils) and binds to UAS_G, a palindromic DNA sequence about 17 bp long. Gal4p has a separate activation domain with many acidic amino acid residues. Experiments that substitute a variety of different peptide sequences for the **acidic activation domain** of Gal4p suggest that the acidic nature of this domain is critical to its function, although its precise amino acid sequence can vary considerably.

Sp1 (M_r 80,000) is a transcription activator for a large number of genes in higher eukaryotes. Its DNA-binding site, the GC box (consensus sequence GGGCGG), is usually quite near the TATA box. The DNA-binding domain of the Sp1 protein is near its carboxyl terminus

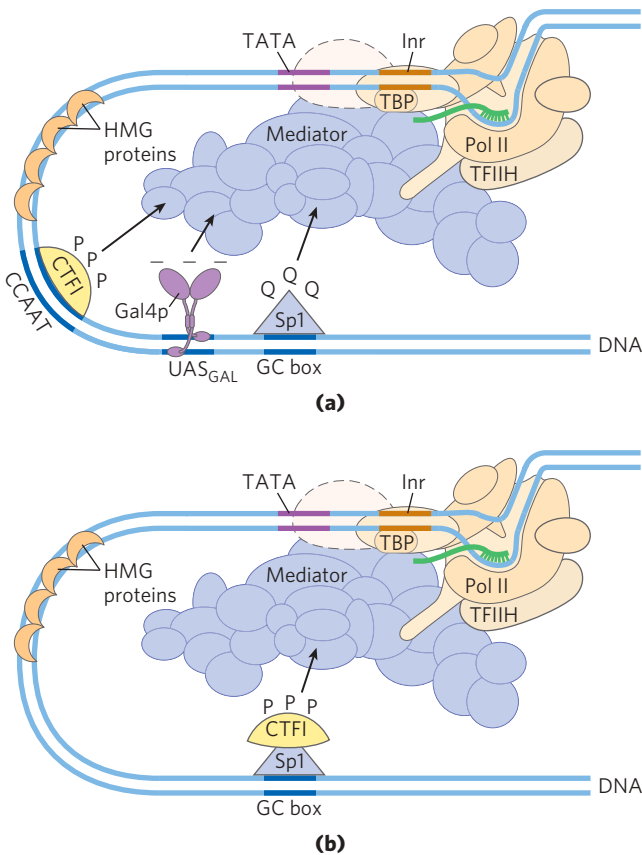


FIGURE 28-31 Transcription activators. (a) Typical activators such as CTF1, Gal4p, and Sp1 have a DNA-binding domain and an activation domain. The nature of the activation domain is indicated by symbols: — — —, acidic; Q Q Q, glutamine-rich; P P P, proline-rich. These proteins generally activate transcription by interacting with coactivator complexes such as Mediator. Note that the binding sites illustrated here are not generally found together near a single gene. (b) A chimeric protein containing the DNA-binding domain of Sp1 and the activation domain of CTF1 activates transcription if a GC box is present.

and contains three zinc fingers. Two other domains in Sp1 function in activation and are notable in that 25% of their amino acid residues are Gln. A wide variety of other activator proteins also have these **glutamine-rich domains**.

CTF1 (*CCAAT-binding transcription factor 1*) belongs to a family of transcription activators that bind a sequence called the CCAAT site (its consensus sequence is TGGN₆GCCAA, where N is any nucleotide). The DNA-binding domain of CTF1 contains many basic amino acid residues, and the binding region is probably arranged as an α helix. This protein has neither a helix-turn-helix nor a zinc finger motif; its DNA-binding mechanism is not yet clear. CTF1 has a **proline-rich activation domain**, with Pro accounting for more than 20% of the amino acid residues.

The discrete activation and DNA-binding domains of regulatory proteins often act completely independently, as has been demonstrated in “domain-swapping” experiments. Genetic engineering techniques (Chapter 9) can join the proline-rich activation domain of CTF1 to

the DNA-binding domain of Sp1 to create a protein that, like intact Sp1, binds to GC boxes on the DNA and activates transcription at a nearby promoter (as in Fig. 28–31b). The DNA-binding domain of Gal4p has similarly been replaced experimentally with the DNA-binding domain of the *E. coli* LexA repressor (of the SOS response; Fig. 28–20). This chimeric protein neither binds at UAS_G nor activates the yeast *GAL* genes (as would intact Gal4p) unless the UAS_G sequence in the DNA is replaced by the LexA recognition site.

Eukaryotic Gene Expression Can Be Regulated by Intercellular and Intracellular Signals

The effects of steroid hormones (and of thyroid and retinoid hormones, which have a similar mode of action) provide additional well-studied examples of the modulation of eukaryotic regulatory proteins by direct interaction with molecular signals (see Fig. 12–30). Unlike other types of hormones, steroid hormones do not have to bind to plasma membrane receptors. Instead, they can interact with intracellular receptors that are themselves transcription activators. Steroid hormones too hydrophobic to dissolve readily in the blood (estrogen, progesterone, and cortisol, for example) travel on specific carrier proteins from their point of release to their target tissues. In the target tissue, the hormone passes through the plasma membrane by simple diffusion. Once inside the cell, the hormones interact with one of two types of steroid-binding nuclear receptor (**Fig. 28–32**). In both cases, the hormone-receptor complex acts by binding to highly specific DNA sequences called **hormone response elements (HREs)**, thereby altering gene expression. Acting at these sites, the receptors act as transcription activators, recruiting coactivators and DNA polymerase II (plus its associated transcription factors) to trigger transcription of the gene.

The DNA sequences (HREs) to which hormone-receptor complexes bind are similar in length and arrangement, but differ in sequence, for the various steroid hormones. Each receptor has a consensus HRE sequence (Table 28–4) to which the hormone-receptor complex binds well, with each consensus consisting of two six-nucleotide sequences, either contiguous or separated by three nucleotides, in tandem or in a palindromic arrangement. The hormone receptors have a highly conserved DNA-binding domain with two zinc fingers (**Fig. 28–33**). The hormone-receptor complex binds to the DNA as a dimer, with the zinc finger domains of each monomer recognizing one of the six-nucleotide sequences. The ability of a given hormone to act through the hormone-receptor complex to alter the expression of a specific gene depends on the exact sequence of the HRE, its position relative to the gene, and the number of HREs associated with the gene.

The ligand-binding region of the receptor protein—always at the carboxyl terminus—is quite specific to the particular receptor. In the ligand-binding region, the

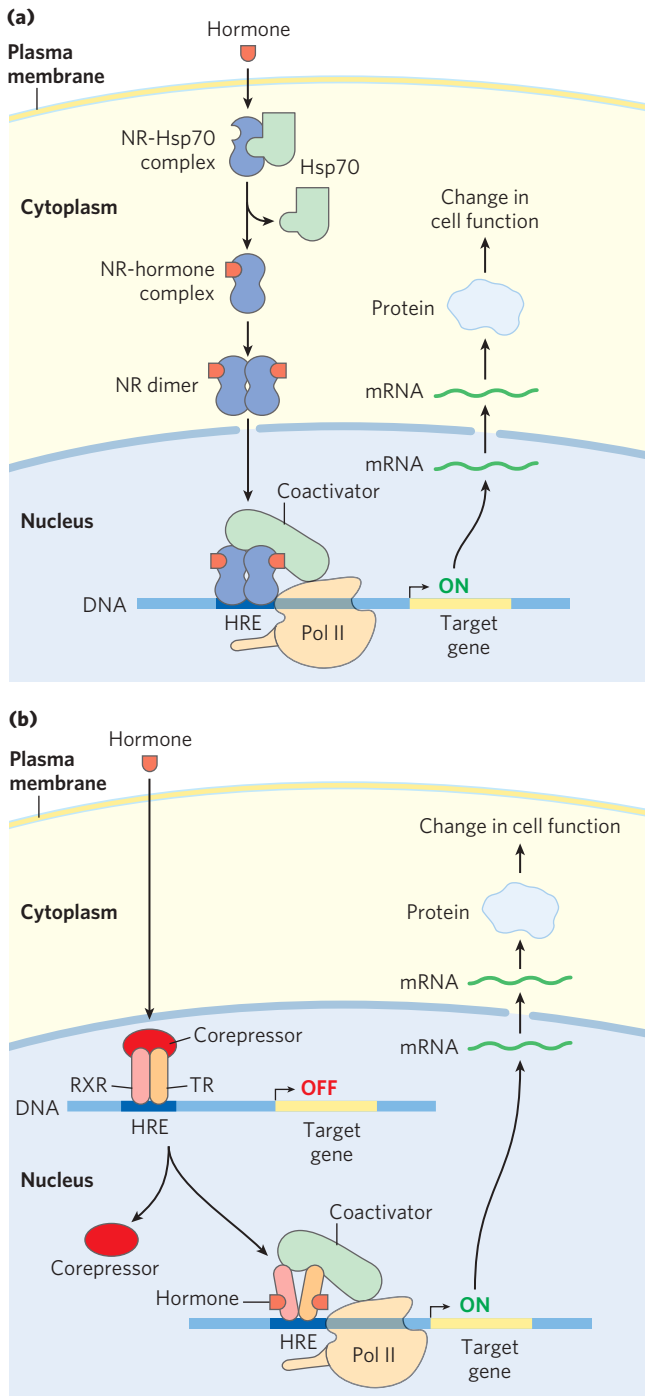


FIGURE 28-32 Mechanisms of steroid hormone receptor function. There are two types of steroid-binding nuclear receptors. **(a)** Monomeric type I receptors (NR) are found in the cytoplasm, in a complex with a heat shock protein (Hsp70). Receptors for estrogen, progesterone, androgens, and glucocorticoids are of this type. When the steroid hormone binds, the Hsp70 dissociates and the receptor dimerizes, exposing a nuclear localization signal. The dimeric receptor, with hormone bound, migrates to the nucleus, where it binds a hormone response element (HRE) and acts as a transcription activator. **(b)** Type II receptors, by contrast, are always in the nucleus, bound to an HRE in the DNA and to a corepressor that renders them inactive. The thyroid hormone receptor (TR) is of this type. The hormone migrates through the cytoplasm and diffuses across the nuclear membrane. In the nucleus it binds to a heterodimer consisting of the thyroid hormone receptor and the retinoid X receptor (RXR). A conformation change leads to dissociation of the corepressor, and the receptor then functions as a transcription activator.

TABLE 28-4 Hormone Response Elements (HREs) Bound by Steroid-Type Hormone Receptors

Receptor	Consensus sequence bound*
Androgen	GG(A/T)ACAN ₂ TGTTCT
Glucocorticoid	GGTACAN ₃ TGTTCT
Retinoic acid (some)	AGGTCAN ₅ AGGTCA
Vitamin D	AGGTCAN ₃ AGGTCA
Thyroid hormone	AGGTCAN ₃ AGGTCA
RX [†]	AGGTCANAGGTCANAG GTCANAGGTCA

*N represents any nucleotide.

