6

Molecular Biology of *Bacteria*

The essence of life is a cell's organization and the orderly replication of its DNA. Seen here, DNA is emerging from a bacterial cell treated to release its chromosome.

Cells may be regarded as chemical machines and coding devices. As chemical machines, cells transform their vast array of macromolecules into new cells. As coding devices, they store, process, and use genetic information. Genes and gene expression are the subject of molecular biology. In particular, the review of molecular biology in this chapter covers the chemical nature of genes, the structure and function of DNA and RNA, and the replication of DNA. We then consider the synthesis of proteins, macromolecules that play important roles in both the structure and the functioning of the cell. Our focus here is on these processes as they occur in *Bacteria*. In particular, *Escherichia coli*, a member of the *Bacteria*, is the model organism for molecular biology and is the main example used. Although *E. coli* was not the first bacterium to have its chromosome sequenced, this organism remains the best characterized of any organism, prokaryote or eukaryote.

I DNA Structure and Genetic Information

6.1 Macromolecules and Genes

The functional unit of genetic information is the **gene**. All life forms, including microorganisms, contain genes. Physically, genes are located on chromosomes or other large molecules known collectively as **genetic elements**. Nowadays, in the "genomics era," biology tends to characterize cells in terms of their complement of genes. Thus, if we wish to understand how microorganisms function we must understand how genes encode information.

Chemically, genetic information is carried by the **nucleic acids** deoxyribonucleic acid, **DNA**, and ribonucleic acid, **RNA**. DNA carries the genetic blueprint for the cell and RNA is the intermediary molecule that converts this blueprint into defined amino acid sequences in proteins. Genetic information consists of the sequence of monomers in the nucleic acids. Thus, in contrast to polysaccharides and lipids, nucleic acids are **informational macromolecules**. Because the sequence of monomers in proteins is determined by the sequence of the nucleic acids that encode them, proteins are also informational macromolecules.

The monomers of nucleic acids are called **nucleotides**, consequently, DNA and RNA are **polynucleotides**. A nucleotide has three components: a pentose sugar, either ribose (in RNA) or deoxyribose (in DNA), a nitrogen base, and a molecule of phosphate, PO_4^{3-}. The general structure of nucleotides of both DNA and RNA is very similar (**Figure 6.1**). The nitrogen bases are either **purines** (*adenine* and *guanine*) which contain two fused heterocyclic rings or **pyrimidines** (*thymine*, *cytosine*, and *uracil*) which contain a single six-membered heterocyclic ring (Figure 6.1*a*). Guanine, adenine, and cytosine are present in both DNA and RNA. With minor exceptions, thymine is present only in DNA and uracil is present only in RNA.

The nitrogen bases are attached to the pentose sugar by a glycosidic linkage between carbon atom 1 of the sugar and a nitrogen atom in the base, either nitrogen 1 (in pyrimidine bases) or 9 (in purine bases). A nitrogen base attached to its sugar, but

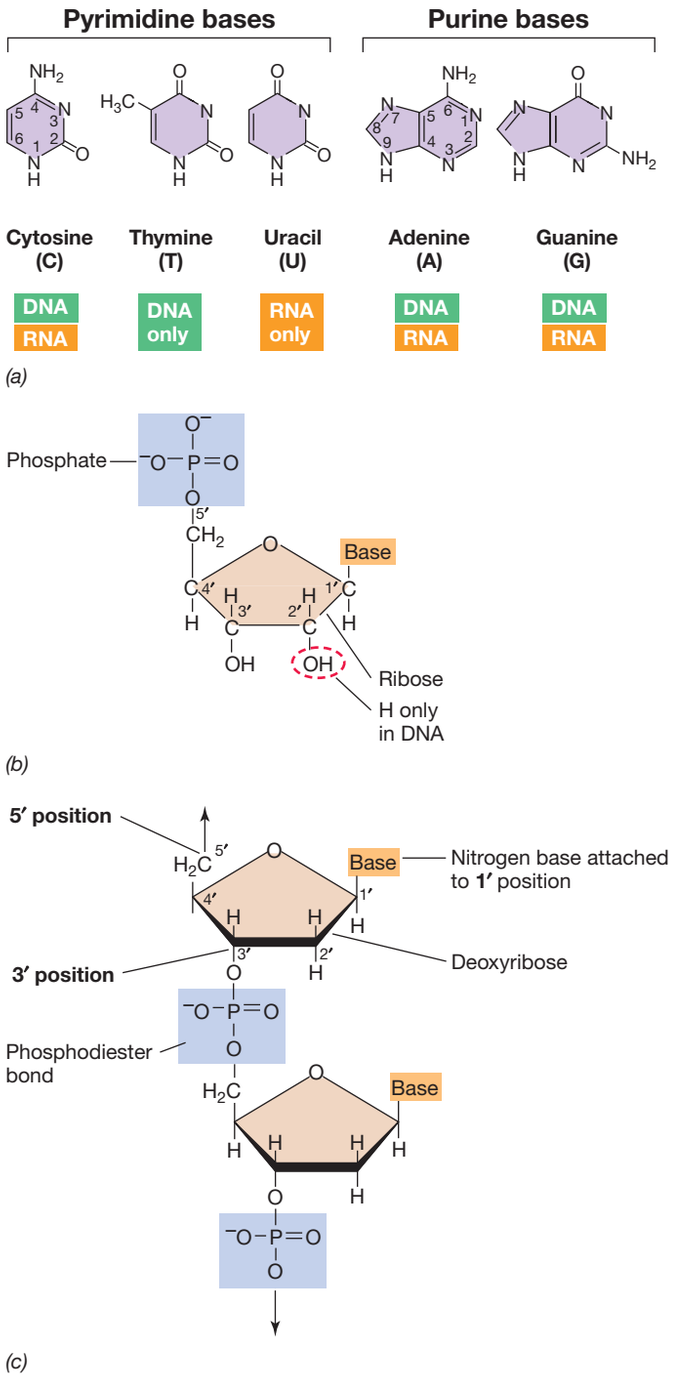


Figure 6.1 **Components of the nucleic acids.** *(a)* The nitrogen bases of DNA and RNA. Note the numbering system of the rings. In attaching itself to the 1′ carbon of the sugar phosphate, a pyrimidine base bonds through N-1 and a purine base bonds at N-9. *(b)* Nucleotide structure. The numbers on the sugar contain a prime (′) after them because the rings of the nitrogen bases are also numbered. In DNA a hydrogen is present on the 2′-carbon of the pentose sugar. In RNA, an OH group occupies this position. *(c)* Part of a DNA chain. The nucleotides are linked by a phosphodiester bond. In addition to the bases shown, transfer RNAs (tRNAs) contain unusual pyrimidines such as pseudouracil and dihydrouracil, and various modified purines not present in other RNAs (see Figure 6.33).

lacking phosphate, is called a **nucleoside**. Nucleotides are nucleosides plus one or more phosphates (Figure 6.1). Nucleotides play other roles in addition to comprising nucleic acids. Nucleotides, especially adenosine triphosphate (ATP) and guanosine triphosphate (GTP), carry chemical energy. Other nucleotides or derivatives function in redox reactions, as carriers of sugars in polysaccharide synthesis, or as regulatory molecules.

The Nucleic Acids, DNA and RNA

The nucleic acid backbone is a polymer of alternating sugar and phosphate molecules. The nucleotides are covalently bonded by phosphate between the 3′- (3 prime) carbon of one sugar and the 5′-carbon of the next sugar. [Numbers with prime marks refer to positions on the sugar ring; numbers without primes to positions on the rings of the bases.] The phosphate linkage is called a **phosphodiester bond** because the phosphate connects two sugar molecules by an ester linkage (Figure 6.1). The sequence of nucleotides in a DNA or RNA molecule is its **primary structure** and the sequence of bases forms the genetic information.

In the genome of cells, DNA is *double-stranded*. Each chromosome consists of two strands of DNA, with each strand containing hundreds of thousands to several million nucleotides linked by phosphodiester bonds. The strands are held together by hydrogen bonds that form between the bases in one strand and those of the other strand. When located next to one another, purine and pyrimidine bases can form hydrogen bonds (**Figure 6.2**). Hydrogen bonding is most stable when guanine (G) bonds with cytosine (C) and adenine (A) bonds with thymine (T). Specific base pairing, A with T and G with C, ensures that the two strands of DNA are *complementary* in base sequence; that is, wherever a G is found in one strand, a C is found in the other, and wherever a T is present in one strand, its complementary strand has an A.

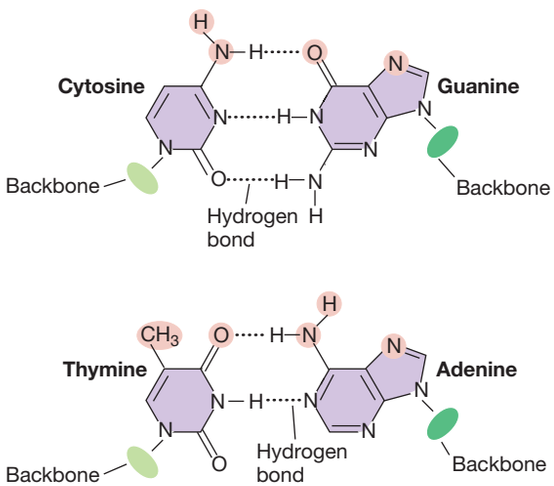


Figure 6.2 Specific pairing between guanine (G) and cytosine (C) and between adenine (A) and thymine (T) via hydrogen bonds. These are the typical base pairs found in double-stranded DNA. Atoms that are found in the major groove of the double helix and that interact with proteins are highlighted in pink. The deoxyribose phosphate backbones of the two strands of DNA are also indicated. Note the different shades of green for the two strands of DNA, a convention used throughout this book.

With a few exceptions, all RNA molecules are *single-stranded*. However, RNA molecules typically fold back upon themselves in regions where complementary base pairing is possible. The term **secondary structure** refers to this folding whereas primary structure refers to the nucleotide sequence. In certain large RNA molecules, such as ribosomal RNA (Section 6.19), some parts of the molecule are unfolded but other regions possess secondary structure. This leads to highly folded and twisted molecules whose biological function depends critically on their final three-dimensional shape.

Genes and the Steps in Information Flow

When genes are expressed, the information stored in DNA is transferred to ribonucleic acid (RNA). Several classes of RNA exist in cells. Three types of RNA take part in protein synthesis. **Messenger RNA** (mRNA) is a single-stranded molecule that carries the genetic information from DNA to the ribosome, the protein-synthesizing machine. **Transfer RNAs** (tRNAs) convert the genetic information on mRNA into the language of proteins. **Ribosomal RNAs** (rRNAs) are important catalytic and structural components of the ribosome. In addition to these, cells contain a variety of *small RNAs* that regulate the production or activity of proteins or other RNAs. The molecular processes of genetic information flow can be divided into three stages (**Figure 6.3**):

1. **Replication**. During replication, the DNA double helix is duplicated, producing two double helices.
2. **Transcription**. Transfer of information from DNA to RNA is called transcription.
3. **Translation**. Synthesis of a protein, using the information carried by mRNA, is known as translation.

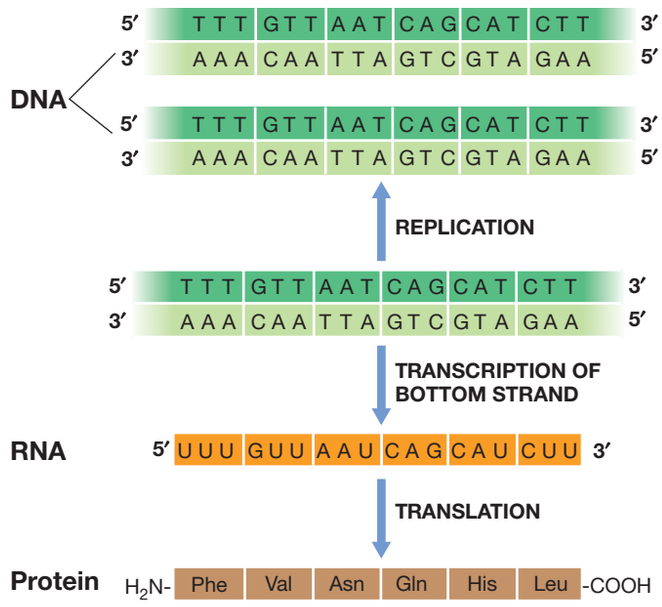


Figure 6.3 Synthesis of the three types of informational macromolecules. Note that for any particular gene only one of the two strands of the DNA double helix is transcribed.

There is a linear correspondence between the base sequence of a gene and the amino acid sequence of a polypeptide. Each group of three bases on an mRNA molecule encodes a single amino acid, and each such triplet of bases is called a **codon**. This genetic code is translated into protein by the ribosomes (which consist of proteins and rRNA), tRNA, and proteins known as translation factors.

The three steps shown in Figure 6.3 are used in all cells and constitute the central dogma of molecular biology (DNA → RNA → protein). Note that many different RNA molecules are each transcribed from a relatively short region of the long DNA molecule. In eukaryotes, each gene is transcribed to give a single mRNA (Chapter 7), whereas in prokaryotes a single mRNA may carry genetic information for several genes, that is, for several protein coding regions. Some viruses violate the central dogma (Chapter 9). Some viruses use RNA as the genetic material and must therefore replicate their RNA using RNA as template. In retroviruses such as HIV—the causative agent of AIDS—an RNA genome is converted to a DNA version by a process called reverse transcription.

MiniQuiz

- What components are found in a nucleotide?
- How does a nucleo*side* differ from a nucleo*tide*?
- Distinguish between the primary and secondary structure of RNA.
- What three informational macromolecules are involved in genetic information flow?
- In all cells there are three processes involved in genetic information flow. What are they?

6.2 The Double Helix

In all cells and many viruses, DNA exists as a double-stranded molecule with two polynucleotide strands whose base sequences are **complementary**. (As discussed in Chapter 9, the genomes of some DNA viruses are single-stranded.) The complementarity of DNA arises because of specific base pairing: adenine always pairs with thymine, and guanine always pairs with cytosine. The two strands of the double-stranded DNA molecule are arranged in an **antiparallel** fashion (**Figure 6.4**, distinguished as two shades of green). Thus, the strand on the left runs 5′ to 3′ from top to bottoms, whereas the other strand runs 5′ to 3′ from bottom to top.

