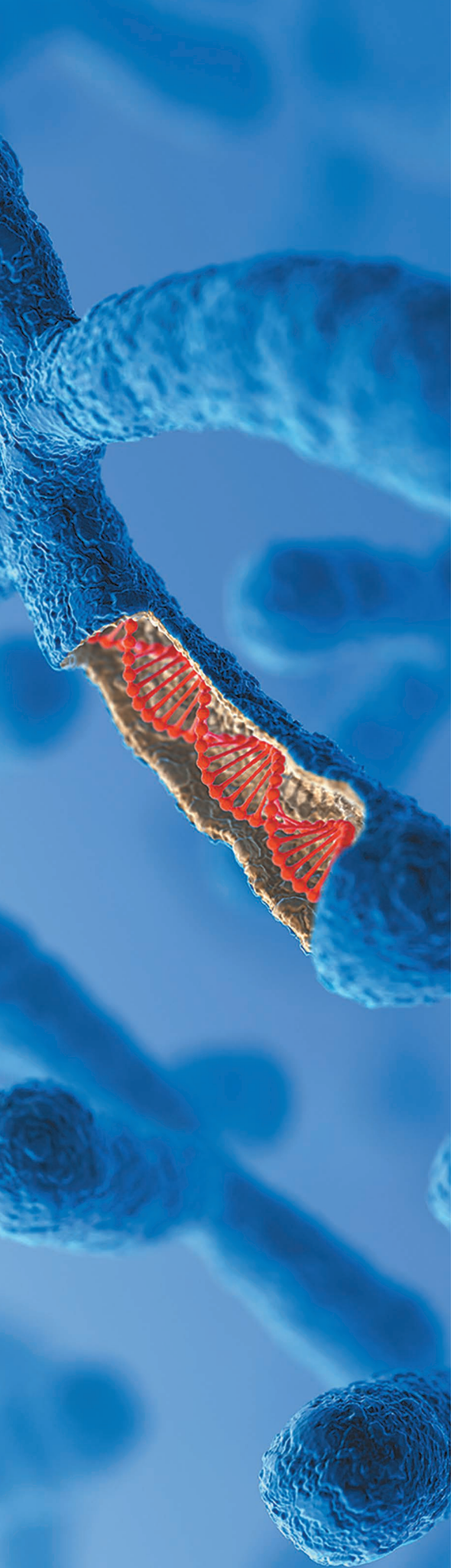


PART I

Genes and Chromosomes

CHAPTER OUTLINE

CHAPTER 1

Genes Are DNA and Encode RNAs and Polypeptides

CHAPTER OUTLINE

1.1 Introduction

genome The complete set of nucleotide sequences in the genetic material of an organism. It includes the sequence of each chromosome plus any DNA in organelles or plasmids.

The hereditary basis of every living organism is its **genome**, the complete sequence of deoxyribonucleic acid (DNA) that provides the full set of hereditary information carried by the organism's cells. The genome includes chromosomal DNA, DNA in plasmids, and in eukaryotes, also includes DNA in mitochondria and chloroplasts. We use the term *information* because the genome does not itself perform an active role in the development of the organism. It is the sequence of the individual subunits, or nucleotides, of the DNA that determines development. By a complex series of interactions, the DNA sequence encodes all the ribonucleic acids (RNAs) and polypeptide subunits of proteins of the organism that are to be produced at the appropriate time and in the appropriate cells. Proteins serve diverse roles in the development and functioning of an organism: they can form parts of the basic structure of the organism; they have the capacity to build the structures; they perform the metabolic reactions necessary for life; and they participate in many aspects of regulation as transcription factors, receptors, key players in signal transduction pathways, and many more. RNAs encoded by the genome but that do not themselves encode proteins also assist with the expression of genes during development, among other roles.

chromosome Discrete unit of the genome carrying many genes. Each consists of a very long molecule of duplex DNA and (in eukaryotes) an approximately equal mass of proteins. It is visible as a morphological entity only during cell division.

Physically, a cellular genome may be divided into a number of different DNA molecules, or **chromosomes**. Functionally, the genome is divided into genes and other regulatory sequences. Each gene is a sequence of DNA that encodes a single type of RNA or polypeptide (although it may encode multiple versions of its product). Each of the discrete chromosomes comprising the genome may contain a large number of genes. Genomes for living organisms may contain as few as about 500 genes (for a mycoplasma, a type of bacterium) to about 20,000 for a human being and as many as 50,000 in some plants.

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

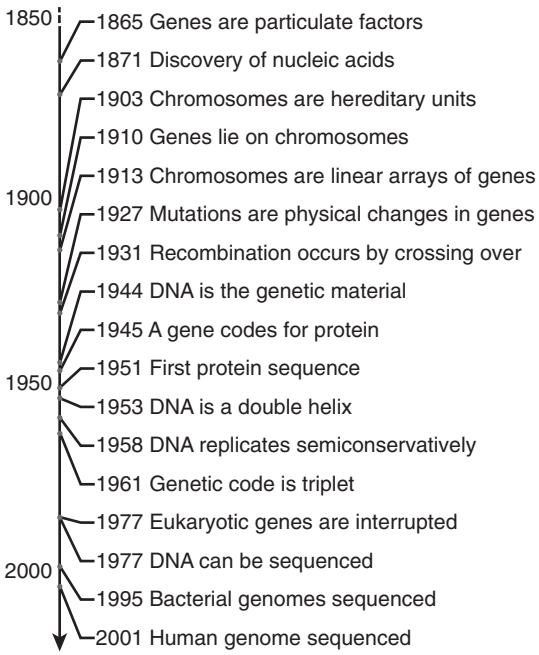


FIGURE 1.1 A brief history of genetics.

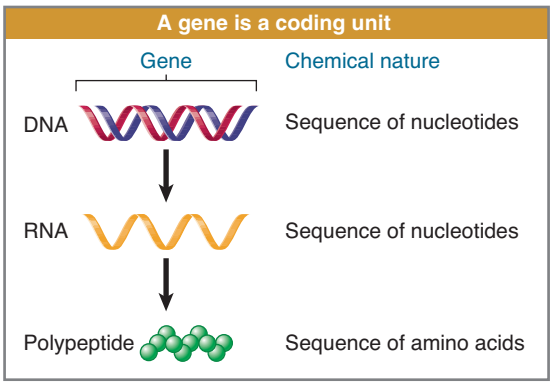


FIGURE 1.2 A gene encodes an RNA, which may encode a polypeptide.

Understanding the process by which a gene is expressed allows us to make a more rigorous definition of its nature. **FIGURE 1.2** shows the basic theme of this text. A gene is a sequence of DNA that is the template for the production of a single strand of another nucleic acid, RNA, with a sequence that is, at least initially, identical to one of the two polynucleotide strands of DNA. In many cases, the RNA is in turn used to direct production of a polypeptide, whereas in other cases (such as rRNA, tRNA, and many other genes), the RNA transcribed from the gene is the functional end product. Thus, a gene is a sequence of DNA that encodes an RNA, and in protein-coding genes, the RNA in turn encodes a polypeptide.

The basic pattern of inheritance of a gene was proposed by Mendel more than a century ago. Summarized in his two major principles of *segregation* and *independent assortment*, the gene was recognized as a "particulate factor" that passes unchanged from parent to progeny. A gene may exist in alternative forms, called **alleles**.

allele One of several alternative forms of a gene occupying a given locus on a chromosome.

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

Each chromosome consists of a linear array of genes and additional sequences, and each gene resides at a particular location on the chromosome. The location is more formally called a genetic **locus**. The alleles of a gene are the different forms that are found at its locus. Although generally there are up to two alleles per locus in an individual, a population may have many alleles.

locus The position on a chromosome at which a particular gene resides. It may be occupied by any one of the alleles for the gene.

The key to understanding the organization of genes into chromosomes was the discovery of genetic **linkage**—the tendency for alleles of different genes on the same chromosome to be inherited together in the progeny instead of assorting independently, as predicted by Mendel's principle. Once the unit of *recombination* (reassortment) was introduced as a measure of linkage, the construction of genetic linkage maps became possible.

linkage The tendency of genes to be inherited together as a result of their location on the same chromosome. This is measured by the percentage of recombination between loci.

The resolution of the linkage map of a multicellular eukaryote is restricted by the small number of progeny that can be obtained from each mating. Recombination occurs so infrequently between nearby points that it is rarely observed between different variable sites in the same gene. As a result, classical linkage maps of eukaryotes can place the genes in order but cannot resolve the locations of variable sites within a gene. By using a microbial system in which a very large number of progeny can be obtained from each genetic cross, researchers could demonstrate that recombination occurs

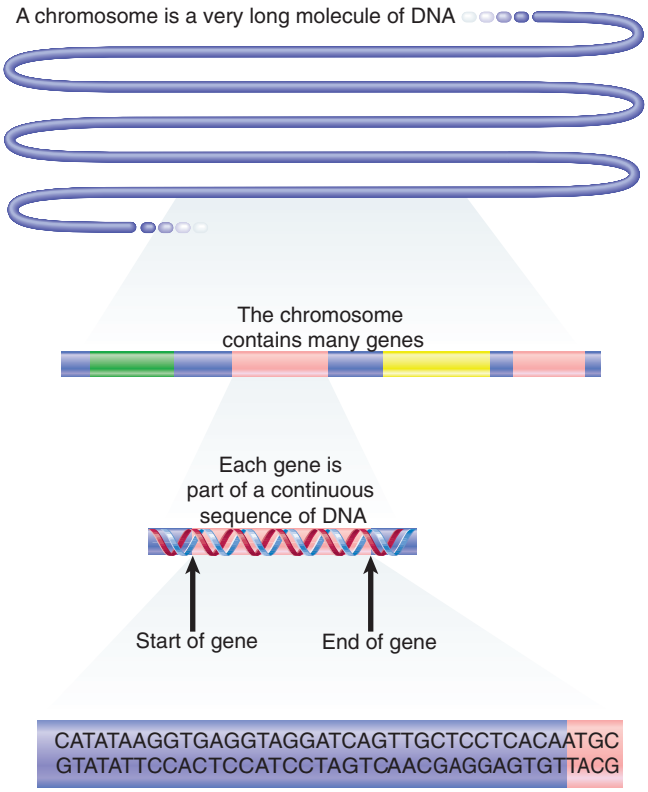


FIGURE 1.3 Each chromosome has a single long molecule of DNA, within which are the sequences of individual genes.

within genes and that it follows the same rules as those for recombination between genes.