[†]Forms a dimer with the retinoic acid receptor or vitamin D receptor.

glucocorticoid receptor is only 30% similar to the estrogen receptor and 17% similar to the thyroid hormone receptor. The size of the ligand-binding region varies dramatically; in the vitamin D receptor it has only 25 amino acid residues, whereas in the mineralocorticoid receptor it has 603 residues. Mutations that change one amino acid in these regions can result in loss of responsiveness to a specific hormone. Some humans unable to respond to cortisol, testosterone, vitamin D, or thyroxine have mutations of this type.

Some hormone receptors, including the human progesterone receptor, activate transcription with the aid of an unusual coactivator—**steroid receptor RNA activator (SRA)**, an ~700 nucleotide RNA. SRA acts

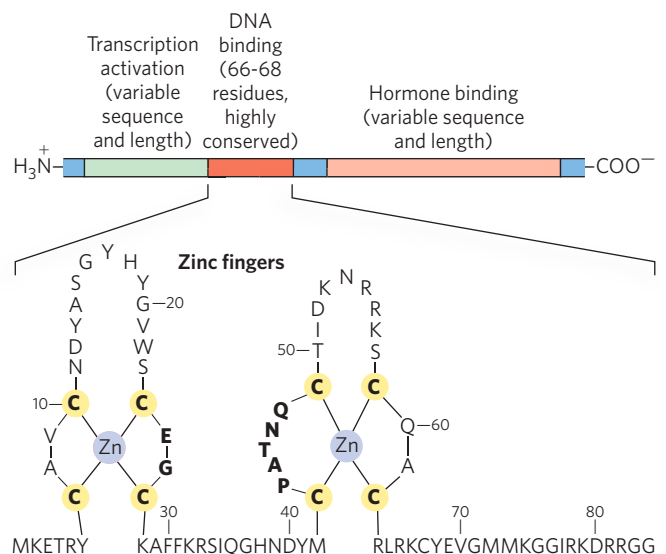


FIGURE 28-33 Typical steroid hormone receptors. These receptor proteins have a binding site for the hormone, a DNA-binding domain, and a region that activates transcription of the regulated gene. The highly conserved DNA-binding domain has two zinc fingers. The sequence shown here is that for the estrogen receptor, but the residues in bold type are common to all steroid hormone receptors.

as part of a ribonucleoprotein complex, but it is the RNA component that is required for transcription coactivation. The detailed set of interactions between SRA and other components of the regulatory systems for these genes remains to be worked out.

Regulation Can Result from Phosphorylation of Nuclear Transcription Factors

We noted in Chapter 12 that the effects of insulin on gene expression are mediated by a series of steps leading ultimately to the activation of a protein kinase in the nucleus that phosphorylates specific DNA-binding proteins and thereby alters their ability to act as transcription factors (see Fig. 12–15). This general mechanism mediates the effects of many nonsteroid hormones. For example, the β -adrenergic pathway that leads to elevated levels of cytosolic cAMP, which acts as a second messenger in eukaryotes as well as in bacteria (see Figs 12–4, 28–17), also affects the transcription of a set of genes, each of which is located near a specific DNA sequence called a **cAMP response element (CRE)**. The catalytic subunit of protein kinase A, released when cAMP levels rise (see Fig. 12–6), enters the nucleus and phosphorylates a nuclear protein, the **CRE-binding protein (CREB)**. When phosphorylated, CREB binds to CREs near certain genes and acts as a transcription factor, turning on the expression of these genes.

Many Eukaryotic mRNAs Are Subject to Translational Repression

Regulation at the level of translation assumes a much more prominent role in eukaryotes than in bacteria and is observed in a range of cellular situations. In contrast to the tight coupling of transcription and translation in bacteria, the transcripts generated in a eukaryotic nucleus must be processed and transported to the cytoplasm before translation. This can impose a significant delay on the appearance of a protein. When a rapid increase in protein production is needed, a translationally repressed mRNA already in the cytoplasm can be activated for translation without delay. Translational regulation may play an especially important role in regulating certain very long eukaryotic genes (a few are measured in the millions of base pairs), for which transcription and mRNA processing can require many hours. Some genes are regulated at both the transcriptional and translational stages, with the latter playing a role in the fine-tuning of cellular protein levels. In some anucleate cells, such as reticulocytes (immature erythrocytes), transcriptional control is entirely unavailable and translational control of stored mRNAs becomes essential. As described below, translational controls can also have spatial significance during development, when the regulated translation of prepositioned mRNAs creates a local gradient of the protein product.

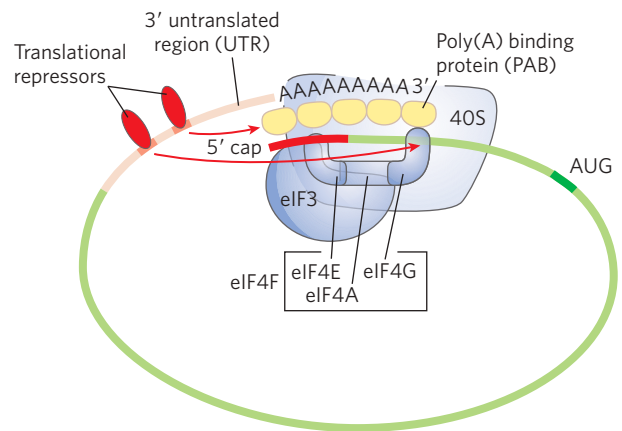


FIGURE 28–34 Translational regulation of eukaryotic mRNA. One of the most important mechanisms for translational regulation in eukaryotes involves the binding of translational repressors (RNA-binding proteins) to specific sites in the 3' untranslated region (3'UTR) of the mRNA. These proteins interact with eukaryotic initiation factors or with the ribosome to prevent or slow translation.

Eukaryotes have at least four main mechanisms of translational regulation.

1. Translation initiation factors are subject to phosphorylation by protein kinases. The phosphorylated forms are often less active and cause a general depression of translation in the cell.
2. Some proteins bind directly to mRNA and act as translational repressors, many of them binding at specific sites in the 3' untranslated region (3'UTR). So positioned, these proteins interact with other translation initiation factors bound to the mRNA or with the 40S ribosomal subunit to prevent translation initiation (**Fig. 28–34**).
3. Binding proteins, present in eukaryotes from yeast to mammals, disrupt the interaction between eIF4E and eIF4G (see Fig. 27–28). The mammalian versions are known as 4E-BPs (eIF4E binding proteins). When cell growth is slow, these proteins limit translation by binding to the site on eIF4E that normally interacts with eIF4G. When cell growth resumes or increases in response to growth factors or other stimuli, the binding proteins are inactivated by protein kinase–dependent phosphorylation.
4. RNA-mediated regulation of gene expression, considered later, often occurs at the level of translational repression.

The variety of translational regulation mechanisms provides flexibility, allowing focused repression of a few mRNAs or global regulation of all cellular translation.

Translational regulation has been particularly well studied in reticulocytes. One such mechanism in these cells involves eIF2, the initiation factor that binds to the initiator tRNA and conveys it to the ribosome; when

Met-tRNA has bound to the P site, the factor eIF2B binds to eIF2, recycling it with the aid of GTP binding and hydrolysis. The maturation of reticulocytes includes destruction of the cell nucleus, leaving behind a plasma membrane packed with hemoglobin. Messenger RNAs deposited in the cytoplasm before the loss of the nucleus allow for the replacement of hemoglobin. When reticulocytes become deficient in iron or heme, the translation of globin mRNAs is repressed. A protein kinase called **HCR (hemin-controlled repressor)** is activated, catalyzing the phosphorylation of eIF2. When phosphorylated, eIF2 forms a stable complex with eIF2B that sequesters the eIF2, making it unavailable for participation in translation. In this way, the reticulocyte coordinates the synthesis of globin with the availability of heme.

Many additional examples of translational regulation have been found in studies of the development of multicellular organisms, as discussed in more detail below.

Posttranscriptional Gene Silencing Is Mediated by RNA Interference

In higher eukaryotes, including nematodes, fruit flies, plants, and mammals, a class of small RNAs called **microRNAs (miRNAs)** mediates the silencing of many genes. In a phenomenon first described and explained by Craig Mello and Andrew Fire, the RNAs function by interacting with mRNAs, often in the 3'UTR, resulting in either mRNA degradation or translation inhibition. In either case, the mRNA, and thus the gene that produces it, is silenced. This form of gene regulation controls developmental timing in at least some organisms. It is also used as a mechanism to protect against invading RNA viruses (particularly important in plants, which lack an immune system) and to control the activity of transposons. In addition, small RNA molecules may play a critical (but still undefined) role in the formation of heterochromatin.

Many miRNAs are present only transiently during development, and these are sometimes referred to as **small temporal RNAs (stRNAs)**. Thousands of different miRNAs have been identified in higher eukaryotes, and they may affect the regulation of a third of mammalian genes. They are transcribed as precursor RNAs about 70 nucleotides long, with internally complementary sequences that form hairpinlike structures. Details of the pathway for processing of miRNAs were described in Chapter 26 (see Fig. 26–27). The precursors are cleaved by endonucleases such as Dro-

sha and Dicer to form short duplexes about 20 to 25 nucleotides long. One strand of the processed miRNA is transferred to the target mRNA (or to a viral or transposon RNA), leading to inhibition of translation or degradation of the RNA (**Fig. 28–35**). Some miRNAs bind to and affect a single mRNA and thus affect expression of only one gene. Others interact with multiple mRNAs and thus form the mechanistic core of regulons that coordinate the expression of multiple genes.

This gene regulation mechanism has an interesting and very useful practical side. If an investigator introduces into an organism a duplex RNA molecule corresponding in sequence to virtually any mRNA, Dicer cleaves the duplex into short segments, called **small interfering RNAs (siRNAs)**. These bind to the mRNA and silence it (**Fig. 28–35b**). The process is known as **RNA interference (RNAi)**. In plants, virtually any gene can be effectively shut down in this way. Nematodes readily ingest entire, functional RNAs, and simply introducing the duplex RNA into the worm's diet produces very effective suppression of the target gene. The technique has rapidly become an important tool in the ongoing efforts to study gene function, because it can disrupt gene function without creating a mutant organism. The procedure can be applied to humans as well. Laboratory-produced siRNAs have been used to block HIV and poliovirus infections in cultured human cells

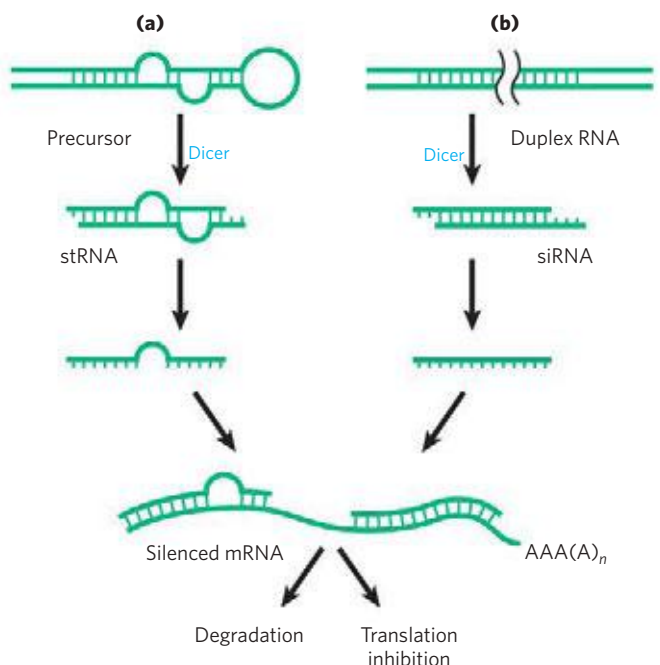


FIGURE 28–35 Gene silencing by RNA interference. (a) Small temporal RNAs (stRNAs, a class of miRNAs) are generated by Dicer-mediated cleavage of longer precursors that fold to create duplex regions. The stRNAs then bind to mRNAs, leading to degradation of mRNA or inhibition of translation. (b) Double-stranded RNAs can be constructed and introduced into a cell. Dicer processes the duplex RNAs into small interfering RNAs (siRNAs), which interact with the target mRNA. Again, either the mRNA is degraded or translation is inhibited.



