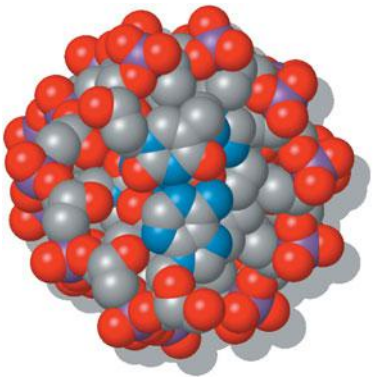



# Nucleic Acids

*We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.*

—J.D. Watson and F.H.C. Crick (1953)

The discovery of the substance that proved to be deoxyribonucleic acid (DNA) was made in 1869 by Friedrich Miescher, a young Swiss physician working in the laboratory of the German physiological chemist Felix Hoppe-Seyler. Miescher treated white blood cells (which came from the pus on discarded surgical bandages) with hydrochloric acid to obtain nuclei for study. When the nuclei were subsequently treated with acid, a precipitate formed that contained carbon, hydrogen, oxygen, nitrogen, and a high percentage of phosphorus. Miescher called the precipitate "nuclein" because it came from nuclei. Later, when it was found to be strongly acidic, its name was changed to nucleic acid. Although he did not know it, Miescher had discovered DNA. Soon afterward, Hoppe-Seyler isolated a similar substance from yeast cells—this substance is now known to be ribonucleic acid (RNA). Both DNA and RNA are polymers of nucleotides, or polynucleotides.

In 1944 Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that DNA is the molecule that carries genetic information. At the time, very little was known about the three-dimensional structure of this important molecule. Over the next few years, the structures of nucleotides were determined and in 1953 James D. Watson and Francis H. C. Crick proposed their model for the structure of double-stranded DNA.

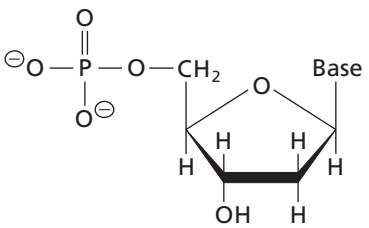
The study of nucleic acid biochemistry has advanced considerably in the past few decades. Today it is possible not only to determine the sequence of your genome but also to synthesize large chromosomes in the laboratory. It has become routine to clone and manipulate DNA molecules. This has led to spectacular advances in our understanding of molecular biology and the ways information contained in DNA is expressed in living cells.

We now know that a living organism contains a set of instructions for every step required to construct a replica of itself. This information resides in the genetic material, or **genome**, of the organism. The genomes of all cells are composed of DNA but some viral genomes are composed of RNA. A genome may consist of a single molecule of DNA, as in many species of bacteria. The genome of eukaryotes is one complete set of DNA molecules found in the nucleus (i.e., the haploid set of chromosomes in diploid organisms). By convention, the genome of a species does not include mitochondrial and chloroplast DNA. With rare exception, no two individuals in a species have exactly

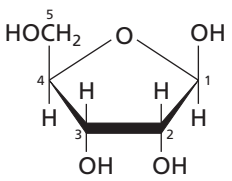
**Top:** Space-filling model of DNA, viewed along the helix axis.

▲ James D. Watson (1928–) (left) and Francis H. C. Crick (1916–2004) (right) describing the structure of DNA in 1953.

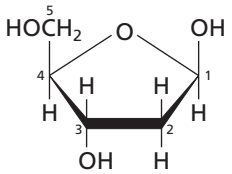
**The distinction between the normal flow of information and the Central Dogma of Molecular Biology is explained in Section 1.1 and the introduction to Chapter 21.**




▲ **Figure 19.1**
**Chemical structure of a nucleotide.** Nucleotides contain a five-carbon sugar, a nitrogenous base, and at least one phosphate group. The sugar can be either deoxyribose (shown here) or ribose.




Ribose
(β-D-Ribofuranose)




Deoxyribose
(2-Deoxy-β-D-ribofuranose)

▲ **Figure 19.2**
**Chemical structures of the two sugars found in nucleotides. (a)** Ribose (β-D-ribofuranose). **(b)** Deoxyribose (2-deoxy-β-D-ribofuranose).

the same genome sequence. If they were alive today, Miescher and Hoppe-Seyler would be astonished to learn that criminals could be convicted by DNA fingerprinting and that we have sequenced the complete genomes of thousands of species, including humans.

In general, the information that specifies the primary structure of a protein is encoded in the sequence of nucleotides in DNA. This information is enzymatically copied during the synthesis of RNA, a process known as transcription. Some of the information contained in the transcribed RNA molecules is translated during the synthesis of polypeptide chains that are then folded and assembled to form protein molecules. Thus, we can generalize that the biological information stored in a cell's DNA flows from DNA to RNA to protein.

Nucleic acids are the fourth major class of macromolecules that we study in this book. Like proteins and polysaccharides, they contain multiple similar monomeric units that are covalently joined to produce large polymers. In this chapter we describe the structure of nucleic acids and how they are packaged in cells. We also describe some of the enzymes that use DNA and RNA as substrates. Many other proteins and enzymes interact with DNA and RNA in order to ensure that genetic information is correctly interpreted. We will consider the biochemistry and the regulation of this flow of information in Chapters 20 to 22.

## 19.1 Nucleotides Are the Building Blocks of Nucleic Acids

Nucleic acids are polynucleotides, or polymers of nucleotides. As we saw in the previous chapter, nucleotides have three components: a five-carbon sugar, one or more phosphate groups, and a weakly basic nitrogenous compound called a base (Figure 19.1). The bases found in nucleotides are substituted pyrimidines and purines. The pentose is either ribose (D-ribofuranose) or 2-deoxyribose (2-deoxy-D-ribofuranose). The pyrimidine or purine *N*-glycosides of these sugars are called nucleosides. Nucleotides are the phosphate esters of nucleosides—the common nucleotides contain from one to three phosphoryl groups. Nucleotides containing ribose are called ribonucleotides and nucleotides containing deoxyribose are called deoxyribonucleotides (Section 18.5).

### A. Ribose and Deoxyribose

The sugar components of the nucleotides found in nucleic acids are shown in Figure 19.2. Both sugars are shown as Haworth projections of the β-conformation of the furanose ring forms (Section 8.2). This is the stable conformation found in nucleotides and polynucleotides. Each of these furanose rings can adopt different conformations such as the envelope forms discussed in Chapter 8. The 2′-endo conformation of deoxyribose predominates in double-stranded DNA (Figure 8.11).

### B. Purines and Pyrimidines

The bases found in nucleotides are derivatives of either pyrimidine or purine (Chapter 18). The structures of these heterocyclic compounds and the numbering systems for the carbon and nitrogen atoms of each base are shown in Figure 19.3. Pyrimidine has a single ring containing four carbon and two nitrogen atoms. Purine has a fused pyrimidine-imidazole ring system. Both types of bases are unsaturated, with conjugated double bonds. This feature makes the rings planar and also accounts for their ability to absorb ultraviolet light.

Substituted purines and pyrimidines are ubiquitous in living cells but the unsubstituted bases are seldom encountered in biological systems. The major pyrimidines that occur in nucleotides are uracil (2,4-dioxopyrimidine, U), thymine (2,4-dioxo-5-methylpyrimidine, T), and cytosine (2-oxo-4-aminopyrimidine, C). The major purines are adenine (6-aminopurine, A) and guanine (2-amino-6-oxopurine, G). The chemical structures of these five major bases are shown in Figure 19.4. Note that thymine can also be called 5-methyluracil because it is a substituted form of uracil (Section 18.6).

Adenine, guanine, and cytosine are found in both ribonucleotides and deoxyribonucleotides. Uracil is found mainly in ribonucleotides and thymine is found mainly in deoxyribonucleotides.

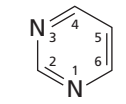
Purines and pyrimidines are weak bases and are relatively insoluble in water at physiological pH. Within cells, however, most pyrimidine and purine bases occur as constituents of nucleotides and polynucleotides and these compounds are highly soluble.

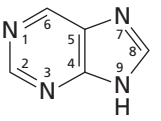
Each heterocyclic base can exist in at least two tautomeric forms. Adenine and cytosine can exist in either amino or imino forms. Guanine, thymine, and uracil can exist in either lactam (keto) or lactim (enol) forms (Figure 19.5). The tautomeric forms of each base exist in equilibrium but the amino and lactam tautomers are more stable and therefore predominate under the conditions found inside most cells. Note that the rings remain unsaturated and planar in each tautomer.

All of the bases in the common nucleotides can participate in hydrogen bonding. The amino groups of adenine and cytosine are hydrogen donors and the ring nitrogen atoms (N-1 in adenine and N-3 in cytosine) are hydrogen acceptors (Figure 19.6). Cytosine also has a hydrogen acceptor group at C-2. Guanine, cytosine, and thymine can form three hydrogen bonds. In guanine, the group at C-6 is a hydrogen acceptor while N-1 and the amino group at C-2 are hydrogen donors. In thymine, the groups at C-4 and C-2 are hydrogen acceptors and N-3 is a hydrogen donor. (Only two of these sites, C-4 and N-3, are used to form base pairs in DNA.) The hydrogen-bonding ability of uracil, a base found in RNA, is similar to that of thymine. The hydrogen-bonding patterns of bases have important consequences for the three-dimensional structure of nucleic acids.

Biochemistry textbooks in the 1940s usually depicted the bases in their imino and lactim forms. These were the structures that Jim Watson was using in 1953 to build a model of DNA. Shortly after being told by Jerry Donohue that the textbooks were wrong, Watson discovered the now-famous A/T and G/C base pairs.

Additional hydrogen bonding occurs in some nucleic acids and in nucleic acid–protein interactions. For example, N-7 of adenine and guanine can be a hydrogen acceptor and both amino hydrogen atoms of adenine, guanine, and cytosine can be donated to form hydrogen bonds.

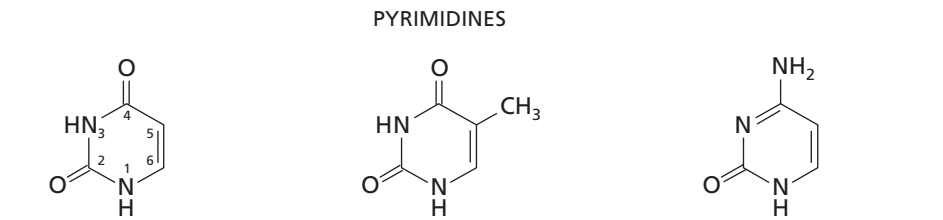


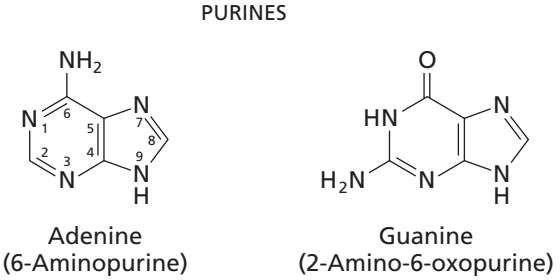




▲ **Figure 19.3**
**Chemical structures of pyrimidine and purine.**

## C. Nucleosides

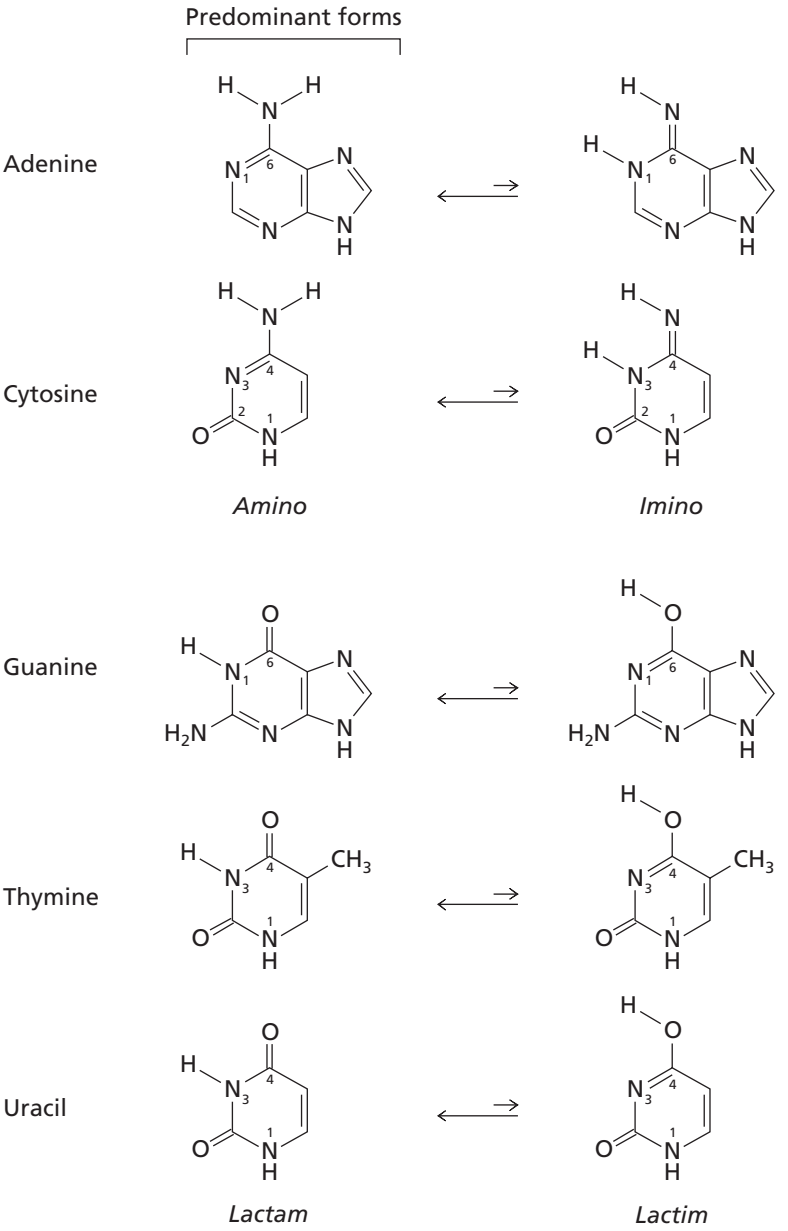
Nucleosides are composed of ribose or deoxyribose and a heterocyclic base. In each nucleoside, a *β*-*N*-glycosidic bond connects C-1 of the sugar to N-1 of the pyrimidine or N-9 of the purine. Nucleosides are therefore *N*-ribosyl or *N*-deoxyribosyl derivatives of pyrimidines or purines. The numbering convention for carbon and nitrogen atoms in

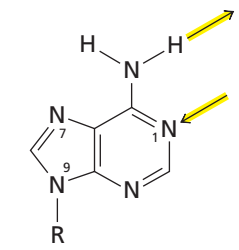


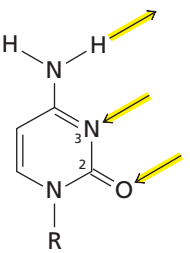



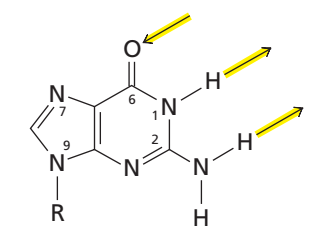

◀ **Figure 19.4**
**Chemical structures of the major pyrimidines and purines.**

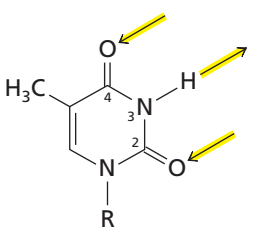
**Figure 19.5 ►**
**Tautomers of adenine, cytosine, guanine, thymine, and uracil.** At physiological pH, the equilibria of these tautomerization reactions lie far in the direction of the amino and lactam forms.





(Deoxy)Adenosine


(Deoxy)Cytidine

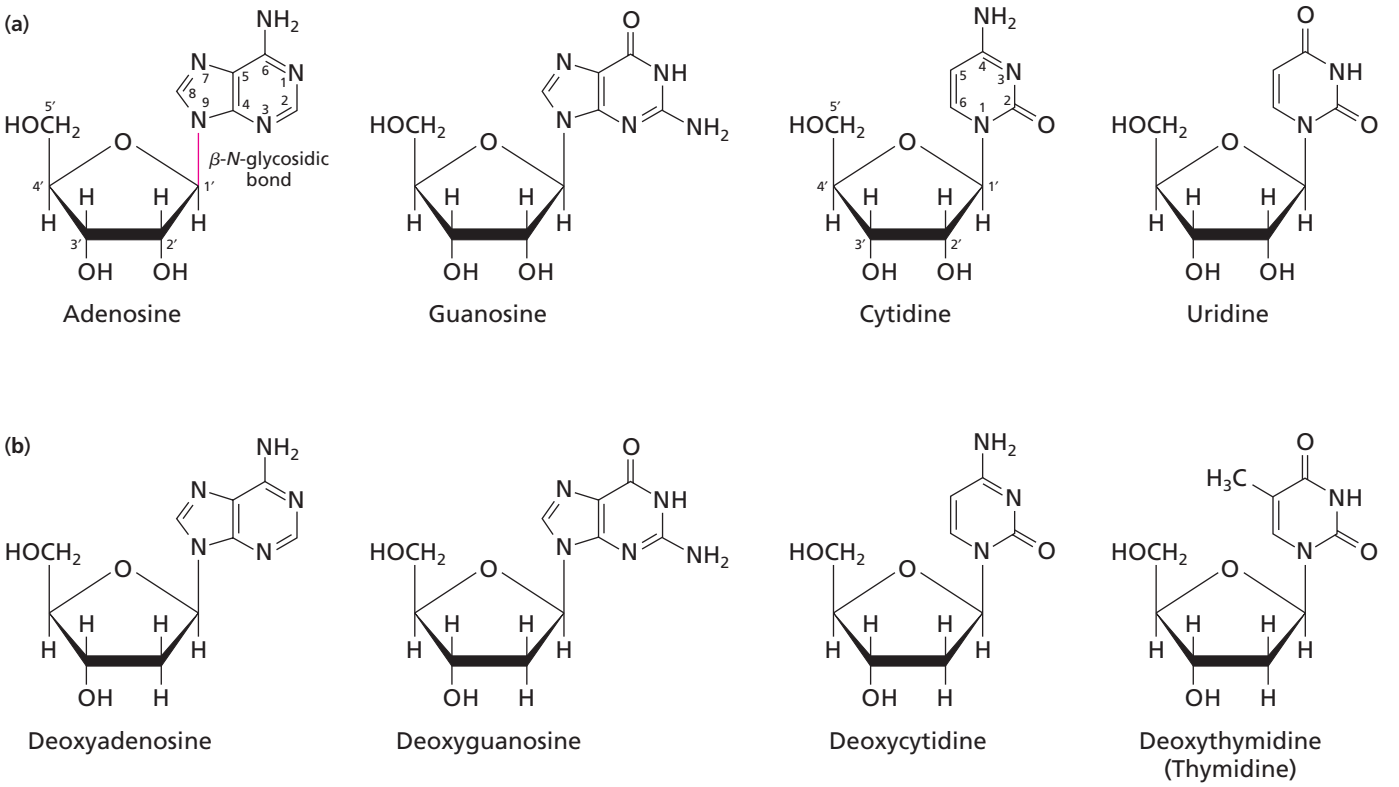

(Deoxy)Guanosine


(Deoxy)Thymidine

nucleosides reflects the fact that they are composed of a base and a five-carbon sugar, each of which has its own numbering scheme. The designation of atoms in the purine or pyrimidine moieties takes precedence. Hence the atoms in the bases are numbered 1, 2, 3, and so on, while those in the furanose ring are distinguished by adding primes (′). Thus, the $\beta$-*N*-glycosidic bond connects the C-1′, or 1′, atom of the sugar moiety to the base. Ribose and deoxyribose differ at the C-2′, or 2′, position. The chemical structures of the major ribonucleosides and deoxyribonucleosides are shown in Figure 19.7.

The names of nucleosides are derived from the names of their bases. The ribonucleoside containing adenine is called adenosine (the systematic name, 9-$\beta$-D-ribofuranosyladenine, is seldom used)

**◄ Figure 19.6**
**Hydrogen bond sites of bases in nucleic acids.** Each base contains atoms and functional groups that can serve as hydrogen donors or acceptors. The common tautomeric forms of the bases are shown. Hydrogen donor and acceptor groups differ in the other tautomers. R represents the sugar moiety.




▲ **Figure 19.7**
**Chemical structures of nucleosides.** Note that the carbon atoms of the sugars are numbered with primes to distinguish them from the atoms of the bases. **(a)** Ribonucleosides. The sugar in ribonucleosides is ribose, which contains a hydroxyl group at C-2′, as shown here. The *β*-*N*-glycosidic bond of adenosine is shown in red. **(b)** Deoxyribonucleosides. In deoxyribonucleosides, there is an additional hydrogen atom at C-2′ instead of a hydroxyl group.

and its deoxy counterpart is called deoxyadenosine. Similarly, the ribonucleosides of guanine, cytosine, and uracil are guanosine, cytidine, and uridine, respectively. The deoxyribonucleosides of guanine, cytosine, and thymine are deoxyguanosine, deoxycytidine, and deoxythymidine, respectively. Deoxythymidine is often simply called thymidine because thymine rarely occurs in ribonucleosides. The single-letter abbreviations for pyrimidine and purine bases are also commonly used to designate ribonucleosides: A, G, C, and U (for adenosine, guanosine, cytidine, and uridine, respectively). The deoxyribonucleosides are abbreviated dA, dG, dC, and dT when it is necessary to distinguish them from ribonucleosides.

Rotation around the glycosidic bonds of nucleosides and nucleotides is sometimes hindered. There are two relatively stable conformations, *syn* and *anti*, that are in rapid equilibrium (Figure 19.8). In the common pyrimidine nucleosides, the *anti* conformation predominates. The *anti* conformations of all nucleotides predominate in nucleic acids, the polymers of nucleotides.

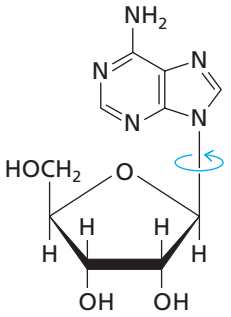
## D. Nucleotides

Nucleotides are phosphorylated derivatives of nucleosides. Ribonucleosides contain three hydroxyl groups that can be phosphorylated (2′, 3′, and 5′), and deoxyribonucleosides contain two such hydroxyl groups (3′ and 5′). The phosphoryl groups in naturally occurring nucleotides are usually attached to the oxygen atom of the 5′-hydroxyl group. By convention, a nucleotide is always assumed to be a 5′-phosphate ester unless otherwise designated.

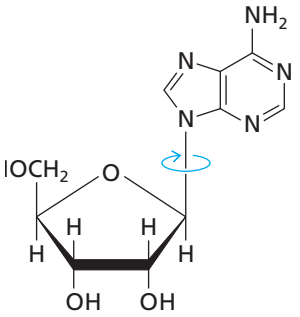
The systematic names for nucleotides indicate the number of phosphate groups present. For example, the 5′-monophosphate ester of adenosine is known as adenosine monophosphate (AMP). It is also simply called adenylate. Similarly, the 5′-monophosphate ester of deoxycytidine can be referred to as deoxycytidine monophosphate (dCMP) or deoxycytidylate. The 5′-monophosphate ester of the deoxyribonucleoside of thymine is usually known as thymidylate but is sometimes called deoxythymidylate to avoid ambiguity. Table 19.1 presents an overview of the nomenclature of bases, nucleosides, and 5′-nucleotides. Nucleotides with the phosphate esterified at the 5′ position are abbreviated AMP, dCMP, and so on. Nucleotides with the phosphate esterified at a position other than 5′ are given similar abbreviations but with position numbers designated (e.g., 3′-AMP).

### KEY CONCEPT

**By convention the numbering of the atoms in the base takes precedence so the carbon atoms in the sugar are numbered 1′ (“one prime”), 2′ (“two prime”), etc.**




*syn* Adenosine




*anti* Adenosine

**▲ Figure 19.8**
***Syn* and *anti* conformations of adenosine.** Some nucleosides assume either the *syn* or *anti* conformation. The *anti* form is usually more stable in pyrimidine nucleosides.

**Table 19.1 Nomenclature of bases, nucleosides, and nucleotides**

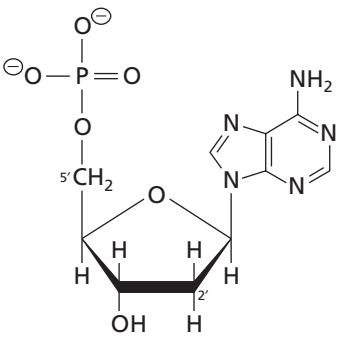
| Base | Ribonucleoside | Ribonucleotide (5′-monophosphate) |
|---|---|---|
| Adenine (A) | Adenosine | Adenosine 5′-monophoshate (AMP); adenylate[a] |
| Guanine (G) | Guanosine | Guanosine 5′-monophosphate (GMP); guanylate[a] |
| Cytosine (C) | Cytidine | Cytidine 5′-monophosphate (CMP); cytidylate[a] |
| Uracil (U) | Uridine | Uridine 5′-monophosphate (UMP); uridylate[a] |
| **Base** | **Deoxyribonucleoside** | **Deoxyribonucleotide (5′-monophosphate)** |
| Adenine (A) | Deoxyadenosine | Deoxyadenosine 5′-monophosphate (dAMP); deoxyadenylate[a] |
| Guanine (G) | Deoxyguanosine | Deoxyguanosine 5′-monophosphate (dGMP); deoxyguanylate[a] |
| Cytosine (C) | Deoxycytidine | Deoxycytidine 5′-monophosphate (dCMP); deoxycytidylate[a] |
| Thymine (T) | Deoxythymidine or thymidine | Deoxythymidine 5′-monophosphate (dTMP); deoxythymidylate[a] or thymidylate[a] |

[a]Anionic forms of phosphate esters predominant at pH 7.4.

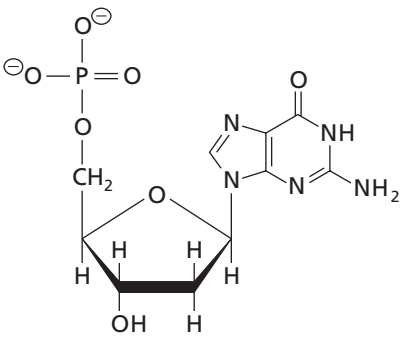
Nucleoside monophosphates, which are derivatives of phosphoric acid, are anionic at physiological pH. They are dibasic acids under physiological conditions since the p$K_a$ values are approximately 1 and 6. The nitrogen atoms of the heterocyclic rings can also ionize.

Nucleoside monophosphates can be further phosphorylated to form nucleoside diphosphates and nucleoside triphosphates. These additional phosphoryl groups are present as phosphoanhydrides. The chemical structures of the deoxyribonucleoside-5′-monophosphates are shown in Figure 19.9. A three-dimensional view of the structure

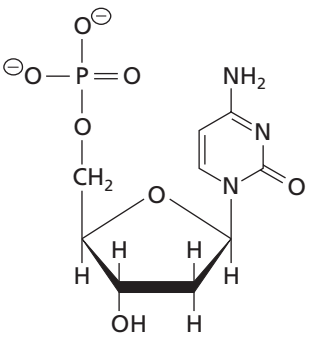
**Figure 19.9 ►**
**Chemical structures of the deoxyribonucleoside-5′-monophosphates.**



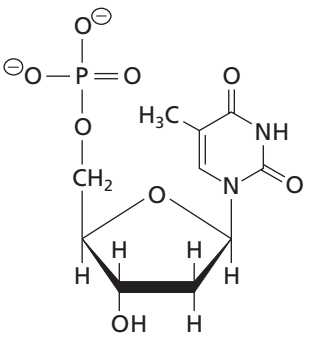

2′-Deoxyadenosine 5′-monophosphate (Deoxyadenylate, dAMP)




2′-Deoxyguanosine 5′-monophosphate (Deoxyguanylate, dGMP)



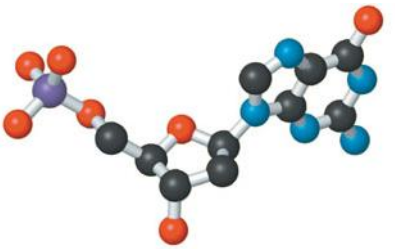

2′-Deoxycytidine 5′-monophosphate (Deoxycytidylate, dCMP)




2′-Deoxythymidine 5′-monophosphate (Thymidylate, dTMP)

of dGMP is shown in Figure 19.10. The base in dGMP is in the *anti* conformation and the sugar ring is puckered. The plane of the purine ring is almost perpendicular to that of the furanose ring. The phosphoryl group attached to the 5′-carbon atom is positioned well above the sugar and far away from the base.

Nucleoside polyphosphates and polymers of nucleotides can also be abbreviated using a scheme in which the phosphate groups are represented by "p" and the nucleosides are represented by their one-letter abbreviations. The position of the "p" relative to the nucleoside abbreviation indicates the position of the phosphate—for a 5′ phosphate, the p precedes the nucleoside abbreviation and for a 3′ phosphate, the "p" follows the nucleoside abbreviation. Thus, 5′-adenylate (AMP) can be abbreviated as pA, 3′-deoxyadenylate as dAp, and ATP as pppA.



▲ **Figure 19.10**
**Deoxyguanosine-5′-monophosphate (dGMP).** Hydrogen atoms have been omitted for clarity. Color key: carbon, black; nitrogen, blue; oxygen, red; phosphorus, purple.

## 19.2 DNA Is Double-Stranded

By 1950 it was clear that DNA is a linear polymer of 2′-deoxyribonucleotide residues linked by 3′–5′ phosphodiesters. Moreover, Erwin Chargaff had deduced certain regularities in the nucleotide compositions of DNA samples obtained from a wide variety of prokaryotes and eukaryotes. Among other things, Chargaff observed that in the DNA of a given cell, A and T are present in equimolar amounts, as are G and C. An example of modern DNA composition data showing these ratios is presented in Table 19.2. Although A = T and G = C for each species, the total mole percent of (G + C) may differ considerably from that of (A + T). The DNA of some organisms, such as the yeast *Saccharomyces cerevisiae*, is relatively deficient in (G + C) whereas the DNA of other organisms, such as the bacterium *Mycobacterium tuberculosis*, is rich in (G + C). In general, the DNAs of closely related species, such as cows, pigs, and humans, have similar base compositions. The data also shows that the ratio of purines to pyrimidines is always 1:1 in the DNA of all species.

The model of DNA proposed by Watson and Crick in 1953 was based on the known structures of the nucleosides and on X-ray diffraction patterns that Rosalind Franklin and Maurice Wilkins obtained from DNA fibers. The Watson–Crick model accounted for the equal amounts of purines and pyrimidines by suggesting that DNA was double-stranded and that bases on one strand paired specifically with bases on the other strand: A with T and G with C. Watson and Crick's proposed structure is now referred to as the B conformation of DNA, or simply B-DNA.

An appreciation of DNA structure is important for understanding the processes of DNA replication (Chapter 20) and transcription (Chapter 21). DNA is the storehouse of biological information. Every cell contains dozens of enzymes and proteins that bind to DNA recognizing certain structural features, such as the sequence of nucleotides. In the following sections we will see how the structure of DNA allows these proteins to gain access to the stored information.

**Table 19.2 Base composition of DNA (mole %) and ratios of bases**

| Source | A | G | C | T | A/T[a] | G/C[a] | (G + C) | Purine/pyrimidine[a] |
|---|---|---|---|---|---|---|---|---|
| *Escherichia coli* | 26.0 | 24.9 | 25.2 | 23.9 | 1.09 | 0.99 | 50.1 | 1.04 |
| *Mycobacterium tuberculosis* | 15.1 | 34.9 | 35.4 | 14.6 | 1.03 | 0.99 | 70.3 | 1.00 |
| *Yeast* | 31.7 | 18.3 | 17.4 | 32.6 | 0.97 | 1.05 | 35.7 | 1.00 |
| *Cow* | 29.0 | 21.2 | 21.2 | 28.7 | 1.01 | 1.00 | 42.4 | 1.01 |
| *Pig* | 29.8 | 20.7 | 20.7 | 29.1 | 1.02 | 1.00 | 41.4 | 1.01 |
| *Human* | 30.4 | 19.9 | 19.9 | 30.1 | 1.01 | 1.00 | 39.8 | 1.01 |

[a]Deviations from a 1:1 ratio are due to experimental variations.

## A. Nucleotides Are Joined by 3′–5′ Phosphodiester Linkages

**A linkage group consists of several different covalent bonds.**

We have seen that the primary structure of a protein refers to the sequence of its amino acid residues linked by peptide bonds. Similarly, the primary structure of a nucleic acid is the sequence of its nucleotide residues connected by 3′–5′ phosphodiester linkages. A tetranucleotide representing a segment of a DNA chain illustrates such linkages (Figure 19.11). The backbone of the polynucleotide chain consists of the phosphoryl groups, the 5′, 4′, and 3′ carbon atoms, and the 3′ oxygen atom of each deoxyribose. These backbone atoms are arranged in an extended conformation. This makes double-stranded DNA a long, thin molecule, unlike polypeptide chains that can easily fold back on themselves.

All the nucleotide residues within a polynucleotide chain have the same orientation. Thus, polynucleotide chains have directionality, like polypeptide chains. One end of a linear polynucleotide chain is said to be 5′ (because no residue is attached to its 5′-carbon) and the other is said to be 3′ (because no residue is attached to its 3′-carbon). By convention, the direction of a DNA strand is defined by reading across the atoms that make up the sugar residue. Thus, going from the top to the bottom of the strand in

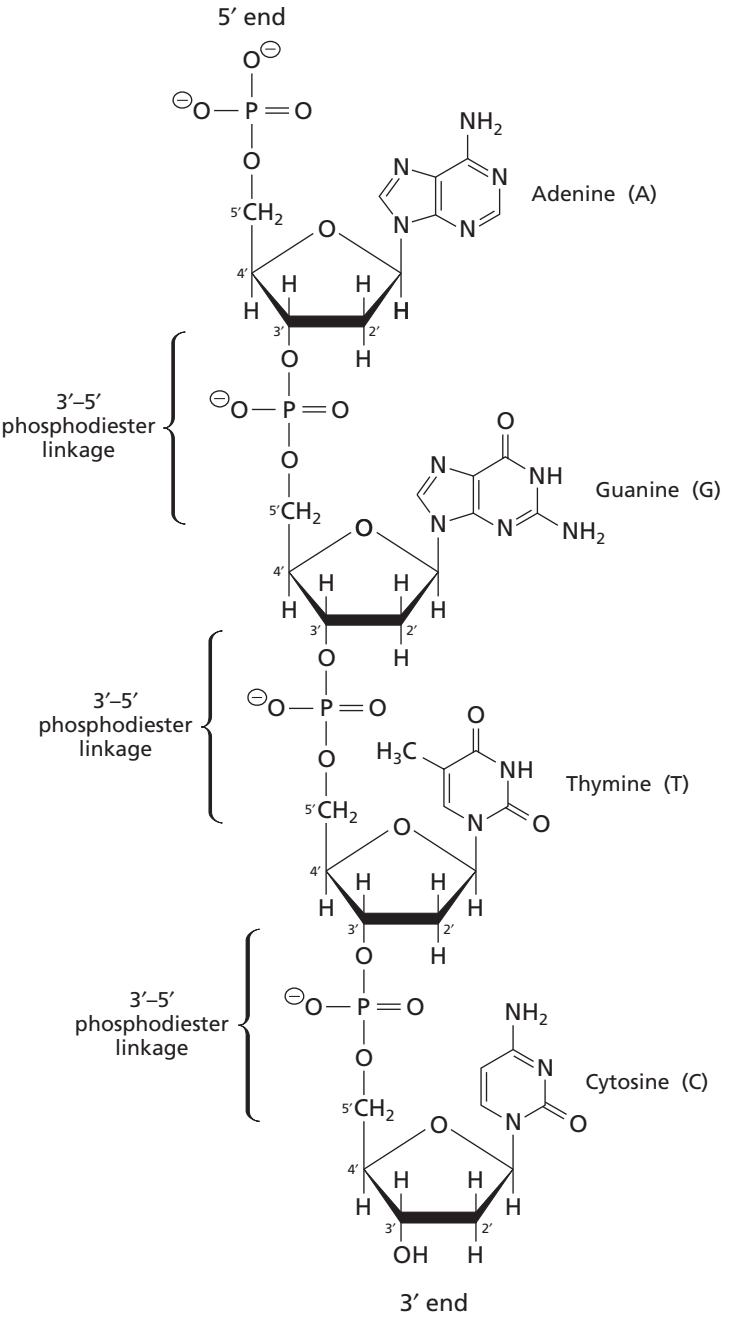



**Figure 19.11 ►**
**Chemical structure of the tetranucleotide pdApdGpdTpdC.** The nucleotide residues are joined by 3′–5′ phosphodiester linkages. The nucleotide with a free 5′-phosphoryl group is called the 5′ end, and the nucleotide with a free 3′-hydroxyl group is called the 3′ end.

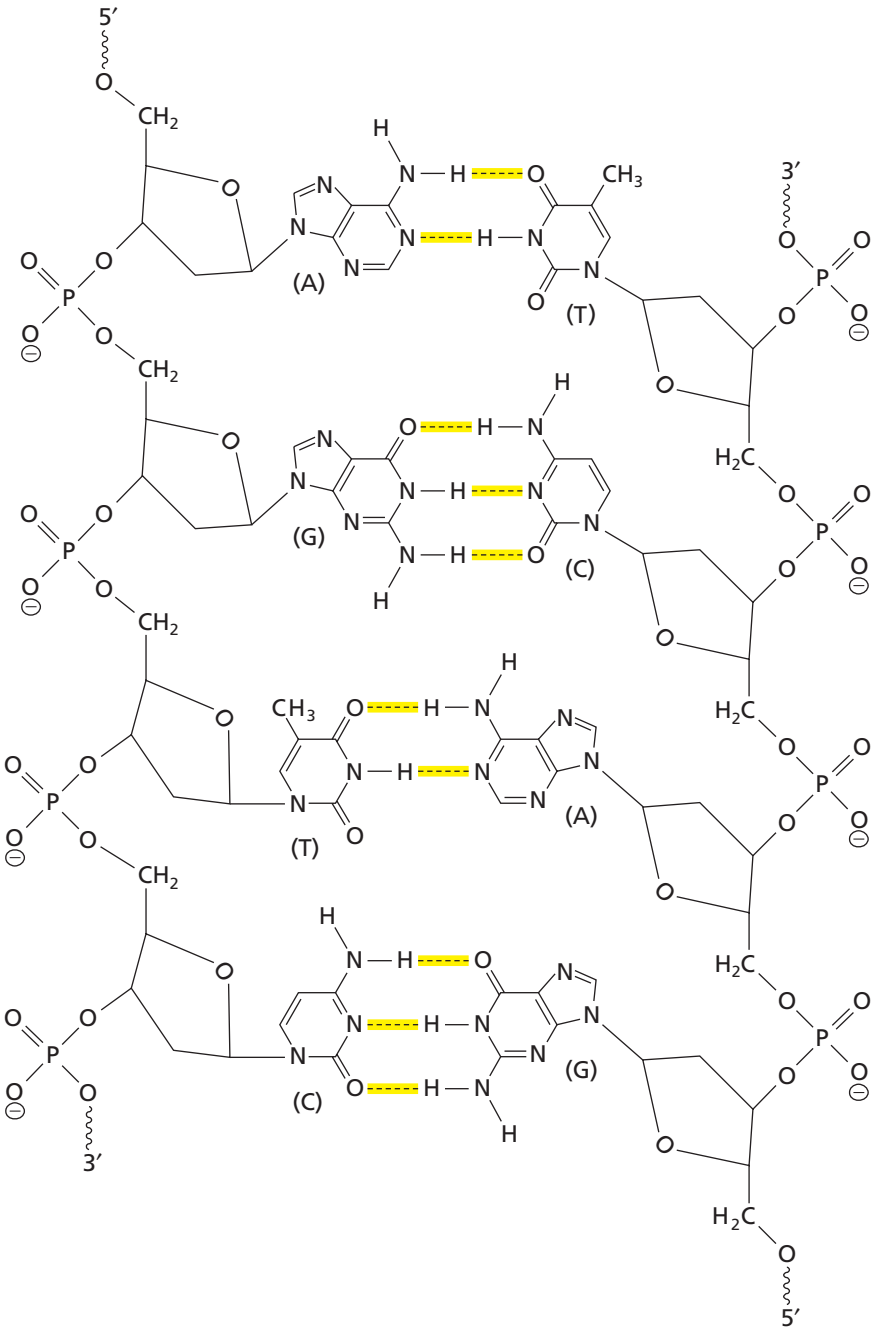
Figure 19.11 is defined as 5′ → 3′ ("five prime to three prime") because one crosses the sugar residue encountering the 5′, 4′, and 3′ carbon atoms, in that order. Similarly, going from the bottom to the top of the strand is moving in the 3′ → 5′ direction.

Structural abbreviations are assumed to read in the 5′ → 3′ direction unless otherwise specified. Because phosphates can be abbreviated as "p," the tetranucleotide in Figure 19.11 can be referred to as pdApdGpdTpdC, or shortened to AGTC when it is clear that the reference is to DNA.

Each phosphate group that participates in a phosphodiester linkage has a p$K_a$ of about 2 and bears a negative charge at neutral pH. Consequently, nucleic acids are polyanions under physiological conditions. Negatively charged phosphate groups are neutralized by small cations and positively charged proteins.

## B. Two Antiparallel Strands Form a Double Helix

Most DNA molecules consist of two strands of polynucleotide. Each of the bases on one strand forms hydrogen bonds with a base of the opposite strand (Figure 19.12). The




**KEY CONCEPT**

**The direction of moving along a DNA or RNA strand can be either 5′ → 3′ or 3′ → 5′. It is defined by the direction of reading across the atoms that make up the sugar residue.**

▲ **Watson and Crick's original DNA model.**

◀ **Figure 19.12**
**Chemical structure of double-stranded DNA.** The two strands run in opposite directions. Adenine in one strand pairs with thymine in the opposite strand, and guanine pairs with cytosine.

most common base pairs occur between the lactam and amino tautomers of the bases. Guanine pairs with cytosine and adenine with thymine in a manner that maximizes hydrogen bonding between potential sites. G/C base pairs have three hydrogen bonds and A/T base pairs have two. This feature of double-stranded DNA accounts for Chargaff's discovery that the ratio of A to T and of G to C is 1:1 for a wide variety of DNA molecules. Because A in one strand pairs with T in the other strand and G pairs with C, the strands are complementary and each one can serve as a template for the other.

The sugar–phosphate backbones of the complementary strands of double-stranded DNA have opposite orientations. In other words, they are antiparallel. This was one of the important new insights contributed by Watson and Crick when they built their model of DNA in 1953.

Each end of double-stranded DNA is made up of the 5′ end of one strand and the 3′ end of another. The distance between the two sugar–phosphate backbones is the same for each base pair. Consequently, all DNA molecules have the same regular structure in spite of the fact that their nucleotide sequences may be quite different.

The actual structure of DNA differs in two important aspects from that shown in Figure 19.12. In a true three-dimensional representation, the two strands wrap around each other to form a two-stranded helical structure, or double helix. Also, the bases are rotated so that the plane of the base pairs is nearly perpendicular to the page. (Recall that the plane of the base in dGMP is nearly perpendicular to that of the sugar, as shown in Figure 19.10.)

## KEY CONCEPT

**The two strands of DNA are anti-parallel.**

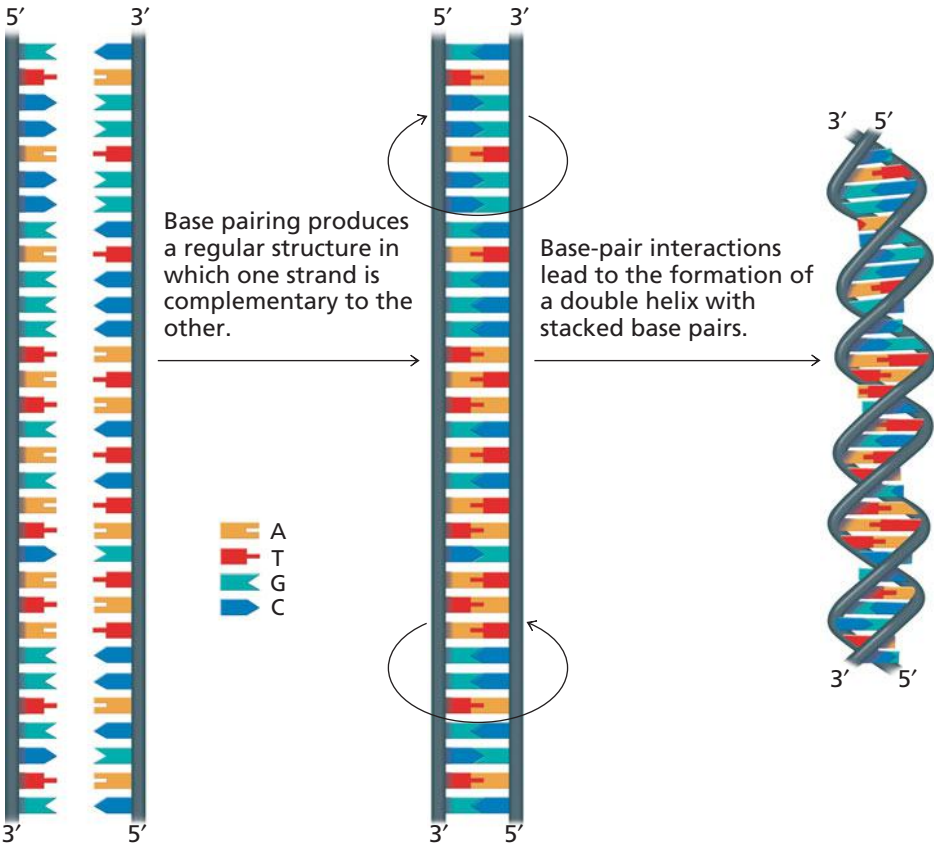
The DNA molecule can be visualized as a "ladder" that has been twisted into a helix. The paired bases represent the rungs of the ladder and the sugar–phosphate backbones represent the supports. Each complementary strand serves as a perfect template for the other. This complementarity is responsible for the overall regularity of the structure of double-stranded DNA. However, complementary base pairing alone does not produce a helix. In B-DNA, the base pairs are stacked one above the other and are nearly perpendicular to the long axis of the molecule. The cooperative, noncovalent interactions between the upper and lower surfaces of each base pair bring the pairs closer together and create a hydrophobic interior that causes the sugar–phosphate backbone to twist. It is these stacking interactions that create the familiar helix (Figure 19.13). Much of the stability of double-stranded DNA is due to the stacking interactions between base pairs.

The two hydrophilic sugar–phosphate backbones wind around the outside of the helix where they are exposed to the aqueous environment. In contrast, the stacked, relatively hydrophobic bases are located in the interior of the helix where they are largely inaccessible to water. This hydrophobic environment makes the hydrogen bonds between bases more stable since they are shielded from competition with water molecules.

The double helix has two grooves of unequal width because of the way the base pairs stack and the sugar–phosphate backbones twist. These grooves are called the **major groove** and the **minor groove** (Figure 19.14). Within each groove, functional groups on the edges of the base pairs are exposed to water. Each base pair has a distinctive pattern of chemical groups projecting into the grooves. Molecules that interact with particular base pairs can identify them by binding in the grooves without disrupting the helix. This is particularly important for proteins that must bind to double-stranded DNA and "read" a specific sequence.

B-DNA is a right-handed helix with a diameter of 2.37 nm. The **rise** of the helix (the distance between one base pair and the next along the helical axis) averages 0.33 nm, and the **pitch** of the helix (the distance to complete one turn) is about 3.40 nm. These values vary to some extent depending on the base composition. Because there are about 10.4 base pairs per turn of the helix, the angle of rotation between adjacent nucleotides within each strand is about 34.6° (360/10.4).

Two views of B-DNA are shown in Figure 19.15. The ball-and-stick model (Figure 19.15a) shows that the hydrogen bonds between base pairs are buried in the interior of the molecule where they are protected from competing interactions with water. The charged phosphate groups (purple and red atoms) are located on the outside surface. This arrangement is more evident in the space-filling model (Figure 19.15b). The space-filling model also clearly shows that functional groups of the base pairs are exposed in




◄ **Figure 19.13**
**Complementary base pairing and stacking in double-stranded DNA.**

the grooves. These groups can be identified by the presence of blue nitrogen atoms and red oxygen atoms.

The length of double-stranded DNA molecules is often expressed in terms of base pairs (bp). For convenience, longer structures are measured in thousands of base pairs, or **kilobase pairs**, commonly abbreviated kb. Most bacterial genomes consist of a single DNA molecule of several thousand kb; for example, the *Escherichia coli* chromosome is 4600 kb in length. The largest DNA molecules in the chromosomes of mammals and flowering plants may be several hundred thousand kb long. The human genome contains 3,200,000 kb ($3.2 \times 10^9$ base pairs) of DNA.

Most bacteria have a single chromosome whose ends are joined to create a circular molecule. DNA in the mitochondria and chloroplasts of eukaryotic cells is also circular. In contrast, the chromosomes in the nucleus of a eukaryotic cell are linear. (Some bacteria also have multiple chromosomes and some have linear chromosomes.)

## C. Weak Forces Stabilize the Double Helix

The forces that maintain the native conformations of complex cellular structures are strong enough to maintain the structures but weak enough to allow conformational flexibility. Covalent bonds between adjacent residues define the primary structures of proteins and nucleic acids but weak forces determine the three-dimensional shapes

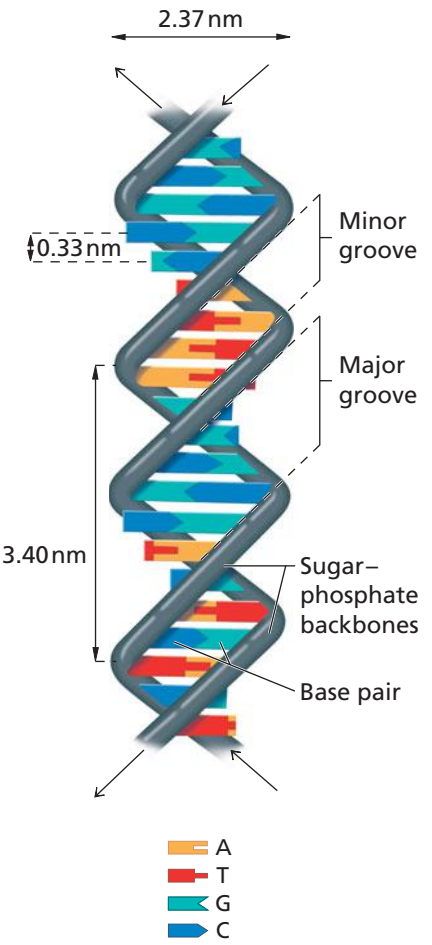
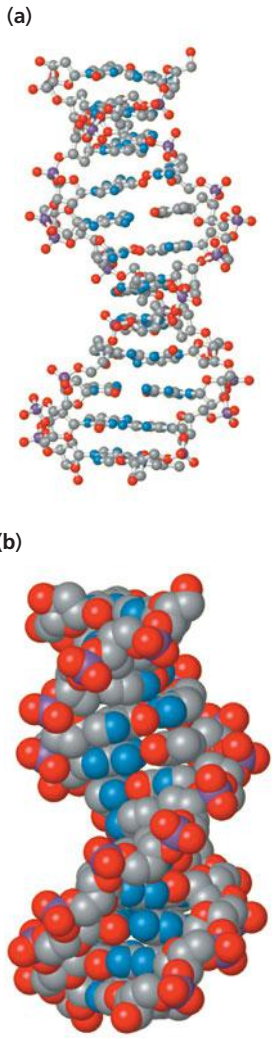



**Figure 19.14** ►
**Three-dimensional structure of B-DNA.** This model shows the orientation of the base pairs and the sugar–phosphate backbone and the relative sizes of the pyrimidine and purine bases. The sugar–phosphate backbone winds around the outside of the helix and the bases occupy the interior. Stacking of the base pairs creates two grooves of unequal width—the major and the minor grooves. The diameter of the helix is 2.37 nm, and the distance between base pairs is 0.33 nm. The distance to complete one turn is 3.40 nm. (For clarity, a slight space has been left between the stacked base pairs and the interactions between complementary bases are shown schematically.)




▲ **Figure 19.15**
**B-DNA. (a)** Ball-and-stick model. The base pairs are nearly perpendicular to the sugar–phosphate backbones. **(b)** Space-filling model. Color key: carbon, gray; nitrogen, blue; oxygen, red; phosphorus, purple. [Nucleic Acids Database BD0001].

of these macromolecules. Four types of interactions affect the conformation of double-stranded DNA.

1. *Stacking interactions.* The stacked base pairs form van der Waals contacts. Although the forces between individual stacked base pairs are weak, they are additive so in large DNA molecules the van der Waals contacts are an important source of stability.
2. *Hydrogen bonds.* Hydrogen bonding between base pairs is a significant stabilizing force.
3. *Hydrophobic effects.* Burying hydrophobic purine and pyrimidine rings in the interior of the double helix increases the stability of the helix.
4. *Charge–charge interactions.* Electrostatic repulsion of the negatively charged phosphate groups of the backbone is a potential source of instability in the DNA helix. However, repulsion is minimized by the presence of cations such as $Mg^{2+}$ and cationic proteins (proteins that contain an abundance of the basic residues arginine and lysine).

The importance of stacking interactions can be illustrated by examining the various stacking energies of the base pairs (Table 19.3). The stacking energy of two base pairs depends on the nature of the base pair (G/C or A/T) and the orientation of each base pair. Typical stacking energies are about 35 kJ $mol^{-1}$. Within the hydrophobic core of stacked double-stranded DNA the hydrogen bonds between base pairs have a strength of about 27 kJ $mol^{-1}$ each (Section 2.5B). However, if the stacking interactions are weakened, the hydrogen bonds in the base pairs are exposed to competition from water molecules and the overall contribution to keeping the strands together diminishes greatly.

Under physiological conditions, double-stranded DNA is thermodynamically much more stable than the separated strands and that explains why the double-stranded form predominates *in vivo*. However, the structure of localized regions of the double helix can sometimes be disrupted by unwinding. Such disruption occurs during DNA replication, repair, recombination, and transcription. Complete unwinding and separation of the complementary single strands is called **denaturation**. Denaturation occurs only *in vitro*.

Double-stranded DNA can be denatured by heat or by a chaotropic agent such as urea or guanidinium chloride. (Recall from Section 4.10 that proteins can also be denatured.) In studies of thermal denaturation, the temperature of a solution of DNA is slowly increased. As the temperature is raised, more and more of the bases become unstacked and hydrogen bonds between base pairs are broken. Eventually, the two strands separate completely. The temperature at which half the DNA has become single-stranded is known as the **melting point** ($T_m$).

Absorption of ultraviolet light can be used to measure the extent of denaturation. Measurements are made at a wavelength of 260 nm—close to the absorbance maximum for nucleic acids. Single-stranded DNA absorbs 12% to 40% more light than double-stranded DNA at 260 nm (Figure 19.16). A plot of the change in absorbance of a DNA solution versus temperature is called a **melting curve** (Figure 19.17). The absorbance increases sharply at the melting point and the transition from double-stranded to single-stranded DNA takes place over a narrow range of temperature.

The sigmoid shape of the melting curve indicates that denaturation is a cooperative process as we saw in the case of protein denaturation (Section 4.10). In this case, cooperativity results from rapid unzippering of the double-stranded molecule as the many stacking interactions and hydrogen bonds are disrupted. The unzippering begins with the unwinding of a short internal stretch of DNA, forming a single-stranded "bubble." This single-stranded bubble rapidly destabilizes the adjacent stacked base pairs and this destabilization is propagated in both directions as the bubble expands.

As shown in Figure 19.17, poly (GC) denatures at a much higher temperature than poly (AT). It is easier to melt A/T-rich DNA than G/C-rich DNA because A/T base pairs have weaker stacking interactions as shown in Table 19.3. It's important to note that the stacking interactions are the first interactions to be disrupted by higher temperature. Once this process begins the hydrogen bonds—although collectively stronger in stacked

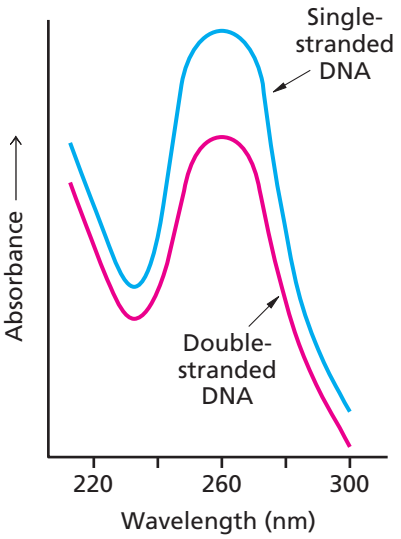



**Figure 19.16 ►**
**Absorption spectra of double-stranded and single-stranded DNA.** At pH 7.0, double-stranded DNA has an absorbance maximum near 260 nm. Denatured DNA absorbs 12% to 40% more ultraviolet light than double-stranded DNA.

DNA—become much weaker because they are exposed to water and the DNA is rapidly destabilized. Naturally occurring DNA is a mixture of regions with varying base compositions but A/T-rich regions are more easily unwound than G/C-rich regions.

At temperatures just below the melting point, a typical DNA molecule contains double-stranded regions that are G/C-rich and local single-stranded regions ("bubbles") that are A/T-rich. These *in vitro* experiments demonstrate an important point—that it is easier to unwind localized regions whose sequence consists largely of A/T base pairs rather than G/C base pairs. We will see in Chapter 21 that the initiation sites for transcription are often A/T-rich.

## D. Conformations of Double-Stranded DNA

Double-stranded DNA can assume different conformations under different conditions. X-ray crystallographic studies of various synthetic oligodeoxyribonucleotides of known sequence indicate that DNA molecules inside the cell do not exist in a "pure" B conformation. Instead, DNA is a dynamic molecule whose exact conformation depends to some extent on the nucleotide sequence. The local conformation is also affected by bends in the DNA molecule and whether it is bound to protein. As a result, the number of base pairs per turn in B-DNA can fluctuate in the range of 10.2–10.6.

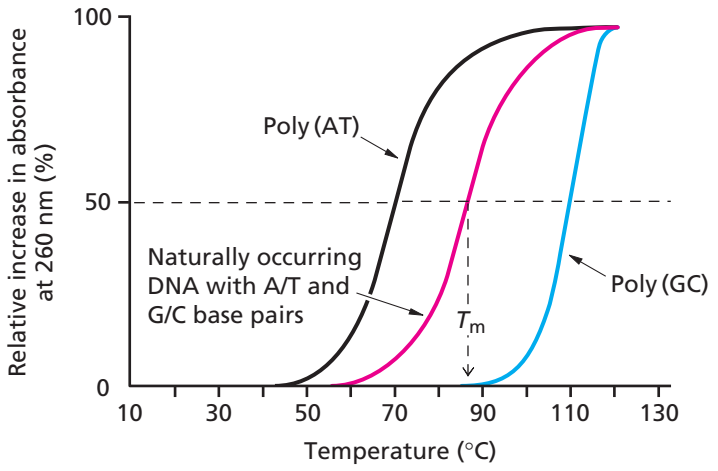
There are two other distinctly different DNA conformations in addition to the various forms of B-DNA. A-DNA forms when DNA is dehydrated and Z-DNA can form when certain sequences are present (Figure 19.18). (The A- and B-DNA forms were discovered by Rosalind Franklin in 1952.) A-DNA is more tightly wound than B-DNA and the

**Table 19.3 Stacking interactions for the ten possible combinations in double-stranded DNA**

| Stacked dimers | | Stacking energies (kJ mol$^{-1}$) |
|---|---|---|
| ↑C-G↓<br>G-C | | –61.0 |
| ↑C-G↓<br>A-T | ↑T-A↓<br>G-C | –44.0 |
| ↑C-G↓<br>T-A | ↑A-T↓<br>G-C | –41.0 |
| ↑G-C↓<br>C-G | | –40.5 |
| ↑G-C↓<br>G-C | ↑C-G↓<br>C-G | –34.6 |
| ↑T-A↓<br>A-T | | –27.5 |
| ↑G-C↓<br>T-A | ↑A-T↓<br>C-G | –27.5 |
| ↑G-C↓<br>T-A | ↑T-A↓<br>C-G | –28.4 |
| ↑A-T↓<br>A-T | ↑T-A↓<br>T-A | –22.5 |
| ↑A-T↓<br>T-A | | –16.0 |

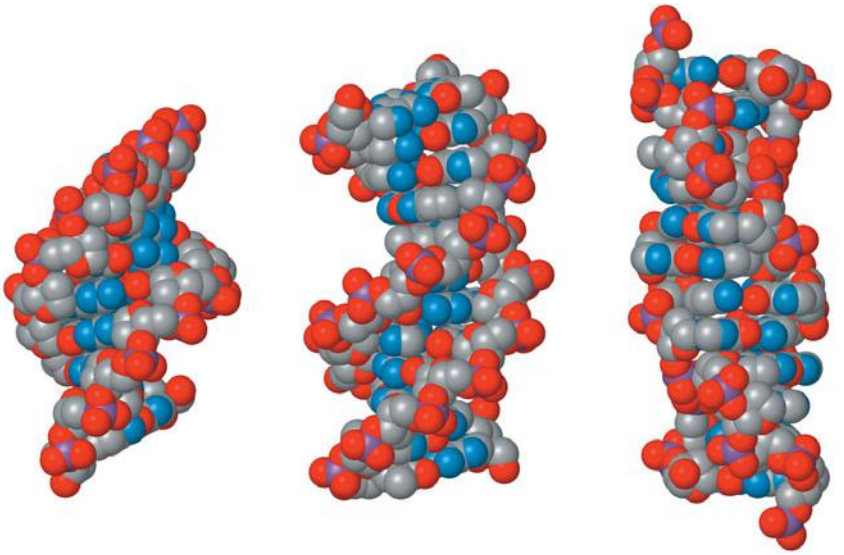
Arrows designate the direction of the sugar-phosphate backbone and point from C-3′ of one sugar unit to C-5′ of the next.

[Adapted from Omstein, R. L., Rein, R., Breen, D. L., and MacElroy, R. D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies: I. Base stacking. *Biopolymers* 17: 2341–2360.]




**◄ Figure 19.17 Melting curve for DNA.**
In this experiment, the temperature of a DNA solution is increased while the absorbance at 260 nm is monitored. The melting point ($T_m$) corresponds to the inflection point of the sigmoidal curve where the increase in absorbance of the sample is one-half the increase in absorbance of completely denatured DNA. Poly (AT) melts at a lower temperature than either naturally occurring DNA or poly (GC) since more energy is required to disrupt stacked G/C base pairs.

**Figure 19.18 ►**
**A-DNA, B-DNA, and Z-DNA.** The A-DNA conformation (left) is favored when DNA is dehydrated [NDB AD0001]. B-DNA (center) is the conformation normally found inside cells [NDB BD0001]. The Z-DNA conformation (right) is favored in certain G/C-rich sequences [NDB ZDJ050].

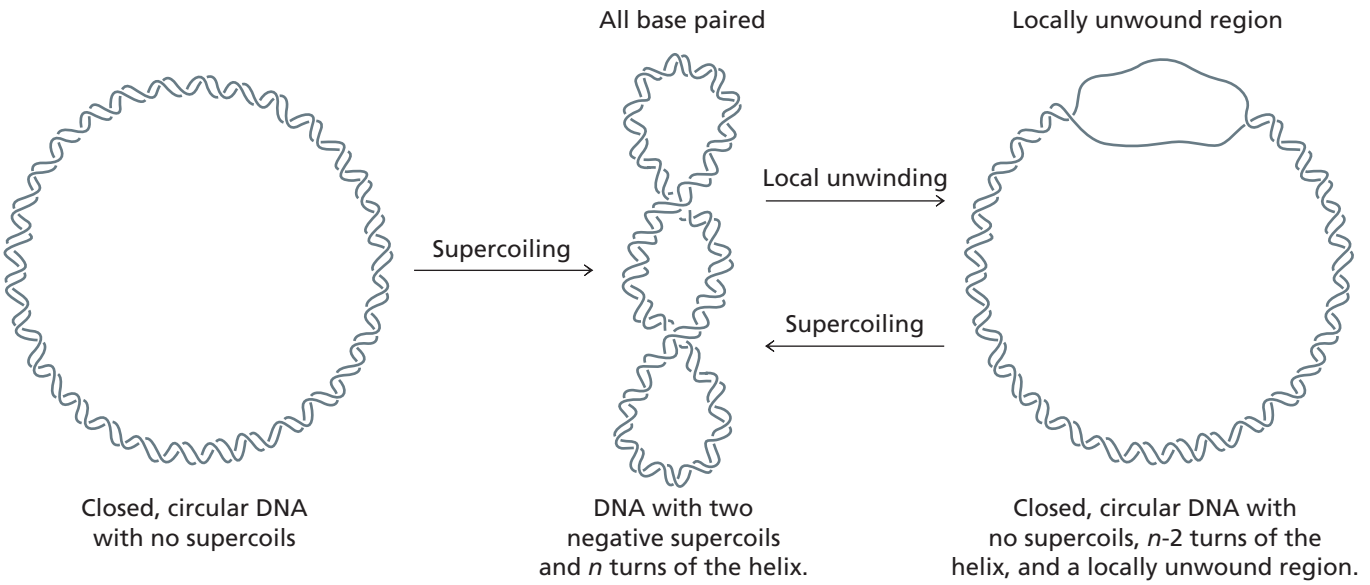


major and minor grooves of A-DNA are similar in width. There are about 11 bp per turn in A-DNA and the base pairs are tilted about 20° relative to the long axis of the helix. Z-DNA differs even more from B-DNA. There are no grooves in Z-DNA and the helix is left-handed, not right-handed. The Z-DNA conformation occurs in G/C-rich regions. Deoxyguanylate residues in Z-DNA have a different sugar conformation (3′-endo) and the base is in the *syn* conformation. A-DNA and Z-DNA conformations exist *in vivo* but they are confined to short regions of DNA.

## 19.3 DNA Can Be Supercoiled

A circular DNA molecule with the B conformation has an average of 10.4 base pairs per turn. It is said to be *relaxed* if such a molecule would lie flat on a surface. This relaxed double helix can be overwound or underwound if the strands of DNA are broken and the two ends of the linear molecule are twisted in opposite directions. When the strands are rejoined to create a circle, there are no longer 10.4 base pairs per turn as required to maintain the stable B conformation. The circular molecule compensates for over- or underwinding by forming supercoils that restore 10.4 base pairs per turn of the double helix (Figure 19.19). A supercoiled DNA molecule would not lie flat on a surface. Each supercoil compensates for one turn of the double helix.

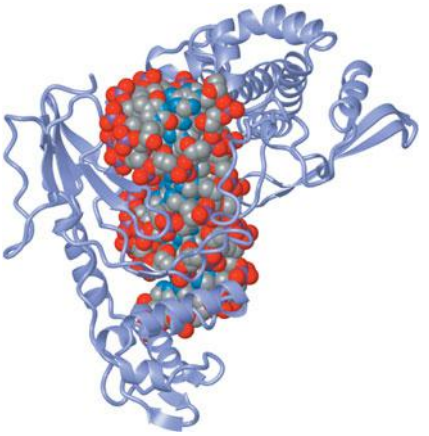
**▼ Figure 19.19**
**Supercoiled DNA.** The DNA molecule on the left is a relaxed closed circle and has the normal B conformation. Breaking the DNA helix and unwinding it by two turns before re-forming the circle produces two supercoils. The supercoils compensate for the underwinding and restore the normal B conformation. The molecule on the right has a locally unwound region of DNA. This conformation is topologically equivalent to negatively supercoiled DNA.



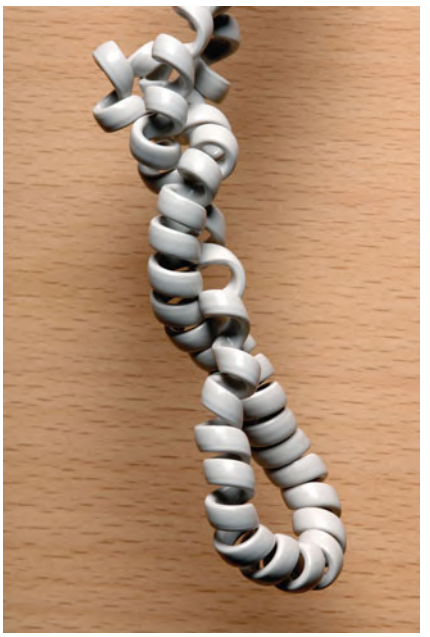
Most circular DNA molecules are supercoiled in cells but even long, linear DNA molecules contain locally supercoiled regions. Bacterial chromosomes typically have about five supercoils per 1000 base pairs of DNA. The DNA in the nuclei of eukaryotic cells is also supercoiled as we will see in Section 19.5. All organisms have enzymes that can break DNA, unwind or overwind the double helix, and rejoin the strands to alter the topology. These enzymes, called topoisomerases, are responsible for adding and removing supercoils. An example of a topoisomerase bound to DNA is shown in Figure 19.20. These remarkable enzymes cleave one or both strands of DNA, unwind or overwind DNA by rotating the cleaved ends, and then rejoin the ends to create (or remove) supercoils.

One of the important consequences of supercoiling is shown in Figure 19.19. If DNA is underwound, it compensates by forming negative supercoils in order to maintain the stable B conformation. (Overwinding produces positive supercoils.) An alternative conformation is shown on the right in Figure 19.19. In this form, most of the DNA is double-stranded but there is a locally unwound region that is due to the slight underwinding. The negatively supercoiled and locally unwound conformations are in equilibrium with the supercoiled form in excess because it is slightly more stable. The difference in free energy between the two conformations is quite small.

Most of the DNA in a cell is negatively supercoiled. This means that it is relatively easy to unwind short regions of the molecule—especially those regions that are A/T-rich. As mentioned earlier, localized unwinding is an essential step in the initiation of DNA replication, recombination, repair, and transcription. Thus, negative supercoiling plays an important biological role in these processes by storing the energy needed for local unwinding. This is why topoisomerases that catalyze supercoiling are essential enzymes in all cells.



▲ **Figure 19.20**
**Human (*Homo sapiens*) topoisomerase I bound to DNA.** [PDB 1A31]



▲ **Supercoiled telephone cords can be very annoying.**

## 19.4 Cells Contain Several Kinds of RNA

RNA molecules participate in several processes associated with gene expression. RNA molecules are found in multiple copies and in several different forms within a given cell. There are four major classes of RNA in all living cells:

1. *Ribosomal RNA* (rRNA) molecules are an integral part of ribosomes (intracellular ribonucleoproteins that are the sites of protein synthesis). Ribosomal RNA is the most abundant class of ribonucleic acid accounting for about 80% of the total cellular RNA.
2. *Transfer RNA* (tRNA) molecules carry activated amino acids to the ribosomes for incorporation into growing peptide chains during protein synthesis. tRNA molecules are only 73 to 95 nucleotide residues long. They account for about 15% of the total cellular RNA.
3. *Messenger RNA* (mRNA) molecules encode the sequences of amino acids in proteins. They are the "messengers" that carry information from DNA to the translation complex where proteins are synthesized. In general, mRNA accounts for only 3% of the total cellular RNA. These molecules are the least stable of the cellular ribonucleic acids.
4. *Small RNA* molecules are present in all cells. Some small RNA molecules have catalytic activity or contribute to catalytic activity in association with proteins. Many of these RNA molecules are associated with processing events that modify RNA after it has been synthesized. Some are required for regulating gene expression.

RNAs are single-stranded molecules, but they often have complex secondary structure. Most single-stranded polynucleotides fold back on themselves to form stable regions of base-paired, double-stranded RNA under physiological conditions. One type of secondary structure is a stem–loop which forms when short regions of complementary sequence form base pairs (Figure 19.21). The structure of the double-stranded regions of such stem–loops resembles the structure of the A form of double-stranded DNA. As we will see in Chapters 21 and 22, such structures are important in transcription and are common features in transfer RNA, ribosomal RNA, and the small RNAs.