Variable nucleotide sites among alleles of a gene can be mapped into a linear order, showing that the gene itself has the same linear construction as the array of genes on a chromosome. In other words, the genetic map is linear within, as well as between, loci as an unbroken sequence of nucleotides. This conclusion leads naturally to the modern view summarized in **FIGURE 1.3** that the genetic material of a chromosome consists of an uninterrupted length of DNA representing many genes.

From the demonstration that a gene consists of DNA and that a chromosome consists of a long stretch of DNA including many genes, we will move to the overall organization of the genome. In the chapter "The Interrupted Gene," we take up in more detail the organization of the gene and its representation in proteins. In the chapter "The Content of the Genome," we consider the total number of genes, and in the chapter "Clusters and Repeats," we discuss other components of the genome and the maintenance of its organization.

CONCEPT AND REASONING CHECKS

1. Why is it accurate to say that a genome has information for the development of an organism but does not directly participate in development?
2. Early work measuring recombination frequencies between genes led to the establishment of "linkage groups," sets of genes that do not assort independently. How does this support the concept that genes are carried on chromosomes?

1.2 DNA as the Genetic Material

The idea that the genetic material is DNA has its roots in the discovery of **transformation** by Frederick Griffith in 1928 (see the accompanying box, "Historical Perspectives: Determining that DNA Is the Genetic Material"). Purification of the "transforming principle" from bacterial cells in 1944 by Avery, MacLeod, and McCarty showed that it is DNA.

transformation In bacteria, the acquisition of new genetic material by incorporation of added DNA.

Having shown that DNA is the genetic material of bacteria, the next step was to demonstrate that DNA is the genetic material in a quite different system. Phage T2 is a bacteriophage virus (or phage) that infects the bacterium *Escherichia coli*. When phage particles are added to bacteria, they attach to the outside surface, some material enters the cell, and then approximately 20 minutes later each cell bursts open, or lyses, to release a large number of progeny phage.

FIGURE 1.4 illustrates the results of an experiment in 1952 by Alfred Hershey and Martha Chase in which bacteria were infected with T2 phages that had been radioactively labeled either in their DNA component (with ^{32}P) or in their protein component (with ^{35}S). The infected bacteria were agitated in a blender, and two fractions were separated by centrifugation. One fraction contained the empty phage "ghosts" that were released from the surface of the bacteria, and the other consisted of the infected bacteria themselves. Previously, it had been shown that phage replication occurs intracellularly, so that the genetic material of the phage would have to enter the cell during infection.

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

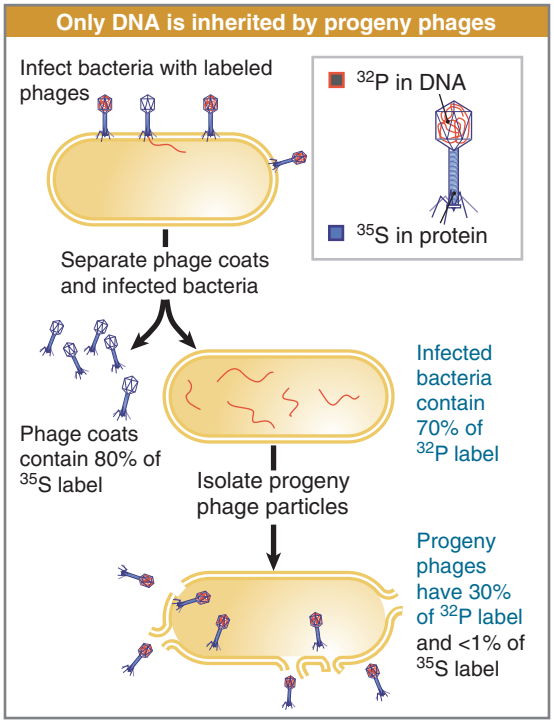


FIGURE 1.4 The genetic material of phage T2 is DNA.

A phage reproduces by commandeering the biochemical machinery of an infected host cell to manufacture more copies of itself. The phage possesses genetic material with properties analogous to those of cellular genomes: its traits are faithfully expressed and are subject to the same rules that govern inheritance of cellular traits. The case of T2 reinforces the general conclusion that DNA is genetic material of the genome of a cell or a virus.

When DNA is added to eukaryotic cells growing in culture under the appropriate conditions, it enters the cells, and in some of them this results in the production of new proteins. When an isolated gene is used, its incorporation leads to the production of a particular protein, as depicted in **FIGURE 1.5**. Although for historical reasons these experiments are described as **transfection** when performed with animal cells, they are a direct counterpart to bacterial transformation. In the process of *stable transfection*, DNA that is introduced into the recipient cell becomes part of its genome and is inherited with it, and expression of the new DNA results in a new trait. (The related method of *transient transfection* introduces DNA that is gradually lost.) At first, these experiments were successful only with individual cells growing in culture, but in later experiments DNA was introduced into mouse eggs by microinjection and became a stable part of the genome of the mouse. Such experiments show directly that DNA is the genetic material in eukaryotes and that it can be transferred between different species and remain functional.

transfection In eukaryotic cells, the acquisition of new genetic material by incorporation of added DNA.

The genetic material of all known organisms and many viruses is DNA. However, some viruses use RNA as the genetic material. Therefore, the general nature of the genetic material is that it is always nucleic acid; specifically, it is DNA, except in the RNA viruses.

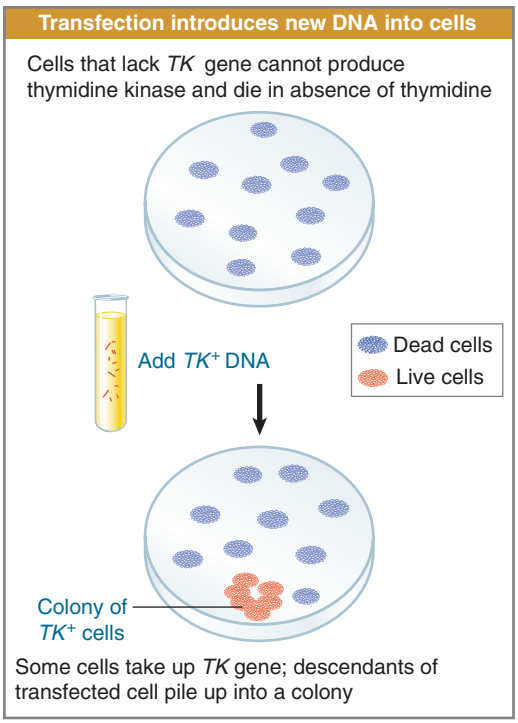


FIGURE 1.5 Eukaryotic cells can acquire a new phenotype as the result of transfection by added DNA.

KEY CONCEPTS

- Bacterial transformation provided the first support that DNA is the genetic material of bacteria. Genetic properties can be transferred from one bacterial strain to another by extracting DNA from the first strain and adding it to the second strain.
- Phage infection showed that DNA is the genetic material of viruses. When the DNA and protein components of bacteriophages are labeled with different radioactive isotopes, only the DNA is transmitted to the progeny phages produced by infecting bacteria.
- DNA can be used to introduce new genetic traits into animal cells or whole animals.
- In some viruses, the genetic material is RNA.

CONCEPT AND REASONING CHECK

If Hershey and Chase had observed that nearly none of the DNA but a substantial fraction of the protein of phage T2 enters *E. coli* cells during infection, what would they have concluded?

HISTORICAL PERSPECTIVES

Determining That DNA Is the Genetic Material

Pneumonia was a leading cause of death in the early part of the 20th century, and much effort was put into understanding the structure and function of the bacteria known to cause it: the pneumococci. Scientists knew that several pneumococcal types existed and that these types could be distinguished by the molecules (capsular polysaccharides) displayed on their surface, but they did not know whether the different types represented individual and stable strains of bacteria or different developmental stages of a single strain. While conducting experiments to distinguish between these possibilities, Frederick Griffith set in motion a series of experiments that ultimately led to the identification of DNA as the genetic material of all cells.

Griffith studied S forms and R forms of pneumococcal bacteria (so-called for the smooth and rough appearance of the bacteria when grown in laboratory culture). The R form, which was generated in the laboratory from the S form, produced no capsular polysaccharides and did not cause pneumonia when injected into mice. The S form was lethal to mice but could be inactivated by exposing the bacteria to high heat before injection. Surprisingly, Griffith found that when the R form was injected into mice alongside heat-killed S-type bacteria, the mice contracted pneumonia and died (**FIGURE A**). What was more, live S-type bacteria—virulent and coated with capsular proteins—could be isolated from the dead mice. R form bacteria had clearly been transformed into a stable S form, and the transformation process required some factor from the heat-killed S sample. Griffith suspected this factor, the so-called "transforming principle," was a protein . . . and he was not alone.