Craig Mello



Andrew Fire

for a week or so at a time. Rapid progress in research into RNA interference makes this a field to watch for future medical advances.

RNA-Mediated Regulation of Gene Expression Takes Many Forms in Eukaryotes

The special-function RNAs in eukaryotes include miRNAs, described above; snRNAs, involved in RNA splicing (see Fig. 26–16); and snoRNAs, involved in rRNA modification (see Fig. 26–25). All RNAs that do not encode proteins, including rRNAs and tRNAs, come under the general designation of **ncRNAs** (noncoding RNAs). Mammalian genomes seem to encode more ncRNAs than coding mRNAs (see Box 26–4). Not surprisingly, additional functional classes of ncRNAs are still being discovered.

Many of the newly found examples interact with proteins rather than with RNAs and affect the function of the bound proteins. The SRA that functions as a coactivator of steroid hormone–responsive genes is one example: it affects the activation of transcription. The heat shock response in human cells provides another example. Heat shock factor 1 (HSF1) is an activator protein that, in nonstressed cells, exists as a monomer bound by the chaperone Hsp90. Under stress conditions, HSF1 is released from Hsp90 and trimerizes. The HSF1 trimer binds to DNA and activates transcription of genes whose products are required to deal with the stress. An ncRNA called HSR1 (heat shock RNA 1; ~600 nucleotides) stimulates the HSF1 trimerization and DNA binding. HSR1 does not act alone; it functions in a complex with the translation elongation factor eEF1A.

Additional RNAs affect transcription in a variety of ways. Besides its role in splicing (see Fig. 26–16), the snRNA U1 directly binds to the transcription factor TFIIF. Its function in this context is not yet clear, but it may regulate TFIIF or affect the coupling between transcription and splicing, or both. A 331 nucleotide ncRNA called 7SK, abundant in mammals, binds to the Pol II transcription elongation factor pTEFb (see Table 26–2) and represses transcript elongation. Another ncRNA,

B2 (~178 nucleotides), binds directly to Pol II during heat shock and represses transcription. The B2-bound Pol II assembles into stable PICs, but transcription is blocked. The B2 RNA thus halts the transcription of many genes during heat shock, and the mechanism that allows HSF1-responsive genes to be expressed in the presence of B2 remains to be worked out.

The recognized roles of ncRNAs in gene expression and in many other cellular processes are rapidly expanding. At the same time, the study of the biochemistry of gene regulation is becoming much less protein-centric.

Development Is Controlled by Cascades of Regulatory Proteins

For sheer complexity and intricacy of coordination, the patterns of gene regulation that bring about development of a zygote into a multicellular animal or plant have no peer. Development requires transitions in morphology and protein composition that depend on tightly coordinated changes in expression of the genome. More genes are expressed during early development than in any other part of the life cycle. For example, in the sea urchin, an oocyte has about 18,500 *different* mRNAs, compared with about 6,000 different mRNAs in the cells of a typical differentiated tissue. The mRNAs in the oocyte give rise to a cascade of events that regulate the expression of many genes across both space and time.

Several organisms have emerged as important model systems for the study of development, because they are easy to maintain in a laboratory and have relatively short generation times. These include nematodes, fruit flies, zebra fish, mice, and the plant *Arabidopsis*. This discussion focuses on the development of fruit flies. Our understanding of the molecular events during development of *Drosophila melanogaster* is particularly well advanced and can be used to illustrate patterns and principles of general significance.

The life cycle of the fruit fly includes complete metamorphosis during its progression from an embryo to an adult (Fig. 28–36). Among the most important characteristics of the embryo are its **polarity** (the

FIGURE 28–36 Life cycle of the fruit fly *Drosophila melanogaster*. *Drosophila* undergoes a complete metamorphosis, which means that the adult insect is radically different in form from its immature stages, a transformation that requires extensive alterations during development. By the late embryonic stage, segments have formed, each containing specialized structures from which the various appendages and other features of the adult fly will develop.

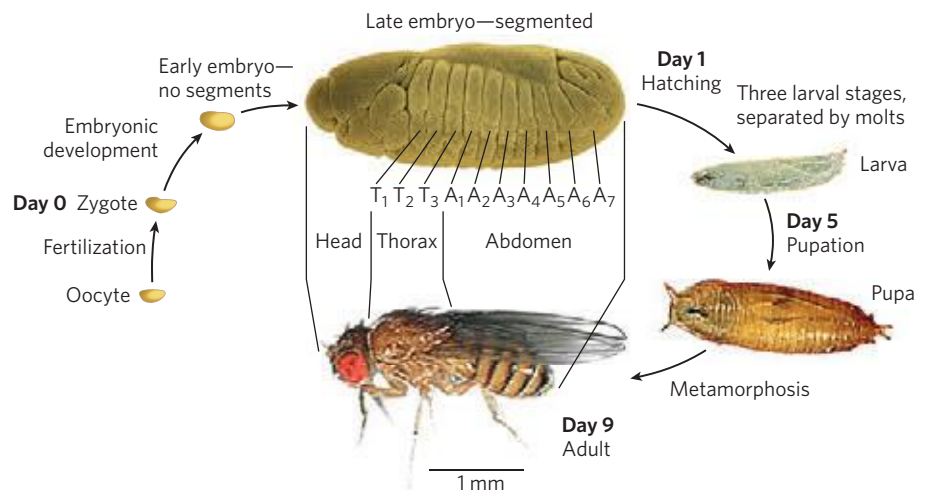


FIGURE 28-37 Early development in *Drosophila*. During development of the egg, maternal mRNAs (including the *bicoid* and *nanos* gene transcripts, discussed in the text) and proteins are deposited in the developing oocyte (unfertilized egg cell) by nurse cells and follicle cells. After fertilization, the two nuclei of the fertilized egg divide in synchrony within the common cytoplasm (syncytium), then migrate to the periphery. Membrane invaginations surround the nuclei to create a monolayer of cells at the periphery; this is the cellular blastoderm stage. During the early nuclear divisions, several nuclei at the far posterior become pole cells, which later become the germ-line cells.

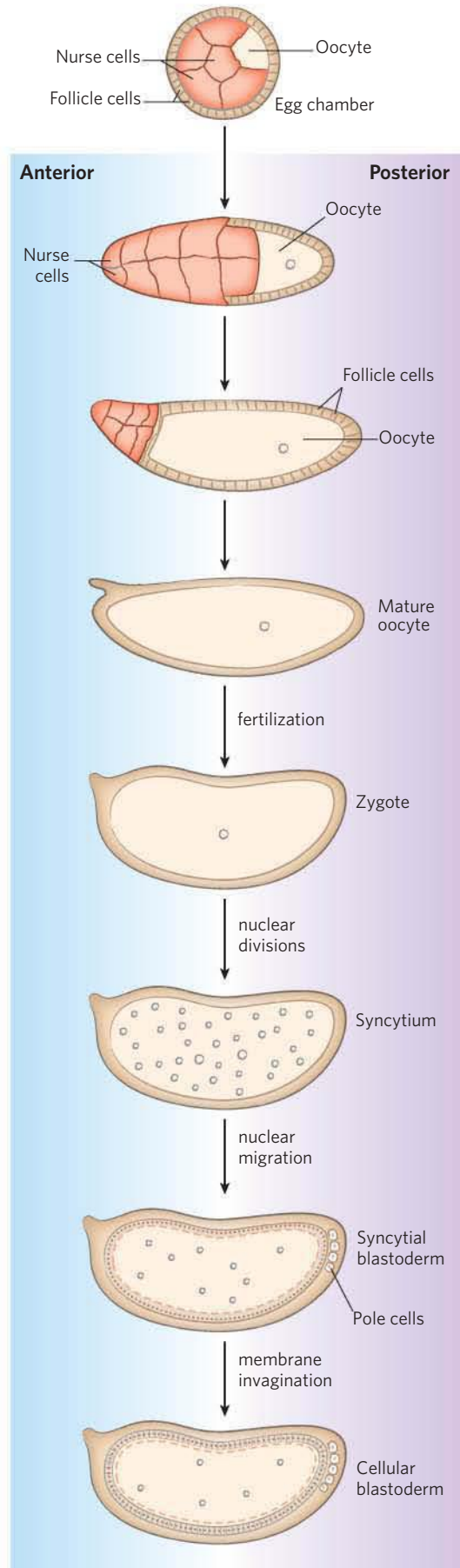
anterior and posterior parts of the animal are readily distinguished, as are its dorsal and ventral parts) and its **metamerism** (the embryo body is made up of serially repeating segments, each with characteristic features). During development, these segments become organized into a head, thorax, and abdomen. Each segment of the adult thorax has a different set of appendages. Development of this complex pattern is under genetic control, and a variety of pattern-regulating genes have been discovered that dramatically affect the organization of the body.

The *Drosophila* egg, along with 15 nurse cells, is surrounded by a layer of follicle cells (Fig. 28-37). As the egg cell forms (before fertilization), mRNAs and proteins originating in the nurse and follicle cells are deposited in the egg cell, where some play a critical role in development. Once a fertilized egg is laid, its nucleus divides and the nuclear descendants continue to divide in synchrony every 6 to 10 min. Plasma membranes are not formed around the nuclei, which are distributed within the egg cytoplasm (forming a syncytium). Between the eighth and eleventh rounds of nuclear division, the nuclei migrate to the outer layer of the egg, forming a monolayer of nuclei surrounding the common yolk-rich cytoplasm; this is the syncytial blastoderm. After a few additional divisions, membrane invaginations surround the nuclei to create a layer of cells that form the cellular blastoderm. At this stage, the mitotic cycles in the various cells lose their synchrony. The developmental fate of the cells is determined by the mRNAs and proteins originally deposited in the egg by the nurse and follicle cells.