### KEY CONCEPT

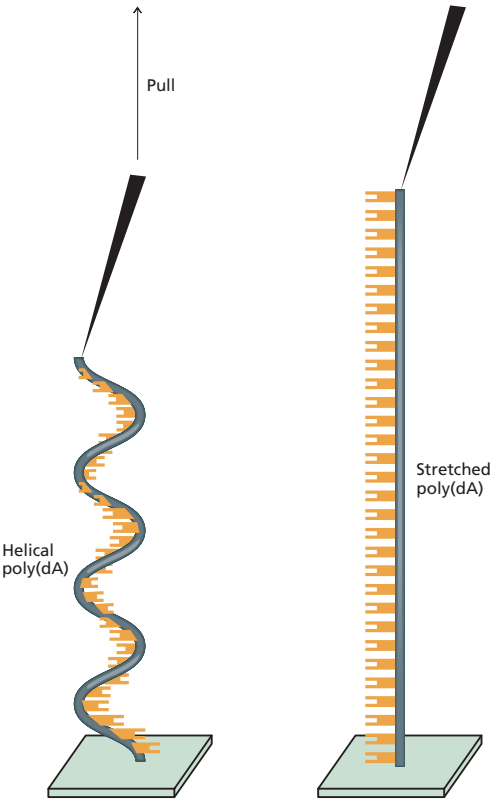
**Single-stranded RNA can fold back on itself to create stable double-stranded helical regions that resemble those in DNA.**

## BOX 19.1 PULLING DNA

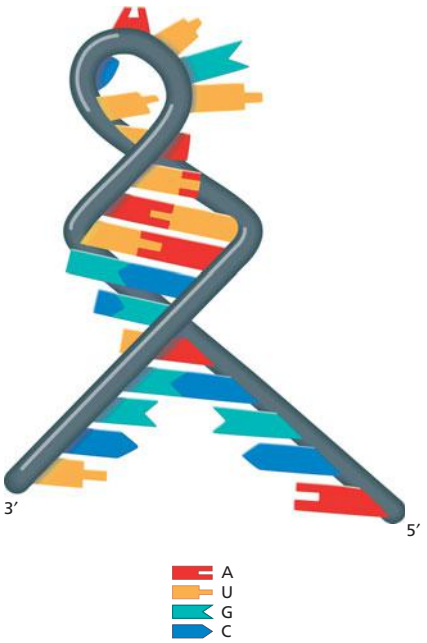
Single-molecule atomic-force spectroscopy is a powerful tool for investigating the properties of single molecules. It has been used to explore the properties of single-stranded DNA. The experiment involves fixing one end of a single-stranded DNA molecule to a solid surface and attaching the other end to a form of molecular tweezer that can be used to pull the molecule and measure its resistance.

When this experiment is done with poly(dT) there is almost no resistance until the molecule is in the fully extended form. This is because poly(dT) has no significant secondary structure. However, when poly(dA) is pulled there is initial resistance followed by a shift to the fully extended form. Poly(dA) is helical in solution because the adenylate residues stack on one another and the initial resistance is due to breaking the helix.

The resistance can be measured and the calculated energy of stacking is 15 kJ $mol^{-1}$, in agreement with other determinations of the stacking interactions of A bases on other A's. The experiment proves that stacking interactions are important in forming helical DNA structures-even with single-stranded polynucleotides.




**Pulling poly(dA).** [Adapted from Ke et al. (2007)] ▶




▲ **Figure 19.21**
**Stem–loop structures in RNA.** Single-stranded polynucleotides, such as RNA, can form stem–loops, or hairpins, when short regions of complementary sequence form base pairs. The stem of the structure consists of base-paired nucleotides, and the loop consists of noncomplementary nucleotides. Note that the strands in the stem are antiparallel.

## 19.5 Nucleosomes and Chromatin

In 1879, ten years after Miescher's discovery of nuclein, Walter Flemming observed banded objects in the nuclei of stained eukaryotic cells. He called the material **chromatin**, from the Greek *chroma*, meaning "color." Chromatin is now known to consist of DNA plus various proteins that package the DNA in a more compact form. Prokaryotic DNA is also associated with protein to form condensed structures inside the cell. These structures differ from those observed in eukaryotes and are usually not called chromatin.

In a normal resting cell, chromatin exists as 30 nm fibers—long, slender threads about 30 nm in diameter. In humans, the nucleus must accommodate 46 such chromatin fibers, or chromosomes. The largest human chromosome is about $2.4 \times 10^8$ bp; it would be about 8 cm long if it were stretched out in the B conformation. During metaphase (when chromosomes are most condensed) the largest chromosome is about 10 $\mu$m long. The difference between the length of the metaphase chromosome and the extended B form of DNA is 8000-fold. This value is referred to as the *packing ratio*.

### A. Nucleosomes

The major proteins of chromatin are known as **histones**. Most eukaryotic species contain five different histones—H1, H2A, H2B, H3, and H4. All five histones are small, basic proteins containing numerous lysine and arginine residues whose positive charges allow the proteins to bind to the negatively charged sugar–phosphate backbone of DNA. The numbers of acidic and basic residues in typical mammalian histones are noted in Table 19.4. Except for H1, the amino acid sequence of each type of histone is highly conserved in all eukaryotes. For example, bovine histone H4 differs from pea histone H4 in only two residues out of 102. Such similarity in primary structure implies a corresponding conservation in tertiary structure and function.

Chromatin unfolds when it is treated with a solution of low ionic strength (<5 mM). The extended chromatin fiber looks like beads on a string in an electron micrograph (Figure 19.22). The "beads" are DNA–histone complexes called **nucleosomes** and the "string" is double-stranded DNA.

**Table 19.4 Basic and acidic residues in mammalian histones**

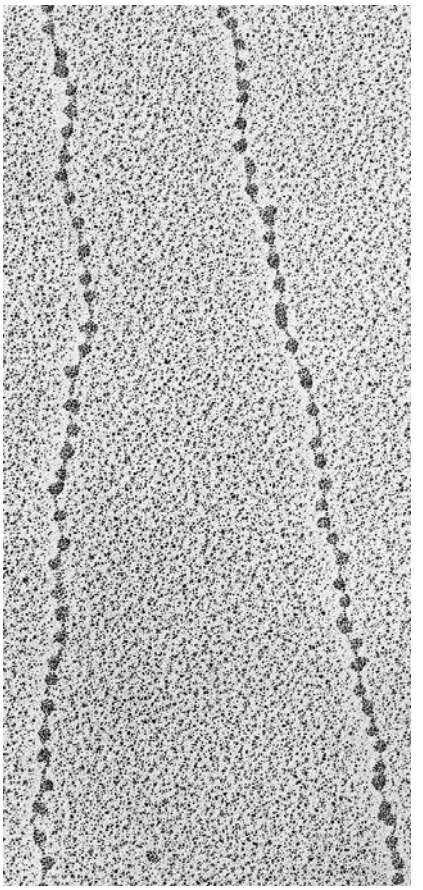
| Type | Molecular weight | Number of residues | Number of basic residues | Number of acidic residues |
|---|---|---|---|---|
| Rabbit thymus H1 | 21,000 | 213 | 65 | 10 |
| Calf thymus H2A | 14,000 | 129 | 30 | 9 |
| Calf thymus H2B | 13,800 | 125 | 31 | 10 |
| Calf thymus H3 | 15,300 | 135 | 33 | 11 |
| Calf thymus H4 | 11,300 | 102 | 27 | 7 |

Each nucleosome is composed of one molecule of histone H1, two molecules each of histones H2A, H2B, H3, and H4, and about 200 bp of DNA (Figure 19.23). The H2A, H2B, H3, and H4 molecules form a protein complex called the histone octamer around which the DNA is wrapped. About 146 bp of DNA are in close contact with the histone octamer forming a **nucleosome core particle**. The DNA between particles is called linker DNA; it is about 54 bp long. Histone H1 can bind to the linker DNA and to the core particle but in the extended beads-on-a-string conformation H1 is often absent. Histone H1 is responsible for higher-order chromatin structures.

The structure of the nucleosome core particle has been determined by X-ray crystallography (Figure 19.24). The eight histone subunits are arranged symmetrically as four dimers: two H2A/H2B dimers and two H3/H4 dimers. The particle is shaped like a flat disk with positively charged grooves that accommodate the sugar–phosphate backbone of DNA.

DNA wraps around the core particle forming about $1\frac{3}{4}$ turns per nucleosome. If this DNA were in an extended conformation it would be about 50 nm in length but when bound to the nucleosome core particle, the overall length is reduced to the width of the disk, about 5 nm. The coils of DNA are topologically equivalent to negative supercoils and that's why eukaryotic DNA becomes supercoiled when histones are removed from chromatin.

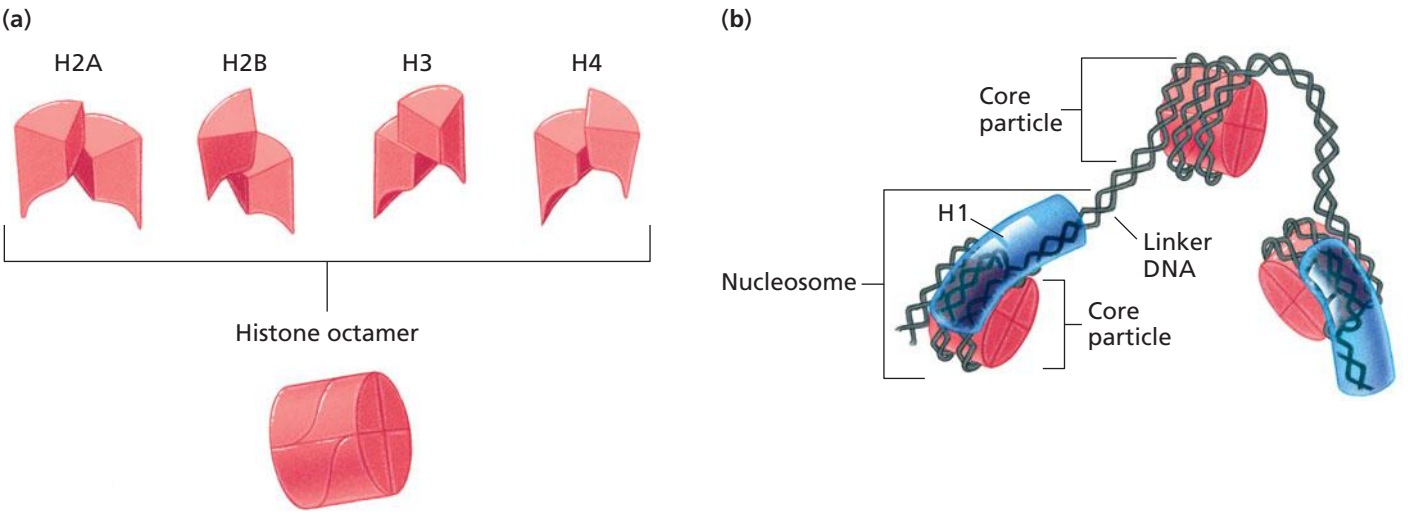
The N-termini of all four core histones are rich in positively charged lysine (K) and arginine (R) residues. These ends extend outward from the core particle where they interact with DNA and negatively charged regions of other proteins (Figure 19.24). These interactions serve to stabilize higher-order chromatin structures such as the 30 nm fiber.



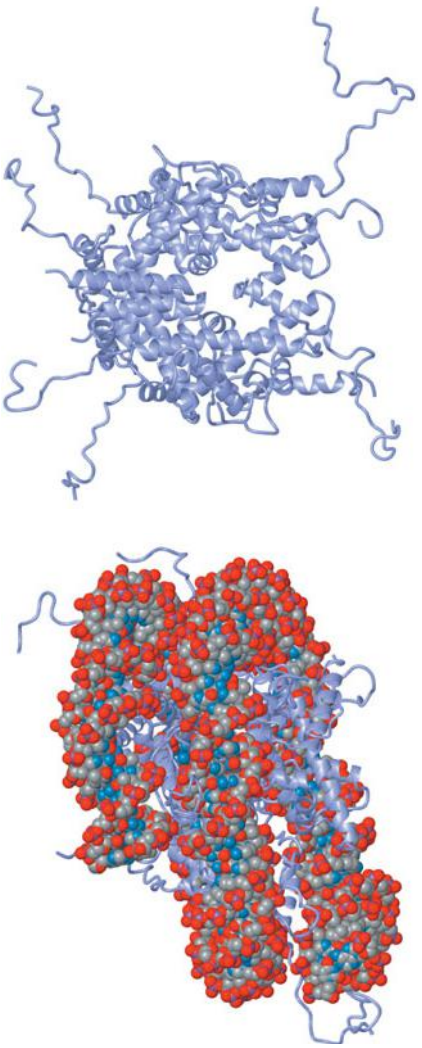
▲ **Figure 19.22**
**Electron micrograph of extended chromatin showing the "beads-on-a-string" organization.**

## KEY CONCEPT

**The vast majority of eukaryotic DNA is bound to nucleosome core particles spaced 200 bp apart.**




▲ **Figure 19.23**
**Diagram of nucleosome structure. (a)** Histone octamer. **(b)** Nucleosomes. Each nucleosome is composed of a core particle plus histone H1 and linker DNA. The nucleosome core particle is composed of a histone octamer and about 146 bp of DNA. Linker DNA consists of about 54 bp. Histone H1 binds to the core particle and to linker DNA.



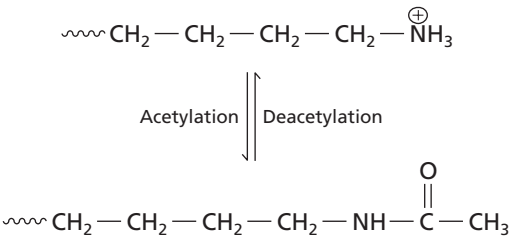
▲ **Figure 19.24**
**Structure of the chicken (*Gallus gallus*) nucleosome core particle. (a)** Histone octamer. **(b)** Histone octamer bound to DNA—side view showing the disk shape of the particle. [PDB 1EQZ].

Specific lysine residues in these N-terminal ends can be acetylated by enzymes known as histone acetyltransferases (HATS). For example, residues 5, 8, 12, 16, and 20 in histone H4 can be modified by acetylation.


**(19.1)**

Acetylation decreases the net positive charge of the histone N-termini and weakens the interactions with other nucleosomes and proteins. The net result is a loosening up of higher-order structures. Acetylation is associated with gene expression. HATS are preferentially directed to sites where chromatin must be unraveled in order to transcribe a gene. The relationship between transcriptional activation and histone acetylation is under active investigation in many laboratories (Section 21.5C).

Histone deacetylases are responsible for removing acetyl groups from lysine residues. This restores the positively charged side chains and allows nucleosomes to adopt the more compact chromatin structure characteristic of regions that are not expressed.




**(19.2)**

## B. Higher Levels of Chromatin Structure

The packaging of DNA into nucleosomes reduces the length of a DNA molecule about tenfold. Further reduction comes from higher levels of DNA packaging. For example, the beads-on-a-string structure is itself coiled into a solenoid to yield the 30 nm fiber. One possible model of the solenoid is shown in Figure 19.25. The 30 nm fiber forms when every nucleosome contains a molecule of histone H1 and adjacent molecules of H1 bind to each other cooperatively bringing the nucleosomes together into a more compact and stable form of chromatin. Condensation of the beads-on-a-string structure into a solenoid achieves a further fourfold reduction in chromosome length.

Finally, 30 nm fibers are themselves attached to an RNA–protein scaffold that holds the fibers in large loops. There may be as many as 2000 such loops on a large chromosome. The RNA–protein scaffold of a chromosome can be seen under an electron microscope when histones have been removed (Figure 19.26). The attachment of DNA loops to the scaffold accounts for an additional 200-fold condensation in the length of DNA.

The loops of DNA are attached to the scaffold at their base. Because the ends are not free to rotate, the loops can be supercoiled. (Some of the supercoils can be seen in Figure 19.26b, but most of the DNA is relaxed because one of the strands is broken during treatment to remove histones.)

## C. Bacterial DNA Packaging

Histones are found only in eukaryotes but prokaryotic DNA is also packaged with proteins in a condensed form. Some of these proteins are referred to as histone like proteins because they resemble eukaryotic histones. In most cases, there are no defined nucleosome-like particles in prokaryotes and much of the DNA is not associated with protein. Bacterial DNA is attached to a scaffold in large loops of about 100 kb. This arrangement converts the bacterial chromosome to a structure known as the nucleoid.

## 19.6 Nucleases and Hydrolysis of Nucleic Acids

Enzymes that catalyze the hydrolysis of phosphodiesters in nucleic acids are collectively known as **nucleases**. There are a variety of different nucleases in all cells. Some of them are required for the synthesis or repair of DNA as we will see in Chapter 20 and others are needed for the production or degradation of cellular RNA (Chapter 21).

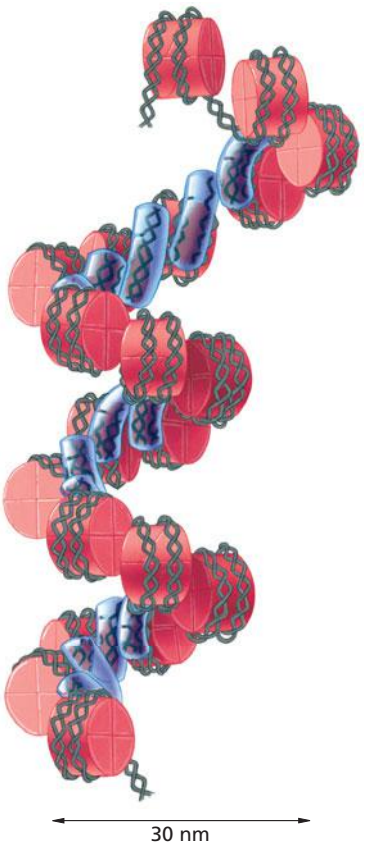
Some nucleases act on both RNA and DNA molecules but many act only on RNA and others only on DNA. The specific nucleases are called ribonucleases (RNases) and deoxyribonucleases (DNases). Nucleases can be further classified as exonucleases or endonucleases. **Exonucleases** catalyze the hydrolysis of phosphodiester linkages to release nucleotide residues from only one end of a polynucleotide chain. The most common exonucleases are the 3′ → 5′ exonucleases but there are some 5′ → 3′ exonucleases. **Endonucleases** catalyze the hydrolysis of phosphodiester linkages at various sites within a polynucleotide chain. Nucleases have a wide variety of specificities for nucleotide sequences.

Nucleases can cleave either the 3′- or the 5′-ester bond of a 3′–5′ phosphodiester linkage. One type of hydrolysis yields a 5′-monophosphate and a 3′-hydroxyl group; the other type yields a 3′-monophosphate and a 5′-hydroxyl group (see Figure 19.27). A given nuclease can catalyze one reaction or the other but not both.

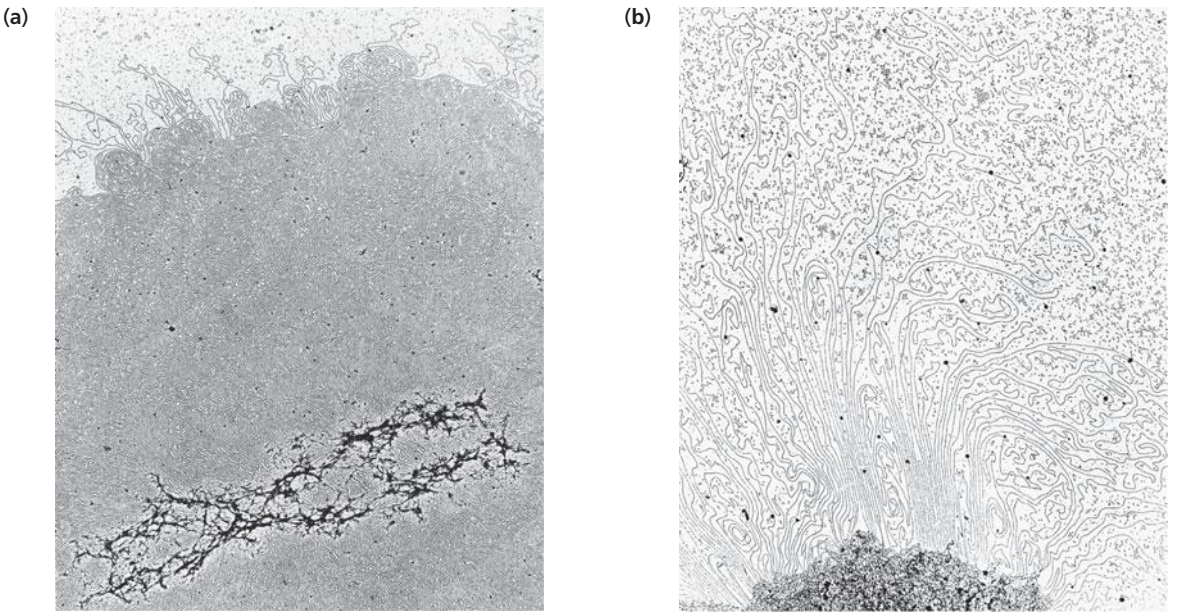
### A. Alkaline Hydrolysis of RNA

The difference between ribose in RNA and 2′-deoxyribose in DNA may seem trivial but it greatly affects the properties of the nucleic acids. The 2′-hydroxyl group of ribose can form hydrogen bonds in some RNA molecules and it participates in certain chemical and enzyme-catalyzed reactions.

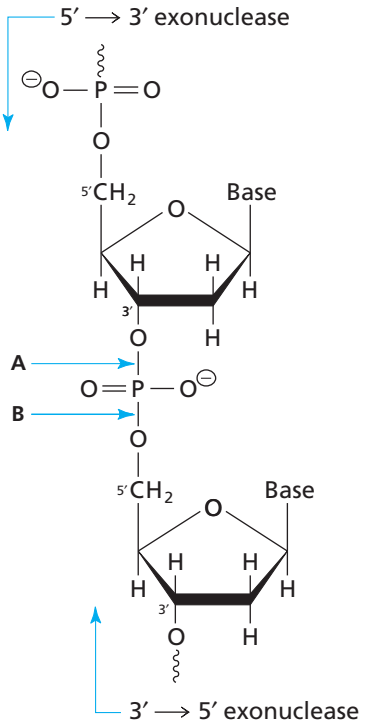
The effect of alkaline solutions on RNA and DNA illustrates the differences in chemical reactivity that result from the presence or absence of the 2′-hydroxyl group. RNA treated with 0.1 M NaOH at room temperature is degraded to a mixture of 2′- and 3′-nucleoside monophosphates within a few hours whereas DNA is stable under the same conditions. Alkaline hydrolysis of RNA (Figure 19.28) requires a 2′-hydroxyl group. In the first and second steps, hydroxide ions act only as catalysts



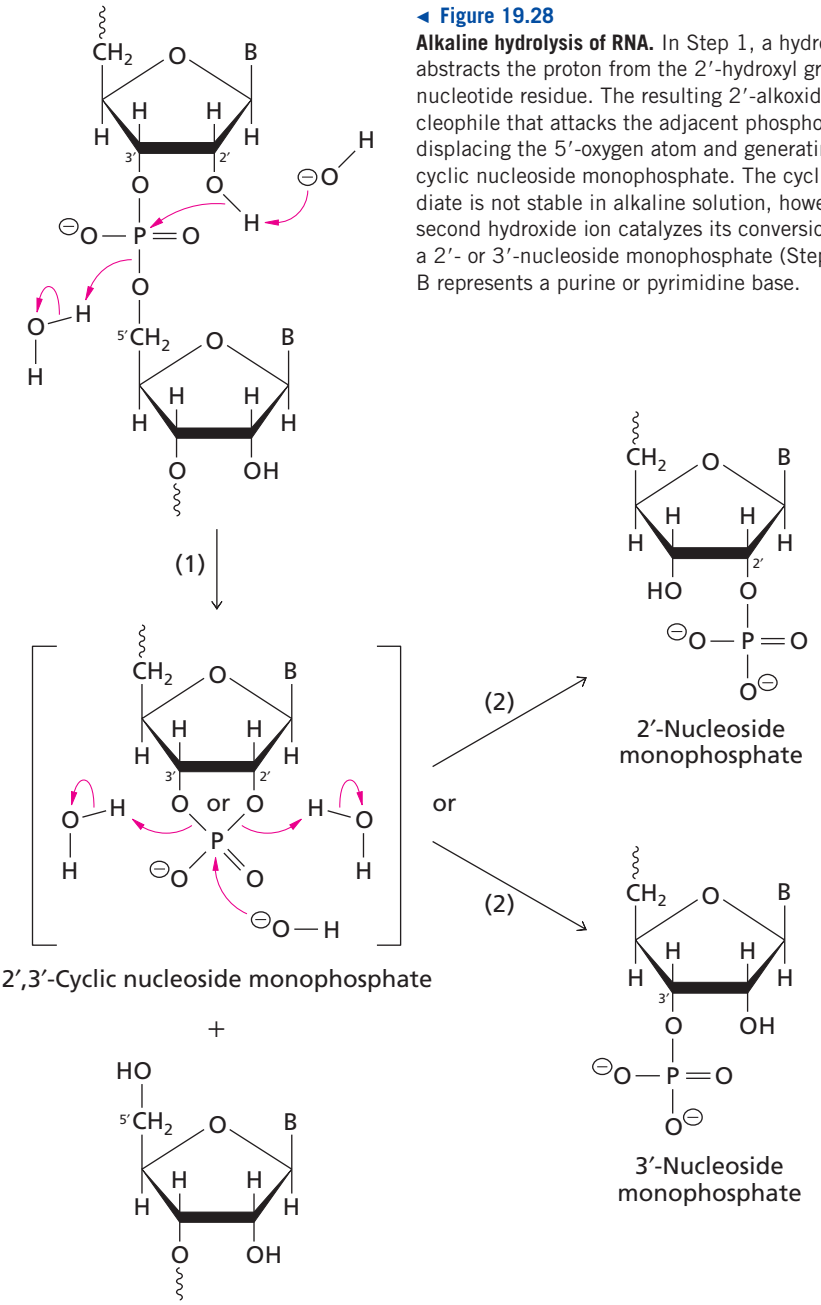

▲ **Figure 19.25**
**A model of the 30 nm chromatin fiber.** In this model the 30 nm fiber is shown as a solenoid, or helix, formed by individual nucleosomes. The nucleosomes associate through contacts between adjacent histone H1 molecules.




▲ **Figure 19.26**
**Electron micrographs of a histone-depleted chromosome. (a)** In this view, the entire protein scaffold is visible. **(b)** In this magnification of a portion of (a), individual loops attached to the protein scaffold can be seen.




▲ **Figure 19.27**
**Nuclease cleavage sites.** Exonucleases act on one free end of a polynucleotide and cleave the next phosphodiester linkage. Endonucleases cleave internal phosphodiester linkages. Cleavage at bond A generates a 5′-phosphate and a 3′-hydroxyl terminus. Cleavage at bond B generates a 3′-phosphate and a 5′-hydroxyl terminus. Both DNA (shown) and RNA are substrates of nucleases.




◀ **Figure 19.28**
**Alkaline hydrolysis of RNA.** In Step 1, a hydroxide ion abstracts the proton from the 2′-hydroxyl group of a nucleotide residue. The resulting 2′-alkoxide is a nucleophile that attacks the adjacent phosphorus atom, displacing the 5′-oxygen atom and generating a 2′,3′-cyclic nucleoside monophosphate. The cyclic intermediate is not stable in alkaline solution, however, and a second hydroxide ion catalyzes its conversion to either a 2′- or 3′-nucleoside monophosphate (Step 2). B represents a purine or pyrimidine base.

since removing a proton from water (to form the 5′-hydroxyl group in the first step or the 2′- or 3′-hydroxyl group in the second) regenerates one hydroxide ion for each hydroxide ion consumed. Note that a 2′,3′-cyclic nucleoside monophosphate intermediate forms. The polyribonucleotide chain rapidly depolymerizes as each phosphodiester linkage is cleaved. DNA is not hydrolyzed under alkaline conditions because it lacks the 2′-hydroxyl group needed to initiate intramolecular transesterification. The greater chemical stability of DNA is an important factor in its role as the primary genetic material.

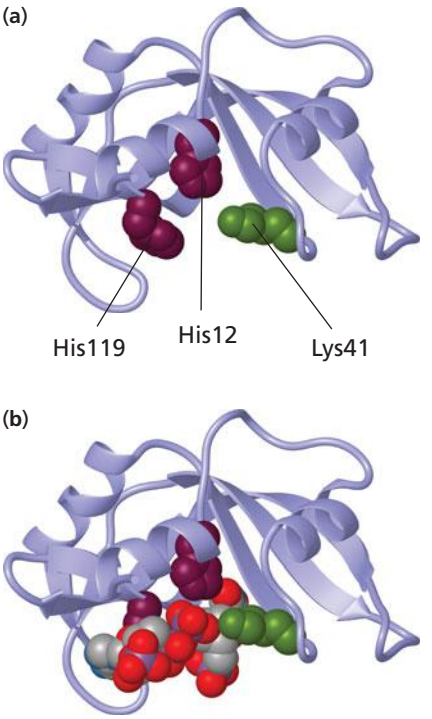
## B. Hydrolysis of RNA by Ribonuclease A

Bovine pancreatic ribonuclease A (RNase A) consists of a single polypeptide chain of 124 amino acid residues cross-linked by four disulfide bridges. (This is the same enzyme that we encountered in Chapter 4 in our discussion of disulfide bond formation

and protein folding.) The enzyme has a pH optimum of about 6. RNase A catalyzes cleavage of phosphodiester linkages in RNA molecules at 5′-ester bonds. Cleavage occurs to the right of pyrimidine nucleotide residues when chains are drawn in the 5′ → 3′ direction. Thus, RNase A catalyzed hydrolysis of a strand with the sequence pApG pUpApCpGpU yields pApGpUp + ApCp + GpU.

RNase A contains three ionic amino acid residues in the active site—Lys-41, His-12, and His-119 (Figure 19.29). Many studies have led to formulation of the mechanism of catalysis shown in Figure 19.30. RNase A uses three fundamental catalytic mechanisms: proximity (in the binding and positioning of a suitable phosphodiester between the two histidine residues); acid–base catalysis (by His-119 and His-12); and transition-state stabilization (by Lys-41). As in alkaline hydrolysis of RNA, hydrolysis produces a leaving group with a 5′-hydroxyl group and a 3′-nucleoside monophosphate product. Water enters the active site on departure of the first product ($P_1$). Note that in the RNase A–catalyzed reaction, the phosphate atom in the transition state is pentacovalent. The pyrimidine binding pocket of the enzyme accounts for the specificity of RNase A.

Alkaline hydrolysis and the reaction catalyzed by RNase A differ in two important ways. First, alkaline hydrolysis can occur at any residue whereas enzyme-catalyzed cleavage occurs only at pyrimidine nucleotide residues. Second, hydrolysis of the cyclic intermediate is random in alkaline hydrolysis (producing mixtures of 2′- and 3′-nucleotides) but specific for RNase A–catalyzed cleavage (producing only 3′-nucleotides).




▲ **Figure 19.29**
**The active site of bovine pancreatic RNase A.** **(a)** The active site of the enzyme has three catalytic residues, His-12, His-119, and Lys-41, whose side chains project into the site where RNA will bind. **(b)** This figure shows RNase A bound to an artificial substrate (3′-phosphothymidine (3′-5′)-pyrophosphate adenosine 3′-phosphate) that mimics RNA. [PDB 1U1B]

## C. Restriction Endonucleases

Restriction endonucleases are an important subclass of endonucleases that act on DNA. The term **restriction endonuclease** is derived from the observation that certain bacteria can block bacteriophage (virus) infections by specifically destroying the incoming bacteriophage DNA. Such bacteria *restrict* the expression of foreign DNA.

Many species of bacteria synthesize restriction endonucleases that bind to and cleave foreign DNA. These endonucleases recognize specific DNA sequences and they cut both strands of DNA at the binding site producing large fragments that are rapidly degraded by exonucleases. The bacteriophage DNA is cleaved and degraded before the genes can be expressed.

The host cell has to protect its own DNA from cleavage by restriction endonucleases. This is accomplished by covalent modification of the bases that make up the potential restriction endonuclease binding site. The most common covalent modification is specific methylation of adenine or cytosine residues within the recognition sequence (Section 18.7). The presence of methylated bases at the potential binding site inhibits cleavage of the host DNA by the restriction endonuclease. Methylation is catalyzed by a specific methylase that binds to the same sequence of DNA recognized by the restriction endonuclease. Thus, cells that contain a restriction endonuclease also contain a methylase with the same specificity.

Normally, all DNA of the host cell is specifically methylated and therefore protected from cleavage. Any unmethylated DNA that enters the cell is cleaved by restriction endonucleases. Following DNA replication, each site in the host DNA is hemimethylated—bases on only one strand are methylated. Hemimethylated sites are high affinity substrates for the methylase but are not recognized by the restriction endonuclease. Thus, hemimethylated sites are rapidly converted to fully methylated sites in the host DNA (Figure 19.31).

Most restriction endonucleases (also called restriction enzymes) can be classified as either type I or type II. Type I restriction endonucleases catalyze both the methylation of host DNA and the cleavage of unmethylated DNA at a specific recognition sequence. Type II restriction endonucleases are simpler in that they can only cleave double-stranded DNA at or near an unmethylated recognition sequence—they do not possess a methylase activity. Separate restriction methylases catalyze methylation of host DNA at the same recognition sequences. The source of the methyl group in these reactions is *S*-adenosylmethionine.

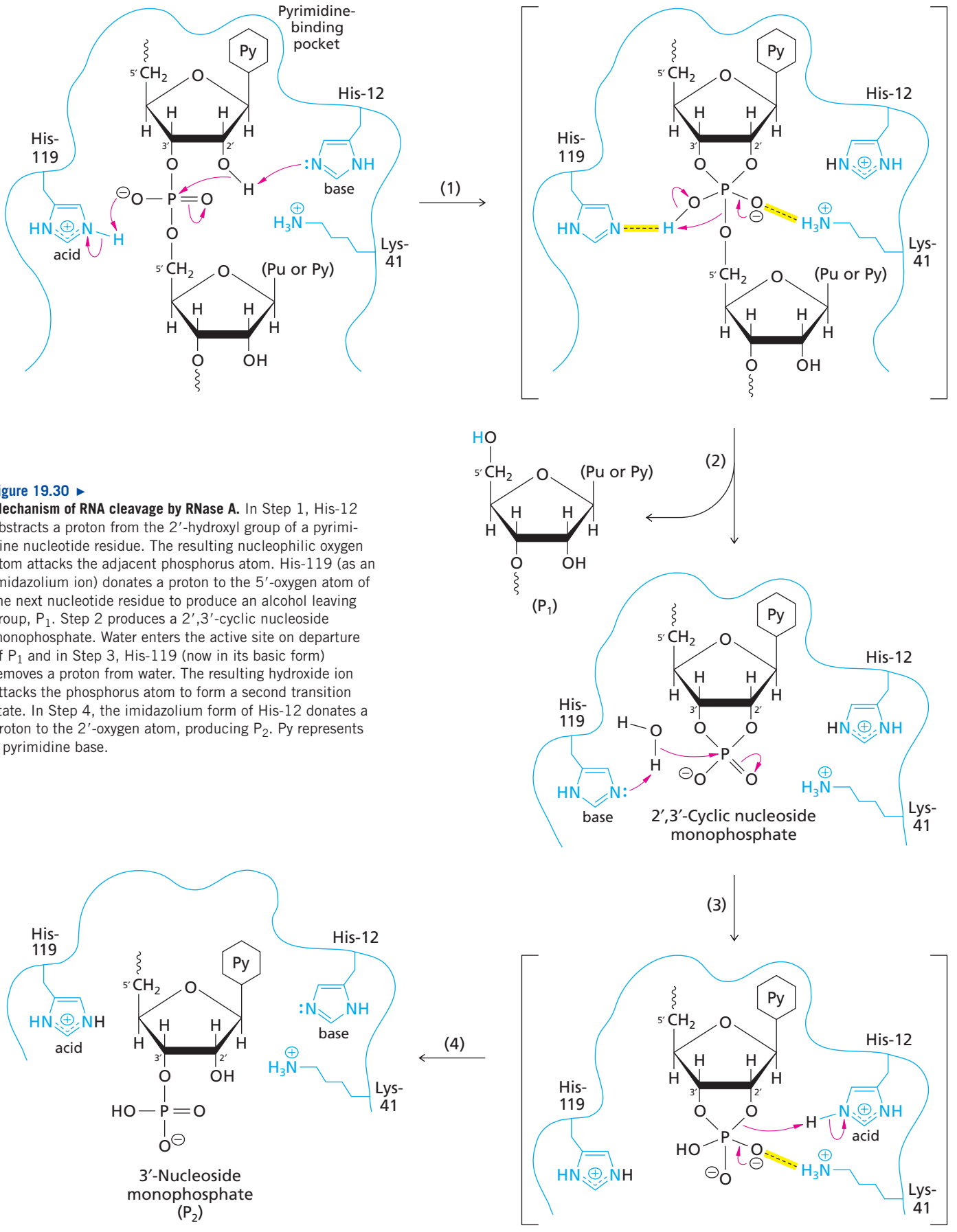



**Figure 19.30** ▶
**Mechanism of RNA cleavage by RNase A.** In Step 1, His-12 abstracts a proton from the 2′-hydroxyl group of a pyrimidine nucleotide residue. The resulting nucleophilic oxygen atom attacks the adjacent phosphorus atom. His-119 (as an imidazolium ion) donates a proton to the 5′-oxygen atom of the next nucleotide residue to produce an alcohol leaving group, $P_1$. Step 2 produces a 2′,3′-cyclic nucleoside monophosphate. Water enters the active site on departure of $P_1$ and in Step 3, His-119 (now in its basic form) removes a proton from water. The resulting hydroxide ion attacks the phosphorus atom to form a second transition state. In Step 4, the imidazolium form of His-12 donates a proton to the 2′-oxygen atom, producing $P_2$. Py represents a pyrimidine base.

**Table 19.5 Specificities of some common restriction endonucleases**

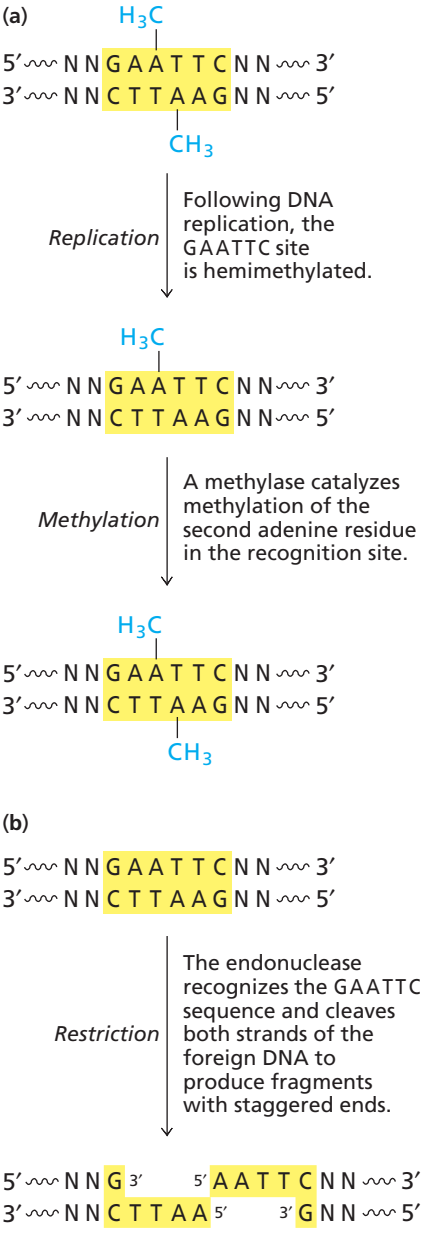
| Source | Enzyme[a] | Recognition sequence[b] |
|---|---|---|
| *Acetobacter pasteurianus* | *Apa*I | GGGCC↓C |
| *Bacillus amyloliquefaciens* H | *Bam*HI | G↓GATCC |
| *Eschericia coli* RY13 | *Eco*RI | G↓AA*TTC |
| *Eschericia coli* R245 | *Eco*RII | ↓CC*TGG |
| *Haemophilus aegyptius* | *Hae*III | GG↓CC |
| *Haemophilus influenzae* $R_d$ | *Hind*III | A*↓AGCTT |
| *Haemophilus parainfluenzae* | *Hpa*II | C↓CGG |
| *Klebsiella pneumoniae* | *Kpn*I | GGTAC↓C |
| *Nocardia otitidis-caviarum* | *Not*I | GC↓GGCCGC |
| *Providencia stuartii* 164 | *Pst*I | CTGCA↓G |
| *Serratia marcescens* $S_b$ | *Sma*I | CCC↓GGG |
| *Xanthomonas badrii* | *Xba*I | T↓CTAGA |
| *Xanthomonas holcicola* | *Xho*I | C↓TCGAG |

[a]The names of restriction endonucleases are abbreviations of the names of the organisms that produce them. Some abbreviated names are followed by a letter denoting the strain. Roman numerals indicate the order of discovery of the enzyme in that strain.
[b]Recognition sequences are written 5′ to 3′. Only one strand is represented. The arrows indicate cleavage sites. Asterisks represent known positions where bases can be methylated.

Hundreds of type I and type II restriction endonucleases have been characterized. The specificities of a few representative enzymes are listed in Table 19.5. In nearly all cases, the recognition sites have a twofold axis of symmetry; that is, the 5′ ⟶ 3′ sequence of residues is the same in both strands of the DNA molecule. Consequently, the paired sequences "read" the same in either direction—such sequences are known as palindromes. (Palindromes in English include BIB, DEED, RADAR, and even MADAM I'M ADAM, provided we ignore punctuation and spacing.)

*Eco*RI was one of the first restriction endonucleases to be discovered. It is present in many strains of *E. coli*. As shown in Table 19.5 and Figure 19.31, *Eco*RI has a palindromic recognition sequence of 6 bp (the 5′ ⟶ 3′ sequence is GAATTC on each strand). *Eco*RI is a homodimer. It possesses a twofold axis of symmetry like its substrate (see next section). In *E. coli*, the companion methylase to *Eco*RI converts the second adenine within the recognition sequence to $N^6$-methyladenine. Any double-stranded DNA molecule with an unmethylated GAATTC sequence is a substrate for *Eco*RI. The endonuclease catalyzes hydrolysis of the phosphodiesters that link G to A in each strand, thus cleaving the DNA.

Some restriction endonucleases (including *Eco*RI, *Bam*HI, and *Hin*dIII) catalyze staggered cleavage, producing DNA fragments with single-stranded extensions (Table 19.5 and Figure 19.31). These single-stranded regions are called sticky ends because they are complementary and can thus re-form a double-stranded structure. Other enzymes, such as *Hae*III and *Sma*I, produce blunt ends with no single-stranded extensions.

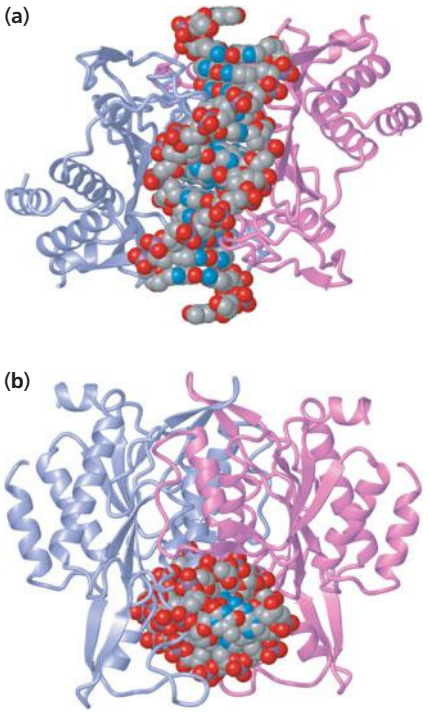



▲ **Figure 19.31**
**Methylation and restriction at the *Eco*RI *site.*** **(a)** Methylation of adenine residues at the recognition site. **(b)** Cleavage of unmethylated DNA to produce sticky ends.

## D. *Eco*RI Binds Tightly to DNA

Restriction endonucleases must bind tightly to DNA in order to recognize a specific sequence and cleave at a specific site. The structure of *Eco*RI bound to DNA has been determined by X-ray crystallography. As shown in Figure 19.32, each half of the *Eco*RI homodimer binds to one side of the DNA molecule so that the DNA molecule is almost surrounded. The enzyme recognizes the specific nucleotide sequence by contacting base pairs in the major groove. The minor groove (in the middle of the structure shown in Figure 19.32) is exposed to the aqueous environment.

Several basic amino acid residues line the cleft that is formed by the two *Eco*RI monomers. The side chains of these residues interact electrostatically with the




▲ **Figure 19.32**
***Eco*RI bound to DNA.** *Eco*RI is composed of two identical subunits (purple and blue). The enzyme is bound to a fragment of DNA with the sequence CGCGAATTCGCG (recognition sequence underlined). **(a)** Side view. **(b)** Top view.

sugar–phosphate backbones of DNA. In addition, two arginine residues (Arg-145 and Arg-200) and one glutamate residue (Glu-144) in each *Eco*RI monomer form hydrogen bonds with base pairs in the recognition sequence thus ensuring specific binding. Other nonspecific interactions with the backbones further stabilize the complex.

*Eco*RI is typical of proteins that recognize and bind to a *specific* DNA sequence. The DNA retains its B conformation although in some cases the helix is slightly bent. Recognition of a specific nucleotide sequence depends on interactions between the protein and the functional groups on the bases that are exposed in the grooves. In contrast, histones are examples of proteins that bind *nonspecifically* to nucleic acids. Binding of such proteins depends largely on weak interactions between the protein and the sugar–phosphate backbones and not on direct contact with the bases. All proteins that bind to specific DNA sequences will also bind non-specifically to DNA with lower affinity (Sections 21.3, 21.7A).

## 19.7 Uses of Restriction Endonucleases

Restriction endonucleases were discovered more than 40 years ago earning Nobel Prizes in 1978 for Werner Arbor, Daniel Nathans, and Hamilton Smith "for the discovery of restriction enzymes and their application to problems of molecular genetics." The first purified enzymes rapidly became important tools used to manipulate DNA in the laboratory.

### A. Restriction Maps

One of the first uses of restriction enzymes was in developing restriction maps of DNA, that is, diagrams of DNA molecules that show specific sites of cleavage. Such maps are useful for identifying fragments of DNA that contain specific genes.

An example of a restriction map of bacteriophage λ DNA is shown in Figure 19.33. The DNA of bacteriophage λ is a linear, double-stranded molecule approximately 48,400 bp (48 kb) long. By treating this DNA with various restriction enzymes and measuring the sizes of the resulting fragments, it is possible to develop a map of the cleavage sites. An example of such restriction digests is shown in Figure 19.34. The information from many restriction digests is combined to produce a complete and accurate map.

### B. DNA Fingerprints

The technology required for mapping restriction endonuclease cleavage sites was developed in the 1970s. It soon became apparent that the procedure could be used to identify the sites of mutations, or variations, in the genome of a population. For example, different strains of bacteriophage λ have slightly different restriction maps because their DNA sequences are not identical. One strain may have the sequence GGGCCC near the left-hand end of its DNA and it is cleaved by *Apa*I, producing the two fragments shown in Figure 19.34. Another strain may have the sequence GGACCC at the same site. Since this sequence is not a cleavage site for *Apa*I, the restriction map of this strain differs from that shown in Figure 19.33.

Variations in DNA sequence can be used to identify individuals in a large heterogeneous population. In humans, for example, regions of the genome that are highly variable give restriction fragments that are as unique as fingerprints. Such DNA fingerprints can be used in paternity disputes or criminal investigations to identify or exonerate suspects.

An example of the use of DNA fingerprinting in a rape case is shown in Figure 19.35. DNAs isolated from the victim, from the evidence (semen), and from two suspects are

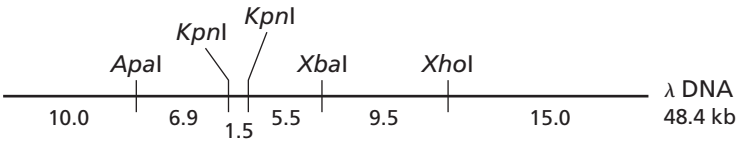
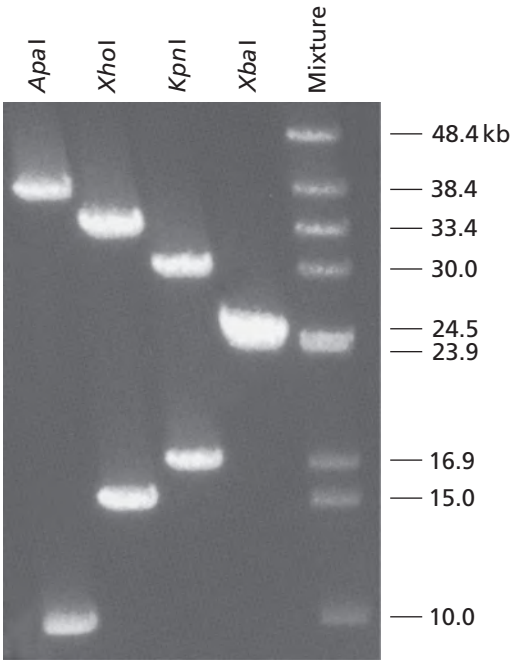
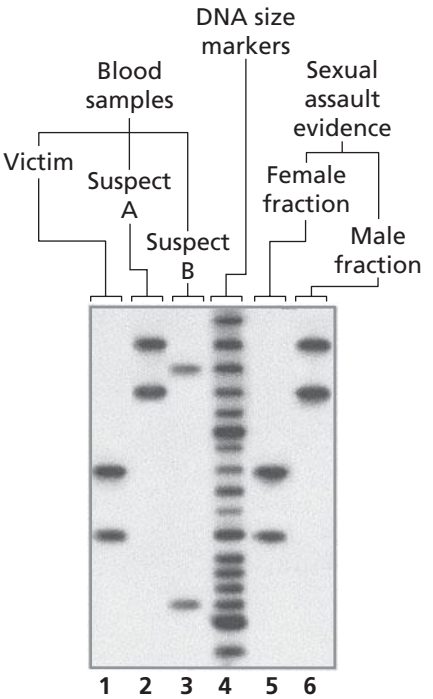



**Figure 19.33 ►**
**Restriction map of bacteriophage λ showing the sites of cleavage by some restriction enzymes.** There is a single site for the enzyme *Apa*I, for example. Digestion of phage λ DNA with this enzyme yields two fragments of 10.0 and 38.4 kb, as shown in the first lane of Figure 19.34.




▲ **Figure 19.34**
**Digestion of bacteriophage λ DNA by four restriction endonucleases.** A solution of DNA is treated with an enzyme and then electrophoresed on an agarose gel, which separates fragments according to size. The smallest fragments move fastest and are found at the bottom of the gel. (A fragment of 1.5 kb is not visible in this figure.) The restriction enzyme for each digest is indicated at the top of the lane. The lane at the right contains intact phage λ DNA and a mixture of fragments from the four digests. In the *Xba*I digest, two fragments of 23.9 and 24.5 kb are not well resolved.




▲ **Figure 19.35**
**DNA fingerprinting.**

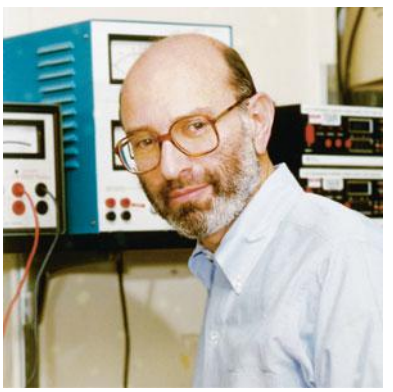
digested with a restriction endonuclease. The fragments are separated on an agarose gel as described in Figure 19.34. This DNA is then transferred (blotted) to a membrane of nylon. The bound DNA is denatured and exposed to small fragments of radioactively labeled DNA from a variable region of the human genome. The labeled DNA probe hybridizes specifically to the restriction fragments on the nylon membrane that are derived from this region. The labeled fragments are identified by autoradiography.

The technique identifies suspect A as the rapist. In actual criminal investigations, a number of different probes are used in combination with different restriction digests in order to ensure that the pattern detected is unique. Modern techniques are powerful and accurate enough to conclusively rule out some suspects and convict others. When combined with polymerase chain reaction (PCR) amplification of DNA (Chapter 22), a fingerprint can be obtained from a hair follicle or a tiny speck of blood.

## C. Recombinant DNA

The discovery of restriction endonucleases soon led to the creation of recombinant DNA molecules by joining, or recombining, different fragments of DNA produced by the enzymes. A common experiment involves excising a DNA fragment containing a target gene of interest and inserting it into a cloning vector. Cloning vectors can be plasmids, bacteriophage, viruses, or even small artificial chromosomes. Most vectors contain sequences that allow them to be replicated autonomously within a compatible host cell.

All cloning vectors have in common at least one unique cloning site, a sequence that can be cut by a restriction endonuclease to allow site-specific insertion of foreign DNA. The most useful vectors have several restriction sites grouped together in a multiple cloning site called a *polylinker*.



► Stanley N. Cohen (1935–) (top) and Herbert Boyer (1936–) (bottom), who constructed the first recombinant DNA using bacterial DNA and plasmids.

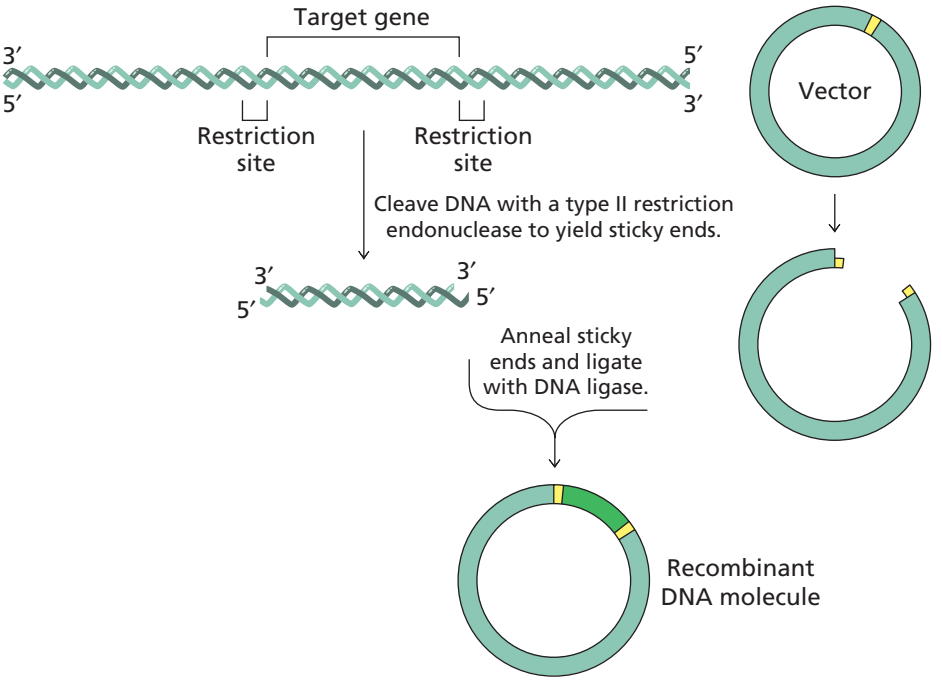



**Figure 19.36 ►**
**Use of restriction enzymes to generate recombinant DNA.** The vector DNA and the target DNA are cleaved by restriction endonucleases to generate ends that can be joined together. In cases where sticky ends are produced, the two molecules join by annealing (base pairing) of the complementary ends. The molecules are then covalently attached to one another in a reaction catalyzed by DNA ligase.

Fragments of DNA to be inserted into a vector can be generated by a variety of means. For example, they can be produced by the mechanical shearing of long DNA molecules or by digesting DNA with type II restriction endonucleases. Unlike shearing, which cleaves DNA randomly, restriction enzymes cleave DNA at specific sequences. For cloning purposes, this specificity offers extraordinary advantages.

The most useful restriction endonucleases produce fragments with single-stranded extensions at their 3′ or 5′ ends. These sticky ends can transiently form base pairs to complementary sticky ends on vector DNA and can be covalently joined to the vector in a reaction catalyzed by DNA ligase (described in Section 20.3C). Thus, the simplest kinds of recombinant DNA are those constructed by digesting both the vector and the target DNA with the same enzyme because the resulting fragments can be joined directly by ligation (Figure 19.36).

## Summary

1. Nucleic acids are polymers of nucleotides that are phosphate esters of nucleosides. The amino and lactam tautomers of the bases form hydrogen bonds in nucleic acids.
2. DNA contains two antiparallel strands of nucleotide residues joined by 3′–5′ phosphodiester linkages. A and G in one strand pair with T and C, respectively, in the other strand.
3. The double-helical structure of DNA is stabilized by hydrogen bonding, hydrophobic effects, stacking interactions, and charge–charge interactions. G/C-rich DNA is more difficult to denature than A/T-rich DNA because the stacking interactions of G/C base pairs are greater than those of A/T base pairs.
4. The most common conformation of DNA is called B-DNA; alternative conformations include A-DNA and Z-DNA.
5. Overwinding or underwinding the DNA helix can produce supercoils that restore the B conformation. Negatively supercoiled DNA exists in equilibrium with DNA that has locally unwound regions.
6. The four major classes of RNA are ribosomal RNA, transfer RNA, messenger RNA, and small RNA. RNA molecules are single-stranded and have extensive secondary structure.
7. Eukaryotic DNA molecules are packaged with histones to form nucleosomes. Further condensation and attachment to the scaffold of a chromosome achieves an overall 8000-fold reduction in the length of the DNA molecule in metaphase chromosomes.
8. The phosphodiester backbones of nucleic acids can be hydrolyzed by the actions of nucleases. Alkaline hydrolysis and RNase A-catalyzed hydrolysis of RNA proceed via a 2′,3′-cyclic nucleoside monophosphate intermediate.
9. Restriction endonucleases catalyze hydrolysis of DNA at specific palindromic nucleotide sequences. Specific methylases protect restriction sites from cleavage.
10. Restriction enzymes are useful for constructing restriction maps of DNA, for DNA fingerprint analysis, and for constructing recombinant DNA molecules.

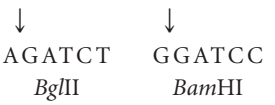
# Problems

1. Compare hydrogen bonding in the $\alpha$ helix of proteins to hydrogen bonding in the double helix of DNA. Include in the answer the role of hydrogen bonding in stabilizing these two structures.
2. A stretch of double-stranded DNA contains 1000 bp and its base composition is 58% (G + C). How many thymine residues are in this region of DNA?
3. (a) Do the two complementary strands of a segment of DNA have the same base composition?
   (b) Does (A + G) equal (C + T)?
4. If one strand of DNA has the sequence

   ATCGCGTAACATGGATTCGG

   write the sequence of the complementary strand using the standard convention.
5. Poly A forms a single-stranded helix. What forces stabilize this structure?
6. The imino tautomer of adenine occurs infrequently in DNA but when it does it can pair with cytosine instead of thymine. Such mispairing can lead to a mutation. Draw the adenine imino tautomer/cytosine base pair.
7. Single-stranded poly-dA can hybridize to single-stranded poly-dT to form Watson–Crick base-paired double-stranded DNA. Under appropriate conditions a second strand of poly-dT can bind in the major groove and form a triple-stranded DNA helix with hydrogen bonds between the thymine and the N7 and amino group in adenine. What would a plot of absorbance at 260 nm vs. temperature look like for this unusual triple-stranded DNA?
8. Write the sequence of the RNA shown in Figure 19.21. Is it a palindrome?
9. Consider a processive exonuclease that binds exclusively to double-stranded DNA and degrades one strand in the 5′ → 3′ direction. In a reaction where the substrate is a 1 kb fragment of linear DNA, what will be the predominant products after the digestion has gone to completion?
10. The average molecular weight of a base pair in double-stranded DNA is approximately 650 kDa. Using the data from Table 19.4, calculate the mass ratio of protein to DNA in a typical 30 nm chromatin fiber.
11. The human haploid genome contains $3.2 \times 10^9$ base pairs. How many nucleosomes did you inherit from your mother?
12. A DNA molecule with the sequence pdApdGpdTpdC can be cleaved by exonucleases. List the products of a single reaction catalyzed by the following enzymes:
   (a) a 3′ → 5′ exonuclease that cleaves the 3′ ester bond of a phosphodiester linkage
   (b) a 5′ → 3′ exonuclease that cleaves the 5′ ester bond of a phosphodiester linkage
   (c) a 5′ → 3′ exonuclease that cleaves the 3′ ester bond of a phosphodiester linkage
13. A non-sequence-specific endonuclease purified from *Aspergillus oryzae* digests single-stranded DNA. Predict the effect of adding this enzyme to a preparation of negatively supercoiled plasmid DNA.
14. One of the proteins in rattlesnake venom is an enzyme named phosphodiesterase. Could polynucleotides be a substrate for this enzyme? Why or why not?
15. RNase T1 cleaves RNA after G residues to leave a 3′ phosphate group. Predict the cleavage products of this substrate:

   pppApCpUpCpApUpApGpCpUpApUpGpApGpU
16. How could bacteriophages escape the effects of bacterial restriction endonucleases?
17. The free-living soil nematode *C. elegans* was the first metazoan to have its entire 100 Mb genome sequenced. Overall, the worm genome is 36% (G + C) and 64% (A + T). The restriction endonuclease *Hin*dIII recognizes and cuts the hexameric palindromic sequence AAGCTT to generate sticky ends. (a) Approximately how many *Hin*dIII sites would you expect to find in the *C. elegans* genome? (b) If the worm genome was actually 25% G and 25% A, approximately how many *Hin*dIII sites would you expect to find?
18. The recognition sites for the restriction endonucleases *Bgl*II and *Bam*HI are shown below. Why is it possible to construct recombinant DNA molecules by combining target DNA cut with *Bgl*II and a vector cut with *Bam*HI?




19. One of the *E. coli* host strains commonly used in recombinant DNA technology carries defective genes for several restriction endonucleases. Why is such a strain useful?

# Selected Readings


## Historical Perspective

Clayton, J., and Denis. C. (eds.) (2003). *50 Years of DNA*. (New York: Nature/Pallgrave/Macmillan).

Judson, H. F. (1996). *The Eighth Day of Creation: Makers of the Revolution in Biology*, expanded ed. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Maddox, B. (2002). *Rosalind Franklin: The Dark Lady of DNA* (New York: Perennial/HarperCollins).

Watson, J. D., and Berry, A. (2003). *DNA: The Secret of Life* (New York: Alfred A. Knopf).

Watson, J. D. (1968). *The Double Helix* (New York: Atheneum).

## Polynucleotide Structure and Properties

Berger, J. M., and Wang, J. C. (1996). Recent developments in DNA topoisomerase II structure and mechanism. *Curr. Opin. Struct. Biol.* 6:84–90.

Ferré-D'Amaré, A. R., and Doudna, J. A. (1999). RNA FOLDS: insights from recent crystal structures. *Annu. Rev. Biophys. Biomol. Struct.* 28:57–73.

Herbert, A., and Rich, A. (1996). The biology of left-handed Z-DNA. *J. Biol. Chem.* 271:11595–11598.

Hunter, C. A. (1996). Sequence-dependent DNA structure. *BioEssays* 18:157–162.

Ke, C., Humeniuk, M., S-Gracz, H., and Marszalek, P. E. (2007). Direct measurements of base stacking interactions in DNA by single-molecule atomic-force spectroscopy. *Phys. Rev. Lett.* 99: 018302.

Kool, E. T., Morales, J. C., and Guckian, K. M. (2000). Mimicking the structure and function of DNA: insights into DNA stability and replication. *Angew. Chem. Int. Ed.* 39:990–1009.

Packer, M. J., and Hunter, C. A. (1998). Sequence-dependent DNA structure: the role of the sugar-phosphate backbone. *J. Mol. Biol.* 280:407–420.

Saenger, W. (1984). *Principles of Nucleic Acid Structure* (New York: Springer-Verlag).

Sharma, A., and Mondragón, A. (1995). DNA topoisomerases. *Curr. Biol.* 5:39–47.

Wang, J. C. (2009). A journey in the world of DNA rings and beyond. *Annu. Rev. Biochem.* 78:31–54.

## Chromatin

Bendich, A. J., and Drlica, K. (2000). Prokaryotic and eukaryotic chromosome: what's the difference? *BioEssays* 22:481–486.

Burlingame, R. W., Love, W. E., Wang, B.-C., Hamlin, R., Xuong, N.-H., and Moudrianakis, E. N. (1985). Crystallographic structure of the octameric histone core of the nucleosome at a resolution of 3.3 Å. *Science* 228:546–553.

Grigoryev, S. A., Arya, G., Correll, S., Woodcock, C. L., and Schlick, T. (2009). Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci. (USA)* 106:13317–13322.

Kornberg, R. D. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome. *Cell* 98:285–294.

Ramakrishnan, V. (1997). Histone structure and the organization of the nucleosome. *Annu. Rev. Biophys. Biomol. Struct.* 26:83–112.

Richmond, T. J., Finch, J. T., Rushton, D., Rhodes, D., and Klug, A. (1984). Structure of the nucleosome core particle at 7 Å resolution. *Nature* 311:532–537.

Van Holde, K., and Zlatanova, J. (1999). The nucleosome core particle: does it have structural and functional relevance? *BioEssays* 21:776–780.

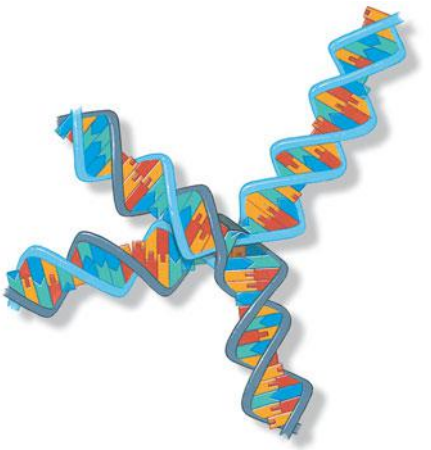
Workman, J. L., and Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.* 67:545–579.

## Restriction Endonucleases

Kovall, R. A., and Mathews, B. W. (1999). Type II restriction endonucleases: structural, functional and evolutionary relationships. *Curr. Opin. Chem. Biol.* 3:587–583.

McClarin, J. A., Frederick, C. A., Wang, B.-C., Greene, P., Boyer, H., Grable, J., and Rosenberg, J. M. (1986). Structure of the DNA-*Eco*RI endonuclease recognition complex at 3 Å resolution. *Science* 234:1526–1541.

Ne, M. (2000). Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Rev.* 64:412–434.

# DNA Replication, Repair, and Recombination

The transfer of genetic information from one generation to the next has puzzled biologists since the time of Aristotle. Today, almost 2500 years later, we can explain why "like begets like." Since genetic information is carried in DNA, it follows that the transfer of information from a parental cell to two daughter cells requires exact duplication of DNA, a process known as DNA replication.

The DNA structure proposed by Watson and Crick in 1953 immediately suggested a method of replication. The nucleotide sequence of one strand automatically specifies the sequence of the other since the two strands of double-helical DNA are complementary. Watson and Crick proposed that the two strands of the helix unwind during DNA replication and that each strand of DNA acts as a template for the synthesis of a complementary strand. In this way, DNA replication produces two double-stranded daughter molecules, each containing one parental strand and one newly synthesized strand. This mode of replication is termed semiconservative because one strand of the parental DNA is conserved in each daughter molecule (Figure 20.1, on the next page).

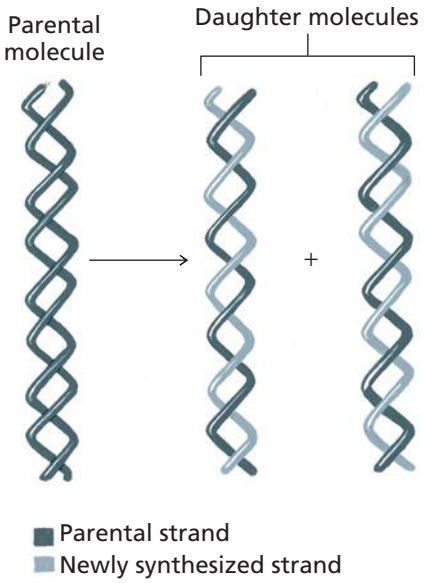
In a series of elegant experiments, Matthew Meselson and Franklin W. Stahl showed in 1958 that DNA was indeed replicated semiconservatively as predicted by Watson and Crick. About the same time, reports of the purification and properties of some of the enzymes involved in replication began to appear. The first DNA polymerase was purified in 1958 by Arthur Kornberg, who was awarded the Nobel Prize for this achievement. More recently, biochemists have isolated and characterized enzymes that catalyze all the steps in DNA replication and have identified the genes that encode these proteins. The actual mechanism of replication is much more complex—and more interesting—than the simple scheme shown in Figure 20.1.

Establishing the steps of the replication mechanism required a combination of both biochemical and genetic analysis. Much of what we know about DNA replication

*The structure of DNA proposed by Watson and Crick brought forth a number of proposals as to how such a molecule might replicate. These proposals make specific predictions concerning the distribution of parental atoms among progeny molecules. The results presented here give a detailed answer to the question of this distribution and simultaneously direct our attention to other problems whose solution must be the next step in progress toward a complete understanding of the molecular basis of DNA duplication.*

—Matthew Meselson and Franklin W. Stahl (1958)

**Top:** Holliday junction, an intermediate formed during recombination between two double-stranded DNA molecules.




◂ **Figure 20.1**
**Semiconservative DNA replication.** Each strand of DNA acts as a template for synthesis of a new strand. Each daughter molecule of DNA contains one parental strand and one newly synthesized strand.

has come from studies of the enzymes in *Escherichia coli* and its bacteriophages. The results of these studies have shown how large numbers of polypeptides assemble into complexes that can carry out a complicated series of reactions. The DNA replication complex is like a machine, or factory, whose parts are made of protein. Some of the component polypeptides are partially active in isolation but others are active only in association with the complete protein machine.

There are three distinct stages in DNA replication. (1) Initiation begins with the correct assembly of the replication proteins at the site where DNA replication is to start. (2) During the elongation stage, DNA is replicated semiconservatively as the complex catalyzes the incorporation of nucleotides into the growing DNA strands. (3) Finally, when replication terminates, the protein machine is disassembled and the daughter molecules separate so that they can segregate into their new cells.

Protein machines that carry out a series of biochemical reactions are not confined to the process of DNA replication but also occur in fatty acid synthesis (Section 16.1), transcription (Chapter 21), and translation (Chapter 22). All four of these processes include initiation, elongation, and termination steps. Furthermore, there is increasing evidence that other processes of cellular metabolism are also carried out by complexes of weakly associated enzymes and other macromolecules.

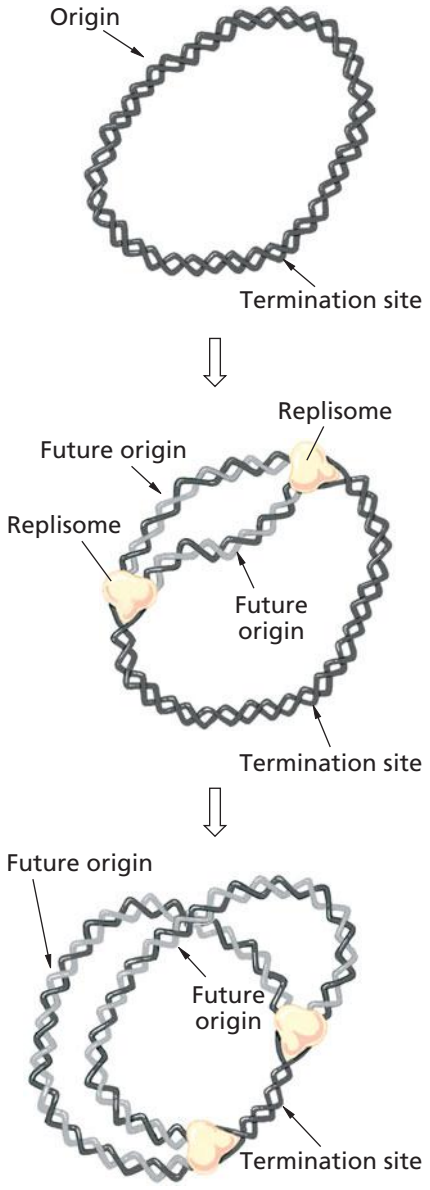
The maintenance of genetic information from generation to generation requires that DNA replication be both rapid (because the entire complement of DNA must be replicated before each cell division) and accurate. All cells have enzymes that correct replication errors and repair damaged DNA. Furthermore, all cells can shuffle pieces of DNA in a process known as genetic recombination. Both repair and recombination use many of the same enzymes and proteins that are required for DNA replication.

The overall strategy of DNA replication, repair, and recombination in prokaryotes and eukaryotes appears to be conserved, although specific enzymes vary among organisms. Just as two different makes of automobile are similar even though individual parts cannot be substituted for one another, so too are the mechanisms of DNA replication, repair, and recombination similar in all organisms, even though the individual enzymes may differ. We are going to focus on the biochemistry of these three processes in *E. coli* because of its many well-characterized enzymes.

## 20.1 DNA Replication Is Bidirectional

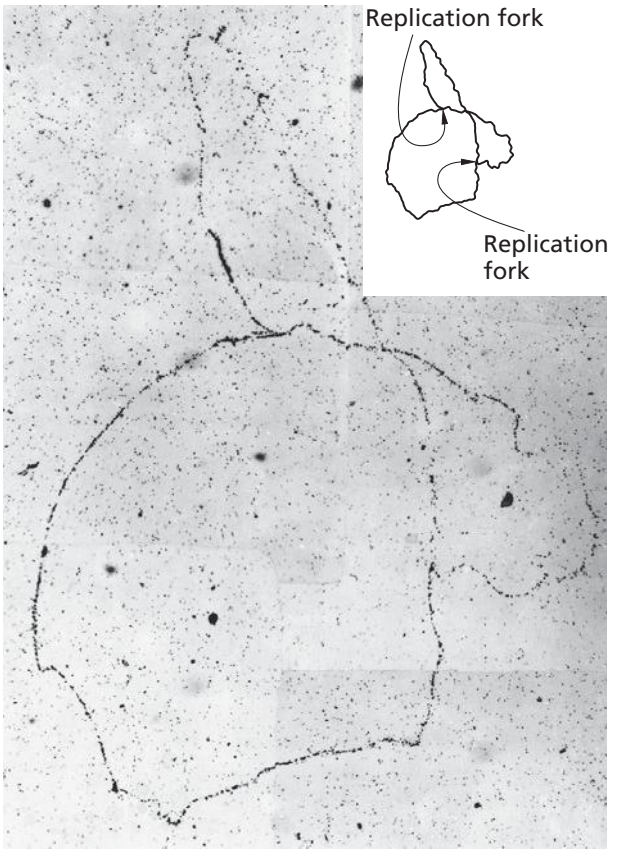
The *E. coli* chromosome is a large, circular, double-stranded DNA molecule of $4.6 \times 10^3$ kilobase pairs (kb). Replication of this chromosome begins at a unique site called the origin of replication and proceeds bidirectionally until the two replication complexes meet at the termination site, where replication stops (Figure 20.2). The protein machine that carries out the polymerization reaction is called a replisome. It contains a number of different proteins that are required for rapid and accurate DNA replication. One replisome is located at each of the two replication forks, the points where the parental DNA is unwound. Figure 20.3 shows an autoradiograph of a replicating *E. coli* chromosome.