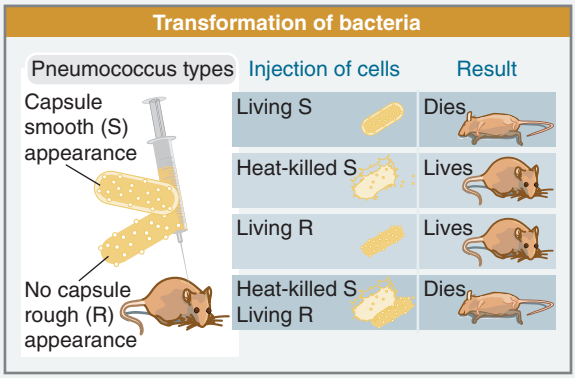


FIGURE A Neither heat-killed S-type nor live R-type bacteria can kill mice, but simultaneous injection of both can kill mice just as effectively as the live S-type.

(*continues*)

HISTORICAL PERSPECTIVES *(continued)*

Determining That DNA Is the Genetic Material

Proteins are by far the most abundant macromolecules in living cells. Their essential components, the amino acids, come in 20 varieties and can be joined in seemingly endless combinations to result in molecules with a tremendous variety of size, shape, and function. By contrast, DNA is a relatively minor component of cells and was known at the time to comprise only four nucleotide types. These nucleotides were thought to be arranged in a specific, repetitive pattern, which contributed to the notion that DNAs were molecules of "monotonous uniformity" that were unlikely to have any biological specificity. And so, it was universally assumed that the transforming principle was a protein.

In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty published the first report on the chemical nature of the transforming principle, and their results were startling. The main component of their active sample of purified transforming principle was not protein at all; it was DNA. To show that the transforming activity of this sample was due to the DNA and not to some minor but active contaminant, the authors treated it with enzymes that degrade protein, RNA, and carbohydrate; in all cases the sample retained the ability to stably transform R bacteria into S bacteria. The authors concluded that the transforming principle "consists principally, if not solely, of a highly polymerized, viscous form of deoxyribonucleic acid."

Avery and his colleagues strengthened their claim that DNA was the transforming (or genetic) material by later showing that DNase I, an enzyme that specifically degrades DNA, completely eliminated the transforming activity of their purified sample, but the scientific community remained skeptical. It was not until 1952, when Alfred D. Hershey and Martha Chase studied the independent roles of protein and DNA in bacteriophage T2 infection, that the notion of DNA as the genetic material finally gained favor (see the section of this chapter titled "DNA Is the Genetic Material of Bacteria, Viruses, and Eukaryotic Cells").

Bacteriophage T2 is a virus that infects bacterial cells in discrete steps: the phage particle attaches itself to the bacterial cell wall, phage material is injected into the cell, the host cell is induced to produce new phage particles, and, finally, the bacterial cell bursts, releasing hundreds of new phage particles. To follow the movement of phage protein and phage DNA during this process, Hershey and Chase radioactively labeled each macromolecule separately: the protein with a radioactive form of sulfur and the DNA with a radioactive form of phosphorous. In their experiment, the majority of the phosphorous label entered the bacterial cell during infection, whereas the majority of the sulfur label stayed outside the cell. The authors went on to show that the new phage particles produced by the infected bacterial cell contained a large percentage of the labeled DNA but virtually none of the labeled protein (see Figure 1.4). These results indicated that it was DNA from the infecting phage, not protein, that entered the bacterial cell and induced genetic changes.

purine Double-ringed nitrogenous base, such as adenine or guanine.

pyrimidine Single ringed nitrogenous base, such as cytosine, thymine, or uracil.

nucleoside Molecule consisting of a purine or pyrimidine base linked to the 1′ carbon of a pentose sugar.

nucleotide Molecule consisting of a purine or pyrimidine base linked to the 1′ carbon of a pentose sugar and a phosphate group linked to either the 5′ or 3′ carbon of the sugar.

1.3 Polynucleotide Chains: Nitrogenous Bases and Sugar–Phosphate Backbone

The essential component of nucleic acids (DNA and RNA) is the nucleotide, which has three parts:

- a nitrogenous base,
- a sugar, and
- one or more phosphates.

The nitrogenous base is a **purine** or **pyrimidine** ring. The base is linked to the 1′ ("one prime") carbon on a pentose sugar by a glycosidic bond from the N_1 of pyrimidines or the N_9 of purines. The pentose sugar linked to a nitrogenous base is called a **nucleoside**. Nucleic acids are named for the type of sugar: DNA has a 2′-deoxyribose, whereas RNA has a 2′-ribose. The difference is that the sugar in RNA has a hydroxyl (-OH) group on the 2′ carbon of the pentose ring. The sugar can be linked by its 5′ or 3′ carbon to a phosphate group. A nucleoside linked to a phosphate is a **nucleotide**.

A **polynucleotide** is a long chain of nucleotides. **FIGURE 1.6** shows that the backbone of the polynucleotide chain consists of an alternating series of pentose (sugar) and phosphate residues. The chain is formed by linking the

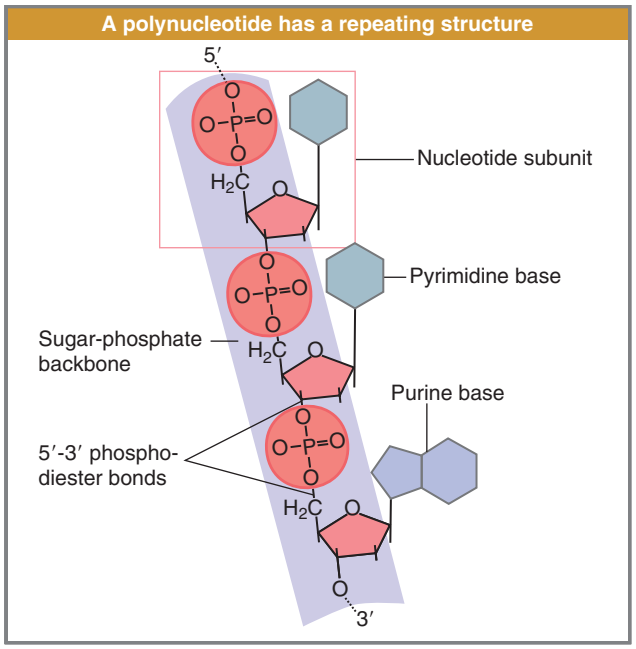


FIGURE 1.6 A polynucleotide chain consists of a series of 5′ to 3′ sugar–phosphate links that form a backbone from which the bases protrude.

5′ carbon of one pentose ring to the 3′ carbon of the next pentose ring via a phosphate group, so the sugar–phosphate backbone is said to consist of 5′ to 3′ phosphodiester linkages. The nitrogenous bases "stick out" from the backbone.

Each nucleic acid contains four types of nitrogenous base. The same two purines, adenine (A) and guanine (G), are present in both DNA and RNA. The two pyrimidines in DNA are cytosine (C) and thymine (T); in RNA, uracil (U) is found instead of thymine. The only difference between uracil and thymine is the presence of a methyl group at position C_5.

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

polynucleotide A chain of nucleotides connected by phosphodiester bonds between the 3′ carbon of one nucleotide and the 5′ carbon of the next nucleotide.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

List the structural differences between DNA and RNA nucleotides.

1.4 DNA Is a Double Helix

By the 1950s, an observation by Erwin Chargaff that the nitrogenous bases are present in different amounts in the genomes of different species led to the concept that the sequence of bases is the form in which genetic information is carried. Given this concept, there were two remaining challenges: working out the structure of DNA and explaining how a sequence of bases in DNA could determine the sequence of amino acids in a protein.

Three pieces of evidence contributed to the construction of the double-helix model for DNA by James Watson and Francis Crick in 1953:

- X-ray diffraction data collected by Rosalind Franklin and Maurice Wilkins showed that the B-form of DNA (crystallized under humid conditions) is a regular helix, making a complete turn every 34 Å (3.4 nm), with a diameter of roughly 20 Å (2 nm). Because the distance between adjacent nucleotides is 3.4 Å (0.34 nm), there must be 10 nucleotides per turn.
- The density of DNA suggests that the helix must contain two polynucleotide chains. The constant diameter of the helix can be explained if the bases in each chain face inward and are restricted so that a purine is always paired with a pyrimidine, avoiding partnerships of purine–purine (which would be too wide) or pyrimidine–pyrimidine (which would be too narrow).
- Chargaff also observed that regardless of the absolute amounts of each base, the proportion of G is always the same as the proportion of C in DNA and the proportion of A is always the same as that of T. Consequently, the composition of any DNA can be described by its G-C content, or the sum of the proportions of G and C bases. (The proportions of A and T bases can be determined by subtracting the G-C content from 1.) G-C content ranges from 0.26 to 0.74 for different species.

Watson and Crick proposed that the two polynucleotide chains in the double helix associate by hydrogen bonding between the nitrogenous bases. Normally, G can hydrogen bond specifically only with C, whereas A can bond specifically only with T. This hydrogen bonding between bases is described as *base pairing*, and the paired bases (G forming three hydrogen bonds with C, or A forming two hydrogen bonds with T) are said to be **complementary**. Base pairing occurs because of the complementary shapes of the correctly matched bases at the interfaces of where they pair, along with the location of just the right functional groups in just the right geometry along those interfaces so that hydrogen bonds can form.

complementary Base pairs that match up in the pairing reactions in double helical nucleic acids (A with T in DNA or with U in RNA, and C with G).

The Watson–Crick model has the two polynucleotide chains running in opposite directions, so they are said to be **antiparallel**, as illustrated in **FIGURE 1.7**. Looking in one direction along the helix, one strand runs in the 5′ to 3′ direction, whereas its complement runs 3′ to 5′. (Double-stranded DNA is often called *duplex DNA*, and is sometimes abbreviated as *dsDNA*.)

antiparallel Strands of the DNA double helix organized in opposite orientation, so that the 5′ end of one strand is aligned with the 3′ end of the other strand.