Proteins that, through changes in local concentration or activity, cause the surrounding tissue to take up a particular shape or structure are sometimes referred to as **morphogens**; they are the products of pattern-regulating genes. As defined by Christiane Nüsslein-Volhard, Edward B. Lewis, and Eric F. Wieschaus, three major classes of pattern-regulating genes—maternal, segmentation, and homeotic genes—function in successive stages of development to specify the basic features of



Christiane Nüsslein-Volhard





Edward B. Lewis, 1918–2004



Eric F. Wieschaus

the *Drosophila* embryo body. **Maternal genes** are expressed in the unfertilized egg, and the resulting **maternal mRNAs** remain dormant until fertilization. These provide most of the proteins needed in very early development, until the cellular blastoderm is formed. Some of the proteins encoded by maternal mRNAs direct the spatial organization of the developing embryo at early stages, establishing its polarity. **Segmentation genes**, transcribed after fertilization, direct the formation of the proper number of body segments. At least three subclasses of segmentation genes act at successive stages: **gap genes** divide the developing embryo into several broad regions, and **pair-rule genes** together with **segment polarity genes** define 14 stripes that become the 14 segments of a normal embryo. **Homeotic genes** are expressed still later; they specify which organs and appendages will develop in particular body segments.

The many regulatory genes in these three classes direct the development of an adult fly, with a head, thorax, and abdomen, with the proper number of segments, and with the correct appendages on each segment. Although embryogenesis takes about a day to complete, all these genes are activated during the first four hours. Some mRNAs and proteins are present for only a few minutes at specific points during this period. Some of the genes code for transcription factors that affect the expression of other genes in a kind of developmental cascade. Regulation at the level of translation also occurs, and many of the regulatory genes encode translational repressors, most of which bind to the 3'UTR of the mRNA (Fig. 28–34). Because many mRNAs are deposited in the egg long before their translation is required, translational repression provides an especially important avenue for regulation in developmental pathways.

Maternal Genes Some maternal genes are expressed within the nurse and follicle cells, and some in the egg itself. In the unfertilized *Drosophila* egg, the maternal gene products establish two axes—anterior-posterior and dorsal-ventral—and thus define which regions of the radially symmetric egg will develop into the head and abdomen and the top and bottom of the adult fly.

If all cells divided to produce two identical daughter cells, multicellular organisms would never be more than a ball of identical cells. The generation of different cell fates requires programmed asymmetric cell divisions. A

key event in very early development is establishment of mRNA and protein gradients along the body axes. Some maternal mRNAs have protein products that diffuse through the cytoplasm to create an asymmetric distribution in the egg. Different cells in the cellular blastoderm therefore inherit different amounts of these proteins, setting the cells on different developmental paths. The products of the maternal mRNAs include transcription activators or repressors as well as translational repressors, all regulating the expression of other pattern-regulating genes. The resulting specific patterns and sequences of gene expression therefore differ between cell lineages, ultimately orchestrating the development of each adult structure.

The anterior-posterior axis in *Drosophila* is defined at least in part by the products of the *bicoid* and *nanos* genes. The *bicoid* gene product is a major anterior morphogen, and the *nanos* gene product is a major posterior morphogen. The mRNA from the *bicoid* gene is synthesized by nurse cells and deposited in the unfertilized egg near its anterior pole. Nüsslein-Volhard found that this mRNA is translated soon after fertilization, and the Bicoid protein diffuses through the cell to create, by the seventh nuclear division, a concentration gradient radiating out from the anterior pole (Fig. 28–38a). The Bicoid protein is a transcription factor that activates the expression of several segmentation genes; the protein contains a homeodomain (p. 1163). Bicoid is also a translational repressor that inactivates certain mRNAs. The amounts of Bicoid protein in various parts of the embryo affect the subsequent expression of other genes in a threshold-dependent manner. Genes are transcriptionally activated or translationally repressed only where the Bicoid protein concentration exceeds the threshold. Changes in the shape of the Bicoid concentration gradient have dramatic effects on the body pattern. Lack of Bicoid protein results in development of an embryo with two abdomens but neither head nor thorax (Fig. 28–38b); however, embryos without Bicoid will develop normally if an adequate amount of *bicoid* mRNA is injected into the egg at the appropriate end. The *nanos* gene has an analogous role, but its mRNA is deposited at the posterior end of the egg and the anterior-posterior protein gradient peaks at the posterior pole. The Nanos protein is a translational repressor.

A broader look at the effects of maternal genes reveals the outline of a developmental circuit. In addition to the *bicoid* and *nanos* mRNAs, which are deposited in the egg asymmetrically, several other maternal mRNAs are deposited uniformly throughout the egg cytoplasm. Three of these mRNAs encode the Pumilio, Hunchback, and Caudal proteins, all affected by *nanos* and *bicoid* (Fig. 28–39). Caudal and Pumilio are involved in development of the posterior end of the fly. Caudal is a transcription activator with a homeodomain; Pumilio is a translational repressor. Hunchback protein plays an important role in development of the anterior end and is also a transcriptional regulator of a variety of

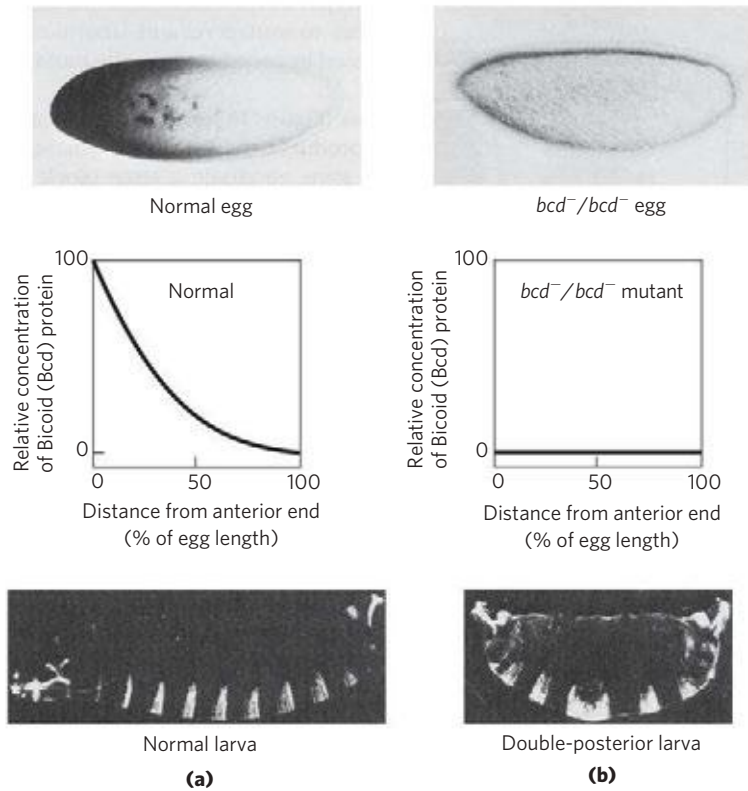
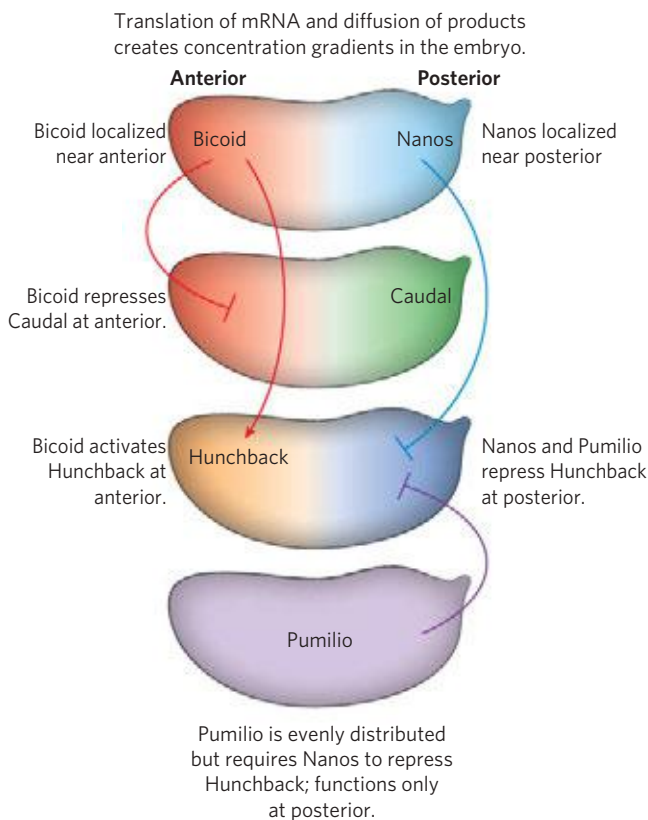


FIGURE 28-38 Distribution of a maternal gene product in a *Drosophila* egg. (a) Micrograph of an immunologically stained egg, showing distribution of the *bicoid* (*bcd*) gene product. The graph measures stain intensity. This distribution is essential for normal development of the anterior structures.

(b) If the *bcd* gene is not expressed by the mother (bcd^-/bcd^- mutant) and thus no *bicoid* mRNA is deposited in the egg, the resulting embryo has two posteriors (and soon dies).



genes, in some cases a positive regulator, in other cases negative. Bicoid suppresses translation of *caudal* at the anterior end and also acts as a transcription activator of *hunchback* in the cellular blastoderm. Because *hunchback* is expressed both from maternal mRNAs and from genes in the developing egg, it is considered both a maternal and a segmentation gene. The result of the activities of Bicoid is an increased concentration of Hunchback at the anterior end of the egg. The Nanos and Pumilio proteins act as translational repressors of *hunchback*, suppressing synthesis of its protein near the posterior end of the egg. Pumilio does not function in the absence of the Nanos protein, and the gradient of Nanos expression confines the activity of both proteins to the posterior region. Translational repression of the

FIGURE 28-39 Regulatory circuits of the anterior-posterior axis in a *Drosophila* egg. The *bicoid* and *nanos* mRNAs are localized near the anterior and posterior poles, respectively. The *caudal*, *hunchback*, and *pumilio* mRNAs are distributed evenly throughout the egg cytoplasm. The gradients of Bicoid (Bcd) and Nanos proteins affect the expression of the *caudal* and *hunchback* mRNAs as shown, leading to accumulation of Hunchback protein in the anterior and Caudal protein in the posterior of the egg. Because Pumilio protein requires Nanos protein for its activity as a translational repressor of *hunchback*, it functions only at the posterior end.

hunchback gene leads to degradation of *hunchback* mRNA near the posterior end. However, lack of Bicoid in the posterior leads to expression of *caudal*. In this way, the Hunchback and Caudal proteins become asymmetrically distributed in the egg.

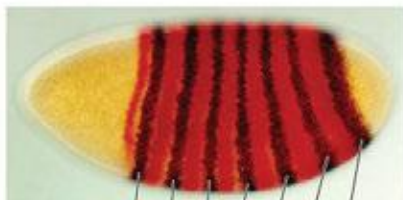
Segmentation Genes Gap genes, pair-rule genes, and segment polarity genes, three subclasses of segmentation genes in *Drosophila*, are activated at successive stages of embryonic development. Expression of the gap genes is generally regulated by the products of one or more maternal genes. At least some of the gap genes encode transcription factors that affect the expression of other segmentation or (later) homeotic genes.