As parental DNA is unwound at a replication fork, each strand is used as a template for the synthesis of a new strand. The rate of movement of a replication fork in *E. coli* is approximately 1000 base pairs (bp) per second. In other words, each of the two new strands is extended at the rate of 1000 nucleotides per second. Since there are two replication forks moving at this rate, the entire *E. coli* chromosome can be duplicated in about 38 minutes.




◂ **Figure 20.2**
**Bidirectional DNA replication in *Escherichia coli*.** Semiconservative DNA replication begins at a unique origin and proceeds in both directions. The synthesis of new strands of DNA (light gray) occurs at the two replication forks where the replisomes are located. The two double-stranded DNA molecules separate when the replication forks meet at the termination site. Note that each daughter molecule consists of one parental strand and one newly synthesized strand.

Eukaryotic chromosomes are linear, double-stranded DNA molecules that are usually much larger than the chromosomes of bacteria. The large chromosomes of the fruit fly *Drosophila melanogaster* for example, are about $5.0 \times 10^4$ kb in size or 10 times larger than the *E. coli* chromosome. Replication in eukaryotes is also bidirectional but whereas the *E. coli* chromosome has a unique origin of replication, eukaryotic chromosomes have multiple sites where DNA synthesis is initiated (Figure 20.4). The rate of fork movement in eukaryotes is slower than in bacteria but the presence of many independent origins of replication enables the larger eukaryotic genomes to be copied in approximately the same amount of time as prokaryotic genomes.

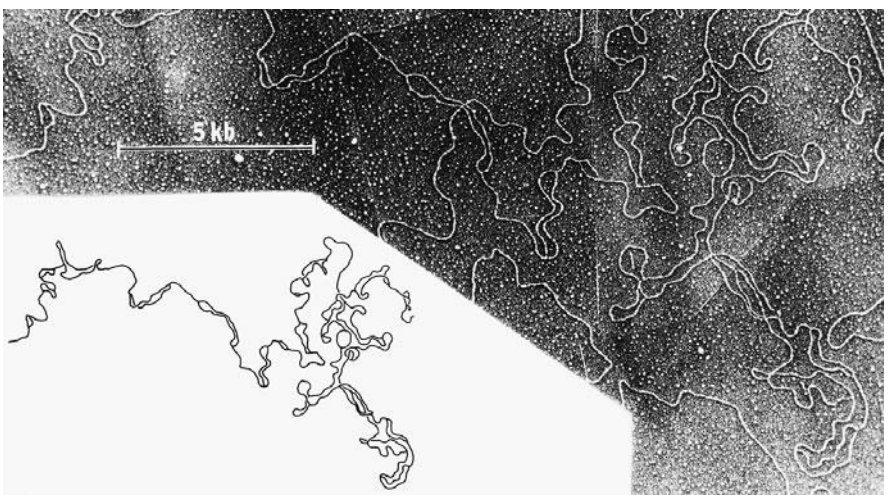



▲ **Figure 20.3**
**Autoradiograph of a replicating *E. coli* chromosome.** The DNA was labeled with $^3$H-deoxythymidine, and the radioactivity detected by overlaying the replicating chromosome with photographic emulsion. The autoradiograph shows that the *E. coli* chromosome has two replication forks.

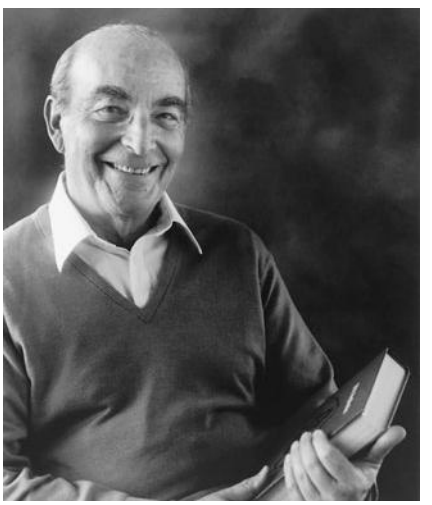
## 20.2 DNA Polymerase

The synthesis of a new strand of DNA is achieved by the successive addition of nucleotides to the end of a growing chain. This polymerization is catalyzed by enzymes known as DNA-directed DNA polymerases, or simply DNA polymerases. *E. coli* cells contain three different DNA polymerases; each protein is identified by a roman numeral according to the order of its discovery. DNA polymerase I repairs DNA and participates in the synthesis of one of the strands of DNA during replication. DNA polymerase II plays a role in DNA repair. DNA polymerase III is the major DNA replication enzyme responsible for chain elongation during DNA replication and is the essential part of the **replisome**.

DNA polymerase III contains ten different polypeptide subunits. It is by far the largest of the three DNA polymerases (Table 20.1). The purified holoenzyme is an asymmetric dimer consisting of two copies of each polypeptide as shown in Figure 20.5. The $\alpha$, $\varepsilon$, and $\theta$ polypeptides combine to form two core complexes that are responsible for the polymerization reactions. The $\beta$ subunits form a sliding clamp that surrounds each of the two DNA strands at the replication fork. Most of the remaining subunits make up the $\gamma$ complex, or "clamp loader" that assists in assembly of the replisome and helps to keep the enzyme bound to parental DNA during successive polymerization reactions.




▲ **Figure 20.4**
**Electron micrograph of replicating DNA from an embryo of the fruit fly *Drosophila melanogaster*.** Note the large number of replication forks at opposite ends of "bubbles" of duplicated DNA.



▲ **Arthur Kornberg (1918–2007).** Kornberg received the Nobel Prize in 1959 for his discovery of DNA polymerase.

## KEY CONCEPT

**Two replication forks move in opposite directions from the origin of replication.**

**Table 20.1 Subunits of DNA polymerase III holoenzyme**

| Subunit | $M_r$ | | Gene | Activity |
|---|---|---|---|---|
| $\alpha$ | 130 000 | core | *polC/dnaE* | Polymerase |
| $\varepsilon$ | 27 000 | core | *dnaQ/mutD* | 3′ → 5′ exonuclease |
| $\theta$ | 8846 | core | *holE* | ? |
| $\beta$ | 40 000 | | *dnaN* | Forms sliding clamp |
| $\tau$ | 71 000 | | *dnaX* | Enhances dimerization of core; ATPase |
| $\gamma$ | 47 000 | $\gamma$ complex | *dnaX* | Enhance processivity; assist in replisome assembly |
| $\delta$ | 38 700 | $\gamma$ complex | *holA* | |
| $\delta'$ | 36 900 | $\gamma$ complex | *holB* | |
| $\chi$ | 16 600 | $\gamma$ complex | *holC* | |
| $\psi$ | 15 174 | $\gamma$ complex | *holD* | |

### A. Chain Elongation Is a Nucleotidyl Group Transfer Reaction

All DNA polymerases, including DNA polymerase III, synthesize DNA by adding one nucleotide at a time to the 3′ end of the growing chain. The nucleotide substrate is a deoxyribonucleoside 5′-triphosphate (dNTP). The specific nucleotide is determined by Watson-Crick base pairing to the template strand; adenine (A) pairs with thymine (T) and guanine (G) pairs with cytosine (C). Since the pool of each dNTP in a cell is approximately equal, this means that on average the enzyme spends three quarters of its time discriminating against incorrect dNTPs that have diffused into the catalytic site where they try to base pair with the template strand.

DNA polymerase III catalyzes the formation of a phosphodiester linkage between the incoming dNTP and the growing chain. The incoming dNTP forms a base pair with a residue of the template strand (Figure 20.6). Once a correct base pair has formed, the free 3′-hydroxyl group of the nascent DNA chain carries out a nucleophilic attack on the $\alpha$-phosphorus atom of the incoming dNTP. This reaction leads to the addition of a nucleoside monophosphate and displacement of pyrophosphate. Subsequent hydrolysis of the pyrophosphate by the abundant enzyme pyrophosphatase makes the polymerization reaction highly favorable in the direction of polymerization. The direction of polymerization (chain growth) is defined as 5′ → 3′, reading across the carbon atoms on the sugar ring of the newly added residue.

**The convention for assigning the direction of DNA strands is described in Section 19.2A.**

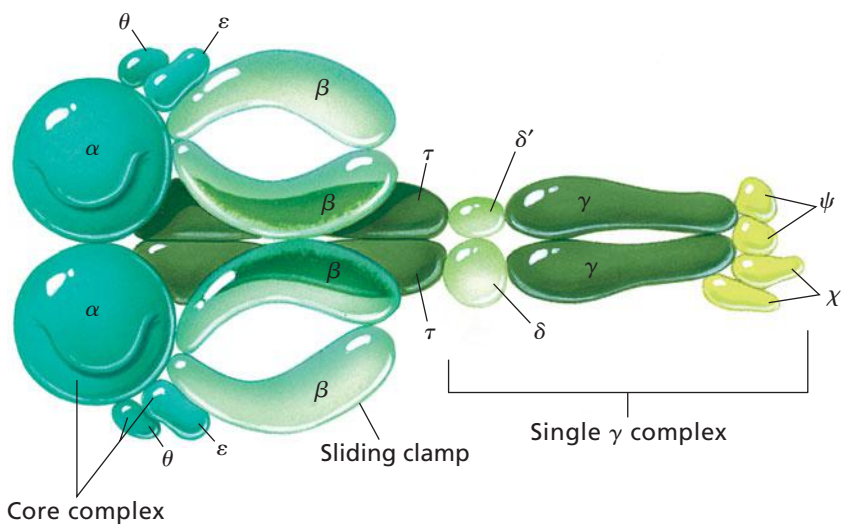
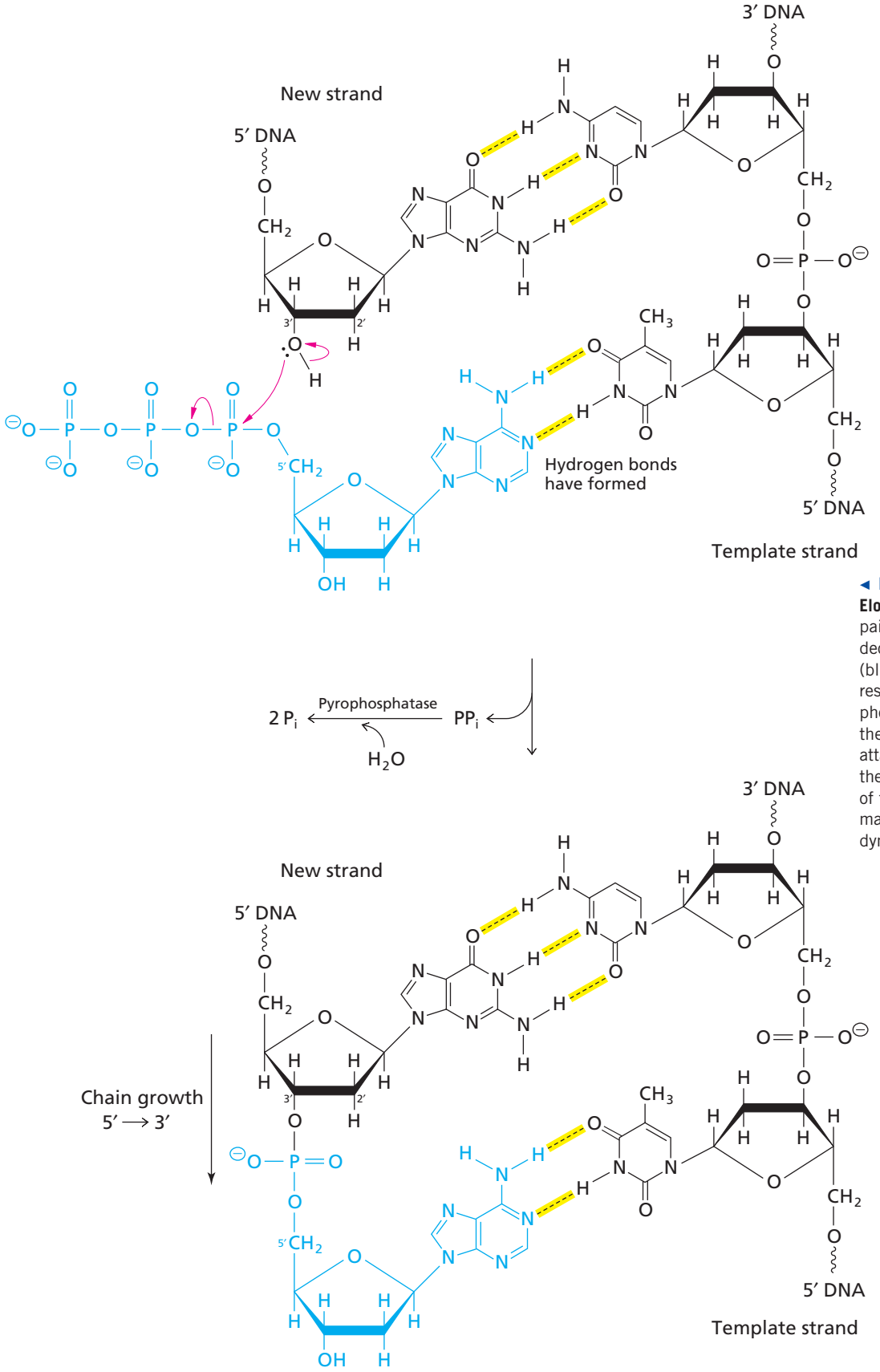



**Figure 20.5 ▸**
**Diagram of the subunit composition of *E. coli* DNA polymerase III.** The holoenzyme consists of two core complexes (containing $\alpha$, $\varepsilon$ and $\theta$), paired copies of $\beta$ and $\tau$, and a single $\gamma$ complex ($\gamma$, $\delta$, $\delta'$, with two copies each of $\psi$, and $\chi$). The structure is thus an asymmetric dimer. Other models of the holoenzyme structure have been proposed. [Adapted from O'Donnell, M. (1992). Accessory protein function in the DNA polymerase III holoenzyme from *E. coli*. *BioEssays* 14:105–111.]




◀ **Figure 20.6**
**Elongation of a DNA chain.** A base pair is created when an incoming deoxynucleoside 5′-triphosphate (blue) forms hydrogen bonds with a residue of the parental strand. A phosphodiester linkage forms when the terminal 3′-hydroxyl group attacks the $\alpha$-phosphorus atom of the incoming nucleotide. Hydrolysis of the released pyrophosphate makes the overall reaction thermodynamically favorable.

DNA polymerase III advances by one residue, after each addition reaction, and another nucleotidyl group transfer reaction occurs. This mechanism ensures that the new chain is extended by the stepwise addition of single nucleotides that are properly aligned by base pairing with the template strand. As expected, DNA polymerase III cannot synthesize DNA in the absence of a template, nor can it add nucleotides in the absence of a 3′ end of a preexisting chain. In other words, DNA polymerase III requires both a template and a primer as substrates for synthesis to occur.

As noted earlier, DNA replication inside the cell proceeds at a rate of approximately 1000 nucleotides per second. This is the fastest known rate of any *in vivo* polymerization reaction. The rate of polymerization catalyzed by purified DNA polymerase III *in vitro*, however, is much slower, indicating that the isolated enzyme lacks some components necessary for full activity. Only when the complete replisome is assembled does polymerization *in vitro* occur at approximately the rate found inside the cell.

## B. DNA Polymerase III Remains Bound to the Replication Fork

Once DNA synthesis has been initiated, the polymerase remains bound at the replication fork until replication is complete. The 3′ end of the growing chain remains associated with the active site of the enzyme while many nucleotides are added sequentially. As part of the replisome, the DNA polymerase III holoenzyme is highly *processive* (see Section 12.5A). This means that only a small number of DNA polymerase III molecules are needed to replicate the entire chromosome. Processivity also accounts for the rapid rate of DNA replication.

The processivity of the DNA polymerase III holoenzyme is due, in part, to the $\beta$ subunits of the enzyme. These subunits have no activity on their own but when assembled into the holoenzyme they form a ring that can completely surround the DNA molecule. The ring is formed by two $\beta$ subunits that form a head-to-tail dimer. Each of the subunits contains three similar domains consisting of a $\beta$ sandwich fold with two $\alpha$ helices at the interior edge that interact with DNA (Figure 20.7). The $\beta$ subunits thus act as a sliding clamp locking the polymerase onto the DNA substrate. Incorporating DNA polymerase III into an even larger protein machine at the replication fork further ensures that the enzyme remains associated with the nascent DNA chains during polymerization. Many other biochemically characterized DNA replication systems have evolved the same strategy to make DNA replication faster (more efficient). For example, two related bacteriophage, $T_4$ and RB69, both encode a replication accessory protein, gp45, that forms a circular clamp (Figure 20.7). This clamp structure locks the phage-encoded DNA polymerases onto their DNA substrates and enhances processivity. Figure 20.8 shows a model for how this is likely to work *in vivo* for bacteriophage DNA polymerase bound to DNA. The sliding clamp surrounds the double-stranded region of DNA and interacts with the subunits containing the polymerase activity that bind to the single-stranded region of the replication fork. Eukaryotic DNA polymerases use the same strategy to clamp onto their substrates (see Section 20.6).

**The elongation reaction in fatty acid synthesis is another example of a processive polymerization reaction catalyzed by a large complex (Section 16.1C). The glycogen synthase reaction is an example of a distributive polymerization reaction (Section 12.5A).**

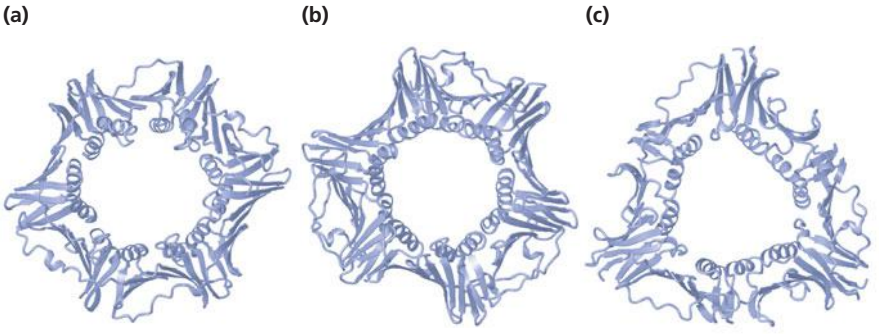
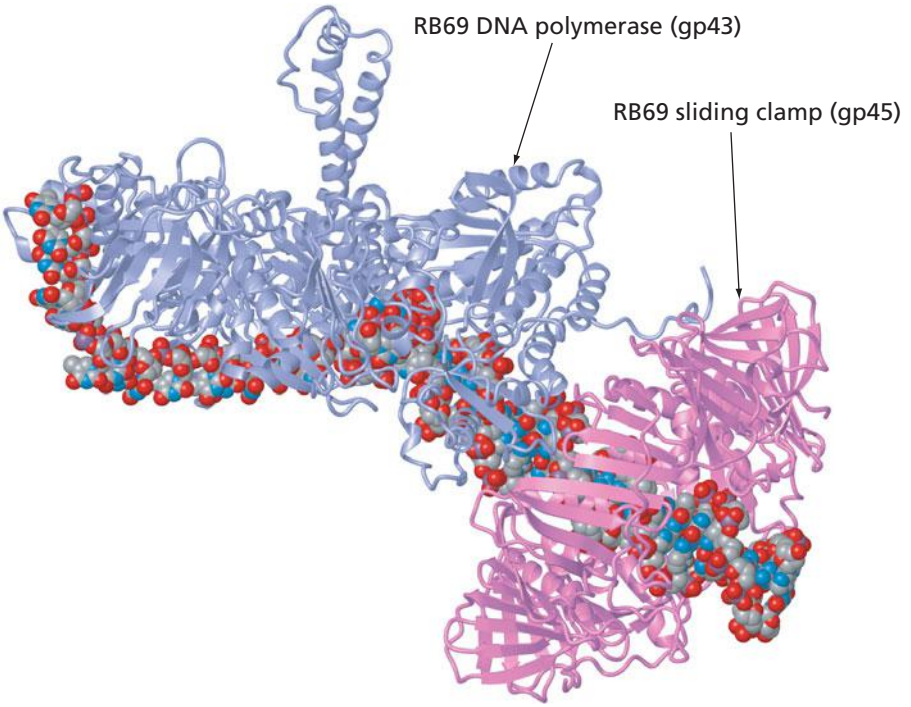



**Figure 20.7 ►**
**DNA polymerases can use sliding ring clamps to increase processivity.** These three crystal structures show the convergent evolution of structure and function; **(a)** the $\beta$ subunit of *E. coli* DNA polymerase III [PDB 1MMI]; **(b)** Proliferating Cell Nuclear Antigen (PCNA) that performs the same function in archaebacteria; [PDB 3LX1] **(c)** gp45 from bacteriophage T4 is also a sliding ring that clamps DNA polymerase to its DNA substrate. [PDB 1CZD]




◀ **Figure 20.8**
**Model of bacteriophage DNA polymerase bound to DNA.** The sliding clamp (purple) surrounds the newly synthesized double-stranded DNA. The subunit containing the active site is shown in blue. The 3′ end of the nascent strand is positioned at the active site and the single-stranded region of the template strand extends leftward. The DNA polymerase will move from right to left as the nascent strand is extended. [PDB 1WAI].

### C. Proofreading Corrects Polymerization Errors

The DNA polymerase III holoenzyme also possesses a 3′ → 5′ exonuclease activity. This exonuclease, whose active site lies primarily within the ε subunit, can catalyze hydrolysis of the phosphodiester linkage that joins the 3′-terminal residue to the rest of the growing polynucleotide chain. Thus, the DNA polymerase III holoenzyme can catalyze both chain elongation and degradation. The exonuclease activity allows the holoenzyme to proofread, or edit, newly synthesized DNA in order to correct any mismatched base pairs. When DNA polymerase III recognizes a distortion in the DNA produced by an incorrectly paired base, the exonuclease activity of the enzyme catalyzes removal of the mispaired nucleotide before polymerization continues.

**Proofreading is possible because the polymerization mechanism is head growth not tail growth (Box 12.3).**

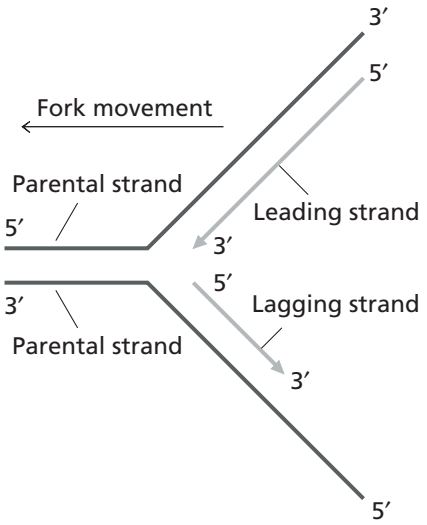
An incorrect base is incorporated about once every $10^5$ polymerization steps for an error rate of about $10^{-5}$. The 3′ → 5′ proofreading exonuclease activity will remove 99% of these incorrect nucleotides. It has an error rate of $10^{-2}$. The combination of these two sequential reactions yields an error rate for polymerization of $10^{-7}$. This is one of the lowest error rates of any enzyme. Most of these replication errors are subsequently repaired by separate DNA repair enzymes (Section 20.7) yielding an overall error rate for DNA replication of between $10^{-9}$ and $10^{-10}$. Despite this impressive accuracy, replication errors are common when large genomes are duplicated. (Recall that the human genome contains $3.2 \times 10^9$ bp, which means that, on average, each time the genome is replicated an error gets transmitted to one of the two daughter cells.) Mistakes that occur during DNA replication are the most common source of mutation. What this means is that most of evolution is due to the inaccuracy of DNA replication!

**KEY CONCEPT**

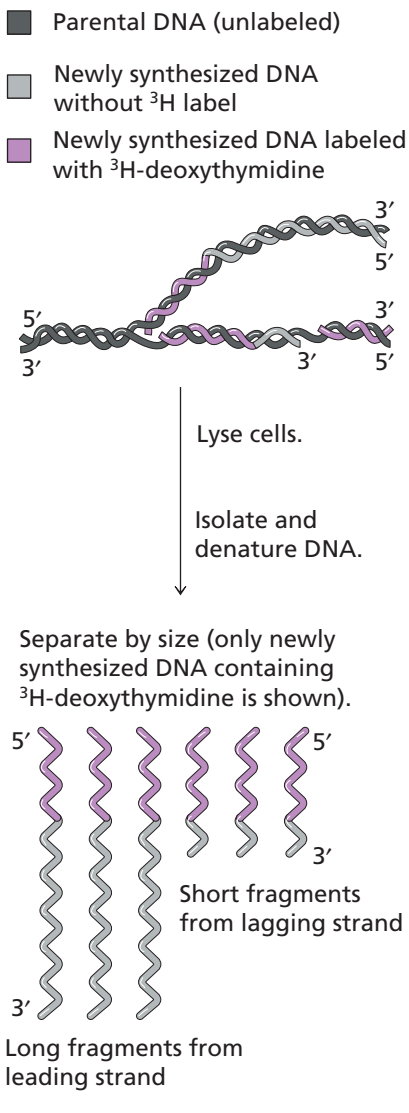
**The accuracy of DNA polymerase combined with proofreading and DNA repair makes DNA replication the most accurate biochemical reaction known.**

## 20.3 DNA Polymerase Synthesizes Two Strands Simultaneously

DNA polymerases catalyze chain elongation exclusively in the 5′ → 3′ direction, as shown in Figure 20.6. Since the two strands of DNA are antiparallel, 5′ → 3′ synthesis using one template strand occurs in the same direction as fork movement but 5′ → 3′ synthesis using the other template strand occurs in the direction opposite fork movement (Figure 20.9). The new strand formed by polymerization in the same direction as




▲ **Figure 20.9**
**Diagram of the replication fork.** The two newly synthesized strands have opposite polarity. On the leading strand, 5′ → 3′ synthesis moves in the same direction as the replication fork; on the lagging strand, 5′ → 3′ synthesis moves in the opposite direction.




fork movement is called the leading strand. The new strand formed by polymerization in the opposite direction is called the lagging strand. Recall that the DNA polymerase III holoenzyme dimer contains two core complexes that can catalyze polymerization. One of these is responsible for synthesis of the leading strand and the other is responsible for synthesis of the lagging strand.

## A. Lagging Strand Synthesis Is Discontinuous

The leading strand is synthesized as one continuous polynucleotide beginning at the origin and ending at the termination site. In contrast, the lagging strand is synthesized discontinuously in short pieces in the direction opposite fork movement. These pieces of lagging strand are then joined by a separate reaction. In Section 20.4, we present a model of the replication fork that explains how one enzyme complex can synthesize both strands simultaneously.

An experiment that illustrates discontinuous DNA synthesis is shown in Figure 20.10. *E. coli* DNA is labeled with a short pulse of $^{3}$H-deoxythymidine. The newly made DNA molecules are then isolated, denatured, and separated by size. The experiment detects two types of labeled DNA molecules: very large DNA molecules that collectively contain about half the radioactivity of the partially replicated DNA and shorter DNA fragments of about 1000 residues that collectively contain the other half of the radioactivity. The large DNA molecules arise from continuous synthesis of the leading strand while the shorter fragments arise from discontinuous synthesis of the lagging strand. The short pieces of lagging strand DNA are named **Okazaki fragments** in honor of their discoverer, Reiji Okazaki. The overall mechanism of DNA replication is called semidiscontinuous to emphasize the different mechanisms for replicating each strand.

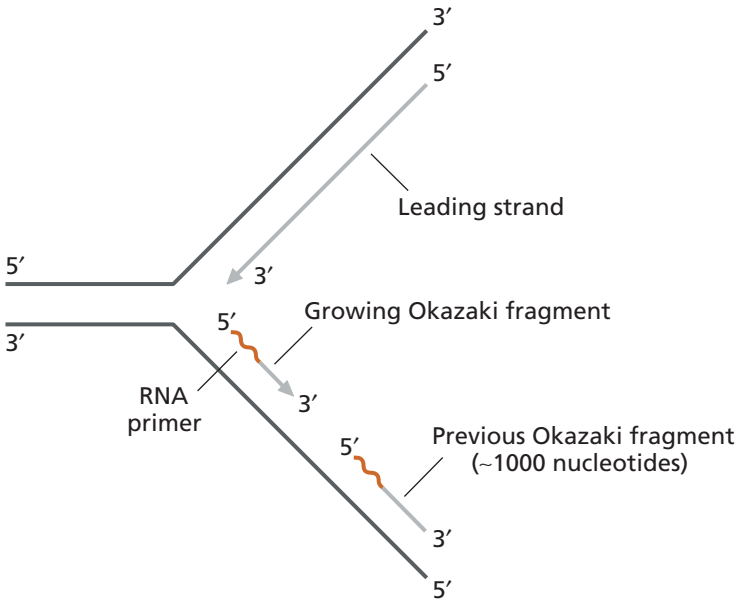
## B. Each Okazaki Fragment Begins with an RNA Primer

It was clear that lagging strand synthesis is discontinuous but it was not obvious how synthesis of each Okazaki fragment is initiated. The problem is that no DNA polymerase can begin polymerization *de novo*; they can only add nucleotides to existing polymers. This limitation presents little difficulty for leading strand synthesis since once DNA synthesis is under way nucleotides are continuously added to a growing chain. However, on the lagging strand the synthesis of each Okazaki fragment requires a new initiation event. This is accomplished by making short pieces of RNA at the replication fork. These RNA primers are complementary to the lagging strand template. Each primer is extended from its 3′ end by DNA polymerase to form an Okazaki fragment as shown in Figure 20.11. (Synthesis of the leading strand also begins with an RNA primer but only one primer is required to initiate synthesis of the entire strand.)

The use of short RNA primers gets around the limitation imposed by the mechanism of DNA polymerase—namely, that it cannot initiate DNA synthesis *de novo*. The primers are synthesized by a DNA-dependent RNA polymerase enzyme called **primase**—the product of the *dnaG* gene in *E. coli*. The three-dimensional crystal structure of the DnaG catalytic domain revealed that its folding and active site are distinct from the well studied polymerases suggesting that it may employ a novel enzyme mechanism. Primase is part of a larger complex called the **primosome** that contains many other polypeptides in addition to primase. The primosome, along with DNA polymerase III, is part of the replisome.

As the replication fork progresses, the parental DNA is unwound and single-stranded DNA becomes exposed. Primase catalyzes the synthesis of a short RNA primer about once every second using this single-stranded DNA as a template. The primers are only a few nucleotides in length. Since the replication fork advances at a rate of about

◀ **Figure 20.10**
**Discontinuous DNA synthesis demonstrated by analysis of newly synthesized DNA.** Nascent DNA molecules are labeled in *E. coli* with a short pulse of $^{3}$H-deoxythymidine. The cells are lysed, the DNA is isolated, and single strands are separated by size. The labeled DNA molecules fall into two classes: long molecules arising from continuous synthesis of the leading strand and short fragments arising from discontinuous synthesis of the lagging strand.




◂ **Figure 20.11**
**Diagram of lagging strand synthesis.** A short piece of RNA (brown) serves as a primer for the synthesis of each Okazaki fragment. The length of the Okazaki fragment is determined by the distance between successive RNA primers.

1000 nucleotides per second, one primer is synthesized for approximately every 1000 nucleotides that are incorporated. DNA polymerase III catalyzes synthesis of DNA in the 5′ → 3′ direction by extending each short RNA primer.

## C. Okazaki Fragments Are Joined by the Action of DNA Polymerase I and DNA Ligase

Okazaki fragments are eventually joined to produce a continuous strand of DNA. The reaction proceeds in three steps: removal of the RNA primer, synthesis of replacement DNA, and sealing of the adjacent DNA fragments. The steps are carried out by the combined action of DNA polymerase I and DNA ligase.

DNA polymerase I of *E. coli* was the enzyme discovered by Arthur Kornberg. It was the first enzyme to be found that could catalyze DNA synthesis using a template strand. In a single polypeptide, DNA polymerase I contains the two activities found in the DNA polymerase III holoenzyme: 5′ → 3′ polymerase activity and 3′ → 5′ proofreading exonuclease activity. In addition, DNA polymerase I has 5′ → 3′ exonuclease activity, an activity not found in DNA polymerase III.

DNA polymerase I can be cleaved with certain proteolytic enzymes to generate a small fragment that contains the 5′ → 3′ exonuclease activity and a larger fragment that retains the polymerization and proofreading activities. The larger fragment consists of the C-terminal 605 amino acid residues, and the smaller fragment contains the remaining N-terminal 323 residues. The large fragment, known as the Klenow fragment, was widely used for DNA sequencing and is still used in many other techniques that require DNA synthesis without 5′ → 3′ degradation. In addition, many studies of the mechanisms of DNA synthesis and proofreading use the Klenow fragment as a model for more complicated DNA polymerases.

Figure 20.12 shows the structure of the Klenow fragment complexed with a fragment of DNA containing a mismatched terminal base pair. The 3′ end of the nascent strand is positioned at the 3′ → 5′ exonuclease site of the enzyme. During polymerization, the template strand occupies the groove at the top of the structure and at least 10 bp of double-stranded DNA are bound by the enzyme as shown in the figure. Many of the amino acid residues involved in binding DNA are similar in all DNA polymerases

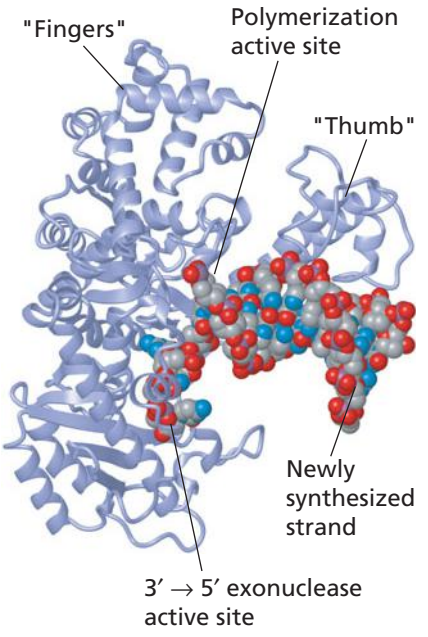
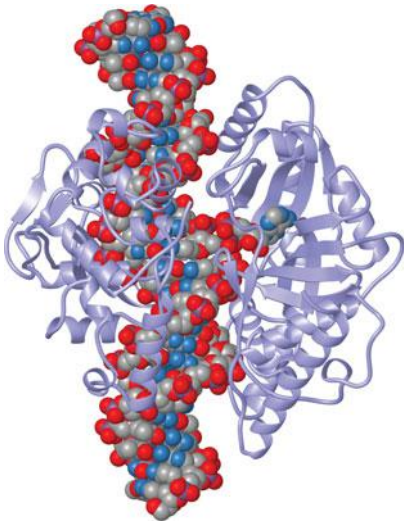
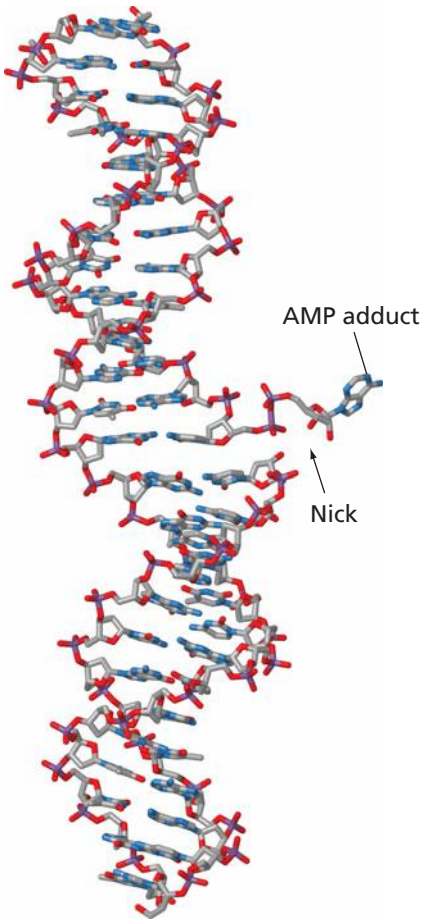



**Figure 20.12** ▸
**Structure of the Klenow fragment with a bound DNA fragment.** The enzyme wraps around the DNA. The 3′ end of the nascent strand is positioned at the 3′ → 5′ exonuclease site (lower left). During DNA synthesis *in vivo* the template strand extends beyond the double-stranded region shown in the crystal structure. [PDB 1KLN].



▲ *E. coli* DNA ligase bound to nicked DNA. [PDB 20W0]




▲ Structure of nicked DNA substrate when bound by DNA ligase [PDB 20W0].

although the enzymes may be otherwise quite different in three-dimensional structure and amino acid sequence.

The unique 5′ → 3′ exonuclease activity of DNA polymerase I removes the RNA primer at the beginning of each Okazaki fragment. (Since it is not part of the Klenow fragment, the 5′ → 3′ exonuclease is not shown in Figure 20.12, but it would be located at the top of the structure next to the groove that accommodates the template strand.) As the primer is removed, the polymerase synthesizes DNA to fill in the region between Okazaki fragments, a process called **nick translation** (Figure 20.13). In nick translation, DNA polymerase I recognizes and binds to the nick between the 3′ end of an Okazaki fragment and the 5′ end of the next primer. The 5′ → 3′ exonuclease then catalyzes hydrolytic removal of the first RNA nucleotide while the 5′ → 3′ polymerase adds a deoxynucleotide to the 3′ end of the DNA chain. In this way, the enzyme moves the nick along the lagging strand. DNA polymerase I dissociates from the DNA after completing 10 or 12 cycles of hydrolysis and polymerization, leaving behind two Okazaki fragments that are separated by a nick in the phosphodiester backbone. The removal of RNA primers by DNA polymerase I is an essential part of DNA replication because the final product must consist entirely of double-stranded DNA.

The last step in the synthesis of the lagging strand of DNA is the formation of a phosphodiester linkage between the 3′-hydroxyl group at the end of one Okazaki fragment and the 5′-phosphate group of an adjacent Okazaki fragment. This step is catalyzed by DNA ligase. The DNA ligases in eukaryotic cells and in bacteriophage-infected cells require ATP as a cosubstrate. In contrast, *E. coli* DNA ligase uses $NAD^{\oplus}$ as a cosubstrate. $NAD^{\oplus}$ is the source of the nucleotidyl group that is transferred, first to the enzyme and then to the DNA, to create an ADP-DNA intermediate. The proposed mechanism of DNA ligase in *E. coli* is shown in Figure 20.14. The net reaction is

$$\text{DNA (nicked)} + NAD^{\oplus} \longrightarrow \text{DNA(sealed)} + NMN^{\oplus} + \text{AMP} \qquad \textbf{(20.1)}$$

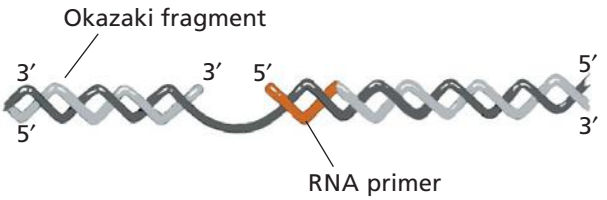
## 20.4 Model of the Replisome

The replisome contains a primosome, the DNA polymerase III holoenzyme, and additional proteins that are required for DNA replication. The assembly of many proteins into a single machine allows coordinated synthesis of the leading and lagging strands at the replication fork.

The template for DNA polymerase III is single-stranded DNA. This means that the two strands of the parental double helix must be unwound and separated during replication. This unwinding is accomplished primarily by a class of proteins called helicases. The helicase DnaB is required for DNA replication in *E. coli*. DnaB is one of the subunits of the primosome that, in turn, is part of the larger replisome. The rate of DNA unwinding is directly coupled to the rate of polymerization as the replisome moves along the chromosome. Unwinding is assisted by the actions of various topoisomerases (Section 19.3) that relieve supercoiling ahead of and behind the replication fork. These enzymes are not part of the replisome but they are required for replication. The most important topoisomerase in *E. coli* is topoisomerase II, or gyrase. Mutants lacking this enzyme cannot replicate their DNA. The end result is the production of two daughter molecules each containing one newly synthesized stand and one parental strand as shown in Figure 20.1. At no time during DNA replication is there a significant stretch of single-stranded DNA other than that found on the lagging strand template.

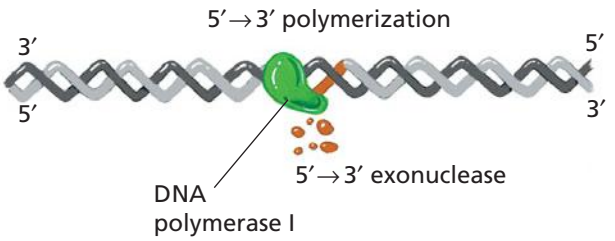
Another protein that is part of the replisome is single-strand binding protein (SSB), also known as helix-destabilizing protein. SSB binds to single-stranded DNA and prevents it from folding back on itself to form double-stranded regions. SSB is a tetramer of four identical small subunits. Each tetramer covers about 32 nucleotides of DNA. Binding of SSB to DNA is cooperative; that is, binding of the first tetramer facilitates binding of the second, and so on. The presence of several adjacent SSB molecules on single-stranded DNA produces an extended, relatively inflexible, DNA conformation. Single-stranded DNA coated with SSB is an ideal template for synthesis of the complementary strand during DNA replication because it is free of secondary structure.

(a) Completion of Okazaki fragment synthesis leaves a nick between the Okazaki fragment and the preceding RNA primer on the lagging strand.

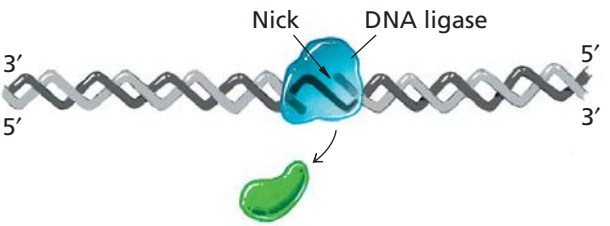



◂ **Figure 20.13**
**Joining of Okazaki fragments by the combined action of DNA polymerase I and DNA ligase.**

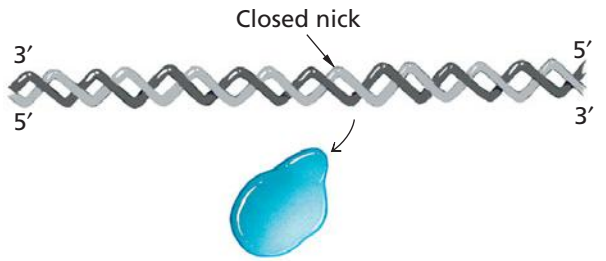
(b) DNA polymerase I extends the Okazaki fragment while its 5′→3′ exonuclease activity removes the RNA primer. This process, called nick translation, results in movement of the nick along the lagging strand.

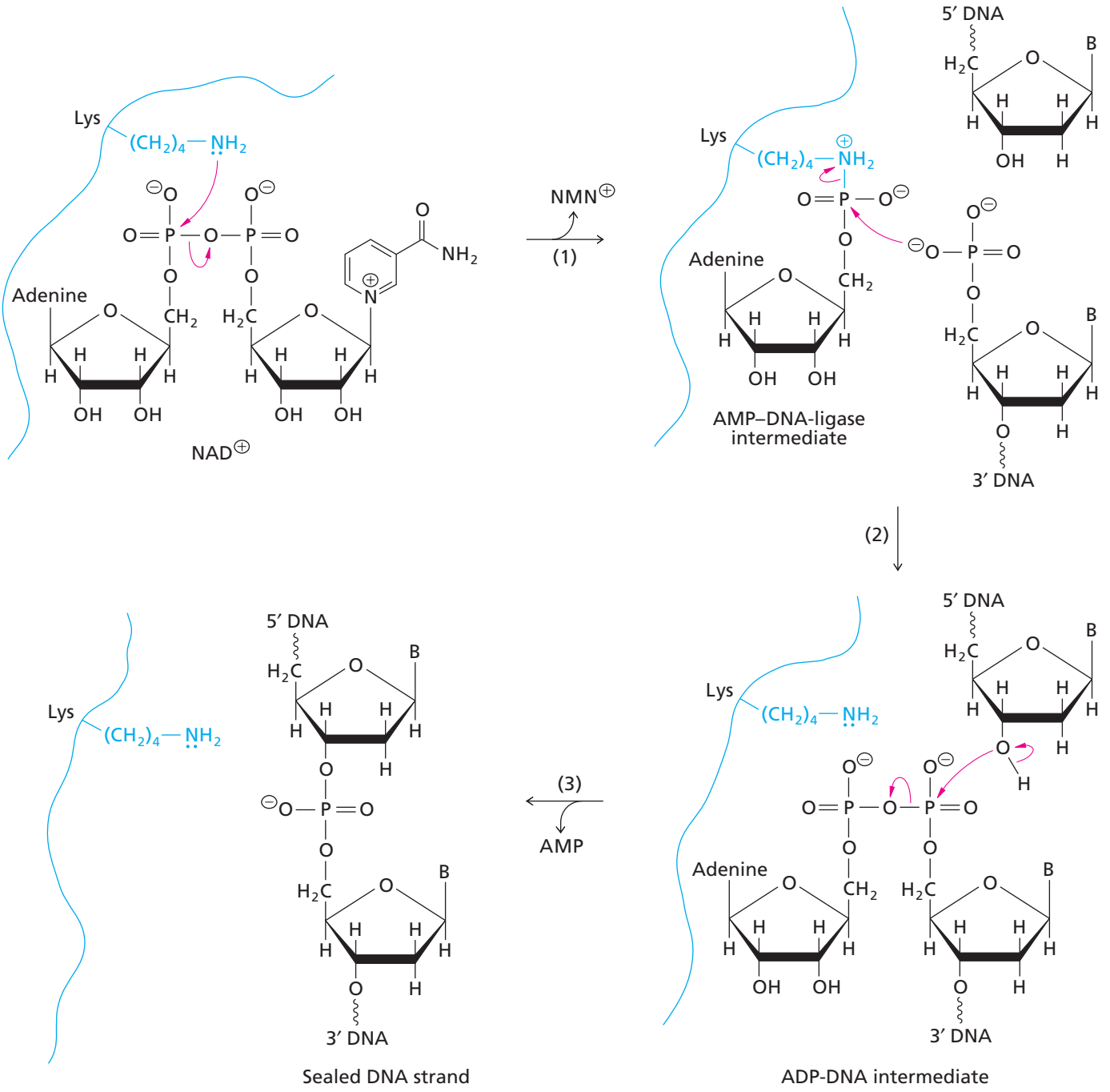



(c) DNA polymerase I dissociates after extending the Okazaki fragment 10–12 nucleotides. DNA ligase binds to the nick.



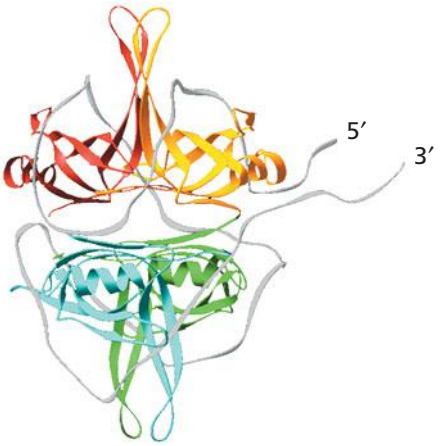

(d) DNA ligase catalyzes formation of a phosphodiester linkage, which seals the nick, creating a continuous lagging strand. The enzyme then dissociates from the DNA.



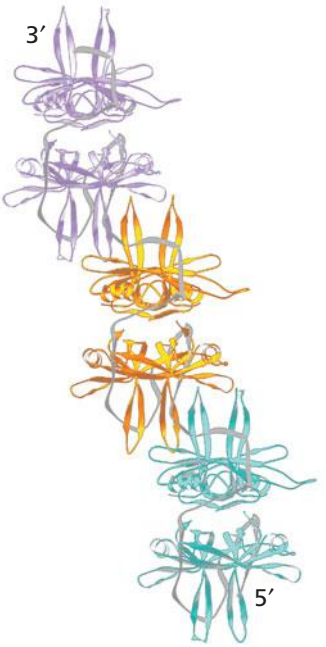



▲ **Figure 20.14**
**Proposed mechanism of DNA ligase in *E. coli*.** Using NAD⊕ as a cosubstrate, *E. coli* DNA ligase catalyzes the formation of a phosphodiester linkage at a nick in DNA. In Step 1, the ε-amino group of a lysine residue of DNA ligase attacks the phosphorus atom bonded to the 5′-oxygen atom of the adenosine moiety of NAD⊕. Nicotinamide mononucleotide (NMN⊕) is displaced, generating an AMP–DNA-ligase intermediate. (With DNA ligases that use ATP as the cosubstrate, pyrophosphate is displaced.) In Step 2, an oxygen atom of the free 5′-phosphate group of the DNA attacks the phosphate group of the AMP-enzyme complex, forming an ADP-DNA intermediate. In Step 3, the nucleophilic 3′-hydroxyl group on the terminal residue of the adjacent DNA strand attacks the activated 5′-phosphate group of ADP-DNA, releasing AMP and generating a phosphodiester linkage that seals the nick in the DNA strand. B represents any base.

A model of DNA synthesis by the replisome is shown in Figure 20.15. The primosome containing the primase and helicase is located at the head of the replication fork, followed by a DNA polymerase III holoenzyme. (In order to simplify the figure, only the core complexes of DNA polymerase III are shown.) Primase synthesizes an RNA primer approximately once every second as the helicase unwinds the DNA. One of the two core complexes in the holoenzyme dimer synthesizes the leading strand continuously



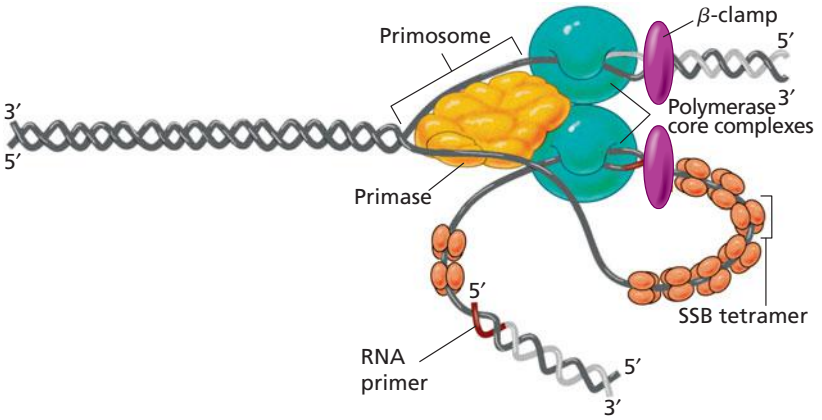

▲ Model for *E. coli* SSB tetramer bound to ssDNA [PDB 1EYG]



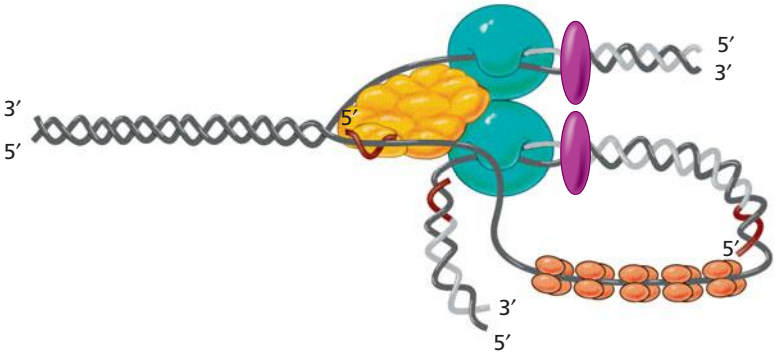

◄ DNA bound to SSB Model for the extended conformation of three SSB tetramers bound cooperatively to ssDNA. [PDB 1EYG]

Source: Nature Structural and Molecular Biology 7:648–652 (2000) Raghunathan et al.

(a) The lagging-strand template loops back through the replisome so that the leading and lagging strands are synthesized in the same direction. SSB binds to single-stranded DNA.



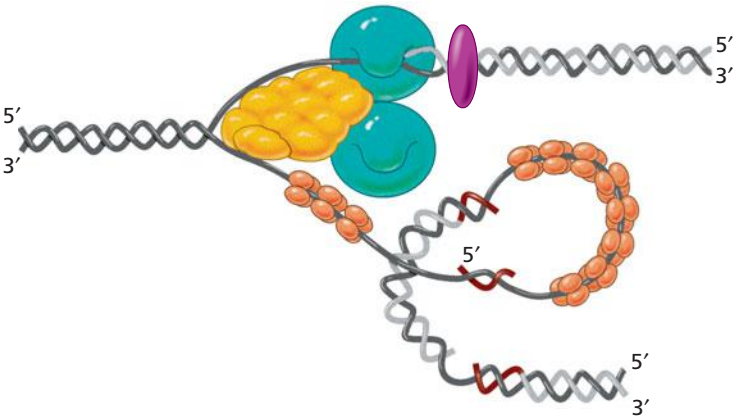

(b) As helicase unwinds the DNA template, primase synthesizes an RNA primer. The lagging-strand polymerase completes an Okazaki fragment.



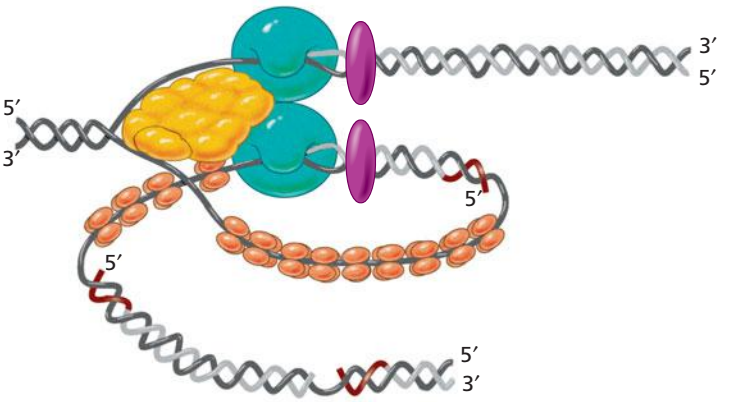

◄ **Figure 20.15**
**Simultaneous synthesis of leading and lagging strands at a replication fork.** The replisome contains the DNA polymerase III holoenzyme (only the core complexes are shown); a primosome containing primase, a helicase, and other subunits; and additional components including single-strand binding protein (SSB). One core complex of the holoenzyme synthesizes the leading strand while the other core complex synthesizes the lagging strand. The lagging-strand template is looped back through the replisome so that the leading and lagging strands can be synthesized in the same direction as fork movement. (c) and (d) continue on the next page.

**Figure 20.15 (Continued) ►**

(c) When the lagging-strand polymerase encounters the preceding Okazaki fragment, it releases the lagging strand.




(d) The lagging-strand polymerase binds to a newly synthesized primer and begins synthesizing another Okazaki fragment.




in the 5′ ⟶ 3′ direction while the other extends the RNA primers to form Okazaki fragments. The lagging-strand template is thought to fold back into a large loop. This configuration allows both the leading and lagging strands to be synthesized in the same direction as fork movement.

The two core complexes of the DNA polymerase III holoenzyme are drawn in the model as equivalent but their roles in DNA replication are not equivalent. One of them remains firmly bound to the leading-strand template whereas the other binds the lagging-strand template until it encounters the RNA primer of the previously synthesized Okazaki fragment. At this point the core complex releases the lagging-strand template. The lagging-strand template reassociates with the holoenzyme at the site of the next primer and synthesis continues (Figure 20.15d). The entire holoenzyme is extremely processive since half of it remains associated with the leading strand from the beginning of replication until termination while the other half processively synthesizes stretches of 1000 nucleotides in the lagging strand. The $\gamma$ complex of the holoenzyme aids in binding and releasing the lagging-strand template by participating in the removal and reassembly of the sliding clamp formed by the $\beta$ subunits.

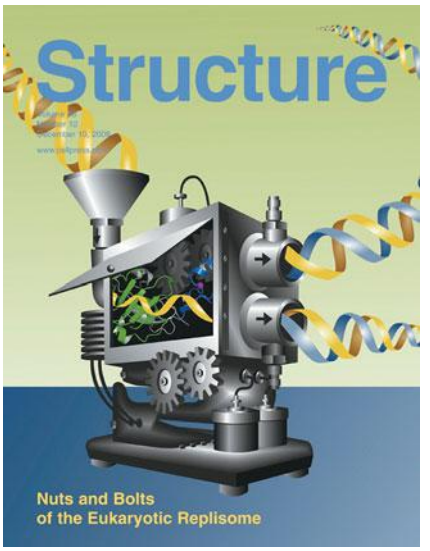
The replisome model explains how synthesis of the leading and lagging strands is coordinated. The structure of the replisome also ensures that all the components necessary for

replication are available at the right time, in the right amount, and in the right place. Complexes of proteins that function together to carry out a biochemical task are frequently called protein machines. The replisome is an example of a protein machine, as are the bacterial flagellum (Chapter 4), the ATP synthase complex (Chapter 14), the photosynthetic reaction center (Chapter 15), and several others that are discussed in the following chapters.

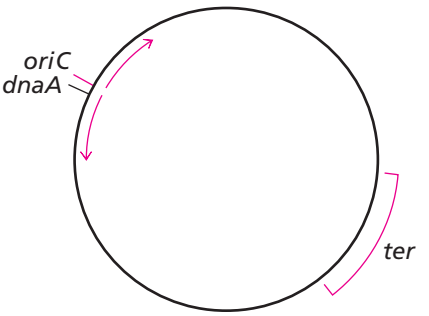
## 20.5 Initiation and Termination of DNA Replication

As noted earlier, DNA replication begins at a specific DNA sequence called the origin. In *E. coli*, this site is called *oriC*, and it is located at about 10 o'clock on the genetic map of the chromosome (Figure 20.16). The initial assembly of replisomes at *oriC* depends on proteins that bind to this site causing local unwinding of the DNA. One of these proteins, DnaA, is encoded by the *dnaA* gene that is located very close to the origin. DnaA helps regulate DNA replication by controlling the frequency of initiation. The initial RNA primers required for leading-strand synthesis are probably made by the primosomes at the origin.

Termination of replication in *E. coli* occurs at the termination site (*ter*), a region opposite the origin on the circular chromosome. This region contains DNA sequences that are binding sites for a protein known as terminator utilization substance (Tus). The structure of Tus bound to a single termination site is shown in Figure 20.17. Regions of $\beta$ strand lie in the major groove of DNA where the amino acid side chains make contact with the base pairs and recognize the *ter* sequence. Tus prevents replication forks from passing through the region by inhibiting the helicase activity of the replisome. The termination site also contains DNA sequences that play a role in the separation of daughter chromosomes when DNA replication is completed.




▲ **Protein machines.** Sometimes the machine metaphor can be taken too literally as in this humorous cover from the Journal *Structure*.




▲ **Figure 20.16**
**Location of the origin (*ori*C) and terminus (*ter*) of DNA replication in *E. coli*.** *dnaA* is the gene for the protein DnaA, which is required to initiate replication. The distance between *oriC* and *dnaA* is about 40 kb. The red arrows indicate the direction of movement of the replication forks.

## 20.6 DNA Replication Technologies

Our understanding of the basic principles of DNA replication has led to the development of some amazing technologies that Watson and Crick could never have anticipated in 1953. We have already encountered site-directed matagenesis (Box 6.1). In this section we explore amplification and sequencing technologies that have transformed biochemistry and, indeed, all biology. These technologies have produced genome sequences of extinct species (e.g., *Homo neanderthalensis*) and to the discovery of the genetic basis of many human traits and diseases.

### A. The Polymerase Chain Reaction Uses DNA Polymerase to Amplify Selected DNA Sequences

The **polymerase chain reaction** (PCR) is a valuable tool for amplifying a small amount of DNA or increasing the proportion of a particular DNA sequence in a population of mixed DNA molecules. The use of PCR technology avoids the need to take large samples of tissue in order to obtain enough DNA to manipulate for sequencing or cloning. The polymerase chain reaction also enables the production of a large number of copies of a gene that has not been isolated but whose sequence is known. It thus can serve as an alternative to cloning for gene amplification.

The PCR technique is illustrated in the figure on page 621. Sequence information from both sides of the desired locus is used to construct oligonucleotide primers that flank the DNA sequence to be amplified. The oligonucleotide primers are complementary to opposite strands and their 3′ ends are oriented toward each other. The DNA from the source (usually representing the entire DNA in a cell) is denatured by heating in the presence of excess oligonucleotides. On cooling, the primers preferentially anneal to their complementary sites, which border the DNA sequence of interest. The primers are then extended using a heat-stable DNA polymerase, such as *Taq* polymerase from the thermophilic bacterium *Thermus aquaticus*. After one cycle of synthesis, the reaction mixture

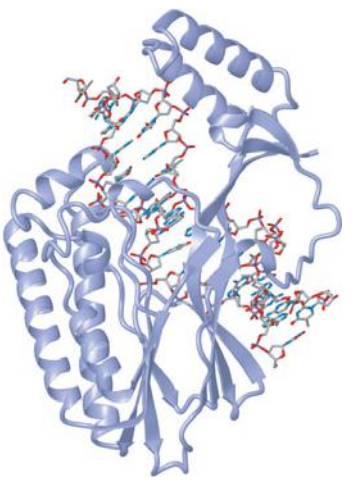


**Figure 20.17 ►**
**Structure of *E. coli* Tus bound to DNA.** Tus binds to specific sequences at the termination site of DNA replication. The bound protein blocks movement of the replisome. [PDB 1ECR].

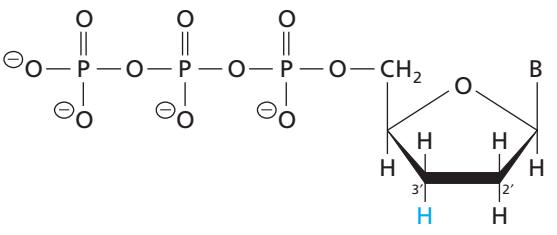
is again heated to dissociate the DNA strands and cooled to reanneal the DNA with the oligonucleotides. The primers are then extended again. In this second cycle, two of the newly synthesized, single-stranded chains are precisely the length of the DNA between the 5′ ends of the primers. The cycle is repeated many times, with reaction time and temperature carefully controlled. With each cycle, the number of DNA strands whose 5′ and 3′ ends are defined by the ends of the primers increases exponentially, whereas the number of DNA strands including sequences outside the region bordered by the primers increases arithmetically. As a result, the desired DNA is preferentially replicated until, after 20 to 30 cycles, it makes up most of the DNA in the test tube. The target DNA sequence can then be cloned, sequenced, or used as a probe for screening a recombinant DNA library.

## B. Sequencing DNA Using Dideoxynucleotides

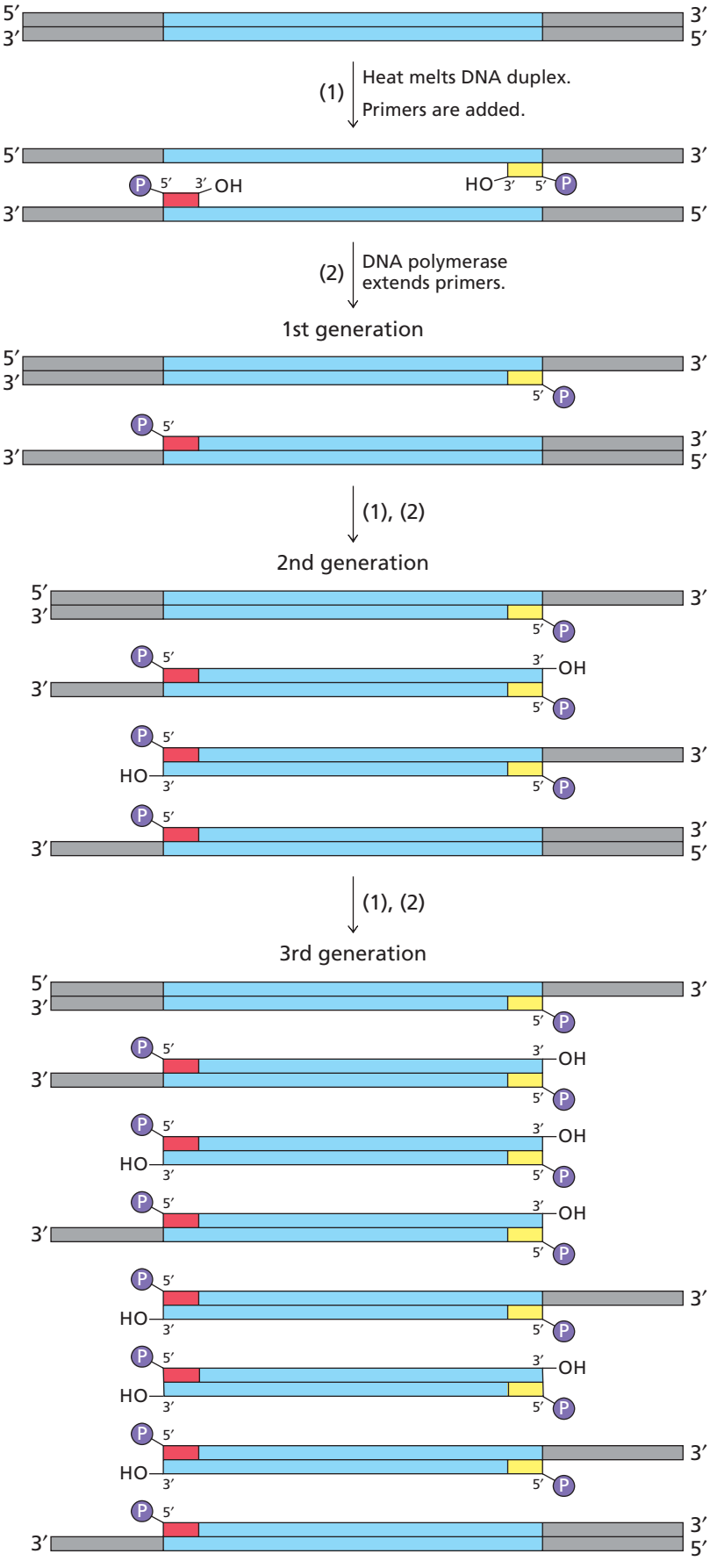
In 1976 Frederick Sanger developed a method for sequencing DNA enzymatically using the Klenow fragment of *E. coli* DNA polymerase I. Sanger was awarded his second Nobel Prize for this achievement (he received his first Nobel Prize for developing a method for sequencing proteins). The advantage of using the Klenow fragment for this type of reaction is that the enzyme lacks 5′ → 3′ exonuclease activity, which could degrade newly synthesized DNA. However, one of the disadvantages is that the Klenow fragment is not very processive and is easily inhibited by the presence of secondary structure in the single-stranded DNA template. This limitation can be overcome by adding SSB or analogous proteins, or more commonly, by using DNA polymerases from bacteria that grow at high temperatures. Such polymerases are active at 60° to 70°C, a temperature at which secondary structure in single-stranded DNA is unstable.

The Sanger sequencing method uses 2′, 3′-dideoxynucleoside triphosphates (ddNTPs), which differ from the deoxyribonucleotide substrates of DNA synthesis by lacking a 3′-hydroxyl group (see below). The dideoxyribonucleotides, which can serve as substrates for DNA polymerase, are added to the 3′ end of the growing chain. Because these nucleotides lack a 3′-hydroxyl group, subsequent nucleotide additions cannot take place and incorporation of a dideoxynucleotide terminates the growth of the DNA chain. When a small amount of a particular dideoxyribonucleotide is included in a DNA synthesis reaction, it is occasionally incorporated in place of the corresponding dNTP, immediately terminating replication. The length of the resulting fragment of DNA identifies the position of the nucleotide that should have been incorporated.

DNA sequencing using ddNTP molecules involves several steps (as shown on page 622). The DNA is prepared as single-stranded molecules and mixed with a short oligonucleotide complementary to the 3′ end of the DNA to be sequenced. This oligonucleotide acts as a primer for DNA synthesis catalyzed by DNA polymerase. The oligonucleotide-primed material is split into four separate reaction tubes. Each tube receives a small amount of an $\alpha$-[$^{32}$P]-labeled dNTP, whose radioactivity allows the newly synthesized DNA to be visualized by autoradiography. Next, each tube receives an excess of the four nonradioactive dNTP molecules and a small amount of one of the four ddNTPs. For example, the A reaction tube receives an excess of nonradioactive dTTP, dGTP, dCTP, and dATP mixed with a small amount of ddATP. DNA polymerase is then added to the reaction mixture. As the polymerase replicates the DNA, it occasionally incorporates a ddATP residue instead of a dATP residue, and synthesis of the growing DNA chain is terminated. Random incorporation of ddATP results in the production of newly synthesized DNA fragments of different lengths, each ending with A (i.e., ddA). The length of each fragment corresponds to the distance from the 5′-end of the primer to one of the adenine residues in the sequence. Adding a different dideoxyribonucleotide

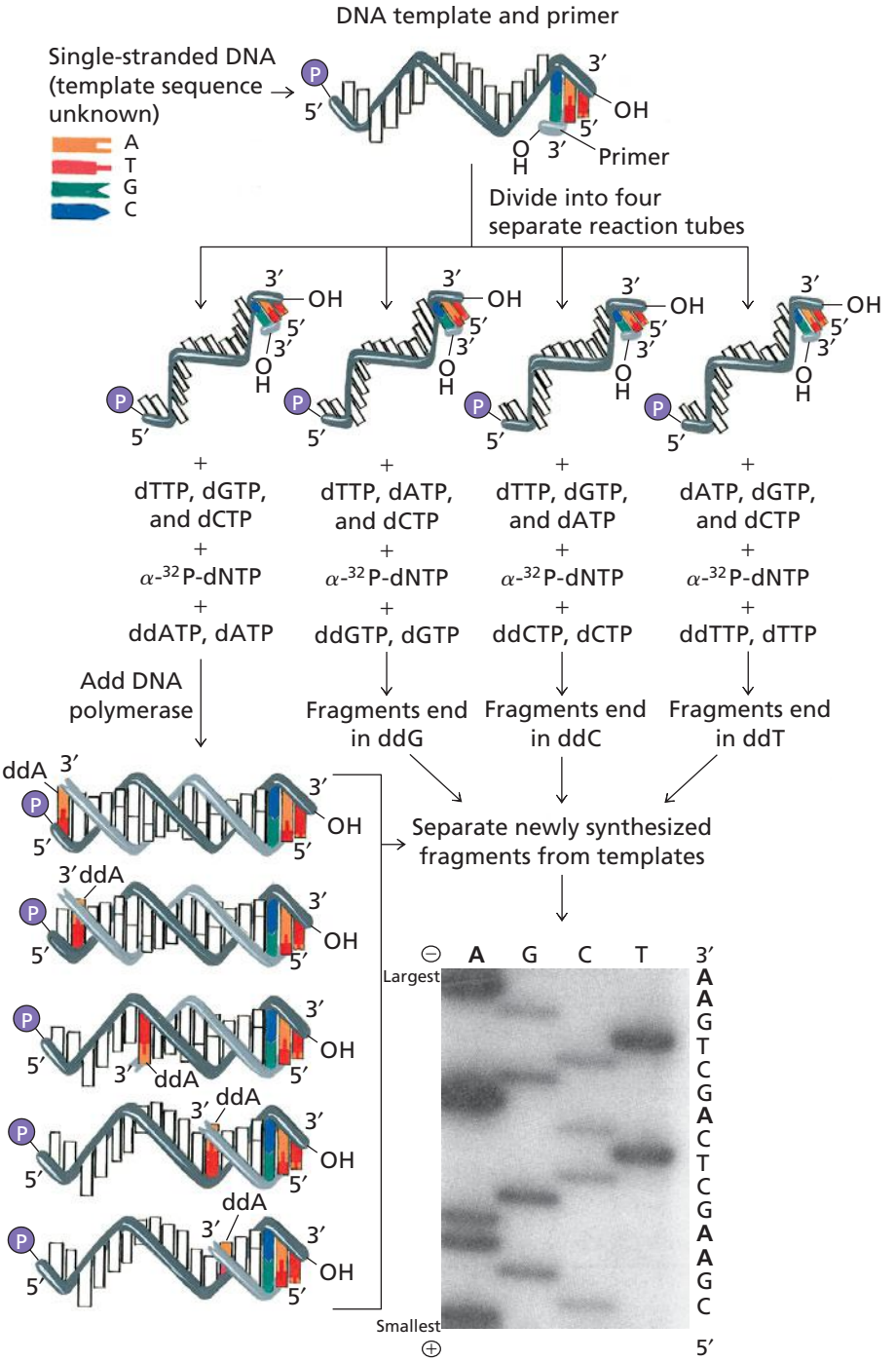



▶ Chemical structure of a 2′,3′-dideoxynucleoside triphosphate. *B* represents any base.