The sugar–phosphate backbones are on the outside of the double helix and carry negative charges on the phosphate groups. When DNA is in solution *in vitro*, the charges are neutralized by the binding of metal ions, typically sodium (Na^+). In the cell, positively charged proteins also provide some of the neutralizing force. These proteins play important roles in determining the organization of DNA in the cell.The base pairs (often abbreviated as bp) are on the inside of the double helix. They are flat and lie perpendicular to the axis of the helix. Using the analogy of the double helix as a spiral staircase, the base pairs form the steps, as illustrated schematically in **FIGURE 1.8**. Proceeding up the helix, bases are stacked above one another like a pile of plates.

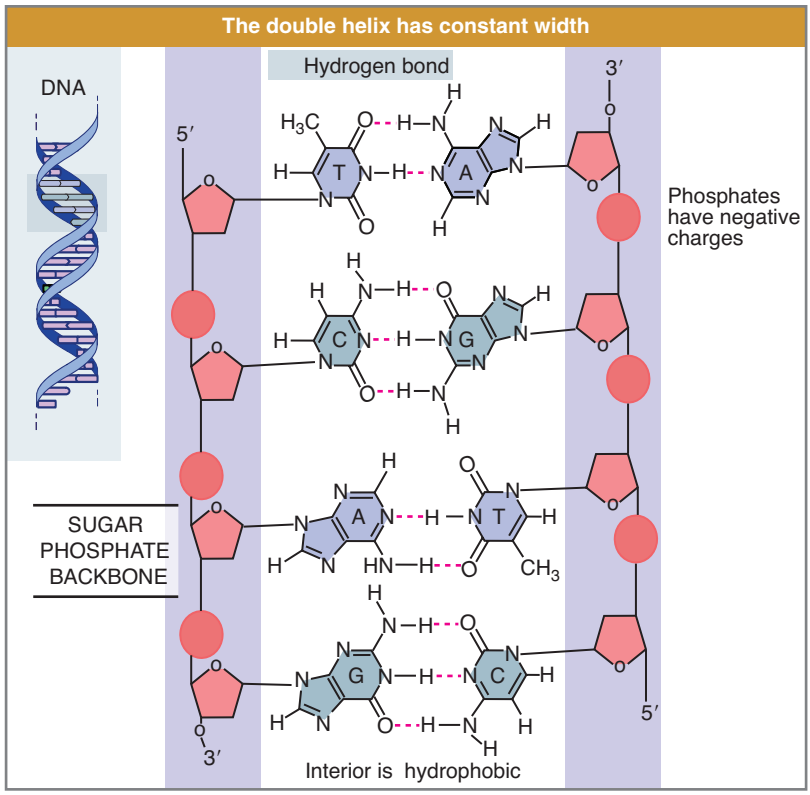


FIGURE 1.7 The DNA double helix maintains a constant width because purines always face pyrimidines in the complementary A-T and G-C base pairs. The sequence in the figure is T-A, C-G, A-T, G-C.

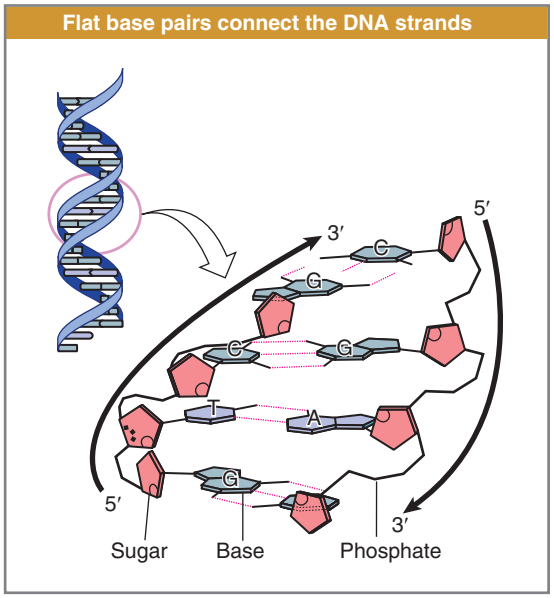


FIGURE 1.8 Flat base pairs lie perpendicular to the sugar–phosphate backbone.

Each base pair is rotated about 36° around the axis of the helix relative to the next base pair, so roughly 10 base pairs (in solution, 10.4 bp) make a complete turn of 360°. The twisting of the two strands around one another forms a double helix with a **minor groove** that is about 12 Å (1.2 nm) across and a **major groove** that is about 22 Å (2.2 nm) across, as can

minor groove Fissure running the length of the DNA double helix that is 12 Å across.

major groove Fissure running the length of the DNA double helix that is 22 Å across.

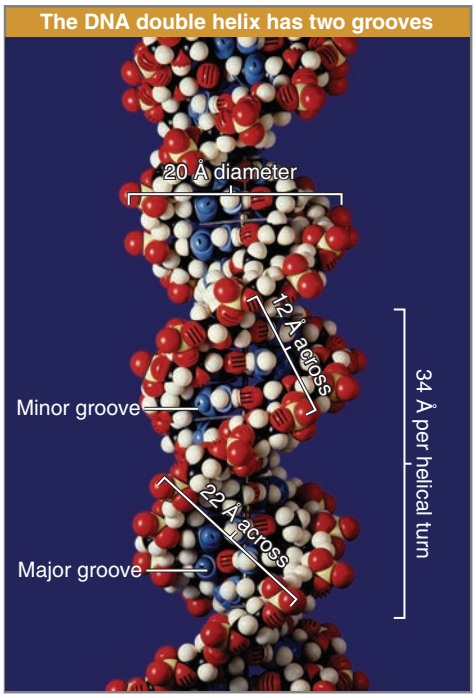


FIGURE 1.9 The two strands of DNA form a double helix.

be seen from the scale model of **FIGURE 1.9**. In B-DNA, the double helix is said to be "right-handed"; the turns run clockwise as viewed along the helical axis.

(The A-form of DNA, found in the absence of water, is also a right-handed helix that is shorter and thicker than the B-form. A third DNA structure, Z-DNA, is longer and narrower than the B-form and is a left-handed helix.)

overwound form DNA that has more than 10.4 base pairs per turn of the helix.

underwound form DNA that has fewer than 10.4 base pairs per turn of the helix.

It is important to realize that the Watson–Crick model of the B-form represents an average structure and that there can be local variations in the precise structure due to sequence. If the double helix has more base pairs per turn than the B-form, it is said to be **overwound**; if it has fewer base pairs per turn, it is **underwound**. The degree of local winding can be affected by the overall conformation of the DNA double helix or by the binding of proteins to specific sites on the DNA.

KEY CONCEPTS

- The B-form of DNA is a double helix consisting of two polynucleotide chains that run antiparallel.
- The nitrogenous bases of each chain are flat purine or pyrimidine rings that face inward and pair with one another by hydrogen bonding to form only A-T or G-C pairs.
- The diameter of the double helix is 20 Å, and there is a complete turn every 34 Å, with 10 base pairs per turn (10.4 bp per turn in solution).
- The double helix has a major (wide) groove and a minor (narrow) groove.

CONCEPT AND REASONING CHECKS

1. Summarize the evidence that Watson and Crick used in proposing their model of the B-form of DNA.
2. If the G-C content of a DNA duplex is 0.44, what are the proportions of each of the four bases?

1.5 Supercoiling Affects the Structure of DNA

The two strands of DNA are wound around each other to form the double helical structure; the double helix can also wind around itself to change the overall conformation, or *topology*, of the DNA molecule in space. This is called **supercoiling**. The effect can be imagined like a rubber band twisted around itself. Supercoiling creates tension in the DNA and, therefore, can occur only if the DNA has no free ends (otherwise, the free ends can rotate to relieve the tension) or in linear DNA if it is anchored to a protein scaffold, as in eukaryotic chromosomes. The simplest example of a DNA with no free ends is a circular molecule. The effect of supercoiling can be seen by comparing the non-supercoiled circular DNA lying flat (**FIGURE 1.10**, center) with the supercoiled circular molecule that forms a twisted (and, therefore, more condensed) shape (Figure 1.10, bottom).

supercoiling The coiling of a closed duplex DNA in space so that it crosses over its own axis.

The consequences of supercoiling depend on whether the DNA is wrapped around itself in the same direction as the two strands within the double helix (clockwise) or in the opposite direction. Wrapping in the same direction produces *positive supercoiling*, which overwinds the DNA so that there are fewer base pairs per turn than in the B-form. Wrapping in the opposite direction produces *negative supercoiling*, or underwinding, so there are more base pairs per turn than in the B-form. Both types of supercoiling of the double helix in space are tensions in the DNA (which is why DNA molecules with no supercoiling are said to be "relaxed"). Negative supercoiling can be thought of as creating tension in the DNA that is relieved by the unwinding of the double helix. The effect of severe negative supercoiling is to generate a region in which the two strands of DNA have separated (technically, zero base pairs per turn).

Topological manipulation of DNA is a central aspect of all its functional activities (recombination, replication, and transcription), as well as of the

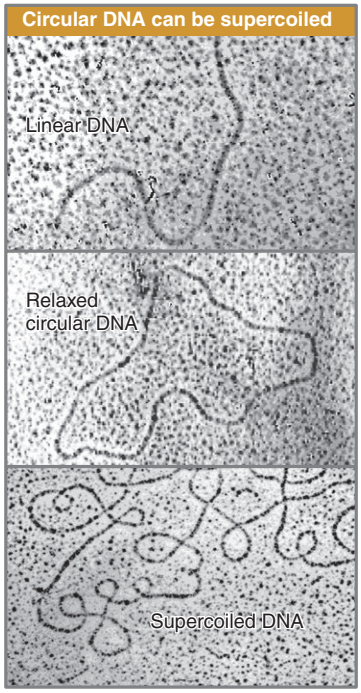


FIGURE 1.10 Linear DNA is extended (top); a circular DNA remains extended if it is relaxed (non-supercoiled) (center), but a supercoiled DNA has a twisted and condensed form (bottom).