One well-characterized segmentation gene is *fushi tarazu* (*ftz*), of the pair-rule subclass. When *ftz* is deleted, the embryo develops 7 segments instead of the normal 14, each segment twice the normal width. The Fushi-tarazu protein (Ftz) is a transcription activator with a homeodomain. The mRNAs and proteins derived from the normal *ftz* gene accumulate in a striking pattern of seven stripes that encircle the posterior two-thirds of the embryo (Fig. 28–40). The stripes demarcate the positions of segments that develop later; these segments are eliminated if *ftz* function is lost. The Ftz protein and a few similar regulatory proteins directly or indirectly regulate the expression of vast numbers of genes in the continuing developmental cascade.

Homeotic Genes A set of 8 to 11 homeotic genes directs the formation of particular structures at specific locations in the body plan. These genes are now more commonly referred to as **Hox genes**, the term derived from “homeobox,” the conserved gene sequence that

encodes the homeodomain and is present in all of these genes. Despite the name, these are not the only development-related proteins to include a homeodomain (for example, the *bicoid* gene product described above has a homeodomain), and “Hox” is more a functional than a structural classification. The *Hox* genes are organized in genomic clusters. *Drosophila* has one such cluster and mammals have four (Fig. 28–41). The genes in these clusters are remarkably similar from nematodes to humans. In *Drosophila*, each of the *Hox* genes is

(a) Side view



(b) Cross-sectional dorsal view

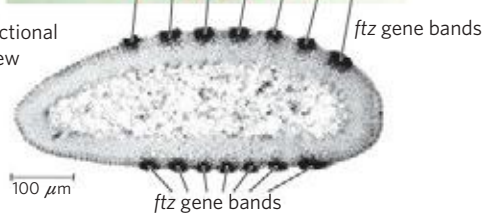
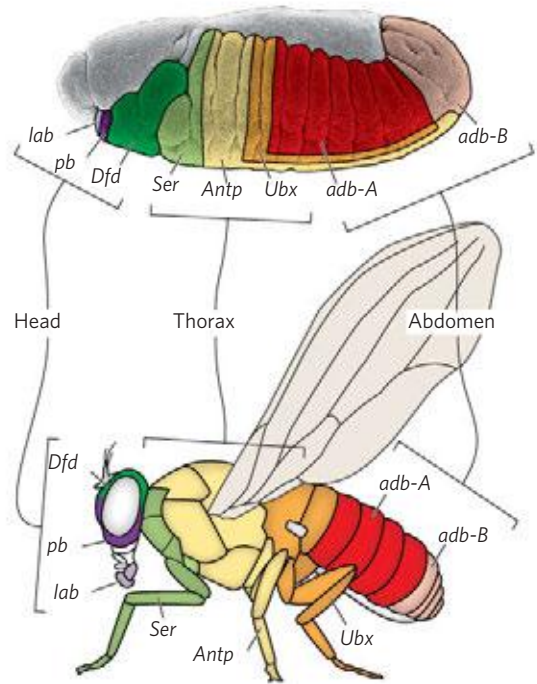


FIGURE 28–40 Distribution of the *fushi tarazu* (*ftz*) gene product in early *Drosophila* embryos. (a) Following a gene-specific staining procedure, the gene product can be detected in seven bands around the circumference of the embryo. These bands (b) appear as dark spots (generated by a radioactive label) in a cross-sectional autoradiograph and demarcate the anterior margins of the segments that will appear in the late embryo.

(a)



(b)

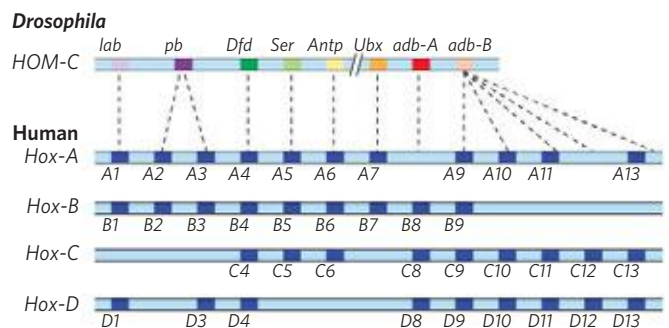


FIGURE 28–41 The *Hox* gene clusters and their effects on development. (a) Each *Hox* gene in the fruit fly is responsible for the development of structures in a defined part of the body and is expressed in defined regions of the embryo, as shown here with color coding. (b) *Drosophila* has one *Hox* gene cluster; the human genome has four. Many of these genes are highly conserved in multicellular animals. Evolutionary relationships, as indicated by sequence alignments, between genes in the *Drosophila* *Hox* gene cluster and those in the mammalian *Hox* gene clusters are shown by dashed lines. Similar relationships among the four sets of mammalian *Hox* genes are indicated by vertical alignment.

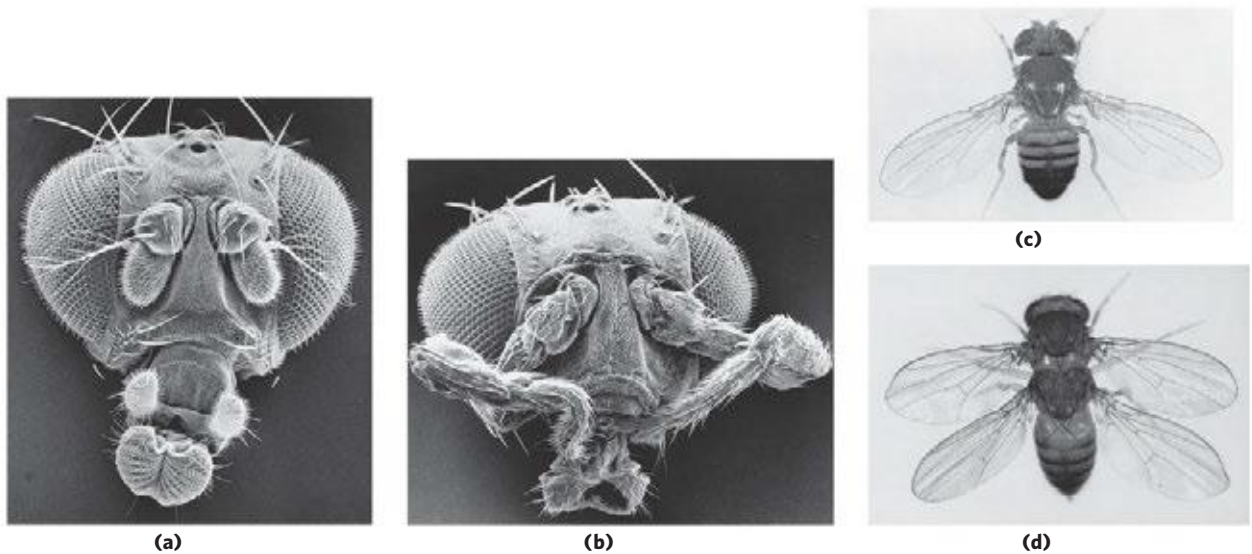


FIGURE 28-42 Effects of mutations in *Hox* genes in *Drosophila*. (a) Normal head. (b) Homeotic mutant (*antennapedia*) in which antennae are replaced by legs. (c) Normal body structure. (d) Homeotic mutant

(*bithorax*) in which a segment has developed incorrectly to produce an extra set of wings.

expressed in a particular segment of the embryo and controls the development of the corresponding part of the mature fly. The terminology used to describe *Hox* genes can be confusing. They have historical names in the fruit fly (for example, *ultrabithorax*), whereas in mammals they are designated by two competing systems based on lettered (A, B, C, D) or numbered (1, 2, 3, 4) clusters.

Loss of *Hox* genes in fruit flies by mutation or deletion causes the appearance of a normal appendage or body structure at an inappropriate body position. An important example is the *ultrabithorax* (*ubx*) gene. When *Ubx* function is lost, the first abdominal segment develops incorrectly, having the structure of the third thoracic segment. Other known homeotic mutations cause the formation of an extra set of wings, or two legs at the position in the head where the antennae are normally found (Fig. 28-42). The *Hox* genes often span long regions of DNA. The *ubx* gene, for example, is 77,000 bp long. More than 73,000 bp of this gene are in introns, one of which is more than 50,000 bp long. Transcription of the *ubx* gene takes nearly an hour. The delay this imposes on *ubx* gene expression is believed to be a timing mechanism involved in the temporal regulation of subsequent steps in development. Many *Hox* genes are further regulated by miRNAs encoded by intergenic regions of the *Hox* gene clusters. All of the *Hox* gene products are themselves transcription factors that regulate the expression of an array of downstream genes. Identification of these downstream targets is ongoing.

Many of the principles of development outlined above apply to other eukaryotes, from nematodes to humans. Some of the regulatory proteins are conserved.

For example, the products of the homeobox-containing genes *HOXA7* in mouse and *antennapedia* in fruit fly differ in only one amino acid residue. Of course, although the molecular regulatory mechanisms may be similar, many of the ultimate developmental events are not conserved (humans do not have wings or antennae). The different outcomes are brought about by differences in the downstream target genes controlled by the *Hox* genes. The discovery of structural determinants with identifiable molecular functions is the first step in understanding the molecular events underlying development. As more genes and their protein products are discovered, the biochemical side of this vast puzzle will be elucidated in increasingly rich detail.

Stem Cells Have Developmental Potential That Can Be Controlled

If we can understand development, and the mechanisms of gene regulation behind it, we can control it. An adult human has many different types of tissues. Many of the cells are terminally differentiated and no longer divide. If an organ malfunctions due to disease or a limb is lost in an accident, the tissues are not readily replaced. Most cells, because of the regulatory processes that are in place, or even the loss of some or all genomic DNA, are not easily reprogrammed. Medical science has made organ transplants possible, but organ donors are a limited resource and organ rejection remains a major medical problem. If humans could regenerate their own organs or limbs or nervous tissue, rejection would no longer be an issue. Real cures for kidney failure or neurodegenerative disorders could become reality.