**◂ Three cycles of the polymerase chain reaction.** The sequence to be amplified is shown in blue. (1) The duplex DNA is melted by heating and cooled in the presence of a large excess of two primers (red and yellow) that flank the region of interest. (2) A heat-stable DNA polymerase catalyzes extension of these primers, copying each DNA strand. Successive cycles of heating and cooling in the presence of the primers allow the desired sequence to be repeatedly copied until, after 20 to 30 cycles, it represents most of the DNA in the reaction mixture.

**Sanger method for sequencing DNA.** ▶ Addition of a small amount of a particular dideoxynucleoside triphosphate (ddNTP) to each reaction mixture causes DNA synthesis to terminate when that dideoxynucleotide is incorporated in place of the normal nucleotide. The positions of incorporated dideoxynucleotides, determined by the lengths of the DNA fragments, indicate the positions of the corresponding nucleotide in the sequence. The fragments generated during synthesis with each ddNTP are separated by size using an electrophoretic sequencing gel, and the sequence of the DNA can be read from an autoradiograph of the gel (as shown by the column of letters to the right of the gel).




to each reaction tube produces a different set of fragments: ddTTP produces fragments that terminate with T, ddGTP with G, and ddCTP with C. The newly synthesized chains from each sequencing reaction are separated from the template DNA. Finally, the mixtures from each sequencing reaction are subjected to electrophoresis in adjacent lanes on a sequencing gel, where the fragments are resolved by size. The sequence of the DNA molecule can then be read from an autoradiograph of the gel.

This technique has also been modified to allow automation for high throughput applications like genomic sequencing. Instead of using radioactivity, automated sequencing relies on fluorescently labeled deoxynucleotides (four colors, one for each base) to detect the different chain lengths. In this system the gel is "read" by a fluorimeter and the data are stored in a computer file. Additionally, the sequencing machine can also provide a graphic chromatogram that shows the location and size of each fluorescent peak on the gel as they passed the detector.

### C. Massively Parallel DNA Sequencing by Synthesis

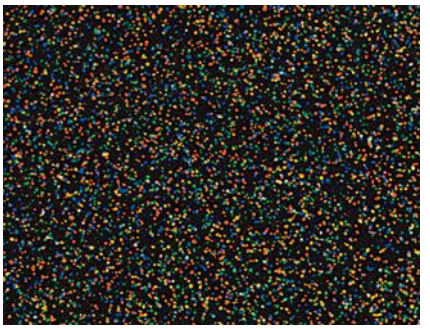
The automated DNA sequencing methods used to sequence the human genome have now been largely supplanted by a variety of so-called "next generation" sequencing technologies. While using slightly different experimental approaches, these devices can all rapidly generate millions (or even billions) of base pairs of sequence at a fraction of the cost of the automated Sanger technology described in the previous section. As an example of this novel approach, we describe the Illumina next-generation sequencing protocol.

In the first step, DNA (typically the entire genome) is randomly fragmented by shearing to yield short double-stranded fragments. The ends of the fragments are enzymatically repaired and a single-stranded oligonucleotide primer is ligated onto each end. Fragments of the desired length are purified from an agarose gel and then amplified using PCR. Oligonucleotides complementary to the PCR primers are covalently attached to the surface of a glass slide. The amplified genomic fragments are denatured into single strands, diluted, and hybridized to the oligonucleotides on the slide.

This creates a slide where millions of individual DNA fragments are bound to the surface. Each one is surrounded by a zone of free oligonucleotides bound to the surface. The individual DNA fragments on the slide's surface are then amplified *in situ* using a bridging technique to yield amplification clusters that are the substrate for the sequencing reaction.

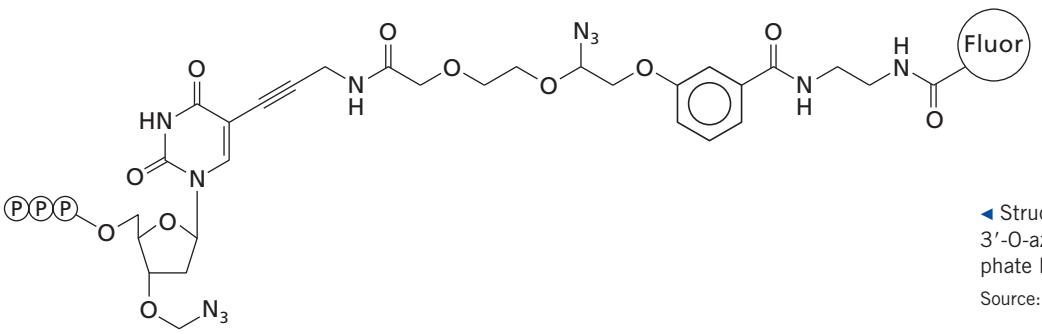
All of the clusters of amplified DNA fragments are sequenced at the same time, in parallel, using a mixture of the four dNTPs that have been labeled with a removable fluorophore (a different dye for each base) and a reversible terminator at the 3′position (see below). To increase the efficiency of this step, a genetically engineered mutant DNA polymerase from the deep hydrothermal vent archeon 9°N–7 that efficiently incorporates these bulky substrates is used. The DNA sequencing primer annealed to the template strands provides the 3′ hydroxyl group and the polymerase incorporates the next labeled nucleotide. The terminator at the 3′position of the incoming base prevents DNA synthesis beyond one single base. The slide is scanned using a laser-scanning confocal microscope to record the base that was incorporated into each growing cluster. The reducing agent TCEP is then added removing both the dye and the terminator to regenerate the 3′-OH. The whole cycle is then repeated. The growing DNA chains can only increase in length via a stepwise process: one base at a time.

The relatively short sequences (less than 100 nucleotides) are not suitable for assembling the genome sequence from a species that has never been sequenced before. However for resequencing a previously sequenced genome, fast computer algorithms can align these short "reads" with high accuracy and detect rare mutations or polymorphisms present in the sample.



▲ **Imaging clusters during the sequencing process.** Part of the image of a flow-cell with a low density of clusters is shown. Since each of the four deoxynucleotide bases is labeled with a different fluophore (each of which fluoresces at a different wavelength), the four separate images have been superimposed (after artificial coloring). After each cycle of DNA synthesis these images provide the raw data that reveal the last base that was incorporated into the growing polynucleotide chain.

Source: Bentley et al. (2008). *Nature* 456:53–59.




◄ Structure of the reversible terminator 3′-O-azidomethyl 2′-deoxythymine triphosphate labeled with a removable fluorophore.

Source: Bentley et al. (2008). *Nature* 456: 53–59.

## 20.7 DNA Replication in Eukaryotes

The mechanisms of DNA replication in prokaryotes and eukaryotes are fundamentally similar. In eukaryotes as in *E. coli*, synthesis of the leading strand is continuous and synthesis of the lagging strand is discontinuous. Furthermore, in both prokaryotes and eukaryotes, synthesis of the lagging strand is a stepwise process involving: primer synthesis, Okazaki fragment synthesis, primer hydrolysis, and gap filling by a polymerase. Eukaryotic primase, like prokaryotic primase, synthesizes a short primer once every second on the lagging-strand template. However, because the replication fork moves more slowly in eukaryotes, each Okazaki fragment is only about 100 to 200 nucleotide

**Table 20.2 Eukaryotic DNA polymerases**

| DNA polymerase | Activities | Role |
|---|---|---|
| $\alpha$ | Polymerase<br>Primase<br>$3' \rightarrow 5'$ Exonuclease[a] | Primer synthesis<br>Repair |
| $\beta$ | Polymerase | Repair |
| $\gamma$ | Polymerase<br>$3' \rightarrow 5'$ Exonuclease | Mitochondrial DNA replication |
| $\delta$ | Polymerase<br>$3' \rightarrow 5'$ Exonuclease | Leading- and lagging-strand synthesis<br>Repair |
| $\varepsilon$ | Polymerase<br>$3' \rightarrow 5'$ Exonuclease<br>$5' \rightarrow 3'$ Exonuclease | Repair<br>Gap filling on lagging strand |

[a]Polymerase $\alpha 3' \rightarrow 5'$ exonuclease activity is not detectable in all species.

residues long, considerably shorter than in prokaryotes. Interestingly, eukaryotic DNA primase does not share significant sequence similarity with the *E. coli* enzyme nor does eukaryotic primase contain some of the classical structural landmarks of DNA polymerases such as the "fingers" or "thumb" domains (Figure 20.12). This lack of homology suggests that the capacity to synthesize an RNA primer for DNA initiation may have evolved independently at least twice.

Most eukaryotic cells contain at least five different DNA polymerases: $\alpha$, $\beta$, $\gamma$, $\delta$, and $\varepsilon$ (Table 20.2). DNA polymerases $\alpha$, $\delta$, and $\varepsilon$ are responsible for the chain elongation reactions of DNA replication and for some repair reactions. DNA polymerase $\beta$ is a DNA repair enzyme found in the nucleus and DNA polymerase $\gamma$ plays a role in replicating mitochondrial DNA. A sixth DNA polymerase is responsible for replicating DNA in chloroplasts.

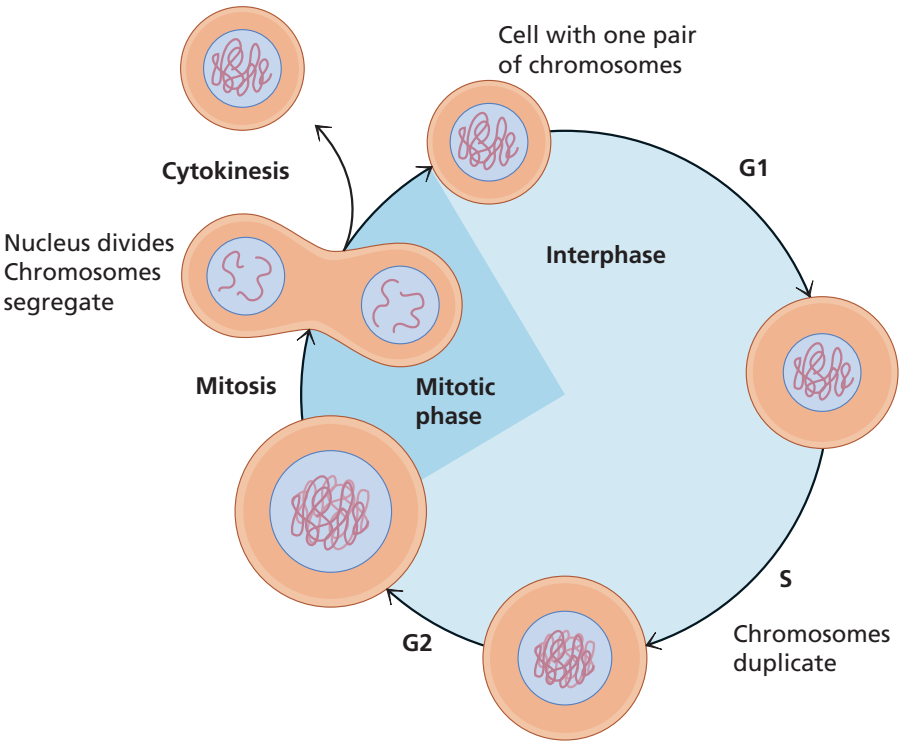
DNA polymerase $\delta$ catalyzes synthesis of the leading strand at the replication fork. This enzyme is composed of two subunits the larger of which contains the polymerase active site. The enzyme also has $3' \rightarrow 5'$ exonuclease activity. DNA replication in eukaryotic cells is extremely accurate. The low error rate indicates that DNA replication in eukaryotes includes an efficient proofreading step.

DNA polymerase $\alpha$ and DNA polymerase $\delta$ cooperate in lagging strand synthesis. DNA polymerase $\alpha$ is a multimeric protein that contains both DNA polymerase and RNA primase activity. The primer made by DNA polymerase $\alpha$ consists of a short stretch of RNA followed by DNA. This two part primer is extended by DNA polymerase $\delta$ to complete an Okazaki fragment.

DNA polymerase $\varepsilon$ is a large, multimeric protein. The largest polypeptide chain includes polymerase activity and $3' \rightarrow 5'$ proofreading exonuclease activity. Like its functional counterpart in *E. coli* (DNA polymerase I), DNA polymerase $\varepsilon$ probably acts as a repair enzyme and also fills gaps between Okazaki fragments.

Several accessory proteins are associated with the replication fork in eukaryotes. These proteins function like some of the proteins in the bacterial replisome. For example, PCNA (proliferating cell nuclear antigen) forms a structure that resembles the $\beta$-subunit sliding clamp of *E. coli* DNA polymerase III (Figure 20.7). The accessory protein RPC (replication factor C) is structurally, functionally, and evolutionarily related to the $\gamma$ complex of DNA polymerase III. Another protein, called RPA (replication factor A), is the eukaryotic equivalent of prokaryotic SSB. In addition, the eukaryotic replication machine includes helicases that unwind DNA at the replication fork.

Each eukaryotic chromosome contains many origins of replication (Section 20.1). The largest chromosome of the fruit fly *Drosophila melanogaster*, for example, contains about 6000 replication forks implying that there are at least 3000 origins. As replication proceeds bidirectionally from each origin the forks move toward one another, merging to form bubbles of ever increasing size (Figure 20.4). Due to the large number of origins, the larger chromosomes of eukaryotes can still be replicated in less than one hour even though the rate of individual fork movement is much slower than in prokaryotes.




◀ **Figure 20.18**
**The eukaryotic cell division cycle coordinates DNA replication and mitosis.** DNA replication occurs exclusively during the synthesis, or S-phase of the cell cycle. There are two gap, or G, phases where a cell grows prior to dividing in the mitosis, or M-phase.

DNA replication in all cells occurs within the context of the cell's programmed cell division cycle. This cell cycle is a highly regulated progression through a series of dependent steps that at a minimum accomplishes two goals: (1) it faithfully duplicates all of the DNA in a cell to produce exactly two copies of each chromosome, and (2) it precisely segregates one copy of each replicated chromosome into one of the two daughter cells. In eukaryotic cells chromosomal segregation occurs at mitosis and this stage is called the mitotic phase, or M-phase (Figure 20.18). The step where DNA is synthesized is called S-phase. The interphase (resting) stage between mitosis and the next round of DNA replication is called G1. There may be a G2 stage between the end of DNA replication and the beginning of mitosis.

Eukaryotic DNA replication origins must be used once, and only once, during S-phase of each cell cycle. We are beginning to understand some of the key players that orchestrate this process. At the end of the previous M-phase and during the subsequent G1-phase, each functional *ori* becomes an assembly site for a conserved multiprotein complex named ORC (origin recognition complex). As the cell progresses through G1 each ORC stimulates the formation of a prereplication complex (pre-RC) that includes a helicase. The pre-RC remains poised until the activity of an S-phase protein kinase (SPK) drops to a critical threshold, whereby the initiation complex recruits waiting replisomes and the origin is said to "fire." The two replication forks are then launched along the chromosome in opposite directions. When SPK activity is high it prevents any new pre-RCs from loading onto the origins, thus preventing multiple rounds of initiation. SPK is proteolytically cleaved at the beginning of the mitotic phase allowing ORC proteins to bind to the origins waiting on each daughter chromosome beginning late in M-phase.

Eukaryotic replication origins do not all fire simultaneously at the beginning of S-phase. Instead, transcribed, or active, regions of a cell's genome tend to be replicated earlier during S-phase while the origins located in quiescent, or repressed, regions of the genome tend to be replicated later in S-phase. It remains to be determined whether this differential timing of replication actually depends on transcription or just reflects that "open" chromatin permits ORC to locate replication origins.

The differences between eukaryotic and prokaryotic DNA replication arise not only from the larger size of the eukaryotic genome but also from the packaging of eukaryotic DNA into chromatin. The binding of DNA to histones and its packaging into nucleosomes, (Section 19.5), is thought to be responsible in part for the slower movement of

**Figure 20.19 ►**
**Photodimerization of adjacent deoxythymidylate residues.** Ultraviolet light causes the bases to dimerize, thus distorting the structure of DNA. For clarity, only a single strand of DNA is shown.

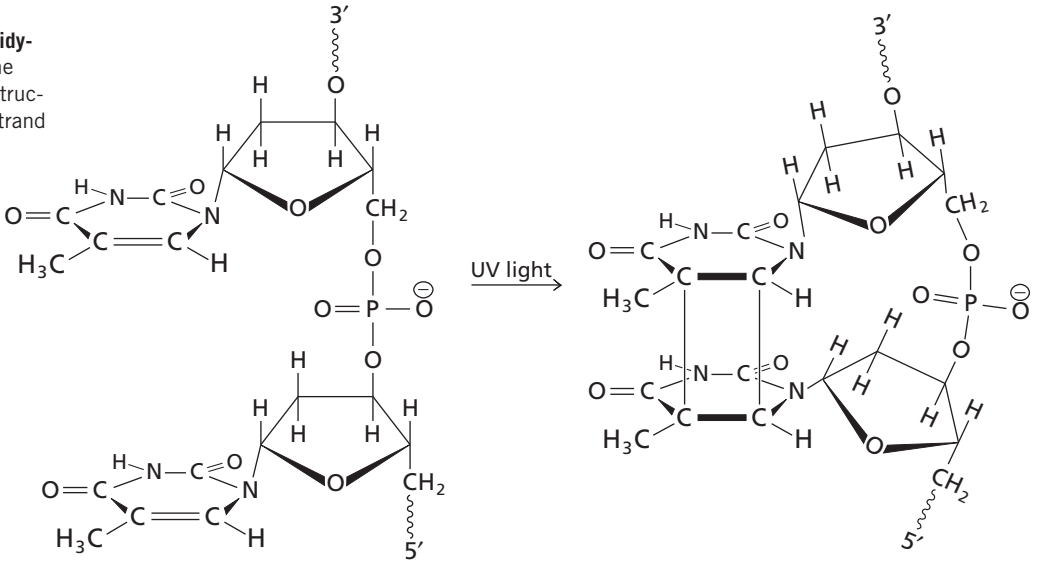
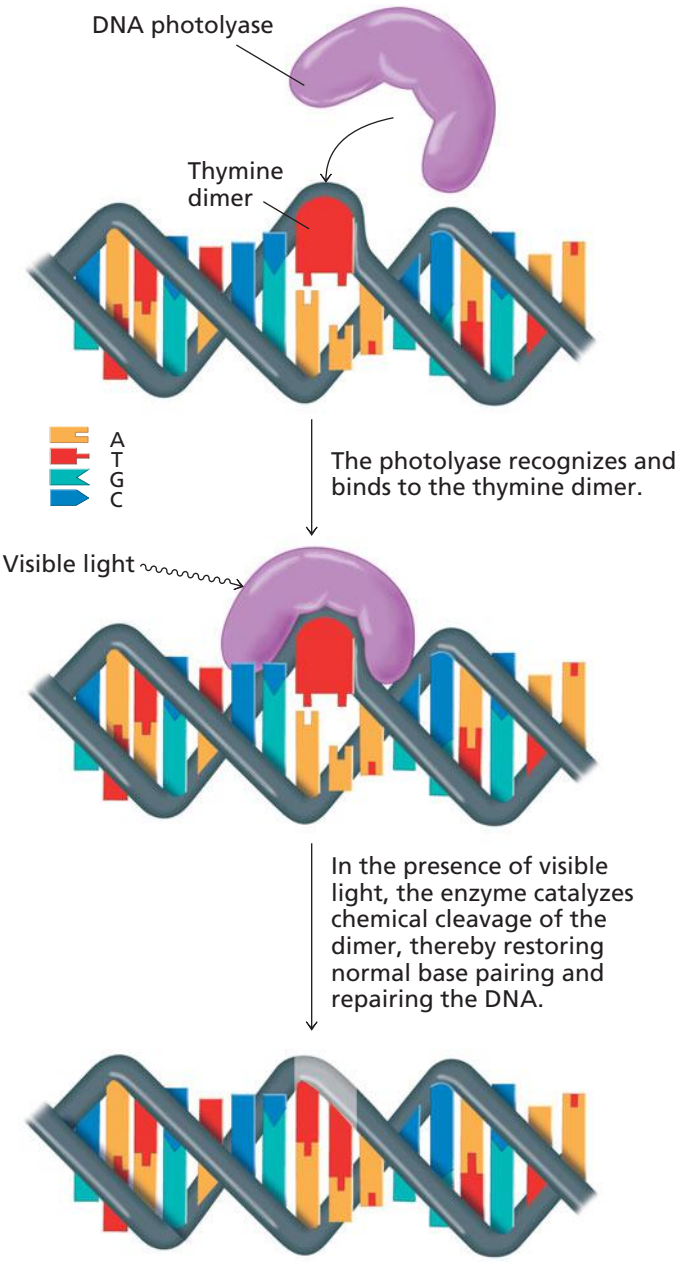



**Figure 20.20 ►**
**Repair of thymine dimers by DNA photolyase.**



the replication fork in eukaryotes. Eukaryotic DNA replication occurs with concomitant synthesis of histones; the number of histones doubles with each round of DNA replication. Histone duplication and DNA replication involve different enzymes acting in different parts of the cell yet both occur at about the same rate. It appears that existing histones remain bound to DNA during replication and that newly synthesized histones bind to DNA behind the replication fork shortly after synthesis of the new strands.

## 20.8 Repair of Damaged DNA

DNA is the only cellular macromolecule that can be repaired. This is probably because the cost to the organism of mutated or damaged DNA far outweighs the energy spent to repair the defect. Repairing other macromolecules is not profitable; for example, little is lost when a defective protein forms as a result of a translation error because the protein is simply replaced by a new, functional protein. When DNA is damaged, however, the entire organism may be in jeopardy if the instructions for synthesizing a critical molecule are altered. In single-celled organisms, damage to a gene encoding an essential protein may kill the organism. Even in multicellular organisms, the accumulation of defects in DNA over time can lead to progressive loss of cellular functions or to deregulated growth such as that seen in cancer cells.

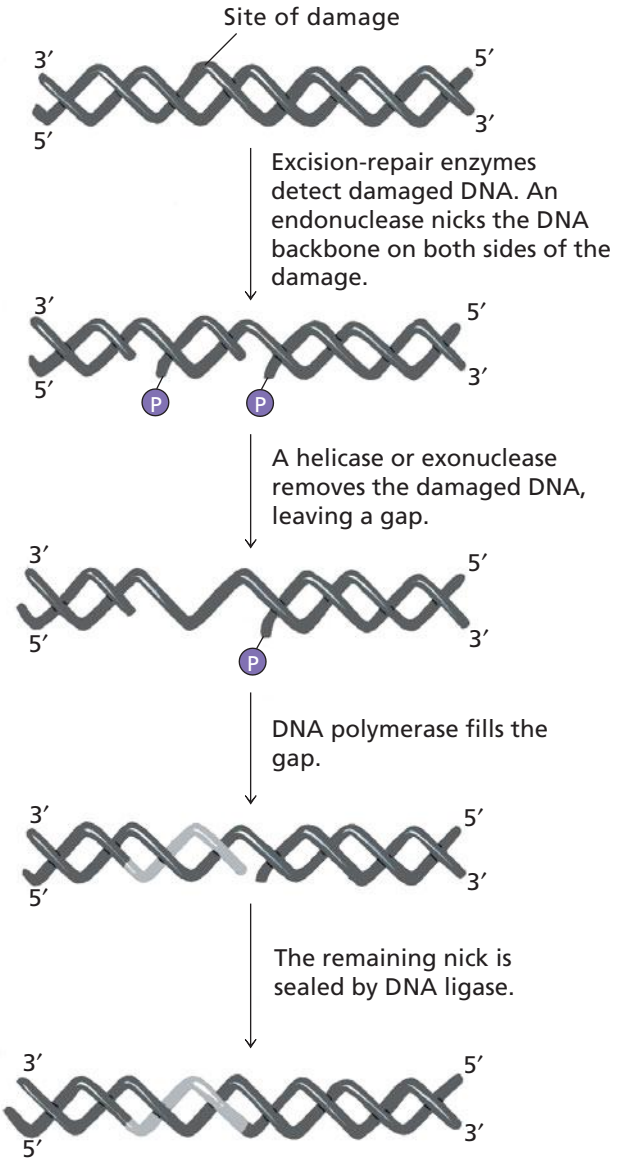
There are several types of DNA damage such as base modifications, nucleotide deletions or insertions, cross-linking of DNA strands, and breakage of the phosphodiester backbone. While some DNA damage is the result of environmental agents (e.g., chemicals or radiation) most DNA damage is the result of errors made during DNA replication. Severe damage may be lethal but much of the damage that occurs *in vivo* is repaired. Many modified nucleotides, as well as mismatched bases that escape the proofreading mechanism of DNA polymerase, are recognized by specific repair enzymes that continually scan DNA in order to detect alterations. Some of the lesions are fixed by **direct repair**, a process that does not require breaking the phosphodiester backbone of DNA. Other repairs require more extensive work.

DNA repair mechanisms protect individual cells as well as subsequent generations. In single-celled organisms, whether prokaryotes or eukaryotes, DNA damage that is not repaired may become a mutation that is passed directly to the daughter cells following DNA replication and cell division. In multicellular organisms, mutations can be passed on to the next generation only if they occur in the germ line. Germ line mutations may have no noticeable effect on the organism that contains them but may have profound effects on the progeny, especially if the mutated genes are important in development. When mutations occur in somatic cells however, while the defects are not transmissible, they can sometimes lead to unrestricted cell growth, or cancer. In spite of the accuracy of DNA replication and the efficiency of repair, the average human accumulates about 130 new mutations every generation. Most of these mutations are neutral and this leads to a huge amount of variation in human populations. It is this variation that makes possible the identification of individuals by DNA fingerprinting.

### A. Repair after Photodimerization: An Example of Direct Repair

Double-helical DNA is susceptible to damage by ultraviolet (UV) light. The most common UV light-induced damage is dimerization of adjacent pyrimidines in a DNA strand. This process is an example of photodimerization. The most common dimers form between adjacent thymines (Figure 20.19). DNA replication cannot occur in the presence of pyrimidine dimers because they distort the template strand. Therefore, removal of pyrimidine dimers is essential for survival.

Many organisms can repair thymine dimer damage using direct repair (notably, humans and all placental mammals lack this repair mechanism—see below). The simplest repair process begins when an enzyme known as DNA photolyase binds to the distorted double helix at the site of the thymine dimer (Figure 20.20). As the DNA-enzyme complex absorbs visible light, the dimer is cleared. The photolyase then dissociates from the repaired DNA and normal A/T base pairs re-form. This process is called photo reactivation; it's an example of direct repair.




▲ **Figure 20.21**
**General excision-repair pathway.**

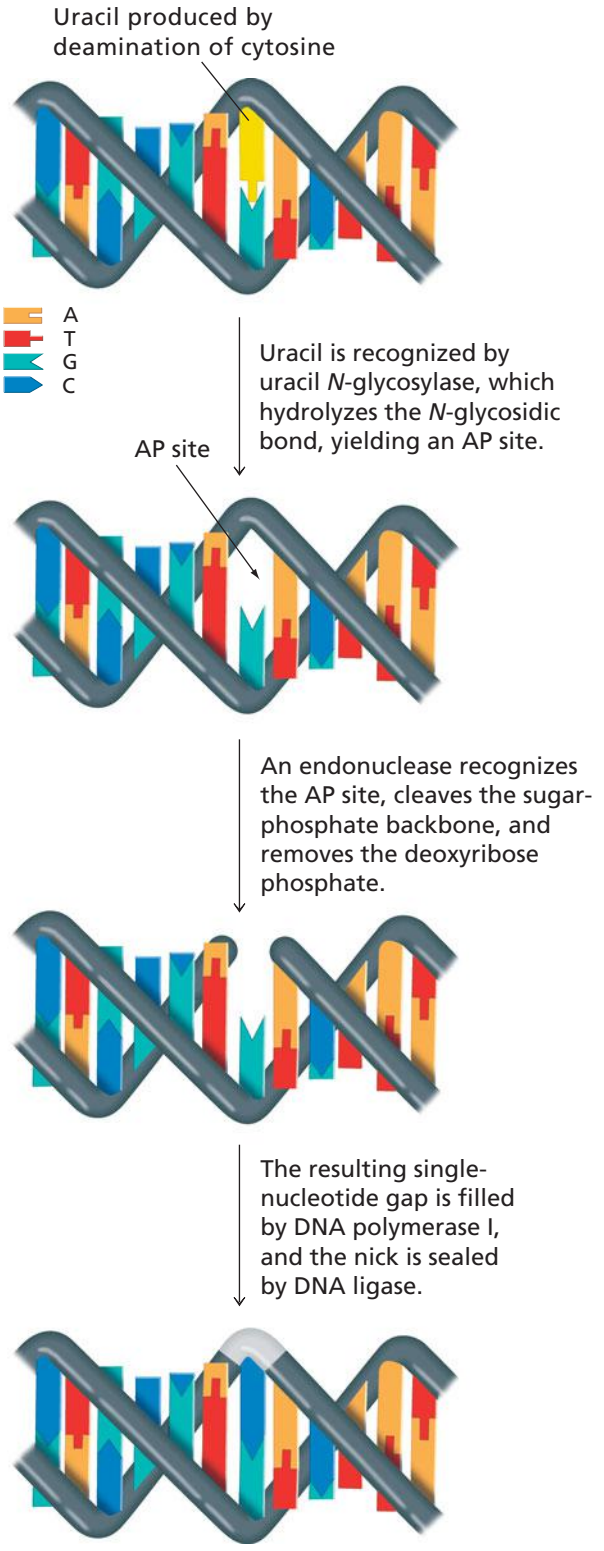
## B. Excision Repair

Other forms of ionizing radiation and naturally occurring chemicals can damage DNA. Some compounds, including acids and oxidizing agents, can modify DNA by alkylation, methylation, or deamination. DNA is also susceptible to spontaneous loss of heterocyclic bases, a process known as depurination or depyrimidization. Many of these defects can be repaired by a general **excision repair pathway** whose overall features are similar in all organisms. The pathway begins when an endonuclease recognizes distorted, damaged DNA and cleaves on both sides of the lesion releasing an oligonucleotide containing 12 to 13 residues. This cleavage is catalyzed by the UvrABC enzyme in *E. coli*. Removal of the DNA oligonucleotide may require helicase activity that is often a component of the excision repair enzyme complex. The result is a single-stranded gap. The gap is then filled in by the action of DNA polymerase I in prokaryotes or repair DNA polymerases in eukaryotes. The nick is sealed by DNA ligase (Figure 20.21).

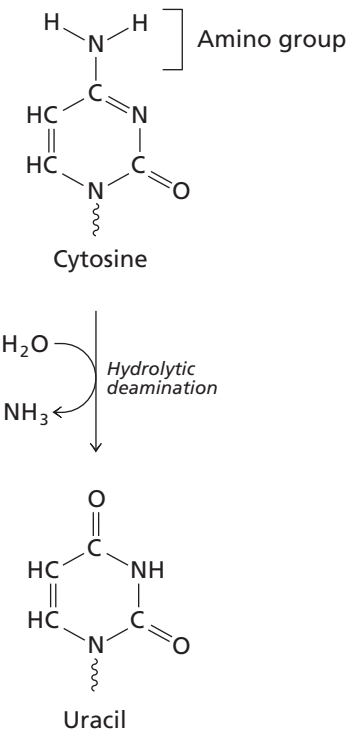
The UvrABC endonuclease also recognizes pyrimidine dimers and modified bases that distort the double helix (this is how thymine dimers are repaired in humans). Other excision-repair enzymes recognize DNA damaged by hydrolytic deamination of adenine, cytosine, or guanine. (Thymine is not subject to deamination because it does not have an amino group.) The deaminated bases can form incorrect base pairs resulting in the incorporation of incorrect bases during the next round of replication. Spontaneous deamination of cytosine is one of the most common types of DNA damage because the product of deamination is uracil that easily forms a base pair with adenine in the next round of replication (Figure 20.22).

Enzymes called DNA glycosylases remove deaminated bases and some other modified bases by catalyzing hydrolysis of the *N*-glycosidic bonds that link the modified bases to the sugars. Let's look at the repair of deaminated cytosine. Repair begins when the enzyme uracil *N*-glycosylase removes the uracil produced by deamination. The enzyme recognizes and binds to the incorrect U/G base pair and flips the uracil base outward, positioning the $\beta$-*N*-glycosidic bond in the active site of the enzyme where it is cleaved from the sugar residue (Figure 20.23). Next, an endonuclease recognizes the site where the base is missing and removes the deoxyribose phosphate, leaving a single-nucleotide gap in the duplex DNA. The endonuclease is called an AP-endonuclease because it recognizes apurinic and apyrimidinic sites (AP sites). Some specific DNA glycosylases are bifunctional enzymes with both glycosylase and AP-endonuclease activities in the same polypeptide chain. Excision repair enzymes with exonuclease activity often extend the gap produced by the endonuclease. In prokaryotes, DNA polymerase I binds to the exposed 3′ end of DNA and fills in the gap. Finally, the strand is sealed by DNA ligase. The steps of the excision repair pathway are summarized in Figure 20.24.

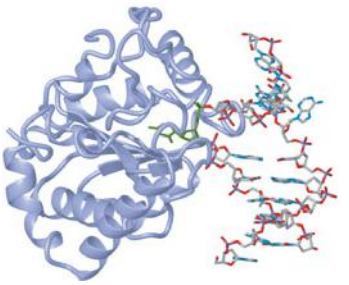
Whereas deamination of adenine or guanine is rare, deamination of cytosine is fairly common and would give rise to large numbers of mutations were it not for the replacement of uracil with thymine in DNA. (Recall that thymine is simply 5-methyluracil.) If uracil were normally found in DNA, as it is in RNA, it would be impossible to distinguish between a correct uridylate residue and one arising from the deamination of cytosine. However, since uracil is not one of the bases in DNA, damage arising from cytosine deamination can be recognized and repaired. Thus, the presence of thymine in DNA increases the stability of genetic information.




▲ **Figure 20.24**
**Repair of damage resulting from the deamination of cytosine.**




▲ **Figure 20.22**
**Hydrolytic deamination of cytosine.** Deamination of cytosine produces uracil, which pairs with adenine rather than guanine.



▲ **Figure 20.23**
**Uracil *N*-glycosylase from human mitochondria.** The enzyme is bound to a uracil-containing nucleotide (green) that has been flipped out of the stacked region of double-stranded DNA. [PDB 1EMH].

### BOX 20.1 THE PROBLEM WITH METHYLCYTOSINE

5-Methylcytosine is common in eukaryotic DNA (Section 18.7). Deamination of 5-methylcytosine produces thymidine giving rise to a T opposite a G in damaged DNA. Repair enzymes cannot recognize which of these bases is incorrect, so the "repair" often results in a T:A base pair. This will also happen if the damaged DNA is replicated before it can be repaired. The cytosines at CG sites are preferentially methylated in mammalian genomes. Frequent loss of the cytosines by deamination of 5-methylcytosine has led to underrepresentation of CG sequences relative to TG, AG, and GG.

## 20.9 Homologous Recombination

Recombination is any event that results in the exchange or transfer of pieces of DNA from one chromosome to another or within a chromosome. Most recombinations are examples of **homologous recombination** because they occur between pieces of DNA that have closely related sequences. Exchanges between paired chromosomes during meiosis are examples of homologous recombination. Recombination between unrelated sequences is called **nonhomologous recombination**. **Transposons** are mobile genetic elements that jump from chromosome to chromosome by taking advantage of nonhomologous recombination mechanisms. Recombination between DNA molecules also occurs when bacteriophages integrate into host chromosomes. When recombination occurs at a specific location it is called **site specific recombination**.

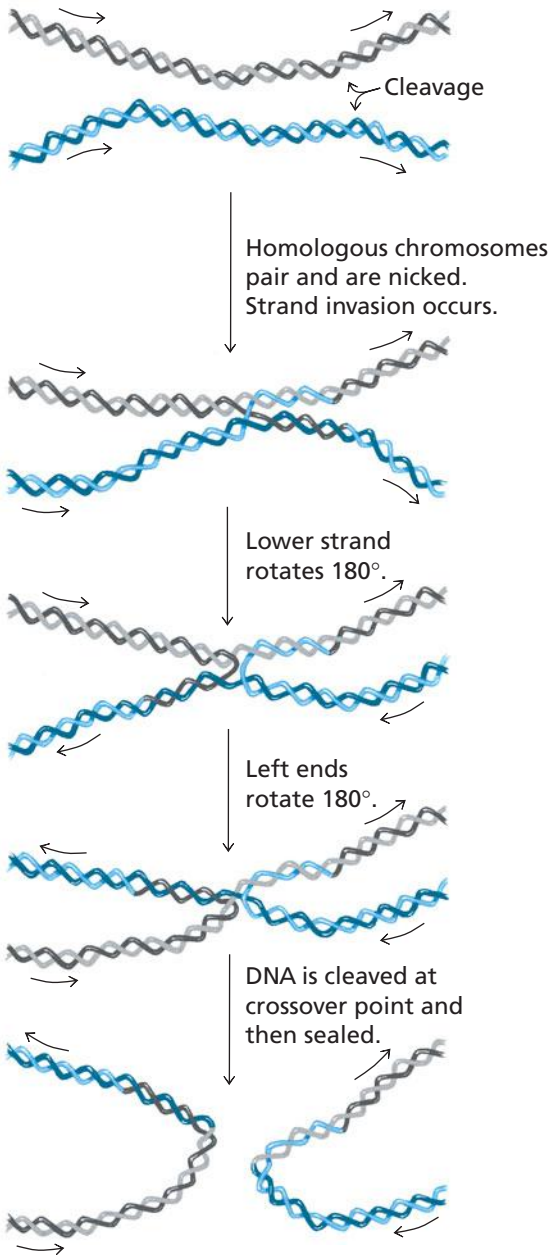
Mutation creates new genetic variation in a population and recombination is a mechanism that creates different combinations of mutations in a genome. Most species have some mechanism for exchanging information between individual organisms. Prokaryotes usually contain only a single copy of their genome (i.e., they are haploid), so this exchange requires recombination. Some eukaryotes are also haploid but most are diploid, having two sets of chromosomes, one contributed by each parent. Genetic recombination in diploids mixes the genes on the chromosomes contributed by each parent so that subsequent generations receive very different combinations of genes. None of your children's chromosomes, for example, will be the same as yours and none of yours are the same as those of your parents. (Although this mixing of alleles is an important consequence of recombination, it is not likely to be the reason why recombination mechanisms evolved in the first place. The problem of why sex evolved is one of the most difficult problems in biology.)

Recombination occurs by many different mechanisms. Many of the proteins and enzymes that participate in recombination reactions are also involved in DNA repair reactions illustrating the close connection between repair and recombination. In this section, we briefly describe the Holliday model of general recombination—a type of recombination that seems to occur in many species.

### A. The Holliday Model of General Recombination

Homologous recombination begins with the introduction of either single-stranded or double-stranded breaks into DNA molecules. Recombination involving single-stranded breaks is often called general recombination. Recombination involving double-stranded breaks is not discussed here, although it is an important mechanism of recombination in some species.

Consider general recombination between two linear chromosomes as an example of recombination in prokaryotes. The exchange of information between the molecules begins with the alignment of homologous DNA sequences. Next, single-stranded nicks are introduced in the homologous regions and single strands exchange in a process called strand invasion. The resulting structure contains a region of strand crossover and




◂ **Figure 20.25**
**Holliday model of general recombination.** Nicks are introduced into a homologous region of each molecule. Subsequent strand invasion, DNA cleavage at the crossover junction, and sealing of nicked strands result in exchange of the ends of the chromosomes.

▲ **Asexual Daphnia**

is known as a Holliday junction after Robin Holliday who first proposed it in 1964 (Figure 20.25).

The chromosomes can be separated at this stage by cleaving the two invading strands at the crossover point. It is important to realize that the ends of the homologous DNA molecules can rotate generating different conformations of the Holliday junction. Rotation followed by cleavage produces two chromosomes that have exchanged ends as shown in Figure 20.25. Recombination in many different organisms probably occurs by a mechanism similar to the one shown in Figure 20.25.

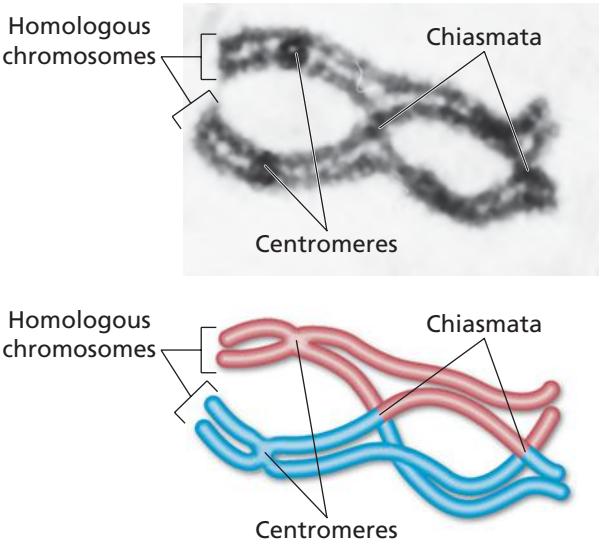
## B. Recombination in *E. coli*

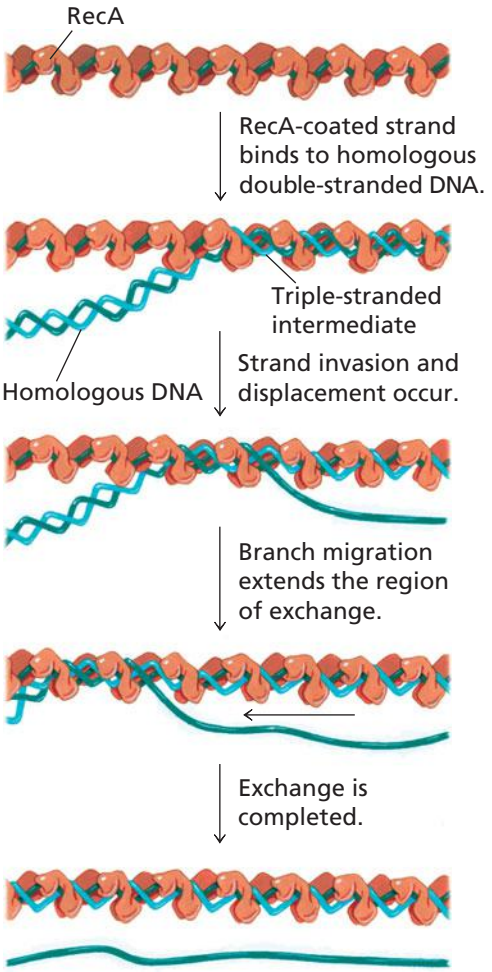
One of the first steps in recombination is the generation of single-stranded DNA with a free 3′ end. In *E. coli*, this step is carried out by RecBCD endonuclease, an enzyme with subunits that are encoded by three genes (*rec*B, *rec*C, and *rec*D) whose products have long been known to play a role in recombination. RecBCD binds to DNA and cleaves

▲ **Male Drosophila melanogaster (no meiotic recombination)**

**Meiotic chisasmata ►**

Source: © 2008 Sinauer Associates Sadava, D. et al. *Life: The Science of Biology*, 8th ed. (Sunderland, MA: Sinauer Associates and W. H. Freeman & Company), 198




▲ **Figure 20.26**
**Strand exchange catalyzed by RecA.**

one of the strands. It then unwinds the DNA in a process coupled to ATP hydrolysis generating single-stranded DNA with a 3′ terminus.

Strand exchange during recombination begins when the single-stranded DNA invades the double helix of a neighboring DNA molecule. Strand exchange is not a thermodynamically favorable event—the invasion must be assisted by proteins that promote recombination and repair. RecA is the prototypical strand exchange protein. It is essential for homologous recombination and for some forms of repair. The protein functions as a monomer that binds cooperatively to single-stranded DNA such as the single-stranded tails produced by the action of RecBCD. Each RecA monomer covers about five nucleotide residues and each successive monomer binds to the opposite side of the DNA strand.

One of the key roles of RecA in recombination is to recognize regions of sequence similarity. RecA promotes the formation of a triple-stranded intermediate between the RecA-coated single strand and a highly similar region of double-stranded DNA. RecA then catalyzes strand exchange in which the single strand displaces the corresponding strand from the double helix.

Strand exchange takes place in two steps: strand invasion, followed by branch migration (Figure 20.26). Both the single-stranded and the double-stranded DNA are in an extended conformation during the exchange reaction. The strands must rotate around each other, a process that is presumably aided by topoisomerases. Strand exchange is a slow process despite the fact that no covalent bonds are broken. (A "slow" process in biochemistry is one that takes several minutes.)

RecA can also promote strand invasion between two aligned, double-stranded DNA molecules. Both molecules must contain single-stranded tails bound to RecA. The tails wind around the corresponding complementary strands in the homologue. This exchange gives rise to a Holliday junction such as the one shown in Figure 20.25. Subsequent branch migration can extend the region of strand exchange. Branch migration can continue even after RecA dissociates from the recombination intermediate.

Branch migration at the double-stranded version of a Holliday junction is driven by a remarkable protein machine found in all species. The bacterial version is made up of RuvA and RuvB subunits. These proteins bind to the junction and

◄ **Bacterial conjugation (or sex).**

promote branch migration as shown in the schematic diagram (Figure 20.27). The two DNA helices are separated when RuvC binds to the Holliday junction and cleaves the crossover strands.

RuvA and RuvB form a complex consisting of four RuvA subunits bound to the Holliday junction and two hexameric rings of RuvB subunits that surround two of the DNA strands (Figure 20.28). The RuvB component is similar to the sliding clamps discussed in the section on DNA replication (Section 20.2B) and it drives branch migration by pulling the strands through the RuvA/Holliday junction complex in a reaction coupled to ATP hydrolysis (Figure 20.29). The rate of RuvAB-mediated branch migration is about 100, 000 bp per second—significantly faster than strand invasion.

RuvC catalyzes cleavage of the crossover strands to resolve Holliday junctions. Two types of recombinant molecules are produced as a result of this cleavage: those in which only single strands are exchanged and those in which the ends of the chromosome have been swapped (Figure 20.25).

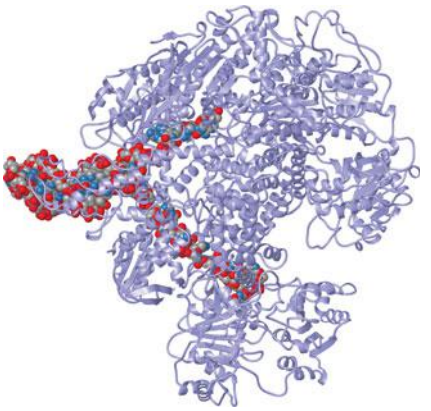
## C. Recombination Can Be a Form of Repair

Since natural selection works predominantly at the level of individual organisms it is difficult to see why recombination would have evolved unless it affected survival of the individual. Recombination enzymes probably evolved because they play a role in DNA repair, which confers a selective advantage. For example, severe lesions in DNA are bypassed during DNA replication, leaving a daughter strand with a single-stranded region. RecA-mediated strand exchange between the homologous daughter chromosomes allows the intact strand from one daughter molecule to act as a template for repairing the broken strand of the other daughter molecule.

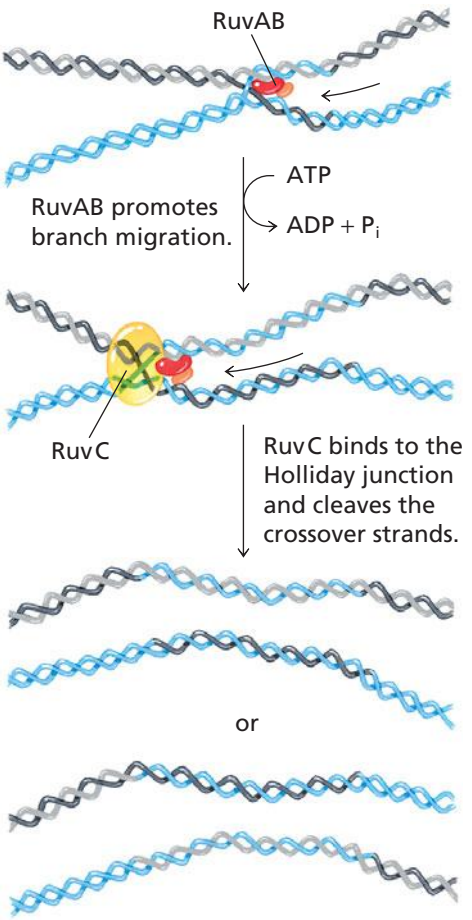
Recombination also creates new combinations of genes on a chromosome and this may be an added bonus for the population and its chances for evolutionary survival. More than 100 *E. coli* genes are required for recombination and repair, and there are twice as many in most eukaryotes.

Most, if not all, of the genes used in recombination play some role in repair as well. Mutations in several human genes give rise to rare genetic defects that result from deficiencies in DNA repair and/or recombination. For example, xeroderma pigmentosum is a hereditary disease associated with extreme sensitivity to ultraviolet light and increased frequency of skin cancer. Excision repair is defective in patients with this disease but the phenotype can be due to mutations in at least eight different genes. One of these genes encodes a DNA glycosylase with AP-endonuclease activity. Other affected genes include some that encode helicases that are required for both repair and recombination.

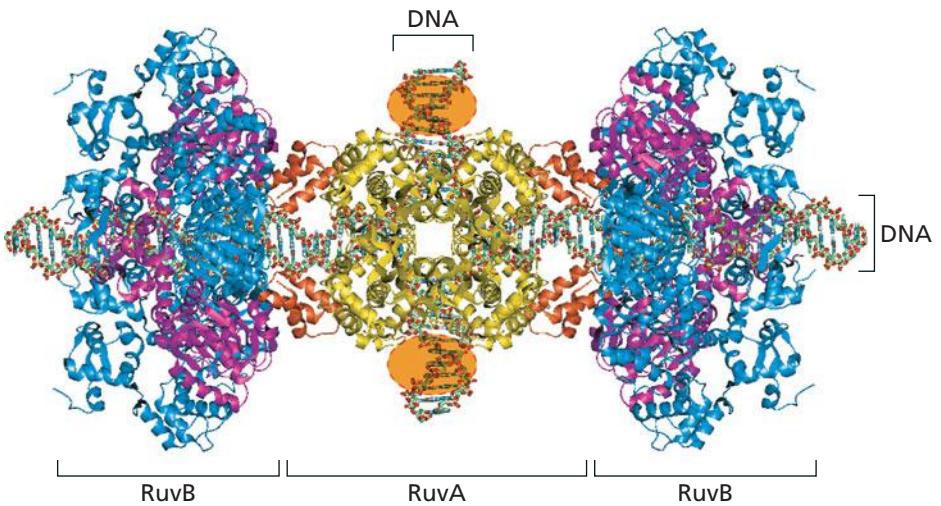
Many other genetic defects related to deficiencies in repair and recombination have not been well characterized. Some of them are responsible for increased incidences of cancer in affected patients.



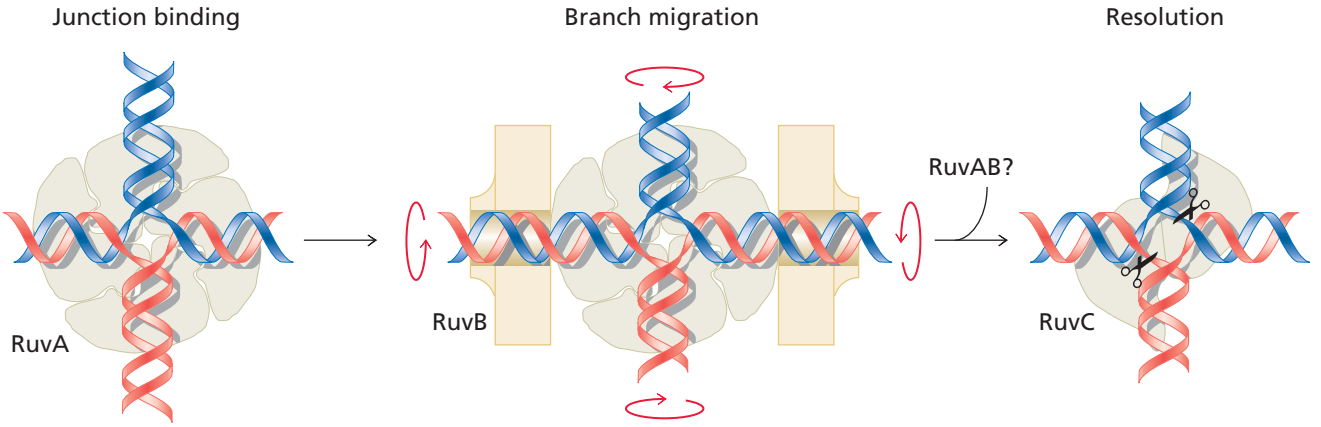
▲ **RecBCD bound to DNA showing separation of strands.** [PDB 3K70]




▲ **Figure 20.27**
**Action of Ruv proteins at Holliday junctions.** RuvAB promotes branch migration in a reaction coupled to ATP hydrolysis. RuvC cleaves Holliday junctions. Two types of recombinant molecules can be generated in this reaction.




◀ **Figure 20.28**
**Model of RuvA and RuvB bound to a Holliday junction.**

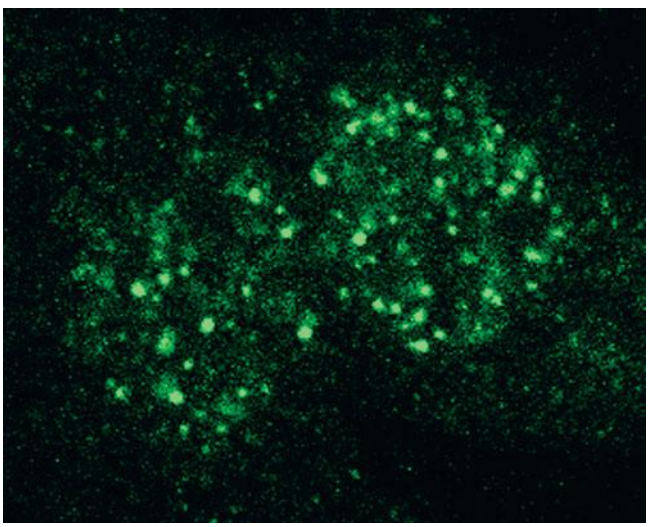



▲ **Figure 20.29**
**Branch migration and resolution.** [Adapted from Rafferty, J. B., et al. (1996). Crystal structure of DNA recombination protein RuvA and a model for its binding to the Holliday junction. *Science* 274:415–421.]

## BOX 20.2 MOLECULAR LINKS BETWEEN DNA REPAIR AND BREAST CANCER

About 180,000 women are diagnosed with breast cancer every year in North America. Approximately one-fifth of these new cases have a familial or genetic component and one-third of these, or 12,000 cases, are due to mutations in one of the two genes named *BRCA1* or *BRCA2* that encode proteins by the same name.

Both of these proteins are required for normal recombinational repair of double strand breaks (DSB). BRCA2 forms a complex with the eukaryotic RecA homologue RAD51. BRCA2 also binds specifically to BRCA1 to form a heterotrimer. Following exposure to ionizing radiation, these three DNA repair proteins are found localized to discrete sites, or foci, inside the interphase nuclei (see figure). These foci are the sites where the proteins are repairing double strand breaks. The BRCA proteins are so vital that cells become susceptible to damage if either copy of the gene is damaged. When one or both copies of the *BRCA1* or *BRCA2* genes are defective, the capacity to repair DSBs is compromised leading eventually to a higher frequency of mutations. Some of these new mutations may allow the cell to escape from the rigorous constraints imposed by the eukaryotic cell cycle, eventually leading to cancer. The BRCA proteins function as sentinels by continually monitoring the genome to identify and correct potential mutagenic lesions. In fact, some humans with a rare autosomal recessive disease called Fanconi's Anemia (FA) have an increased sensitivity to several mutagenic compounds and a genetic predisposition to many different types of cancers. It has been shown that FA patients are affected in one of seven different genes that are presumably important for DNA repair. One of these genes is *BRCA2*, underscoring its essential role in the repair process.



▲ Ionizing radiation induces nuclear foci of the DNA repair protein BRCA1. Energetic γ-rays can induce double-stranded breaks in DNA and trigger DNA repair. This tissue culture cell nucleus was exposed to IR and then treated with antibodies that recognize BRCA1 (stained green).

## Summary

1. DNA replication is semiconservative; each strand of DNA serves as the template for synthesis of a complementary strand. The products of replication are two double-stranded daughter molecules consisting of one parental strand and one newly synthesized strand. DNA replication is bidirectional, proceeding in both directions from an origin in replication.
2. DNA polymerases add nucleotides to a growing DNA chain by catalyzing nucleotidyl-group–transfer reactions. DNA synthesis proceeds in the 5′ → 3′ direction. Errors in DNA synthesis are removed by the 3′ → 5′ exonuclease activity of the polymerase. Some DNA polymerases contain an additional 5′ → 3′ exonuclease activity.
3. The leading strand of DNA is synthesized continuously but the lagging strand is synthesized discontinuously producing Okazaki fragments. Synthesis of the leading strand and of each Okazaki fragment begins with an RNA primer. In *E. coli*, the primer is removed and replaced with DNA by the action of DNA polymerase I. The action of DNA ligase joins the separate fragments of the lagging strand.
4. The replisome is a complex protein machine that is assembled at the replication fork. The replisome contains two DNA polymerase molecules plus additional proteins such as helicase and primase.
5. Assembly of the replisome ensures simultaneous synthesis of two strands of DNA. In *E. coli*, a helicase unwinds the parental DNA and SSB binds to the single strands. The lagging-strand template is looped through the replisome so that the synthesis of both strands proceeds in the same direction as replication fork movement. Because it is part of the replisome, DNA polymerase is highly processive.
6. Initiation of DNA replication occurs at specific DNA sequences (e.g., *oriC* in *E. coli*) and depends on the presence of additional proteins. In bacteria, termination of DNA replication also occurs at specific sites and requires additional proteins.
7. Several new technologies such as PCR and DNA sequencing are based on an understanding of DNA replication.
8. Eukaryotic DNA replication resembles prokaryotic DNA replication except that eukaryotic chromosomes contain multiple origins of replication and eukaryotic Okazaki fragments are smaller. The slower movement of the replication fork in eukaryotes than in prokaryotes is due to the presence of nucleosomes.
9. DNA damaged by radiation or chemical agents can be repaired by direct-repair mechanisms or by a general excision-repair pathway. Excision-repair mechanisms also remove misincorporated nucleotides. Specific enzymes recognize damaged or misincorporated nucleotides.
10. Recombination can occur when a single strand of DNA exchanges with a homologous strand in double-stranded DNA producing a Holliday junction. Strand invasion is promoted by RecA in *E. coli*. Branch migration and resolution of Holliday junctions are catalyzed by RuvABC in *E. coli*.
11. Repair and recombination are similar processes and use many of the same enzymes. Defects in human genes required for repair and recombination cause sensitivity to ultraviolet light and increased risks of cancer.

## Problems

1. The chromosome of a certain bacterium is a circular, double-stranded DNA molecule of $5.2 \times 10^6$ base pairs. The chromosome contains one origin of replication and the rate of replication-fork movement is 1000 nucleotides per second.
   (a) Calculate the time required to replicate the chromosome.
   (b) Explain how the bacterial generation time can be as short as 25 minutes under extremely favorable conditions.
2. In many DNA viruses the viral genes can be divided into two nonoverlapping groups: early genes, whose products can be detected prior to replication of the viral genome; and late genes, whose products accumulate in the infected cell after replication of the viral genome. Some viruses, like bacteriophage T4 and T7, encode their own DNA polymerase enzymes. Would you expect the gene for T4 DNA polymerase to be in the early or late class? Why?
3. (a) Why does the addition of SSB to sequencing reactions often increase the yield of DNA?
   (b) What is the advantage of carrying out sequencing reactions at 65°C using a DNA polymerase isolated from bacteria that grow at high temperatures?
4. How does the use of an RNA primer rather than a DNA primer affect the fidelity of DNA replication in *E. coli*?
5. Both strands of DNA are synthesized in the 5′ → 3′ direction.
   (a) Draw a hypothetical reaction mechanism for synthesis of DNA in the 3′ → 5′ direction using a 5′-dNTP and a growing chain with a 5′-triphosphate group.
   (b) How would DNA synthesis be affected if the hypothetical enzyme had proofreading activity?
6. Ciprofloxacin is an antimicrobial used in the treatment of a wide variety of bacterial infections. One of the targets of ciprofloxacin in *E. coli* is topoisomerase II. Explain why the inhibition of topoisomerase II is an effective target to treat infections by *E. coli*.
7. The entire genome of the fruit fly *D. melanogaster* consists of $1.65 \times 10^8$ bp. If replication at a single replication fork occurs at the rate of 30 bp per second, calculate the minimum time required to replicate the entire genome if replication were initiated
   (a) at a single bidirectional origin
   (b) at 2000 bidirectional origins
   (c) In the early embryo, replication can require as few as 5 minutes. What is the minimum number of origins necessary to account for this replication time?
8. Ethyl methane sulfonate (EMS) is a reactive alkylating agent that ethylates the O-6 residue of guanine in DNA. If this modified G is not excised and replaced with a normal G, what would be the outcome of one round of DNA replication?
9. Why do cells exposed to visible light following irradiation with ultraviolet light have a greater survival rate than cells kept in the dark after irradiation with ultraviolet light?
10. *E. coli* uses several mechanisms to prevent the incorporation of the base uracil into DNA. First, the enzyme dUTPase, encoded by the *dut* gene, degrades dUTP. Second, the enzyme uracil *N*-glycosylase,

encoded by the *ung* gene, removes uracils that have found their way into DNA. The resulting apyrimidinic sites have to be repaired.

(a) If we examine the DNA from a strain carrying a mutation in the *dut* gene, what will we find?

(b) What if we examine the DNA from a strain in which both the *dut* and *ung* genes are mutated?

**11.** Explain why uracil *N*-glycosylase cannot repair the damage when 5-methylcytosine is deaminated to thymine.

**12.** Why are high rates of mutation observed in regions of DNA that contain methylcytosine?

**13.** Explain why the overall error rate for DNA replication in *E. coli* is approximately $10^{-9}$ although the rate of misincorporation by the replisome is about $10^{-5}$.

**14.** Will DNA repair in *E. coli* be dependent on the enzymatic cofactor $NAD^{\oplus}$?

**15.** Describe two methods that can be used to repair pyrimidine dimers in *E. coli*.

**16.** Damage to a single strand of DNA is readily repaired through a variety of mechanisms while damage to bases on both strands of DNA is more difficult for the cell to repair. Explain.

**17.** Why does homologous recombination occur only between DNAs with identical, or almost identical, sequences?

**18.** Why are two different DNA polymerase enzymes required to replicate the *E. coli* chromosome?

# Selected Readings


## General

Adams, R. L. P., Knowler, J. T., and Leader, D. P. (1992). *The Biochemistry of the Nucleic Acids*, 11th ed. (New York: Chapman and Hall).

Aladjem, M. I. (2007). Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nat. Rev. Genet.* 8:588–600.

Bentley, D. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59.

Kornberg, A., and Baker, T. (1992). *DNA Replication*, 2nd ed. (New York: W. H. Freeman).

## DNA Replication

Beese, L. S., Derbyshire, V., and Steitz, T. A. (1993). Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science* 260: 352–355.

Bell, S. P. (2002). The origin recognition complex: from simple origins to complex functions. *Genes & Devel.* 16:659–672.

Davey, M. J., Jeruzalmi, D., Kuriyan, J., and O'Donnell, M. (2002). Motors and switches: AAA + machines within the replisome. *Nat. Rev. Mol. Cell Biol.* 3:1–10.

Gilbert, D. M. (2001). Making sense of eukaryotic DNA replication origins. *Science* 294:96–100.

Keck, J. L., and Berger, J. M. (2001). Primus inter pares (First among equals). *Nat. Struct. Biol.* 8:2–4.

Kong, X.-P., Onrust, R., O'Donnell, M., and Kuriyan, J. (1992). Three-dimensional structure of the $\beta$ subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell* 69:425–437.

Kunkel, T. A., and Bebenek, K. (2000). DNA replication fidelity. *Annu. Rev. Biochem.* 69:497–529.

Marians, K. J. (1992). Prokaryotic DNA replication. *Annu. Rev. Biochem.* 61:673–719.

McHenry, C. S. (1991). DNA polymerase III holoenzyme. *J. Biol. Chem.* 266:19127–19130.

Meselson, M., and Stahl, F. W. (1958). The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 44:671–682.

Radman, M. (1998). DNA replication: one strand may be more equal. *Proc. Natl. Acad. Sci. USA* 95:9718–9719.

Waga, S., and Stillman, B. (1998). The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* 67:721–751.

Wake, R. G., and King, G. F. (1997). A tale of two terminators: crystal structures sharpen the debate on DNA replication fork arrest mechanisms. *Structure* 5:1–5.

Wyman, C., and Botchan, M. (1995). A familiar ring to DNA polymerase processivity. *Curr. Biol.* 5:334–337.

## DNA Repair

Echols, H., and Goodman, M. F. (1991). Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* 60:477–511.

Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell. Biol.* 9:958–970.

Kogoma, T. (1997). Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol. Mol. Biol. Rev.* 61:212–238.

McCullough, A. K., Dodson, M. L., and Lloyd, R. S. (1999). Initiation of base excision repair: glycosylase mechanisms and structures. *Annu. Rev. Biochem.* 68:255–285.

Mol, C. D., Parikh, S. S., Putnam, C. D., Lo, T. P., and Taylor, J. A. (1999). DNA repair mechanisms for the recognition and removal of damaged DNA bases. *Annu. Rev. Biophys. Biomol. Struct.* 28:101–128.

Tainer, J. A., Thayer, M. M., and Cunningham, R. P. (1995). DNA repair proteins. *Curr. Opin. Struct. Biol.* 5:20–26.

Yang, W. (2000). Structure and function of mismatch repair proteins. *Mutat. Res.* 460:245–256.

## Recombination

Ortiz-Lombardia, M., González, A., Ertja, R., Aymami, J., Azorin, F., and Coll, M. (1999). Crystal structure of a Holliday junction. *Nat. Struct. Biol.* 6:913–917.

Rafferty, J. B., Sedelnikove, S. E., Hargreaves, D., Artmiuk, P. J., Baker, P. J., Sharples, G. J., Mahdi, A. A., Lloyd, R. G., and Rice, D. W. (1996). Crystal structure of DNA recombination protein RuvA and a model for its binding to the Holliday junction. *Science* 274:415–421.

Rao, B. J., Chiu, S. K., Bazemore, L. R., Reddy, G., and Radding, C. M. (1995). How specific is the first recognition step of homologous recombination? *Trends Biochem. Sci.* 20:109–113.

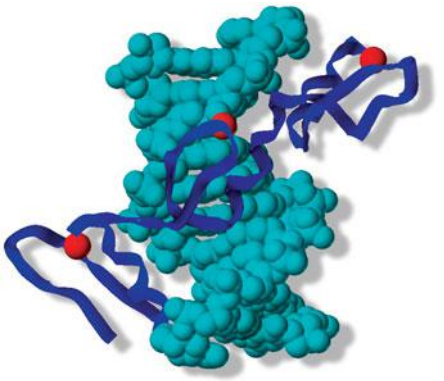
West, S. C. (1996). The RuvABC proteins and Holliday junction processing in *Escherichia coli*. *J. Bacteriol.* 178:1237–1241.

West, S. C. (1997). Processing of recombination intermediates by the RuvABC proteins. *Annu. Rev. Genet.* 31:213–244.

West, S. C. (2003). Molecular views of recombination proteins and their control. *Nat. Rev. Mol. Cell Biol.* 4:1–11.

White, M. F., Giraud-Panis, M.-J. E., Pöhler, J. R. G., and Lilley, D. M. J. (1997). Recognition and manipulation of branched DNA structure by junction-resolving enzymes. *J. Mol. Biol.* 269:647–664.

Wuethrich, B. (1998). Why sex? *Science* 281:1980–1982.

# Transcription and RNA Processing

As we have seen, the structure of DNA proposed by Watson and Crick in 1953 immediately suggested a means of replicating DNA to transfer genetic information from one generation to the next but it did not reveal how an organism makes use of the information stored in its genetic material.

Based on studies of the bread mold *Neurospora crassa*, George Beadle and Edward Tatum proposed that a single unit of heredity, or gene, directed the production of a single enzyme. A full demonstration of the relationship between genes and proteins came in 1956 when Vernon Ingram showed that hemoglobin from patients with the heritable disease sickle-cell anemia differed from normal hemoglobin by the replacement of a single amino acid. Ingram's results indicated that genetic changes can manifest themselves as changes in the amino acid sequence of a protein. By extension, the information contained in the genome must specify the primary structure of each protein in an organism.

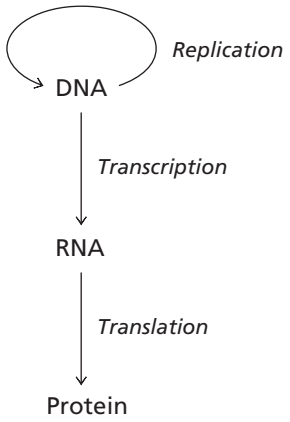
We define a **gene** as a DNA sequence that is transcribed. This definition includes genes that do not encode proteins (not all transcripts are messenger RNA). The definition normally excludes regions of the genome that control transcription but are not themselves transcribed. We will encounter some exceptions to our definition of a gene—surprisingly, there is no definition that is entirely satisfactory.

Many prokaryotic genomes contain several thousand genes, although some simple bacteria have only 500 to 600 genes. Most of these are "housekeeping genes" that encode proteins or RNA molecules that are essential for the normal activities of all living cells. For example, the enzymes involved in the basic metabolic processes of glycolysis and the synthesis of amino acids and DNA are encoded by such housekeeping genes, as are transfer RNAs and ribosomal RNAs. The number of housekeeping genes in unicellular eukaryotes, such as yeast and some algae, is similar to the number in complex prokaryotes.

*"This fraction (which we shall designate "messenger RNA" or M-RNA) amounts to only about 3% of the total RNA. . . . The property attributed to the structural messenger of being an unstable intermediate is one of the most specific and novel implications of this scheme. . . . This leads to a new concept of the mechanism of information transfer, where the protein synthesizing centers (ribosomes) play the role of non-specific constituents which can synthesize different proteins, according to specific instructions which they receive from the genes through M-RNA."*

—François Jacob and Jacques Monod, 1961

**Top:** A portion of the mouse transcription factor Zif268 (dark blue) bound to DNA (light blue). Side chains from three zinc-containing domains interact with base pairs in DNA.




▲ **Figure 21.1**
**Biological information flow**. The normal flow of biological information is from DNA to RNA to protein.

## KEY CONCEPT

**Before a cell can access the genetic information stored in its DNA, the DNA must be transcribed into RNA.**

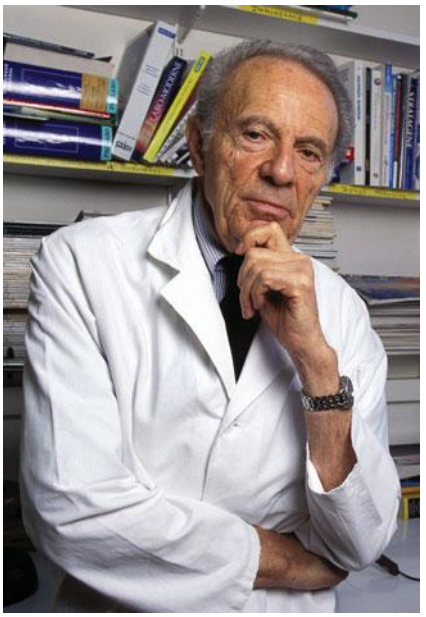
In addition to housekeeping genes, all cells contain genes that are expressed only in special circumstances, such as during cell division. Multicellular organisms also contain genes that are expressed only in certain types of cells. For example, all cells in a maple tree contain the genes for the enzymes that synthesize chlorophyll but these genes are expressed only in cells that are exposed to light, such as cells on the surface of a leaf. Similarly, all cells in mammals contain insulin genes, but only certain pancreatic cells produce insulin. The total number of genes in multicellular eukaryotes ranges from as few as 15,000 in *Drosophila melanogaster* to more than 50,000 in some other animals.

In this chapter and the next, we will examine how the information stored in DNA directs the synthesis of proteins. A general outline of this flow of information is summarized in Figure 21.1. In this chapter, we describe transcription (the process where information stored in DNA is copied into RNA thereby making it available for either protein synthesis or other cellular functions) and RNA processing (the post-transcriptional modification of RNA molecules). We also briefly examine how gene expression is regulated by factors that affect the initiation of transcription. In Chapter 22, we will examine translation (the process where information coded in mRNA molecules directs the synthesis of individual proteins).

One feature of the complete pathway outlined in Figure 21.1 is that it is irreversible. In particular, the information contained in the amino acid sequence of a protein cannot be translated back into nucleic acid. This irreversibility of information flow is known as the "Central Dogma" of molecular biology and was predicted by Francis Crick in 1958, many years before the mechanisms of transcription and translation were worked out (see Section 1.1). The original version of the Central Dogma did not rule out information flow from RNA to DNA. Such a pathway was eventually discovered in retrovirus-infected cells; it is known as reverse transcription.