Photos courtesy of Nirupam Roy Choudhury, International Centre for Genetic Engineering and Biotechnology (ICGEB).

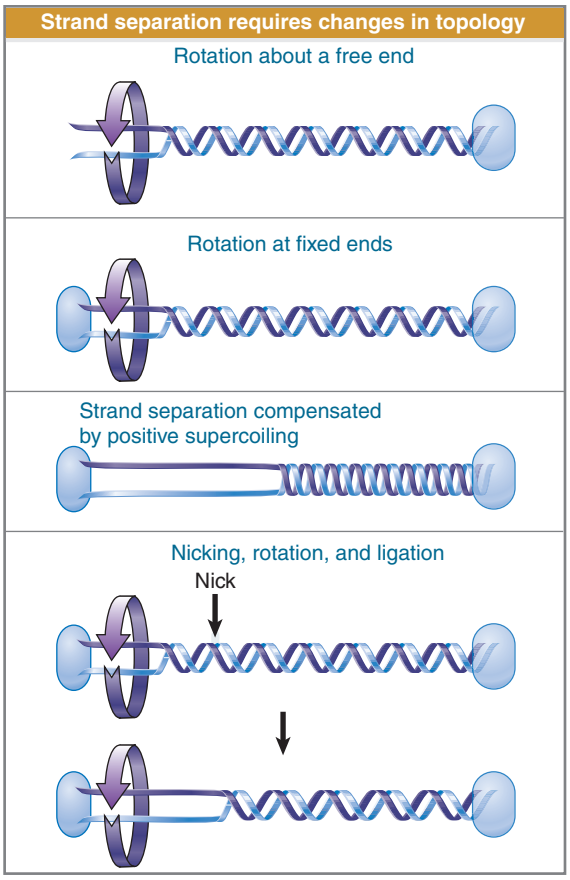


FIGURE 1.11 Separation of the strands of a DNA double helix can be achieved in several ways.

organization of its higher-order structure. All synthetic activities involving double-stranded DNA require the strands to separate. The strands do not simply lie side by side, though; they are intertwined. Their separation therefore requires the strands to rotate about each other in space. Some possibilities for the unwinding reaction are illustrated in **FIGURE 1.11**.

Unwinding a short linear DNA presents no problems, as the DNA ends are free to spin around the axis of the double helix to relieve any tension. However, DNA in a typical chromosome is not only extremely long but is also coated with proteins that serve to anchor the DNA at numerous points. Therefore, even a linear eukaryotic chromosome does not functionally possess free ends.

Consider the effects of separating the two strands in a molecule whose ends are not free to rotate. When two intertwined strands are pulled apart from one end, the result is to *increase* their winding about each other farther along the molecule, resulting in positive supercoiling elsewhere in the molecule to balance the underwinding generated in the single-stranded region. The problem can be overcome by introducing a transient nick in one strand. An internal free end allows the nicked strand to rotate about the intact strand, after which the nick can be sealed. Each repetition of the nicking and sealing reaction releases one superhelical turn. The topoisomerase enzymes that perform these reactions to control supercoiling in the cell will be discussed in the section titled "Topoisomerases Relax or Introduce Supercoils in DNA" in the "Homologous, Somatic and Site-Specific Recombination" chapter.

KEY CONCEPTS

- Supercoiling occurs only in "closed" DNA with no free ends.
- Closed DNA is either circular DNA or linear DNA in which the ends are anchored so that they are not free to rotate.

CONCEPT AND REASONING CHECK

Why does negative supercoiling facilitate unwinding of DNA but positive supercoiling inhibit unwinding?

1.6 DNA Replication Is Semiconservative

It is crucial that DNA be reproduced accurately. The two polynucleotide strands are joined only by hydrogen bonds, so they are able to separate without breaking covalent bonds. The specificity of base pairing allows both of the separated parental strands to act as template strands for the synthesis of complementary daughter strands. **FIGURE 1.12** shows that each new daughter strand is assembled using the information from one parental strand. The sequence of the daughter strand is determined by the parental strand: an A in the parental strand causes a T to be placed in the daughter strand, a parental G directs incorporation of a daughter C, and so on.

The top part of Figure 1.11 shows an unreplicated parental duplex with the original two parental strands. The lower part shows the two daughter duplexes produced by complementary base pairing. Each of the daughter duplexes is identical in sequence to the original parent duplex, containing one parental strand and one newly synthesized strand. The structure of DNA carries the information needed for its own replication. The consequences of this mode of replication, called **semiconservative replication**, are illustrated in **FIGURE 1.13**. The unit conserved from one generation to the next is one of the two individual strands comprising the parental duplex.

semiconservative replication DNA replication accomplished by separation of the strands of a parental duplex, with each strand then acting as a template for synthesis of a complementary strand.

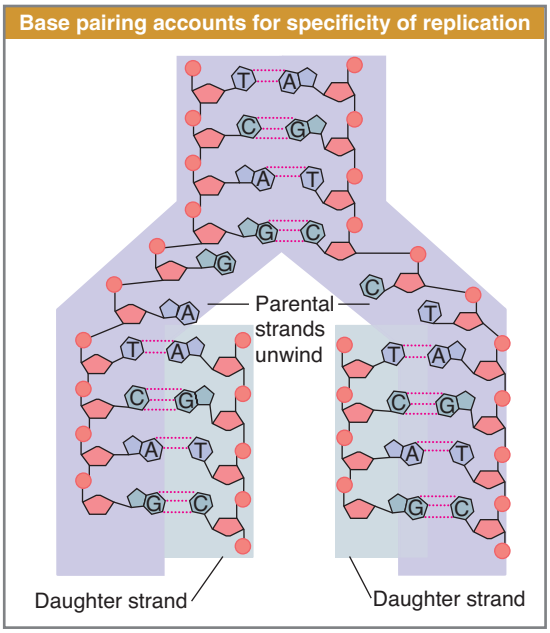


FIGURE 1.12 Base pairing provides the mechanism for replicating DNA.

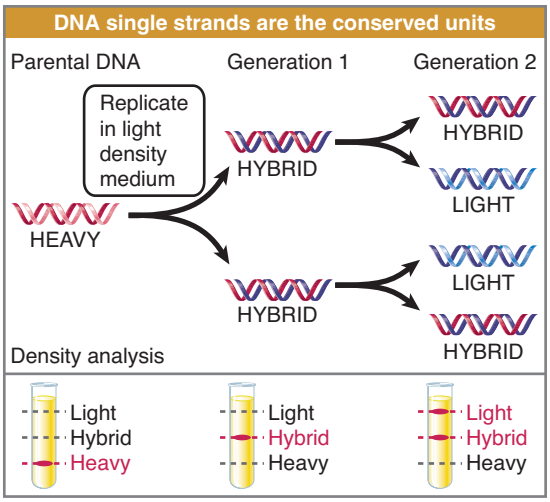


FIGURE 1.13 Replication of DNA is semiconservative.

Figure 1.13 illustrates a prediction that was originally used to test the semiconservative model. If the parental DNA carries a "heavy" density label because the organism has been grown in medium containing a suitable isotope (such as ^{15}N), its strands can be distinguished from those that are synthesized when the organism is transferred to a medium containing "light" isotopes. The parental DNA is a duplex of two "heavy" strands (red). After one generation of growth in "light" medium, the duplex DNA is *hybrid* in density—it consists of one heavy parental strand (red) and one light daughter strand (blue). After a second generation, the two strands of each hybrid duplex have separated. Each strand gains a light partner, so that now one half of the duplex DNA remains hybrid and the other half is entirely light (both strands are blue).

The individual strands of these duplexes are entirely heavy or entirely light. This pattern was confirmed experimentally by Matthew Meselson and Franklin Stahl in 1958. Meselson and Stahl followed the semiconservative replication of DNA through three generations of growth of *E. coli*. When DNA was extracted from bacteria and separated in a density gradient by centrifugation, the DNA formed bands corresponding to its density—heavy for parental, hybrid for the first generation, and half-hybrid bands and half-light bands in the second generation, indicating that a single parental strand is retained in the daughter molecule. (See the box titled "Historical Perspectives: The Meselson-Stahl Experiment" in the chapter titled "The Replicon: Initiation of Replication" for more detail on this experiment.)

KEY CONCEPTS

- The Meselson–Stahl experiment used "heavy" isotope labeling to show that the single polynucleotide strand is the unit of DNA that is conserved during replication.
- Each strand of a DNA duplex acts as a template for synthesis of a daughter strand.
- The sequences of the daughter strands are determined by complementary base pairing with the separated parental strands.

CONCEPT AND REASONING CHECK

What is the expected result of an experiment similar to that of Meselson and Stahl, but beginning with "light" DNA and culturing cells in a "heavy" medium?

1.7 Polymerases Act on Separated DNA Strands

Replication requires the two strands of the parental duplex to undergo separation, or **denaturation** (this is sometimes called "melting"). The disruption of the duplex, however, is only transient and is reversed, or undergoes **renaturation**, as the daughter duplex is formed. Only a small stretch of the duplex DNA is denatured at any moment during replication.

The helical structure of a molecule of DNA during replication is illustrated in **FIGURE 1.14**. The unreplicated region consists of the parental duplex, opening into the replicated region where the two daughter duplexes have formed. The duplex is disrupted at the junction between the two regions, which is called the **replication fork**. Replication involves movement of the replication fork along the parental DNA, so that there is continuous denaturation of the parental strands and formation of daughter duplexes.