The two strands of DNA are wrapped around each other to form a double helix (**Figure 6.5**) that forms two distinct grooves, the *major groove* and the *minor groove*. Most proteins that interact specifically with DNA bind in the major groove, where there is plenty of space. Because the double helix is a regular structure, some atoms of each base are always exposed in the major groove (and some in the minor groove). Key regions of nucleotides that are important in interactions with proteins are shown in Figure 6.2.

Several double-helical structures are possible for DNA. The Watson and Crick double helix is known as the B-form or *B-DNA* to distinguish it from the A- and Z-forms. The A-form is shorter and fatter than the B-form. It has 11 base pairs per turn, and the

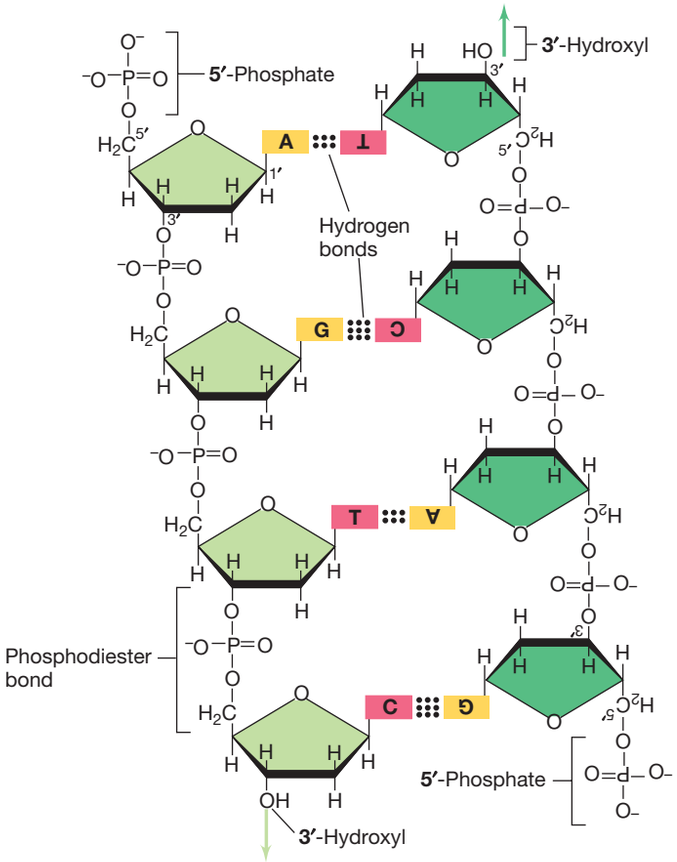


Figure 6.4 DNA structure. Complementary and antiparallel nature of DNA. Note that one chain ends in a 5′-phosphate group, whereas the other ends in a 3′-hydroxyl. The red bases represent the pyrimidines cytosine (C) and thymine (T), and the yellow bases represent the purines adenine (A) and guanine (G).

major groove is narrower and deeper. Double-stranded RNA or hybrids of one RNA plus one DNA strand often form the A-helix. The Z-DNA double helix has 12 base pairs per turn and is left-handed. Its sugar–phosphate backbone is a zigzag line rather than a smooth curve. Z-DNA is found in GC- or GT-rich regions, especially when negatively supercoiled. Occasional enzymes and regulatory proteins bind Z-DNA preferentially.

Size and Shape of DNA Molecules

The size of a DNA molecule is expressed as the number of nucleotide bases or base pairs per molecule. Thus, a DNA molecule with 1000 bases is 1 kilobase (kb) of DNA. If the DNA is a double helix, then *kilobase pairs* (kbp) is used. Thus, a double helix 5000 base pairs in size would be 5 kbp. The bacterium *Escherichia coli* has about 4640 kbp of DNA in its chromosome. When dealing with large genomes the term *megabase pair* (Mbp) for a million base pairs is used. The genome of *E. coli* is thus 4.64 Mbp.

Each base pair takes up 0.34 nanometer (nm) in length along the double helix, and each turn of the helix contains approximately 10 base pairs. Therefore, 1 kbp of DNA is 0.34 μm long with 100 helical turns. The *E. coli* genome is thus 4640 × 0.34 = 1.58 mm

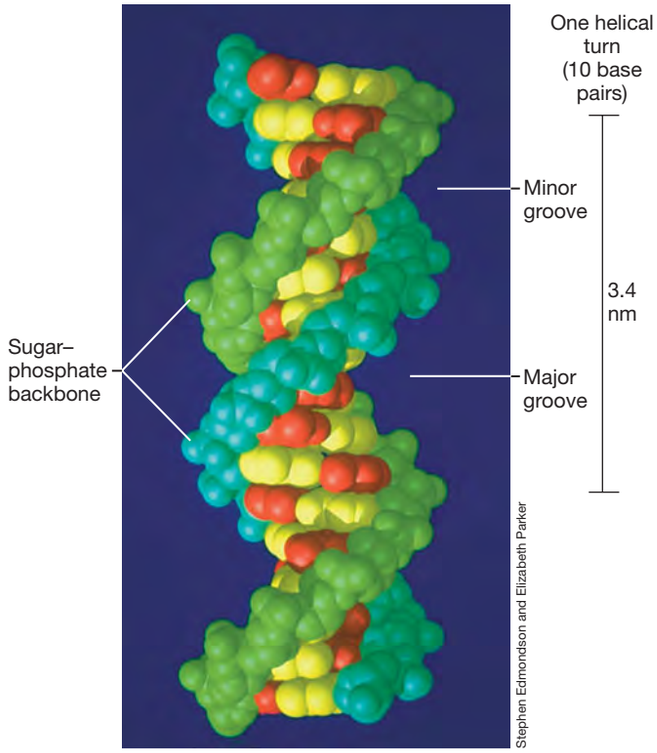


Figure 6.5 **A computer model of a short segment of DNA showing the overall arrangement of the double helix.** One of the sugar–phosphate backbones is shown in blue and the other in green. The pyrimidine bases are shown in red and the purines in yellow. Note the locations of the major and minor grooves (compare with Figure 6.2). One helical turn contains 10 base pairs.

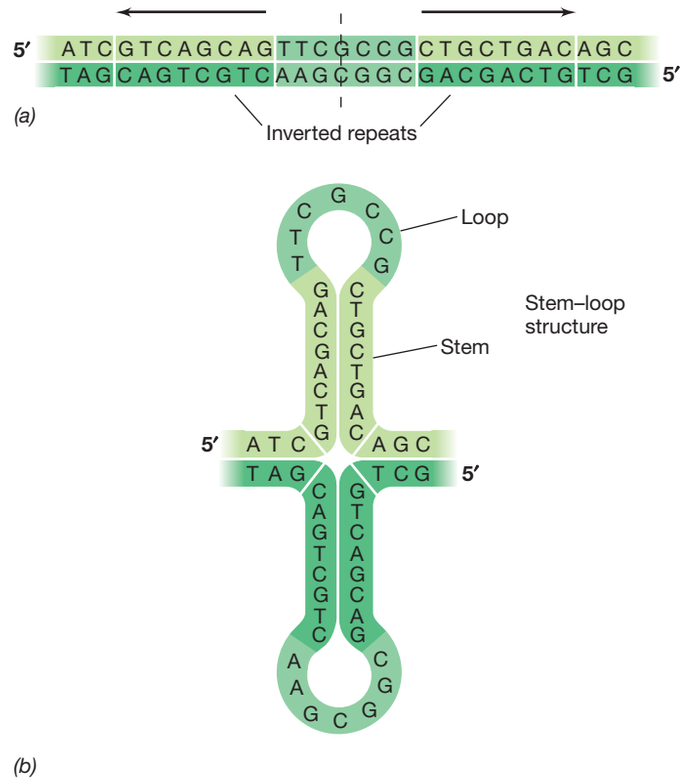


Figure 6.6 **Inverted repeats and the formation of a stem–loop.** *(a)* Nearby inverted repeats in DNA. The arrows indicate the symmetry around the imaginary axis (dashed line). *(b)* Formation of stem–loop structures by pairing of complementary bases on the same strand.

long. Since cells of *E. coli* are about 2 μm long, the chromosome is several hundred times longer than the cell itself!

Long DNA molecules are quite flexible, but stretches of DNA less than 100 base pairs are more rigid. Some short segments of DNA can be bent by proteins that bind them. However, certain base sequences themselves cause DNA to bend. Such sequences usually have several runs of five or six adenines, each separated by four or five other bases.

Inverted Repeats and Stem–Loop Structures

Short, repeated sequences are often found in DNA molecules. Many proteins bind to regions of DNA containing inverted repeat sequences (Chapter 8). As shown in **Figure 6.6**, nearby inverted repeats can form stem–loop structures. The stems are short double-helical regions with normal base pairing. The loop contains the unpaired bases between the two repeats.

The formation of stem–loop structures in DNA itself is relatively rare. However, the production of stem–loop structures in the RNA produced from DNA following transcription is common. Such secondary structures formed by base pairing within a single strand of RNA are found in transfer RNA (Section 6.18) and ribosomal RNA (Section 6.19). Even when a stem–loop does not form, inverted repeats in DNA are often binding sites for DNA-binding proteins that regulate transcription (Chapter 8) or for endonucleases that cut DNA (↻ Section 11.1).

The Effect of Temperature on DNA Structure

Although individual hydrogen bonds are very weak, the large number of such bonds between the base pairs of a long DNA molecule hold the two strands together effectively. There may be millions or even hundreds of millions of hydrogen bonds in a long DNA molecule, depending on the number of base pairs. Recall that each adenine–thymine base pair has *two* hydrogen bonds, while each guanine–cytosine base pair has *three*. This makes GC pairs stronger than AT pairs.

When isolated from cells and kept near room temperature and at physiological salt concentrations, DNA remains double-stranded. However, if the temperature is raised, the hydrogen bonds will break but the covalent bonds holding a chain together will not, and so the two DNA strands will separate. This process is called denaturation (melting) and can be measured experimentally because single-stranded and double-stranded nucleic acids differ in their ability to absorb ultraviolet radiation at 260 nm (**Figure 6.7**).

DNA with a high percentage of GC pairs melts at a higher temperature than a similar-sized molecule with more AT pairs. If the heated DNA is allowed to cool slowly, the double-stranded DNA can re-form, a process called annealing. This can be used not only to re-form native DNA but also to form hybrid molecules whose two strands come from different sources. Hybridization, the artificial assembly of a double-stranded nucleic acid by complementary base pairing of two single strands, is a powerful technique in molecular biology (↻ Section 11.2).

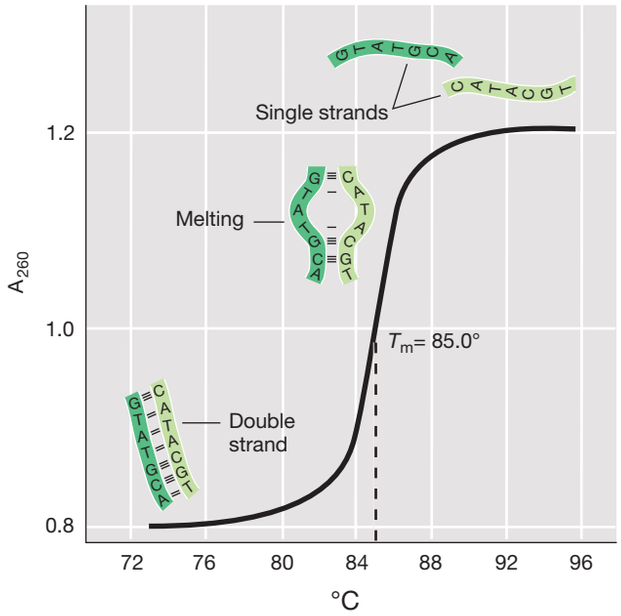


Figure 6.7 **Thermal denaturation of DNA.** DNA absorbs more ultraviolet radiation at 260 nm as the double helix is denatured. The transition is quite abrupt, and the temperature of the midpoint, T_m, is proportional to the GC content of the DNA. Although the denatured DNA can be renatured by slow cooling, the process does not follow a similar curve. Renaturation becomes progressively more complete at temperatures well below the T_m and then only after a considerable incubation time.

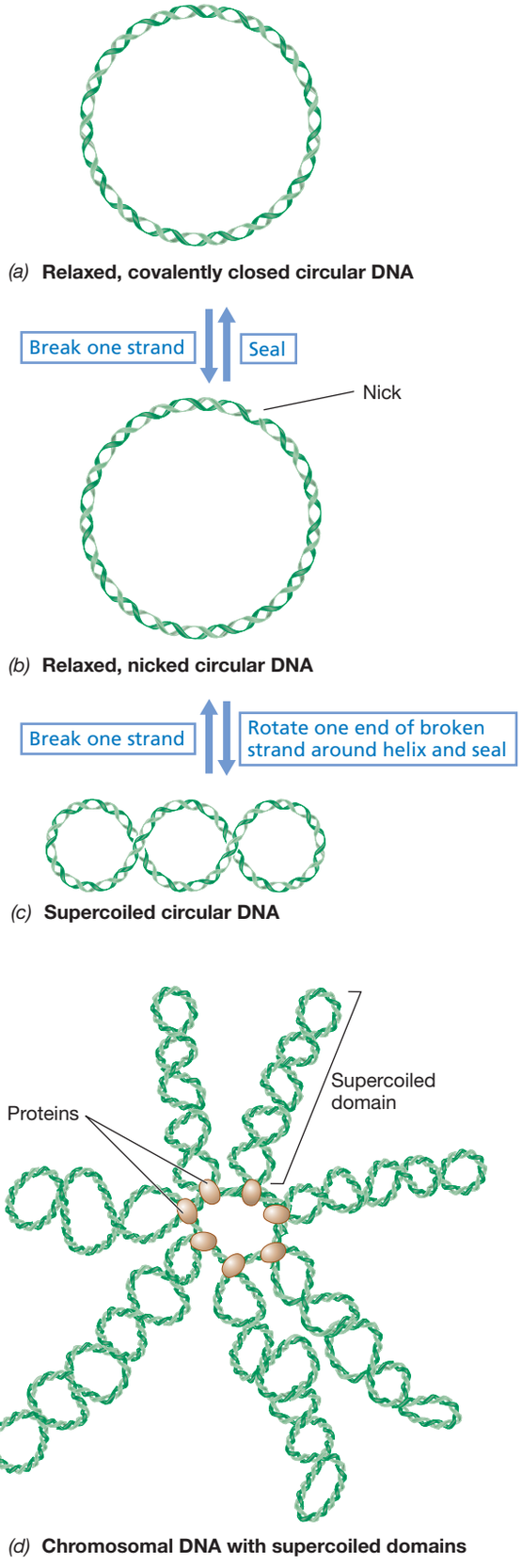
MiniQuiz

- What does antiparallel mean in terms of the structure of double-stranded DNA?
- Define the term complementary when used to refer to two strands of DNA.
- Define the terms denaturation, reannealing, and hybridization as they apply to nucleic acids.
- Why do GC-rich molecules of DNA melt at higher temperatures than AT-rich molecules?