## 21.1 Types of RNA

Several classes of RNA molecules have been discovered. *Transfer RNA* (tRNA) carries amino acids to the translation machinery. *Ribosomal RNA* (rRNA) makes up much of the ribosome. A third class of RNA is *messenger RNA* (mRNA), whose discovery was due largely to the work of François Jacob, Jacques Monod, and their collaborators at the Pasteur Institute in Paris. In the early 1960s, these researchers showed that ribosomes participate in protein synthesis by translating unstable RNA molecules (mRNA). Jacob and Monod also discovered that the sequence of an mRNA molecule is complementary to a segment of one of the strands of DNA. A fourth class of RNA consists of small RNA molecules that participate in various metabolic events, including RNA processing. Many of these small RNA molecules have catalytic activity. Some of these small RNAs are regulatory molecules that can bind specifically to mRNAs and down-regulate that messenger and the protein it encodes.



▲ **François Jacob (1920–).** Jacob and Monod received the Nobel Prize in Physiology or Medicine in 1965 for their work on the genetic control of enzyme synthesis.

A large percentage of the total RNA in a cell is ribosomal RNA, and only a small percentage is mRNA. But if we compare the rates at which the cell synthesizes RNA rather than the steady state levels of RNA, we see a different picture (Table 21.1). Even though mRNA accounts for only 3% of the total RNA in *Escherichia coli*, the bacterium devotes almost one-third of its capacity for RNA synthesis to the production of mRNA. This value may increase to about 60% when the cell is growing slowly and does not need to replace ribosomes and tRNA. The discrepancy between steady state levels of various RNA molecules and the rates at which they are synthesized can be explained by the differing stabilities of the RNA molecules: rRNA and tRNA molecules are extremely stable, whereas mRNA is rapidly degraded after translation. Half of all newly synthesized mRNA is degraded by nucleases within three minutes in bacterial cells. In eukaryotes, the average half-life of mRNA is about ten times longer. The relatively high stability of eukaryotic mRNA results from processing events that prevent eukaryotic mRNA from being degraded during transport from the nucleus, where transcription occurs, to the cytoplasm, where translation occurs.

**Table 21.1 The RNA content of an *E. coli* cell**

| Type | Steady state level | Synthetic type capacity[a] |
|---|---|---|
| rRNA | 83% | 58% |
| tRNA | 14% | 10% |
| mRNA | 03% | 32% |
| RNA primers[b] | <1% | <1% |
| Other RNA molecules[c] | <1% | <1% |

[a]Relative amount of each type of RNA being synthesized at any instant.
[b]RNA primers are those used in DNA replication; they are not synthesized by RNA polymerase.
[c]Other RNA molecules include several RNA enzymes, such as the RNA component of RNase P.

[Adapted from Bremer, H., and Dennis, P. P. (1987). Modulation of chemical composition and other parameters of the cell by growth rate. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, Vol. 2, F. C. Neidhardt, ed. (Washington, DC: American Society for Microbiology), pp. 1527–1542.]

## 21.2 RNA Polymerase

About the time that mRNA was identified, researchers in several laboratories independently discovered an enzyme that catalyzes the synthesis of RNA when provided with ATP, UTP, GTP, CTP, and a template DNA molecule. The newly discovered enzyme was RNA polymerase. This enzyme catalyzes DNA-directed RNA synthesis, or **transcription**.

RNA polymerase was initially identified by its ability to catalyze polymerization of ribonucleotides but further study of the enzyme revealed that it does much more. RNA polymerase is the core of a larger transcription complex just as DNA polymerase is the core of a larger replication complex (Section 20.4). This complex assembles at one end of a gene when transcription is initiated. During initiation, the template DNA partially unwinds and a short piece of RNA is synthesized. In the elongation phase of transcription, RNA polymerase catalyzes the processive elongation of the RNA chain while the DNA is continuously unwound and rewound. Finally, the transcription complex responds to specific transcription termination signals and disassembles.

Although the composition of the transcription complex varies considerably among different organisms, all transcription complexes catalyze essentially the same types of reactions. We introduce the general process of transcription by discussing the reactions catalyzed by the well-characterized transcription complex in *E. coli*. The more complicated eukaryotic transcription complexes are presented in Section 21.5.

### A. RNA Polymerase Is an Oligomeric Protein

Core RNA polymerase is isolated from *E. coli* cells as a multimeric protein with four different types of subunits (Table 21.2). Five of these subunits combine with a stoichiometry of $\alpha_2\beta\beta'\omega$ to form the core enzyme that participates in many of the transcription reactions. The large $\beta$ and $\beta'$ subunits make up the active site of the enzyme; the $\beta'$ subunit contributes to DNA binding, whereas the $\beta$ subunit contains part of the polymerase active site. The $\alpha$ subunits are the scaffold for assembly of the other subunits and they also interact with many proteins that regulate transcription. The role of the small $\omega$ subunit is not well characterized.

The structure of RNA polymerase holoenzyme from the bacterium *Thermus aquaticus* complexed with DNA is shown in Figure 21.2. The $\beta$ and $\beta'$ subunits form a large groove at one end. This is where DNA binds and polymerization takes place. The groove is large enough to accommodate about 16 base pairs of double-stranded B-DNA and is shaped like the DNA-binding sites of DNA polymerases (such as DNA polymerase I; Figure 20.12). The pair of $\alpha$ subunits is located at the "back end" of the molecule. This region also contacts DNA when the polymerase is actively transcribing a gene. The $\omega$ subunit is bound to the outer surface of the $\beta'$ subunit. We will see later that various transcription factors interact with RNA polymerase by binding to the $\alpha$ subunits.

**Table 21.2 Subunits of *E. coli* RNA polymerase holoenzyme**

| Subunit | $M_r$ |
|---|---|
| $\beta'$[a] | 155,600 |
| $\beta$ | 150,600 |
| $\sigma$[b] | 70,300[c] |
| $\alpha$ | 36,500 |
| $\omega$ | 11,000 |

[a]The $\beta$ and $\beta'$ subunits are unrelated despite the similarity of their names.
[b]This subunit is not part of the core RNA polymerase.
[c]The molecular weight given is for the $\sigma$ subunit found in the most common form of the holoenzyme.

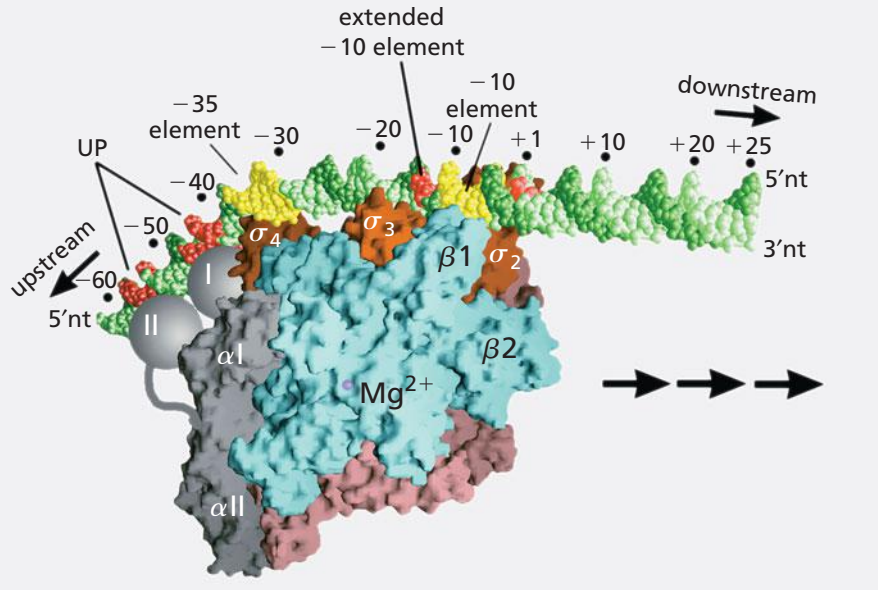



**Figure 21.2 ►**
***Thermus aquaticus (taq)* RNA polymerase holoenzyme/promoter DNA closed complex.** The template strand is dark green and the coding strand is light green; both the –10 and –35 elements are yellow. The transcription start site is shown in red and labeled +1. Once the open complex forms, then transcription will proceed downstream, to the right as shown by the arrows. The $\alpha$ and $\omega$ subunits are shown in gray; the $\beta$ subunit is cyan, while the $\beta''$ subunit is pink. The $\sigma$ subunit is orange.

The $\sigma$ subunit of the holoenzyme plays an important role in transcription initiation. Bacteria contain several different types of $\sigma$ subunits. The major form of the holoenzyme in *E. coli* contains the subunit $\sigma^{70}$ ($M_r$ 70,300). The $\sigma$ subunits contact DNA during transcription initiation and bind to the core enzyme in the region of the $\omega$ subunit. The overall dimensions of RNA polymerase are 10 × 10 × 16 nm. This makes it considerably larger than a nucleosome but smaller than a ribosome or a replisome.

## B. The Chain Elongation Reaction

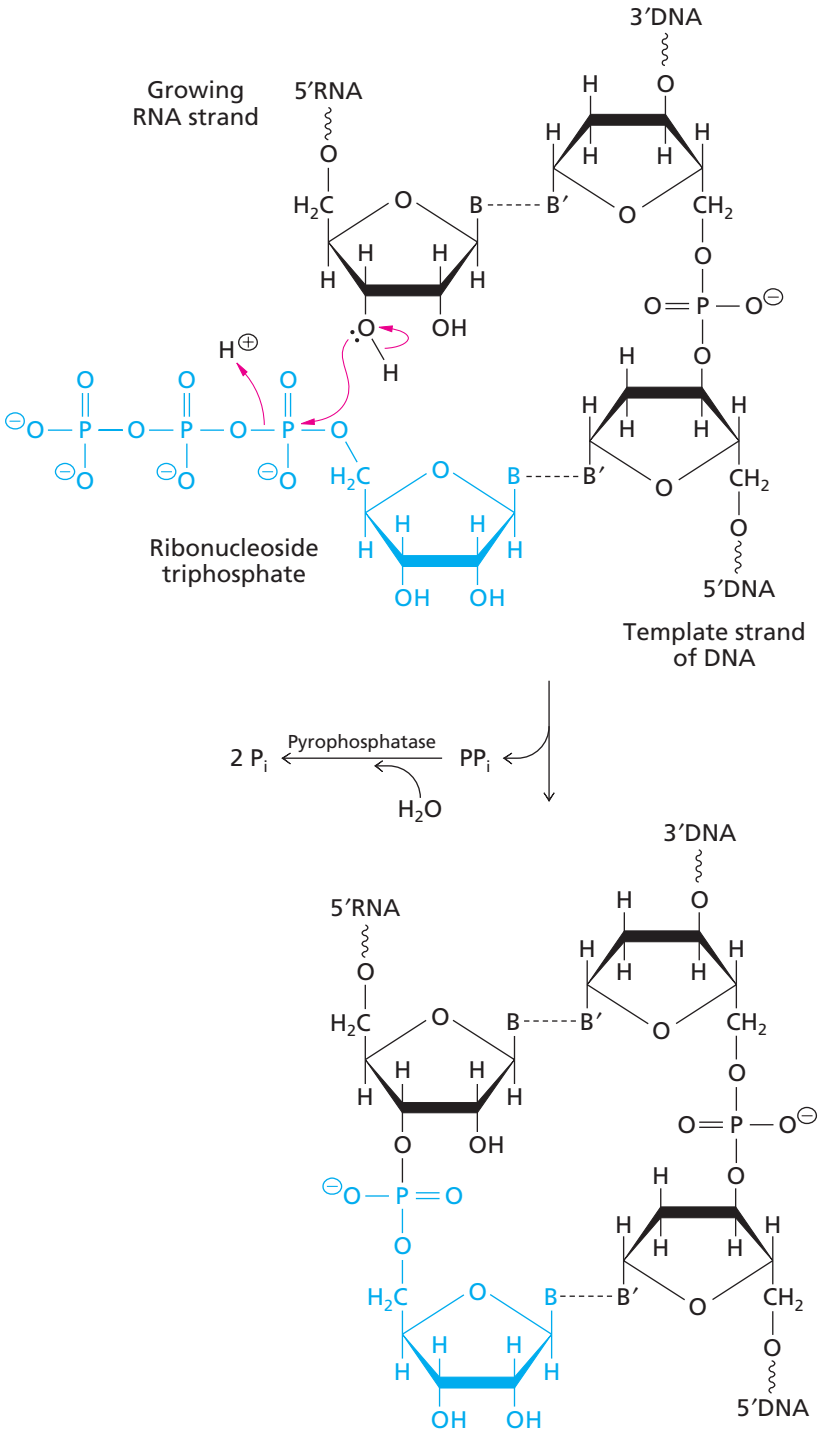
RNA polymerase catalyzes chain elongation by a mechanism almost identical to that used by DNA polymerase (Figure 20.6). Part of the growing RNA chain is base-paired to the DNA template strand, and incoming ribonucleoside triphosphates are tested in the active site of the polymerase for correct hydrogen bonding to the next unpaired nucleotide of the template strand. When the incoming nucleotide forms correct hydrogen bonds, RNA polymerase catalyzes a nucleotidyl-group–transfer reaction, resulting in formation of a new phosphodiester linkage and the release of pyrophosphate (Figure 21.3).

Like DNA polymerase III, RNA polymerase catalyzes polymerization in the 5′ ⟶ 3′ direction and is highly processive when it is bound to DNA as part of a transcription complex. The overall reaction of RNA synthesis can be summarized as

$$\text{RNA}_n - \text{OH} + \text{NTP} \longrightarrow \text{RNA}_{n+1} - \text{OH} + \text{PP}_i \qquad \textbf{(21.1)}$$

The Gibbs free energy change for this reaction is highly favorable because of the high concentration of NTPs relative to RNA. In addition, the RNA polymerase reaction like the DNA polymerase reaction is thermodynamically assisted by the subsequent hydrolysis of pyrophosphate inside the cell. Thus, two phosphoanhydride linkages are expended for every nucleotide added to the growing chain.

RNA polymerase differs from DNA polymerase in using ribonucleoside triphosphates (UTP, GTP, ATP, and CTP) as substrates rather than deoxyribonucleoside triphosphates (dTTP, dGTP, dATP, and dCTP). Another difference is that the growing RNA strand only interacts with the template strand over a short distance (see below). The final product of transcription is single-stranded RNA, not an RNA-DNA duplex. Transcription is much slower than DNA replication. In *E. coli*, the rate of transcription ranges from 30 to 85 nucleotides per second, or less than one-tenth the rate of DNA replication.




◀ **Figure 21.3**
**Reaction catalyzed by RNA polymerase.** When an incoming ribonucleoside triphosphate correctly pairs with the next unpaired nucleotide on the DNA template strand, RNA polymerase catalyzes a nucleophilic attack by the 3′-hydroxyl group of the growing RNA strand on the $\alpha$-phosphorus atom of the incoming ribonucleoside triphosphate. As a result, a phosphodiester forms and pyrophosphate is released. The subsequent hydrolysis of pyrophosphate catalyzed by pyrophosphatase provides additional thermodynamic driving force for the reaction. (B and B′ represent complementary bases, and hydrogen bonding between bases is indicated by a single dashed line.)

RNA polymerase catalyzes the formation of a new phosphodiester linkage only when the incoming ribonucleoside triphosphate fits the active site of the enzyme precisely. A precise fit requires base stacking and appropriate hydrogen bonding between the incoming ribonucleoside triphosphate and the template nucleotide.

Despite the requirement for an accurate fit, RNA polymerase does make mistakes. The error rate of RNA synthesis is $10^{-6}$ (one mistake for every 1 million nucleotides incorporated). This rate is higher than the overall error rate of DNA synthesis because, in contrast to most DNA polymerases, RNA polymerase does not possess an exonuclease proofreading activity. Extreme precision in DNA replication is necessary to minimize mutations that could be passed on to progeny but accuracy in RNA synthesis is not as crucial to survival.

## 21.3 Transcription Initiation

The elongation reactions of RNA synthesis are preceded by a distinct initiation step in which a transcription complex assembles at an initiation site and a short stretch of RNA is synthesized. The regions of DNA that serve as sites of transcription initiation are called **promoters**. In bacteria, several genes are often co-transcribed from a single promoter; such a transcription unit is called an **operon**. In eukaryotic cells, each gene usually has its own promoter. There are hundreds of promoters in bacterial cells and thousands in eukaryotic cells.

The frequency of transcription initiation at any given promoter is usually related to the need for that gene's particular product. For example, in cells that are dividing rapidly, the genes for ribosomal RNA are usually transcribed frequently. Every few seconds a new transcription complex begins transcribing at the promoter. This process gives rise to structures such as those seen in Figure 21.4 showing multiple transcription complexes on one *E. coli* ribosomal RNA operon. Transcripts of increasing length are arrayed along the genes because many RNA polymerases transcribe the genes at the same time. In contrast, some bacterial genes are transcribed only once every two generations. In these cases initiation may occur only once every few hours. (Outside of the laboratory, the average generation time of most bacteria is many hours.)

### A. Genes Have a 5′ → 3′ Orientation

In Section 19.2A, we introduced the convention that single-strand nucleic acid sequences are written from left to right in the 5′ → 3′ direction. When a sequence of double-stranded DNA is displayed, the sequence of the top strand is written 5′ → 3′ and the sequence of the bottom, antiparallel, strand is written 3′ → 5′ (left to right).

Since our operational definition of a gene is a DNA sequence that is transcribed, a gene begins at the point where transcription starts (designated +1) and ends at the point where transcription terminates. The beginning of a gene is called the 5′ end, corresponding to the convention for writing sequences. Moving along a gene in the 5′ → 3′ direction is described as moving "downstream" and moving in the 3′ → 5′ direction is moving "upstream." RNA polymerization proceeds in the 5′ → 3′ direction. Consequently, in accordance with the convention for writing DNA sequences, the transcription start site of a gene is shown on the left of a diagram of double-stranded DNA and the termination site is on the right. The top strand is often called the **coding strand** because its sequence corresponds to the DNA version of the mRNA that encodes the amino acid sequence of a protein. The bottom strand is called the **template strand** because it is the strand used as a template for RNA synthesis (Figure 21.5). Alternatively, the top strand may be called the **sense strand** to indicate that translating ribosomes attempting to "read" the codons in an mRNA with this sequence will make the correct protein. Therefore the bottom strand becomes the **antisense strand** because an mRNA with this sequence will not make the correct protein. Note that RNA is synthesized in

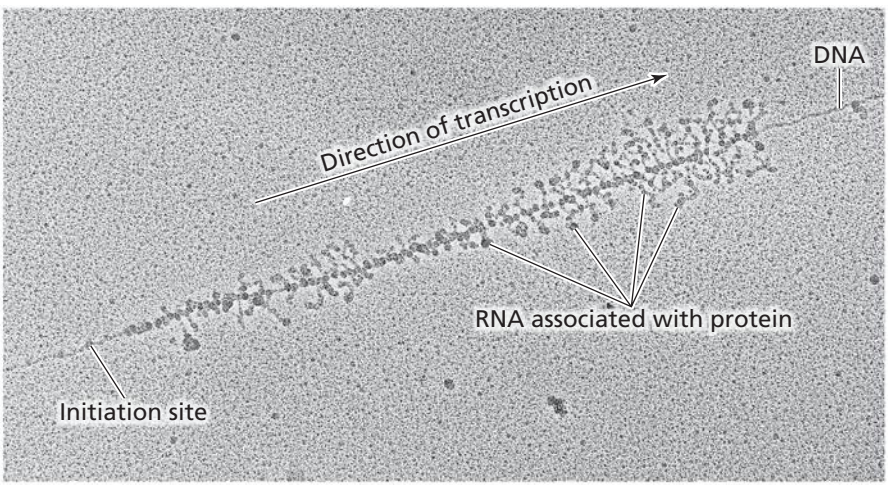
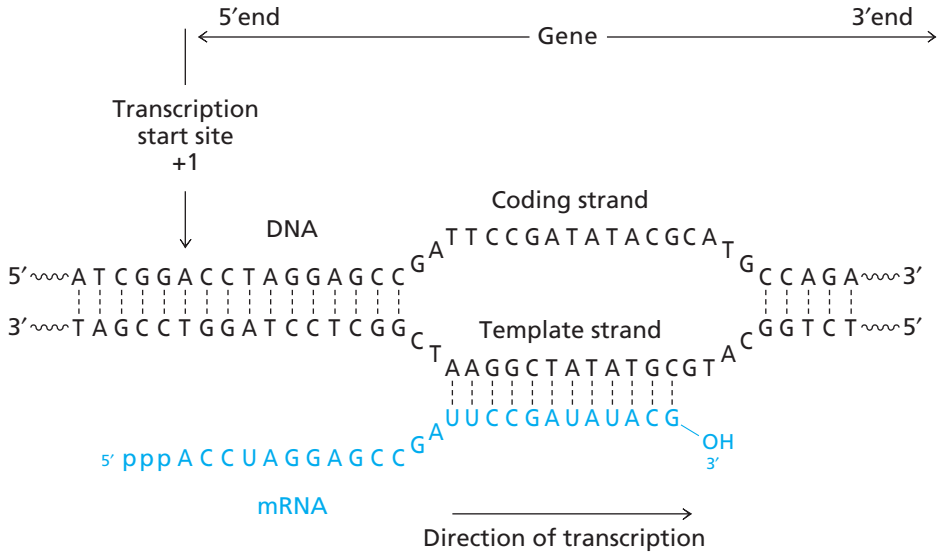



**Figure 21.4 ▸**
**Transcription of *E. coli* ribosomal RNA genes.** The genes are being transcribed from left to right. The nascent rRNA product associates with proteins and is processed by nucleolytic cleavage before transcription is complete.




◀ **Figure 21.5**
**Orientation of a gene.** The sequence of a hypothetical gene and the RNA transcribed from it are shown. By convention, the gene is said to be transcribed from the 5′ end to the 3′ end but the template strand of DNA is copied from the 3′ end to the 5′ end. Growth of the ribonucleotide chain proceeds 5′ → 3′.

the 5′ → 3′ direction but the template strand is copied from its 3′ end to its 5′ end. Also note that the RNA product is identical in sequence to the coding strand except that U replaces T.

## B. The Transcription Complex Assembles at a Promoter

A transcription complex forms when one or more proteins bind to the promoter sequence and also to RNA polymerase. These DNA-binding proteins direct RNA polymerase to the promoter site. In bacteria, the $\sigma$ subunit of RNA polymerase is required for promoter recognition and formation of the transcription complex.

The nucleotide sequence of a promoter is one of the most important factors affecting the frequency of transcription of a gene. Soon after the development of DNA-sequencing technology, many different promoters were examined. The start sites, the points at which transcription actually begins, were identified, and the regions upstream of these sites were sequenced to learn whether the promoter sequences of different genes were similar. This analysis revealed a common pattern called a **consensus sequence**—a hypothetical sequence made up of the nucleotides found most often in each position.

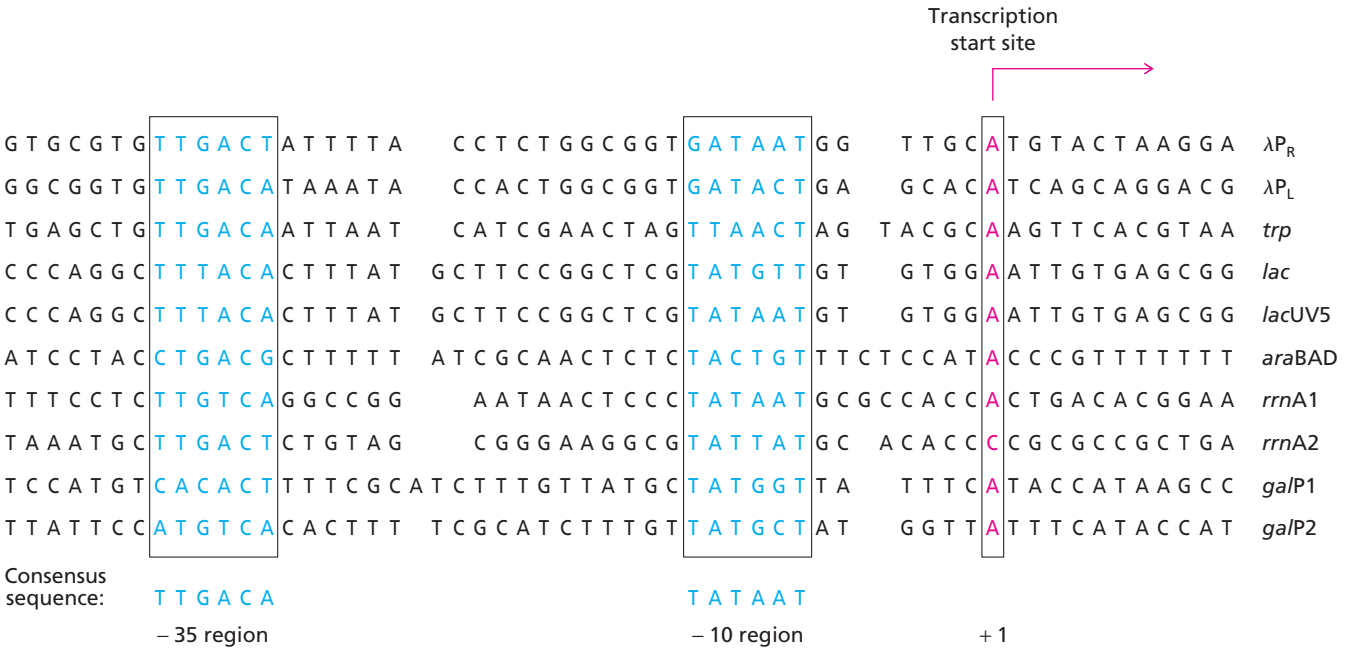
The consensus sequence of the most common type of promoter in *E. coli* is shown in Figure 21.6. This promoter is bipartite, which means that there are two separate regions of sequence similarity. The first region is 10 bp upstream of the transcription start site and is rich in A/T base pairs. The consensus sequence is TATAAT. The second part of the promoter sequence is centered approximately 35 bp upstream of the start site. The consensus sequence in this region is TTGACA. The average distance between the two parts of the promoter is 17 bp.

The −10 region is known as a **TATA box**, and the −35 region is simply referred to as the **–35 region**. Together, the two regions define the promoter for the *E. coli* holoenzyme containing $\sigma^{70}$, the most common $\sigma$ subunit in *E. coli* cells. The $\sigma^{70}$-containing holoenzyme binds specifically to sequences that resemble the consensus sequence. Other *E. coli* $\sigma$ subunits recognize and bind to promoters with quite different consensus sequences (Table 21.3). Orthologous $\sigma$ subunits from other prokaryotic species may recognize different promoter consensus sequences.

A consensus sequence is not an exact sequence but indicates the nucleotides most commonly found at each position. Very few promoters match their consensus sequence exactly. In some cases, the match is quite poor, with G or C found at positions normally occupied by A or T. Such promoters are known as weak promoters and are usually associated with genes that are transcribed infrequently. Strong promoters, such as the promoters for ribosomal RNA operons, resemble the consensus sequence quite closely. These operons are transcribed very efficiently. Observations such as these suggest that the consensus sequence describes the most efficient promoter sequence for the RNA polymerase holoenzyme.

## KEY CONCEPT

**Promoter sequences contain the information that instructs transcription complexes: "Initiate a transcript here."**




▲ **Figure 21.6**

**Promoter sequences from ten bacteriophage and bacterial genes.** All these promoter sequences are recognized by the $\sigma^{70}$ subunit in *E. coli*. The nucleotide sequences are aligned so that their +1, –10, and –35 regions are in register. Note the degree of sequence variation at each position. The consensus sequence was derived from a much larger database of more than 300 well-characterized promoters.

The promoter sequence of each gene has likely been optimized by natural selection to fit the requirements of the cell. An inefficient promoter is ideal for a gene whose product is not needed in large quantities whereas an efficient promoter is necessary for producing large amounts of a gene product.

## C. The $\sigma$ Subunit Recognizes the Promoter

The effect of $\sigma$ subunits, also called $\sigma$ factors, on promoter recognition can best be explained by comparing the DNA-binding properties of core polymerase versus the holoenzyme containing $\sigma^{70}$. The core polymerase, which lacks a $\sigma$ subunit, binds to DNA nonspecifically; it has no greater affinity for promoters than for any other DNA sequence (the association constant, $K_a$, is approximately $10^{10}\ M^{-1}$). Once formed, this DNA-protein complex dissociates slowly ($t_{1/2} \approx 60$ minutes). In contrast, the holoenzyme, which contains the $\sigma^{70}$ subunit, binds more tightly to promoter sequences ($K_a \approx 2 \times 10^{11}\ M^{-1}$) than the core polymerase and forms more stable complexes ($t_{1/2} \approx 2$ to 3 hours). Although the holoenzyme binds preferentially to promoter sequences, it also has appreciable affinity for the rest of the DNA in a cell

**Table 21.3 *E. coli* $\sigma$ subunits**

| Subunit | Gene | Genes transcribed | Consensus | |
|---|---|---|---|---|
| | | | –35 | –10 |
| $\sigma^{70}$ | *rpoD* | Many | TTGACA | TATAAT |
| $\sigma^{54}$ | *rpoN* | Nitrogen metabolism | None | CTGGCACNNNNNTTGCA[a] |
| $\sigma^{38}$ | *rpoS* | Stationary phase | ? | TATAAT |
| $\sigma^{28}$ | *flaI* | Flagellar synthesis and chemotaxis | TAAA | GCCGATAA |
| $\sigma^{32}$ | *rpoH* | Heat shock | CTTGAA | CCCATNTA[a] |
| $\sigma^{gp55}$ | gene 55 | Bacteriophage T4 | None | TATAAATA |

[a]N represents any nucleotide.

($K_a \approx 5 \times 10^6\ M^{-1}$). The complex formed by nonspecific binding of the holoenzyme to DNA dissociates rapidly ($t_{1/2} \approx 3$ seconds). These binding parameters reveal the functions of the $\sigma^{70}$ subunit. One of the roles of $\sigma^{70}$ is to decrease the affinity of the core polymerase for nonpromoter sequences. Another equally important role is to increase the affinity of the core polymerase for specific promoter sequences.

The association constants do not tell us how the RNA polymerase holoenzyme finds the promoter. We might expect the holoenzyme to search for the promoter by continuously binding and dissociating until it encounters a promoter sequence. Such binding would be a second-order reaction, and its rate would be limited by the rate at which the holoenzyme diffuses in three dimensions. However, promoter binding is 100 times faster than the maximum theoretical value for a diffusion-limited second-order reaction. This remarkable rate is achieved by one-dimensional diffusion of RNA polymerase along the length of the DNA molecule. During the short period of time that the enzyme is bound nonspecifically, it can scan 2000 bp in its search for a promoter sequence. Several other sequence-specific DNA-binding proteins, such as restriction enzymes (Section 19.6C), locate their binding sites in a similar manner.

**The binding properties of RNA polymerase tell us that many RNA polymerase molecules will be located on random stretches of DNA that may, or may not, resemble a promoter sequence.**

## D. RNA Polymerase Changes Conformation

Initiation of transcription is slow, even though the holoenzyme searches for and binds to the promoter very quickly. In fact, initiation is often the rate limiting step in transcription because it requires unwinding of the DNA helix and synthesis of a short stretch of RNA that serves as a primer for subsequent chain elongation. During DNA replication these steps are carried out by a helicase and a primase but in transcription these steps are carried out by the RNA polymerase holoenzyme itself. Unlike DNA polymerases, RNA polymerases can initiate polynucleotide synthesis on their own in the presence of initiation factors such as $\sigma^{70}$ (when a DNA template and rNTPs are available as substrates).

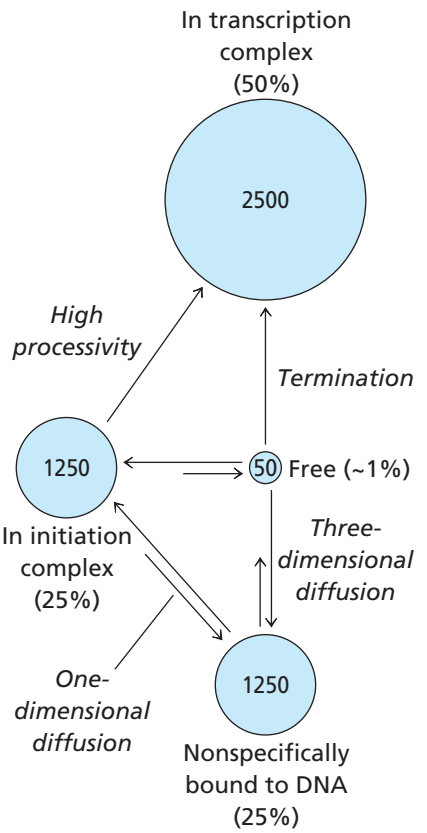
The unwinding of DNA at the initiation site is an example of a conformational change in which RNA polymerase (R) and the promoter (P) shift from a closed complex ($RP_c$) to an open complex ($RP_o$). In the closed complex, the DNA is double-stranded. In the open complex, 18 bp of DNA are unwound, forming a transcription bubble. Formation of the open complex is usually the slowest step of the initiation events.

Once the open complex forms, the template strand is positioned at the polymerization site of the enzyme. In the next step, a phosphodiester linkage forms between two ribonucleoside triphosphates that have diffused into the active site and formed hydrogen bonds with the +1 and +2 nucleotides of the template strand. This initiation reaction is slower than the analogous polymerization reaction during chain elongation where one of the substrates (the growing RNA chain) is held in place by the formation of a short RNA-DNA helix.

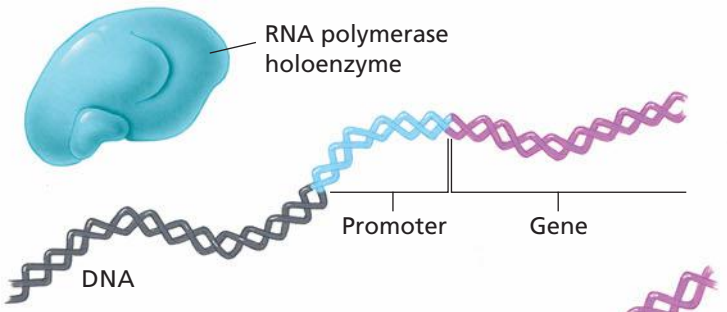
Additional nucleotides are then added to the dinucleotide to create a short RNA that is paired with the template strand. When this RNA is approximately ten nucleotides long, the RNA polymerase holoenzyme undergoes a transition from the initiation to the elongation mode, and the transcription complex moves away from the promoter along the DNA template. This step is called promoter clearance. The initiation reactions can be summarized as

$$R + P \overset{K_{assoc}}{\longleftrightarrow} RP_c \underset{\text{conformational change}}{\longrightarrow} RP_o \underset{\text{promoter clearance}}{\longrightarrow} \qquad (21.2)$$

As noted earlier, the holoenzyme containing $\sigma$ factor has a much greater affinity for the promoter sequence than for any other DNA sequence. Because of this tight binding, it resists moving away from the initiation site. However, during elongation, the core polymerase binds nonspecifically to all DNA sequences to form a highly processive complex. The transition from initiation to chain elongation is associated with a conformational change in the holoenzyme that causes release of the $\sigma$ subunit. Without $\sigma$, the enzyme no longer binds specifically to the promoter and is able to leave, or exit, the initiation site. At this time, several accessory proteins bind to the




▲ **RNA polymerase distribution.** Estimate of the distribution of the approximately 5000 RNA polymerase molecules typically found in an *E. coli* cell. Very few molecules are free in the cytosol, yet only half of all RNA polymerases are actively transcribing.

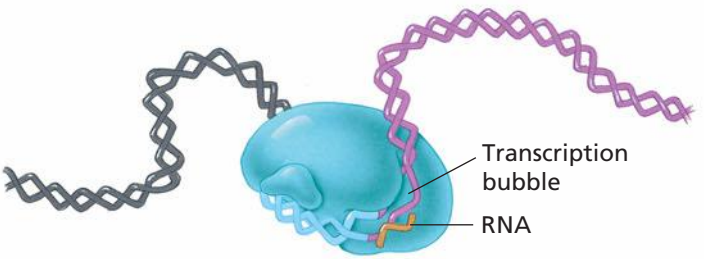



◀ **Figure 21.7**
**Initiation of transcription in *E. coli.***

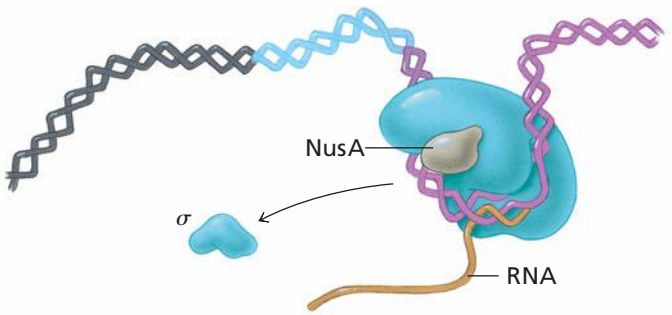
(**a**) RNA polymerase holoenzyme binds nonspecifically to DNA.

(**b**) The holoenzyme conducts a one-dimensional search for a promoter.

(**c**) When a promoter is found, the holoenzyme and the promoter form a closed complex.




(**d**) A conformational change from the closed complex to an open complex produces a transcription bubble at the initiation site. A short stretch of RNA is then synthesized.




(**e**) The $\sigma$ subunit dissociates from the core enzyme, and RNA polymerase clears the promoter. Accessory proteins, including NusA, bind to the polymerase.

core polymerase to create the complete protein machine required for RNA chain elongation. The binding of one of these accessory proteins, NusA, helps convert RNA polymerase to the elongation form. The elongation complex is responsible for most of the synthesis of RNA. NusA also interacts with other accessory proteins and plays a role in termination. Transcription initiation in *E. coli* is summarized in Figure 21.7.

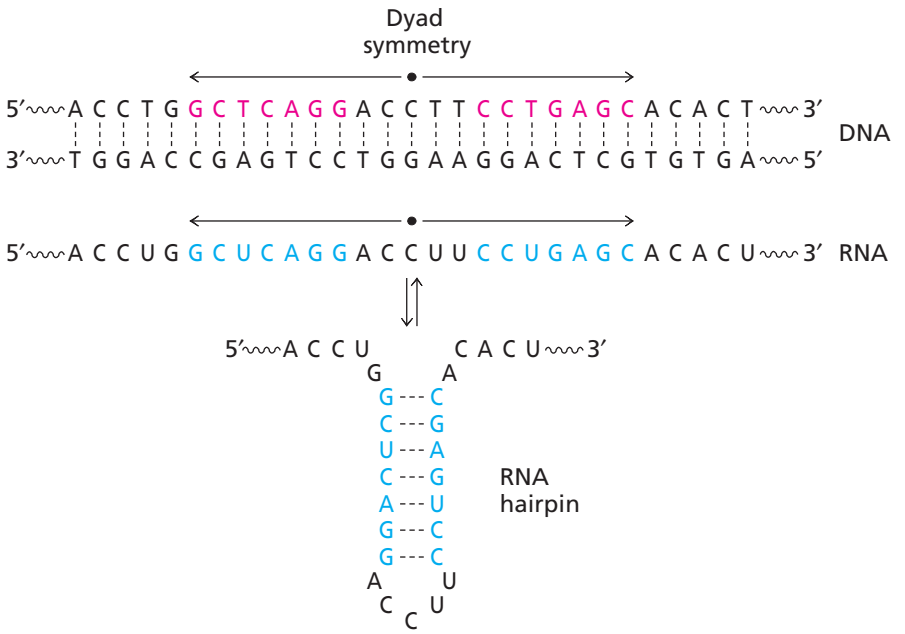
## 21.4 Transcription Termination

Only certain regions of DNA are specifically transcribed. Transcription complexes assemble at promoters and, in bacteria, disassemble at the 3′ end of genes at specific sequences called **termination sequences**. There are two types of transcription termination sequences. The simplest form of termination occurs at certain DNA sequences where the elongation complex is unstable, and the transcription complex spontaneously disassembles. The other type of termination requires a specific protein named *rho* that facilitates disassembly of the transcription complex, template, and mRNA.

Transcription termination often occurs near **pause sites**. These are regions of the gene where the rate of elongation slows down or stops temporarily. For example, because it is more difficult to melt G/C base pairs than it is to melt A/T base pairs, a transcription complex pauses when it encounters a GC-rich region.

Pausing is exaggerated at sites where the DNA sequence is palindromic, or has dyad symmetry (Section 19.6C). When the DNA is transcribed, the newly synthesized RNA can form a hairpin (Figure 21.8). (A three-dimensional representation of such a structure is shown in Figure 19.21.) Formation of an RNA hairpin may destabilize the RNA-DNA hybrid in the elongation complex by prematurely stripping off part of the newly transcribed RNA. This partial disruption of the transcription bubble probably causes the transcription complex to cease elongation until the hybrid re-forms. NusA increases pausing at palindromic sites, perhaps by stabilizing the hairpin. The transcription complex may pause for 10 seconds to 30 minutes, depending on the structure of the hairpin.

Some of the strong pause sites in *E. coli* are termination sequences. Such termination sites are found at the 3′ end of a gene beyond the region that encodes the polypeptide chain (for protein-encoding genes) or the complete functional RNA (for other genes). These sites specify an RNA hairpin structure that is weakly bound to the template




◀ **Figure 21.8**
**Formation of an RNA hairpin.** The transcribed DNA sequence contains a region of dyad symmetry. Complementary sequences in RNA can base-pair to form a hairpin.

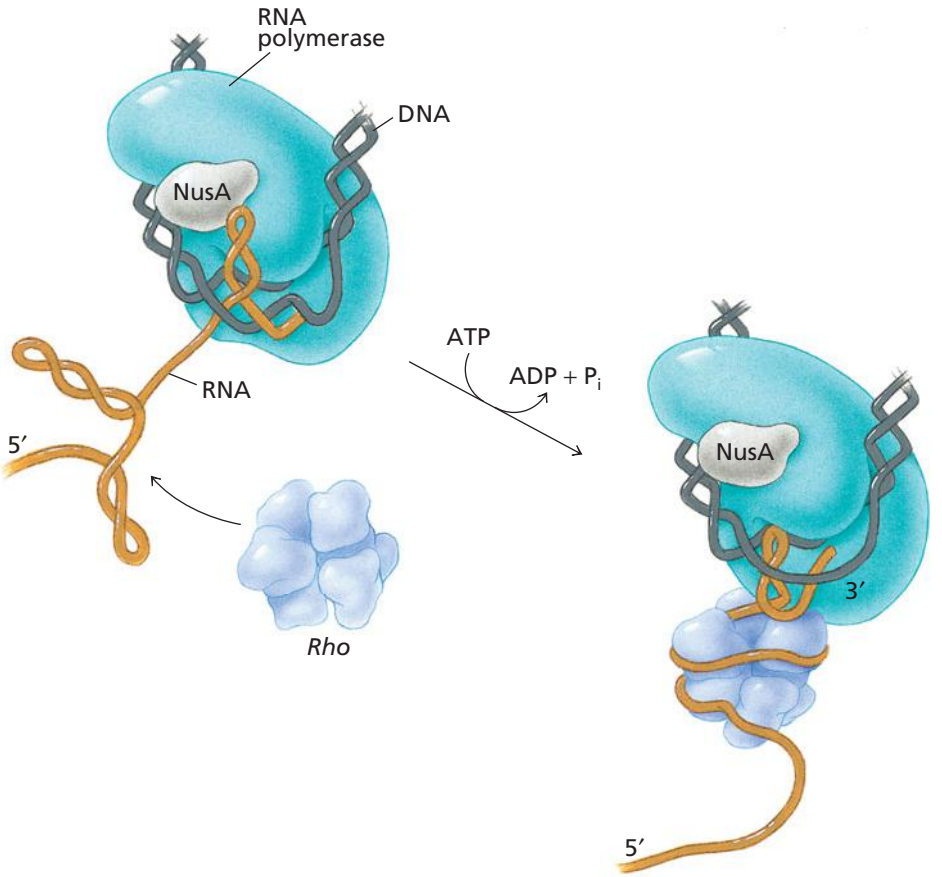



**Figure 21.9 ►**
***Rho*-dependent termination of transcription in *E. coli*.** RNA polymerase is stalled at a pause site where *rho* binds to newly synthesized RNA. This binding is accompanied by ATP hydrolysis. *Rho* probably wraps the nascent RNA chain around itself, thereby destabilizing the RNA-DNA hybrid and terminating transcription.

[Adapted from Platt, T. (1986). Transcription termination and the regulation of gene expression. *Annu. Rev. Biochem.* 55:339–372.]

strand by a short stretch of A/U base pairs. These are the weakest possible base pairs (Table 19.3) and they are easily disrupted during pausing. Disruption leads to release of RNA from the transcription complex.

The other type of bacterial termination sequences are said to be *rho*-dependent. *Rho* also triggers disassembly of transcription complexes at some pause sites. It is a hexameric protein with a potent ATPase activity and an affinity for single-stranded RNA. *Rho* may also act as an RNA-DNA helicase. It binds to single-stranded RNA that is exposed behind a paused transcription complex in a reaction coupled to hydrolysis of ATP. Approximately 80 nucleotides of RNA wrap around the protein, causing the transcript to dissociate from the transcription complex (Figure 21.9). *Rho*-dependent termination results from both destabilization of the RNA-DNA hybrid and direct contact between the transcription complex and *rho* as *rho* binds RNA. *Rho* can also bind to accessory proteins, such as NusA. This interaction may cause the RNA polymerase to change conformation and dissociate from the template DNA.

*Rho*-dependent termination requires exposure of single-stranded RNA. In bacteria, RNA transcribed from protein-encoding genes is typically bound by translating ribosomes that interfere with *rho* binding. Single-stranded RNA only becomes exposed to *rho* when transcription passes beyond the point where protein synthesis terminates. Transcription terminates at the next available pause site. In other words, *rho*-dependent termination does not occur at pause sites within the coding region but can occur at pause sites past the translation termination codon. The net effect is to couple transcription termination to translation. The advantages of such a coupling mechanism are that synthesis of an mRNA coding region is not interrupted (which would prevent protein synthesis) and that there is minimal wasteful transcription downstream of the coding region.

## 21.5 Transcription in Eukaryotes

**KEY CONCEPT**

**Eukaryotic transcription complexes tend to have more factors than the analogous bacterial complexes.**

The same processes carried out by a single RNA polymerase in *E. coli* are carried out in eukaryotes by several similar enzymes. The activities of eukaryotic transcription complexes also require many more accessory proteins than those seen in bacteria.

### A. Eukaryotic RNA Polymerases

Three different RNA polymerases transcribe nuclear genes in eukaryotes. Other RNA polymerases are found in mitochondria and chloroplasts. Each nuclear enzyme transcribes a different class of genes (Table 21.4). RNA polymerase I transcribes genes that encode large ribosomal RNA molecules (class I genes). RNA polymerase II transcribes genes that encode proteins and a few that encode small RNA molecules (class II genes). RNA polymerase III transcribes genes that encode a number of small RNA molecules, including tRNA and 5S rRNA (class III genes). (Some of the RNA molecules listed in the table are discussed in subsequent sections.)

The mitochondrial version of RNA polymerase is a monomeric enzyme encoded by the nuclear genome. It is substantially similar in amino acid sequence to the RNA polymerases of T3 and T7 bacteriophages. This similarity suggests that these enzymes share a common ancestor. It is likely that the gene for mitochondrial RNA polymerase was transferred to the nucleus from the primitive mitochondrial genome.

Chloroplast genomes often contain genes that encode their own RNA polymerase. The genes encoding the chloroplast RNA polymerase are similar in sequence to those of RNA polymerase in cyanobacteria. This is further evidence that chloroplasts, like mitochondria, originated from bacterial endosymbionts in ancestral eukaryotic cells.

The three nuclear RNA polymerases are complex multisubunit enzymes. They differ in subunit composition, although they share several small polypeptides in common. The exact number of subunits in each polymerase varies among organisms but there are always 2 large subunits and 7 to 12 smaller ones (Figure 21.10). RNA polymerase II transcribes all protein-coding genes as well as some genes that encode small RNA molecules. The protein-coding RNA synthesized by this enzyme was originally called heterogeneous nuclear RNA (hnRNA) but it is now more commonly referred to as mRNA precursor, or pre-mRNA. The processing of this precursor into mature mRNA is described in Section 21.9.

About 40,000 molecules of RNA polymerase II are found in large eukaryotic cells; the activity of this enzyme accounts for roughly 20% to 40% of all cellular RNA synthesis. The two largest subunits of each nuclear eukaryotic RNA polymerase are similar in sequence to the $\beta$ and $\beta'$ subunits of *E. coli* RNA polymerase indicating that they share a common ancestor. Like their prokaryotic counterparts, the core eukaryotic RNA

**Table 21.4 Eukaryotic RNA polymerases**

| Polymerase | Location | Copies per cell | Products | Polymerase activity of cell |
|---|---|---|---|---|
| RNA polymerase I | Nucleolus | 40,000 | 35–47S pre-rRNA | 50%–70% |
| RNA polymerase II | Nucleoplasm | 40,000 | mRNA precursors<br>U1, U2, U4, and U5 snRNA | 20%–40% |
| RNA polymerase III | Nucleoplasm | 20,000 | 5S rRNA<br>tRNA<br>U6 snRNA<br>7S RNA<br>Other small RNA molecules | 10% |
| Mitochondrial RNA polymerase | Mitochondrion | ? | Products of all mitochondrial genes | <1% |
| Chloroplast RNA polymerase | Chloroplast | ? | Products of all chloroplast genes | <1% |

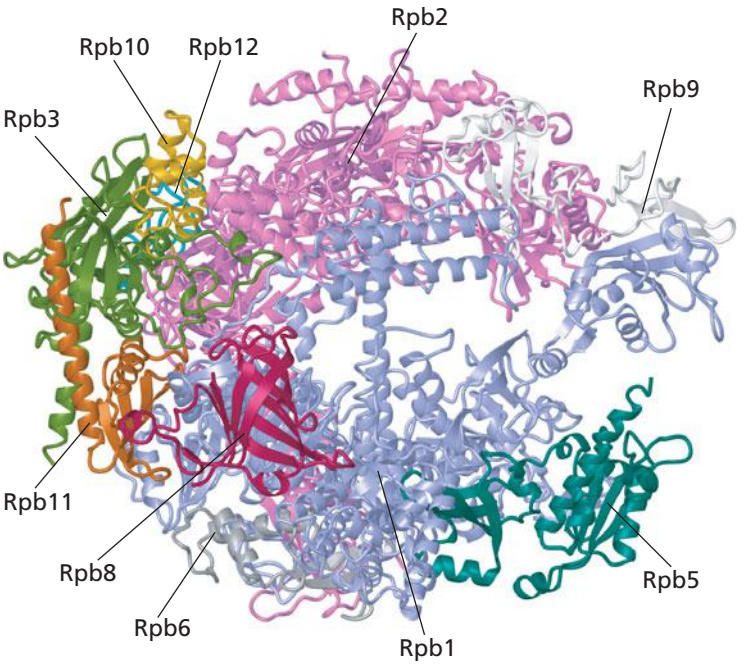



**Figure 21.10 ▸**
**RNA polymerase II from the yeast *Saccharomyces cerevisiae*.** The large subunit colored purple (Rpb2) is the homolog of the $\beta$ subunit of the prokaryotic enzyme shown in Figure 21.2. [PDB 1EN0].

polymerases do not bind on their own to promoters. RNA polymerase II requires five different biochemical activities, or factors, to form a basal transcription complex capable of initiating transcription on a minimal eukaryotic promoter (Figure 21.11). These general transcription factors (GTFs) are: TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Table 21.5).

Many class II genes contain an A/T-rich region, also called a TATA box, that is functionally similar to the prokaryotic TATA box discussed above (recall that A/T-rich regions are more easily unwound to create an open complex, especially if the DNA is negatively supercoiled (Section 19.3)). This eukaryotic A/T-rich region is located 19 to 27 bp upstream of the transcription start site and serves to recruit RNA polymerase II to the DNA during assembly of the initiation complex.

The general transcription factor TFIID is a multisubunit factor and one of its subunits, TATA-binding protein (TBP), binds to the region containing the TATA box. The structure of TBP from the plant *Arabidopsis thaliana* is shown in Figure 21.12. TBP forms a saddle-shaped molecular clamp that almost surrounds the DNA at the TATA box. The main contacts between TBP and DNA are due to interactions between acidic amino acid side chains in $\beta$ strands and the edges of base pairs in the minor groove. When TBP binds to DNA, the promoter DNA is bent so that it no longer resembles the standard B-DNA conformation. This is an unusual interaction for DNA-binding proteins. The TBP subunit of TFIID is also required to initiate transcription of class I and class III genes by RNA polymerases I and III, respectively.

The eukaryotic RNA polymerase II subunit homologous to the prokaryotic RNA polymerase $\beta'$ subunit has an unusual carboxy-terminal domain (CTD) or "tail" that

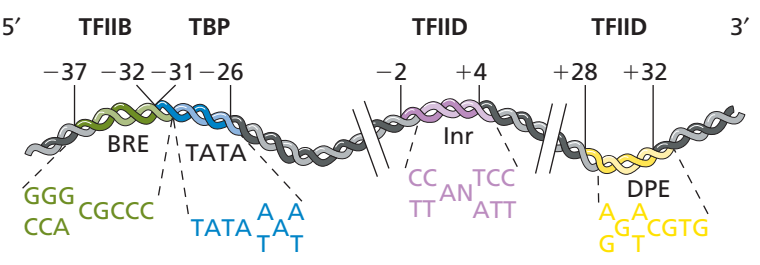
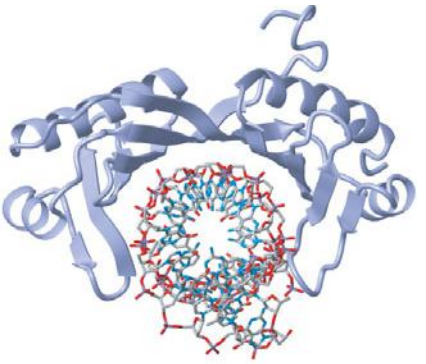



**Figure 21.11 ▸**
**A generic eukaryotic promoter showing the basal or "core" promoter elements.** The TATA box is described in the text. The BRE is the TFIIB recognition element, while Inr stands for the initiator element. The DPE is the downstream promoter element. The names of the factors that bind to each site are shown above the promoter, and the consensus recognition sequences for each site are shown below the schematic promoter fragment.

**Table 21.5 Some representative RNA polymerase II transcription factors**

| Factor | Characteristics |
|---|---|
| TFIIA | Binds to TFIID; can interact with TFIID in the absence of DNA |
| TFIIB | Interacts with RNA polymerase II |
| TFIID | RNA polymerase II initiation factor |
| TBP | TATA-binding protein; subunit of TFIID |
| TAFs | TBP-associated factors; many subunits |
| TFIIE | Interacts with RNA polymerase II |
| TFIIH | Required for initiation; helicase activity; couples transcription to DNA repair |
| TFIIS[a] | Binds to RNA polymerase II; elongation factor |
| TFIIF | Binds to RNA polymerase II; two subunits—RAP30 and RAP74 |
| SP1 | Binds to GC-rich sequence |
| CTF[b] | Family of different proteins that recognize the core sequence CCAAT |

[a]Also known as sII or RAP38
[b]Also known as NP1.



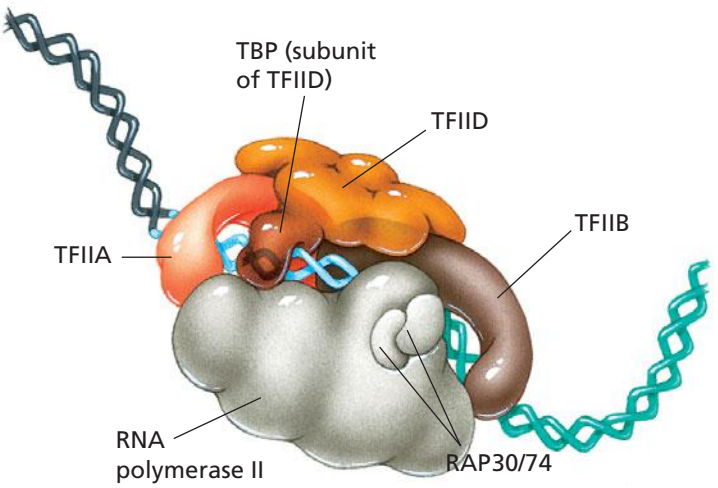
▲ **Figure 21.12**
***Arabidopsis thaliana* TATA-binding protein (TBP) bound to DNA.** TBP (blue) is bound to a double-stranded DNA fragment with a sequence corresponding to a TATA box (5′-TATAAAG-3′) DNA is shown as a wire-frame model. Note that the $\beta$ sheet of TBP lies in the minor groove of the DNA fragment. [PDB 1VOL].

consists of multiple repeats of the amino acid heptamer PTSPSYS. The Ser and Thr residues in the tail are phosphorylation targets for nuclear protein kinases. RNA polymerase II molecules with a hyperphosphorylated CTD are typically transcriptionally active, or engaged, while the cellular pol II with hypophosphorylated CTDs are usually quiescent.

Although it has proven possible to purify RNA polymerase II and each GTF and use them to reconstitute accurate transcription initiation *in vitro*, these basal transcription complexes are not competent to recognize the many different types of *trans*-acting factors and *cis*-acting sequences that are known to play important roles *in vivo*. Searching for cellular constituents that could respond to transcriptional activators *in vitro* led to the discovery of a large preformed RNA pol II holoenzyme that contains not only the five GTFs but also many other polypeptides that mediate interactions between pol II and sequence-specific DNA-binding proteins. This eukaryotic holoenzyme is analogous to the core + $\sigma$ holoenzyme in *E. coli.*

## B. Eukaryotic Transcription Factors

TFIIA and TFIIB are essential components of the RNA polymerase II holoenzyme complex. Neither TFIIA nor TFIIB can bind to DNA in the absence of TFIID. TFIIF (also known as Factor 5 or RAP30/74) binds to RNA polymerase II during initiation (Figure 21.13). TFIIF plays no direct role in recognizing the promoter but it is analogous to bacterial $\sigma$ factors in two ways: it decreases the affinity of RNA polymerase II for nonpromoter




◀ **Figure 21.13**
**RNA polymerase II holoenzyme complex bound to a promoter.** This model shows various transcription factors bound to RNA polymerase II at a promoter. The transcription factors are often larger and more complex than those shown in this diagram.

DNA, and it helps form the open complex. TFIIH, TFIIE, and other, less well-characterized factors, are also part of the transcription initiation complex.

Once the initiation complex assembles at the site of the promoter, the next steps are similar to those in bacteria. An open complex is formed, a short stretch of RNA is synthesized, and the transcription complex clears the promoter. Most transcription factors dissociate from DNA and RNA polymerase II once elongation begins. However, TFIIF may remain bound and a specific elongation factor, TFIIS (also called sII or RAP38), associates with the transcribing polymerase. TFIIS may play a role in pausing and transcription termination that is similar to the role of NusA in bacteria.

With the exception of TBP, the transcription factors that interact with the other two eukaryotic RNA polymerases are not the same as those required by RNA polymerase II.

### C. The Role of Chromatin in Eukaryotic Transcription

As described in Chapter 19, the eukaryotic genome is packaged using small, ubiquitous building blocks, called nucleosomes, that contain an octamer of the four core histone proteins. It is estimated that approximately 35% of the mammalian genome is transcribed into protein-coding genes (including the introns) and so most of a cell's DNA is relatively inert. But even within that 35%, which contains about 20,000 protein-coding genes, the majority of the sequences are quiescent. In any single cell, the primary determinant of whether a gene is competent to be transcribed resides in the state of its chromatin. This status is modulated by two mechanisms. The first involves implementing or removing post-translational modifications on the flexible amino-terminal arms of the four core histones (Section 19.5B). Specific Lys residues are targeted for methylation or acetylation, specific Arg residues may also be methylated, while Ser and Thr side chains can be phosphorylated. Different modifications serve as signals to recruit either activators or repressors to the chromatin. The second mechanism for specifying the transcriptional status of a eukaryotic gene involves nucleosome positioning and remodeling.

Nontranscribed genes are relatively inaccessible in the nucleus while transcribed genes are relatively accessible to transcription factors, pol II holoenzyme, and other nuclear proteins. How does a gene move between these two conflicting states? The answer lies with large multiprotein complexes that use the energy from hydrolyzing ATP to physically remodel a gene's nucleosomes and allow proteins to have access to the DNA. Some of the remodeling complexes actually contain histone-modifying enzymes like histone acetylase (HAT) or histone deacetylase (HDAC).

## 21.6 Transcription of Genes Is Regulated

As noted at the beginning of this chapter, many genes are expressed in every cell. The expression of these housekeeping genes is said to be ***constitutive***. In general, such genes have strong promoters and are transcribed efficiently and continuously. Genes whose products are required at low levels usually have weak promoters and are transcribed infrequently. In addition to constitutively expressed genes, cells contain genes that are expressed at high levels in some circumstances and not at all in others. Such genes are said to be regulated.

Regulation of gene expression can occur at any point in the flow of biological information but occurs most often at the level of transcription. Various mechanisms have evolved that allow cells to program gene expression during differentiation and development and to respond to environmental stimuli.

The initiation of transcription of regulated genes is controlled by regulatory proteins that bind to specific DNA sequences. Transcriptional regulation can be negative or positive. Transcription of a negatively regulated gene is prevented by a regulatory protein called a **repressor**. A negatively regulated gene can be transcribed only in the absence of an active repressor. Transcription of a positively regulated gene can be activated

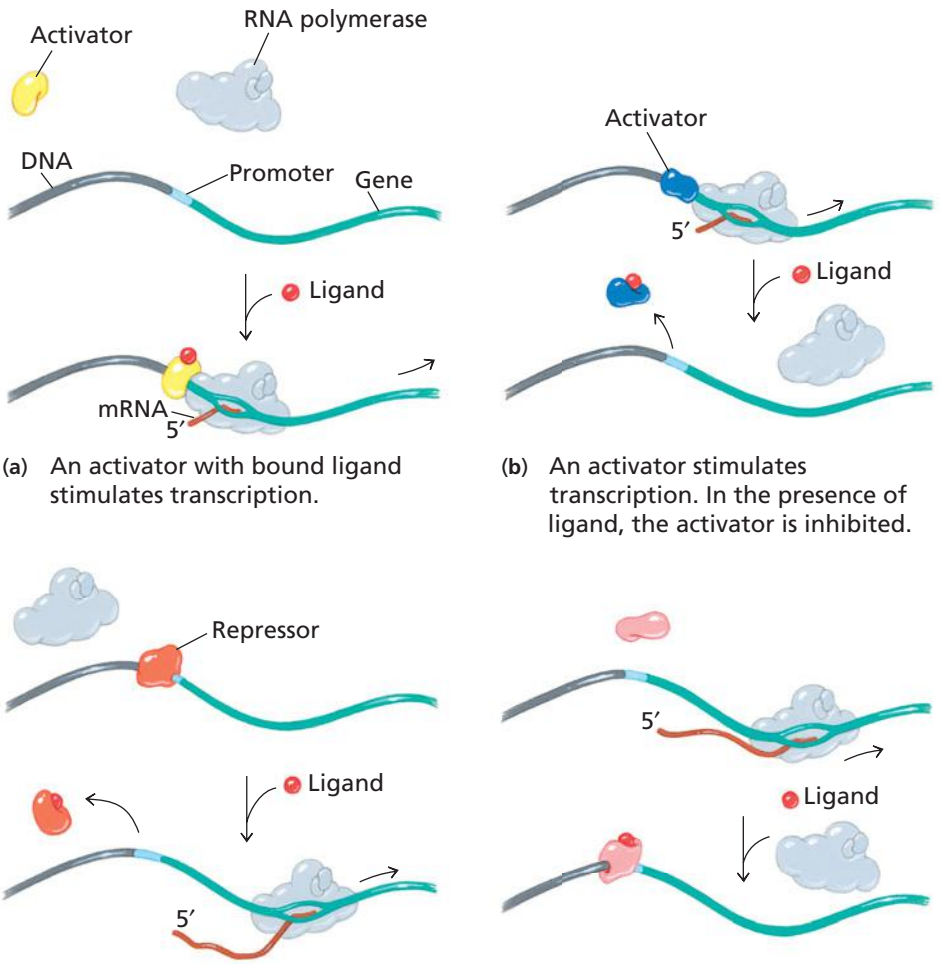
by a regulatory protein called an **activator**. A positively regulated gene is transcribed poorly or not at all in the absence of the activator.

Repressors and activators are often allosteric proteins whose function is modified by ligand binding. In general, a ligand alters the conformation of the protein and affects its ability to bind to specific DNA sequences. For example, some repressors control the synthesis of enzymes for a catabolic pathway. In the absence of substrate for these enzymes, the genes are repressed. When substrate is present, it binds to the repressor, causing the repressor to dissociate from the DNA and allowing the genes to be transcribed. Ligands that bind to and inactivate repressors are called **inducers** because they induce transcription of the genes controlled by the repressors. In contrast, some repressors that control the synthesis of enzymes for a biosynthetic pathway bind to DNA only when associated with a ligand. The ligand is often the end product of the biosynthetic pathway. This regulatory mechanism ensures that the genes in the pathway are turned off as product of the pathway accumulates. Ligands that bind to and activate repressors are called **corepressors**. The DNA-binding activity of allosteric activators can also be affected in two ways by ligand binding. Four general strategies for regulating transcription are illustrated in Figure 21.14. Examples of all four strategies have been identified.

Few regulatory systems are as simple as those described above. For example, the transcription of many genes is regulated by a combination of repressors and activators or by multiple activators. Elaborate mechanisms for regulating transcription

## KEY CONCEPT

**Cells don't synthesize a specific protein until it is required (e.g., the *lac* operon is not transcribed until the intracellular concentration of lactose inactivates the *lac* repressors).**




(a) An activator with bound ligand stimulates transcription.

(b) An activator stimulates transcription. In the presence of ligand, the activator is inhibited.

(c) A repressor prevents transcription. Binding of ligand (inducer) to the repressor inactivates the repressor and allows transcription.

(d) In the absence of ligand, the repressor does not bind to DNA. Repression occurs only when ligand (corepressor) is present.

◀ **Figure 21.14**
**Strategies for regulating transcription initiation by regulatory proteins.**

have evolved to meet the specific requirements of individual organisms. A greater range of cellular responses is possible when transcription is regulated by a host of mechanisms acting together. By examining how the transcription of a few particular genes is controlled, we can begin to understand how positive and negative mechanisms can be combined to produce the remarkably sensitive regulation seen in bacterial cells.

## 21.7 The *lac* Operon, an Example of Negative and Positive Regulation

Some bacteria obtain the carbon they need for growth by metabolizing five- or six-carbon sugars via glycolysis. For example, *E. coli* preferentially uses glucose as a carbon source but can also use other sugars, including β-galactosides such as lactose. The enzymes required for β-galactoside uptake and catabolism are not synthesized unless a β-galactoside substrate is available. Even in the presence of their substrate, these enzymes are synthesized in limited amounts when the preferred carbon source (glucose) is also present. Synthesis of the enzymes required for β-galactoside utilization is regulated at the level of transcription initiation by a repressor and an activator.

The uptake and catabolism of β-galactosides requires three proteins. The product of the *lacY* gene is lactose permease, a symport transporter that is responsible for the uptake of β-galactosides. Most β-galactosides are subsequently hydrolyzed to metabolizable hexoses by the activity of β-galactosidase, a large enzyme with four identical subunits encoded by the *lacZ* gene. β-Galactosides that cannot be hydrolyzed are acetylated by the activity of thiogalactoside transacetylase, the product of the *lacA* gene. Acetylation helps to eliminate toxic compounds from the cell.

The three genes—*lacZ*, *lacY*, and *lacA*—form an operon that is transcribed from a single promoter to produce a large mRNA molecule containing three separate protein-coding regions. In this case, we refer to a protein-coding region as a gene, a definition that differs from our standard use of the term. The arrangement of genes with related functions in an operon is efficient because the concentrations of a set of proteins can be controlled by transcribing from a single promoter. Operons composed of protein-coding genes are common in *E. coli* and other prokaryotes but were thought to be extremely rare in eukaryotes. We now realize that operons are also quite common in the model organism *C. elegans*, a nematode or round worm, and are likely widespread in this large phylum. Operons are also common in mitochondrial and chloroplast genomes.

### A. *lac* Repressor Blocks Transcription

Expression of the three genes of the *lac* operon is controlled by a regulatory protein called *lac* repressor, a tetramer of identical subunits. The repressor is encoded by a fourth gene, *lacI*, which is located just upstream of the *lac* operon but is transcribed from a separate promoter (Figure 21.15).

*lac* repressor binds simultaneously to two sites near the promoter of the *lac* operon. Repressor-binding sites are called **operators**. One operator ($O_1$) is adjacent to the promoter, and the other ($O_2$) is within the coding region of *lacZ*. When bound to both operators, the repressor causes the DNA to form a stable loop that can be seen

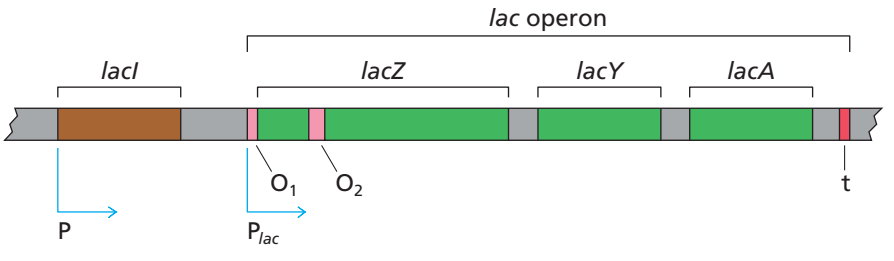
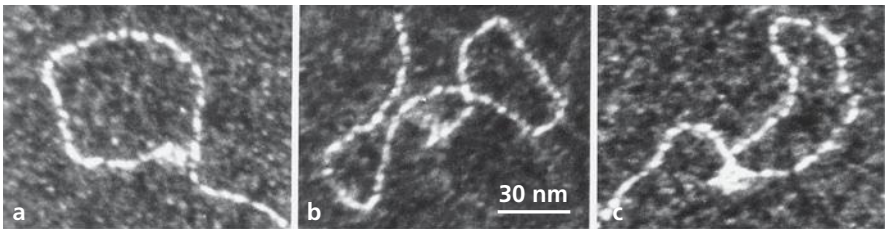



**Figure 21.15 ▸**
**Organization of the genes that encode proteins required to metabolize lactose.** The coding regions for three proteins—LacZ, LacY, and LacA—constitute the *lac* operon and are co-transcribed from a single promoter ($P_{lac}$). The gene that encodes *lac* repressor, *lacI*, is located upstream of the *lac* operon and has its own promoter, P; *lac* repressor binds to the operators $O_1$ and $O_2$ near $P_{lac}$; t denotes the transcription termination sequence.



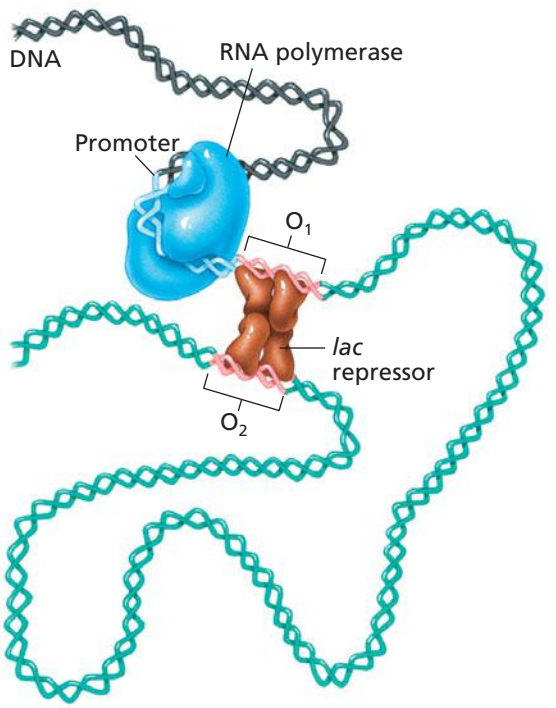

◀ **Figure 21.16**
**Electron micrographs of DNA loops.** These loops were formed by mixing *lac* repressor with a fragment of DNA bearing two synthetic *lac* repressor–binding sites. One binding site is located at one end of the DNA fragment, and the other is 535 bp away. DNA loops 535 bp in length form when the tetrameric repressor binds simultaneously to the two sites.

in electron micrographs of the complex formed between *lac* repressor and DNA (Figure 21.16). The interaction of *lac* repressor with the operator sequences may block transcription by preventing the binding of RNA polymerase to the *lac* promoter. However, it is now known that, in some cases, both *lac* repressor and RNA polymerase can bind to the promoter at the same time. Thus, the repressor may also block transcription initiation by preventing formation of the open complex and promoter clearance. A schematic diagram of *lac* repressor bound to DNA in the presence of RNA polymerase is shown in Figure 21.17. The diagram illustrates the relationship between the operators and the promoter and the DNA loop that forms when the repressor binds to DNA.

**At any given time, one molecule of repressor is bound to the operator and nine molecules are bound non-specifically to DNA.**

The repressor locates an operator by binding nonspecifically to DNA and searching by sliding or hopping in one dimension. The non-specific equilibrium constant is about $10^6\ M^{-1}$—comparable to that of RNA polymerase (Section 21.3C). (Recall from Section 21.3C that RNA polymerase also uses this kind of searching mechanism.) The equilibrium association constant for the specific binding of *lac* repressor to $O_1$ *in vitro* is very high ($K_a \approx 10^{13}M^{-1}$). As a result, the repressor blocks transcription very effectively. (*lac* repressor binds to the $O_2$ site with lower affinity.) A bacterial cell contains only about ten molecules of *lac* repressor but the repressor searches for and finds an operator so rapidly that when a repressor dissociates spontaneously from the operator, another occupies the site within a very short time. However, during this brief interval, one transcript of the operon can be made since RNA polymerase is poised at the promoter. This low level of transcription, called escape synthesis, ensures that small amounts of lactose permease and β-galactosidase are present in the cell.