The key to tissue regeneration lies in **stem cells**—cells that have retained the capacity to differentiate into various tissues. In humans, after an egg is fertilized, the first few cell divisions create a ball of **totipotent** cells (the morula), cells that have the capacity to differentiate individually into any tissue or even into a complete organism (**Fig. 28–43**). Continued cell

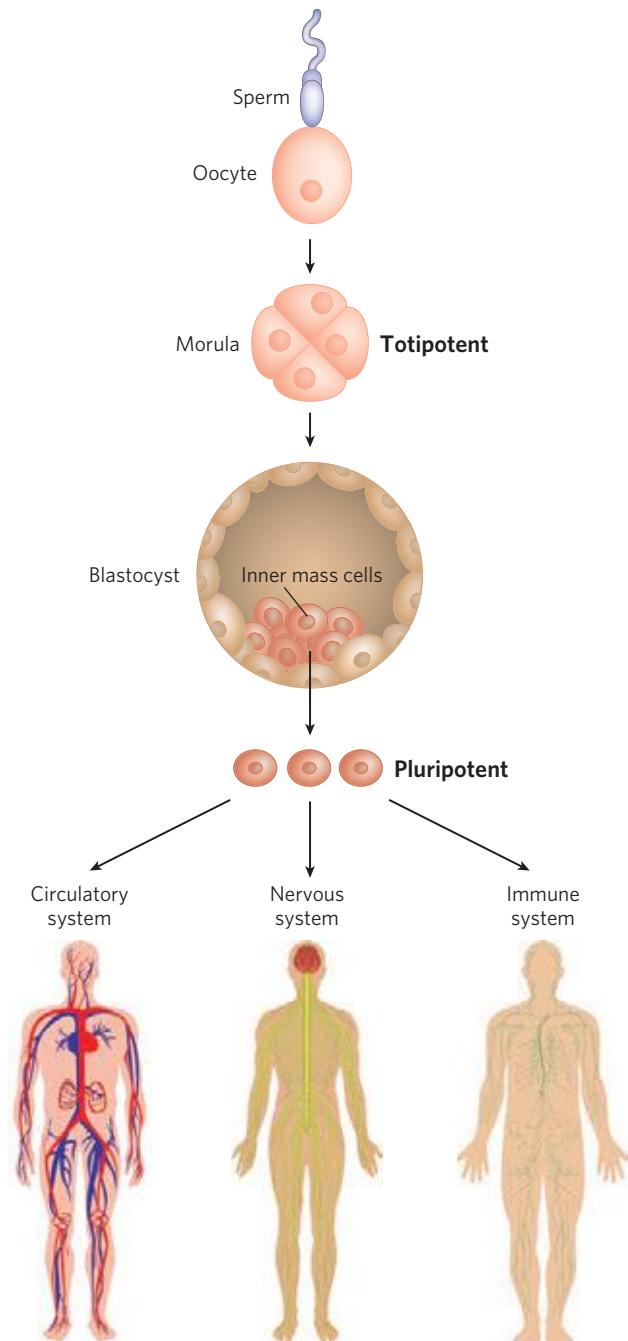


FIGURE 28–43 Totipotent and pluripotent stem cells. Cells of the morula stage are totipotent and have the capacity to differentiate into a complete organism. The source of pluripotent embryonic stem cells is the inner mass cells of the blastocyst. Pluripotent cells give rise to many tissue types but cannot form complete organisms.

division produces a hollow ball, a blastocyst. The outer cells of the blastocyst eventually form the placenta. The inner layers form the germ layers of the developing fetus—the ectoderm, mesoderm, and endoderm. These cells are **pluripotent**: they can give rise to cells of all three germ layers and can differentiate into many types of tissues. However, they cannot differentiate into a complete organism. Some of these cells are **unipotent**: they can develop into only one type of cell and/or tissue. It is the pluripotent cells of the blastocyst, the **embryonic stem cells**, that are currently used in embryonic stem cell research.

Stem cells have two functions: to replenish themselves and, at the same time, provide cells that can differentiate. These tasks are accomplished in multiple ways (**Fig. 28–44a**). All or parts of the stem cell population can, in principle, be involved in replenishment, differentiation, or both.

Other types of stem cells can potentially be used for medical benefit. In the adult organism, **adult stem cells**, as products of additional differentiation, have a more limited potential for further development than do embryonic stem cells. For example, the hematopoietic stem cells of bone marrow can give rise to many types of blood cells and also to cells with the capacity to regenerate bone. They are referred to as **multipotent**. However, these cells cannot differentiate into a liver or kidney or neuron. Adult stem cells are often said to have a **niche**, a microenvironment that promotes stem cell maintenance while allowing differentiation of some daughter cells as replacements for cells in the tissue they serve (**Fig. 28–44b**). Hematopoietic stem cells in the bone marrow occupy a niche in which signaling from neighboring cells and other cues maintain the stem cell lineage. At the same time, some daughter cells differentiate to provide needed blood cells. Understanding the niche in which stem cells operate, and the signals the niche provides, is essential in efforts to harness the potential of stem cells for tissue regeneration.

All stem cells have problems with respect to human medical applications. Adult stem cells have a limited capacity to regenerate tissues, are generally present in small numbers, and are hard to isolate from an adult human. Embryonic stem cells have much greater differentiation potential and can be cultured to generate large numbers of cells. However, their use is accompanied by ethical concerns related to the necessary destruction of human embryos. Identifying a source of plentiful and medically useful stem cells that does not raise concerns remains a major goal of medical research.

Our ability to culture stem cells (i.e., maintain them in an undifferentiated state), and to manipulate them to grow and differentiate into particular tissues, is very much a function of our understanding of developmental biology. The identification and culturing of pluripotent

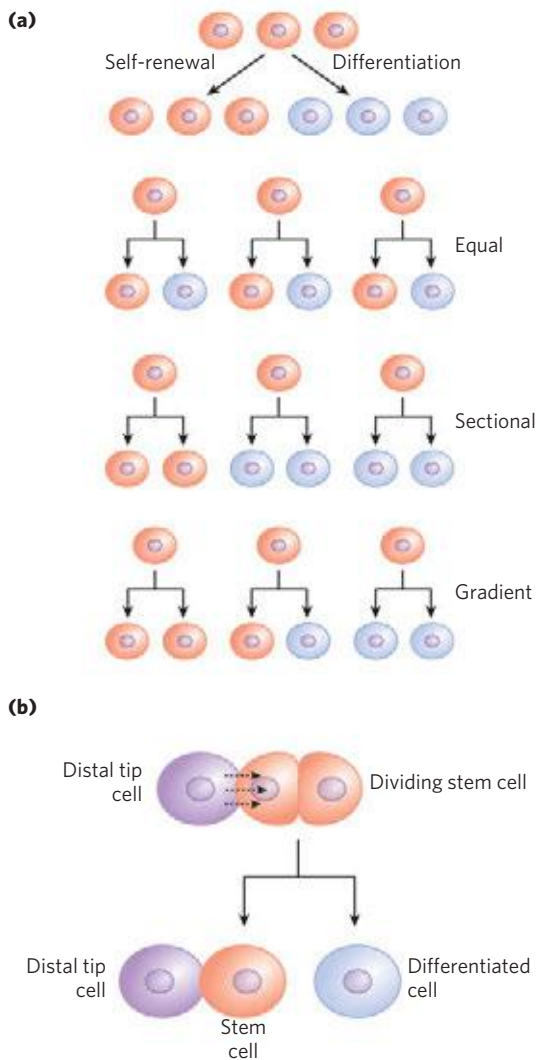


FIGURE 28-44 Stem cell proliferation versus differentiation and development. Stem cells must strike a balance between self-renewal and differentiation. **(a)** Some possible cell division patterns that allow for replenishment of stem cells and production of some differentiated cells. Each cell may produce one stem cell and one differentiated cell, or two differentiated cells, or two stem cells in defined parts of the tissue or culture. Or a gradient of growth conditions can be established, with cell fates differing from one end of the gradient to the other. **(b)** Establishing a developmental niche through stem cell contact with a cell or group of cells. Molecular signals provided by the niche cells (in this case, for plants, a distal tip cell) help orient the mitotic spindle for stem cell division and ensure that one daughter cell retains stem cell properties.



James Thomson

stem cells from human blastocysts was reported by James Thomson and colleagues in 1998. This advance led to the long-term availability of established cell lines for research.

Thus far, mouse and human embryonic stem cells have been used for most research. Although both types

of stem cells are pluripotent, they require very different culture conditions, optimized to allow cell division indefinitely without differentiation. Mouse embryonic stem cells are grown on a layer of gelatin and require the presence of leukemia inhibitory factor (LIF). Human embryonic stem cells are grown on a feeder layer of mouse embryonic fibroblasts and require basic fibroblast growth factor (bFGF or FGF2). The use of a feeder cell layer implies that the mouse cells are providing a diffusible product or some surface signal, not yet known, that is needed by human stem cells to either promote cell division or prevent differentiation.

A significant advance, reported in 2007, centers on success in reversing differentiation. In effect, skin cells—first from mice, then from humans—have been reprogrammed to take on the characteristics of pluripotent stem cells. The reprogramming involves manipulations to get the cells to express at least four transcription factors (Oct4, Sox2, Nanog, and Lin28), all of which are known to help maintain the stem cell-like state. Gradual improvements in this technology may make the harvesting of embryonic stem cells unnecessary and provide a source of stem cells that is genetically matched to a prospective patient.

Our discussion of developmental regulation and stem cells brings us full circle, back to a biochemical beginning—both figuratively and literally. Evolution appropriately provides the first and last words of text in this book. If evolution is to generate the kind of changes in an organism that we associate with a different species, it is the developmental program that must be affected. Developmental and evolutionary processes are closely allied, each informing the other (Box 28-1). The continuing study of biochemistry has everything to do with enriching the future of humanity and understanding our origins.

SUMMARY 28.3 Regulation of Gene Expression in Eukaryotes

- ▶ In eukaryotes, positive regulation is more common than negative regulation, and transcription is accompanied by large changes in chromatin structure.
- ▶ Promoters for Pol II typically have a TATA box and Inr sequence, as well as multiple binding sites for transcription activators. The latter sites, sometimes located hundreds or thousands of base pairs away from the TATA box, are called upstream activator sequences in yeast and enhancers in higher eukaryotes.
- ▶ Large complexes of proteins are generally required to regulate transcriptional activity. The effects of transcription activators on Pol II are mediated by coactivator protein complexes such

BOX 28–1 METHODS Of Fins, Wings, Beaks, and Things

South America has several species of seed-eating finches, commonly called grassquits. About 3 million years ago, a small group of grassquits, of a single species, took flight from the continent's Pacific coast. Perhaps driven by a storm, they lost sight of land and traveled nearly 1,000 km. Small birds such as these might easily have perished on such a journey, but the smallest of chances brought this group to a newly formed volcanic island in an archipelago later to be known as the Galápagos. It was a virgin landscape with untapped plant and insect food sources, and the newly arrived finches survived. Over the years, new islands formed and were colonized by new plants and insects—and by the finches. The birds exploited the new resources on the islands, and groups of birds gradually specialized and diverged into new species. By the time Charles Darwin stepped onto the islands in 1835, there were many different finch species to be found on the various islands of the archipelago, feeding on seeds, fruits, insects, pollen, or even blood.

The diversity of living creatures was a source of wonder for humans long before scientists sought to understand its origins. The extraordinary insight handed down to us by Darwin, inspired in part by his encounter with the Galápagos finches, provided a broad explanation for the existence of organisms with a vast array of appearances and characteristics. It also gave rise to many questions about the mechanisms underlying evolution. Answers to those questions have started to appear, first through the study of genomes and nucleic acid metabolism in the last half of the twentieth century, and more recently through an emerging field nicknamed *evo-devo*—a blend of evolutionary and developmental biology.