6.3 Supercoiling

If linearized, the *Escherichia coli* chromosome would be over 1 mm in length, about 700 times longer than the *E. coli* cell itself. How is it possible to pack so much DNA into such a little space? The solution is the imposition of a "higher-order" structure on the DNA, in which the double-stranded DNA is further twisted in a process called *supercoiling*. **Figure 6.8** shows how supercoiling occurs in a circular DNA duplex. If a circular DNA molecule is linearized, any supercoiling is lost and the DNA becomes "relaxed." When relaxed, a DNA molecule has exactly the number of turns of the helix predicted from the number of base pairs.

Supercoiling puts the DNA molecule under torsion, much like the added tension to a rubber band that occurs when it is twisted. DNA can be supercoiled in either a positive or a negative manner. In positive supercoiling the double helix is overwound, whereas in negative supercoiling the double helix is underwound. Negative supercoiling results when the DNA is twisted about its



(a) **Relaxed, covalently closed circular DNA**

(b) **Relaxed, nicked circular DNA**

(c) **Supercoiled circular DNA**

(d) **Chromosomal DNA with supercoiled domains**

Figure 6.8 **Supercoiled DNA.** *(a–c)* Relaxed, nicked, and supercoiled circular DNA. A nick is a break in a phosphodiester bond of one strand. *(d)* In fact, the double-stranded DNA in the bacterial chromosome is arranged not in one supercoil but in several supercoiled domains, as shown here.

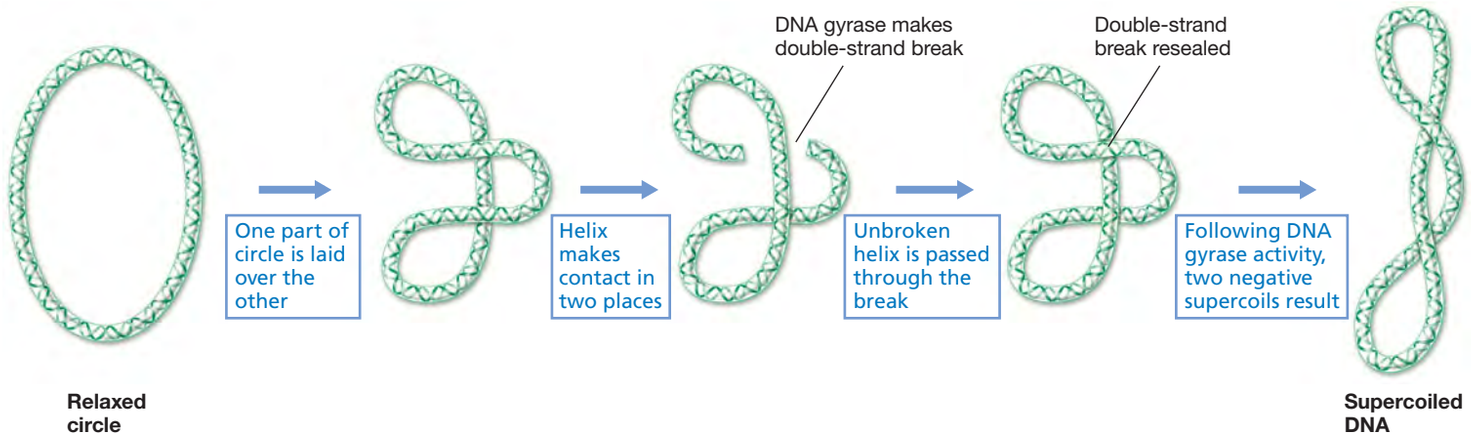


Figure 6.9 DNA gyrase. Introduction of negative supercoiling into circular DNA by the activity of DNA gyrase (topoisomerase II), which makes double-strand breaks.

axis in the opposite sense from the right-handed double helix. Negatively supercoiled DNA is the form predominantly found in nature. However, certain species of *Archaea* (Chapter 7) that grow at very high temperatures do contain positively supercoiled DNA. In *Escherichia coli* more than 100 supercoiled domains are thought to exist, each of which is stabilized by binding to specific proteins.

Topoisomerases: DNA Gyrase

Supercoils are inserted or removed by enzymes known as topoisomerases. Two major classes of topoisomerase exist with different mechanisms. Class I topoisomerases make a single-stranded break in the DNA that allows the rotation of one strand of the double helix around the other. Each rotation adds or removes a single supercoil. After this, the nick is resealed. For example, surplus supercoiling in DNA is generally removed by the class I enzyme, topoisomerase I. As shown in Figure 6.8, a break in the backbone (a nick) of either strand allows DNA to lose its supercoiling. However, to prevent the entire bacterial chromosome from becoming relaxed every time a nick is made, the chromosome contains supercoiled domains as shown in Figure 6.8*d*. A nick in the DNA of one domain does not relax DNA in the others. It is unclear precisely how these domains are formed, although specific DNA-binding proteins are involved.

Class II topoisomerases make double-stranded breaks, pass the double helix through the break, and reseal the break (**Figure 6.9**). Each such operation adds or removes two supercoils. Inserting supercoils into DNA requires energy from ATP, whereas releasing supercoils does not. In *Bacteria* and most *Archaea*, the class II topoisomerase, **DNA gyrase**, inserts negative supercoils into DNA. Some antibiotics inhibit the activity of DNA gyrase. These include the quinolones (such as nalidixic acid), the fluoroquinolones (such as ciprofloxacin), and novobiocin.

Through the activity of topoisomerases, a DNA molecule can be alternately supercoiled and relaxed. Supercoiling is necessary for packing the DNA into the cell and relaxation is necessary for DNA replication and transcription. In most prokaryotes, the level of negative supercoiling results from the balance between the activity of DNA gyrase and topoisomerase I. Supercoiling also affects gene expression. Certain genes are more actively transcribed when DNA is supercoiled, whereas transcription of other genes is inhibited by supercoiling.

MiniQuiz

- Why is supercoiling important?
- What mechanism is used by DNA gyrase?
- What function do topoisomerases serve inside cells?

6.4 Chromosomes and Other Genetic Elements

Structures containing genetic material (DNA in most organisms, but RNA in some viruses) are called *genetic elements*. The **genome** is the total complement of genes in a cell or virus. Although the main genetic element in prokaryotes is the **chromosome**, other genetic elements are found and play important roles in gene function in both prokaryotes and eukaryotes (**Table 6.1**). These include virus genomes, plasmids, organellar genomes, and transposable elements. A typical prokaryote has a single circular chromosome containing all (or most) of the genes found inside the cell. Although a single chromosome is the rule among prokaryotes, there are exceptions. A few prokaryotes contain two chromosomes. Eukaryotes have multiple chromosomes making up their genome (↻ Section 7.5). Also, the DNA in all known eukaryotic chromosomes is linear in contrast to most prokaryotic chromosomes, which are circular DNA molecules.

Viruses and Plasmids

Viruses contain genomes, *either* of DNA or RNA, that control their own replication and their transfer from cell to cell. Both linear and circular viral genomes are known. In addition, the nucleic acid in viral genomes may be single-stranded or double-stranded. Viruses are of special interest because they often cause disease. We discuss viruses in Chapters 9 and 21 and a variety of viral diseases in later chapters.

Table 6.1 *Kinds of genetic elements*

Organism	*Element*	*Type of nucleic acid*	*Description*
Prokaryote	Chromosome	Double-stranded DNA	Extremely long, usually circular
Eukaryote	Chromosome	Double-stranded DNA	Extremely long, linear
All organisms	Plasmid[a]	Double-stranded DNA	Relatively short circular or linear, extrachromosomal
All organisms	Transposable element	Double-stranded DNA	Always found inserted into another DNA molecule
Mitochondrion or chloroplast	Genome	Double-stranded DNA	Medium length, usually circular
Virus	Genome	Single- or double-stranded DNA or RNA	Relatively short, circular or linear

[a]Plasmids are uncommon in eukaryotes.

Plasmids are genetic elements that replicate separately from the chromosome. The great majority of plasmids are double-stranded DNA, and although most plasmids are circular, some are linear. Most plasmids are much smaller than chromosomes. Plasmids differ from viruses in two ways: (1) They do not cause cellular damage (generally they are beneficial), and (2) they do not have extracellular forms, whereas viruses do. Although only a few eukaryotes contain plasmids, one or more plasmids have been found in most prokaryotic species and can be of profound importance. Some plasmids contain genes whose protein products confer important properties on the host cell, such as resistance to antibiotics.

What is the difference, then, between a large plasmid and a chromosome? A chromosome is a genetic element that contains genes whose products are necessary for essential cellular functions. Such essential genes are called *housekeeping genes*. Some of these encode essential proteins, such as DNA and RNA polymerases, and others encode essential RNAs, such as ribosomal and transfer RNA. In contrast to the chromosome, plasmids are usually expendable and rarely contain genes required for growth under all conditions. There are many genes on a chromosome that are unessential as well, but the presence of *essential* genes is necessary for a genetic element to be classified as a chromosome.

Transposable Elements

Transposable elements are segments of DNA that can move from one site on a DNA molecule to another site, either on the same molecule or on a different DNA molecule. Transposable elements are not found as separate molecules of DNA but are inserted into other DNA molecules. Chromosomes, plasmids, virus genomes, and any other type of DNA molecule may act as host molecules for transposable elements. Transposable elements are found in both prokaryotes and eukaryotes and play important roles in genetic variation. In prokaryotes there are three main types of transposable elements: insertion sequences, transposons, and some special viruses. Insertion sequences are the simplest type of transposable element and carry no genetic information other than that required for them to move about the chromosome. Transposons are larger and contain other genes. We discuss both of these in more detail in Chapter 10. In Chapter 21 we discuss a bacterial virus, Mu, that is itself a transposable element. The unique feature common to all transposable elements is that they replicate as part of some other molecule of DNA.

MiniQuiz

- What is a genome?
- What are viruses and plasmids?
- What genetic material is found in all cellular chromosomes?
- What defines a chromosome in prokaryotes?

II Chromosomes and Plasmids

6.5 The *Escherichia coli* Chromosome

Today, many bacterial genomes, including that of *Escherichia coli*, have been completely sequenced, thus revealing the number and location of the genes they possess. However, the genes of *E. coli* were initially mapped long before sequencing was performed, using conjugation and transduction (↻ Sections 10.8 and 10.9). The genetic map of *E. coli* strain K-12 is shown in **Figure 6.10**. Map distances are given in "minutes" of transfer that derive from conjugation experiments, with the entire chromosome containing 100 minutes (or centisomes). Zero is arbitrarily set at *thrABC* (the threonine operon), because the *thrABC* genes were the first shown to be transferred by conjugation in *E. coli*. The genetic map in Figure 6.10 shows only a few of the several thousand genes in the *E. coli* chromosome. The size of the chromosome is given in both minutes and in kilobase pairs of DNA.

The strain of *E. coli* whose chromosome was originally sequenced, strain MG1655, is a derivative of *E. coli* K-12, the traditional strain used for genetics. Wild-type *E. coli* K-12 has bacteriophage lambda integrated into its chromosome (↻ Section 9.10) and also contains the F plasmid. However, strain MG1655 had both of these removed before sequencing (lambda by radiation and the F plasmid by acridine treatment). The chromosome of strain MG1655 contains 4,639,221 bp. Analysis revealed 4288 possible protein-encoding genes that account for about 88% of the genome. Approximately 1% of the genome consists of genes encoding tRNAs and rRNAs. Regulatory sequences—promoters, operators, origin and terminus of DNA replication, and so on—comprise around 10% of the genome. The remaining 0.5% consists of noncoding, repetitive sequences.