The synthesis of DNA is aided by specific enzymes, **DNA polymerases**, that recognize the template strand and catalyze the addition of nucleotide subunits to the polynucleotide chain that is being synthesized. They are accompanied in DNA replication by ancillary enzymes such as helicases that unwind the DNA duplex, a primase that synthesizes an RNA primer required by DNA polymerase, and DNA ligase that connects discontinuous DNA fragments. Degradation of nucleic acids also requires specific enzymes: deoxyribonucleases (**DNases**) degrade DNA, and ribonucleases (**RNases**) degrade RNA. The nucleases (see the chapter titled "Methods in Molecular Biology and Genetic Engineering") fall into the general classes of **exonucleases** and **endonucleases**:

- Endonucleases break individual phosphodiester linkages within RNA or DNA molecules, generating discrete fragments. Some DNases cleave both strands of a duplex DNA at the target site, whereas others cleave only one of the two strands. Endonucleases are involved in cutting reactions, as shown in **FIGURE 1.15**.

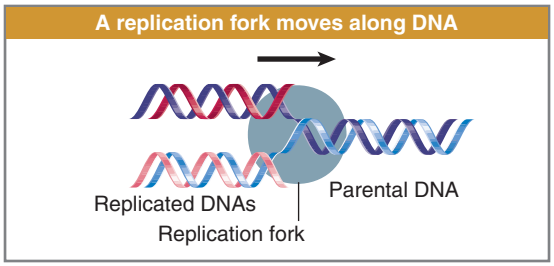


FIGURE 1.14 The replication fork is the region of DNA in which there is a transition from the unwound parental duplex to the newly replicated daughter duplexes.

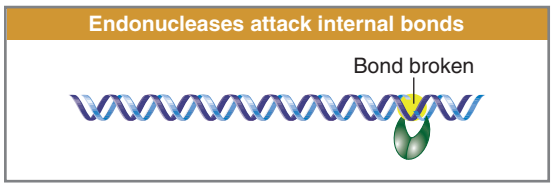


FIGURE 1.15 An endonuclease cleaves a bond within a nucleic acid. This example shows an enzyme that attacks one strand of a DNA duplex.

denaturation A molecule's conversion from the physiological conformation to some other (inactive) conformation. In DNA, this involves the separation of the two strands due to breaking of hydrogen bonds between bases.renaturation The reassociation of denatured complementary single strands of a DNA double helix.

replication fork The point at which strands of parental duplex DNA are separated so that replication can proceed. A complex of proteins, including DNA polymerase, is found there.

DNA polymerase An enzyme that synthesizes a daughter strand(s) of DNA (under direction from a DNA template). Any particular enzyme may be involved in repair or replication (or both).

DNase An enzyme that degrades DNA.

RNase An enzyme that degrades RNA.

exonuclease An enzyme that cleaves nucleotides one at a time from the end of a polynucleotide chain; it may be specific for either the 5′ or 3′ end of DNA or RNA.

endonuclease An enzyme that cleaves bonds within a nucleic acid chain; it may be specific for RNA or for single-stranded or double-stranded DNA.

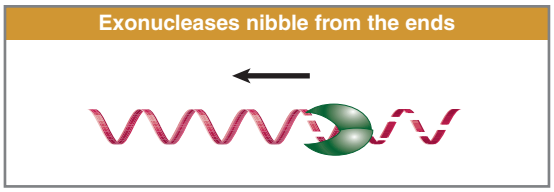


FIGURE 1.16 An exonuclease removes bases one at a time by cleaving the last bond in a polynucleotide chain.

- Exonucleases remove nucleotide residues one at a time from the end of the molecule, generating mononucleotides. They always function on a single nucleic acid strand, and each exonuclease proceeds in a specific direction, that is, starting either at a 5′ or at a 3′ end and proceeding toward the other end. They are involved in trimming reactions, as shown in **FIGURE 1.16**.

KEY CONCEPTS

- Replication of DNA is undertaken by a complex of enzymes that separate the parental strands and synthesize the daughter strands.
- The replication fork is the point at which the parental strands are separated.
- The enzymes that synthesize DNA are called DNA polymerases.
- Nucleases are enzymes that degrade nucleic acids; they include DNases and RNases and can be categorized as endonucleases or exonucleases.

CONCEPT AND REASONING CHECK

What are the functions of DNA polymerases, DNases, and RNases in living cells?

1.8 Genetic Information Can Be Provided by DNA or RNA

central dogma Information cannot be transferred from polypeptide to polypeptide or polypeptide to nucleic acid, but can be transferred between nucleic acids and from nucleic acid to polypeptide.

RNA polymerase An enzyme that synthesizes RNA using a DNA template (formally described as DNA-dependent RNA polymerases).

The **central dogma** is the dominant paradigm of molecular biology. Protein-coding genes exist as sequences of nucleic acid, but they function by being expressed in the form of polypeptides. Replication makes possible the inheritance of genetic information, while transcription and translation are responsible for its expression to another form. The central dogma includes several observations about the processes of replication, transcription, and translation (see Figure 1.30):

- Transcription of DNA by a DNA-dependent **RNA polymerase** generates RNA molecules. Messenger RNAs (mRNAs) are translated to polypeptides. Other types of RNA, such as rRNAs and tRNAs, are functional themselves and are not translated.
- A genetic system may involve either DNA or RNA as the genetic material. Cells use only DNA. Some viruses use RNA, and replication of viral RNA by an RNA-dependent RNA polymerase occurs in the infected cell.
- The expression of cellular genetic information is usually unidirectional. Transcription of DNA generates RNA molecules; the exception is the reverse transcription of retroviral RNA to DNA that occurs when retroviruses infect cells (see below). Generally polypeptides cannot be retrieved for use as genetic information; translation of RNA into polypeptide is always irreversible.

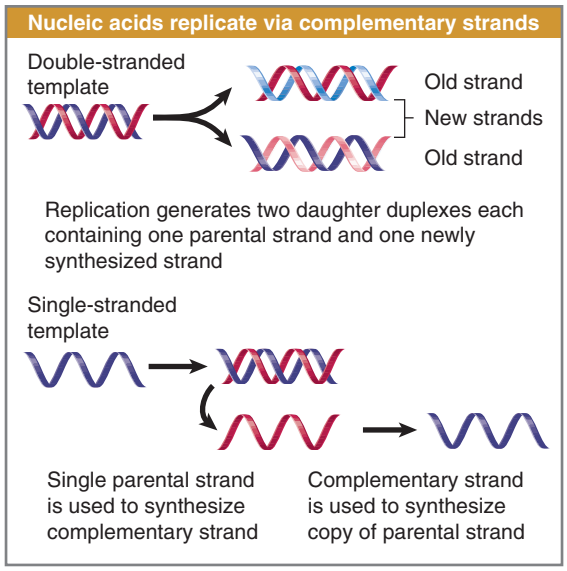


FIGURE 1.17 Double-stranded and single-stranded nucleic acids both replicate by synthesis of complementary strands governed by the rules of base pairing.

These mechanisms are equally effective for the cellular genetic information of prokaryotes or eukaryotes and for the information carried by viruses. The genomes of all living organisms consist of duplex DNA. Viruses have genomes that consist of DNA or RNA, and there are examples of each type that are double-stranded (dsDNA or dsRNA) or single-stranded (ssDNA or ssRNA). Details of the mechanism used to replicate the nucleic acid vary among viruses, but the principle of replication via synthesis of complementary strands remains the same, as illustrated in **FIGURE 1.17**. The restriction of a unidirectional transfer of information from DNA to RNA in cells is not absolute. It is broken by the retroviruses, which have genomes consisting of a single-stranded RNA molecule. During the retroviral cycle of infection, the RNA is converted into a single-stranded DNA by the process of **reverse transcription**, which is accomplished by the enzyme *reverse transcriptase*, an RNA-dependent DNA polymerase. The resulting ssDNA is in turn converted into dsDNA. This duplex DNA becomes part of the genome of the host cell and is inherited like any other gene. So reverse transcription allows a sequence of RNA to be retrieved and used as DNA in a cell.

reverse transcription Synthesis of DNA from a template of RNA by the enzyme reverse transcriptase.

The existence of RNA replication and reverse transcription establishes the general principle that information in the form of either type of nucleic acid sequence can be converted into the other type. In the usual course of events, however, the cell relies on the processes of DNA replication, transcription, and translation. But on rare occasions (possibly mediated by an RNA virus), information from a cellular RNA is converted into DNA and inserted into the genome. Although retroviral reverse transcription is not necessary for the regular operations of the cell, it becomes a mechanism of potential importance when we consider the evolution of the genome (see the chapter titled "Genome Sequences and Evolution").

The same principles for the perpetuation of genetic information apply to the massive genomes of plants or amphibians, as well as the tiny genomes of mycoplasma and the even smaller genomes of DNA or RNA viruses. **FIGURE 1.18** presents some examples that illustrate the range of genome types and sizes. The reasons for such variation in genome size and gene

Genomes vary greatly in size		
Genome	Gene Number	Base Pairs
Organisms		
Plants	<50,000	$<10^{11}$
Mammals	30,000	~3 x 10^9
Worms	14,000	~10^8
Flies	12,000	1.6 x 10^8
Fungi	6,000	1.3 x 10^7
Bacteria	2–4,000	$<10^7$
Mycoplasma	500	$<10^6$
dsDNA Viruses		
Vaccinia	<300	187,000
Papova (SV40)	~6	5,226
Phage T4	~200	165,000
ssDNA Viruses		
Parvovirus	5	5,000
Phage fX174	11	5,387
dsRNA Viruses		
Reovirus	22	23,000
ssRNA Viruses		
Coronavirus	7	20,000
Influenza	12	13,500
TMV	4	6,400
Phage MS2	4	3,569
STNV	1	1,300
Viroids		
PSTV RNA	0	359

FIGURE 1.18 The size of the genome varies over an enormous range.

number will be explored in the "Content of the Genome" and "Genome Sequences and Evolution" chapters.