In its modern synthesis, the theory of evolution has two main elements: mutations in a population generate genetic diversity; natural selection then acts on this diversity to favor individuals with more useful genomic tools and to disfavor others. Mutations occur at significant rates in every individual's genome, in every cell (see Section 8.3 and Fig. 25–20). Advantageous mutations in single-celled organisms or in the

germ line of multicellular organisms can be inherited, and they are more likely to be inherited (that is, are passed on to greater numbers of offspring) if they confer an advantage. It is a straightforward scheme. But many have wondered whether it is enough to explain, say, the many different beak shapes in the Galápagos finches or the diversity of size and shape among mammals. Until recent decades, there were several widely held assumptions about the evolutionary process: that many mutations and new genes would be needed to bring about a new physical structure, that more-complex organisms would have larger genomes, and that very different species would have few genes in common. All of these assumptions were wrong.

Modern genomics has revealed that the human genome contains fewer genes than expected—not many more than the fruit fly genome and fewer than some amphibian genomes. The genomes of every mammal, from mouse to human, are surprisingly similar in the number, types, and chromosomal arrangement of genes. Meanwhile, *evo-devo* is telling us how complex and very different creatures can evolve within these genomic realities.

The kinds of mutant organisms shown in Figure 28–42 were studied by the English biologist William Bateson in the late nineteenth century. Bateson used his observations to challenge the Darwinian notion that evolutionary change would have to be gradual. Recent studies of the genes that control organismal development have put an exclamation point on Bateson's ideas. Subtle changes in regulatory patterns during development, reflecting just one or a few mutations, can result in startling physical changes and fuel surprisingly rapid evolution.

The Galápagos finches provide a wonderful example of the link between evolution and development. There are at least 14 (some specialists list 15) species of Galápagos finches, distinguished in large measure by their beak structure. The ground finches, for example, have broad, heavy beaks adapted to crushing large, hard seeds. The cactus finches have longer, slender beaks ideal for probing cactus fruits

as Mediator. The modular structures of the activators have distinct activation and DNA-binding domains. Other protein complexes, including histone acetyltransferases and ATP-dependent complexes such as SWI/SNF and NURF, reversibly remodel and modify chromatin structure.

- ▶ Hormones affect the regulation of gene expression in one of two ways. Steroid

hormones interact directly with intracellular receptors that are DNA-binding regulatory proteins; binding of the hormone has either positive or negative effects on the transcription of genes targeted by the hormone. Nonsteroid hormones bind to cell surface receptors, triggering a signaling pathway that can lead to phosphorylation of a regulatory protein, affecting its activity.

and flowers (Fig. 1). Clifford Tabin and colleagues carefully surveyed a set of genes expressed during avian craniofacial development. They identified a single gene, *Bmp4*, whose expression level correlated with formation of the more robust beaks of the ground finches. More robust beaks were also formed in chicken embryos when high levels of *Bmp4* were artificially expressed in the appropriate tissues, confirming the importance of *Bmp4*. In a similar study, the formation of long, slender beaks was linked to the expression of calmodulin (see Fig. 12–11) in particular tissues at appropriate developmental stages. Thus, major changes in the shape and function of the beak can be brought about by subtle changes in the expression of just two genes involved in developmental regulation. Very few mutations are required, and the needed mutations affect regulation. New genes are *not* required.

The system of regulatory genes that guides development is remarkably conserved among all vertebrates. Elevated expression of *Bmp4* in the right tissue at the right time leads to more robust jaw parts in zebrafish. The same gene plays a key role in tooth

development in mammals. The development of eyes is triggered by the expression of a single gene, *Pax6*, in fruit flies and in mammals. The mouse *Pax6* gene will trigger the development of fruit fly eyes in the fruit fly, and the fruit fly *Pax6* gene will trigger the development of mouse eyes in the mouse. In each organism, these genes are part of the much larger regulatory cascade that ultimately creates the correct structures in the correct locations in each organism. The cascade is ancient; for example, the *Hox* genes (described in the text) have been part of the developmental program of multicellular eukaryotes for more than 500 million years. Subtle changes in the cascade can have large effects on development, and thus on the ultimate appearance, of the organism. These same subtle changes can fuel remarkably rapid evolution. For example, the 400 to 500 described species of cichlids (spiny-finned fish) in Lake Malawi and Lake Victoria on the African continent are all derived from one or a few populations that colonized each lake in the past 100,000 to 200,000 years. The Galápagos finches simply followed a path of evolution and change that living creatures have been traveling for billions of years.

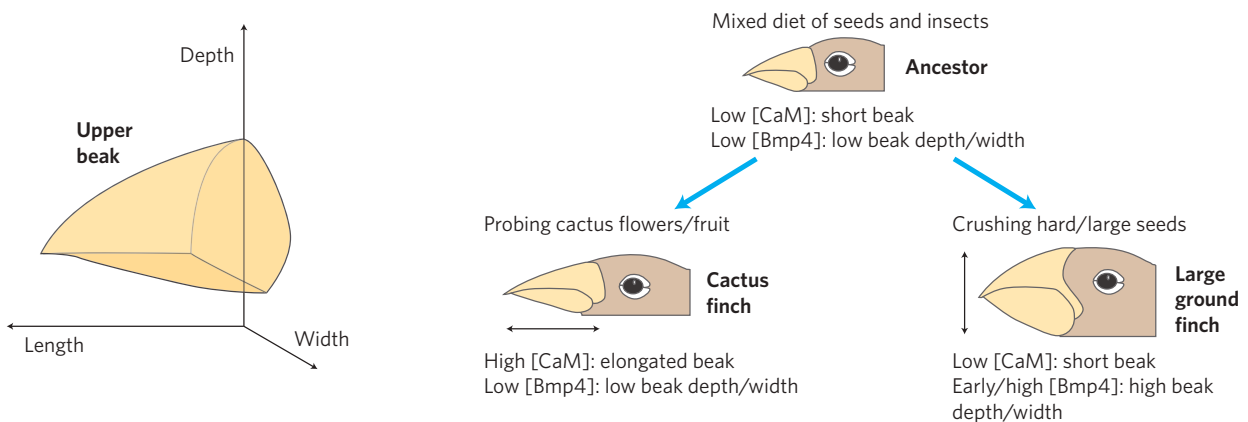


FIGURE 1 Evolution of new beak structures to exploit new food sources. In the Galápagos finches, the different beak structures of the cactus finch and the large ground finch, which feed on different,

specialized food sources, were produced to a large extent by a few mutations that altered the timing and level of expression of just two genes: those encoding calmodulin (CaM) and *Bmp4*.

- ▶ RNA-mediated regulation plays an important role in eukaryotic gene expression, with the range of known mechanisms expanding.
- ▶ Development of a multicellular organism presents the most complex regulatory challenge. The fate of cells in the early embryo is determined by establishment of anterior-posterior and dorsal-ventral gradients of proteins that act as transcription activators or translational repressors,

regulating the genes required for the development of structures appropriate to a particular part of the organism. Sets of regulatory genes operate in temporal and spatial succession, transforming given areas of an egg cell into predictable structures in the adult organism.

- ▶ The differentiation of stem cells into functional tissues can be controlled by extracellular signals and conditions.

Key Terms

Terms in bold are defined in the glossary

housekeeping genes 1156	upstream activator sequences (UASs) 1178
induction 1156	transcription
repression 1156	activators 1178
specificity factor 1157	coactivators 1178
repressor 1157	basal transcription factors 1178
activator 1157	high mobility group (HMG) 1179
operator 1157	Mediator 1179
negative regulation 1157	preinitiation complex (PIC) 1179
positive regulation 1157	TATA-binding protein (TBP) 1179
architectural regulator 1158	hormone response elements (HREs) 1182
operon 1159	microRNAs (miRNAs) 1185
helix-turn-helix 1162	RNA interference (RNAi) 1185
zinc finger 1162	ncRNAs 1186
homeodomain 1163	polarity 1186
homeobox 1163	metamerism 1187
leucine zipper 1163	maternal genes 1188
basic helix-loop-helix 1163	maternal mRNAs 1188
combinatorial control 1164	segmentation genes 1188
cAMP receptor protein (CRP) 1165	gap genes 1188
regulon 1166	pair-rule genes 1188
transcription	segment polarity genes 1188
attenuation 1167	homeotic genes 1188
translational repressor 1170	<i>Hox</i> genes 1190
stringent response 1171	totipotent 1192
riboswitch 1172	pluripotent 1192
phase variation 1173	unipotent 1192
chromatin remodeling 1175	embryonic stem cells 1192
SWI/SNF 1175	
histone acetyltransferases (HATs) 1176	
enhancers 1178	

Further Reading

General

Carroll, S.B. (2005) *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom*, W. W. Norton & Company, New York.

A fascinating look at how developmental biology informs evolutionary biology.

Neidhardt, F.C. (ed.). *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* (Curtiss, R., Ingraham, J.L., Lin, E.C.C., Magasanik, B., Low, K.B., Reznikoff, W.S., Riley, M., Schaechter, M., & Umberger, H.E., vol. eds), American Society for Microbiology, Washington, DC.

A good source for information on bacterial gene regulation and many other topics. Last published in a print edition in 1996, this is now an important and continuously updated online resource at <http://ecosal.org>.

Regulation of Gene Expression in Bacteria

Babitske, P., Baker, C.S., & Romeo, T. (2009) Regulation of translation initiation by RNA binding proteins. *Annu. Rev. Microbiol.* **63**, 27–44.

Jacob, F. & Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356.

The operon model and the concept of messenger RNA, first proposed in the *Proceedings* of the French Academy of Sciences in 1960, are presented in this historic paper.

Osterberg, S., del Peso-Santos, T., & Shingler, V. (2011) Regulation of alternative sigma factor use. *Annu. Rev. Microbiol.* **65**, 37–55.

Roth, A. & Breaker, R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.* **78**, 305–334.

Regulation of Gene Expression in Eukaryotes

Berezikov, E. (2011) Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* **12**, 846–860.

Conaway, R.C. & Conaway, J.W. (2011) Function and regulation of the Mediator complex. *Curr. Opin. Genet. Dev.* **21**, 225–230.

Fabian, M.R., Sonenberg, N., & Filipowicz, W. (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79**, 351–379.

Groppo, R. & Richter, J.D. (2009) Translational control from head to tail. *Curr. Opin. Cell Biol.* **21**, 444–451.

Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**, 533–543.

Krol, J., Loedige, I., & Filipowicz, W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* **11**, 597–610.

Pasquinelli, A.E. (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.* **13**, 271–282.

Prud'homme, B., Gompel, N., & Carroll, S.B. (2007) Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. USA* **104**, 8605–8612.

Rinn, J.L. & Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166.

Shahbazian, M.D. & Grunstein, M. (2007) Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem.* **76**, 75–100.

Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**, 1–4.

Talbert, P.B. & Henikoff, S. (2010) Histone variants—ancient wrap artists of the epigenome. *Nat. Rev. Mol. Cell Biol.* **11**, 264–275.