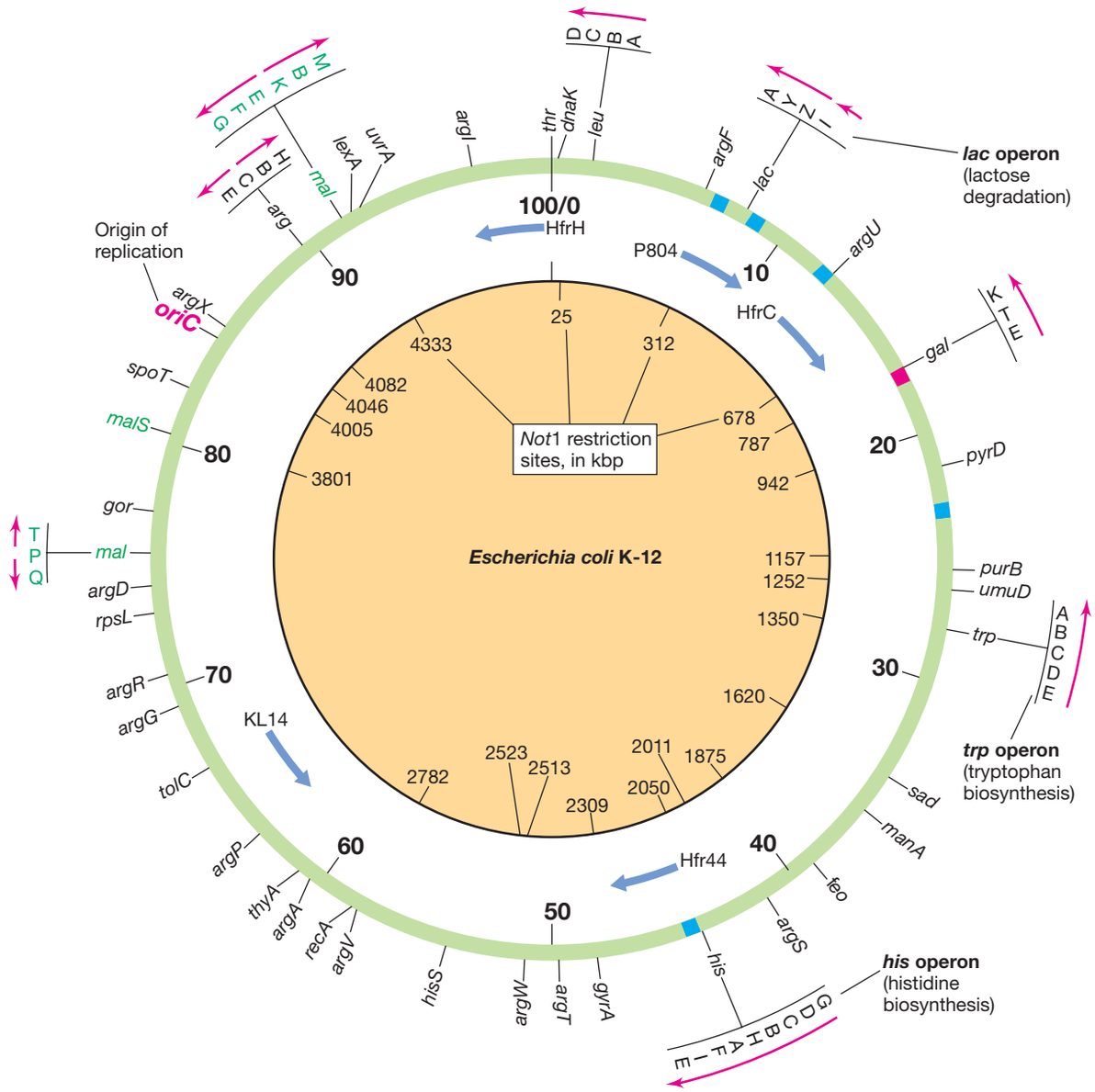


Figure 6.10 **The chromosome of *Escherichia coli* strain K-12.** The *E. coli* chromosome contains 4,639,221 base pairs and 4288 open reading frames (an indicator of genes; Section 6.17). On the outer edge of the map, the locations of a few genes are indicated. A few operons are also shown, with their directions of transcription. Around the inner edge, the numbers from 0 to 100 refer to map position in minutes. Note that 0 is located by convention at the *thr* locus. Replication proceeds bidirectionally from the origin of DNA replication, *oriC*, at 84.3 min. The inner circle shows the locations, in kilobase pairs, of the sites where the restriction enzyme *Not*I cuts. The origins and directions of transfer of a few Hfr strains are also shown (arrows). The locations of five copies of the transposable element IS3 found in a particular strain are shown in blue. The site where bacteriophage lambda integrates is shown in red. If lambda were present, it would add an extra 48.5 kbp (slightly over 1 min) to the map. The genes of the maltose regulon, which includes several operons, are indicated by green labels. The maltose genes are abbreviated *mal* except for *lamB*, which encodes an outer membrane protein for maltose uptake that is also the receptor for bacteriophage lambda. The gene *rpsL* (73 min) encodes a ribosomal protein. This gene was once called *str* because mutations in it lead to streptomycin resistance.

Arrangement of Genes on the *Escherichia coli* Chromosome

Genetic mapping of the genes that encode the enzymes of a single biochemical pathway in *E. coli* has shown that these genes are often clustered. On the genetic map in Figure 6.10, a few such clusters are shown. Notice, for instance, the *gal* gene cluster at 18 min, the *trp* gene cluster at about 28 min, and the *his* cluster at 44 min. Each of these gene clusters constitutes an *operon* that is transcribed as a single mRNA carrying multiple coding sequences, that is, a *polycistronic* mRNA (Section 6.15).

Genes for some other biochemical pathways in *E. coli* are not clustered. For example, genes for arginine biosynthesis (*arg* genes) are scattered throughout the chromosome. The early discovery of multigene operons and their use in studying gene

regulation (for example, the *lac* operon; ↻ Section 8.5), often gives the impression that such operons are the rule in prokaryotes. However, sequence analysis of the *E. coli* chromosome has shown that over 70% of the 2584 predicted or known transcriptional units contain only a single gene. Only about 6% of the operons have four or more genes.

In *E. coli* the transcription of some genes proceeds clockwise around the chromosome, whereas transcription of others proceeds counterclockwise. This means that some coding sequences are on one strand of the chromosome whereas others are on the opposite strand. There are about equal numbers of genes on both strands. The direction of transcription of a few multigene operons is shown by the arrows in Figure 6.10. Many genes that are highly expressed in *E. coli* are oriented so that they are transcribed in the same direction that the DNA replication fork moves through them. The two replication forks start at the origin, *oriC* located at about 84 min, and move in opposite directions around the circular chromosome toward the terminus, which is located at approximately 34 min. All seven of the rRNA operons of *E. coli* and 53 of its 86 tRNA genes are transcribed in the same direction as replication. Presumably, this arrangement for highly expressed genes allows RNA polymerase to avoid collision with the replication fork, because this moves in the same direction as the RNA polymerase.

Almost 2000 *E. coli* proteins, or genes encoding proteins, were identified by classical genetic analyses before its chromosome was sequenced. Sequence analyses indicate that approximately 4225 different proteins may be encoded by the *E. coli* chromosome. Around 30% of these proteins are of unknown function or are hypothetical. The average *E. coli* protein contains slightly more than 300 amino acid residues, but many proteins are smaller and many are much larger. The largest gene in *E. coli* encodes a protein of 2383 amino acids that is still uncharacterized. This giant protein shows similarities to proteins found in pathogenic enteric bacteria closely related to *E. coli* and may thus play some role in infection.

Although sequence analysis yields much information, to understand the function, particularly of regulatory sequences, it is still necessary to isolate mutants, map the mutations, and use biochemical and physiological analyses to determine their effects on the organism. This is especially true of the 20–40% of genes that show up in all genomic analyses (↻ Section 12.3) as encoding proteins of unknown function. This huge repository of hypothetical proteins doubtless holds new biochemical secrets that will expand the known metabolic capabilities of prokaryotes. In addition, because many prokaryotic genes have homologs in eukaryotes including humans, understanding gene function in prokaryotes aids our understanding of human genetics.

Although *E. coli* has very few duplicate genes, computer analyses have shown that many of its protein-encoding genes arose by gene duplication during evolutionary history (↻ Section 12.10). The *E. coli* genome also contains some large gene families—groups of genes with related sequences encoding products with related functions. For example, there is a family of 70 genes that all encode membrane transport proteins. Gene families are common, both within a species and across broad taxonomic lines. Thus gene duplication plays a major role in evolution.

Insertions within the *Escherichia coli* Chromosome and Horizontal Gene Transfer

Several other genetic elements are inserted into the *E. coli* chromosome and are consequently replicated with it. There are multiple copies of several different insertion sequences (IS elements), including seven copies of IS*2* and five of IS*3*. Both of these IS elements are also found on the F plasmid, and both take part in the formation of Hfr strains (↻ Section 10.10). There are several defective integrated viruses that vary from nearly complete virus genomes to small fragments. Three of these are related to bacteriophage lambda.

E. coli obtained part of its genome by *horizontal* (*lateral*) gene transfer from other organisms. Horizontal transfer contrasts with *vertical* gene transfer in which genes move from mother cell to daughter cell. In fact, it has been estimated that nearly 20% of the *E. coli* genome originated from horizontal transfers. Horizontally transferred segments of DNA can often be detected because they have significantly different GC ratios (the ratio of guanine–cytosine base pairs to adenine–thymine base pairs) or codon distributions (codon bias, Section 6.17) from those of the host organism.

Horizontal gene transfer may cause large-scale changes in a genome. For example, strains of *E. coli* are known that contain virulence genes located on large, unstable regions of the chromosome called *pathogenicity islands* that can be acquired by horizontal transfer (↻ Sections 12.12 and 12.13). Horizontal transfer does not necessarily result in an ever-larger genome size. Many genes acquired in this way provide no selective advantage and so are lost by deletion. This keeps the chromosome of a given species at roughly the same size over time. For example, comparisons of genome sizes of several strains of *E. coli* have shown them all to be about 4.5–5.5 Mbp, despite the fact that prokaryotic genomes can vary from under 0.5 to over 10 Mbp. Genome size is therefore a species-specific trait.

MiniQuiz

- Genetic maps of bacterial chromosomes are now typically made using only molecular cloning and DNA sequencing. Why were other methods also used for *E. coli*?
- How large is an average bacterial protein?
- Approximately how large is the *E. coli* genome in base pairs? How many genes does it contain?

6.6 Plasmids: General Principles

Many prokaryotic cells contain other genetic elements, in particular, **plasmids**, in addition to the chromosome. Plasmids are genetic elements that replicate independently of the host chromosome, in the sense of possessing their own origin of replication. However, they do rely on chromosomally encoded enzymes for their replication. Unlike viruses, plasmids do not have an extracellular form and exist inside cells as free, typically circular, DNA. Plasmids differ from chromosomes in carrying only nonessential (but often very helpful) genes. Essential genes reside on chromosomes. Thousands of different plasmids are known. Indeed, over 300 different naturally occurring plasmids have

been isolated from strains of *Escherichia coli* alone. In this section we discuss their basic properties.

Plasmids have been widely exploited in genetic engineering. Countless new, artificial plasmids have been constructed in the laboratory. Genes from a wide variety of sources have been incorporated into such plasmids, thus allowing their transfer across any species barrier. The only requirements for artificial plasmids are that they carry genes controlling their own replication and are stably maintained in the host of choice. This topic is discussed further in Chapter 11.

Physical Nature and Replication of Plasmids

Almost all known plasmids consist of double-stranded DNA. Most are circular, but many linear plasmids are also known. Naturally occurring plasmids vary in size from approximately 1 kbp to more than 1 Mbp. Typical plasmids are circular double-stranded DNA molecules less than 5% the size of the chromosome (**Figure 6.11**). Most plasmid DNA isolated from cells is supercoiled, this being the most compact form that DNA takes within the cell (Figure 6.8). Some bacteria may contain several different types of plasmids. For example, *Borrelia burgdorferi* (the Lyme disease pathogen, Section 34.4) contains 17 different circular and linear plasmids!

The enzymes that replicate plasmids are normal cell enzymes. The genes carried by the plasmid itself are concerned primarily with controlling the initiation of replication and with partitioning replicated plasmids between daughter cells. Different plasmids are present in cells in different numbers; this is called the *copy number*. Some plasmids are present in the cell in only 1–3 copies, whereas others may be present in over 100 copies. Copy number is controlled by genes on the plasmid and by interactions between the host and the plasmid.

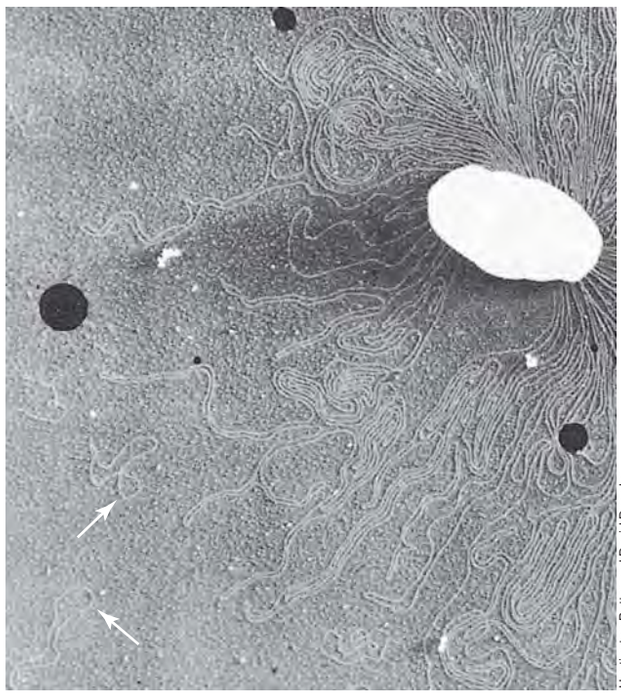


Figure 6.11 The bacterial chromosome and bacterial plasmids, as seen in the electron microscope. The plasmids (arrows) are the circular structures and are much smaller than the main chromosomal DNA. The cell (large, white structure) was broken gently so the DNA would remain intact.

Most plasmids in gram-negative *Bacteria* replicate in a manner similar to that of the chromosome. This involves initiation at an origin of replication and bidirectional replication around the circle, giving a theta intermediate (Section 6.10). However, some small plasmids have unidirectional replication, with just a single replication fork. Because of the small size of plasmid DNA relative to the chromosome, plasmids replicate very quickly, perhaps in a tenth or less of the total time of the cell division cycle.

Most plasmids of gram-positive *Bacteria*, plus a few from gram-negative *Bacteria* and *Archaea*, replicate by a rolling circle mechanism similar to that used by bacteriophage ϕX174 (Section 21.2). This mechanism proceeds via a single-stranded intermediate. Most linear plasmids replicate by using a protein bound to the 5′ end of each strand to prime DNA synthesis (Section 7.7).

Plasmid Incompatibility and Plasmid Curing

Many bacterial cells contain multiple different plasmids. However, when two different plasmids are closely related genetically, they cannot both be maintained in the same cell. The two plasmids are then said to be *incompatible*. When a plasmid is transferred into a cell that already carries another related and incompatible plasmid, one or the other will be lost during subsequent cell replication. A number of incompatibility (Inc) groups exist. Plasmids belonging to the same Inc group exclude each other but can coexist with plasmids from other groups. Plasmids within each Inc group are related in sharing a common mechanism of regulating their replication. Therefore, although a bacterial cell may contain different kinds of plasmids, each is genetically distinct.

Some plasmids, called *episomes*, can integrate into the chromosome. Under such conditions their replication comes under control of the chromosome. This situation is analogous to that of several viruses whose genomes can integrate into the host genome (Section 9.10).