Among the various living organisms, with genomes varying in size over a 100,000-fold range, a common principle prevails: the DNA encodes all of the polypeptides that the cell(s) of the organism must synthesize, and the polypeptides in turn (directly or indirectly) provide the functions needed for survival. A similar principle describes the function of the genetic information of viruses, whether DNA or RNA: the nucleic acid encodes the polypeptide(s) needed to package the genome and for any other functions in addition to those provided by the host cell that are needed to reproduce the virus. (The smallest virus—the satellite tobacco necrosis virus [STNV]—cannot replicate independently. It requires the presence of a "helper" virus—the tobacco necrosis virus [TNV], which is itself a normally infectious virus.)

KEY CONCEPTS

- Cellular genomes are DNA, but viruses may have genomes of RNA.
- DNA is converted into RNA by transcription, and RNA may be converted into DNA by reverse transcription.
- The translation of RNA into protein is unidirectional.

CONCEPT AND REASONING CHECK

What types of enzymes would be necessary to replicate ssDNA, ssRNA, dsDNA, and dsRNA genomes to produce exact copies of the same type of nucleic acid?

1.9 Nucleic Acids Hybridize by Base Pairing

A crucial property of the double helix is the capacity to separate the two strands without disrupting the covalent bonds that form the polynucleotides and at the (very rapid) rates needed to sustain genetic functions. The specificity of the processes of denaturation and renaturation is determined by complementary base pairing.

The concept of base pairing is central to all processes involving nucleic acids. Disruption of the base pairs is crucial to the function of a double-stranded nucleic acid, whereas the ability to form base pairs is essential for the activity of a single-stranded nucleic acid. **FIGURE 1.19** shows that base pairing enables complementary single-stranded nucleic acids to form a duplex.

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

Formation of duplex regions from single-stranded nucleic acids is most prevalent in RNA but is also important for single-stranded viral DNA genomes. Base pairing between independent complementary single strands is not restricted to DNA–DNA or RNA–RNA but can also occur between DNA and RNA.

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

Denaturation of DNA occurs over a narrow temperature range and results in striking changes in many of its physical properties. The midpoint of the temperature range over which the strands of DNA separate is called the **melting temperature** (T_m) and depends on the G-C content of the duplex. Because each G-C base pair has three hydrogen bonds, it is more stable than an A-T base pair, which has only two hydrogen bonds. The more G-C base pairs in a DNA, the greater the energy that is needed to separate the two strands. In solution under physiological conditions, a DNA that is 40% G-C

melting temperature The midpoint of the temperature range over which the strands of DNA separate.

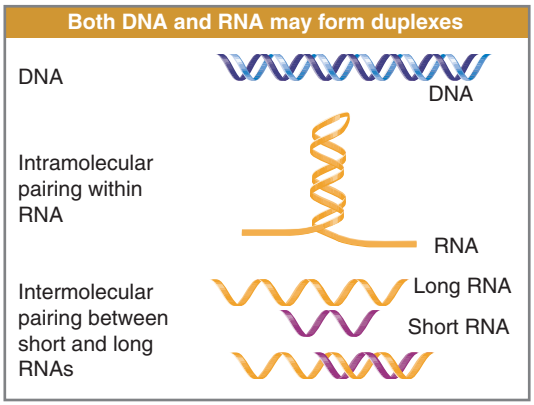


FIGURE 1.19 Base pairing occurs in duplex DNA and in intra- and intermolecular interactions in single-stranded RNA (or DNA).

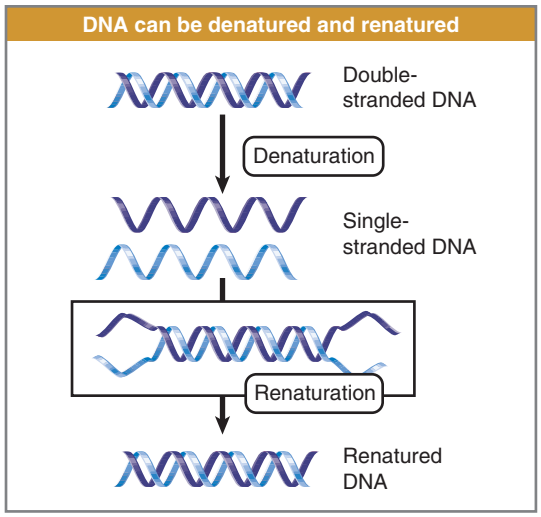


FIGURE 1.20 Denatured single strands of DNA can renature to give the duplex form.

(a value typical of mammalian genomes) denatures with a T_m of about 87°C, so duplex DNA is stable at the temperature of the cell.

The denaturation of DNA is reversible under appropriate conditions. Renaturation depends on specific base pairing between the complementary strands. **FIGURE 1.20** shows that the reaction takes place in two stages. First, single strands of DNA in the solution encounter one another by chance; if their sequences are complementary, the two strands base pair to generate a short double-stranded region. This region of base pairing then extends along the molecule, much like a zipper, to form a lengthy duplex.

annealing The renaturation of a duplex structure from single strands that were obtained by denaturing duplex DNA.

Complete renaturation restores the properties of the original double helix. The property of renaturation applies to any two complementary nucleic acid sequences. This is sometimes called **annealing**, but the reaction is more generally called **hybridization** whenever nucleic acids from different sources are involved, as in the case when DNA hybridizes to RNA. The ability of two nucleic acids to hybridize constitutes a precise test for their complementarity because only complementary sequences can form a duplex (although under certain conditions, imperfect matches can be tolerated).

hybridization The pairing of complementary RNA and DNA strands to give an RNA–DNA hybrid.

The principle of the hybridization reaction is to combine two single-stranded nucleic acids in solution and then to measure the amount of double-stranded material that forms. **FIGURE 1.21** illustrates a procedure in which a DNA preparation is denatured and the single strands are attached to a filter. Then a second denatured DNA (or RNA) preparation is added. The filter is treated so that the second preparation can attach to it only if it is able to base pair with the DNA that was originally attached. Usually, the second preparation is labeled so that the hybridization reaction can be measured as the amount of label retained by the filter. Alternatively, hybridization in solution can be measured as the change in ultraviolet (UV) absorbance of a nucleic acid solution at 260 nm as detected via spectrophotometry. As DNA denatures to single strands with increasing temperature, UV absorbance of the DNA solution increases; UV absorbance consequently decreases as ssDNA hybridizes to complementary DNA or RNA with decreasing temperature.

Two sequences need not be perfectly complementary to hybridize. If they are similar but not identical, an imperfect duplex is formed in which base pairing is interrupted at positions where the two single strands are not complementary.

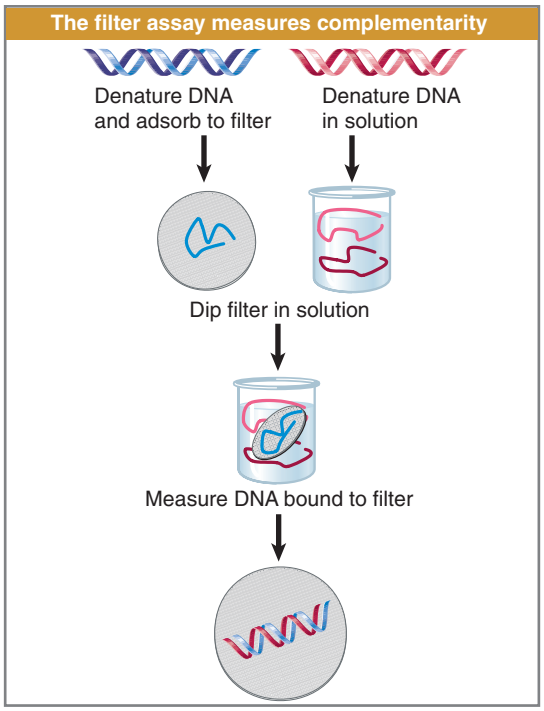


FIGURE 1.21 Filter hybridization establishes whether a solution of denatured DNA (or RNA) contains sequences complementary to the strands immobilized on the filter.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Describe how measures of hybridization between DNA from different species can be used to estimate the evolutionary relationships between those species.

1.10 Mutations Change the Sequence of DNA

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

All organisms undergo a certain number of mutations as the result of normal cellular operations or random interactions with the environment.

spontaneous mutations Mutations that occur in the absence of any added reagent to increase the mutation rate; these result from unrepaired errors in replication or random changes to the chemical structure of bases.

mutagens Substances that increase the rate of mutation by inducing changes in DNA sequence, either directly or indirectly.

induced mutations Mutations that result from the action of a mutagen. The mutagen may act directly on the bases in DNA, or it may act indirectly to trigger a pathway that leads to a change in DNA sequence.

These are called **spontaneous mutations**, and the rate at which they occur (the "background level") is characteristic for any particular organism. Mutations are rare events, and of course those that have deleterious effects are selected against during evolution. It is therefore difficult to observe large numbers of spontaneous mutants from natural populations.

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

Mutation rates can be measured at several levels of resolution: mutation across the whole genome (as the rate per genome per generation), mutation in a gene (as the rate per locus per generation), or mutation at a specific nucleotide site (as the rate per base pair per generation). These rates correspondingly decrease as smaller units are observed.