Werner, F. & Grohmann, D. (2011) Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98.

Zhou, Q., Li, T., & Price, D.H. (2012) RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**, 119–143.

Problems

1. Effect of mRNA and Protein Stability on Regulation
E. coli cells are growing in a medium with glucose as the sole carbon source. Tryptophan is suddenly added. The cells continue to grow, and divide every 30 min. Describe (qualitatively) how the amount of tryptophan synthase activity in the cells changes with time under the following conditions:

(a) The *trp* mRNA is stable (degraded slowly over many hours).

(b) The *trp* mRNA is degraded rapidly, but tryptophan synthase is stable.

(c) The *trp* mRNA and tryptophan synthase are both degraded rapidly.

2. The Lactose Operon A researcher engineers a *lac* operon on a plasmid but inactivates all parts of the *lac* operator (*lacO*) and the *lac* promoter, replacing them with the binding site for the LexA repressor (which acts in the SOS response) and a promoter regulated by LexA. The plasmid is introduced into *E. coli* cells that have a *lac* operon with an inactive *lacZ* gene. Under what conditions will these transformed cells produce β -galactosidase?

3. Negative Regulation Describe the probable effects on gene expression in the *lac* operon of a mutation in (a) the *lac* operator that deletes most of O_1 , (b) the *lacI* gene that inactivates the repressor, and (c) the promoter that alters the region around position -10 .

4. Specific DNA Binding by Regulatory Proteins A typical bacterial repressor protein discriminates between its specific DNA-binding site (operator) and nonspecific DNA by a factor of 10^4 to 10^6 . About 10 molecules of repressor per cell are sufficient to ensure a high level of repression. Assume that a very similar repressor existed in a human cell, with a similar specificity for its binding site. How many copies of the repressor would be required to elicit a level of repression similar to that in the bacterial cell? (Hint: The *E. coli* genome contains about 4.6 million bp; the human haploid genome has about 3.2 billion bp.)

5. Repressor Concentration in *E. coli* The dissociation constant for a particular repressor-operator complex is very low, about 10^{-13} M. An *E. coli* cell (volume 2×10^{-12} mL) contains 10 copies of the repressor. Calculate the cellular concentration of the repressor protein. How does this value compare with the dissociation constant of the repressor-operator complex? What is the significance of this answer?

6. Catabolite Repression *E. coli* cells are growing in a medium containing lactose but no glucose. Indicate whether each of the following changes or conditions would increase, decrease, or not change the expression of the *lac* operon. It may be helpful to draw a model depicting what is happening in each situation.

(a) Addition of a high concentration of glucose

(b) A mutation that prevents dissociation of the Lac repressor from the operator

(c) A mutation that completely inactivates β -galactosidase

(d) A mutation that completely inactivates galactoside permease

(e) A mutation that prevents binding of CRP to its binding site near the *lac* promoter

7. Transcription Attenuation How would transcription of the *E. coli trp* operon be affected by the following manipulations of the leader region of the *trp* mRNA?

(a) Increasing the distance (number of bases) between the leader peptide gene and sequence 2

(b) Increasing the distance between sequences 2 and 3

(c) Removing sequence 4

(d) Changing the two Trp codons in the leader peptide gene to His codons

(e) Eliminating the ribosome-binding site for the gene that encodes the leader peptide

(f) Changing several nucleotides in sequence 3 so that it can base-pair with sequence 4 but not with sequence 2

8. Repressors and Repression How would the SOS response in *E. coli* be affected by a mutation in the *lexA* gene that prevented autocatalytic cleavage of the LexA protein?

9. Regulation by Recombination In the phase variation system of *Salmonella*, what would happen to the cell if the Hin recombinase became more active and promoted recombination (DNA inversion) several times in each cell generation?

10. Initiation of Transcription in Eukaryotes A new RNA polymerase activity is discovered in crude extracts of cells derived from an exotic fungus. The RNA polymerase initiates transcription only from a single, highly specialized promoter. As the polymerase is purified its activity declines, and the purified enzyme is completely inactive unless crude extract is added to the reaction mixture. Suggest an explanation for these observations.

11. Functional Domains in Regulatory Proteins A biochemist replaces the DNA-binding domain of the yeast Gal4 protein with the DNA-binding domain from the Lac repressor and finds that the engineered protein no longer regulates transcription of the *GAL* genes in yeast. Draw a diagram of the different functional domains you would expect to find in the Gal4 protein and in the engineered protein. Why does the engineered protein no longer regulate transcription of the *GAL* genes? What might be done to the DNA-binding site recognized by this chimeric protein to make it functional in activating transcription of *GAL* genes?

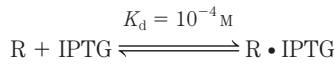
12. Nucleosome Modification during Transcriptional Activation To prepare genomic regions for transcription, certain histones in the resident nucleosomes are acetylated and methylated at specific locations. Once transcription is no longer needed, these modifications need to be reversed. In mammals, the methylation of Arg residues in histones is reversed by peptidylarginine deiminases (PADIs). The reaction promoted by these enzymes does not yield unmethylated arginine. Instead, it produces citrulline residues in the histone. What is the other product of the reaction? Suggest a mechanism for this reaction.

13. Inheritance Mechanisms in Development A *Drosophila* egg that is *bcd*⁻/*bcd*⁻ may develop normally, but the adult fruit fly will not be able to produce viable offspring. Explain.

Data Analysis Problem

14. Engineering a Genetic Toggle Switch in *Escherichia coli* Gene regulation is often described as an “on or off” phenomenon—a gene is either fully expressed or not expressed at all. In fact, repression and activation of a gene involve ligand-binding reactions, so genes can show intermediate levels of expression when intermediate levels of regulatory molecules are present. For example, for the *E. coli lac* operon, consider the binding equilibrium of the Lac repressor, operator

DNA, and inducer (see Fig. 28–8). Although this is a complex, cooperative process, it can be approximately modeled by the following reaction (R is repressor; IPTG is the inducer isopropyl- β -D-thiogalactoside):



Free repressor, R, binds to the operator and prevents transcription of the *lac* operon; the R • IPTG complex does not bind to the operator and thus transcription of the *lac* operon can proceed.

(a) Using Equation 5–8, we can calculate the relative expression level of the proteins of the *lac* operon as a function of [IPTG]. Use this calculation to determine over what range of [IPTG] the expression level would vary from 10% to 90%.

(b) Describe qualitatively the level of *lac* operon proteins present in an *E. coli* cell before, during, and after induction with IPTG. You need not give the amounts at exact times—just indicate the general trends.

Gardner, Cantor, and Collins (2000) set out to make a “genetic toggle switch”—a gene-regulatory system with two key characteristics of a light switch. (A) *It has only two states*: it is either fully on or fully off; it is not a dimmer switch. In biochemical terms, the target gene or gene system (operon) is either fully expressed or not expressed at all; it cannot be expressed at an intermediate level. (B) *Both states are stable*: although you must use a finger to flip the light switch from one state to the other, once you have flipped it and removed your finger, the switch stays in that state. In biochemical terms, exposure to an inducer or some other signal changes the expression state of the gene or operon, and it remains in that state once the signal is removed.

(c) Explain how the *lac* operon lacks both characteristics A and B.

To make their “toggle switch,” Gardner and coworkers constructed a plasmid from the following components:

OP_{lac} The operator-promoter region of the *E. coli lac* operon

OP_{λ} The operator-promoter region of λ phage

lacI The gene encoding the *lac* repressor protein, LacI. In the absence of IPTG, this protein strongly represses OP_{lac} ; in the presence of IPTG, it allows full expression from OP_{lac} .

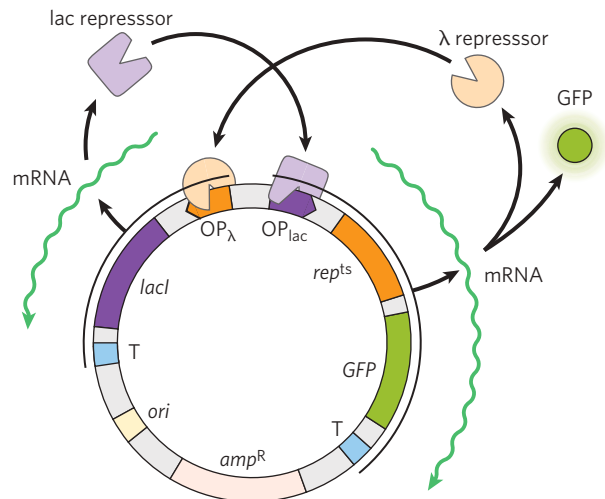
rep^{ts} The gene encoding a temperature-sensitive mutant λ repressor protein, rep^{ts} . At 37°C this protein strongly represses OP_{λ} ; at 42°C it allows full expression from OP_{λ} .

GFP The gene for green fluorescent protein (GFP), a highly fluorescent reporter protein (see Fig. 9–16)

T Transcription terminator

The investigators arranged these components (see figure below) so that the two promoters were reciprocally repressed: OP_{lac} controlled expression of rep^{ts} , and OP_{λ} controlled expression of *lacI*. The state of this system was reported by the expression level of *GFP*, which was also under the control of OP_{lac} .

(d) The constructed system has two states: GFP-on (high level of expression) and GFP-off (low level of expression). For each state, describe which proteins are present and which promoters are being expressed.

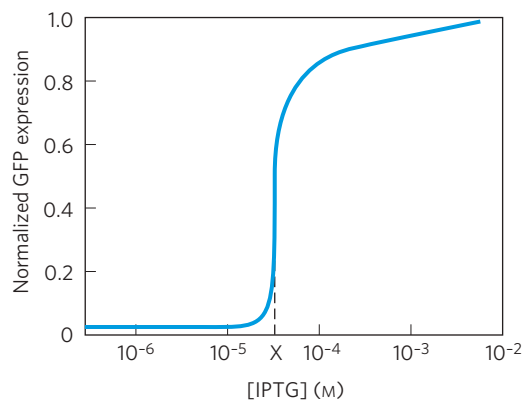


(e) Treatment with IPTG would be expected to toggle the system from one state to the other. From which state to which? Explain your reasoning.

(f) Treatment with heat (42°C) would be expected to toggle the system from one state to the other. From which state to which? Explain your reasoning.

(g) Why would this plasmid be expected to have characteristics A and B as described above?

To confirm that their construct did indeed exhibit these characteristics, Gardner and colleagues first showed that, once switched, the GFP expression level (high or low) was stable for long periods of time (characteristic B). Next, they measured GFP level at different concentrations of the inducer IPTG, with the following results.



They noticed that the average GFP expression level was intermediate at concentration X of IPTG. However, when they measured the GFP expression level *in individual cells* at [IPTG] = X, they found either a high level or a low level of GFP—no cells showed an intermediate level.

(h) Explain how this finding demonstrates that the system has characteristic A. What is happening to cause the bimodal distribution of expression levels at [IPTG] = X?

Reference

Gardner, T.S., Cantor, C.R., & Collins, J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342.