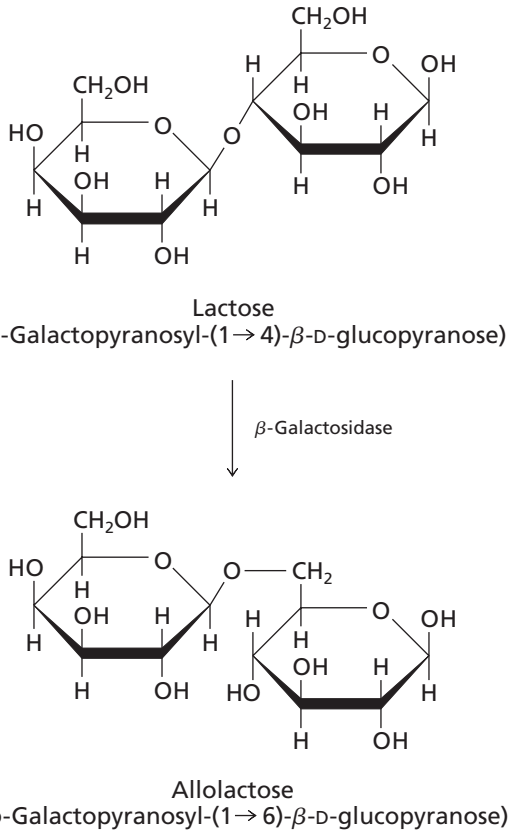
In the absence of lactose, *lac* repressor blocks expression of the *lac* operon, but when β-galactosides are available as potential carbon sources, the genes are transcribed. Several β-galactosides can act as inducers. If lactose is the available carbon source, the inducer is allolactose, which is produced from lactose by the action of β-galactosidase (Figure 21.18). Allolactose binds tightly to *lac* repressor and causes a conformational change that reduces the affinity of the repressor for the operators ($K_a \approx 10^{10}M^{-1}$). In the presence of the inducer, *lac* repressor dissociates from the DNA, allowing RNA polymerase to initiate transcription. (Note that because of escape synthesis, lactose can be taken up and converted to allolactose even when the operon is repressed.)




▲ **Figure 21.17**
**Binding of *lac* repressor to the *lac* operon.** The tetrameric *lac* repressor interacts simultaneously with two sites near the promoter of the *lac* operon. As a result, a loop of DNA forms. RNA polymerase can still bind to the promoter in the presence of the *lac* repressor–DNA complex.

## B. The Structure of *lac* Repressor

The role of *lac* repressor in regulating expression of the *lac* operon has been known since the 1960s. However, the structure of this important protein was solved only in the 1990s after the development of new techniques for determining the structure of large molecules. The structure of part of the *lac* repressor bound to one operator sequence is shown in Figure 21.19. The complete protein contains four identical subunits arranged as two pairs, and each pair of subunits binds to a different operator sequence. Inside the cell these two fragments of DNA are part of a single DNA molecule—and repressor binding forms a loop of DNA at the 5′ end of the *lac* operon.




**▲ Figure 21.18**
**Formation of allolactose from lactose, catalyzed by β-galactosidase.** This is a minor or side reaction. The main enzymatic activity of β-galactosidase is to cleave disaccharides into monomers that can be converted into substrates for glycolysis.

The subunits are joined together at a hinge region. The X-ray crystallographic structure reveals that the two pairs of subunits are stacked on top of one another (Figure 21.17) and not extended away from the hinge region as was expected. This makes a more compact protein that is less symmetric than many other tetrameric proteins.

Each subunit contains a helix-turn-helix motif at the ends farthest from the hinge region. When bound to DNA, one of the α helices lies in the major groove where amino acid side chains interact directly with the specific base pairs of the operator sequence. The two helices from each pair of subunits are positioned about one turn of DNA apart (about 10 bp), and each one interacts with half of the operator sequence. This binding strategy is similar to that of restriction endonuclease *Eco*RI (Section 19.6C).

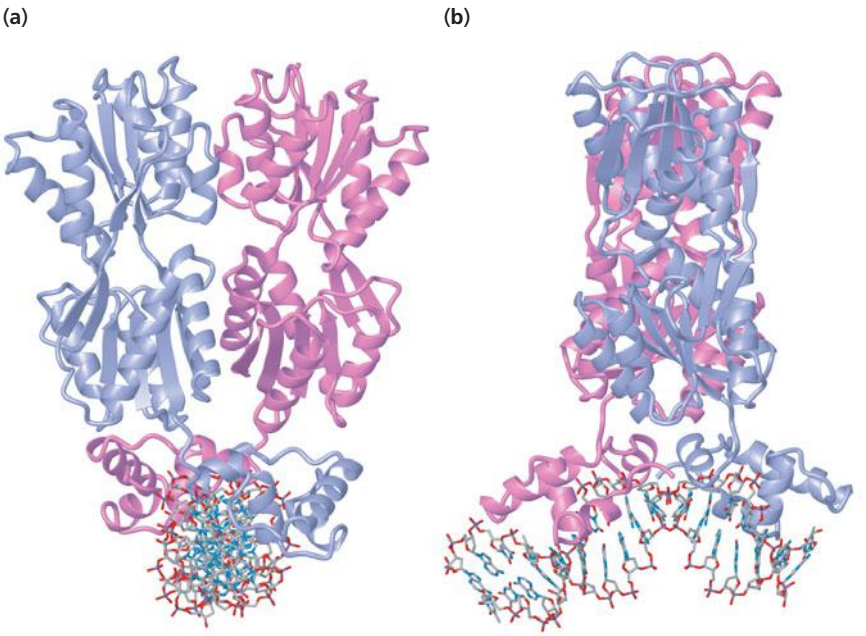
In the absence of DNA the distal regions of the *lac* repressor subunits are disordered (Section 4.7D). This is one reason why it took such a long time to work out the structure. The structure of the helix-turn-helix motif can only be seen when the protein is bound to DNA. There are now many examples of such interactions in which the stable structure of the protein is significantly altered by ligand binding. In the presence of inducers, such as allolactose or IPTG, the repressor adopts a slightly different conformation and can no longer bind to the DNA operators.

## C. cAMP Regulatory Protein Activates Transcription

Transcription of the *lac* operon in *E. coli* depends not only on the presence of β-galactosides but also on the concentration of glucose in the external medium. The *lac* operon is transcribed maximally when β-galactosides, such as lactose, are the only carbon source; transcription is reduced 50-fold when glucose is also present. The decreased rate of transcription of operons when glucose is present is termed catabolite repression.

Catabolite repression is a feature of many operons encoding metabolic enzymes. These operons characteristically have weak promoters from which transcription is initiated inefficiently in the presence of glucose. In the absence of glucose, however, the rate of transcription initiation increases dramatically due to an activator that converts the relatively weak promoter to a stronger one. No repressor is involved, despite the

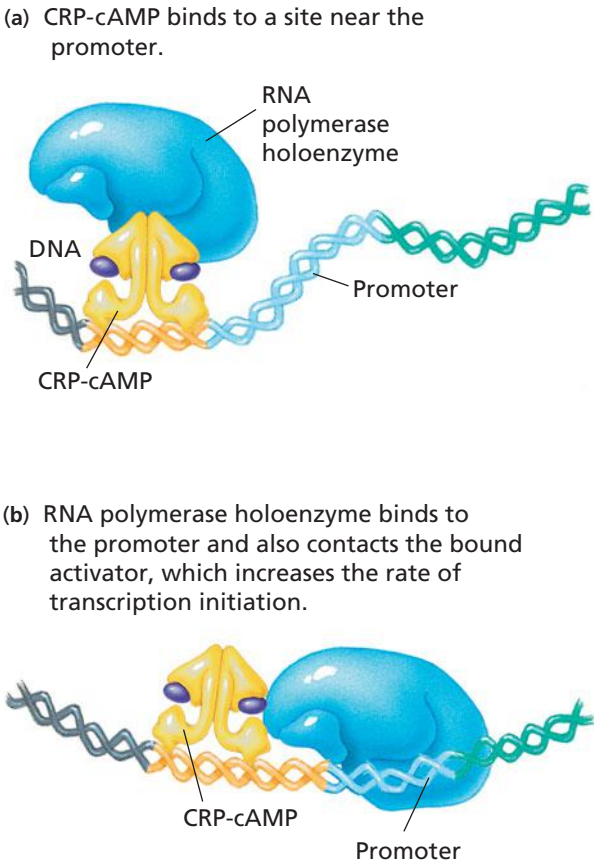
**Figure 21.19 ►**
**Structure of *E. coli lac* repressor.** This figure shows a dimer of *lac* repressor subunits bound to DNA. *Lac* repressor is a tetramer *in vivo*, containing two DNA-binding sites. **(a)** End-on view of the DNA molecule. **(b)** Side view showing the *lac* repressor α helix in the major groove. [PDB 1EFA].



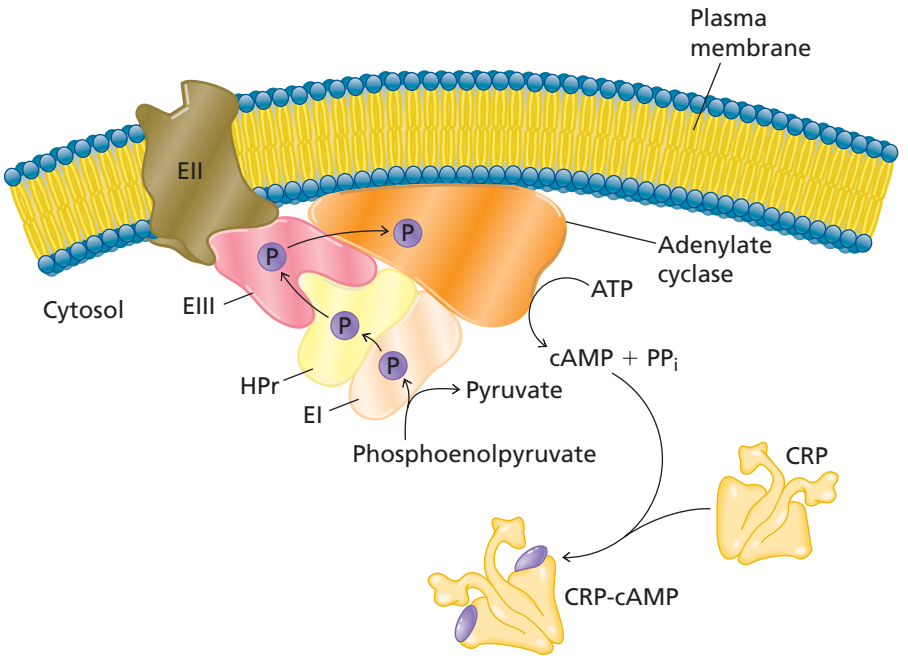
use of the term *catabolite repression*. In fact, this is a well-studied example of an activation mechanism.

The activator is cyclic AMP regulatory (or receptor) protein (CRP), also known as catabolite activator protein (CAP). CRP is a dimeric protein whose activity is modulated by cyclic AMP. In the absence of cAMP, CRP has low affinity for DNA but when cAMP is present it binds to CRP and converts it to a sequence-specific DNA-binding protein. The CRP-cAMP complex interacts with specific DNA sequences near the promoters of more than 30 genes including the *lac* operon. Because the genome contains many more binding sites for CRP-cAMP than for *lac* repressor, it is not surprising that there are at least 1000 molecules of CRP per cell compared to only about 10 molecules of *lac* repressor. The CRP-cAMP binding sites are often just upstream of the −35 regions of the promoters they activate. While bound to DNA, CRP-cAMP can contact RNA polymerase at the promoter site, leading to increased rates of transcription initiation (Figure 21.20). Most of the protein–protein interactions are between bound CRP-cAMP and the $\alpha$ subunits of RNA polymerase. This is typical of most interactions between activators and RNA polymerase. (There are many different transcriptional activators in bacterial cells.) The net effect of CRP-cAMP is to increase the production of enzymes that can use substrates other than glucose. In the case of the *lac* operon, activation by CRP-cAMP occurs only when $\beta$-galactosides are available. At other times, transcription of the operon is repressed.

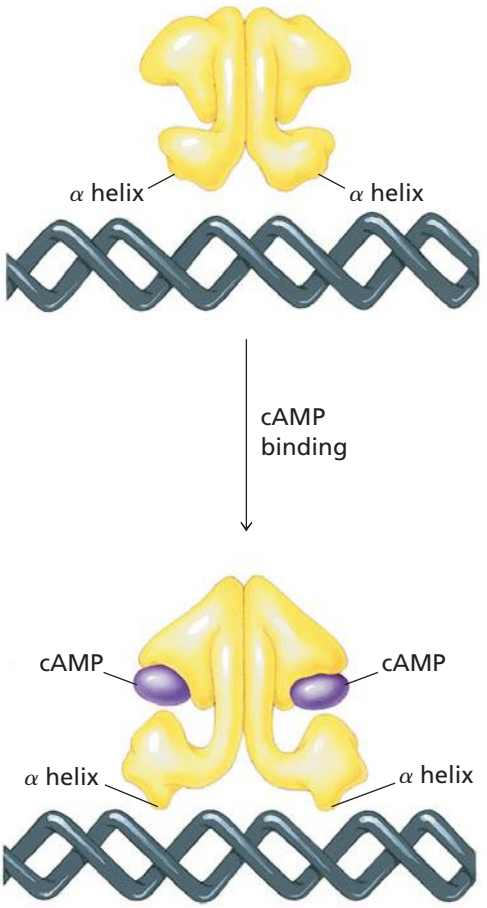
The concentration of cAMP inside an *E. coli* cell is controlled by the concentration of glucose outside the cell. When glucose is available, it is imported into the cell and phosphorylated by a complex of transport proteins collectively known as the phosphoenolpyruvate-dependent sugar phosphotransferase system. When glucose is not available, one of the glucose transport enzymes, enzyme III, catalyzes the transfer of a phosphoryl group, ultimately derived from phosphoenolpyruvate, to adenylate cyclase, leading to its activation (Figure 21.21).




▲ **Figure 21.20**
**Activation of transcription initiation at the *lac* promoter by CRP-cAMP.**




◀ **Figure 21.21**
**cAMP production**. In the absence of glucose, enzyme III (EIII) transfers a phosphoryl group, originating from phosphoenolpyruvate, to membrane-bound adenylate cyclase. Phosphorylated adenylate cyclase catalyzes the conversion of ATP to cAMP. cAMP binds to CRP, and CRP-cAMP activates the transcription of a number of genes encoding enzymes that compensate for the lack of glucose as a carbon source.




▲ **Figure 21.22**
**Conformational changes in CRP caused by cAMP binding.** Each monomer of the CRP dimer contains a helix-turn-helix motif. In the absence of cAMP, the $\alpha$ helices cannot fit into adjacent sections of the major groove of DNA and cannot recognize the CRP-cAMP binding site. When cAMP binds to CRP, the two $\alpha$ helices assume the proper conformation for binding to DNA.

Adenylate cyclase (also know as adenylyl cyclase) then catalyzes the conversion of ATP to cAMP thereby increasing the levels of cAMP in the cell. As molecules of cAMP are produced, they bind to CRP stimulating transcription initiation at promoters that respond to catabolite repression. Similar mechanisms for responding to external stimuli operate in eukaryotes where molecules such as cAMP act as second messengers (Section 9.12B).

Each subunit of the CRP dimer contains a helix-turn-helix DNA binding motif. In the presence of cAMP, two helices—one from each monomer—fit into adjacent sections of the major groove of DNA and contact the nucleotides of the CRP-cAMP binding site. This is the same general binding strategy used by *lac* repressor and *Eco*RI. In the absence of cAMP, the conformation of CRP changes so that the two $\alpha$ helices can no longer bind to the major groove (Figure 21.22). When CRP-cAMP is bound to the activator sequence, the DNA is bent slightly to conform to the surface of the protein (Figure 21.23).

## 21.8 Post-transcriptional Modification of RNA

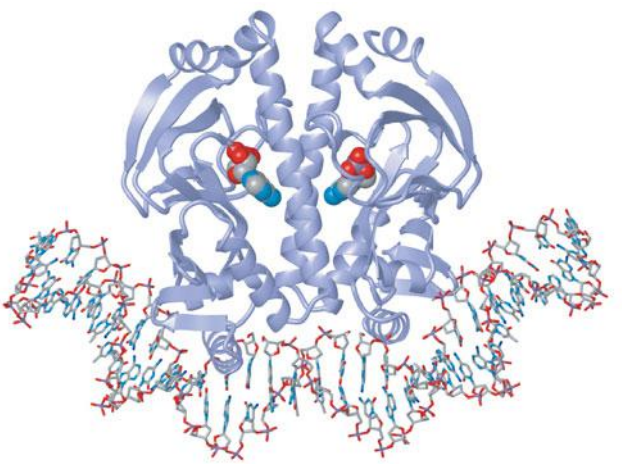
In many cases, RNA transcripts must be extensively altered before they can adopt their mature structures and functions. These alterations fall into three general categories: (1) removal of nucleotides from primary RNA transcripts; (2) addition of nucleotides not encoded by the corresponding genes; and (3) covalent modification of certain bases. The reactions that transform a primary RNA transcript into a mature RNA molecule are referred to collectively as **RNA processing**. RNA processing is crucial for the function of most RNA molecules and is an integral part of gene expression.

### A. Transfer RNA Processing

Mature tRNA molecules are generated in both eukaryotes and prokaryotes by processing primary transcripts. In prokaryotes, the primary transcript often contains several tRNA precursors. These precursors are cleaved from the large primary transcripts and trimmed to their mature lengths by ribonucleases, or RNases. Figure 21.24 summarizes the processing of prokaryotic tRNA precursors.

The endonuclease RNase P catalyzes the initial cleavage of most tRNA primary transcripts. The enzyme cleaves the transcript on the 5′ side of each tRNA sequence, releasing monomeric tRNA precursors with mature 5′ ends. Digestion with RNase P *in vivo* is rapid and occurs while the transcript is still being synthesized.

**Figure 21.23 ►**
**Structure of a complex between CRP-cAMP and DNA.** Both subunits contain a molecule of cAMP bound at the allosteric site. Each subunit has an $\alpha$ helix positioned in the major groove of DNA at the CRP-cAMP binding site. Note that binding induces a slight bend in the DNA. [PDB 1CGP].



RNase P was one of the first specific ribonucleases studied in detail and much is known about its structure. The enzyme is actually a ribonucleoprotein. In *E. coli,* it is composed of a 377-nucleotide RNA molecule ($M_r$ 130,000) and a small polypeptide ($M_r$ 18,000). In the absence of protein the RNA component is catalytically active *in vitro* (under certain conditions). It was one of the first RNA molecules shown to have enzymatic activity and is an example of the fourth class of RNA molecules described in Section 21.1. The protein component of RNase P helps maintain the three-dimensional structure of the RNA. Sidney Altman was awarded the Nobel Prize in 1989 for showing that the RNA component of RNase P had catalytic activity.

Other endonucleases cleave tRNA precursors near their 3′ ends. Subsequent processing of the 3′ end of a tRNA precursor requires the activity of an exonuclease, such as RNase D. This enzyme catalyzes the sequential removal of nucleotides from the 3′ end of a monomeric tRNA precursor until the 3′ end of the tRNA is reached.

All mature prokaryotic and eukaryotic tRNA molecules must contain the sequence CCA as the final three nucleotides at their 3′ ends. In some cases, these nucleotides are added post-transcriptionally after all other types of processing at the 3′ end have been completed. The addition of these three nucleotides is catalyzed by tRNA nucleotidyltransferase and is one of the few examples of the addition of nucleotides that are not encoded by a gene.

Processing of tRNA precursors also involves covalently modifying some of the nucleotide bases. Mature tRNA molecules exhibit a greater diversity of covalent modifications than any other class of RNA molecule. Typically 26 to 30 of the approximately 80 nucleotides in a tRNA molecule are covalently modified. Each type of covalent modification usually occurs in only one location on each molecule. Some examples of the sites of modification of nucleotides are shown in Figure 21.25.

### B. Ribosomal RNA Processing

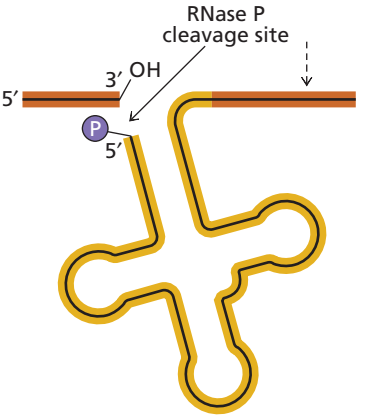
Ribosomal RNA molecules in all organisms are produced as large primary transcripts that require subsequent processing, including methylation and cleavage by endonucleases, before the mature molecules can adopt their active forms. This processing of ribosomal RNA is coupled to ribosome assembly.

The primary transcripts of prokaryotic rRNA molecules are about 30S in size and contain one copy each of the 16S, 23S, and 5S rRNAs. The transcripts also contain interspersed tRNA precursors. (Note that S is the symbol for the Svedberg unit, a measure of the rate at which particles move in the gravitational field established in an ultracentrifuge. Large S values are associated with large masses. The relationship between S and mass is not linear; therefore, S values are not additive.) Since the three rRNAs are derived from a single transcript, this processing ensures that there are equimolar amounts of each of the mature ribosomal RNAs.

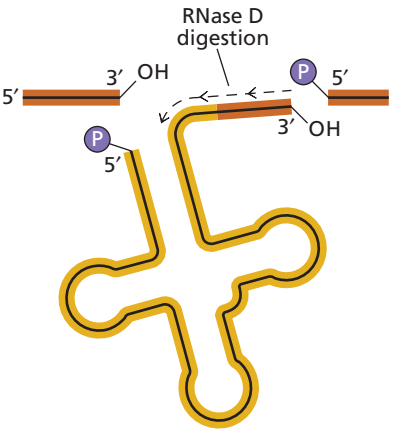
The 5′ and 3′ ends of each mature rRNA molecule are usually found in base-paired regions in the primary transcript. In prokaryotes, the endonuclease RNase III binds to these regions and cleaves the precursor near the ends of the 16S and 23S rRNAs. Following the initial cleavage, the ends of the rRNA molecules are trimmed by the actions of specific endonucleases (Figure 21.26).

Eukaryotic ribosomal RNAs are also produced by processing a larger precursor. The primary transcripts are between 35S and 47S in size and contain a copy of each of three eukaryotic rRNA species: 18S, 5.8S, and 28S. (The fourth eukaryotic rRNA, 5S rRNA, is transcribed as a monomer by RNA polymerase III and is processed separately.) The primary transcripts are synthesized in the region of the nucleus called the nucleolus, where initial processing occurs. Each rRNA precursor partially folds up and binds to some of its ribosomal protein partners before the processing cleavages take place.

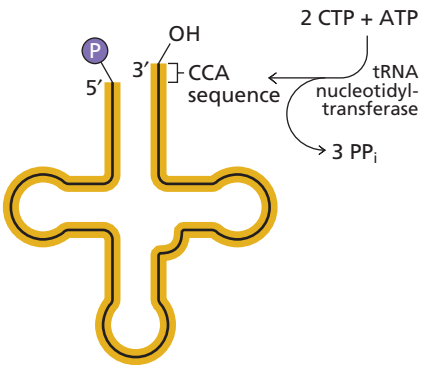
(a) RNase P and other endonucleases cleave the primary transcript.




(b) RNase D trims the 3′ end.




(c) tRNA nucleotidyl transferase adds CCA to the 3′ end.



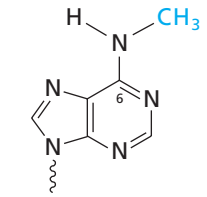

▲ **Figure 21.24**
**Summary of prokaryotic tRNA processing.**

## 21.9 Eukaryotic mRNA Processing

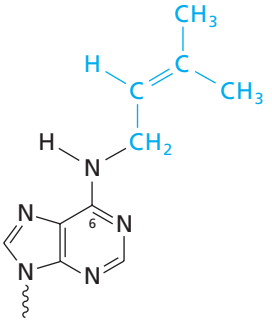
The processing of mRNA precursors is one of the biochemical features that distinguishes prokaryotes from eukaryotes. In prokaryotes, the primary mRNA transcripts are translated directly, often initiating translation before transcription is

**KEY CONCEPT**

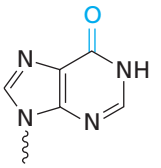
**Unmodified mRNAs are inherently unstable in a cell and would be rapidly degraded by ribonucleases.**



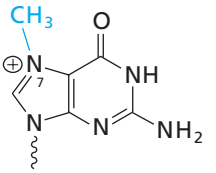
$N^6$-Methyladenylate
($m^6$-A)



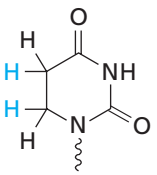
$N^6$-Isopentenyladenylate
($i^6$-A)



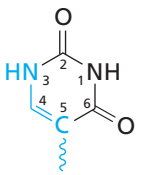
Inosinate
(I)



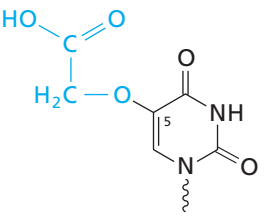
7-Methylguanylate
($m^7$G)



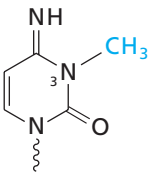
Dihydrouridylate
(D)



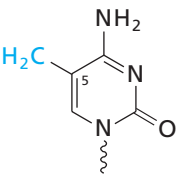
Pseudouridylate
($\psi$)
(ribose at C-5)



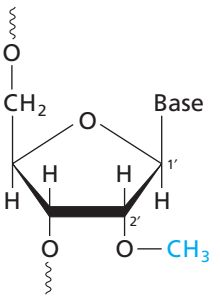
Uridylate
5-oxyacetic acid
($cmo^5$-U)



3-Methylcytidylate
($m^3$C)



5-Methylcytidylate
($m^5$C)



2′-*O*-Methylated
nucleotide
(Nm)

**▲ Figure 21.25**
**Examples of common covalent modifications found in tRNA molecules (the modifications are shown in blue).**

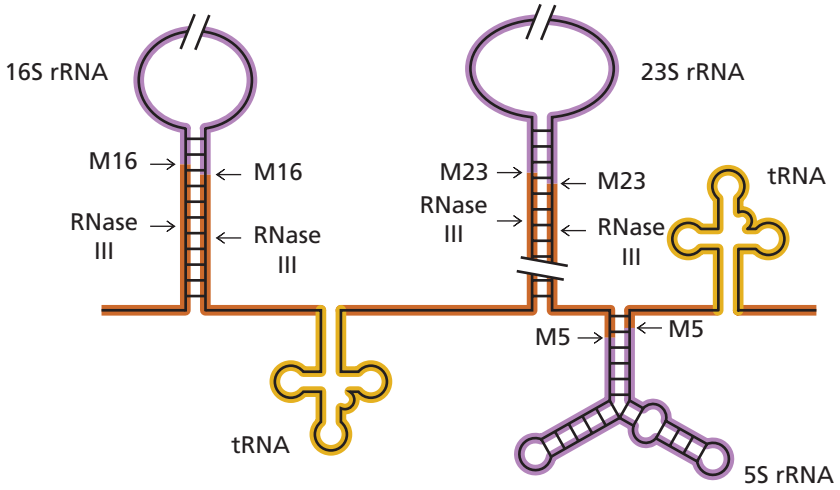



**Figure 21.26 ►**
**Endonucleolytic cleavage of ribosomal RNA precursors in *E. coli*.** The primary transcript contains a copy of each of the three ribosomal RNAs and may also contain several tRNA precursors. The large rRNA precursors are cleaved from the large primary transcript by the action of RNase III. The ends of the 16S, 23S, and 5S rRNAs are trimmed by the action of endonucleases M16, M23, and M5, respectively. (Slash marks indicate that portions of the rRNA primary transcript have been deleted for clarity.)

complete. In eukaryotes, on the other hand, transcription occurs in the nucleus, and translation takes place in the cytoplasm. This compartmentalization of functions in eukaryotic cells allows nuclear processing of mRNA precursors without disrupting translation.

Mature eukaryotic mRNA molecules are often derived from much larger primary transcripts. Subsequent processing of these primary transcripts includes some of the same steps that we saw in the previous section, namely: cleavage of a precursor, addition of terminal nucleotides, and covalent modification of nucleotides. Often, specific nucleotides (called intervening sequences, or introns) from the middle of an mRNA primary transcript are actually excised, or removed, and the resulting fragments are ligated together to produce the mature mRNA. This step, called **splicing**, is common in most eukaryotic species. Splicing also occurs during the processing of some eukaryotic tRNA and rRNA precursors (although these post-transcriptional modifications use a different splicing mechanism).

## A. Eukaryotic mRNA Molecules Have Modified Ends

All eukaryotic mRNA precursors undergo modifications that increase the stability of the mature mRNAs and make them better substrates for translation. One way to increase the stability of mRNAs is to modify their ends so that they are no longer susceptible to cellular exonucleases that degrade RNA.

The 5′ ends are modified while the mRNA precursors are still being synthesized. The 5′ end of the primary transcript is a nucleoside triphosphate residue (usually a purine) that was the first nucleotide incorporated by RNA polymerase II. Modification of this end begins when the gamma-phosphate group is removed by the action of a phosphohydrolase (Figure 21.27). The resulting 5′-diphosphate group then reacts with the *a*-phosphorus atom of a GTP molecule to create a 5′–5′ triphosphate linkage. This reaction is catalyzed by guanylyltransferase and produces a structure called a **cap**. The cap is often further modified by methylating the newly added guanylate. The 2′-hydroxyl groups of the first two nucleotides in the original transcript may also be methylated. Methyl groups for these reactions are donated by *S*-adenosylmethionine (Section 7.3).

**KEY CONCEPT**

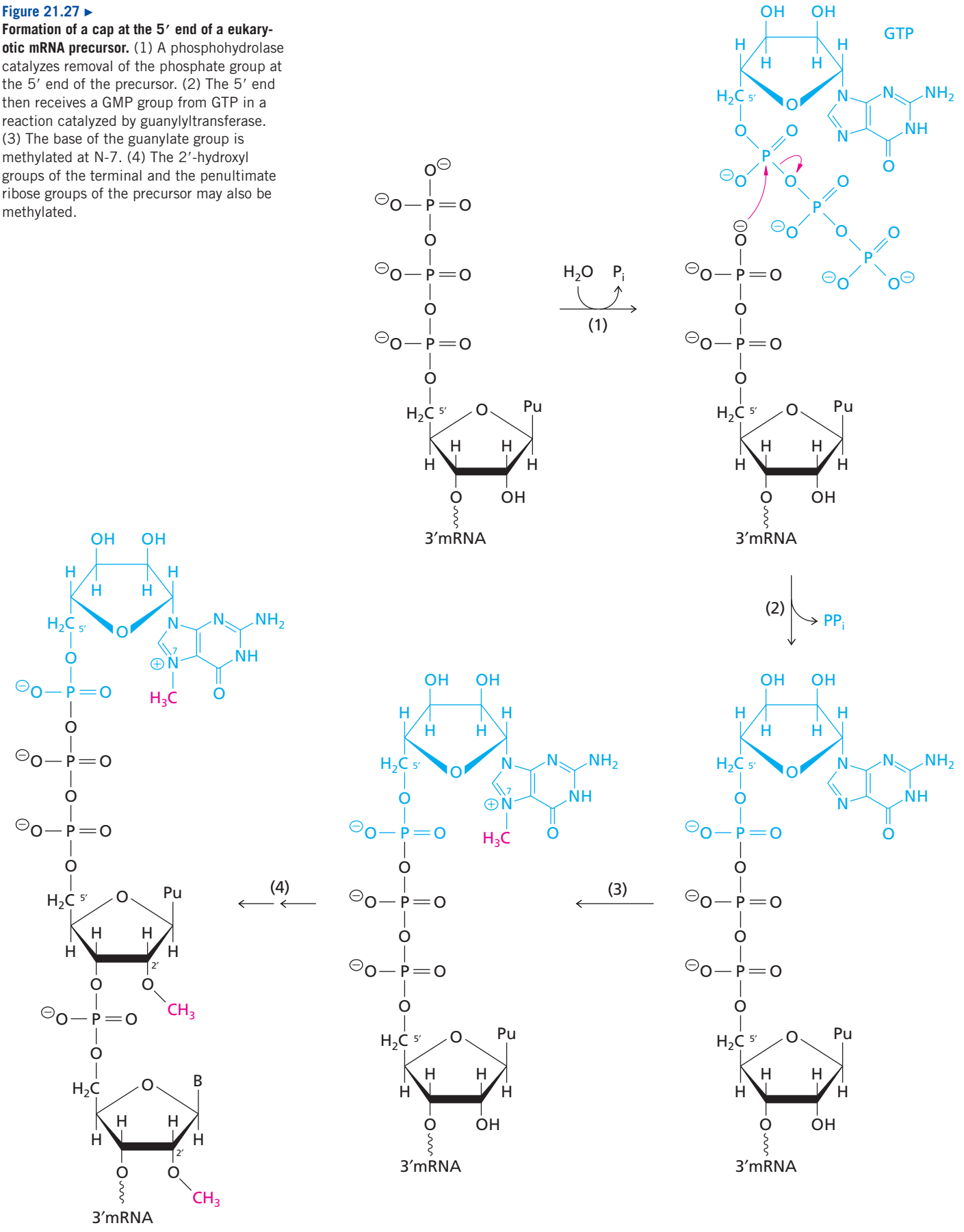
**Many eukaryotic coding sequences are interrupted by introns.**

The 5′–5′ triphosphate linkage protects the mRNA molecule from 5′ exonucleases by blocking its 5′ end. The cap also converts mRNA precursors into substrates for other processing enzymes in the nucleus, such as those that catalyze splicing. In mature mRNA, the cap is the site where ribosomes attach during protein synthesis. Capping is a cotranscriptional process that is confined to the nucleus. The capping enzymes shown in Figure 21.27 interact directly with RNA polymerase II transcription complexes but not with RNA polymerase I or RNA polymerase III complexes, ensuring that mRNA precursors are the only capped RNAs (i.e., tRNA and rRNA are not substrates for capping).

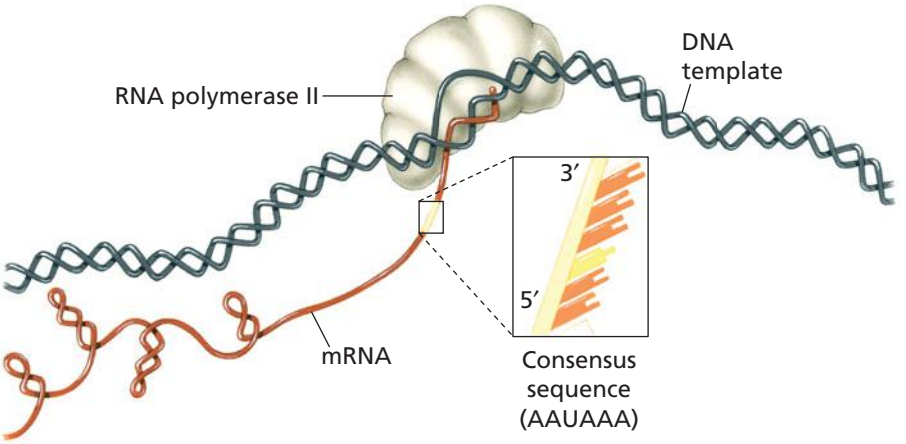
Eukaryotic mRNA precursors are also modified at their 3′ ends. Once RNA polymerase II has transcribed past the 3′ end of the coding region of DNA, the newly synthesized RNA is cleaved by an endonuclease downstream of a specific site whose consensus recognition sequence is AAUAAA. This sequence is bound by a cleavage and polyadenylation specificity factor (CPSF), a protein that also interacts with the endonuclease and a polymerase (Figure 21.28). After cleaving the RNA, the endonuclease dissociates and multiple adenylate residues are added to the newly generated 3′ end of the molecule. The addition reactions are catalyzed by poly A polymerase, which adds adenylate residues using ATP as a substrate. Up to 250 nucleotides can be added to form a stretch of polyadenylate known as a **poly A tail**.

With a few rare exceptions, all mature mRNA molecules in eukaryotes contain poly A tails. The length of the tail varies, depending on the species and possibly on the type of mRNA and the developmental stage of the cell. The length also depends on the age of the mRNA since the poly A tail is progressively shortened by the action of 3′ exonucleases. In fact, the tail has already been shortened by 50 to 100 nucleotides by the time the mature mRNA reaches the nuclear pores. The presence of the poly A tail increases the time required for the exonucleases to reach the coding region.

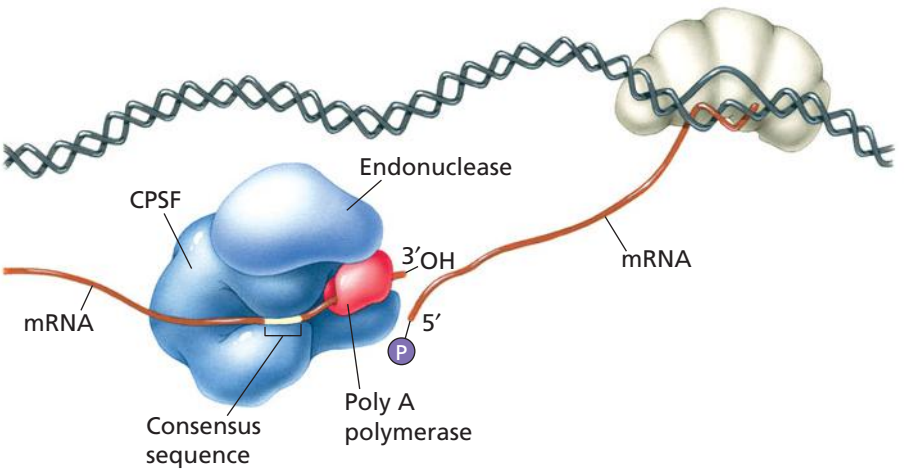
**Figure 21.27 ▸**
**Formation of a cap at the 5′ end of a eukaryotic mRNA precursor.** (1) A phosphohydrolase catalyzes removal of the phosphate group at the 5′ end of the precursor. (2) The 5′ end then receives a GMP group from GTP in a reaction catalyzed by guanylyltransferase. (3) The base of the guanylate group is methylated at N-7. (4) The 2′-hydroxyl groups of the terminal and the penultimate ribose groups of the precursor may also be methylated.



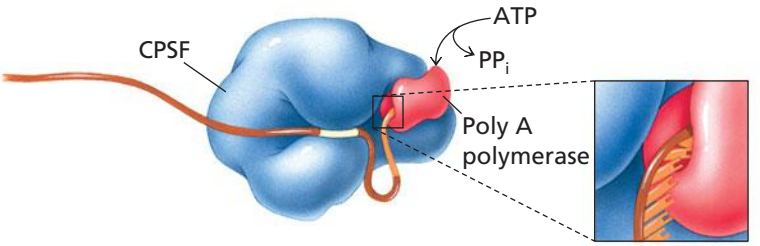
(a) Polyadenylation begins when RNA polymerase II transcription complex transcribes through a polyadenylation signal at the 3′ end of an mRNA precursor.




(b) CPSF binds to the consensus sequence and forms a complex containing an RNA endonuclease. The endonuclease catalyzes cleavage of the transcript downstream of the polyadenylation sequence, forming a new 3′ end. Poly A polymerase can then bind to the end of the mRNA precursor.




(c) The endonuclease dissociates and the new 3′ end of the RNA is polyadenylated by the activity of poly A polymerase.

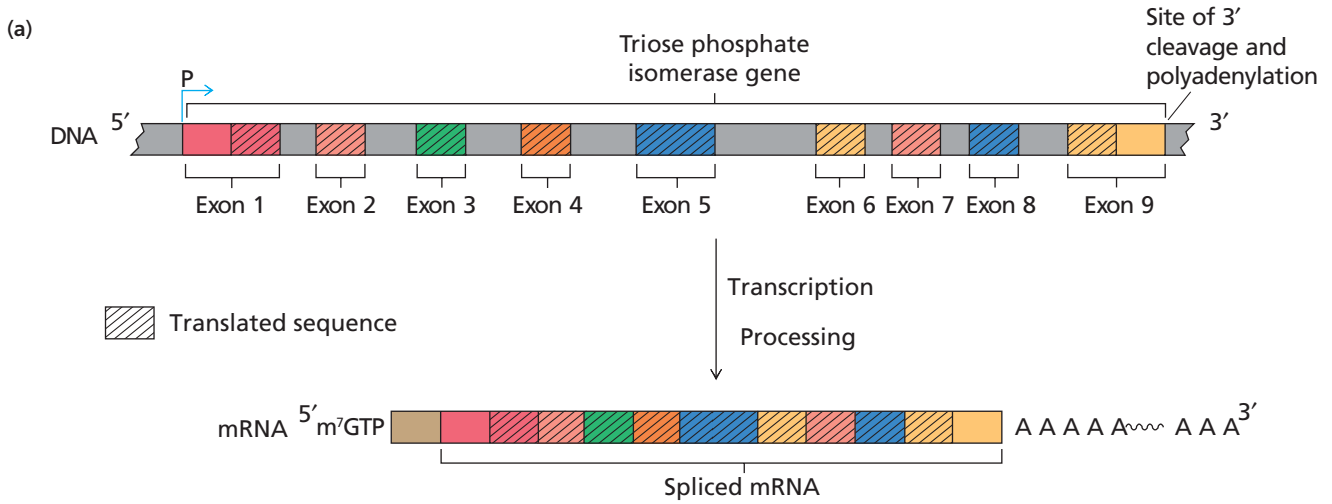



▲ **Figure 21.28**
**Polyadenylation of a eukaryotic mRNA precursor.**

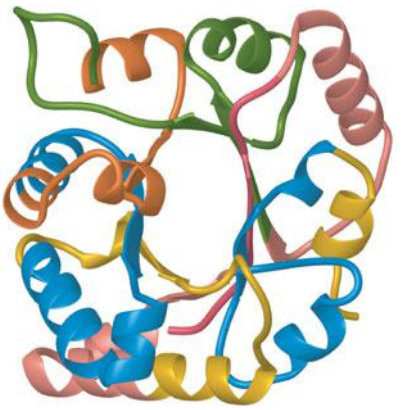
## B. Some Eukaryotic mRNA Precursors Are Spliced

Splicing is rare in prokaryotes but it is the rule in animals and flowering plants. Internal sequences that are removed from the primary RNA transcript are called **introns**. Sequences that are present in the primary RNA transcript and in the mature RNA molecule are called **exons**. The words *intron* and *exon* also refer to the regions of the gene (DNA) that encode corresponding RNA introns and exons. Since DNA introns are transcribed, they are considered part of the gene. The junctions of introns and exons are known as **splice sites** since these are the sites where the mRNA precursor is cut and joined.

Because of the loss of introns, mature mRNA is often a fraction of the size of the primary transcript. For example, the gene for triose phosphate isomerase from maize contains nine exons and eight introns and spans over 3400 bp of DNA. The mature




(b)

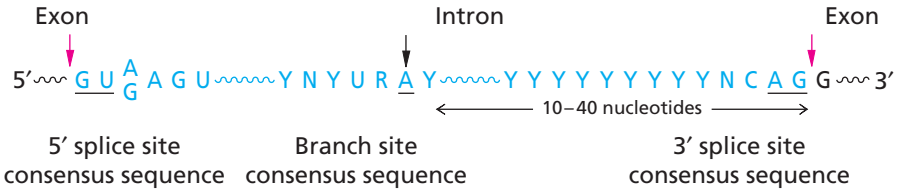


▲ **Figure 21.29**
**Triose phosphate isomerase gene from maize and the encoded enzyme. (a)** Diagram of the gene showing nine exons and eight introns. Some exons contain both translated and untranslated sequences. **(b)** Three-dimensional structure of the protein showing the parts of the protein encoded by each exon.

mRNA, which includes a poly A tail, is only 1050 nucleotides long (Figure 21.29). The enzyme itself contains 253 amino acid residues.

It used to be thought that there was a correlation between the intron/exon organization of a gene and the structure of the protein that the gene encodes. According to this hypothesis, exons encode protein domains and the presence of introns reflects the primitive organization of the gene. In other words, introns arose early in evolution. However, as shown in Figure 21.29b, there is no obvious correlation between exons and protein structure. Most biochemists and molecular biologists now believe that introns have been inserted at random locations during the evolution of a gene. The "introns late" hypothesis states that most primitive genes did not have introns and postulates that introns arose much later during the evolution of eukaryotes.

Introns can vary in length from as few as 42 bp to as many as 10,000 bp (the lower limit varies with each species; for example, most *C. elegans* introns are too small to be accurately spliced in either a vertebrate cell or cell-free extract). The nucleotide sequences at splice sites are similar in all mRNA precursors, but the sequence of the rest of the intron is not conserved. The vertebrate consensus sequences at the two splice sites are shown in Figure 21.30. Another short consensus sequence is found within the intron near the 3′ end. This sequence, known as the **branch site** or the branch-point sequence, also plays an important role in splicing.




◀ **Figure 21.30**
**Consensus sequences at splice sites in vertebrates**. Highly conserved nucleotides are underlined. Y represents a pyrimidine (U or C), R represents a purine (A or G), and N represents any nucleotide. The splice sites, where the RNA precursor is cut and joined, are indicated by red arrows, and the branch site is indicated by a black arrow. The intron is highlighted in blue.

The splicing of an mRNA precursor to remove a single intron requires two transesterification reactions: one between the 5′ splice site and the branch-site adenylate residue, and one between the 5′ exon and the 3′ splice site. The products of these two reactions are (1) the joined exons and (2) the excised intron in the form of a lariat-shaped molecule. These splicing reactions are catalyzed by a large RNA-protein complex called the **spliceosome**. The spliceosome helps to not only retain the intermediate splicing products but also position the splice sites so that the exons can be precisely joined (Figure 21.31).

The spliceosome is a large, multisubunit complex. It contains over 100 proteins and five molecules of RNA whose total length is about 5000 nucleotides. These RNA molecules are called small nuclear RNA (snRNA) molecules and are associated with proteins to form small nuclear ribonucleoproteins, or snRNPs (pronounced "snurps"). snRNPs are important not only in the splicing of mRNA precursors but also in other cellular processes.

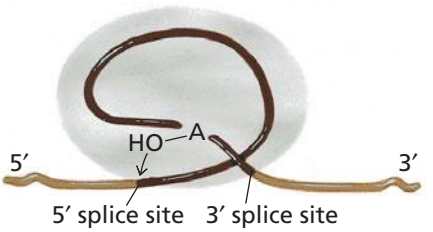
There are five different types of snRNAs—U1, U2, U4, U5, and U6. (U stands for uracil, a common base in these small RNA molecules.)—and a diploid vertebrate nucleus contains more than 100,000 total copies of snRNA. All five snRNA molecules are extensively base-paired and contain modified nucleotides. Each snRNP contains one or two snRNAs plus a number of proteins. Some of these proteins are common to all snRNPs; others are found in only one class of snRNP.

Biochemical experiments *in vitro* using purified components have led to a sequential model for spliceosome assembly (Figure 21.32). Spliceosome formation begins when a U1 snRNP binds to the newly synthesized 5′ splice site of the mRNA precursor. This interaction depends on base pairing between the 5′ splice site and a complementary sequence near the 5′ end of the U1 snRNA. A U2 snRNP then binds to the branch site of the intron, forming a stable complex that covers about 40 nucleotides. Next, a U5 snRNP associates with the 3′ splice site. Finally, a U4/U6 snRNP joins the complex, and all snRNPs are drawn together to form the spliceosome. Because several groups have now discovered that these same snRNPs are found preassembled in a much larger complex, prior to splicing, this pathway may not accurately reflect the splicing cycle *in vivo*.

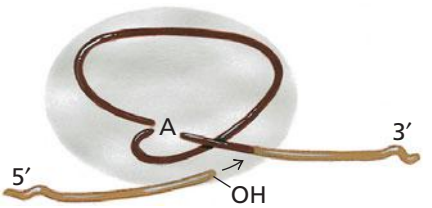
Binding of the U1, U2, and U5 snRNPs to consensus sequences at the 5′ splice site, branch site, and 3′ splice site of the intron positions these three interactive sites properly for the splicing reaction. The spliceosome then prevents the 5′ exon from diffusing away after cleavage and positions it to be joined to the 3′ exon. Once a spliceosome has formed at an intron, it is quite stable and can be purified from cell extracts.

Since spliceosomes can be observed on nascent transcripts, it is thought that intron removal is the rate limiting step in RNA processing. Since the spliceosome, which is almost as large as a ribosome, is too large to fit through the nuclear pores, the mRNA precursors are prevented from leaving the nucleus before processing is complete. Once an intron is excised, the spliceosome gets recycled and will repeat the catalytic cycle on the next intron it encounters.

(a) The spliceosome positions the adenylate residue at the branch site near the 5′ splice site. The 2′-hydroxyl group of the adenylate attacks the 5′ splice site.




(b) The 2′-hydroxyl group is attached to the 5′ end of the intron, and the newly created 3′-hydroxyl group of the exon attacks the 3′ splice site.




(c) As a result, the ends of the exons are joined, and the intron, a lariat-shaped molecule, is released.

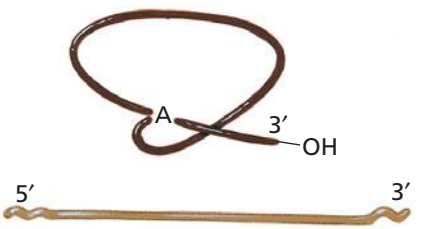
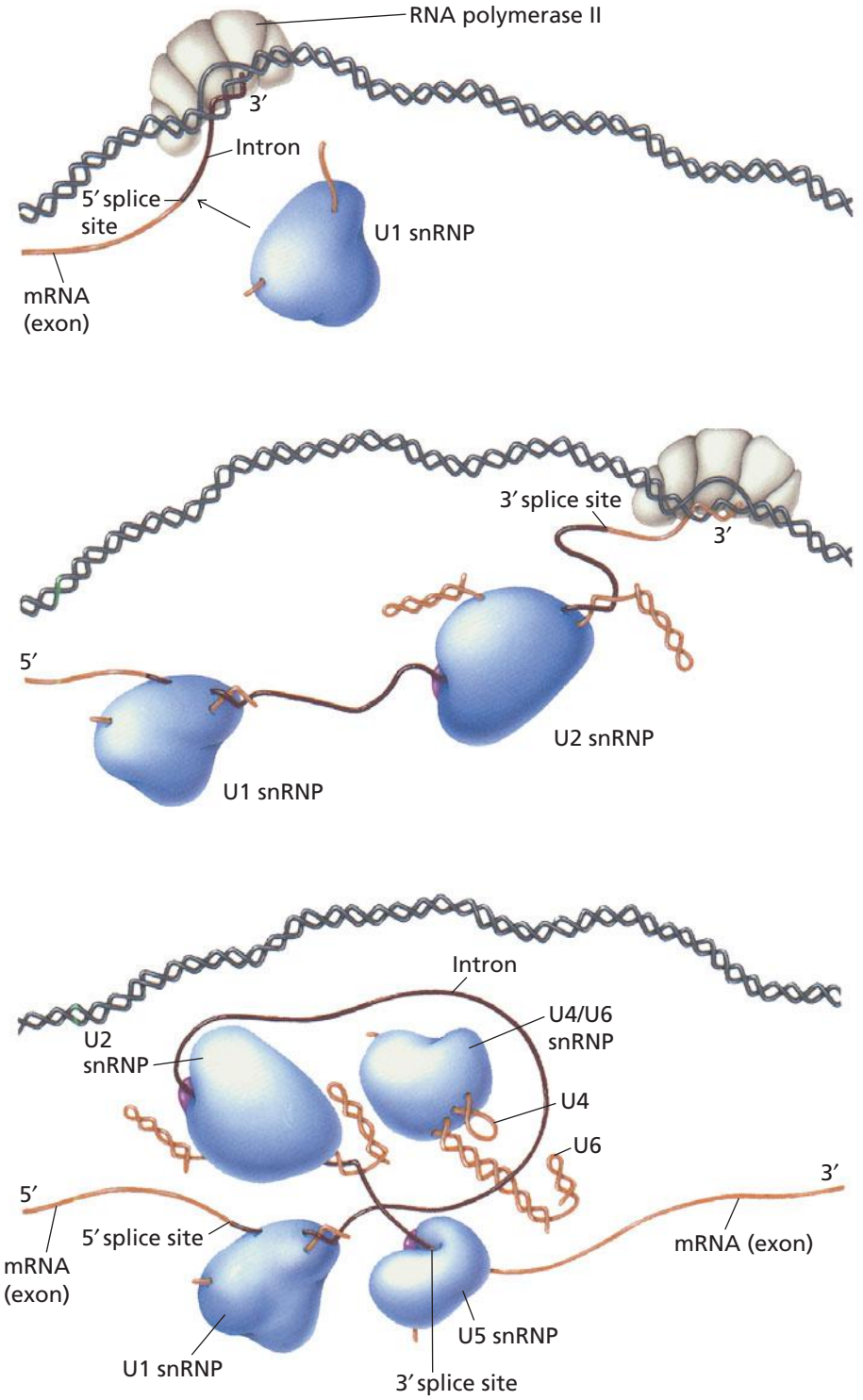



**Figure 21.31** ▶
**Intron removal in mRNA precursors**. The spliceosome, a multicomponent RNA-protein complex, catalyzes splicing.
[Adapted from Sharp, P. A. (1987). Splicing of messenger RNA precursors. Science 235:766–771.]




◀ **Figure 21.32**
**Formation of a spliceosome.**

**(a)** As soon as the 5′ splice site exits the transcription complex, a U1 snRNP binds to it.

**(b)** Next, a U2 snRNP binds to the branch site within the intron.

**(c)** When the 3′ splice site emerges from the transcription complex, a U5 snRNP binds, and the complete spliceosome assembles around a U4/U6 snRNP.

## Summary

1. A gene is a sequence of DNA that is transcribed. Housekeeping genes encode proteins and RNA molecules that are essential for normal cellular activities.
2. Cells contain several types of RNA, including transfer RNA, ribosomal RNA, messenger RNA, and small RNA molecules.
3. DNA-directed RNA synthesis, or transcription, is catalyzed by RNA polymerase. Ribonucleoside triphosphates are added in nucleotidyl-group–transfer reactions using a DNA strand as a template.
4. Transcription begins at a promoter sequence and proceeds in the 5′ → 3′ direction. A promoter consensus sequence indicates the nucleotides most commonly found at each position. The $\sigma$ subunit of *E. coli* RNA polymerase increases the affinity of the core polymerase for a promoter and decreases the affinity for nonpromoter sequences. During initiation, a transcription bubble forms and a short stretch of RNA is synthesized. The $\sigma$ subunit dissociates in the transition from initiation to chain elongation.

5. Transcription termination in *E. coli* occurs near pause sites, often when the RNA forms a hairpin structure. Some terminations require *rho*, which binds to single-stranded RNA.
6. In eukaryotes, several different RNA polymerases carry out transcription. Transcription factors interact with the promoter and RNA polymerase to initiate transcription.
7. Some genes are expressed constitutively, but the transcription of other genes is regulated. Transcription may be regulated by a repressor or an activator. These are often allosteric proteins.
8. Transcription of the three genes of the *lac* operon is blocked when *lac* repressor binds to two operators near the promoter. The repressor dissociates from the DNA when it binds the inducer allolactose. Transcription is activated by a complex of cAMP and CRP (cAMP regulatory protein).
9. RNA transcripts are frequently modified by processing, which includes the removal, addition, or modification of nucleotide residues. Primary transcripts of prokaryotic tRNA and rRNA are processed by nucleolytic cleavage and covalent modification.
10. Processing of mRNA in eukaryotes includes the addition of a 5′ cap and a 3′ poly A tail to protect the molecule from nuclease digestion. In some cases, introns are removed by splicing. The two transesterification reactions of splicing are catalyzed by the spliceosome, a complex containing small nuclear ribonucleoproteins (snRNPs).

# Problems

1. A bacterial RNA polymerase elongates RNA at a rate of 70 nucleotides per second, and each transcription complex covers 70 bp of DNA.
   (a) What is the maximum number of RNA molecules that can be produced per minute from a gene of 6000 bp? (Assume that initiation is not rate limiting.)
   (b) What is the maximum number of transcription complexes that can be bound to this gene at one time?
2. The *E. coli* genome is approximately 4600 kb in size and contains about 4000 genes. The mammalian genome is approximately $33 \times 10^6$ kb in size and contains at most 30,000 genes. An average gene in *E. coli* is 1000 bp long.
   (a) Calculate the percentage of *E. coli* DNA that is not transcribed.
   (b) Although many mammalian genes are larger than bacterial genes, most mammalian gene products are the same size as bacterial gene products. Calculate the percentage of DNA in exons in the mammalian genome.
3. There are a variety of methods that will allow you to introduce an intact eukaryotic gene (e.g., the triose phosphate isomerase gene) into a prokaryotic cell. Would you expect this gene to be properly transcribed by prokaryotic RNA polymerase? What about the converse situation, where an intact prokaryotic gene is introduced into a eukaryotic cell; will it be properly transcribed by a eukaryotic transcription complex?
4. Assume that, in a rare instance, a typical eukaryotic triose phosphate isomerase gene contains the correct sequences to permit accurate transcription in a prokaryotic cell. Would the resulting RNA be properly translated to yield the intact enzyme?
5. Describe how the rate of transcription of the *lac* operon is affected when *E. coli* cells are grown in the presence of (a) lactose plus glucose, (b) glucose alone, and (c) lactose alone.
6. In the promoter of the *E. coli lac* operon the −10 region has the sequence 5′-TATGTT-3′. A mutation named UV5 changes this sequence to 5′-TATAAT-3′ (see Figure 21.6). Transcription from the *lac* UV5 promoter is no longer dependent on the CRP-cAMP complex. Why?
7. When $\beta$-3$^{32}$P4-ATP is incubated with a eukaryotic cell extract that is capable of transcription and RNA processing, where does the label appear in mRNA?
8. Unlike DNA polymerase, RNA polymerase does not have proofreading activity. Explain why the lack of proofreading activity is not detrimental to the cell.
9. Mature mRNA from eukaryotic cells is often purified from other components in the cell with the use of columns containing oligo (dT) cellulose. These columns contain short segments of single-stranded deoxyribose thymidylate residues, oligo(dT), attached to a cellulose matrix. Explain the rationale for use of these columns to purify mature mRNA from a mixture of components.
10. Rifampicin is a semisynthetic compound made from rifamycin B, an antibiotic isolated from *Streptomyces mediterranei*. Rifampicin is an approved anti-mycobacterial drug that is a standard component of combination regimens for treating tuberculosis and staphylococci infections that resist penicillin. Recent studies have suggested that rifampicin-resistant tuberculosis is becoming more common. For example, 2% of samples from a survey in Botswana were found to be resistant to the drug. The table below gives some results from wild type *E. coli* and *E. coli* with a single amino acid change in the $\beta$ subunit of RNA polymerase (Asp to Tyr at amino acid position 516) and their growth response to media that contained rifampicin. [Severinov, K., Soushko, M., Goldfarb, A., and Nikiforov, V. (1993). Rifampicin region revisited. *J. Biol. Chem.* 268:14820–14825].

| ***E. coli*** | **Rifampicin[a] ($\mu$g/ml)** |
|---|---|
| Wild type | <5 |
| Asp516Tyr in $\beta$ subunit | >50 |

[a]Rifampicin concentration at the point of growth arrest of the *E. coli*.

   (a) What is your interpretation of the data?
   (b) What role does the $\beta$ subunit have in RNA polymerase?
   (c) Describe one mechanism for rifampicin-resistant bacteria.
11. A segment of DNA from the middle of an *E. coli* gene has the sequence below. Write the mRNA sequences that can be produced by transcribing this segment in either direction.

CCGGCTAAGATCTGACTAGC

12. Does the definition of a gene given on page 638 5e [first page of Chapter 21] apply to the rRNA and tRNA genes whose primary transcript is shown in Figure 21.26?

13. In general, if we know the genomic DNA sequence of a gene we can reliably predict the nucleotide sequence of the RNA encoded by that gene. Is this statement also true for tRNAs in prokaryotes? What about tRNAs in eukaryotes?
14. Assume that a spliceosome assembles at the first intron of the gene for triose phosphate isomerase in maize (Figure 21.29) almost as soon as the intron is transcribed (i.e., after about 500 nucleotides of RNA have been synthesized). How long must the spliceosome be stable if the splicing reaction cannot occur until transcription terminates? Assume that the rate of transcription by RNA polymerase II in maize is 30 nucleotides per second.
15. CRP-cAMP represses transcription of the *crp* gene. Predict the location of the CRP-cAMP binding site relative to the promoter of the *crp* gene.
16. Why are mutations within an intron of a protein-coding gene sometimes detrimental?
17. A deletion in one of the introns in the gene for the triose phosphate isomerase moves the branch site to a new location seven nucleotides away from the 3′-splice acceptor sequence. Will this deletion have any affect on splicing of the gene?

# Selected Readings


## General

Alberts, B., Johnson, A., Lewis, J., and Raff, M. (2007). *Molecular Biology of the Cell*, 5th ed. (New York: Garland).

Krebs, J., Goldstein, L., and Kilpatrick, S. (2009). *Lewin's Genes X* (New York: Jones & Bartlett).

## RNA Polymerases and Transcription

Ardehali, M. B., and Lis, J. T. (2009). Tracking rates of transcription and splicing *in vivo*. *Nature Structural & Molecular Biology* 16:1123–1124.

Bushnell, D. A., and Kornberg, R. D. (2003). Complete, 12-subunit RNA polymerase II and 4.1-A resolution: implications for the initiation of transcription. *Proc. Natl. Acad. Sci. (U.S.A.)* 100:6969–6973.

Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends Cell Biol.* 9:M46–M49.

Lisser, S., and Margalit, H. (1993). Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res.* 21:1507–1516.

Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., and Darst, S. A. (2002). Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* 296:1285–1290.

Richardson, J. P. (1993). Transcription termination. *Crit. Rev. Biochem.* 28:1–30.

## Regulation of Transcription

Becker, P. B., and Horz, W. (2002). ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.* 71:247–273.

Bushman, F. D. (1992). Activators, deactivators and deactivated activators. *Curr. Biol.* 2:673–675.

Fuda, N. J., Behfar, M., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription *in vivo*. *Nature* 461:186–192.

Harrison, S. C., and Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* 59:933–969.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318–356.

Kolb, A., Busby, S., Buc, H., Garges, S., and Adhya, S. (1993). Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.* 62:749–795.

Myers, L. C., and Kornberg, R. D. (2000). Mediator of transcriptional regulation. *Annu. Rev. Biochem.* 69:729–749.

Pan, Y., Tsai, C.-J., Ma, B., and Nussinov, R. (2009). How do transcription factors select specific binding sites in the genome? *Nature Structural & Molecular Biology* 16:1118–1120.

Wolfe, A. P., and Guschin, D. (2000). Review: chromatin structural features and targets that regulate transcription. *J. Struct. Biol.* 129:102–122.

Workman, J. L., and Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.* 67: 545–579.

## RNA Processing

Apirion, D., and Miczak, A. (1993). RNA processing in prokaryotic cells. *BioEssays* 15:113–120.

Collins, C. A., and Guthrie, C. (2000). The question remains: is the spliceosome a ribozyme? *Nature Struct. Biol.* 7: 850–854.

James, B. D., Olsen, G. J., Liu, J., and Pace, N. R. (1988). The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. *Cell* 52:19–26.

Jurica, M. S., and Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Molecular Cell* 12:5–14.

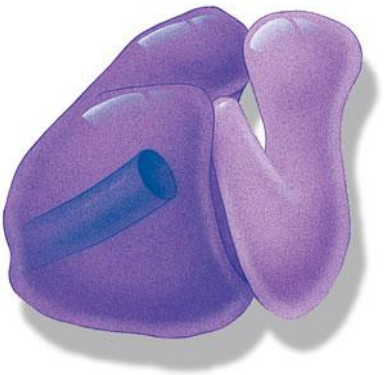
McKeown, M. (1993). The role of small nuclear RNAs in RNA splicing. *Curr. Biol.* 5:448–454.

Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *BioEssays* 25:1147–1149.

Proudfoot, N. (2000). Connecting transcription to messenger RNA processing. *Trends Biochem. Sci.* 25:290–293.

Shatkin, A. J., and Manley, J. L. (2000). The ends of the affair: capping and polyadenylation. *Nature Struct. Biol.* 7: 838–842.

Wahle, E. (1992). The end of the message: 3′-end processing leading to polyadenylated messenger RNA. *BioEssays* 14:113–118.

# Protein Synthesis

We are now ready to examine the final stage of biological information flow: the translation of mRNA and the polymerization of amino acids into proteins. The essential features of the biochemistry of protein synthesis were worked out in the decade between 1955 and 1965. It was clear that there was a genetic code that had to be used to translate a nucleotide sequence into a sequence of amino acids. In 1955, Francis Crick proposed that the first step in this process was the attachment of an amino acid to a small adapter RNA. Shortly after that, the adapters, now known as transfer RNAs, were identified. Ribosomes and the other essential components of the translation machinery were discovered by fractionating cells and reconstituting protein synthesis *in vitro*. Workers in several laboratories demonstrated that messenger RNA is one of the key intermediates in the flow of information from DNA to protein. By 1961, the most important missing ingredient was the nature of the genetic code.

We begin this chapter with a discussion of the genetic code and tRNA structure. Next, we examine how mRNA, tRNA, ribosomes, and accessory proteins participate in protein synthesis. We will also present some examples of the regulation of translation and post-translational processing.

*The results indicate that polyuridylic acid contains the information for the synthesis of a protein having many of the characteristics of poly-L-phenylalanine. . . . One or more uridylic acid residues therefore appear to be the code for phenylalanine. Whether the code is of the singlet, triplet, etc., type has not yet been determined. Polyuridylic acid seemingly functions as a synthetic template or messenger RNA, and this stable, cell-free* E. coli *system may well synthesize any protein corresponding to meaningful information contained in added RNA.*

—M. Nirenberg and H. Matthaei, 1961

## 22.1 The Genetic Code

George Gamow first proposed the basic structural units of the genetic code. He reasoned that since the DNA "alphabet" consists of only four "letters" (A, T, C, and G) and since these four letters encode 20 amino acids, the genetic code might contain "words," or **codons**, with a uniform length of three letters. Two-letter words constructed from any combination of the four letters produce a vocabulary of only 16 words ($4^2$), not enough for all 20 amino acids. In contrast, four-letter words produce a vocabulary of 256 words ($4^4$), far more than are needed. Three-letter words allow a possible vocabulary of 64 words ($4^3$), more than sufficient to specify each of the 20 amino acids but not excessive.

**Top:** *Escherichia coli* ribosome. The ribosome, a complex of RNA and protein, is the site where genetic information is translated into protein.


▲ **The enigma cryptography machine used by German armed forces during the Second World War.** This mechanical typewriter permitted the user to adjust its three large dials to encrypt outgoing messages before being sent by telegraph. The recipients could decode the message by setting the dials on their enigma machine to match. This type of encryption is extremely difficult to decipher, but when Allied forces were able to capture an intact enigma machine they could listen in on all their enemy's transmissions.

The "cracking" of the genetic code began with a chance observation by Marshall Nirenberg and J. Heinrich Matthaei. They discovered that polyuridylate (poly U) could direct the synthesis of polyphenylalanine *in vitro*. By showing that UUU encodes phenylalanine, they identified the first codon.

Between 1962 and 1965, the rest of the code was deciphered by a number of workers, chiefly Nirenberg and H. Gobind Khorana. Overall, it took ten years of hard work to learn how mRNA encodes proteins. The development of methods for sequencing genes and proteins has allowed direct comparison of the primary sequences of proteins with the nucleotide sequences of their corresponding genes. Each time a new protein and its gene are characterized, the genetic code is confirmed.

Transfer RNA (tRNA) plays an important role in interpreting the genetic code and translating a nucleotide sequence into an amino acid sequence. tRNAs are the adapters between mRNA and proteins. One region of a tRNA molecule is covalently linked to a specific amino acid, while another region on the same tRNA molecule interacts directly with an mRNA codon by complementary base pairing. It is this processive joining of the amino acids specified by an mRNA template that allows the precise synthesis of proteins.

In principle, a genetic code made up of three-letter words can be either overlapping or nonoverlapping (Figure 22.1). If the codons overlap, then each letter is part of more than one word and mutating a single letter changes several words simultaneously. For example, in the sequence shown in Figure 22.1a, each letter is part of three different words in an overlapping code. One of the advantages of a nonoverlapping code (Figure 22.1b) is that each letter is part of only one word; therefore, mutating a single nucleotide affects only one codon. All living organisms use a nonoverlapping genetic code.

Even with a nonoverlapping code, a sequence can be translated in many different ways, depending on where translation begins. (We will see later that translation does not typically begin with the very first nucleotide in an mRNA.) Each potential translation initiation point defines a unique sequence of three-letter words, or **reading frame**, in the mRNA. The correct translation of the "message" transcribed, or written, in the genetic code depends on establishing the correct reading frame for translation (Figure 22.2).

The standard genetic code is shown in Figure 22.3. With a few minor exceptions, all living organisms use this genetic code, suggesting that all modern species are descended from a common ancestor that also used the standard genetic code. This ancestral species probably lived billions of years ago, making the genetic code one of the most ancient remnants of early life.

By convention, all nucleotide sequences are written in the 5′ → 3′ direction. Thus, UAC specifies tyrosine, and CAU specifies histidine. The term *codon* usually refers to triplets of nucleotides in mRNA but it can also apply to triplets of nucleotides in the DNA sequence of a gene. For example, one DNA codon for tyrosine is TAC.

Codons are always translated 5′ → 3′, beginning near the 5′ end of the message (i.e., the end synthesized first) and proceeding to the end of the coding region that is

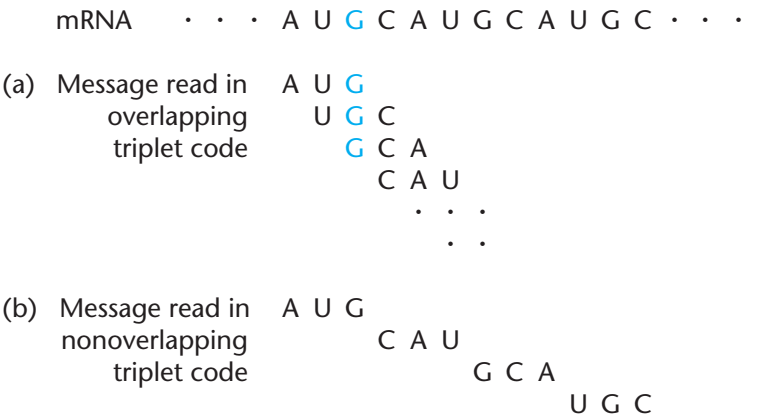


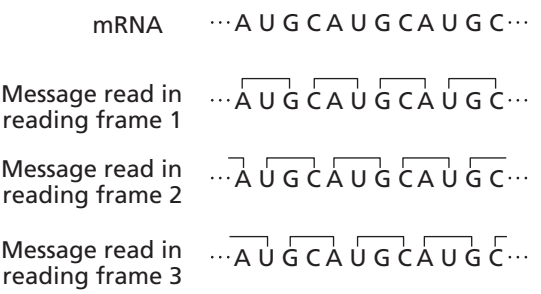

**Figure 22.1 ►**
**Message read in (a) overlapping and (b) nonoverlapping three-letter codes.** In an overlapping code, each letter is part of three different three-letter words (as indicated for the letter G in blue); in a nonoverlapping code, each letter is part of only one three-letter word.

usually near the 3′ end of the mRNA. The correct reading frame is specified by special punctuation signals that mark the beginning and the end.

The standard genetic code has several prominent features:

1. The genetic code is unambiguous. In a particular organism or organelle each codon corresponds to one, and only one, amino acid.
2. There are multiple codons for most amino acids. For example, leucine is the most abundant amino acid found in proteins (Table 3.3) and has six codons. Because of the existence of several codons for most amino acids, the genetic code is said to be **degenerate**. Different codons that specify the same amino acid (e.g., UCU and CGU both specify Ser; ACA, ACC, ACG, and ACU all specify Thr) are known as **synonymous codons**.
3. The first two nucleotides of a codon are often enough to specify a given amino acid. For example, the four codons for glycine (GGU, GGC, GGA, and GGG) all begin with GG.
4. Codons with similar sequences specify chemically similar amino acids. For example, the codons for threonine differ from four of the codons for serine by only a single nucleotide at the 5′ position and the codons for aspartate and glutamate begin with GA and differ only at the 3′ position. In addition, codons with pyrimidines at the second position usually encode hydrophobic amino acids. Therefore, mutations that alter either the 5′ or the 3′ position of these codons usually result in the incorporation of a chemically similar amino acid into the protein.
5. Only 61 of the 64 codons specify amino acids. The three remaining codons (UAA, UGA, and UAG) are **termination codons**, or **stop codons**. Termination codons are not normally recognized by any tRNA molecules in the cell. Instead, they are recognized by specific proteins that cause newly synthesized peptides to be released from the translation machinery. The methionine codon, AUG, also specifies the initiation site for protein synthesis and is often called the **initiation codon**.