Plasmids can sometimes be eliminated from host cells by various treatments. This removal, called *curing*, results from inhibition of plasmid replication without parallel inhibition of chromosome replication. As a result, the plasmid is diluted out during cell division. Curing may occur spontaneously, but is greatly increased by treatments with certain chemicals such as acridine dyes, which insert into DNA, or other treatments that interfere more with plasmid replication than with chromosome replication.

Cell-to-Cell Transfer of Plasmids

How do plasmids manage to infect new host cells? Some prokaryotic cells can take up free DNA from the environment (Section 10.7). Consequently, plasmids released by the death and disintegration of their previous host cell may be taken up by a new host. However, few bacterial species have this ability, and it is unlikely to account for much plasmid transfer. The main

mechanism of plasmid transfer is *conjugation*, a function encoded by some plasmids themselves that involves cell-to-cell contact (↻ Section 10.9).

Plasmids capable of transferring themselves by cell-to-cell contact are called *conjugative*. Not all plasmids are conjugative. Transfer by conjugation is controlled by a set of genes on the plasmid called the *tra* (for transfer) region. These genes encode proteins that function in DNA transfer and replication and others that function in mating pair formation. If a plasmid possessing a *tra* region becomes integrated into the chromosome, the plasmid can then mobilize the chromosomal DNA, which may be transferred from one cell to another (↻ Section 10.10).

Most conjugative plasmids can only move between closely related species of bacteria. However, some conjugative plasmids from *Pseudomonas* have a broad host range. This means that they are transferable to a wide variety of other gram-negative *Bacteria*. Such plasmids can transfer genes between distantly related organisms. Conjugative plasmids have been shown to transfer between gram-negative and gram-positive *Bacteria*, between *Bacteria* and plant cells, and between *Bacteria* and fungi. Even if the plasmid cannot replicate independently in the new host, transfer of the plasmid itself could have important evolutionary consequences if genes from the plasmid recombine with the genome of the new host.

MiniQuiz

- How does a plasmid differ from a virus?
- How can a large plasmid be differentiated from a small chromosome?
- What function do the *tra* genes of the F plasmid carry out?

6.7 The Biology of Plasmids

Clearly, all plasmids must carry genes that ensure their own replication. In addition, some plasmids also carry genes necessary for conjugation. Although plasmids do not carry genes that are essential to the host, plasmids may carry genes that profoundly influence host cell physiology. In some cases plasmids encode properties fundamental to the ecology of the bacterium. For example, the ability of *Rhizobium* to interact with plants and form nitrogen-fixing root nodules requires certain plasmid functions (↻ Section 25.8). Other plasmids confer special metabolic properties on bacterial cells, such as the ability to degrade toxic pollutants. Indeed, plasmids are a major mechanism for conferring special properties on bacteria and for mobilizing these properties by horizontal gene flow. Some special properties conferred by plasmids are summarized in **Table 6.2**.

Table 6.2 *Examples of phenotypes conferred by plasmids in prokaryotes*

Phenotype class	*Organisms*[a]
Antibiotic production	*Streptomyces*
Conjugation	Wide range of bacteria
Metabolic functions	
Degradation of octane, camphor, naphthalene	*Pseudomonas*
Degradation of herbicides	*Alcaligenes*
Formation of acetone and butanol	*Clostridium*
Lactose, sucrose, citrate, or urea utilization	Enteric bacteria
Pigment production	*Erwinia, Staphylococcus*
Gas vesicle production	*Halobacterium*
Resistance	
Antibiotic resistance	Wide range of bacteria
Resistance to toxic metals	Wide range of bacteria
Virulence	
Tumor production in plants	*Agrobacterium*
Nodulation and symbiotic nitrogen fixation	*Rhizobium*
Bacteriocin production and resistance	Wide range of bacteria
Animal cell invasion	*Salmonella, Shigella, Yersinia*
Coagulase, hemolysin, enterotoxin	*Staphylococcus*
Toxins and capsule	*Bacillus anthracis*
Enterotoxin, K antigen	*Escherichia coli*

Resistance Plasmids

Among the most widespread and well-studied groups of plasmids are the resistance plasmids, usually just called *R plasmids*, which confer resistance to antibiotics and various other growth inhibitors. Several antibiotic resistance genes can be carried by a single R plasmid, or, alternatively, a cell may contain several R plasmids. In either case, the result is multiple resistance. R plasmids were first discovered in Japan in the 1950s in strains of enteric bacteria that had acquired resistance to sulfonamide antibiotics. Since then they have been found throughout the world. The emergence of bacteria resistant to antibiotics is of considerable medical significance and is correlated with the increasing use of antibiotics for treating infectious diseases (↻ Section 26.12). Soon after these resistant strains were isolated, it was shown that they could transfer resistance to sensitive strains via cell-to-cell contact. The infectious nature of conjugative R plasmids permitted their rapid spread through cell populations.

In general, resistance genes encode proteins that either inactivate the antibiotic or protect the cell by some other mechanism. Plasmid R100, for example, is a 94.3-kbp plasmid (**Figure 6.12**) that carries genes encoding resistance to sulfonamides, streptomycin, spectinomycin, fusidic acid, chloramphenicol, and tetracycline. Plasmid R100 also carries several genes conferring resistance to mercury. Plasmid R100 can be transferred between enteric bacteria of the genera *Escherichia, Klebsiella, Proteus, Salmonella*, and *Shigella*, but does not transfer to gram-negative bacteria outside the enteric group. Different R plasmids with genes for resistance to most antibiotics are known. Many drug-resistant modules on R plasmids, such as those on R100, are also transposable elements (↻ Section 12.11), and this, combined with the fact that many of these plasmids are conjugative, have made them a serious threat to traditional antibiotic therapy.

Plasmids Encoding Virulence Characteristics

Pathogenic microorganisms possess a variety of characteristics that enable them to colonize hosts and establish infections. Here we note two major characteristics of the virulence (disease-causing

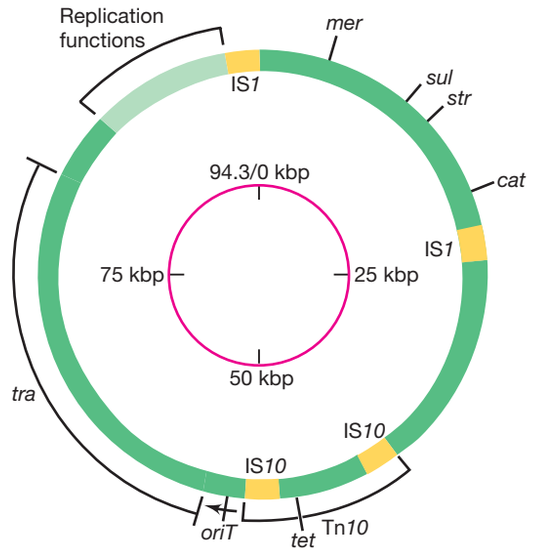


Figure 6.12 Genetic map of the resistance plasmid R100. The inner circle shows the size in kilobase pairs. The outer circle shows the location of major antibiotic resistance genes and other key functions: *mer*, mercuric ion resistance; *sul*, sulfonamide resistance; *str*, streptomycin resistance; *cat*, chloramphenicol resistance; *tet*, tetracycline resistance; *oriT*, origin of conjugative transfer; *tra*, transfer functions. The locations of insertion sequences (IS) and the transposon Tn*10* are also shown. Genes for plasmid replication are found in the region from 88 to 92 kbp.

ability) of pathogens that are often plasmid encoded: (1) the ability of the pathogen to attach to and colonize specific host tissue and (2) the production of toxins, enzymes, and other molecules that cause damage to the host.

Enteropathogenic strains of *Escherichia coli* are characterized by the ability to colonize the small intestine and to produce a toxin that causes diarrhea. Colonization requires a cell surface protein called *colonization factor antigen*, encoded by a plasmid. This protein confers on bacterial cells the ability to attach to epithelial cells of the intestine. At least two toxins in enteropathogenic *E. coli* are encoded by plasmids: the hemolysin, which lyses red blood cells, and the enterotoxin, which induces extensive secretion of water and salts into the bowel. It is the enterotoxin that is responsible for diarrhea (↻ Section 27.11).

Some virulence factors are encoded on plasmids. Other virulence factors are encoded by other mobile genetic elements, such as transposons and bacteriophages. Some virulence factors are chromosomal. Several examples are known in which multiple virulence genes are present on different genetic elements within the same cell. For instance, the genes encoding the virulence determinants of Shiga toxin–producing strains of *E. coli* (↻ Section 36.9) are distributed among the chromosome, a bacteriophage, and a plasmid.

Bacteriocins

Many bacteria produce proteins that inhibit or kill closely related species or even different strains of the same species. These agents are called **bacteriocins** to distinguish them from antibiotics. Bacteriocins have a narrower spectrum of activity than antibiotics. The genes encoding bacteriocins and the proteins needed for processing and transporting them and for conferring immunity on the producing organism are usually carried on plasmids or transposons. Bacteriocins are often named after the species of organism that produces them. Thus, *E. coli* produces *colicins*; *Yersinia pestis* produces *pesticins*, and so on.

The Col plasmids of *E. coli* encode various colicins. Col plasmids can be either conjugative or nonconjugative. Colicins released from the producer cell bind to specific receptors on the surface of susceptible cells. The receptors for colicins are typically proteins whose normal function is to transport growth factors or micronutrients across the outer membrane of the cell. Colicins kill cells by disrupting some critical cell function. Many colicins form channels in the cell membrane that allow potassium ions and protons to leak out, leading to loss of the ability to generate energy. Another major group of colicins are nucleases and degrade DNA or RNA. For example, colicin E2 is a DNA endonuclease that cleaves DNA, and colicin E3 is a ribonuclease that cuts at a specific site in 16S rRNA and therefore inactivates ribosomes.

The bacteriocins or bacteriocin-like agents of gram-positive bacteria are quite different from the colicins but are also often encoded by plasmids; some even have commercial value. For instance, lactic acid bacteria produce the bacteriocin nisin A, which strongly inhibits the growth of a wide range of gram-positive bacteria and is used as a preservative in the food industry.

MiniQuiz

- What properties does an R plasmid confer on its host cell?
- What properties does a virulence plasmid typically confer on its host cell?
- How do bacteriocins differ from antibiotics?

III DNA Replication

DNA replication is necessary for cells to divide, whether to reproduce new organisms, as in unicellular microorganisms, or to produce new cells as part of a multicellular organism. DNA replication must be sufficiently accurate that the daughter cells are genetically identical to the mother cell (or almost so). This involves a host of special enzymes and processes.

6.8 Templates and Enzymes

DNA exists in cells as a double helix with complementary base pairing. If the double helix is opened up, a new strand can be synthesized as the complement of each parental strand. As shown in **Figure 6.13**, replication is **semiconservative**, meaning that the two resulting double helices consist of one new strand and one parental strand. The DNA strand that is used to make a complementary daughter strand is called the template, and in DNA replication each parental strand is a template for one newly synthesized strand (Figure 6.13).

The precursor of each new nucleotide in the DNA strand is a deoxynucleoside 5′-triphosphate. The two terminal phosphates are removed and the innermost phosphate is then attached

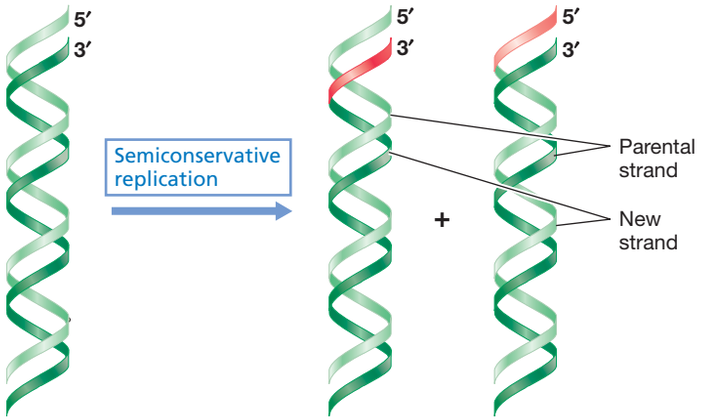


Figure 6.13 **Overview of DNA replication.** DNA replication is a semiconservative process in all cells. Note that the new double helices each contain one new strand (shown topped in red) and one parental strand.

covalently to a deoxyribose of the growing chain (**Figure 6.14**). This addition of the incoming nucleotide requires the presence of a free hydroxyl group, which is available only at the 3′ end of the molecule. This leads to the important principle that

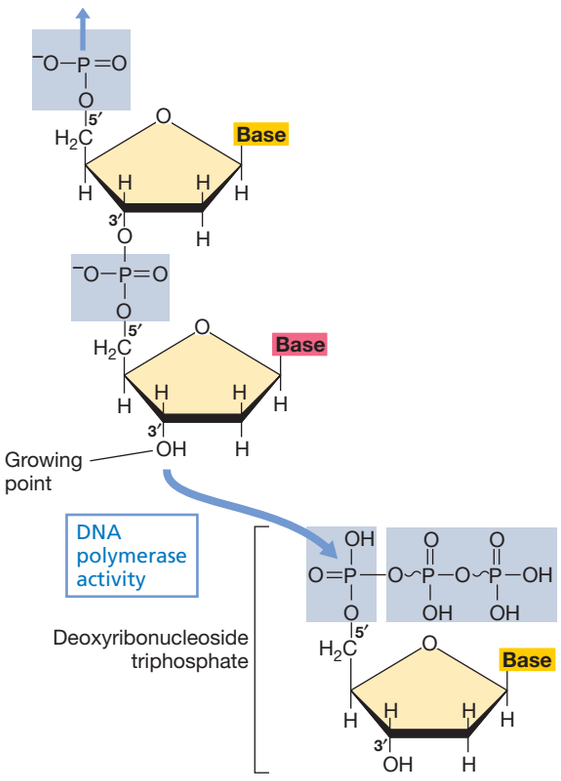


Figure 6.14 **Extension of a DNA chain by adding a deoxyribonucleoside triphosphate at the 3′ end.** Growth proceeds from the 5′-phosphate to the 3′-hydroxyl end. DNA polymerase catalyzes the reaction. The four precursors are deoxythymidine triphosphate (dTTP), deoxyadenosine triphosphate (dATP), deoxyguanosine triphosphate (dGTP), and deoxycytidine triphosphate (dCTP). Upon nucleotide insertion, the two terminal phosphates of the triphosphate are split off as pyrophosphate (PP_i). Thus, two energy-rich phosphate bonds are consumed when adding each nucleotide.