Since the completion of the first draft of the human genome in 2000, it has been common to read in the popular press of "deciphering the code of life" or "unlocking the human genetic code." Strictly speaking, the information in the human genome is encoded using the same "universal" genetic code discovered 50 years ago. Sequencing projects actually reveal the *messages* encoded by the genes and not the code itself.




▲ **Figure 22.2**
**One mRNA contains three different reading frames.** The same string of letters read in three different reading frames will be translated into three different "messages" or protein sequences. Thus, translation of the correct message requires selecting the correct reading frame.

**INTERNATIONAL MORSE CODE**
Time of dash equals three dots

| | | |
|---|---|---|
| A • — | N — • | 1 • — — — — |
| B — • • • | O — — — | 2 • • — — — |
| C — • — • | P • — — • | 3 • • • — — |
| D — • • | Q — — • — | 4 • • • • — |
| E • | R • — • | 5 • • • • • |
| F • • — • | S • • • | 6 — • • • • |
| G — — • | T — | 7 — — • • • |
| H • • • • | U • • — | 8 — — — • • |
| I • • | V • • • — | 9 — — — — • |
| J • — — — | W • — — | 0 — — — — — |
| K — • — | X — • • — | |
| L • — • • | Y — • — — | |
| M — — | Z — — • • | |

▲ **Morse code permitted text to be sent by telegraph.** Messages written in the Latin alphabet and/or Arabic numerals could be transmitted via electrical wires using a code invented by Samuel Morse. In the Morse code the most common letters in English language text are coded by the shortest sequence of dashes and dots (allowing messages to be sent with the fewest number of symbols).

| First position (5′ end) | Second position U | C | A | G | Third position (3′ end) |
|---|---|---|---|---|---|
| U | Phe | Ser | Tyr | Cys | U |
| | Phe | Ser | Tyr | Cys | C |
| | Leu | Ser | STOP | STOP | A |
| | Leu | Ser | STOP | Trp | G |
| C | Leu | Pro | His | Arg | U |
| | Leu | Pro | His | Arg | C |
| | Leu | Pro | Gln | Arg | A |
| | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U |
| | Ile | Thr | Asn | Ser | C |
| | Ile | Thr | Lys | Arg | A |
| | Met | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U |
| | Val | Ala | Asp | Gly | C |
| | Val | Ala | Glu | Gly | A |
| | Val | Ala | Glu | Gly | G |

◀ **Figure 22.3**
**Standard genetic code.** The standard genetic code is composed of 64 triplet codons. The left-hand column indicates the nucleotide found at the first (5′) position of the codon; the top row indicates the nucleotide found at the second (middle) position of the codon; and the right column indicates the nucleotide found at the third (3′) position of the codon. The codon AUG specifies methionine (Met) and is also used to initiate protein synthesis. STOP indicates a termination codon.

## 22.2 Transfer RNA

Transfer RNA molecules are the interpreters of the genetic code. They are the crucial link between the sequence of nucleotides in mRNA and the sequence of amino acids in the corresponding polypeptide. In order for tRNA to fulfill this role, every cell must contain at least 20 different tRNA species (one for every amino acid) and each tRNA must recognize at least one codon.

### A. The Three-Dimensional Structure of tRNA

The nucleotide sequences of different tRNA molecules from many organisms have been determined. The sequences of almost all these molecules are compatible with the secondary structure shown in Figure 22.4. This "cloverleaf" structure contains several arms that are composed of a loop or a loop with a hydrogen-bonded stem. The double-stranded region of each arm forms a short, stacked, right-handed helix similar to that of double-stranded DNA.

The 5′ end and the region near the 3′ end of the tRNA molecule are base-paired to each other forming the *acceptor stem* (or amino acid stem). The activated amino acid will be covalently attached to tRNA on the 3′ end of this stem. The amino acid's carboxyl group gets linked to the terminal adenylate's ribose on either its 2′- or 3′-hydroxyl group (Recall from Section 21.8A that mature tRNA molecules are produced by processing a larger primary transcript and that the nucleotides at the 3′ end of a mature tRNA molecule are invariably CCA.) All tRNA molecules have a phosphorylated nucleotide on the 5′ end.

The single-stranded loop opposite the acceptor stem in the cloverleaf structure is called the anticodon loop. It contains the **anticodon**, the three-base sequence that binds to a complementary codon in mRNA. The arm of the tRNA molecule that contains the anticodon is called the *anticodon arm.* The remaining two arms of the tRNA molecule are named for the covalently modified nucleotides found within them. (See Figure 21.25

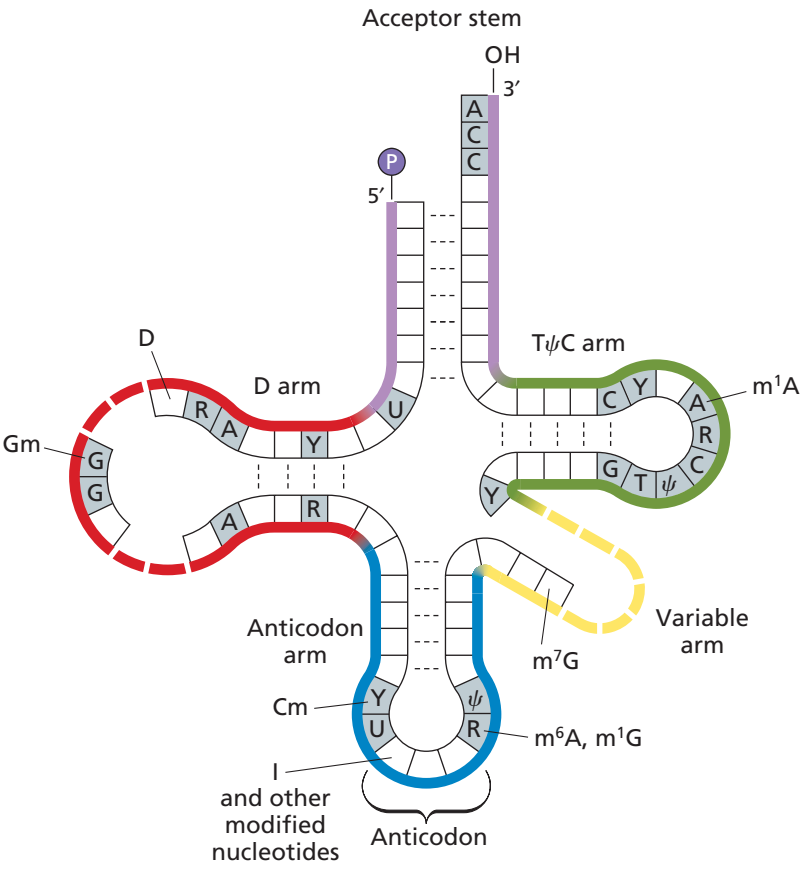
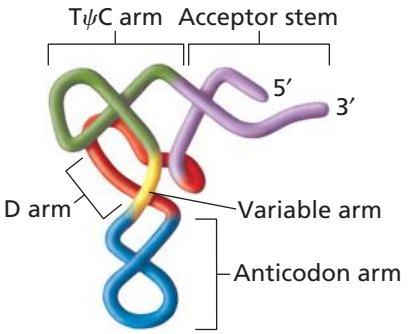



**Figure 22.4 ►**
**Cloverleaf secondary structure of tRNA.** Watson-Crick base pairing is indicated by dashed lines between nucleotide residues. The molecule is divided into an acceptor stem and four arms. The acceptor stem is the site of amino acid attachment, and the anticodon arm is the region of the tRNA molecule that interacts with mRNA codons. The D and TψC arms are named for modified nucleotides that are conserved within these arms. The number of nucleotide residues in each arm is more or less constant (except in the variable arm). Conserved bases (gray) and positions of common modified nucleotides are noted. Abbreviations other than standard nucleotides: R, a purine nucleotide; Y, a pyrimidine nucleotide; $m^1A$, 1-methyladenylate; $m^6A$, $N^6$-methyladenylate; Cm, 2′-O-methylcytidylate; D, dihydrouridylate; Gm, 2′-O-methylguanylate; $m^1G$, 1-methylguanylate; $m^7G$, 7-methylguanylate; I, inosinate; ψ, pseudouridylate; T, thymine ribonucleotide.

for the structures of these nucleotides.) One of the arms, called the TψC arm, always contains the triplet sequence ribothymidylate (T), pseudouridylate (ψ), and cytidylate (C). Dihydrouridylate (D) residues lend their name to the *D arm*. tRNA molecules also have a *variable arm* between the anticodon arm and the *TψC arm*. The variable arm ranges in length from about 3 to 21 nucleotides. With a few rare exceptions, tRNA molecules are between 73 and 95 nucleotides long.

The cloverleaf diagram of tRNA is a two-dimensional representation of a three-dimensional molecule. In three dimensions, the tRNA molecule is folded into a sideways "L" shape (Figures 22.5 and 22.6). The acceptor stem is at one end of the L-shaped molecule, and the anticodon is located in a loop at the opposite end. The resulting structure is compact and very stable, in part because of hydrogen bonds between the nucleotides in the D, TψC, and variable arms. This base pairing differs from normal Watson-Crick base pairing. Most of the nucleotides in tRNA are part of two perpendicular stacked helices. The interactions between the adjacent stacked base pairs are additive and make a major contribution to tRNA stability (analogous to the role of base stacking interactions in the 3D structure of double-stranded DNA we described in Section 19.2C).

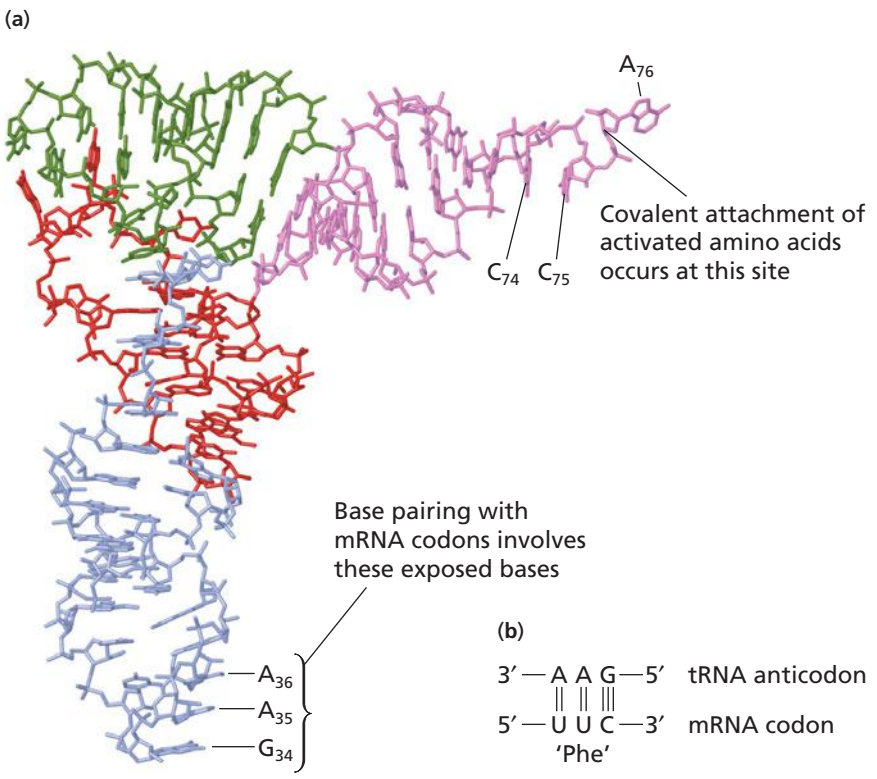



▲ **Figure 22.5**
**Tertiary structure of tRNA.** The cloverleaf-shaped molecule shown in Figure 22.4 actually folds up into this three-dimensional shape. The tertiary structure of tRNA results from base pairing between the TψC loop and the D loop, and two stacking interactions that **(a)** align the TψC arm with the acceptor arm, and **(b)** align the D arm with the anticodon arm. For clarity, only the ribose-phosphate backbone is shown here.

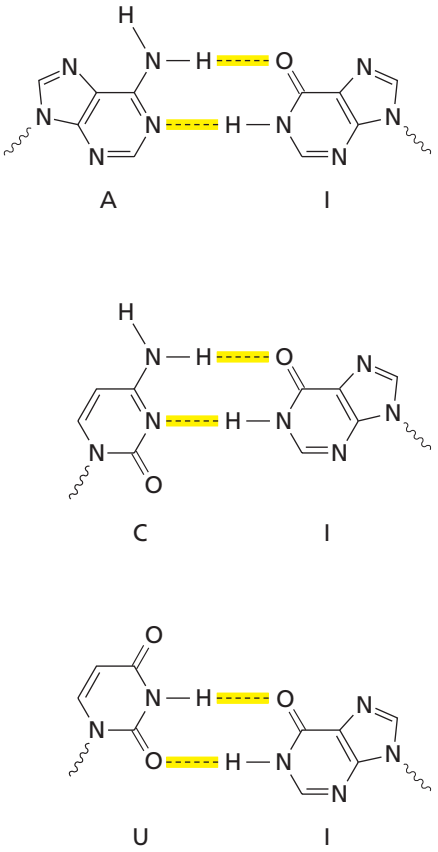
## B. tRNA Anticodons Base-Pair with mRNA Codons

tRNA mediated decoding of the information stored in mRNA molecules requires base-pairing interactions between tRNA anticodons and complementary mRNA codons. The anticodon of a tRNA molecule therefore determines where the amino acid attached to its acceptor stem is added to a growing polypeptide chain. Transfer RNA molecules are named for the amino acid they carry. For example, the tRNA molecule shown in Figure 22.6 has the anticodon GAA that binds to the phenylalanine codon UUC. Prior to protein synthesis, phenylalanine is covalently attached to the acceptor stem of this tRNA. The molecule is therefore designated $tRNA^{Phe}$.

Much of the base pairing between the codon and the anticodon is governed by the rules of Watson-Crick base pairing: A pairs with U, G pairs with C, and the strands in the base-paired region are antiparallel. However, some exceptions to these rules led Francis




◀ **Figure 22.6**
**Structure of $tRNA^{Phe}$ from the yeast *Saccharomyces cerevisiae*.** **(a)** Stick model showing base pairs and the position of the D arm (red) relative to the TψC arm (green). Note that there are two double-stranded RNA helices arrayed at right angles to each other to form an L-shaped structure. **(b)** Diagram showing the complement base-pairing between $tRNA^{Phe}$ and a phe codon to generate a double-stranded, antiparallel RNA helix during decoding. [NDB TRNA10].




▲ **Figure 22.7**
**Inosinate base pairs.** Inosinate (I) is often found at the 5′ (wobble) position of a tRNA anticodon. Inosinate can form hydrogen bonds with A, C, or U. This versatility in hydrogen bonding allows a tRNA carrying a single anticodon to recognize more than one synonymous codon.

**Table 22.1 Predicted base pairing between the 5′ (wobble) position of the anticodon and the 3′ position of the codon**

| Nucleotide at 5′ (wobble) position of anticodon | Nucleotide at 3′ position of codon |
|---|---|
| C | G |
| A | U |
| U | A or G |
| G | U or C |
| I[a] | U, A, or C |

[a] I = Inosinate.

Crick to propose that complementary Watson-Crick base pairing is required for only two of the three base pairs formed. The codon must form Watson-Crick base pairs with the 3′ and middle bases of the anticodon but other types of base pairing are permitted at the 5′ position of the anticodon. This alternate pairing suggests that the 5′ position is conformationally flexible. Crick dubbed this flexibility "wobble" and the 5′ position of the anticodon is sometimes called the wobble position.

Table 22.1 summarizes the allowable base pairs between the wobble position of an anticodon and the third nucleotide of an mRNA codon. When G is at the wobble position, for example, it can pair with either C or U (!). The base at the wobble position of many anticodons is covalently modified permitting additional flexibility in codon recognition. For example, in several tRNA molecules, G at the 5′ anticodon position is deaminated at C-2 to form inosinate (I), which can hydrogen-bond with A, C, or U (Figure 22.7). The presence of I at the 5′ position of the anticodon explains why $tRNA^{Ala}$ with the anticodon IGC can bind to three different codons specifying alanine: GCU, GCC, and GCA (Figure 22.8).

Wobble allows some tRNA molecules to recognize more than one codon but several different tRNA molecules are often required to recognize all synonymous codons. Different tRNA molecules that can attach to the same amino acid are called **isoacceptor tRNA molecules**. The term *isoacceptor* describes not only tRNA molecules with different anticodons that are covalently attached to the same activated amino acid but also tRNA molecules with the same anticodon but different primary sequences. Isoacceptor tRNAs are identified by Roman numerals or by the codons they recognize (i.e., $tRNA_I^{Ala}$, $tRNA_{II}^{Ala}$, or $tRNA_{GCG}^{Ala}$).

Genome sequencing data reveal that bacterial genomes encode 30 to 60 different tRNAs and that eukaryotic genomes have genes for as many as 80 different tRNA molecules. Many of the eukaryotic tRNA genes are present in multiple copies, especially those genes that encode abundant tRNAs used most frequently in protein synthesis.

## 22.3 Aminoacyl-tRNA Synthetases

Like DNA and RNA synthesis, protein synthesis can be divided into three distinct stages: initiation, chain elongation, and termination. However, our description of translation includes a step prior to the initiation of polymerization, namely, aminoacylation of tRNA. The activation of amino acids is considered part of the overall translation process because it is such an important part of the flow of biological information from nucleic acid to protein.

Each of the 20 amino acids is covalently attached to the 3′ end of its respective tRNA molecules. The product of this reaction is called an **aminoacyl-tRNA**. The amino acid is said to be "activated" for subsequent transfer to a growing polypeptide chain because the aminoacyl-tRNA is a "high-energy" molecule. A specific aminoacyl-tRNA molecule is identified by naming both the tRNA and the attached amino acid;

for example, aminoacylated $tRNA^{Ala}$ is called alanyl-$tRNA^{Ala}$. The various enzymes that catalyze the aminoacylation reaction are called aminoacyl-tRNA synthetases (e.g., alanyl-tRNA synthetase).

Most species have at least 20 different aminoacyl-tRNA synthetases in each cell since there are 20 different amino acids. A few species have two different aminoacyl-tRNA synthetases for the same amino acid. Some bacteria don't have glutaminyl- or asparaginyl-tRNA synthetases. In these species, the glutaminyl- and asparaginyl-tRNAs are synthesized by modifying glutamate and aspartate residues after they have been covalently attached to $tRNA^{Gln}$ and $tRNA^{Asn}$ by glutamyl- and aspartyl-tRNA synthetases (Glutamate and aspartate residues that are bound to their proper tRNAs are not modified.)

Although each synthetase is specific for a particular amino acid, it can recognize many isoacceptor tRNA molecules. For example, there are six codons for serine and several different isoacceptor $tRNA^{Ser}$ molecules. All these different $tRNA^{Ser}$ molecules are recognized by the organism's single seryl-tRNA synthetase enzyme. The accuracy of protein synthesis depends on the ability of aminoacyl-tRNA synthetases to catalyze attachment of the correct amino acid to its corresponding tRNA.

## A. The Aminoacyl-tRNA Synthetase Reaction

The activation of an amino acid by its specific aminoacyl-tRNA synthetase requires ATP. The overall reaction is:

$$\text{Amino Acid} + \text{tRNA} + \text{ATP} \longrightarrow \text{Aminoacyl-tRNA} + \text{AMP} + PP_i \qquad \textbf{(22.1)}$$

The amino acid is covalently attached to the tRNA molecule by the formation of an ester linkage between the carboxylate group of the amino acid and a hydroxyl group of the ribose at the 3′ end of the tRNA molecule. Since all tRNAs end in —CCA, the attachment site is always an adenylate residue.

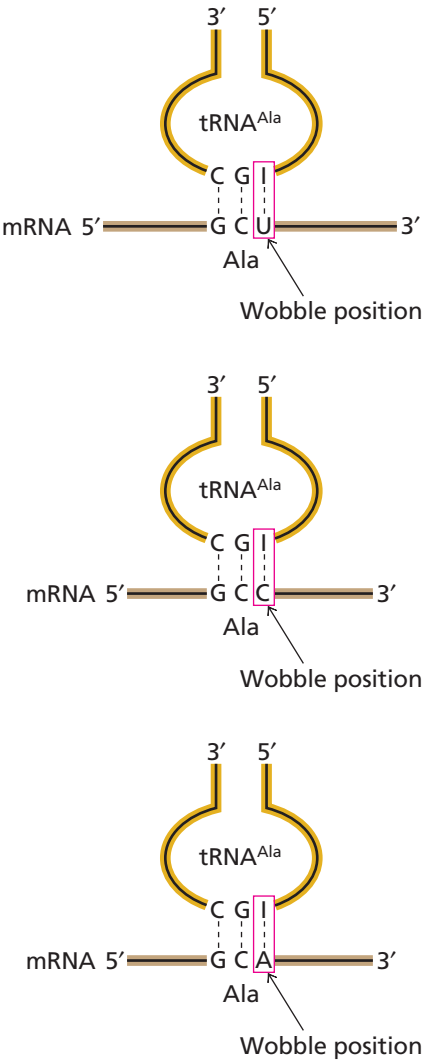
Aminoacylation proceeds in two discrete steps (Figure 22.9). In the first step, the amino acid is activated by formation of a reactive aminoacyl-adenylate intermediate. The intermediate remains tightly but noncovalently bound to the aminoacyl-tRNA synthetase. Rapid hydrolysis of the liberated pyrophosphate strongly favors the forward reaction. The second step of aminoacyl-tRNA formation is aminoacyl-group transfer from the aminoacyl-adenylate intermediate to tRNA. The amino acid is attached to either the 2′- or the 3′-hydroxyl group of the terminal adenylate residue of tRNA, depending on the specific aminoacyl-tRNA synthetase catalyzing the reaction. If the amino acid is initially attached to the 2′-hydroxyl group, it is shifted to the 3′-hydroxyl group in an additional step. The amino acid must be attached to the 3′ position to function as a protein synthesis substrate.

Formation of the aminoacyl-tRNA is favored under cellular conditions and the intracellular concentration of free tRNA is very low. The Gibbs free energy of hydrolysis of an aminoacyl-tRNA is approximately equivalent to that of a phosphoanhydride bond in ATP. The energy stored in the aminoacyl-tRNA is ultimately used in the formation of a peptide bond during protein synthesis. Note that the two ATP equivalents consumed during each aminoacylation reaction contribute to the energetic cost of protein synthesis.

## B. Specificity of Aminoacyl-tRNA Synthetases

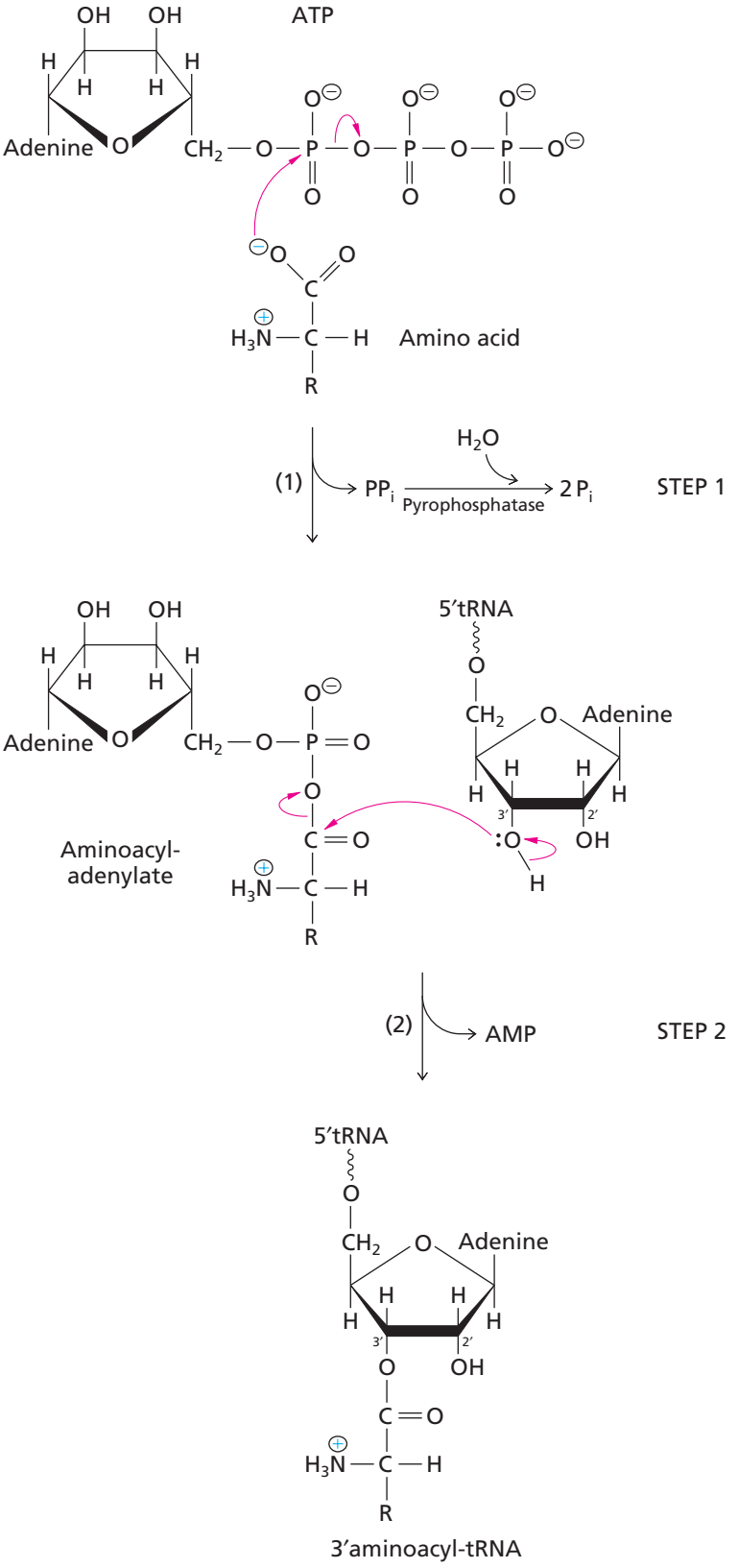
Attaching a specific amino acid to its corresponding tRNA is a crucial step in translating a genetic message. If there are errors at this step, the wrong amino acid could be incorporated into a protein.

Each aminoacyl-tRNA synthetase binds ATP and selects the proper amino acid based on its charge, size, and hydrophobicity. This initial selection eliminates most of the other amino acids. For example, tyrosyl-tRNA synthetase almost always binds tyrosine but rarely phenylalanine or any other amino acid. The synthetase then selectively binds a specific tRNA molecule. The proper tRNA is distinguished by features unique to its structure. In particular, the part of the acceptor stem that lies on the inner surface of




▲ **Figure 22.8**
**Base pairing at the wobble position.** The $tRNA^{Ala}$ molecule with the anticodon IGC can bind to any one of three codons specifying alanine (GCU, GCC, or GCA) because I can pair with U, C, or A. Note that the RNA strand containing the codon and the strand containing the anticodon are antiparallel. The wobble position is boxed in each example.

**Figure 22.9 ►**
**Synthesis of an aminoacyl-tRNA molecule catalyzed by its specific aminoacyl-tRNA synthetase.** In the first step, the nucleophilic carboxylate group of the amino acid attacks the $\alpha$-phosphorus atom of ATP, displacing pyrophosphate and producing an aminoacyl-adenylate intermediate. In the second step, nucleophilic attack by the 3′-hydroxyl group of the terminal residue of tRNA leads to displacement of AMP and formation of an aminoacyl-tRNA molecule.

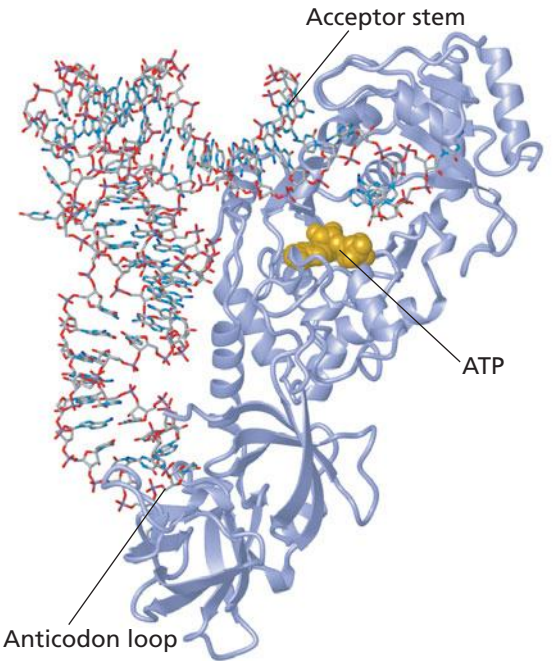



the L-shaped tRNA molecule is implicated in the binding of tRNA to the aminoacyl-tRNA synthetase (Figure 22.10).

In some cases, the synthetase enzyme recognizes not only the the acceptor stem of the tRNA but also the anticodon. For example, the glutaminyl-tRNA synthetase's ability to recognize Gln-tRNAs and to discriminate against the other 19 types of tRNAs ensures

that glutamine is specifically attached to the correct tRNA (shown in Figure 22.10). Note that glutaminyl-tRNA synthetase contacts both the acceptor stem and the anticodon region of $tRNA^{Gln}$. The crystal structure also shows a molecule of ATP bound in the active site near the 3′ end of the tRNA.

Half of the 20 different aminoacyl-tRNA synthetases resemble glutaminyl-tRNA synthetase. These enzymes bind the anticodon and aminoacylate tRNA at the 2′-hydroxyl group. A subsequent chemical rearrangement shifts the aminoacyl group to the 3′-hydroxyl group. Such enzymes are known as class I synthetases. Class II aminoacyl-tRNA synthetases are often more complex, multisubunit enzymes and they aminoacylate tRNA at the 3′-hydroxyl group. In all cases, the net effect of the interaction between tRNA and synthetase is to position the 3′ end of the tRNA molecule in the active site of the enzyme.




▲ **Figure 22.10**
**Structure of *E. coli* $tRNA^{Gln}$ bound to glutaminyl-tRNA synthetase.** The 3′ end of the tRNA is buried in a pocket on the surface of the enzyme. A molecule of ATP is also bound at this site. The enzyme interacts with both the tRNA acceptor stem and anticodon. [PDB 1QRS].

### C. Proofreading Activity of Aminoacyl-tRNA Synthetases

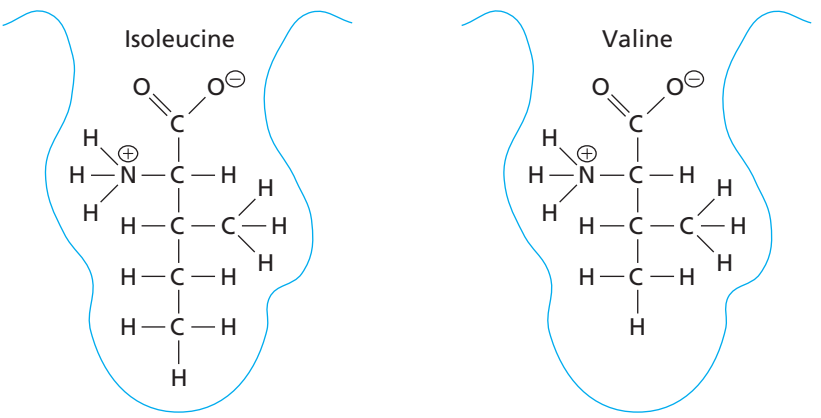
The error rate for most aminoacyl-tRNA synthetases is low because they make multiple contacts with a specific tRNA and a specific amino acid. However, isoleucine and valine are chemically similar amino acids, and both can be accommodated in the active site of isoleucyl-tRNA synthetase (Figure 22.11). Isoleucyl-tRNA synthetase mistakenly catalyzes the formation of the valyl-adenylate intermediate about 1% of the time. On the basis of this observation, we might expect valine to be attached to isoleucyl-tRNA and incorporated into protein in place of isoleucine about 1 time in 100 but the observed substitution of valine for isoleucine in polypeptide chains is only about 1 time in 10,000. This lower level of valine incorporation suggests that isoleucyl-tRNA synthetase also discriminates between the two amino acids after aminoacyl-adenylate formation. In fact, isoleucyl-tRNA synthetase carries out proofreading in the next step of the reaction. Although isoleucyl-tRNA synthetase may mistakenly catalyze the formation of valyl-adenylate, it usually catalyzes hydrolysis of the incorrect valyl-adenylate to valine and AMP or the hydrolysis of valyl-$tRNA^{Ile}$. The overall error rate of the reaction is $10^{-5}$ for most amino acyl-tRNA synthetases.

**KEY CONCEPT**

**The accuracy of information flow from nucleic acids to protein depends, in part, on the accuracy of the amino acyl-tRNA synthetase reaction.**

## 22.4 Ribosomes

Protein synthesis requires assembling four components that form an elaborate translation complex: the ribosome, which catalyzes peptide bond formation; its accessory protein factors, which help the ribosome in each step of the process; the mRNA, which carries the information specifying the protein's sequence; and the aminoacyl-tRNAs that carry the activated amino acids. Initiation involves assembly of the translation complex at the first codon in the mRNA. During polypeptide chain elongation the ribosome and associated components move, or translocate, along the template mRNA in the 5′ → 3′ direction.




◀ **Figure 22.11**
**Model of the substrate-binding site in isoleucyl-tRNA synthetase.** Despite the similar size and charge of isoleucine and valine, isoleucyl-tRNA synthetase binds to isoleucine about 100 times more readily than it binds to valine. A subsequent proofreading step also helps prevent the formation of valyl-$tRNA^{Ile}$.

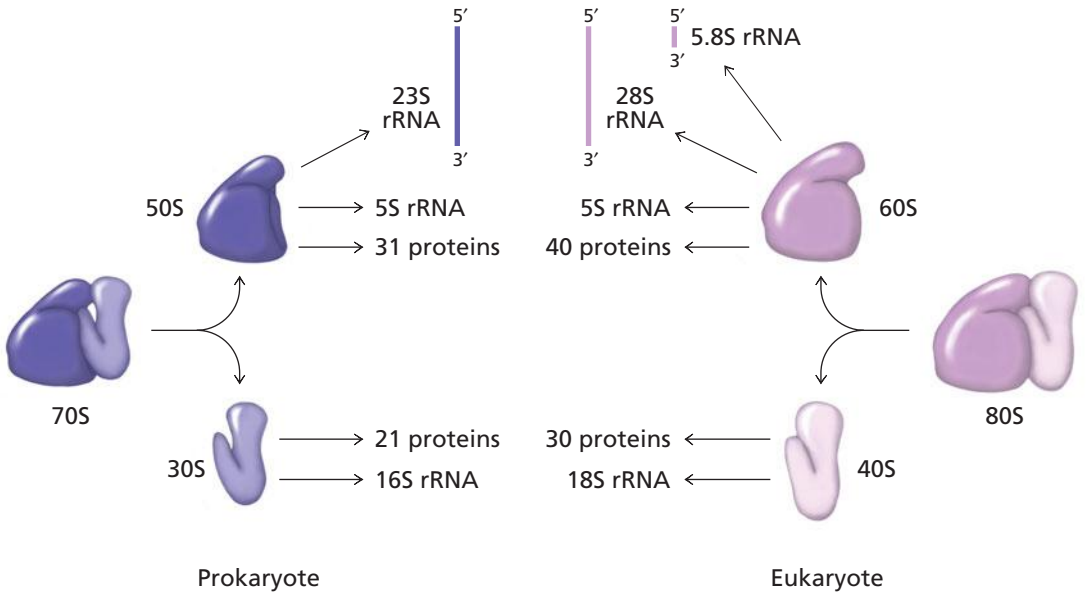
The polypeptide is synthesized from the N-terminus to its C-terminus. Finally, when synthesis of the protein is complete, the translation complex disassembles in a separate termination step. An important function of disassembly is to release the two ribosomal subunits from the mRNA so that they can participate in further rounds of translation.

## A. Ribosomes Are Composed of Both Ribosomal RNA and Protein

All ribosomes contain two subunits of unequal size. In *E. coli,* the small subunit is called the 30S subunit and the large subunit is called the 50S subunit. (The terms 30S and 50S originally referred to the sedimentation rate of these subunits.) The 30S subunit is elongated and asymmetric, with overall dimensions of 5.5 × 22 × 22.5 nm. A narrow neck separates the head from the base and a protrusion extends from the base forming a cleft where the mRNA molecule appears to rest. The 50S ribosomal subunit is wider than the 30S subunit and has several protrusions; its dimensions are about 15 × 20 × 20 nm. The 50S subunit also contains a tunnel about 10 nm long and 2.5 nm in diameter. This tunnel extends from the site of peptide bond formation and accommodates the growing polypeptide chain during protein synthesis. The 30S and 50S subunits combine to form an active 70S ribosome.

In *E. coli,* the RNA component of the 30S subunit is a 16S rRNA of 1542 nucleotides. Although its exact length varies among species, the 16S rRNA contains extensive regions of secondary structure that are highly conserved in the ribosomes of all living organisms. There are 21 ribosomal proteins in the 30S subunit. The 50S subunit of the *E. coli* ribosome contains two molecules of ribosomal RNA: one 5S rRNA of 120 nucleotides and one 23S rRNA of 2904 nucleotides. There are 31 different proteins associated with the 5S and 23S rRNA molecules in the 50S subunit (Figure 22.12).

Eukaryotic ribosomes are similar in shape to bacterial ribosomes but they tend to be somewhat larger and more complex. Intact vertebrate ribosomes are designated 80S and are made up of 40S and 60S subunits (Figure 22.12). The small 40S subunit is analogous to the 30S subunit of the prokaryotic ribosome; it contains about 30 proteins and a single molecule of 18S rRNA. The large 60S subunit contains about 40 proteins and three ribosomal RNA molecules: 5S rRNA, 28S rRNA, and 5.8S rRNA. The 5.8S rRNA is about 160 nucleotides long and its sequence is similar to that of the 5′ end of prokaryotic 23S rRNA. This similarity implies that the 5.8S rRNA and the 5′ end of prokaryotic 23S




▲ **Figure 22.12**
**Comparison of prokaryotic and eukaryotic ribosomes.** Both types of ribosomes consist of two subunits, each of which contains ribosomal RNA and proteins. The large subunit of the prokaryotic ribosome contains two molecules of rRNA: 5S and 23S. The large subunit of almost all eukaryotic ribosomes contains three molecules of rRNA: 5S, 5.8S, and 28S. The sequence of the eukaryotic 5.8S rRNA is similar to the sequence of the 5′ end of the prokaryotic 23S rRNA.

rRNA are derived from a common ancestor and that there has been a fusion or splitting of rRNA genes during their evolution.

Both prokaryotic and eukaryotic genomes contain multiple copies of ribosomal RNA genes. The combination of a large number of copies and strong promoters for these genes allows cells to maintain a high level of ribosome synthesis. Eukaryotic ribosomal RNA genes, which are transcribed by RNA polymerase I (Section 21.5A), occur as tandem arrays of hundreds of copies. In most eukaryotes, these genes are clustered in the nucleolus, where processing of ribosomal RNA precursors and ribosome assembly occur (Section 21.8B). This processing is coupled to ribosome assembly, as shown in Figure 22.13 for the *E. coli* 30S subunit. Many of the ribosomal proteins contact RNA and bind specifically to regions of secondary structure in 16S rRNA. Others form protein–protein contacts and assemble into the complex only when other ribosomal proteins are present.

The structure of the 30S ribosomal subunit from the bacterium *Thermus thermophilus* is shown in Figure 22.14 on page 676. Note that most of the mass of the 30S subunit is due to the 16S ribosomal RNA, which forms a compact structure made up of multiple regions of double-stranded RNA. The ribosomal proteins bind to the surface of the RNA or to grooves and crevices between regions of RNA secondary structure.

Similarly, the assembly of the bacterial 50S subunit and of the 40S and 60S eukaryotic subunits are also coupled to the processing of their ribosomal RNA precursors. The structure of the 50S subunit from the archeon *Haloarcula marismortui* is also shown in Figure 22.14.

### B. Ribosomes Contain Two Aminoacyl-tRNA Binding Sites

As discussed in Section 22.3, the substrates for peptide bond formation are not free amino acids but relatively large aminoacyl-tRNA molecules. A ribosome must align two adjacent aminoacyl-tRNA molecules so that their anticodons interact with the correct mRNA codons. The aminoacylated ends of these two tRNAs are positioned at the site of peptide bond formation. The ribosome must also hold the mRNA and the growing polypeptide chain, and it must accommodate the binding of several protein factors during protein synthesis. The ability to accomplish these tasks simultaneously explains, in part, why the ribosome is so large and complex.

The orientation of the two tRNA molecules during protein synthesis is shown in Figure 22.15 on page 677. The growing polypeptide chain is covalently attached to the tRNA positioned at the peptidyl site (P site), forming peptidyl-tRNA. The second aminoacyl-tRNA is bound at the aminoacyl site (A site). As the polypeptide chain is synthesized, it passes through the tunnel of the large ribosomal subunit and emerges on the outer surface of the ribosome.

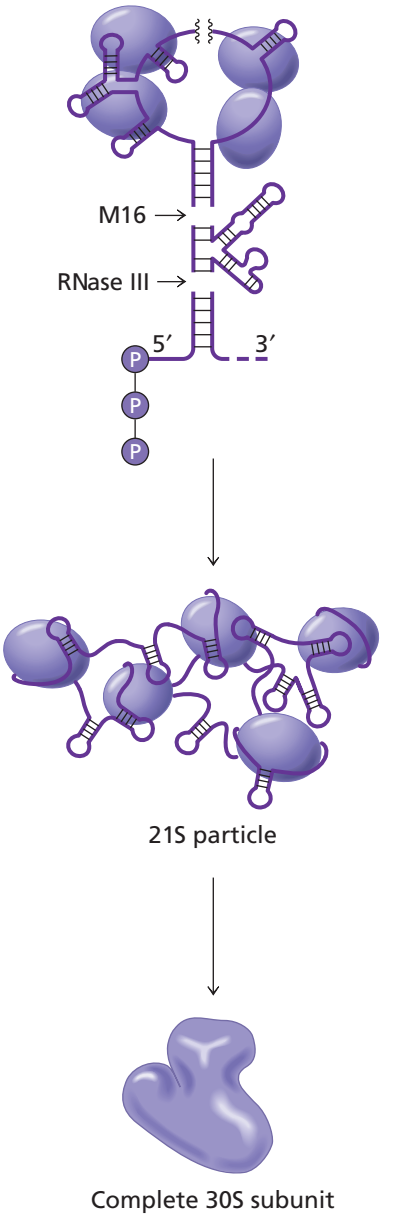
## 22.5 Initiation of Translation

The initiation of protein synthesis involves assembling a translation complex at the beginning of an mRNA's coding sequence. This complex consists of the two ribosomal subunits, an mRNA template to be translated, an initiator tRNA molecule, and several accessory proteins called *initiation factors*. This crucial initiation step ensures that the proper initiation codon (and therefore the correct reading frame) is selected before translation begins.

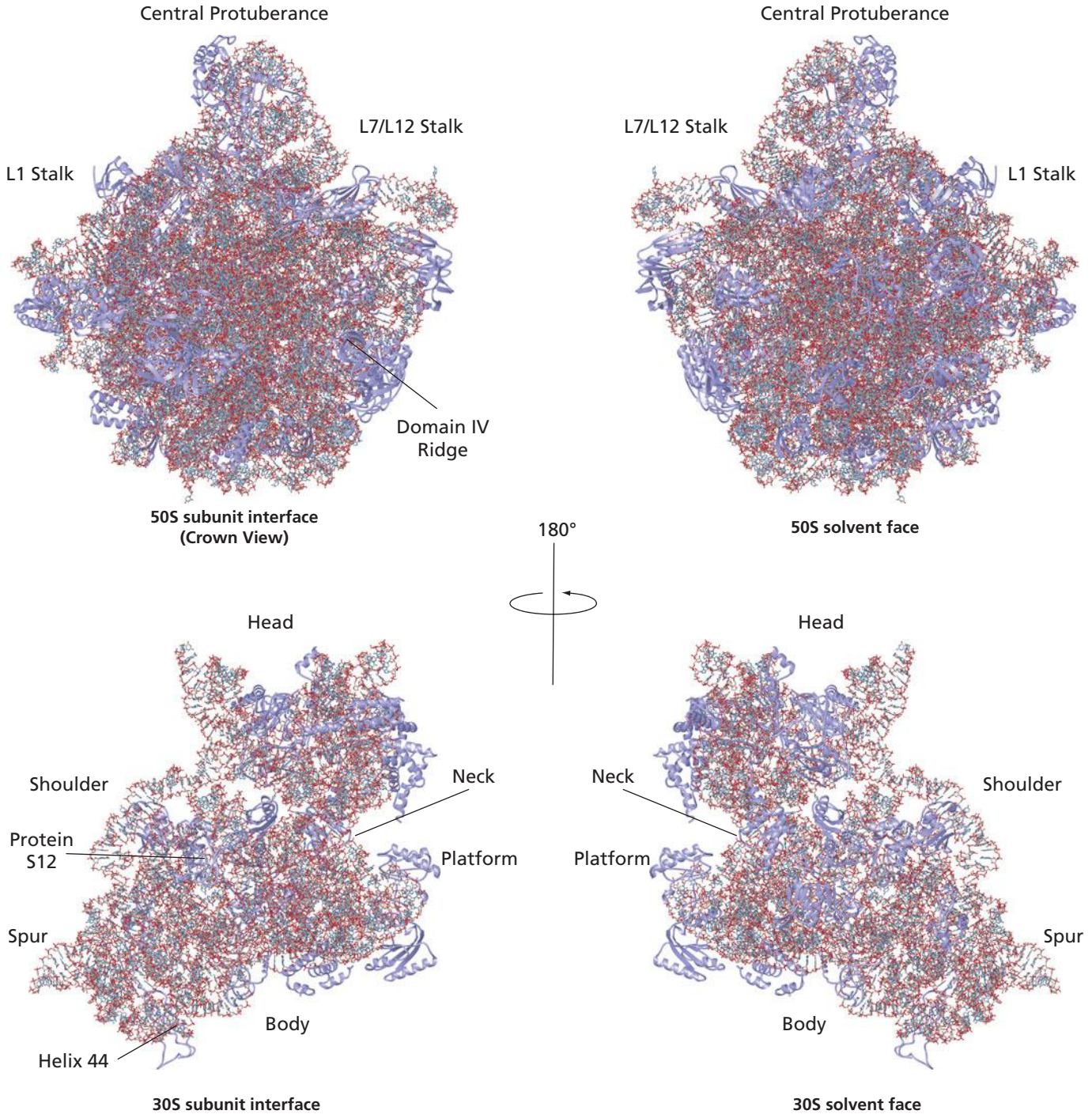
### A. Initiator tRNA

As mentioned in Section 22.1, the first codon translated is usually AUG. Every cell contains at least two types of methionyl-tRNA$^{Met}$ molecules that can recognize AUG codons. One type is used exclusively at initiation codons and is called the initiator tRNA. The other type only recognizes internal methionine codons. Although these two tRNA$^{Met}$ molecules have different primary sequences, and distinct functions, both of them are aminoacylated by the same methionyl-tRNA synthetase.

In bacteria, the initiator tRNA is called tRNA$_f^{Met}$. The charged initiator tRNA (methionyl-tRNA$_f^{Met}$) is the substrate for a formyltransferase that catalyzes addition of a formyl group from 10-formyltetrahydrofolate to the methionine residue producing




▲ **Figure 22.13**
**Assembly of the 30S ribosomal subunit and maturation of 16S rRNA in *E. coli*.** Assembly of the 30S ribosomal subunit begins when six or seven ribosomal proteins bind to the 16S rRNA precursor as it is being transcribed, thereby forming a 21S particle. The 21S particle undergoes a conformational change, and the 16S rRNA molecule is processed to its final length. During this processing, the remaining ribosomal proteins of the 30S subunit bind (recall that M16 is a site-specific endonuclease involved in RNA processing that we discussed in Chapter 21).




▲ **Figure 22.14**
**Three-dimensional structures of the *H. marismortui* 50S subunit (top) and the *T. thermophilus* 30S subunit (bottom).**

*N*-formylmethionyl-$tRNA_f^{Met}$ ($fMet\text{-}tRNA_f^{Met}$) as shown in Figure 22.16 on page 681. In eukaryotes and archaebacteria, the initiator tRNA is called $tRNA_i^{Met}$. The methionine that begins protein synthesis in eukaryotes is not formylated.

*N*-Formylmethionine in bacteria—or methionine in other organisms—is the first amino acid incorporated into proteins. After protein synthesis is under way, the N-terminal methionine can be either deformylated or removed from the polypeptide chain altogether.

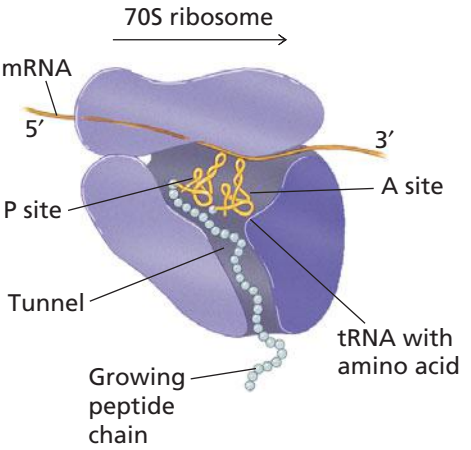
## B. Initiation Complexes Assemble Only at Initiation Codons

There are three possible reading frames in an mRNA molecule but only one of them is correct. Establishing the correct reading frame during the initiation of translation is

critical for the accurate decoding of information from mRNA into protein. Shifting the reading frame by even a single nucleotide would alter the sequence of the entire polypeptide and result in a nonfunctional protein. The translation machinery must therefore accurately locate the initiation codon that serves as the start site for protein synthesis.

The ribosome needs to distinguish between the single correct initiation codon and all the other incorrect AUGs. These other AUGs specify either internal methionine residues in the correct reading frame or irrelevant methionine codons in the two other incorrect reading frames. It is important to appreciate that the initiation codon is not simply the first three nucleotides of the mRNA. Initiation codons can be located many nucleotides downstream of the 5′-end of the mRNA molecule.

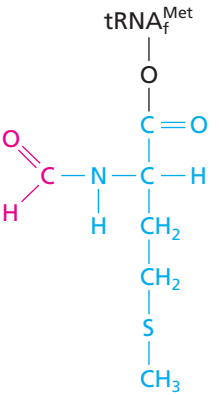
In prokaryotes, the selection of an initiation site depends on an interaction between the small subunit of the ribosome and the mRNA template. The 30S subunit binds to the mRNA template at a purine-rich region just upstream of the correct initiation codon. This region, called the Shine-Dalgarno sequence, is complementary to a pyrimidine-rich stretch at the 3′ end of the 16S rRNA molecule. During formation of the initiation complex, these complementary nucleotides pair to form a double-stranded RNA structure that binds the mRNA to the ribosome. The result of this interaction is to position the initiation codon at the P site on the ribosome (Figure 22.17). The initiation complex assembles exclusively at initiation codons because Shine-Dalgarno sequences are not found immediately upstream of internal methionine codons.



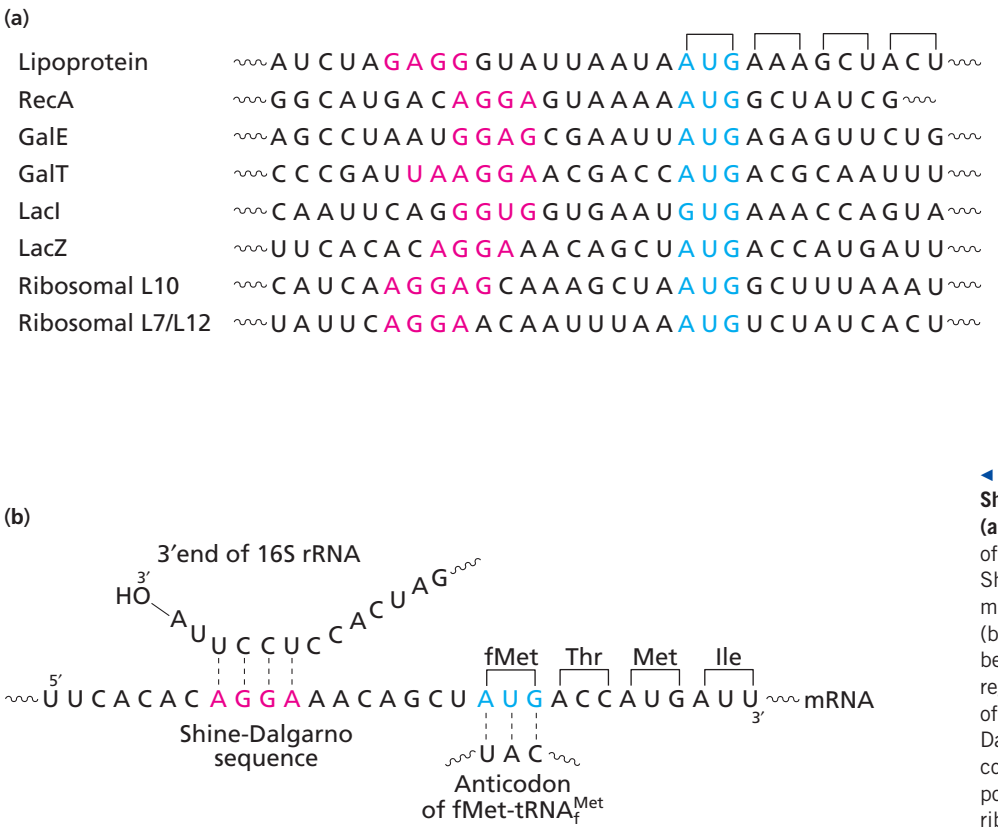

▲ **Figure 22.15**
**Sites for tRNA binding in prokaryotic ribosomes.** During protein synthesis, the P site is occupied by the tRNA molecule attached to the growing polypeptide chain, and the A site holds an aminoacyl-tRNA. The growing polypeptide chain passes through the tunnel of the large subunit.

## C. Initiation Factors Help Form the Initiation Complex

Formation of the initiation complex requires several **initiation factors** in addition to ribosomes, initiator tRNA, and mRNA. Prokaryotes contain three initiation factors, designated IF-1, IF-2, and IF-3. There are at least eight eukaryotic initiation factors (eIF's). In both prokaryotes and eukaryotes, the initiation factors catalyze assembly of the protein synthesis complex at the initiation codon.




▲ **Figure 22.16**
**Chemical structure of fMet-tRNA$_f^{Met}$.** A formyl group (red) is added to the methionyl moiety (blue) of methionyl-tRNA$_f^{Met}$ in a reaction catalyzed by a formyltransferase.

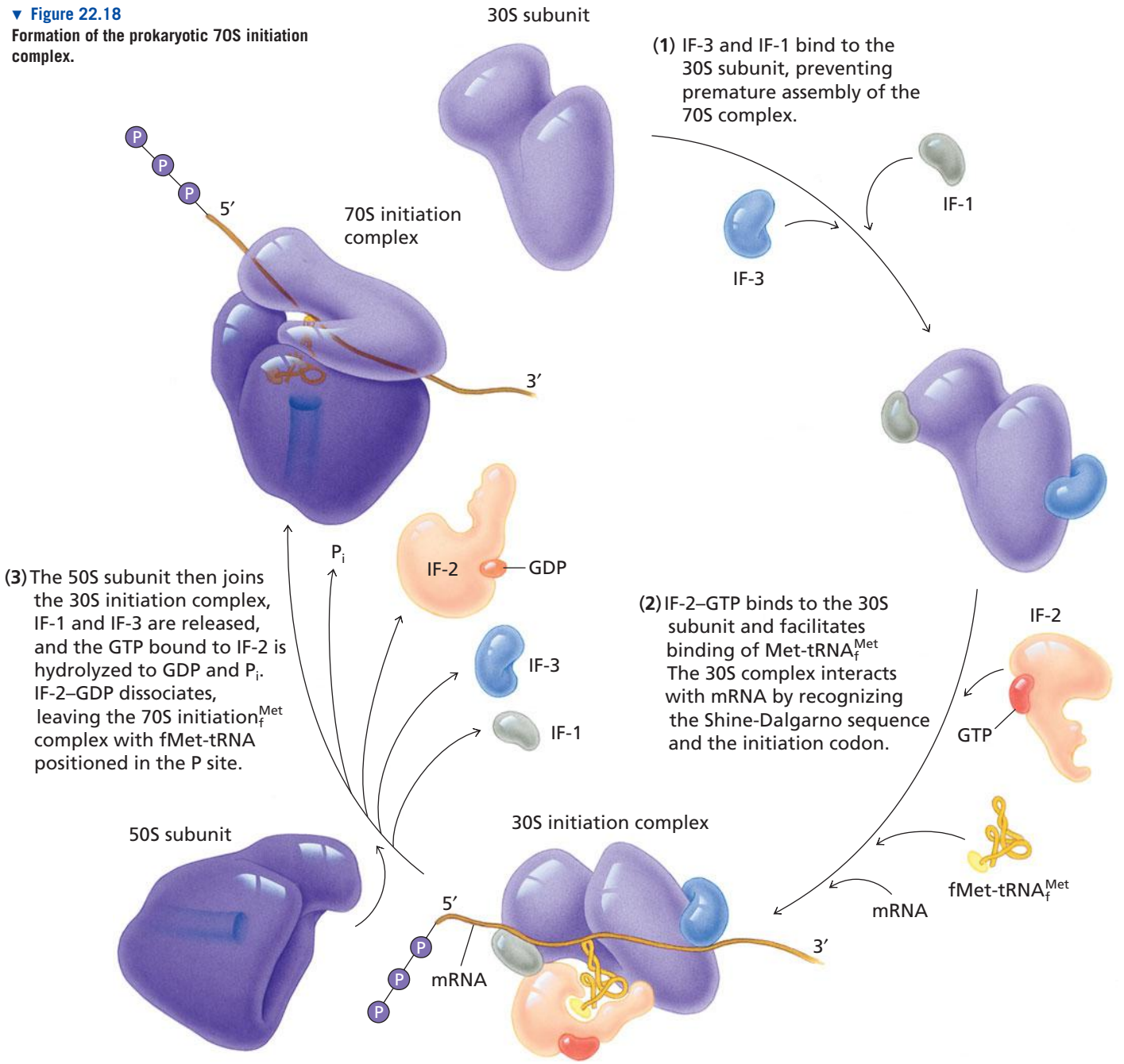



◀ **Figure 22.17**
**Shine-Dalgarno sequences in *E. coli* mRNA.** **(a)** Ribosome-binding sites at the 5′ end of mRNA for several *E. coli* proteins. The Shine-Dalgarno sequences (red) occur immediately upstream of initiation codons (blue). **(b)** Complementary base pairing between the 3′ end of 16S rRNA and the region near the 5′ end of an mRNA. Binding of the 3′ end of the 16S rRNA to the Shine-Dalgarno sequence helps establish the correct reading frame for translation by positioning the initiation codon at the ribosome's P site.

One of the roles of IF-3 is to maintain the ribosomal subunits in their dissociated state by binding to the small subunit. The ribosomal subunits bind separately to the initiation complex and the association of IF-3 with the 30S subunit prevents the 30S and 50S subunits from forming the 70S complex prematurely. IF-3 also helps position fMet-tRNA$_f^{Met}$ and the initiation codon at the P site of the ribosome. IF-2 selects the initiator tRNA from the pool of aminoacylated tRNA molecules in the cell. It binds GTP forming an IF-2–GTP complex that specifically recognizes the initiator tRNA and rejects all other aminoacyl-tRNA molecules. The third initiation factor, IF-1, binds to the 30S subunit and facilitates the actions of IF-2 and IF-3.

Once the 30S complex has been formed at the initiation codon, the 50S ribosomal subunit binds to the 30S subunit. Next, the GTP bound to IF-2 is hydrolyzed and $P_i$ is released. The initiation factors dissociate from the complex when GTP is hydrolyzed. IF-2–GTP is regenerated when the bound GDP is exchanged for GTP. The steps in the formation of the 70S initiation complex are summarized in Figure 22.18.

▼ **Figure 22.18**
**Formation of the prokaryotic 70S initiation complex.**



The role of the prokaryotic initiation factors is to ensure that the aminoacylated initiator tRNA (fMet-tRNA$_i^{Met}$) is correctly positioned at the initiation codon. The initiation factors also mediate the formation of a complete initiation complex by reconstituting a 70S ribosome such that the initiation codon is positioned in the P site.

### D. Translation Initiation in Eukaryotes

Eukaryotic mRNAs do not have distinct Shine-Dalgarno sequences that serve as ribosome binding sites. Instead, the first AUG codon in the message usually serves as the initiation codon. eIF-4 (eukaryotic initiation factor 4), also known as cap binding protein (CBP), binds specifically to the 7-methylguanylate cap (Figure 21.26) at the 5′ end of eukaryotic mRNA. Binding of eIF-4 to the cap structure leads to the formation of a preinitiation complex consisting of the 40S ribosomal subunit, an aminoacylated initiator tRNA, and several other initiation factors. The preinitiation complex then scans along the mRNA in the 5′ → 3′ direction until it encounters an initiation codon. When the search is successful, the small ribosomal subunit is positioned so that Met-tRNA$_i^{Met}$ interacts with the initiation codon in the P site. In the final step, the 60S ribosomal subunit binds to complete the 80S initiation complex and all the initiation factors dissociate. The dissociation of eIF-2—the eukaryotic counterpart of bacterial IF-2—is accompanied by GTP hydrolysis.

Most eukaryotic mRNA molecules encode only a single polypeptide since the normal mechanism of selecting the initiation codon by scanning along the mRNA from the 5′ end permits only one initiation codon per mRNA. In contrast, prokaryotic mRNAs often contain several coding regions. Each coding region begins with an initiation codon that is associated with its own upstream Shine-Dalgarno sequence. mRNA molecules that encode several polypeptides are said to be **polycistronic**.

## 22.6 Chain Elongation During Protein Synthesis Is a Three-Step Microcycle

At the end of the initiation step, the mRNA is positioned so that the next codon can be translated during the elongation stage of protein synthesis. The initiator tRNA occupies the P site in the ribosome and the A site is ready to receive an incoming aminoacyl-tRNA. During chain elongation each additional amino acid is added to the nascent polypeptide chain in a three-step microcycle. The steps in this microcycle are (1) positioning the correct aminoacyl-tRNA in the A site of the ribosome, (2) forming the peptide bond, and (3) shifting, or translocating, the mRNA by one codon relative to the ribosome (the two tRNAs in the ribosome's P and A sites also translocate).

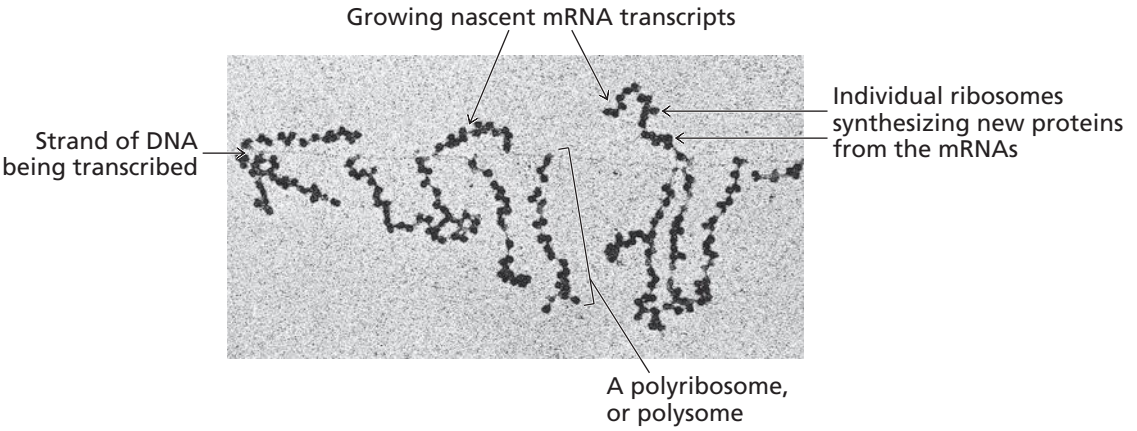
The translation machinery works relatively slowly compared to the enzyme systems that catalyze DNA replication. Proteins are synthesized at a rate of only 18 amino acid residues per second, whereas bacterial replisomes synthesize DNA at a rate of 1000 nucleotides per second. This difference in rates reflects, in part, the difference between polymerizing four types of nucleotides to make nucleic acids and polymerizing 20 types of amino acids to make proteins. Testing and rejecting all of the incorrect aminoacyl-tRNA molecules also takes time and slows protein synthesis.

The rate of transcription in prokaryotes is approximately 55 nucleotides per second. This corresponds to about 18 codons per second or the same rate at which the mRNA is translated. In bacteria, translation initiation occurs as soon as the 5′ end of an mRNA is synthesized and translation and transcription are coupled (Figure 22.19 on page 680). This tight coupling is not possible in eukaryotes because transcription and translation are carried out in separate compartments of the cell (the nucleus and the cytoplasm, respectively). Eukaryotic mRNA precursors must be processed in the nucleus (e.g., capped, polyadenylated, spliced) before they are exported to the cytoplasm for translation.

An *E. coli* cell contains about 20,000 ribosomes. Many large eukaryotic cells have several hundred thousand ribosomes. Large mRNA molecules can be translated simultaneously by many protein synthesis complexes forming a polyribosome or **polysome**, as

**KEY CONCEPT**

**The A site of an actively translating ribosome spends the vast majority of its time bound to one of the 19 types of incorrect aminoacyl-tRNAs as it randomly samples the pool of charged tRNAs, seeking the correct tRNA.**




▲ **Figure 22.19**
**Coupled transcription and translation of an *E. coli* gene.** The gene is being transcribed from left to right. Ribosomes bind to the 5′ end of the mRNA molecules as soon as they are synthesized. The large polysomes on the right are released from the gene when transcription terminates.

seen in Figure 22.19. The number of ribosomes bound to an mRNA molecule depends on the length of the mRNA and the efficiency of initiation of protein synthesis. At maximal efficiency the spacing between each translation complex in the polysome is about 100 nucleotides. On average, each mRNA molecule in an *E. coli* cell is translated 30 times, effectively amplifying the information it encodes by 30-fold.

## A. Elongation Factors Dock an Aminoacyl-tRNA in the A Site

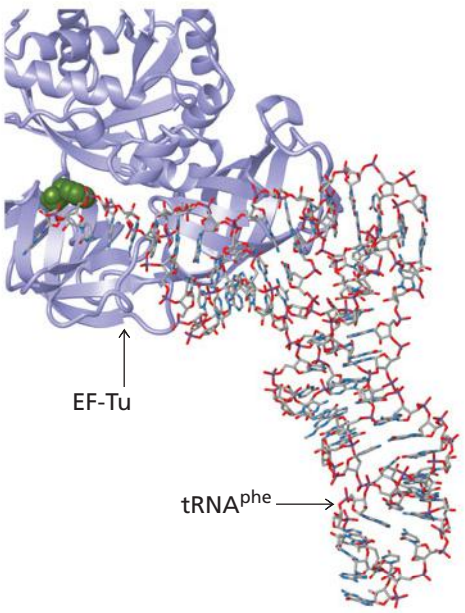
At the start of the first chain elongation microcycle, the A site is empty and the **P site** is occupied by the aminoacylated initiator tRNA. The first step in chain elongation is insertion of the correct aminoacyl-tRNA into the A site of the ribosome. In bacteria, this step is catalyzed by an elongation factor called EF-Tu. EF-Tu is a monomeric protein that contains a binding site for GTP. Each *E. coli* cell has about 135,000 molecules of EF-Tu, making it one of the most abundant proteins in the cell (emphasizing the importance of protein synthesis to a cell).

EF-Tu–GTP associates with an aminoacyl-tRNA molecule to form a ternary complex that fits into the A site of a ribosome. Almost all aminoacyl-tRNA molecules *in vivo* are found in such ternary complexes (Figure 22.20). The structure of EF-Tu is similar to that of IF-2 (which also binds GTP) and other G proteins (Section 9.12A), suggesting that they all evolved from a common ancestral protein.

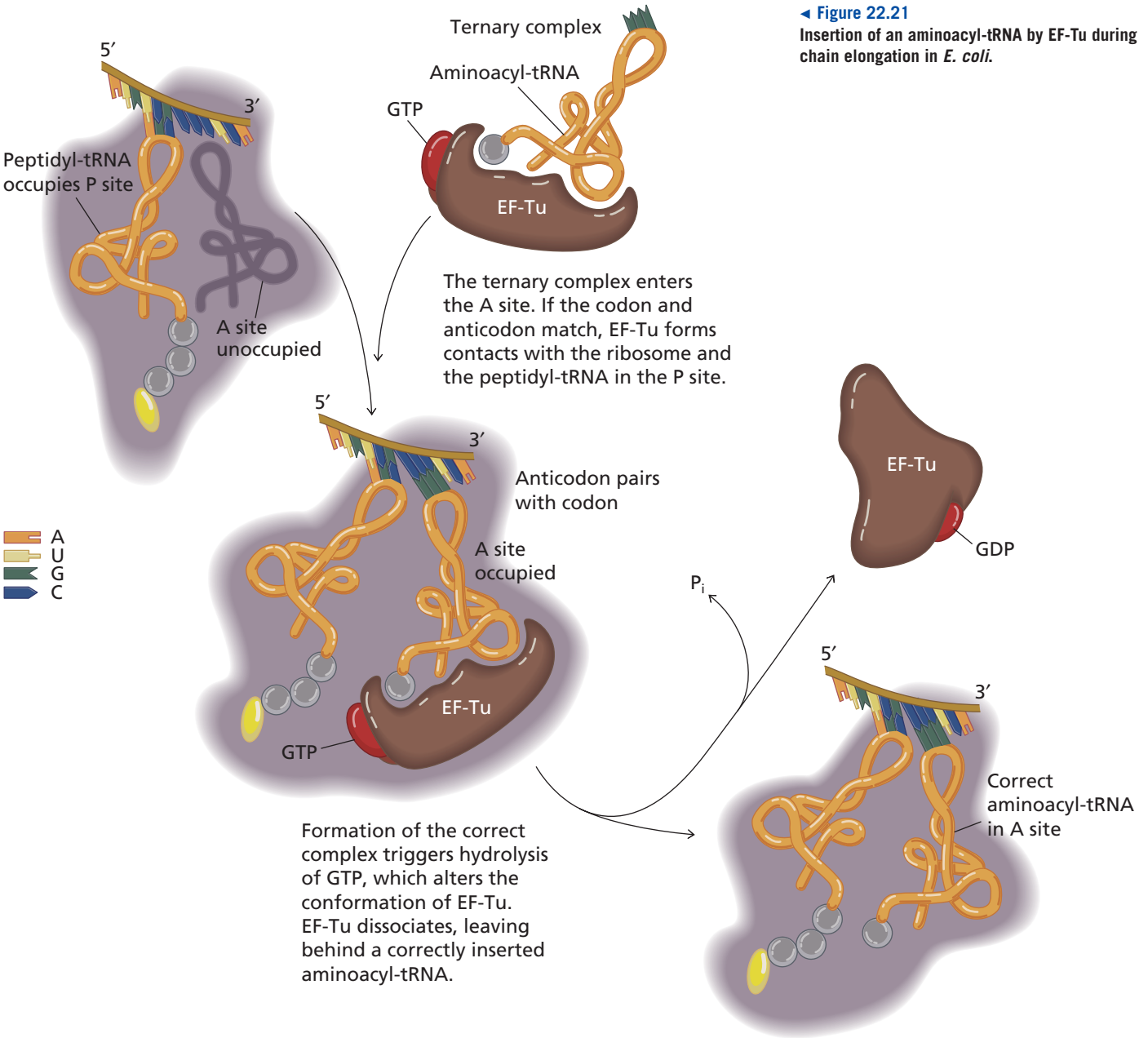
The EF-Tu–GTP complex recognizes common features of the tertiary structure of tRNA molecules and binds tightly to all aminoacyl-tRNA molecules except fMet-tRNA$_f^{Met}$. The fMet-tRNA$_f^{Met}$ molecule is distinguished from all other aminoacyl-tRNA molecules by the distinctive secondary structure of its acceptor stem.

A ternary complex of EF-Tu–GTP–aminoacyl-tRNA can diffuse freely into the A site in the ribosome. When correct base pairs form between the anticodon of the aminoacyl-tRNA and the mRNA codon in the A site, the complex is stabilized. EF-Tu–GTP can then contact sites in the ribosome as well as the tRNA in the P site (Figure 22.21, on page 681). These contacts trigger hydrolysis of GTP to GDP and $P_i$ causing a conformational change in EF-Tu–GDP that releases the bound aminoacyl-tRNA. EF-Tu–GDP then dissociates from the chain elongation complex. The aminoacyl-tRNA remains in the A site where it is positioned for peptide bond formation.

EF-Tu–GDP cannot bind another aminoacyl-tRNA molecule until GDP dissociates. An additional elongation factor called EF-Ts catalyzes the exchange of bound GDP for GTP (Figure 22.22, on page 682). Note that one GTP molecule is hydrolyzed for every aminoacyl-tRNA that is successfully inserted into the A site.