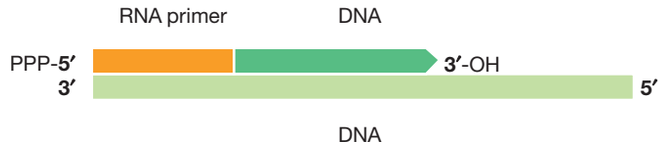


Figure 6.15 **The RNA primer.** Structure of the RNA–DNA hybrid formed during initiation of DNA synthesis.

DNA replication always proceeds from the 5′ end to the 3′ end, the 5′-phosphate of the incoming nucleotide being attached to the 3′-hydroxyl of the previously added nucleotide.

Enzymes that catalyze the addition of deoxynucleotides are called **DNA polymerases**. Several such enzymes exist, each with a specific function. There are five different DNA polymerases in *Escherichia coli*, called DNA polymerases I, II, III, IV, and V. DNA polymerase III (Pol III) is the primary enzyme for replicating chromosomal DNA. DNA polymerase I (Pol I) is also involved in chromosomal replication, though to a lesser extent (see below). The other DNA polymerases help repair damaged DNA (↻ Section 10.4).

All known DNA polymerases synthesize DNA in the 5′ → 3′ direction. However, no known DNA polymerase can initiate a new chain; all of these enzymes can only add a nucleotide onto a preexisting 3′-OH group. To start a new chain, a **primer**, a nucleic acid molecule to which DNA polymerase can attach the first nucleotide, is required. In most cases this primer is a short stretch of RNA.

When the double helix is opened at the beginning of replication, an RNA-polymerizing enzyme makes the RNA primer. This enzyme, called **primase**, synthesizes a short stretch of RNA of around 11–12 nucleotides that is complementary in base pairing to the template DNA. At the growing end of this RNA primer is a 3′-OH group to which DNA polymerase can add the first deoxyribonucleotide. Continued extension of the molecule thus occurs as DNA rather than RNA. The newly synthesized molecule has a structure like that shown in **Figure 6.15**. The primer will eventually be removed and replaced with DNA, as described later.

MiniQuiz

- To which end (5′ end or 3′ end) of a newly synthesized strand of DNA does polymerase add a base?
- Why is a primer required for DNA replication?

6.9 The Replication Fork

Much of our understanding of the details of DNA replication has been obtained from studying the bacterium *Escherichia coli*, and the following discussion deals primarily with this organism. However, DNA replication is probably quite similar in all *Bacteria*. By contrast, although most species of *Archaea* have circular chromosomes, many events in DNA replication resemble those in eukaryotic cells more than those in *Bacteria* (Chapter 7). This again reflects the phylogenetic affiliation between *Archaea* and *Eukarya*.

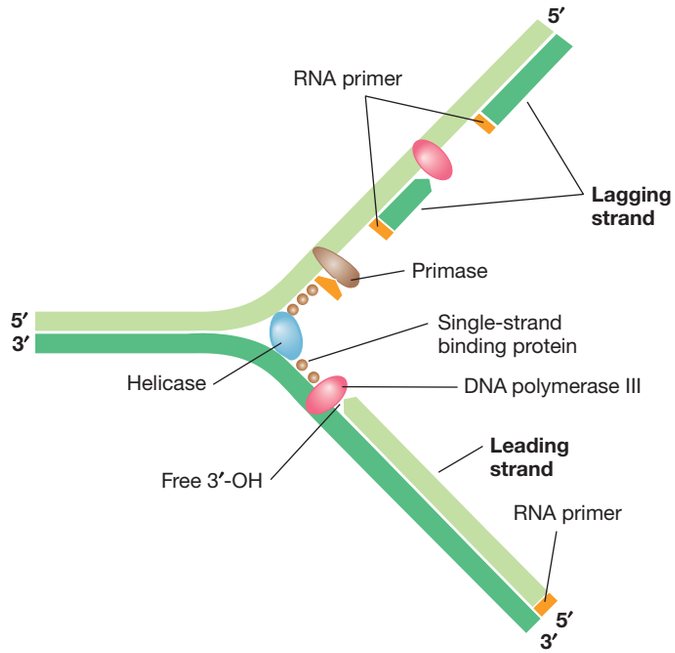


Figure 6.16 **Events at the DNA replication fork.** Note the polarity and antiparallel nature of the DNA strands.

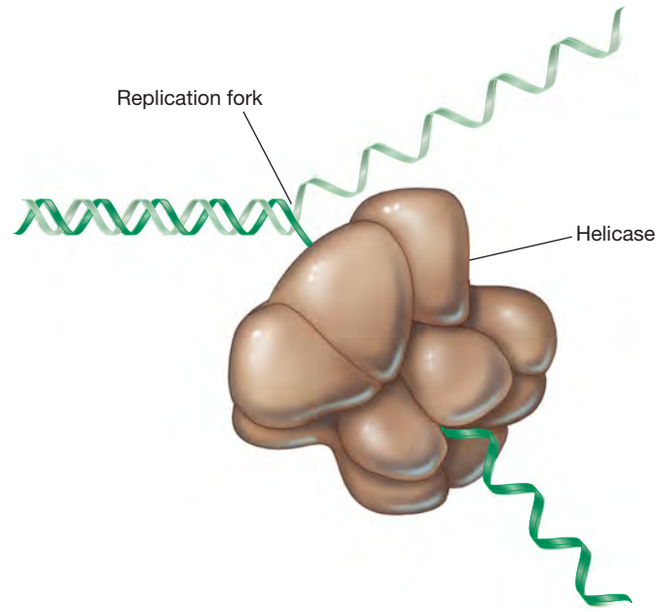


Figure 6.17 **DNA helicase unwinding a double helix.** In this figure, the protein and DNA molecules are drawn to scale. Simple diagrams often give the incorrect impression that most proteins are relatively small compared to DNA. Although DNA molecules are generally extremely long, they are relatively thin compared to many proteins.

Initiation of DNA Synthesis

Before DNA polymerase can synthesize new DNA, the double helix must be unwound to expose the template strands. The zone of unwound DNA where replication occurs is known as the **replication fork**. The enzyme DNA helicase unwinds the double helix, using energy from ATP, and exposes a short single-stranded region (**Figures 6.16** and **6.17**). Helicase moves along the DNA and separates the strands just in advance of the replication fork. The single-stranded region is covered by single-strand binding protein. This stabilizes the single-stranded DNA and prevents the double helix from re-forming.

Unwinding of the double helix by helicase generates positive supercoils ahead of the advancing replication fork. To counteract this, DNA gyrase travels along the DNA ahead of the replication fork and inserts negative supercoils to cancel out the positive supercoiling.

Bacteria possess a single location on the chromosome where DNA synthesis is initiated, the origin of replication (*oriC*). This consists of a specific DNA sequence of about 250 bases that is recognized by initiation proteins, in particular a protein called DnaA (**Table 6.3**), which binds to this region and opens up the double helix. Next to assemble is the helicase (known as DnaB), which is helped onto the DNA by the helicase loader protein (DnaC). Two helicases are loaded, one onto each strand, facing in opposite directions. Next, two primase and then two DNA polymerase enzymes are loaded onto the DNA behind the helicases. Initiation of DNA replication then begins on the two single strands. As replication proceeds, the replication fork appears to move along the DNA (Figure 6.16). www.microbiologyplace.com
Online Tutorial 6.1: DNA Replication

Leading and Lagging Strands

Figure 6.16 shows an important distinction in replication between the two DNA strands due to the fact that replication always proceeds from 5′ → 3′ (always adding a new nucleotide to the 3′-OH of the growing chain). On the strand growing from the 5′-PO_4^{2-} to the 3′-OH, called the **leading strand**, DNA synthesis occurs continuously because there is always a free 3′-OH at the replication fork to which a new nucleotide can be added. But on the opposite strand, called the **lagging strand**, DNA synthesis occurs discontinuously because there is no 3′-OH at the replication fork to which a new nucleotide can attach. Where is the 3′-OH on this strand? It is located at the opposite end, away from the replication fork. Therefore, on the lagging strand, RNA primers must be synthesized by primase multiple times to provide free 3′-OH groups. By contrast, the leading strand is primed only once, at the origin. As a result, the lagging strand is made in short segments, called *Okazaki fragments*, after their discoverer, Reiji Okazaki. These fragments are joined together later to give a continuous strand of DNA.

Synthesis of the New DNA Strands

After synthesizing the RNA primer, primase is replaced by Pol III. This enzyme is a complex of several proteins (Table 6.3), including the polymerase core enzyme itself. Each polymerase is held on the DNA by a sliding clamp, which encircles and slides along the single template strands of DNA. Consequently, the replication fork contains two polymerase core enzymes and two sliding clamps, one set for each strand. However, there is only a single clamp-loader complex. This is needed to assemble the sliding clamps onto the DNA. After assembly on the lagging

Table 6.3 ***Major enzymes involved in DNA replication in* Bacteria**

Enzyme	*Encoding genes*	*Function*
DNA gyrase	*gyrAB*	Replaces supercoils ahead of replisome
Origin-binding protein	*dnaA*	Binds origin of replication to open double helix
Helicase loader	*dnaC*	Loads helicase at origin
Helicase	*dnaB*	Unwinds double helix at replication fork
Single-strand binding protein	*ssb*	Prevents single strands from annealing
Primase	*dnaG*	Primes new strands of DNA
DNA polymerase III		Main polymerizing enzyme
Sliding clamp	*dnaN*	Holds Pol III on DNA
Clamp loader	*holA–E*	Loads Pol III onto sliding clamp
Dimerization subunit (Tau)	*dnaX*	Holds together the two core enzymes for the leading and lagging strands
Polymerase subunit	*dnaE*	Strand elongation
Proofreading subunit	*dnaQ*	Proofreading
DNA polymerase I	*polA*	Excises RNA primer and fills in gaps
DNA ligase	*ligA*, *ligB*	Seals nicks in DNA
Tus protein	*tus*	Binds terminus and blocks progress of the replication fork
Topoisomerase IV	*parCE*	Unlinking of interlocked circles

strand, the elongation component of Pol III, DnaE, then adds deoxyribonucleotides until it reaches previously synthesized DNA (**Figure 6.18**). At this point, Pol III stops.

The next enzyme to take part, Pol I, has more than one enzymatic activity. Besides synthesizing DNA, Pol I has a 5′ → 3′ exonuclease activity that removes the RNA primer preceding it (Figure 6.18). When the primer has been removed and replaced with DNA, Pol I is released. The last phosphodiester bond is made by an enzyme called **DNA ligase**. This enzyme seals nicks in DNAs that have an adjacent 5′-PO_4^{2-} and 3′-OH (something that Pol III is unable to do), and along with Pol I, it also participates in DNA repair. DNA ligase is also important for sealing genetically manipulated DNA during molecular cloning (↻ Section 11.3).

MiniQuiz

- Why are there leading and lagging strands?
- What recognizes the origin of replication?
- What enzymes take part in joining the fragments of the lagging strand?

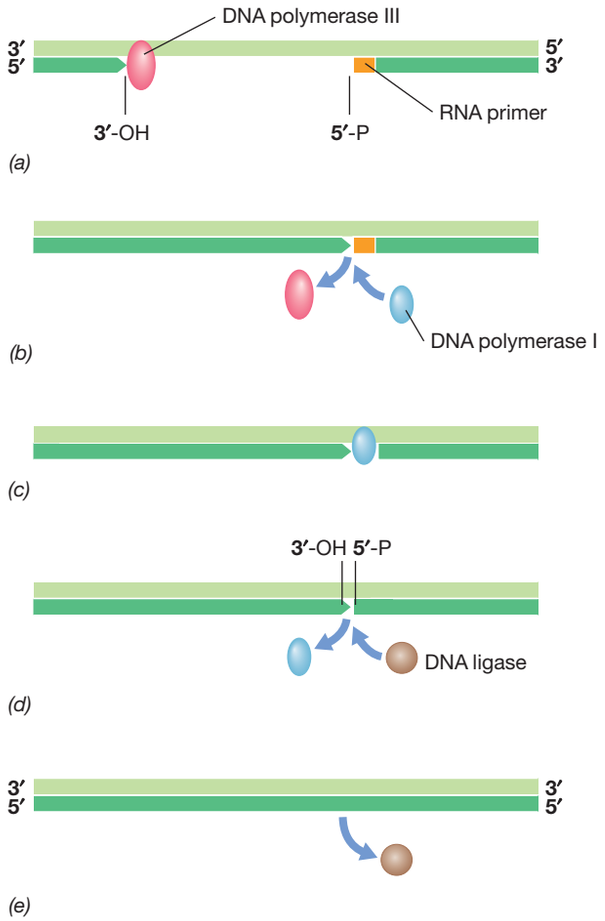


Figure 6.18 **Sealing two fragments on the lagging strand.** *(a)* DNA polymerase III is synthesizing DNA in the 5′ → 3′ direction toward the RNA primer of a previously synthesized fragment on the lagging strand. *(b)* On reaching the fragment, DNA polymerase III leaves and is replaced by DNA polymerase I. *(c)* DNA polymerase I continues synthesizing DNA while removing the RNA primer from the previous fragment. *(d)* DNA ligase replaces DNA polymerase I after the primer has been removed. *(e)* DNA ligase seals the two fragments together.

6.10 Bidirectional Replication and the Replisome

The circular nature of the chromosome of *Escherichia coli* and most other prokaryotes creates an opportunity for speeding up replication. In *E. coli*, and probably in all prokaryotes with circular chromosomes, replication is *bidirectional* from the origin of replication, as shown in **Figures 6.19** and **6.20**. There are thus two replication forks on each chromosome moving in opposite directions. These are held together by the two Tau protein subunits. In circular DNA, bidirectional replication leads to the formation of characteristic shapes called theta structures (Figure 6.19). Most large DNA molecules, whether from prokaryotes or eukaryotes, have bidirectional replication from fixed origins. In fact, large eukaryotic chromosomes have multiple origins (↻ Section 7.7). During bidirectional replication, synthesis occurs in both a leading and lagging fashion on each template strand (Figure 6.20).

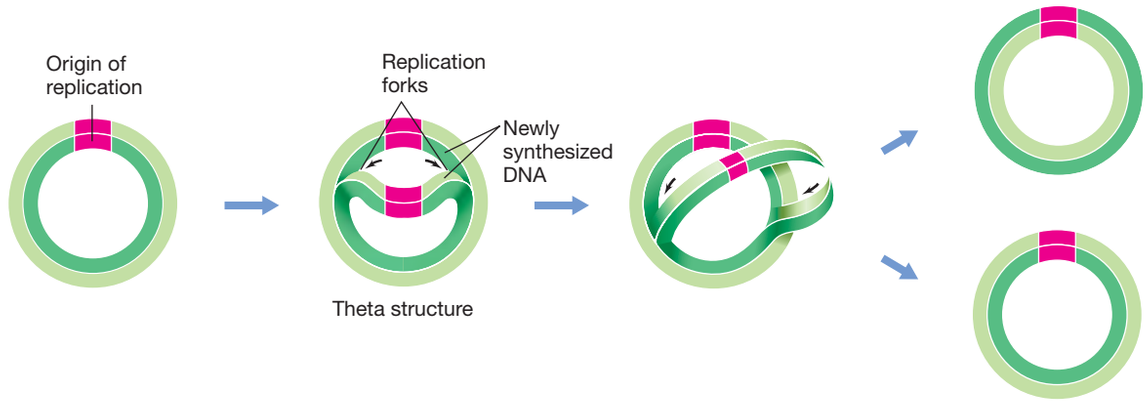


Figure 6.19 **Replication of circular DNA: the theta structure.** In circular DNA, bidirectional replication from an origin forms an intermediate structure resembling the Greek letter theta (θ).

Bidirectional DNA synthesis around a circular chromosome allows DNA to replicate as rapidly as possible. Even taking this into account and considering that Pol III can add nucleotides to a growing DNA strand at the rate of about 1000 per second, chromosome replication in *E. coli* still takes about 40 min. Interestingly, under the best growth conditions, *E. coli* can grow with a doubling time of about 20 min. However, even under these conditions, chromosome replication still takes 40 min. The solution to this conundrum is that cells of *E. coli* growing at doubling times shorter than 40 min contain multiple DNA replication forks. That is, a new round of DNA replication begins before the last round has been completed (**Figure 6.21**). Only in this way can a generation time shorter than the chromosome replication time be maintained.

The Replisome

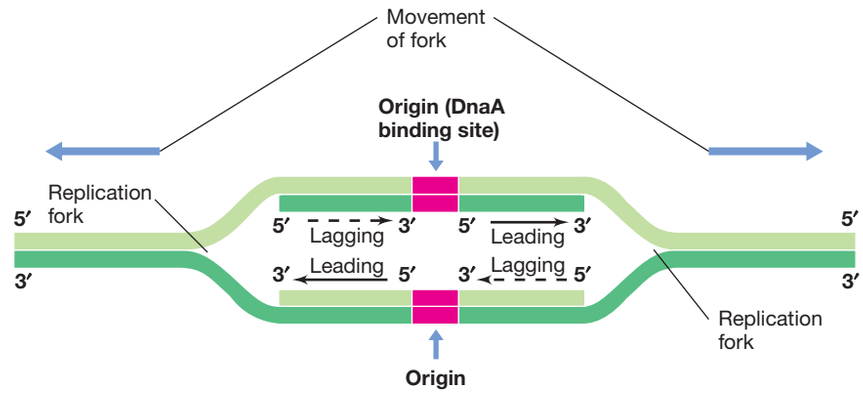
Figure 6.16 shows the differences in replication of the leading and the lagging strands and the enzymes involved. From such a simplified drawing it would appear that each replication fork contains a host of different proteins all working independently. Actually, this is not so. These proteins aggregate to form a large replication complex called the *replisome* (**Figure 6.22**). The lagging strand of DNA loops out to allow the replisome to move smoothly along both strands, and the replisome literally pulls the DNA template through it as replication occurs. Therefore, it is the DNA, rather than DNA polymerase, that moves during replication. Note also how helicase and primase form a subcomplex, called the *primosome*, which aids their working in close association during the replication process.

In summary, in addition to Pol III, the replisome contains several key replication proteins: (1) DNA gyrase, which removes supercoils; (2) DNA helicase and primase (the primosome), which unwind and prime the DNA; and (3) single-strand binding protein, which prevents the separated template strands from reforming a double helix (Figure 6.22). Table 6.3 summarizes the properties of proteins essential for DNA replication.

Fidelity of DNA Replication: Proofreading

DNA replicates with a remarkably low error rate. Nevertheless, when errors do occur, a backup mechanism exists to detect and correct them. Errors in DNA replication introduce mutations, changes in DNA sequence. Mutation rates in cells are remarkably low, between 10^{-8} and 10^{-11} errors per base pair inserted. This accuracy is possible partly because DNA polymerases get two chances to incorporate the correct base at a given site. The first chance comes when complementary bases are inserted opposite the bases on the template strand by Pol III according to the base-pairing rules, A with T and G with C. The second chance depends upon a second enzymatic activity of both Pol I and Pol III, called *proofreading* (**Figure 6.23**). In Pol III, a separate protein subunit,

Figure 6.20 **Dual replication forks in the circular chromosome.** At an origin of replication that directs bidirectional replication, two replication forks must start. Therefore, two leading strands must be primed, one in each direction. In *Escherichia coli*, the origin of replication is recognized by a specific protein, DnaA. Note that DNA synthesis is occurring in both a leading and a lagging manner on each of the new daughter strands. Compare this figure with the description of the replisome shown in Figure 6.22.



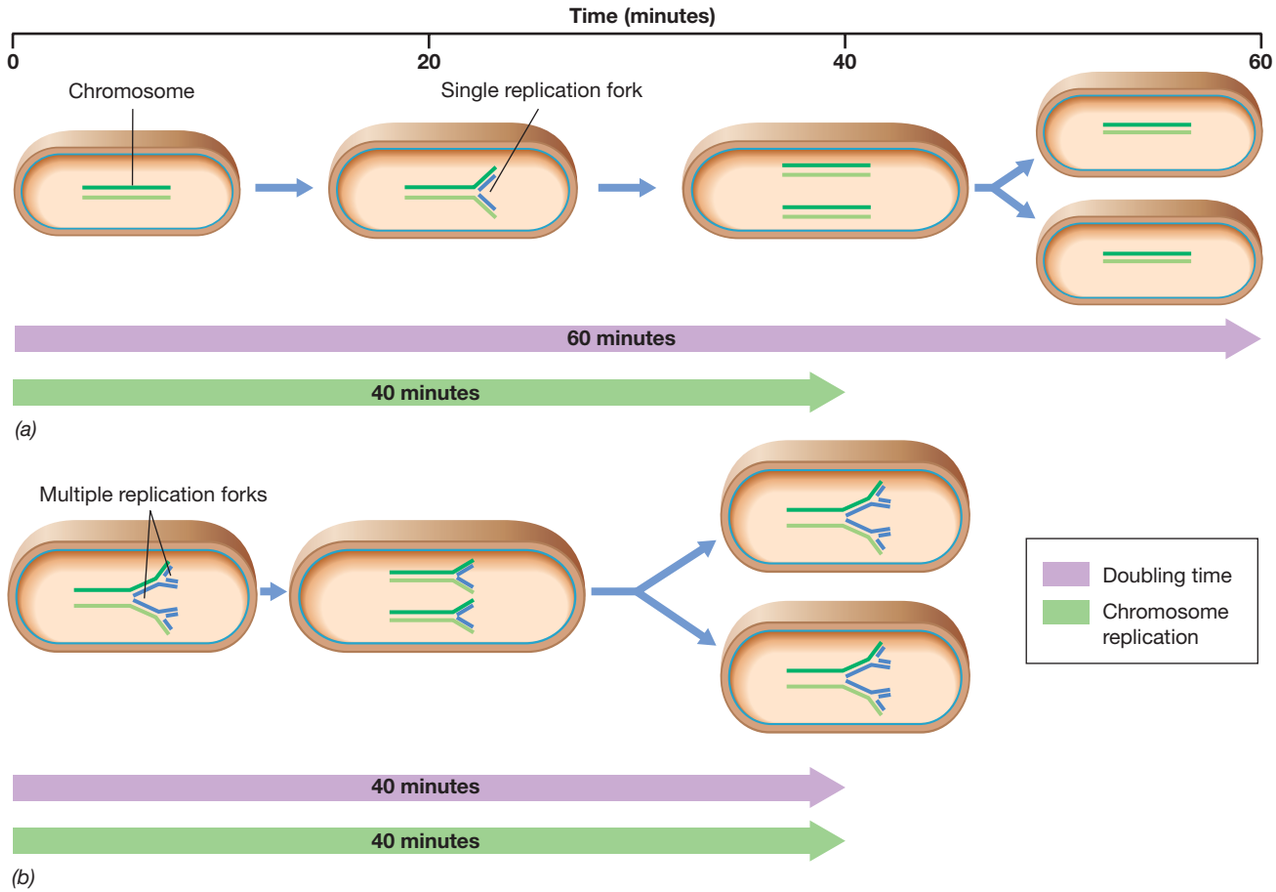


Figure 6.21 **Cell division versus chromosome duplication.** *(a)* Cells of *Escherichia coli* take approximately 40 min to replicate the chromosome and an additional 20 min for cell division. *(b)* When cells double in less than 60 min, a new round of chromosome replication must be initiated before the previous round is finished.

DnaQ, performs the proofreading, whereas in Pol I a single protein performs all functions.

Proofreading activity occurs if an incorrect base has been inserted because this creates a mismatch in base pairing. Both Pol I and Pol III possess a 3′ → 5′ exonuclease activity that can remove such wrongly inserted nucleotides. The polymerase senses the mistake because incorrect base pairing causes a slight distortion in the double helix. After the removal of a mismatched nucleotide, the polymerase then gets a second chance to insert the correct nucleotide (Figure 6.23). The proofreading exonuclease activity is distinct from the 5′ → 3′ exonuclease activity of Pol I that removes the RNA primer from both the leading and lagging strands. Only Pol I has this latter activity. Exonuclease proofreading occurs in prokaryotes, eukaryotes, and viral DNA replication systems. However, many organisms have additional mechanisms for reducing errors made during DNA replication, which operate after the replication fork has passed by. We will discuss some of these in Chapter 10.

Termination of Replication

Eventually the process of DNA replication is finished. How does the replisome know when to stop? On the opposite side of the circular chromosome from the origin is a site called the terminus of replication. Here the two replication forks collide as the new circles of DNA are completed. The details of termination are not fully known. However, in the terminus region there are several DNA sequences called *Ter* sites that are recognized by a protein called Tus, whose function is to block progress of the replication forks. When replication of the circular chromosome is complete, the two circular molecules are linked together, much like the links of a chain. They are unlinked by another enzyme, topoisomerase IV. Obviously, it is critical that, after DNA replication, the DNA is partitioned so that each daughter cell receives a copy of the chromosome. This process may be assisted by the important cell division protein FtsZ, which helps orchestrate several key events of cell division (↻ Section 5.2).

MiniQuiz

- What is the replisome and what are its components?
- How can *Escherichia coli* carry out cell division in less time than it takes to duplicate its chromosome?
- How is proofreading carried out during DNA replication?
- What brings the replication forks to a halt in the terminus region of the chromosome?

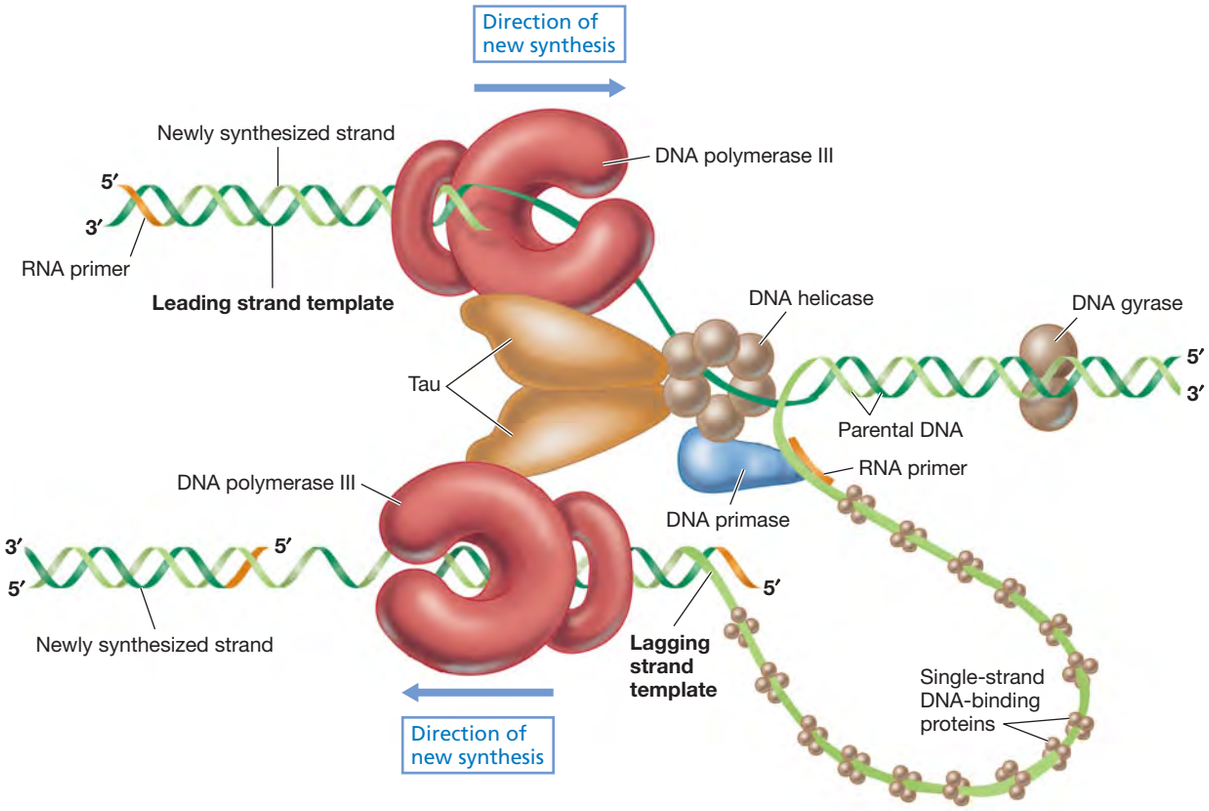


Figure 6.22 **The replisome.** The replisome consists of two copies of DNA polymerase III, plus helicase and primase (together forming the primosome), and many copies of single-strand DNA-binding protein. The tau subunits hold the two DNA polymerase assemblies and helicase together. Just upstream of the replisome, DNA gyrase removes supercoils in the DNA to be replicated. Note that the two polymerases are replicating the two individual strands of DNA in opposite directions. Consequently, the lagging-strand template loops around so that the whole replisome moves in the same direction along the chromosome.