▲ **Figure 22.20**
**EF-Tu binds aminoacylated tRNAs.** The EF-Tu–GTP complex binds to the acceptor end of aminoacylated tRNA (in this case phenylalanyl-tRNA$^{Phe}$). The phenylalanine residue is shown in green. This is how charged tRNAs commonly exist inside a cell.




◀ **Figure 22.21**
**Insertion of an aminoacyl-tRNA by EF-Tu during chain elongation in *E. coli*.**

## B. Peptidyl Transferase Catalyzes Peptide Bond Formation

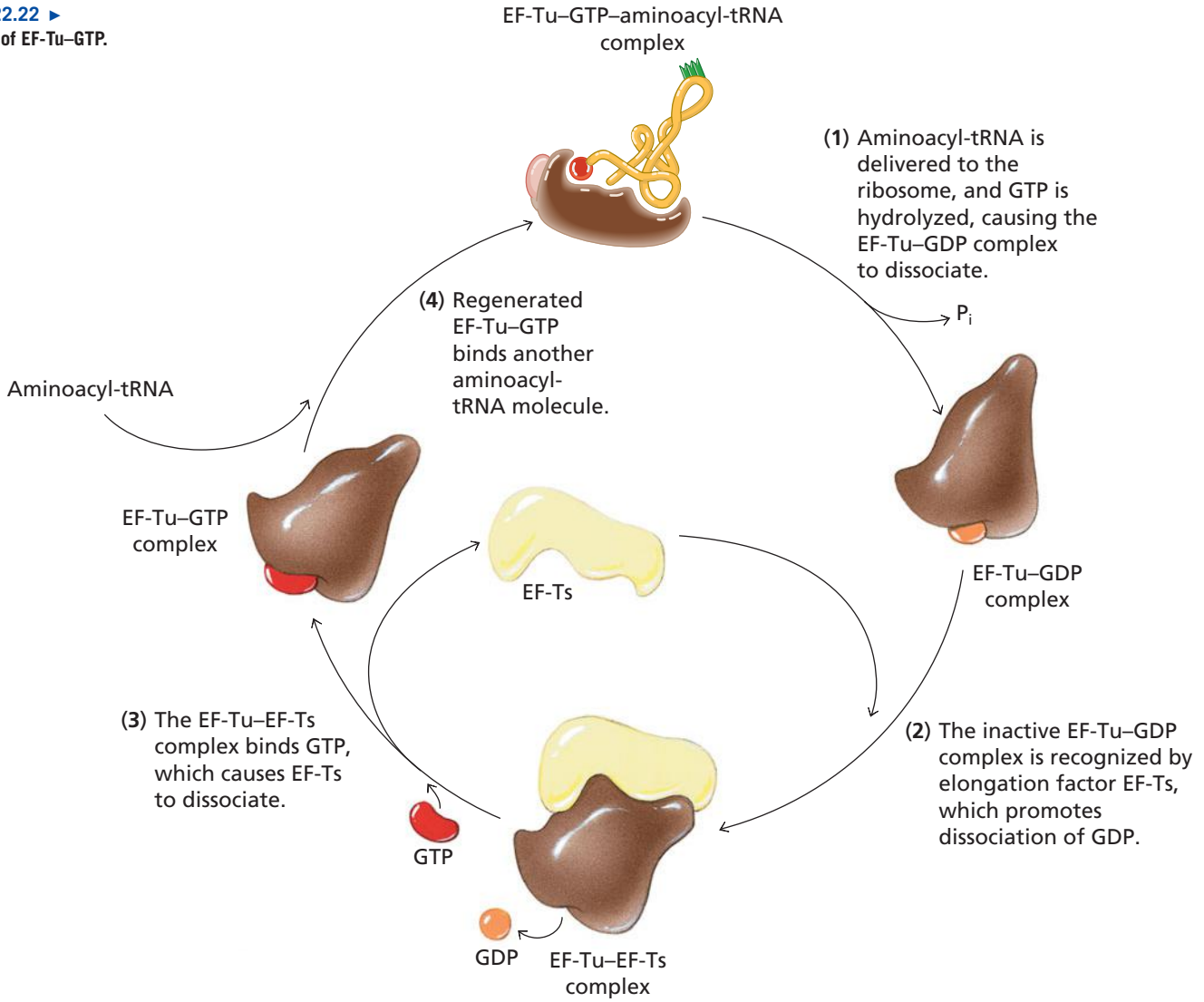
Binding of a correct aminoacyl-tRNA in the A site aligns the activated amino acid's $\alpha$-amino group next to the ester bond's carbonyl on the peptidyl-tRNA in the neighboring P site. The nitrogen atom's lone pair of electrons execute a nucleophilic attack on the carbonyl carbon, resulting in the formation of a peptide bond via a displacement reaction. While it is straightforward to visualize how the ribosome's active site aligns these substrates, we do not understand precisely how the ribosome enhances the rate of this reaction. The peptide chain, now one amino acid longer, is transferred from the tRNA in the P site to the tRNA in the A site (Figure 22.23, on page 683). Formation of the peptide bond requires hydrolysis of the energy-rich peptidyl-tRNA linkage. Note that the growing polypeptide chain is covalently attached to the tRNA in the A site, forming a peptidyl-tRNA.

The enzymatic activity responsible for formation of the peptide bond is referred to as **peptidyl transferase**. This activity is contained within the large ribosomal subunit. Both the 23S rRNA molecule and the 50S ribosomal proteins contribute to the substrate binding sites, but the catalytic activity is localized to the RNA component. Thus, peptidyl transferase is yet another example of an RNA-catalyzed reaction.

### KEY CONCEPT

**Formation of the new peptide bond involves physically transferring the polypeptide attached to the P site tRNA onto the amino-terminus of the aminoacyl-tRNA bound in the ribosome's A site.**

**Figure 22.22 ►**
**Cycling of EF-Tu–GTP.**

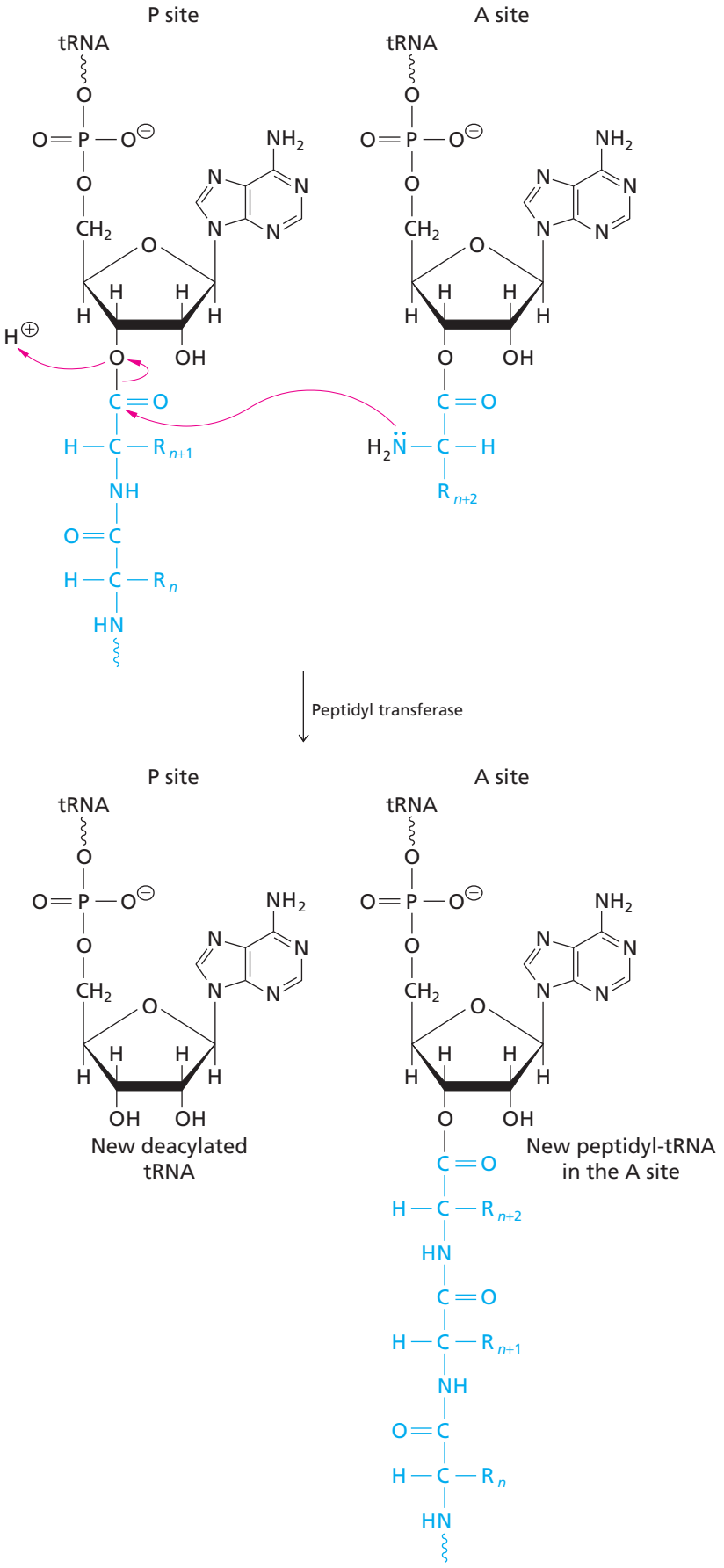



## C. Translocation Moves the Ribosome by One Codon

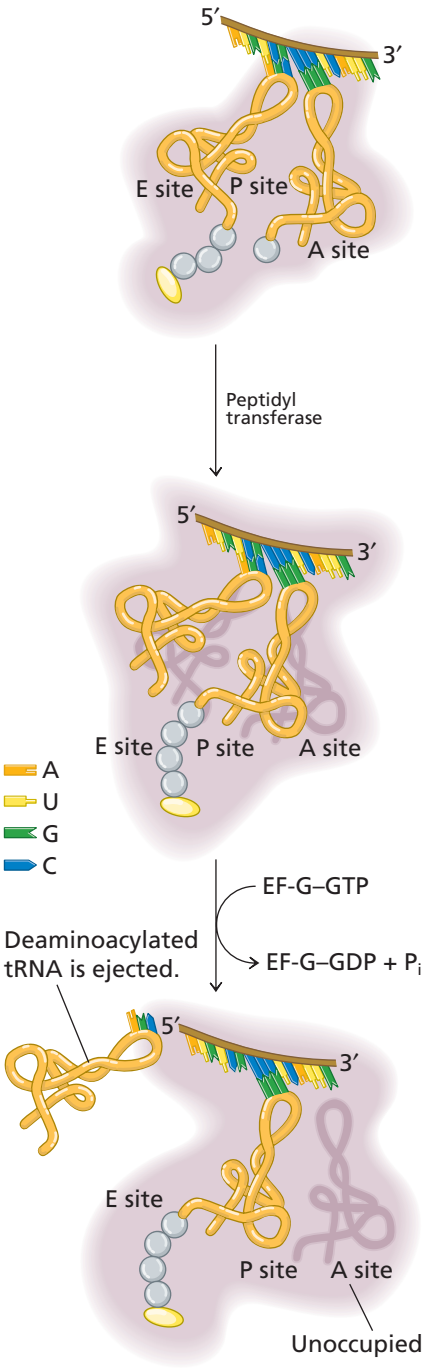
After the peptide bond has formed, the newly created peptidyl-tRNA is partially in the A site and partially in the P site (Figure 22.24, on page 684). The deaminoacylated tRNA has been displaced somewhat from the P site. It now occupies a position on the ribosome that is referred to as the exit site, or **E site**. Before the next codon can be translated, the deaminoacylated tRNA must be released and the peptidyl-tRNA must be completely transferred from the A site to the P site. At the same time, the mRNA must shift by one codon relative to the ribosome. This **translocation** is the third step in the chain elongation microcycle.

In prokaryotes, the translocation step requires a third elongation factor, EF-G. Like the other elongation factors, EF-G is an abundant protein; an *E. coli* cell contains approximately 20,000 molecules of EF-G, or roughly one for every ribosome. Like EF-Tu, EF-G has a binding site for GTP. Binding of EF-G–GTP to the ribosome completes the translocation of the peptidyl-tRNA from the A site to the P site and releases the deaminoacylated tRNA from the E site. EF-G itself is released from the ribosome only when its bound GTP is hydrolyzed to GDP and $P_i$ is released. The dissociation of EF-G–GDP leaves the ribosome free to begin another microcycle of chain elongation.

The growing polypeptide chain extends from the peptidyl-tRNA in the P site through a tunnel in the 50S subunit, to exit on the exterior surface of the ribosome




◀ **Figure 22.23**
**Formation of a peptide bond.** The carbonyl carbon of the peptidyl-tRNA undergoes nucleophilic attack by the nitrogen atom of the amino group. This aminoacyl-group–transfer reaction results in growth of the peptide chain by one residue and transfer of the nascent peptide to the tRNA in the A site.




▲ **Figure 22.24**
**Translocation during protein synthesis in prokaryotes.**

top: Aminoacyl-tRNA is positioned in the A site.

middle: Following synthesis of the peptide bond, the newly formed peptidyl-tRNA is partly in the A site and partly in the P site.

bottom: Translocation shifts the peptidyl-tRNA completely into the P site, leaving the A site empty and ejecting the deaminoacylated tRNA from the E site.

(Figure 22.15). Each translocation step helps push the chain through the tunnel. The newly synthesized polypeptide doesn't begin to fold into its final shape until it emerges from the tunnel. This folding is assisted by chaperones, such as HSP70, that are associated with the translation machinery (Section 4.10D).

The elongation microcycle is repeated for each new codon in the mRNA being translated, resulting in the synthesis of a polypeptide chain that may be several hundred residues long. Eventually, the translation complex reaches the final codon at the end of the coding region, where translation is terminated.

The elongation reactions in eukaryotes are very similar to those in *E. coli*. Three accessory protein factors participate in chain elongation in eukaryotes: EF-1$\alpha$, EF-1$\beta$, and EF-2. EF-1$\alpha$ docks the aminoacyl-tRNA in the A site; its activity thus parallels that of *E. coli* EF-Tu. EF-1$\beta$ acts like bacterial EF-Ts, recycling EF-1$\alpha$. EF-2 carries out translocation in eukaryotes. EF-Tu and EF-1$\alpha$ are highly conserved, homologous proteins, as are EF-G and EF-2. Eukaryotic and prokaryotic ribosomal RNAs are also very similar in sequence and in secondary structure. These similarities indicate that the common ancestor of prokaryotes and eukaryotes carried out protein synthesis in a manner similar to that seen in modern organisms. Thus, protein synthesis is one of the most ancient and fundamental biochemical reactions.

## 22.7 Termination of Translation

*E. coli* has three release factors (RF-1, RF-2, and RF-3) that participate in the termination of protein synthesis. After formation of the final peptide bond, the peptidyl-tRNA is translocated from the A site to the P site, as usual. The translocation positions one of the three termination codons (UGA, UAG, or UAA) in the A site. These termination codons are not recognized by any tRNA molecules so protein synthesis stalls at the termination codon. Eventually, one of the release factors diffuses into the A site. RF-1 recognizes UAA and UAG and RF-2 recognizes UAA and UGA. RF-3 binds GTP and enhances the effects of RF-1 and RF-2.

When the release factors recognize a termination codon, they cause hydrolysis of the peptidyl-tRNA. Release of the completed polypeptide is probably accompanied by GTP hydrolysis and dissociation of the release factors from the ribosome. At this point, the ribosomal subunits dissociate from the mRNA and initiation factors bind to the 30S subunit in preparation for the next round of protein synthesis.

## 22.8 Protein Synthesis Is Energetically Expensive

Protein synthesis is very expensive—it uses a large fraction of all ATP equivalents that are available in a cell. Where does all this energy go?

For each amino acid added to a polypeptide chain, four phosphoanhydride bonds are cleaved: ATP is hydrolyzed to AMP + 2 $P_i$ during activation of the amino acid and two GTP molecules are hydrolyzed to 2 GDP + 2 $P_i$ during chain elongation. The hydrolysis of GTP is coupled to conformational changes in the translation machinery. In this sense, GTP and GDP act as allosteric modulators. However, unlike most conformational changes induced by allosteric modulators, the conformational changes that occur during protein synthesis are associated with a considerable consumption of energy.

The hydrolysis of four phosphoanhydride bonds represents a large Gibbs free energy change—much more than is required for the formation of a single peptide bond. Most of the "extra" energy compensates for the loss of entropy during protein synthesis. The decrease in entropy is due primarily to the specific ordering of 20 different kinds of amino acids into a polypeptide chain. In addition, entropy is lost when an amino acid is linked to a particular tRNA and when an aminoacyl-tRNA associates with a specific codon.

## 22.9 Regulation of Protein Synthesis

Ribosomes moving
on messenger RNA
synthesize proteins
*haiku* by Sydney Brenner (2002)

One way gene expression can be regulated is by controlling the translation of mRNA into protein. Translation can be controlled at initiation, elongation, or termination. In general, translational control of gene expression is used to regulate the production of proteins that assemble into multisubunit complexes and proteins whose expression in the cell must be strictly and quickly controlled.

The rate of translation depends to some extent on the sequence of the template. An mRNA containing an abundance of rare codons, for example, is translated less rapidly (and therefore less frequently) than one containing the most frequently used codons. In addition, the rate of translation initiation varies with the nucleotide sequence at the initiation site. A strong ribosome binding site in bacterial mRNA leads to more efficient initiation. There is also evidence that the nucleotide sequence surrounding the initiation codon in eukaryotic mRNA influences the rate of initiation.

One difference between the initiation of translation and the initiation of transcription is that the formation of a translation complex can be influenced by secondary structure in the message. For example, the formation of intramolecular double-stranded regions in mRNA can mask ribosome binding sites and the initiation codon. Although structural properties can determine whether a given mRNA molecule is translated frequently or infrequently, this is not regulation in the strict sense. We use the term *translational regulation* to refer to cases where extrinsic factors modulate the frequency of mRNA translation.

Polypeptide synthesis is an example of head growth (Box 12.5).

**KEY CONCEPT**

**mRNA codons in the ribosome's A site are also being continually tested by randomly diffusing release factors, which are seeking translation termination codons.**

### A. Ribosomal Protein Synthesis Is Coupled to Ribosome Assembly in *E. coli*

Every *E. coli* ribosome contains at least 52 ribosomal proteins. The genes encoding these ribosomal proteins are scattered throughout the genome in 13 operons and seven isolated genes. When multiple copies of genes encoding some of these ribosomal proteins are inserted into *E. coli*, the concentrations of the respective mRNAs increase sharply, yet the overall rate of ribosomal protein synthesis scarcely changes. Furthermore, the relative concentrations of ribosomal proteins remain unchanged even though the various mRNA molecules for ribosomal proteins are present in unequal amounts. These findings suggest that the synthesis of ribosomal proteins is tightly regulated at the level of translation.

Translational regulation of ribosomal protein synthesis is crucial since ribosomes cannot assemble unless all the proteins are present in the proper stoichiometry. The production of ribosomal proteins is controlled by regulating the efficiency with which their mRNAs are translated. Each of the large operons containing ribosomal protein genes encodes one ribosomal protein that inhibits translation of its own polycistronic mRNA by binding near the initiation codon of one of the first genes of the operon.

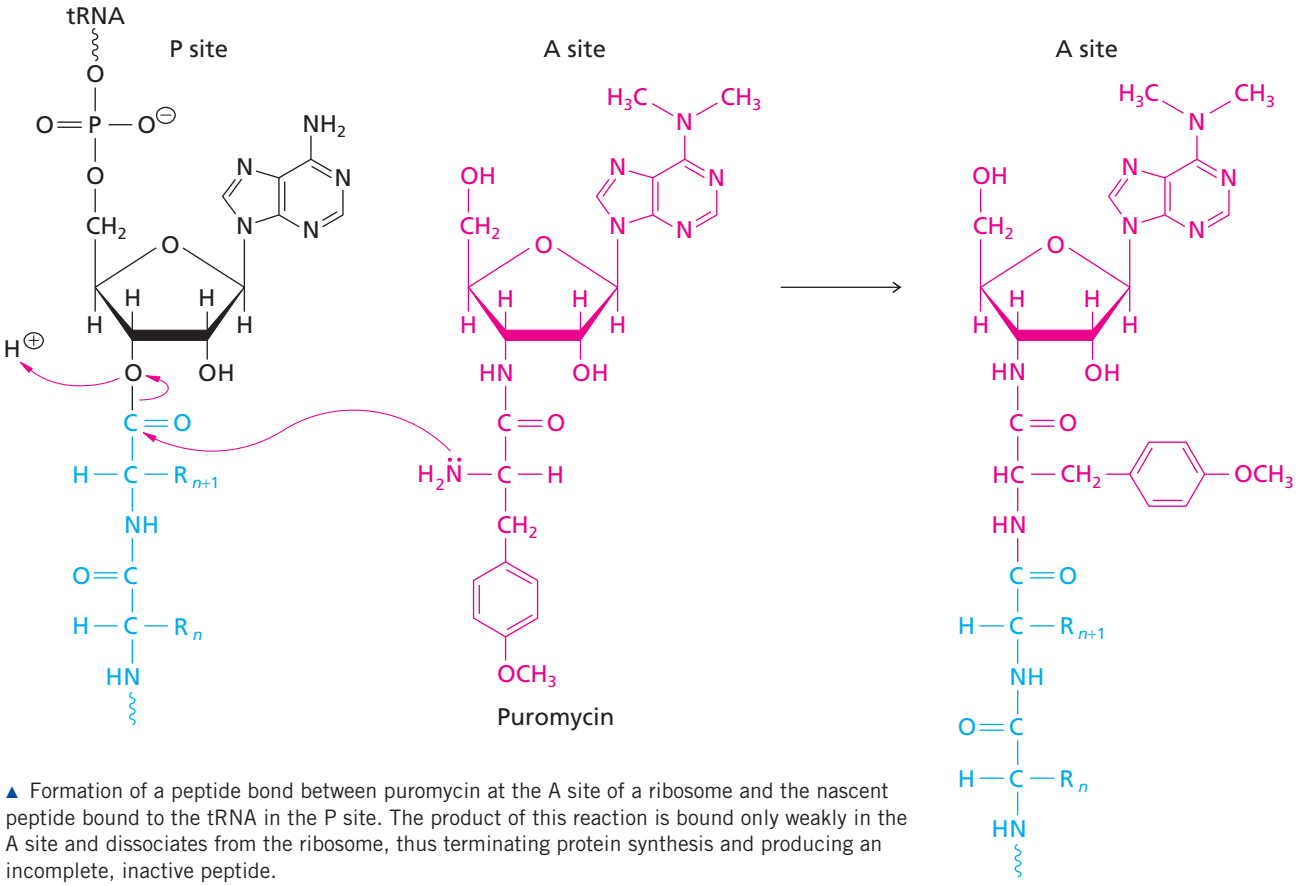
The interactions between the inhibiting ribosomal proteins and their mRNAs may resemble the interactions between these proteins and the ribosomal RNA to which they bind when assembled into mature ribosomes. For example, the mRNA transcript of the *str* operon, which includes the coding region for the ribosomal protein S7, contains some regions of RNA sequence that are identical to the S7 binding site of 16S rRNA. Moreover, the proposed secondary structure of the *str* mRNA resembles the proposed secondary structure of the 16S rRNA S7 binding site (Figure 22.25). S7 binds to this region of the *str* mRNA molecule and inhibits translation. It is likely that S7 recognizes analogous structural features in both RNA molecules. Similar mechanisms regulate the translation of mRNAs that encode the other ribosomal proteins.

The ribosomal proteins that inhibit translation bind more tightly to ribosomal RNA than to the similar sites on messenger RNA. Thus, the mRNA continues to be translated as long as newly synthesized ribosomal proteins are incorporated into ribosomes. However, as soon as ribosome assembly slows and the concentration of free ribosomal proteins increases within the cell, the inhibiting ribosomal proteins bind to their own mRNA molecules and block additional protein synthesis. In this way, synthesis of ribosomal proteins is coordinated with ribosome assembly.

## BOX 22.1 SOME ANTIBIOTICS INHIBIT PROTEIN SYNTHESIS

Many microorganisms produce antibiotics, which they use as a chemical defense against competitors. Some antibiotics prevent bacterial growth by inhibiting the formation of peptide bonds. For example, the structure of the antibiotic puromycin closely resembles the structure of the 3′ end of an aminoacyl-tRNA molecule. Because of this similarity, puromycin can enter the A site of a ribosome. Peptidyl transferase then catalyzes the transfer of the nascent polypeptide to the free amino group of puromycin (see figure below). The peptidyl-puromycin is bound weakly in the A site and soon dissociates from the ribosome, thereby terminating protein synthesis.

Although puromycin effectively blocks protein synthesis in prokaryotes, it is not clinically useful since it also blocks protein synthesis in eukaryotes and is therefore poisonous to humans. Clinically important antibiotics, which include streptomycin, chloramphenicol, erythromycin, and tetracycline, are specific for bacteria and have little or no effect on eukaryotic protein synthesis. Streptomycin binds to one of the ribosomal proteins in the 30S subunit and inhibits the initiation of translation. Chloramphenicol interacts with the 50S subunit and inhibits peptidyl transferase. Erythromycin binds to the 50S subunit, inhibiting the translocation step. Tetracycline binds to the 30S subunit, preventing the binding of aminoacyl-tRNA molecules to the A site.




▲ Formation of a peptide bond between puromycin at the A site of a ribosome and the nascent peptide bound to the tRNA in the P site. The product of this reaction is bound only weakly in the A site and dissociates from the ribosome, thus terminating protein synthesis and producing an incomplete, inactive peptide.

### B. Globin Synthesis Depends on Heme Availability

The synthesis of hemoglobin, the major protein in red blood cells, requires globin chains and heme in stoichiometric amounts (Section 4.12). One way globin synthesis is controlled is by regulation of translation initiation. Hemoglobin is initially synthesized in immature erythrocytes called rubriblasts. Mammalian rubriblasts lose their nuclei during maturation and eventually become reticulocytes, which are the

immediate precursors of erythrocytes. Hemoglobin continues to be synthesized in reticulocytes that are packed with processed, stable mRNA molecules encoding globin polypeptides.

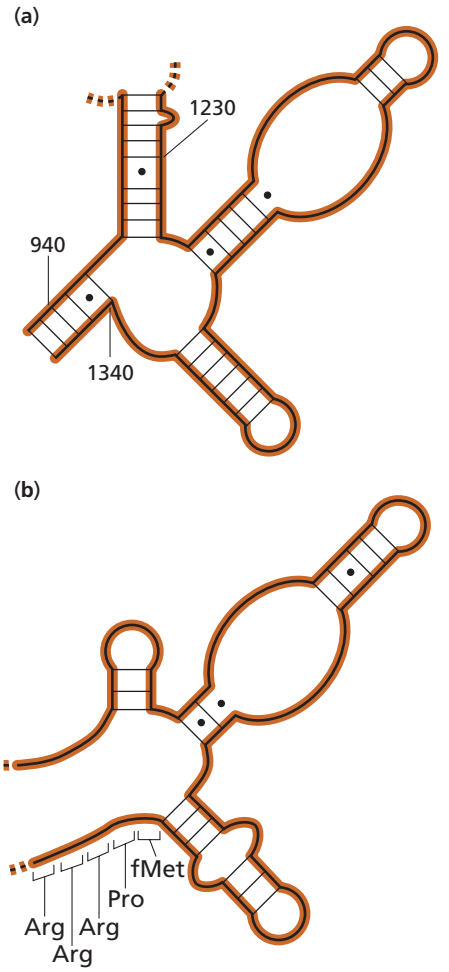
The rate of globin synthesis in reticulocytes is determined by the concentration of heme. When the concentration of heme decreases, the translation of globin mRNA is inhibited. The effect of heme on globin mRNA translation is mediated by a protein kinase called heme-controlled inhibitor (HCI) (Figure 22.26). Active HCI catalyzes transfer of a phosphoryl group from ATP to the translation initiation factor eIF-2. Phosphorylated eIF-2 is unable to participate in translation initiation and protein synthesis in the cell is inhibited.

During the initiation of translation, eIF-2 binds methionyl-$tRNA_i^{Met}$ and GTP. When the preinitiation complex encounters an initiation codon, methionyl-$tRNA_i^{Met}$ is transferred from eIF-2 to the initiation codon of the mRNA. This transfer reaction is accompanied by the hydrolysis of GTP and the release of eIF-2-GDP. An enzyme called guanine nucleotide exchange factor (GEF) catalyzes the replacement of GDP with GTP on eIF-2 and the attachment of another methionyl-$tRNA_i^{Met}$ to eIF-2. GEF binds very tightly to phosphorylated eIF-2–GDP, preventing the nucleotide exchange reaction. Protein synthesis is completely inhibited when all the GEF in the cell is bound because the active eIF-2–GTP complex cannot be regenerated.

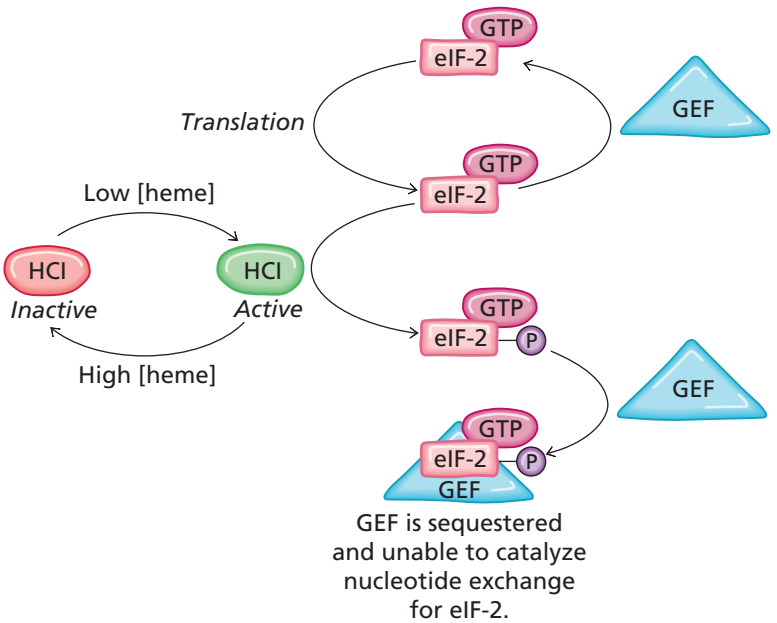
Heme regulates the synthesis of globin by interfering with the activation of HCI. When heme is abundant, HCI is inactive and globin mRNA can be translated. When heme is scarce, however, HCI is activated and translation of all mRNA within the cell is inhibited (Figure 22.26). Phosphorylation of eIF-2 also appears to regulate the translation of mRNA in other mammalian cell types. For example, during infection of human cells by RNA viruses, the presence of double-stranded RNA leads to the production of interferon, which in turn activates a protein kinase that phosphorylates eIF-2. This reaction inhibits protein synthesis in the virus-infected cell.

## C. The *E. coli trp* Operon Is Regulated by Repression and Attenuation

The *trp* operon in *E. coli* encodes the proteins necessary for the biosynthesis of tryptophan. Most organisms synthesize their own amino acids but can also obtain them by degrading exogenous proteins. For this reason, most organisms have evolved mechanisms




▲ **Figure 22.25**
**Comparison of proposed secondary structures of S7 binding sites. (a)** S7 binding site on 16S rRNA. **(b)** S7 binding site on the *str* mRNA molecule.

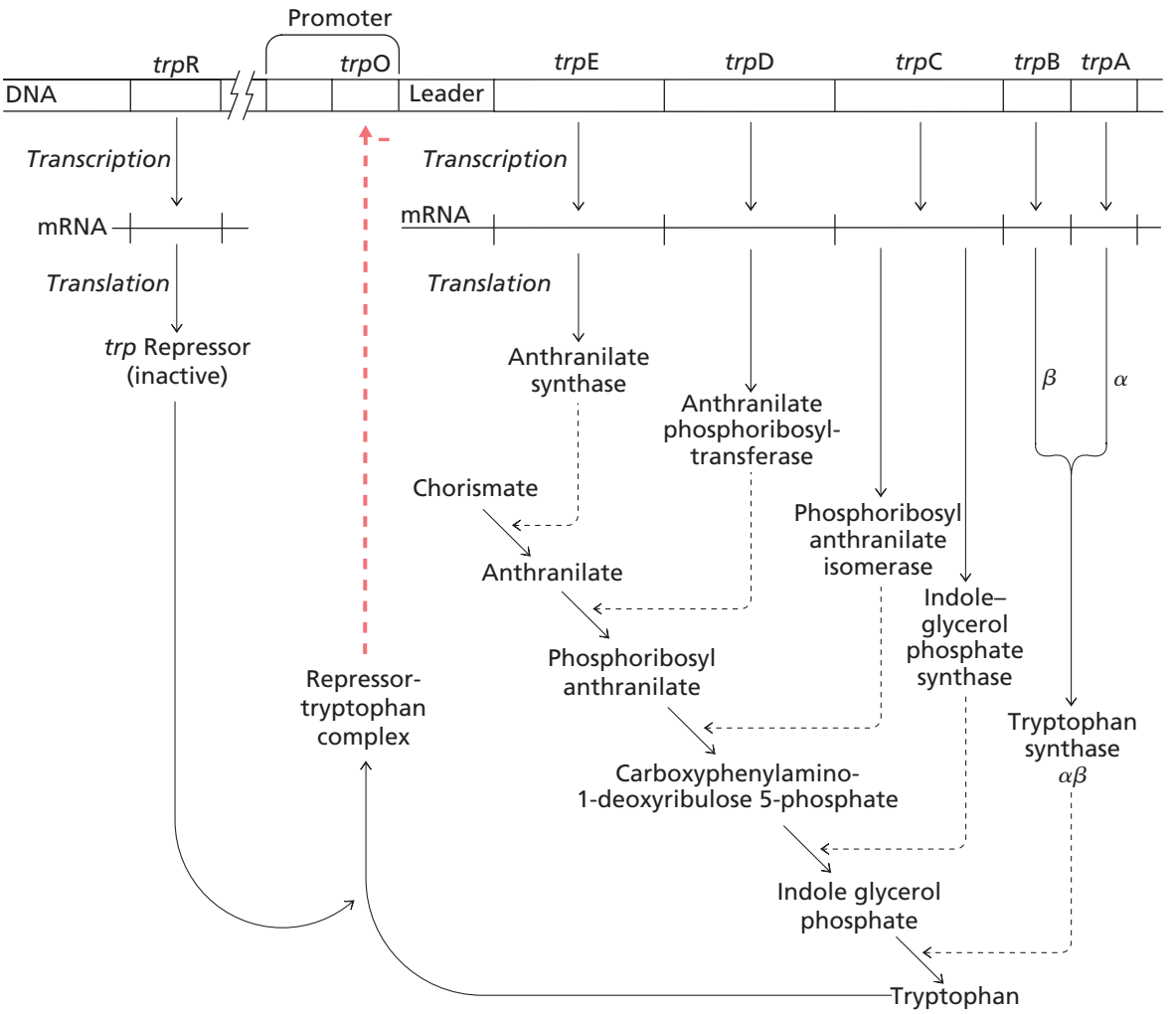



◄ **Figure 22.26**
**Inhibition of protein synthesis by phosphorylation of eIF-2 in reticulocytes.** When the concentration of heme is high, HCI is inactive and translation proceeds normally. When the concentration of heme is low, HCI catalyzes the phosphorylation of eIF-2. Phosphorylated eIF-2 binds the limiting amounts of GEF in the cell very tightly, sequestering the GEF and preventing translation of cellular mRNAs (including the globins).

to repress the synthesis of the enzymes required for *de novo* amino acid biosynthesis when the amino acid is available from exogenous sources. For example, in *E. coli,* tryptophan is a negative regulator of its own biosynthesis. In the presence of tryptophan, the *trp* operon is not expressed (Figure 22.27). Expression of the *trp* operon is inhibited in part by *trp* repressor, a dimer of two identical subunits. *trp* repressor is encoded by the *trpR* gene, which is located elsewhere on the bacterial chromosome and is transcribed separately. When tryptophan is abundant, a repressor-tryptophan complex binds to the operator *trpO*, which lies within the promoter. The bound repressor-tryptophan complex prevents RNA polymerase from binding to the promoter. Tryptophan is thus a corepressor of the *trp* operon.

Regulation of the *E. coli trp* operon is supplemented and refined by a second, independent mechanism called **attenuation**. This second mechanism depends on translation and helps determine whether transcription of the *trp* operon proceeds or terminates prematurely. The movement of RNA polymerase from the promoter into the *trpE* gene is governed by a 162 nucleotide sequence that lies between the promoter and *trpE*. This sequence, called the leader region (Figure 22.27), includes a stretch of 45 nucleotides that encodes a 14 amino acid peptide called the *leader peptide*. The mRNA transcript of the leader region contains two consecutive codons specifying tryptophan near the end of the coding region for the leader peptide. In addition, the

**▼ Figure 22.27**
**Repression of the *E. coli trp* operon.** The *trp* operon is composed of a leader region and five genes required for the biosynthesis of tryptophan from chorismate. The *trpR* gene, located upstream of the *trp* operon (*trpO*), encodes *trp* repressor, which is inactive in the absence of its corepressor, tryptophan. When tryptophan is present in excess, it binds to *trp* repressor, and the repressor-tryptophan complex binds to the *trp* operator (*trpO*). Once bound to the operator, the repressor-tryptophan complex prevents further transcription of the *trp* operon by excluding RNA polymerase from the promoter.



leader region contains four GC-rich sequences. The codons that specify tryptophan and the four GC-rich sequences regulate the synthesis of mRNA by affecting transcription termination.

When transcribed into mRNA, the four GC-rich sequences of the leader region can base-pair to form one of two alternative secondary structures (Figure 22.28, on the next page). The first possible secondary structure includes two RNA hairpins. These hairpins form between the sequences labeled 1 and 2 and between those labeled 3 and 4 in Figure 22.28a. The 1-2 hairpin is a typical transcription pause site. The 3-4 hairpin is followed by a string of uridylate residues, which is a typical rho-independent termination signal (Section 21.4). This particular termination signal is unusual, however, because it occurs upstream of the first gene in the *trp* operon. The other possible secondary structure includes a single RNA hairpin between sequences 2 and 3. This hairpin, which is more stable than the 3-4 hairpin, forms only when sequence 1 is not available for hairpin formation with sequence 2.

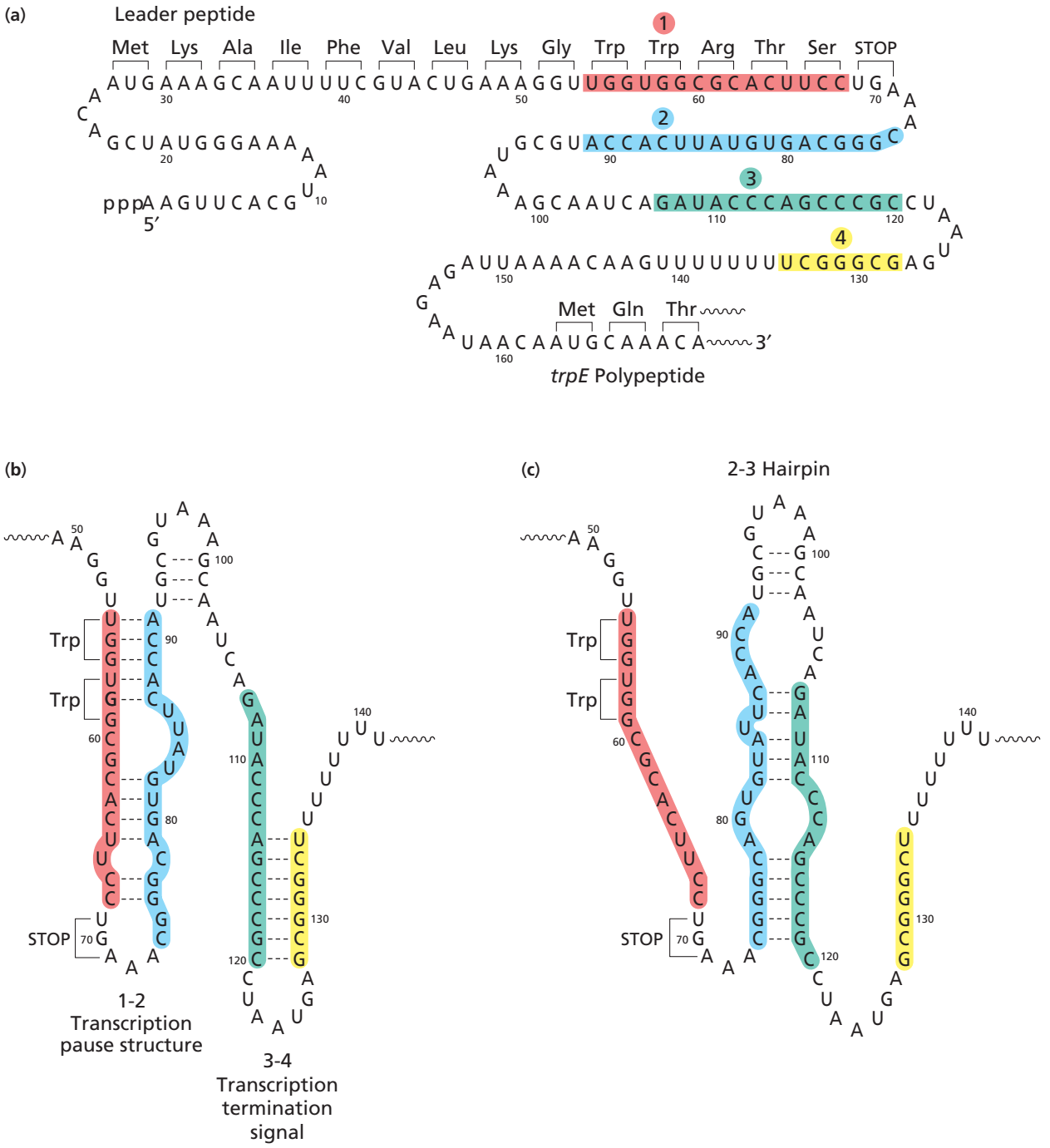
During transcription of the leader region, RNA polymerase pauses when the 1-2 hairpin forms. While RNA polymerase pauses, a ribosome initiates translation of the mRNA encoding the leader peptide. This coding region begins just upstream of the 1-2 RNA hairpin. Sequence 1 encodes the C-terminal amino acids of the leader peptide and also contains a termination codon. As the ribosome translates sequence 1, it disrupts the 1-2 hairpin, thereby releasing the paused RNA polymerase, which then transcribes sequence 3. In the presence of tryptophanyl-$tRNA^{Trp}$, the ribosome and RNA polymerase move at about the same rate. When the ribosome encounters the termination codon of the *trp* leader mRNA, it dissociates and the 1-2 hairpin re-forms. After the ribosome has disassembled, RNA polymerase transcribes sequence 4, which forms a transcription termination hairpin with sequence 3. This termination signal causes the transcription complex to dissociate from the DNA template before the genes of the *trp* operon have been transcribed.

When tryptophan is scarce, however, the ribosome and RNA polymerase do not move synchronously. When the concentration of cellular tryptophan falls, the cell becomes deficient in tryptophanyl-$tRNA^{Trp}$. Under these circumstances, the ribosome pauses when it reaches the two codons specifying tryptophan in sequence 1 of the mRNA molecule. RNA polymerase, which has already been released from the 1-2 pause site, transcribes sequences 3 and 4. While the ribosome is stalled and sequence 1 is covered, sequence 2 forms a hairpin loop with sequence 3. Since the 2-3 hairpin is more stable than the 3-4 hairpin, sequence 3 does not pair with sequence 4 to form the transcription termination hairpin. Under these conditions, RNA polymerase passes through the potential termination site (UGA in Figure 22.28a), and the rest of the *trp* operon is transcribed.

Attenuation appears to be a regulatory mechanism that has evolved relatively recently and is found only in enteric bacteria, such as *E. coli.* (Attenuation cannot occur in eukaryotes because transcription and translation take place in different parts of the cell.) Several *E. coli* operons, including the *phe, thr, his, leu,* and *ile* operons, are regulated by attenuation. Some operons, such as the *trp* operon, combine attenuation with repression, whereas others, such as the *his* operon, are regulated solely by attenuation. The leader peptides of operons whose genes are involved in amino acid biosynthesis may contain as many as seven codons specifying a particular amino acid.

## 22.10 Post-Translational Processing

As the translation complex moves along the mRNA template in the 5′ → 3′ direction, the nascent polypeptide chain lengthens. The 30 or so most recently polymerized amino acid residues remain buried in the ribosome, but amino acid residues closer to the N-terminus are extruded from the ribosome. The N-terminal residues start to fold into the native protein structure even before the C-terminus of the protein has




▲ **Figure 22.28**

***trp* leader region.** **(a)** mRNA transcript of the *trp* leader region. This 162 nucleotide mRNA sequence includes four GC-rich sequences and the coding region for a 14 amino acid leader peptide. The coding region includes two consecutive codons specifying tryptophan. The four GC-rich sequences can base-pair to form one of two alternative secondary structures. **(b)** Sequence 1 (red) and sequence 2 (blue) are complementary and, when base-paired, form a typical transcription pause site. Sequence 3 (green) and sequence 4 (yellow) are complementary and, when base-paired, form a rho-independent termination site. **(c)** Sequences 2 and 3 are also complementary and can form an RNA hairpin that is more stable than the 3-4 hairpin. This structure forms only when sequence 1 is not available for hairpin formation with sequence 2.

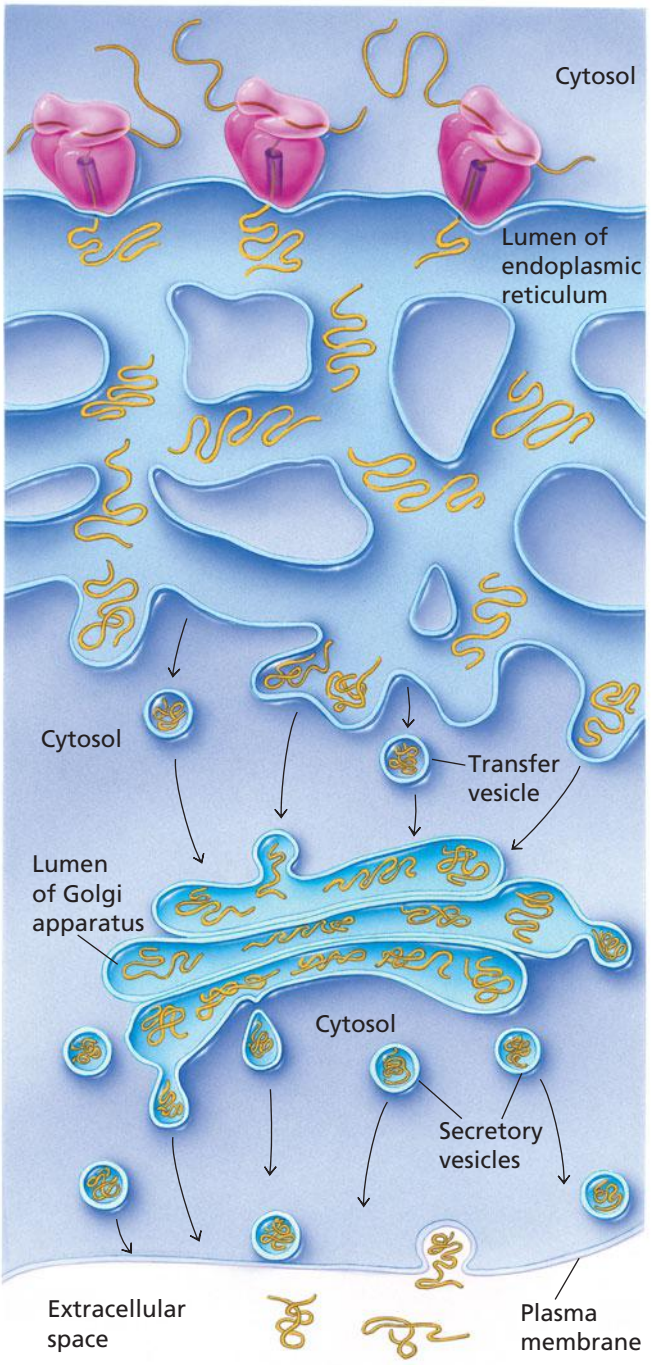
been synthesized. As these residues fold, they are acted on by enzymes that modify the nascent chain.

Modifications that occur before the polypeptide chain is complete are said to be **cotranslational**, whereas those that occur after the chain is complete are said to be **post-translational**. Some examples from the multitude of cotranslational and post-translational modifications include deformylation of the N-terminal residue in prokaryotic

proteins, removal of the N-terminal methionine from prokaryotic and eukaryotic proteins, formation of disulfide bonds, cleavage by proteinases, phosphorylation, addition of carbohydrate residues, and acetylation.

One of the most important events that occurs co- and post-translationally is the processing and transport of proteins through membranes. Protein synthesis occurs in the cytosol, but the mature forms of many proteins are embedded in membranes or are inside membrane bounded compartments. For example, many receptor proteins are embedded in the external membrane of the cell, with the bulk of the protein outside the cell. Other proteins are secreted from cells, and still others reside in lysosomes and other organelles inside eukaryotic cells. In each case, the protein synthesized in the cytosol must be transported across a membrane barrier. In fact, such proteins are synthesized by membrane bound ribosomes that are attached to the plasma membrane in bacteria and to the endoplasmic reticulum in eukaryotic cells.

The best-characterized transport system is the one that carries proteins from the cytosol to the plasma membrane for secretion (Figure 22.29). In eukaryotes, proteins destined for secretion are transported across the membrane of the endoplasmic reticulum into the lumen, which is topologically equivalent to the cell exterior. Once the protein has been transported into the endoplasmic reticulum, it can be transported by vesicles through the Golgi apparatus to the plasma membrane for release outside the cell.




▲ **Figure 22.29**
**Secretory pathway in eukaryotic cells.** Proteins whose synthesis begins in the cytosol are transported into the lumen of the endoplasmic reticulum. After further modification in the Golgi apparatus, the proteins are secreted.

## A. The Signal Hypothesis

Secreted proteins are synthesized on the surface of the endoplasmic reticulum, and the newly synthesized protein is passed through the membrane into the lumen. In cells that make large amounts of secreted protein, the endoplasmic reticulum membranes are covered with ribosomes (Figure 22.30, on the next page).

The clue to the process by which many proteins cross the membrane of the endoplasmic reticulum appears in the first 20 or so residues of the nascent polypeptide chain. In most membrane bound and secreted proteins, these residues are present only in the nascent polypeptide, not in the mature protein. The N-terminal sequence of residues that is proteolytically removed from the protein precursor is called the **signal peptide** since it is the portion of the precursor that signals the protein to cross a membrane. Signal peptides vary in length and composition, but they are typically from 16 to 30 residues long and include 4 to 15 hydrophobic residues (Figure 22.31, on the next page).

In eukaryotes, many proteins destined for secretion appear to be translocated across the endoplasmic reticulum by the pathway shown in Figure 22.32 on page 693. In the first step, an 80S initiation complex—including a ribosome, a Met-tRNA$_i^{Met}$ molecule, and an mRNA molecule—forms in the cytosol. Next, the ribosome begins translating the mRNA and synthesizing the signal peptide at the N-terminus of the precursor. Once the signal peptide has been synthesized and extruded from the ribosome, it binds to a protein-RNA complex called a signal recognition particle (SRP).

SRP is a small ribonucleoprotein containing a 300 nucleotide RNA molecule called 7SL RNA and four proteins. SRP recognizes and binds to the signal peptide as it emerges from the ribosome. When SRP binds, further translation is blocked. The SRP-ribosome complex then binds to an SRP receptor protein (also known as docking protein) on the cytosolic face of the endoplasmic reticulum. The ribosome is anchored to the membrane of the endoplasmic reticulum by ribosome binding proteins called translocons, and the signal peptide is inserted into the membrane at a pore that is part

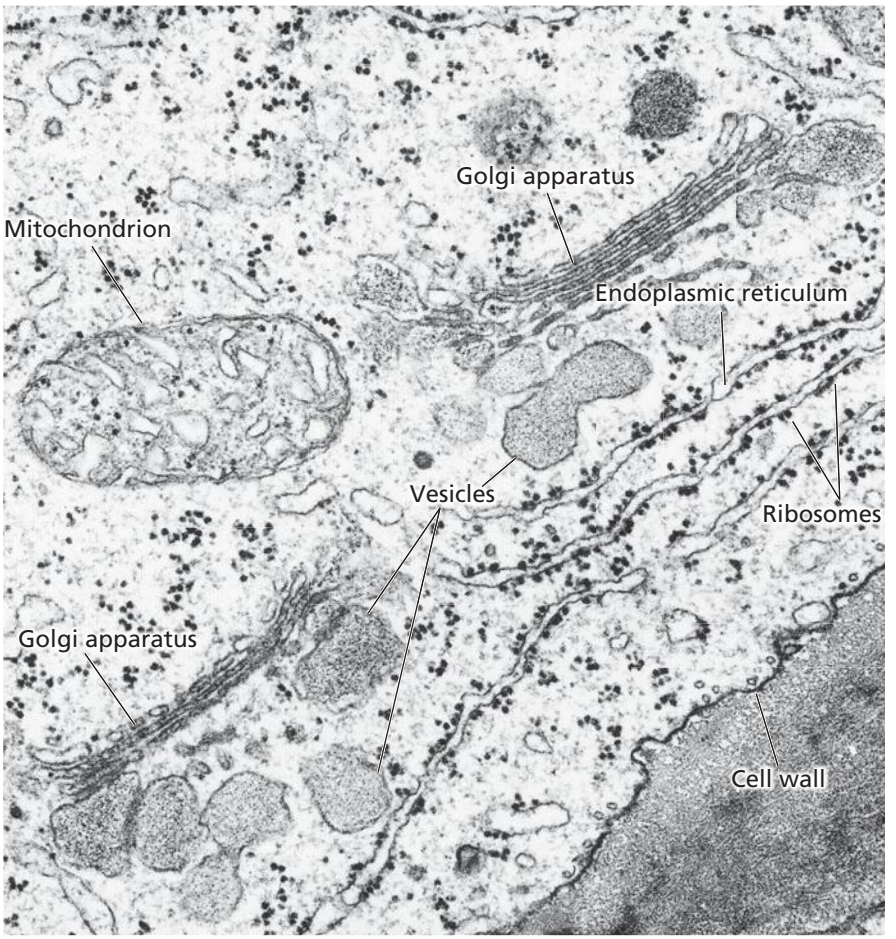



**Figure 22.30 ►**
**Secretory vesicles in a maize rootcap cell.** Large secretory vesicles containing proteins are budding off the Golgi apparatus (center). Note the abundance of ribosomes bound to the endoplasmic reticulum.

of the complex formed by the endoplasmic reticulum proteins at the docking site. Once the ribosome-SRP complex is bound to the membrane, the inhibition of translation is relieved and SRP dissociates in a reaction coupled to GTP hydrolysis. Thus, the role of SRP is to recognize nascent polypeptides containing a signal peptide and to target the translation complex to the surface of the endoplasmic reticulum.

Once the translation complex is bound to the membrane, translation resumes and the new polypeptide chain passes through the membrane. The signal peptide is then cleaved from the nascent polypeptide by a signal peptidase, an integral membrane protein associated with the pore complex. The transport of proteins across the membrane

**Figure 22.31 ▼**
**Signal peptides from secreted proteins.** Hydrophobic residues are shown in blue, and arrows mark the sites where the signal peptide is cleaved from the precursor. (OmpA is a bacterial membrane protein.)

Prelysozyme

$H_3\overset{\oplus}{N}$–Met–Arg–Ser–Leu–Leu–Ile–Leu–Val–Leu–Cys–Phe–Leu–Pro–Leu–Ala–Ala–Leu–Gly↓–Gly ~~

Preproalbumin

$H_3\overset{\oplus}{N}$–Met–Lys–Trp–Val–Thr–Phe–Leu–Leu–Leu–Leu–Phe–Ile–Ser–Gly–Ser–Ala–Phe–Ser↓–Arg ~~

Alkaline phosphatase

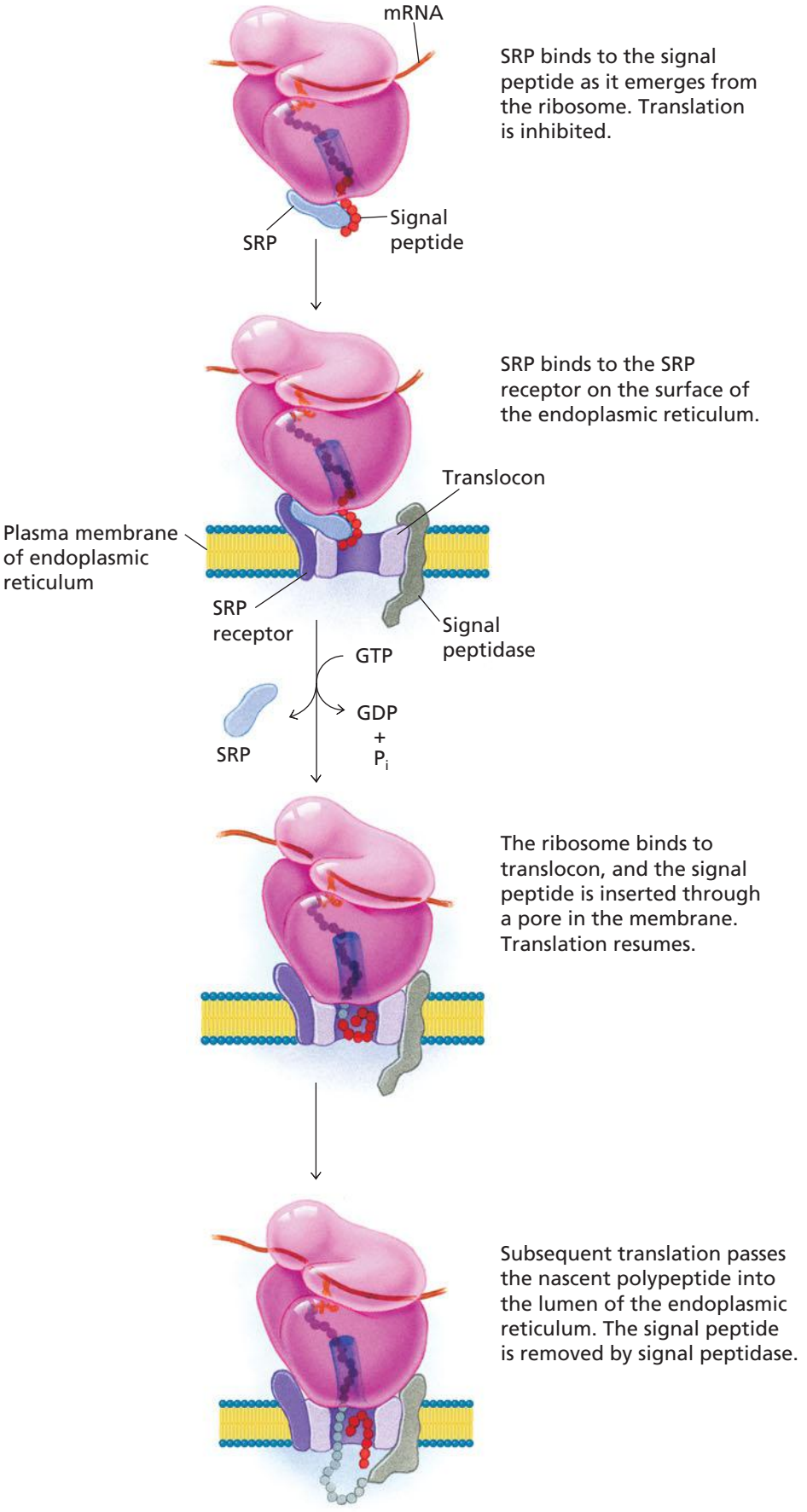
$H_3\overset{\oplus}{N}$–Met–Lys–Gln–Ser–Thr–Ile–Ala–Leu–Ala–Leu–Leu–Pro–Leu–Leu–Phe–Thr–Pro–Val–Thr–Lys–Ala↓–Arg ~~

Maltose-binding protein

$H_3\overset{\oplus}{N}$–Met–Lys–Ile–Lys–Thr–Gly–Ala–Arg–Ile–Leu–Ala–Leu–Ser–Ala–Leu–Thr–Thr–Met–Met–Phe–Ser–Ala–Ser–Ala–Leu–Ala↓–Lys ~~

OmpA

$H_3\overset{\oplus}{N}$–Met–Lys–Lys–Thr–Ala–Ile–Ala–Ile–Ala–Val–Ala–Leu–Ala–Gly–Phe–Ala–Thr–Val–Ala–Gln–Ala↓–Ala ~~




◀ **Figure 22.32**
**Translocation of eukaryotic proteins into the lumen of the endoplasmic reticulum.**

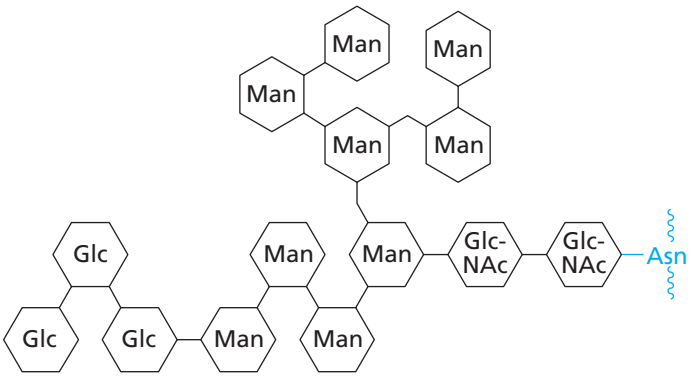



**Figure 22.33 ►**
**Structure of a complex oligosaccharide linked to an asparagine residue.** Abbreviations: Glc, glucose; GlcNAc, *N*-acetylglucosamine; Man, mannose.

is assisted by chaperones in the lumen of the endoplasmic reticulum. In addition to their role in protein folding, chaperones are required for translocation, and their activity requires ATP hydrolysis. When protein synthesis terminates, the ribosome dissociates from the endoplasmic reticulum, and the translation complex disassembles.

### B. Glycosylation of Proteins

Many integral membrane proteins and secretory proteins contain covalently bound oligosaccharide chains. The addition of these chains to proteins is called protein glycosylation (Section 8.7C). Protein glycosylation is one of the major metabolic activities of the lumen of the endoplasmic reticulum and of the Golgi apparatus and is an extension of the general process of protein biosynthesis. A glycoprotein can contain dozens, indeed hundreds, of monosaccharide units. The mass of the carbohydrate portion may account for as little as 1% or as much as 80% of the mass of the glycoprotein.

A common glycosylation reaction involves the covalent attachment of a complex oligosaccharide to the side chain of an asparagine residue (Figure 22.33). During subsequent transit through the endoplasmic reticulum and the Golgi apparatus, proteins may be covalently modified in many ways, including the formation of disulfide bonds and proteolytic cleavage. The complex oligosaccharides attached to the proteins are likewise modified during transit. A variety of different oligosaccharides can be covalently bound to proteins. In some cases, the structure of the oligosaccharide acts as a signal to target proteins to a specific location. For example, lysosomal proteins contain sites for the attachment of an oligosaccharide that targets these proteins to the lysosome. By the time they have traversed the Golgi apparatus, the proteins and their oligosaccharides are usually fully modified.

## Summary

1. The genetic code consists of nonoverlapping, three-nucleotide codons. The code is unambiguous and degenerate; the first two nucleotides of the three-letter code are often sufficient; codons with similar sequences specify chemically similar amino acids; and there are special codons for the initiation and termination of peptide synthesis.
2. tRNA molecules are the adapters between mRNA codons and amino acids in proteins. All tRNA molecules have a similar cloverleaf secondary structure with a stem and three arms. The tertiary structure is L-shaped. The anticodon loop is at one end of the structure, and the acceptor stem is at the other. The anticodon in tRNA base-pairs with a codon in mRNA. The 5′ (wobble) position of the anticodon is conformationally flexible.
3. An aminoacyl-tRNA synthetase catalyzes the addition of a specific amino acid to the acceptor stem of the appropriate tRNA, producing an aminoacyl-tRNA. Some aminoacyl-tRNA synthetases carry out proofreading.
4. Ribosomes are the RNA-protein complexes that catalyze the polymerization of amino acids bound to aminoacyl-tRNA molecules. All ribosomes are composed of two subunits: prokaryotic ribosomes contain three rRNA molecules, and eukaryotic ribosomes contain four. The growing polypeptide chain is attached to a tRNA in the peptidyl (P) site of the ribosome, and the aminoacyl-tRNA molecule bearing the next amino acid to be added to the nascent polypeptide chain docks in the aminoacyl (A) site.
5. Translation begins with the formation of an initiation complex consisting of an initiator tRNA, the mRNA template, the ribosomal subunits, and several initiation factors. In prokaryotes, initiation occurs just downstream of Shine-Dalgarno sequences; in eukaryotes, initiation usually occurs at the initiation codon closest to the 5′ end of the mRNA.
6. The elongation step of translation requires accessory proteins called elongation factors. The three steps of elongation are (1) positioning of the correct aminoacyl-RNA in the A site,

(2) formation of the peptide bond by peptidyl transferase, and (3) translocation of the ribosome by one codon.

7. Release factors recognize termination codons and catalyze the termination of protein synthesis and disassembly of the translation complex.
8. Protein synthesis requires the energy of four phosphoanhydride bonds per residue.
9. The regulation of translation includes the formation of secondary structure in mRNA that influences the rate of initiation. Ribosomal RNA proteins can inhibit translation of their own mRNA by binding to such sites. Phosphorylation of an initiation factor regulates globin synthesis. Regulation of expression of the *E. coli trp* operon involves attenuation, in which translation of a leader mRNA governs transcription of the operon.
10. Many proteins are post-translationally modified. Some eukaryotic proteins destined for secretion contain N-terminal signals for transport into the endoplasmic reticulum. Many membrane and secreted proteins are glycosylated.

# Problems

1. The standard genetic code is read in codons that are three nucleotides long. How many potential reading frames are there on a single piece of double-stranded DNA? If instead the genetic code was read in codons that were four nucleotides long, how many reading frames would there be on the same piece of double-stranded DNA?
2. Examine the sequences of the mRNAs transcribed from the DNA sequence in Problem 11 in Chapter 21. Assuming that the DNA segment is from the middle of a protein-coding gene, which of the possible mRNAs is most likely to be the actual transcript? What is the sequence of the encoded peptide?
3. Calculate the number of phosphoanhydride bonds that are hydrolyzed during synthesis of a 600 amino acid residue protein in *E. coli*. Do not include the energy required to synthesize the amino acids, mRNA, tRNA, or the ribosomes.
4. Polypeptide chain elongation on the ribosome can be broken down into three discrete steps (the microcycle): (1) binding of the correct aminoacyl-tRNA in the ribosome's A site, (2) peptide bond formation, and (3) translocation. What, specifically, is it that gets translocated in the third step of this cycle?
5. A prokaryotic mRNA may contain many AUG codons. How does the ribosome distinguish AUG codons specifying initiation from AUG codons specifying internal methionine?
6. Given that the genetic code is universal, would a plant mRNA be correctly translated in a prokaryotic cell like *E. coli*?
7. Bacterial genomes usually contain multiple copies of the genes for rRNA. These are transcribed very efficiently in order to produce large amounts of rRNA for assembly into ribosomes. In contrast, the genes that encode ribosomal proteins are present only as single copies. Explain the difference in the number of copies of rRNA and ribosomal protein genes.
8. Suppressor mutations suppress the effects of other mutations. For example, mutations that produce the stop codon UAG in the middle of a gene are suppressed by an additional mutation in a tRNA gene that gives rise to a mutant anticodon with the sequence CUA. Consequently, an amino acid is inserted at the mutant stop codon, and a protein is synthesized (although it may be only partially active). List all the tRNA species that could be mutated to a suppressor of UAG mutations by a single base change in the anticodon. How can a cell with a suppressor tRNA survive?
9. Transfer RNAs are absolutely essential for polypeptide synthesis. After reviewing the material in this chapter, name five different cellular components that can bind to (interact with) tRNA molecules.
10. On rare occasions, the translation machinery encounters a codon that cannot be quickly interpreted because of the lack of a particular tRNA or release factor. In these cases, the ribosome may pause and then shift by a single nucleotide and begin translating a different reading frame. Such an occurrence is known as translational frameshifting. The *E. coli* release factor RF-2, which is translated from mRNA that contains an internal UGA stop codon, is produced by translational frameshifting. Explain how this phenomenon might regulate RF-2 production.
11. The mechanism of attenuation requires the presence of a leader region. Predict the effect of the following changes on regulation of the *trp* operon:
    (a) The entire leader region is deleted.
    (b) The sequence encoding the leader peptide is deleted.
    (c) The leader region, an AUG codon, is mutated.
12. In Chapter 21, you learned of many different regulatory mechanisms that control transcription of the *lac* operon in *E. coli*. In Chapter 22, one of the mechanisms of translational regulation discussed was called attenuation. Would you predict that in some other bacterial species the *lac* operon might have evolved such that an attenuation mechanism was used to regulate expression levels from this operon?
13. In the operons that contain genes for isoleucine biosynthesis, the leader regions that precede the genes contain multiple codons that specify not only isoleucine but valine and leucine as well. Suggest a reason why this is so.
14. Suggest the steps involved in the synthesis and processing of a glycosylated, eukaryotic integral membrane protein with a C-terminal cytosolic domain and an N-terminal extracellular domain.
15. In Chapter 23, you will learn about recombinant DNA techniques that allow genes to be cut and pasted at will. If you could remove the coding region for a secretion signal sequence from one protein and place it such that it will now occupy the N-terminus of a cytosolic protein (e.g., $\beta$-galactosidase), would you expect the new hybrid protein to enter the cell's secretory pathway?
16. In some species of bacteria, the codon GUG initiates protein synthesis (e.g., LacI, Figure 22.17a). The completed proteins always contain methionine at the N-terminus. How can the initiator tRNA base-pair with the codon GUG? How is this phenomenon related to wobble?

# Selected Readings


## Aminoacyl-tRNA Synthetases

Carter, C. W., Jr. (1993). Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. *Annu. Rev. Biochem.* 62:715–748.

Ibba, M., and Söll, D. (2000). Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* 69:617–650.

Jakubowski, H., and Goldman, E. (1992). Editing of errors in selection of amino acids for protein synthesis. *Microbiol. Rev.* 56:412–429.

Kurland, C. G. (1992). Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* 26:29–50.

Schimmel, P., and Ribas de Pouplana, L. (2000). Footprints of aminoacyl-tRNA synthetases are everywhere. *Trends Biochem. Sci.* 25:207–209.

## Ribosomes and Translation

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4Å resolution. *Science* 289:905–919.

Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* 407:340–348.

Garrett, R. A., Douthwate, S. R., Matheson A. T., Moore, P. B., and Noller, H. F., eds. (2000). *The Ribosome: Structure, Function, Antibiotics and Cellular Interactions* (Washington, DC: American Society for Microbiology).

Hanawa-Suetsugu, K., Sekine, S., Sakai, H., Hori-Takemoto, C., Tevader, T., Unzai, S., Tame, J.R.H., Kuramitsu, S., Shirouzu, M., and Yokoyama, S. (2004). Crystal structure of elongation factor P from *Thermus thermophilus* HB8. *Proc. Natl. Acad. Sci.* 101:9595–9600.

Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., and Leberman, R. (1996). The structure of the *Escherichia coli* EF-Tu • EF-Ts complex at 2.5 Å resolution. *Nature* 379:511–518.

Moore, P. B., and Steitz, T. A. (2003). The structural basis of large ribosomal subunit function. *Annu. Rev. Biochem.* 72:813–850.

Nirenberg, M.W., and Matthaei, J.H., (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* 47:1588–1602.

Noller, H. F. (1993). Peptidyl transferase: protein, ribonucleoprotein, or RNA? *J. Bacteriol.* 175:5297–5300.

Pestova, T. V., and Hellen, C. U. T. (1999). Ribosome recruitment and scanning: what's new? *Trends Biochem. Sci.* 24:85–87.

Ramakrishnan, V. (2009). Unravelling the structure of the ribosome. Nobel Lecture 135–160.

Selmer, M., Al-Karadaghi, S., Hirokawa, G., Kaji, A., and Liljas, A. (1999). Crystal Structure of *Thermotoga maritima* ribosome recycling factor: A tRNA mimic. *Science* 286:2349–2352.

Steitz, T.A. (2009). From the structure and function of the ribosome to new antibiotics. Nobel Lecture 179–204.

## Regulation of Translation

Kozak, M. (1992). Regulation of translation in eukaryotic systems. *Annu. Rev. Cell Biol.* 8:197–225.

McCarthy, J. E. G., and Gualerzi, C. (1990). Translational control of prokaryotic gene expression. *Trends Genet.* 6:78–85.

Merrick, W. C. (1992). Mechanism and regulation of eukaryotic protein synthesis. *Microbiol. Rev.* 56:291–315.

Rhoads, R. E. (1993). Regulation of eukaryotic protein synthesis by initiation factors. *J. Biol. Chem.* 268:3017–3020.

Samuel, C. E. (1993). The eIF-2a protein kinases, regulators of translation in eukaryotes from yeasts to humans. *J. Biol. Chem.* 268:7603–7606.

## Post-translational Modification

Hurtley, S. M. (1993). Hot line to the secretory pathway. *Trends Biochem. Sci.* 18:3–6.

Parodi, A. J. (2000). Protein glycosylation and its role in protein folding. *Annu. Rev. Biochem.* 69:69–93.