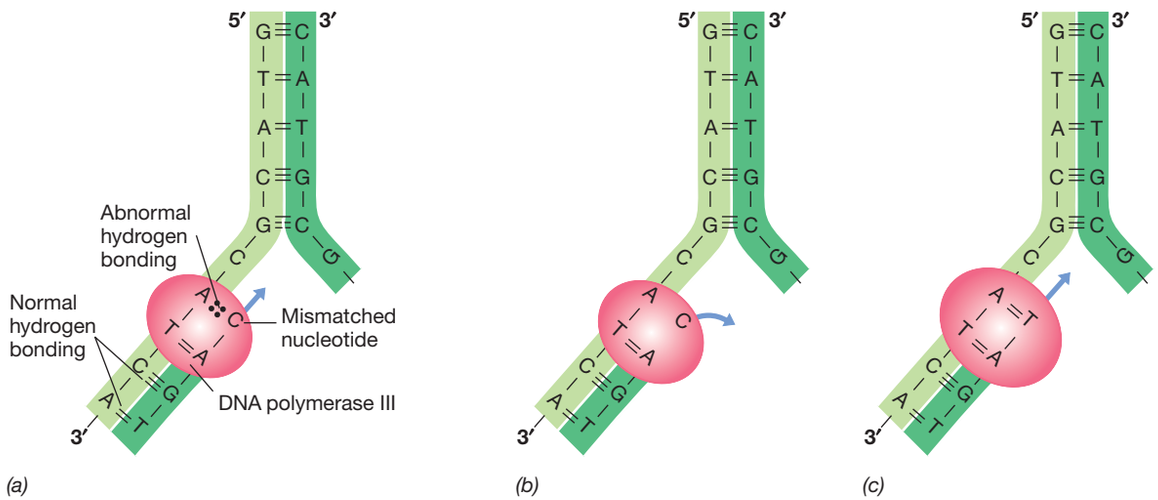


Figure 6.23 **Proofreading by the 3′ → 5′ exonuclease activity of DNA polymerase III.** *(a)* A mismatch in base pairing at the terminal base pair causes the polymerase to pause briefly. This signals the proofreading activity to *(b)* cut out the mismatched nucleotide, after which *(c)* the correct base is inserted by the polymerase activity.

6.11 The Polymerase Chain Reaction (PCR)

The **polymerase chain reaction (PCR)** is essentially DNA replication *in vitro*. The PCR can copy segments of DNA by up to a billionfold in the test tube, a process called *amplification*. This yields large amounts of specific genes or other DNA segments that may be used for a host of applications in molecular biology. PCR uses the enzyme DNA polymerase, which naturally copies DNA molecules (Section 6.8). Artificially synthesized primers (↻ Section 11.4) are used to initiate DNA synthesis, but are made of DNA (rather than RNA like the primers used by cells). PCR does not actually copy whole DNA molecules but amplifies stretches of up to a few thousand base pairs (the *target*) from within a larger DNA molecule (the *template*). PCR was conceived by Kary Mullis, who received a Nobel Prize for this achievement.

The steps in PCR amplification of DNA can be summarized as follows (**Figure 6.24**):

1. The template DNA is denatured by heating.
2. Two artificial DNA oligonucleotide primers flanking the target DNA are present in excess. This ensures that most template strands anneal to a primer, and not to each other, as the mixture cools (Figure 6.24*a*).
3. DNA polymerase then extends the primers using the original DNA as the template (Figure 6.24*b*).
4. After an appropriate incubation period, the mixture is heated again to separate the strands. The mixture is then cooled to allow the primers to hybridize with complementary regions of newly synthesized DNA, and the whole process is repeated (Figure 6.24*c*).

The power of PCR is that the products of one primer extension are templates for the next cycle. Consequently, each cycle doubles the amount of the original target DNA. In practice, 20–30 cycles are usually run, yielding a 10^6-fold to 10^9-fold increase in the target sequence (Figure 6.24*d*). Because the technique consists of several highly repetitive steps, PCR machines, called *thermocyclers*, are available that run through the heating and cooling cycles automatically. Because each cycle requires only about 5 min, the automated procedure gives large amplifications in only a few hours.

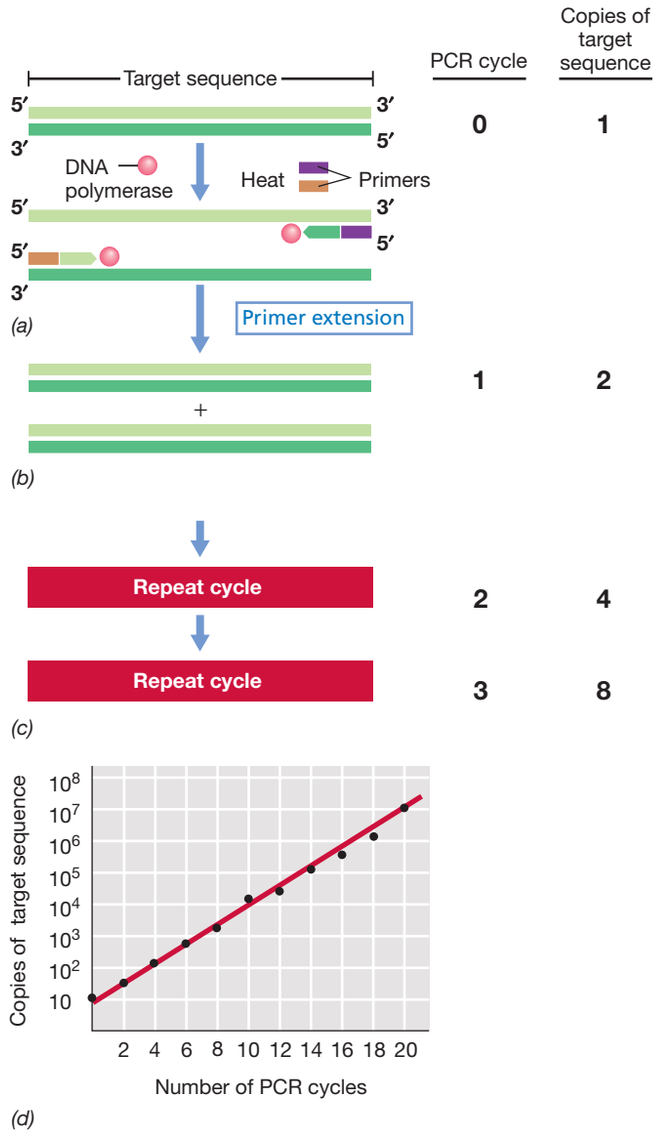


Figure 6.24 The polymerase chain reaction (PCR). The PCR amplifies specific DNA sequences. *(a)* Target DNA is heated to separate the strands, and a large excess of two oligonucleotide primers, one complementary to each strand, is added along with DNA polymerase. *(b)* Following primer annealing, primer extension yields a copy of the original double-stranded DNA. *(c)* Two additional PCR cycles yield four and eight copies, respectively, of the original DNA sequence. *(d)* Effect of running 20 PCR cycles on a DNA preparation originally containing ten copies of a target gene. Note that the plot is semilogarithmic.

PCR at High Temperature

The original PCR technique employed the DNA polymerase *Escherichia coli* Pol III, but because of the high temperatures needed to denature the double-stranded copies of DNA, the enzyme was also denatured and had to be replenished every cycle. This problem was solved by employing a thermostable DNA polymerase isolated from the thermophilic hot spring bacterium *Thermus aquaticus*. DNA polymerase from *T. aquaticus*, called *Taq polymerase*, is stable to 95°C and thus is unaffected by the denaturation step employed in the PCR. The use of Taq DNA polymerase also increased the specificity of the PCR because the DNA is copied at 72°C rather than 37°C. At such high temperatures, nonspecific hybridization of primers to nontarget DNA is rare, thus making the products of Taq PCR more homogeneous than those obtained using the *E. coli* enzyme. On the other hand, the primer hybridization step is often carried out at lower temperatures, which may allow some nonspecific binding.

DNA polymerase from *Pyrococcus furiosus*, a hyperthermophile with a growth temperature optimum of 100°C (↻ Section 19.5) is called *Pfu polymerase* and is even more thermostable than Taq polymerase. Moreover, unlike Taq polymerase, Pfu polymerase has proofreading activity (Section 6.10), making it especially useful when high accuracy is crucial. Thus, the error rate for Taq

polymerase under standard conditions is 8.0×10^{-6} (per base duplicated), whereas for Pfu polymerase it is only 1.3×10^{-6}. To supply the commercial demand for thermostable DNA polymerases, the genes for these enzymes have been cloned into *E. coli*, allowing the enzymes to be produced in large quantities. www.microbiologyplace.com **Online Tutorial 6.2: Polymerase Chain Reaction (PCR)**

Applications and Sensitivity of PCR

PCR is a powerful tool. It is easy to perform, extremely sensitive and specific, and highly efficient. During each round of amplification the amount of product doubles, leading to an exponential increase in the DNA. This means not only that a large amount of amplified DNA can be produced in just a few hours, but that only a few molecules of target DNA need be present in the sample to start the reaction. The reaction is so specific that, with primers of 15 or so nucleotides and high annealing temperatures, there is almost no "false priming," and therefore the PCR product is virtually homogeneous.

PCR is extremely valuable for obtaining DNA for cloning genes or for sequencing purposes because the gene or genes of interest can easily be amplified if flanking sequences are known. PCR is also used routinely in comparative or phylogenetic studies to amplify genes from various sources. In these cases the primers are made for regions of the gene that are conserved in sequence across a wide variety of organisms. Because 16S rRNA, a molecule used for phylogenetic analyses, has both highly conserved and highly variable regions, primers specific for the 16S rRNA gene from various taxonomic groups can be synthesized. These may be used to survey different groups of organisms in any specific habitat. This technique is in widespread use in microbial ecology and has revealed the enormous diversity of the microbial world, much of it not yet cultured (↻ Section 22.5).

Because it is so sensitive, PCR can be used to amplify very small quantities of DNA. For example, PCR has been used to amplify and clone DNA from sources as varied as mummified human remains and fossilized plants and animals. The ability of PCR to amplify and analyze DNA from cell mixtures has also made it a common tool of diagnostic microbiology. For example, if a clinical sample shows evidence of a gene specific to a particular pathogen, then it can be assumed that the pathogen was present in the sample. Treatment of the patient can then begin without the need to culture the organism, a time-consuming and often fruitless process. PCR has also been used in forensics to identify human individuals from very small samples of their DNA.

MiniQuiz

- Why is a primer needed at each end of the DNA segment being amplified by PCR?
- From which organisms are thermostable DNA polymerases obtained?
- How has PCR improved diagnostic clinical medicine?

IV RNA Synthesis: Transcription

Transcription is the synthesis of ribonucleic acid (RNA) using DNA as a template. There are three key differences in the chemistry of RNA and DNA: (1) RNA contains the sugar ribose instead of deoxyribose; (2) RNA contains the base uracil instead of thymine; and (3) except in certain viruses, RNA is not double-stranded. The change from deoxyribose to ribose affects the chemistry of a nucleic acid; enzymes that act on DNA usually have no effect on RNA, and vice versa. However, the change from thymine to uracil does not affect base pairing, as these two bases pair with adenine equally well.

RNA plays several important roles in the cell. Three major types of RNA are involved in protein synthesis: **messenger RNA (mRNA)**, **transfer RNA (tRNA)**, and **ribosomal RNA (rRNA)**. Several other types of RNA also occur that are mostly involved in regulation (Chapter 8). These RNA molecules all result from the transcription of DNA. It should be emphasized that RNA operates at two levels, genetic and functional. At the genetic level, mRNA carries genetic information from the genome to the ribosome. In contrast, rRNA has both a functional and a structural role in ribosomes and tRNA has an active role in carrying amino acids for protein synthesis. Indeed, some RNA molecules including rRNA have enzymatic activity (ribozymes, ↻ Section 7.8). Here we focus on how RNA is synthesized in the *Bacteria*, using *Escherichia coli* as our model organism.

6.12 Overview of Transcription

Transcription is carried out by the enzyme **RNA polymerase**. Like DNA polymerase, RNA polymerase catalyzes the formation of phosphodiester bonds but between ribonucleotides rather than deoxyribonucleotides. RNA polymerase uses DNA as a template. The precursors of RNA are the ribonucleoside triphosphates ATP, GTP, UTP, and CTP. The mechanism of RNA synthesis is much like that of DNA synthesis. During elongation of an RNA chain, ribonucleoside triphosphates are added to the 3′-OH of the ribose of the preceding nucleotide. Polymerization is driven by the release of energy from the two energy-rich phosphate bonds of the incoming ribonucleoside triphosphates. In both DNA replication and RNA transcription the overall direction of chain growth is from the 5′ end to the 3′ end; thus the new strand is antiparallel to the template strand. Unlike DNA polymerase, however, RNA polymerase can initiate new strands of nucleotides on its own; consequently, no primer is necessary.

RNA Polymerases

The template for RNA polymerase is a double-stranded DNA molecule, but only one of the two strands is transcribed for any given gene. Nevertheless, genes are present on both strands of DNA and thus DNA sequences on both strands are transcribed, although at different locations. Although these principles are true for transcription in all organisms, there are significant differences among RNA polymerase from *Bacteria, Archaea*, and *Eukarya*. The following discussion deals only with RNA polymerase from *Bacteria*, which has the simplest structure and about which most is known (RNA polymerase in *Archaea* and *Eukarya* is discussed in Chapter 7).