Spontaneous mutations that inactivate gene function occur in bacteriophages and bacteria at a relatively constant rate of 3 to 4×10^{-3} per genome per generation. Given the large variation in genome sizes between bacteriophages and bacteria, this corresponds to great differences in the mutation rate per base pair. This suggests that the overall rate of mutation has been subject to selective forces that have balanced the deleterious effects of most mutations against the advantageous effects of some mutations. This conclusion is strengthened by the observation that an archaean that lives under harsh conditions of high temperature and acidity (which are expected to damage DNA) does not show an elevated mutation rate but in fact has an overall mutation rate just below the average range. **FIGURE 1.22** shows that in bacteria, the mutation rate corresponds to about 10^{-6} events per locus per generation, or to an average rate of change per base pair of about 10^{-9} to 10^{-10} per generation. The rate of individual base pairs varies very widely, over a 10,000-fold range. We have no accurate measurement of the rate of mutation in eukaryotes, although usually it is thought to be somewhat similar to that

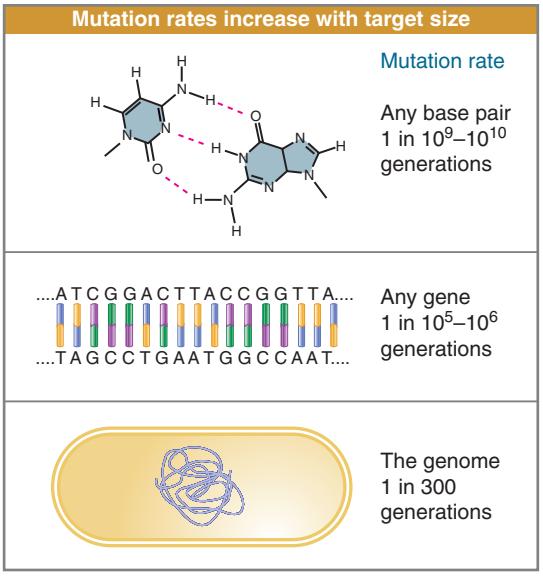


FIGURE 1.22 A base pair is mutated at a rate of 10^{-9} to 10^{-10} per generation, a gene of 1,000 bp is mutated at about 10^{-6} per generation, and a bacterial genome is mutated at 3×10^{-3} per generation.

of bacteria on a per locus per generation basis. One reason that mutation rates vary across species is that the activity and efficacy of DNA repair systems also vary. DNA repair systems will be discussed in the "Repair Systems" chapter.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

What are the advantages of maintaining a nonzero mutation rate?

1.11 The Effects of Mutations

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

point mutation A change in the sequence of DNA involving a single base pair.

- Chemical modification of DNA directly changes one base into a different base.
- An error during the replication of DNA causes the wrong base to be inserted into a polynucleotide.

Point mutations can be divided into two types, depending on the nature of the base substitution:

- The most common class is the **transition**, resulting from the substitution of one pyrimidine by the other or of one purine by the other. If the substitution is not repaired, after a DNA replication event, a G-C pair is replaced with an A-T pair, or vice versa.
- The less common class is the **transversion**, in which a purine is replaced by a pyrimidine, or vice versa, so that following replication (for example), an A-T pair becomes a T-A or C-G pair.

transition Mutation in which one pyrimidine is replaced by the other or in which one purine is replaced by the other.

transversion Mutation in which a purine is replaced by a pyrimidine, or vice versa.

As shown in **FIGURE 1.23**, the mutagen nitrous acid performs an oxidative deamination that converts cytosine into uracil, resulting in a transition. In the replication cycle following the transition, the U pairs with an A instead of the G with which the original C would have paired. So the C-G pair is replaced by a T-A pair when the A pairs with the T in the next replication cycle. (Nitrous acid can also deaminate adenine, causing the reverse transition from A-T to G-C.)

Transitions are also caused by base mispairing, when noncomplementary bases pair instead of the usual Watson–Crick pairs. Base mispairing usually occurs as an aberration resulting from the incorporation into DNA of an abnormal base that has flexible pairing properties. **FIGURE 1.24** shows the example of the mutagen bromouracil (BrdU), an analog of thymine that contains a bromine atom in place of thymine's methyl group and that can be incorporated into DNA in place of thymine. However, BrdU has flexible pairing properties because the presence of the bromine atom allows a *tautomeric shift* from a keto (=O) form to an enol (-OH) form. The enol form of BrdU can pair with guanine, which after replication leads to substitution of the original A-T pair by a G-C pair. Tautomeric shifts can also occur when a proton shifts position within a normal base and results in an anomalous, but more stable,

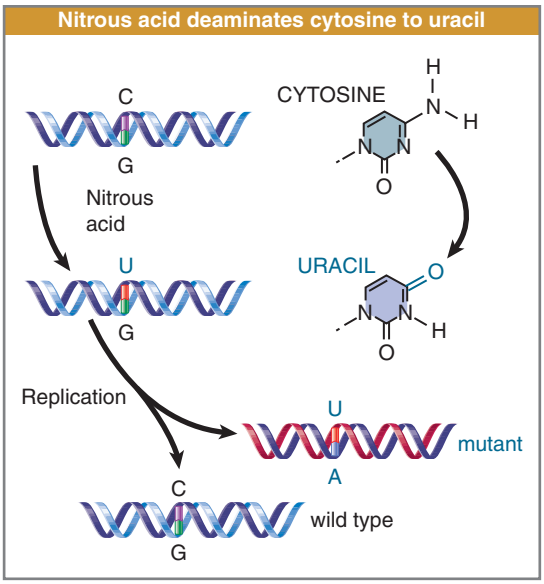


FIGURE 1.23 Mutations can be induced by chemical modification of a base.

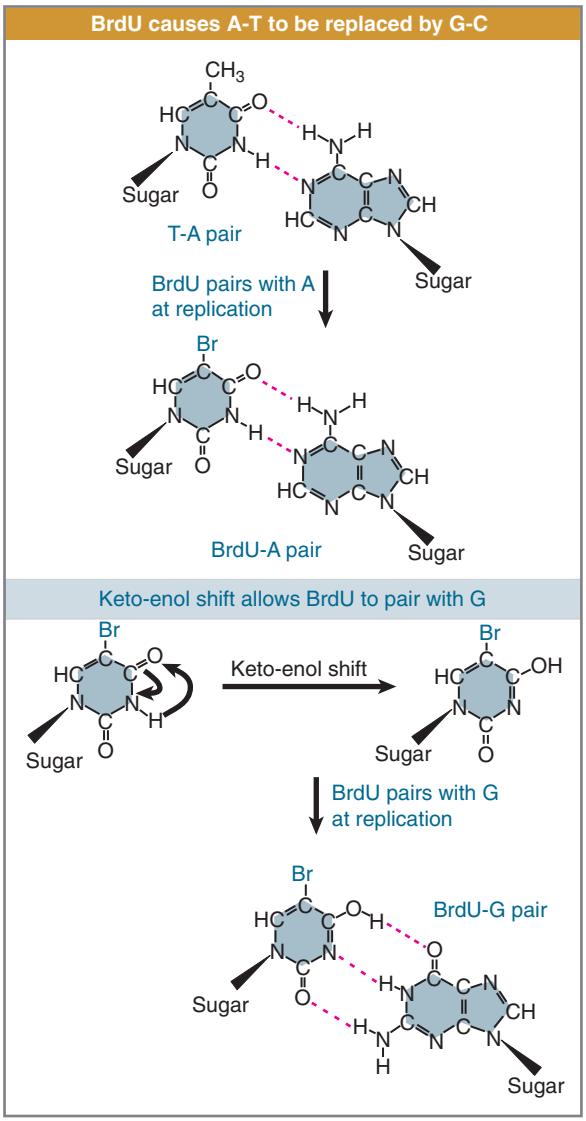


FIGURE 1.24 Mutations can be induced by the incorporation of base analogs into DNA.

base pairing. For example, while the common keto form of guanine pairs stably with cytosine, the rare enol form of guanine pairs stably with thymine.

Transversions are rarer than transitions, since they require a temporary purine–purine or pyrimidine–pyrimidine pairing that would alter the diameter of the DNA duplex. However, one cause of transversions is a proton shift followed by a 180° rotation of the base around the glycosidic bond. For example, a rotated synadenine (produced by a proton shift in adenine) will pair stably with a normal adenine.

Point mutations were once considered the principal means of change in individual genes. However, we now know that insertions and deletions ("indels") of short sequences are quite frequent. Some mutagens, such as intercalating agents, can cause the insertion or deletion of a single base pair. An intercalating agent will insert itself between two adjacent base pairs in a DNA duplex, so that the duplex is distorted and a DNA polymerase can either skip or add a base during DNA replication. If this occurs in the coding sequence of a gene, a frameshift mutation will result (see the section later in this chapter titled "The Genetic Code Is Triplet"). Often, the insertions are the result of transposable elements, which are sequences of DNA with the ability to move from one site to another (see the chapter titled "Transposable Elements and Retroviruses"). An insertion within a coding region usually abolishes the activity of the gene. However, both insertions and deletions of short sequences can occur by other mechanisms—for example, those involving errors during replication or recombination. In addition, mutagens belonging to a class called acridines introduce very small insertions and deletions.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Why are transitions more common than transversions? Consider how the DNA repair mechanisms might recognize errors and the effects of these mutations on DNA structure in your answer.

1.12 The Effects of Mutations Can Be Reversed

FIGURE 1.25 shows that the possibility of reversion mutations, or **revertants**, is an important characteristic that distinguishes point mutations and insertions from deletions. Mutations that inactivate a gene are called **forward mutations**. Their effects are reversed by **back mutations**, which are of two types: true reversions and second-site reversions.

- A point mutation can revert either by a true reversion or a second-site reversion.
- An insertion can revert by deletion of the inserted sequence.
- A deletion of a sequence cannot revert in the absence of some mechanism to restore the lost sequence.

revertants Reversions of a mutant cell or organism to the wild-type phenotype.

forward mutation A mutation that inactivates a functional gene.

back mutations A mutation that reverses the effect of a mutation in a given gene, restoring the original sequence or function of the gene product.

FIGURE 1.25 Point mutations and insertions can revert, but deletions cannot revert.

true reversion A mutation that restores the original sequence of the DNA.

An exact reversal of the original mutation is called a **true reversion**. For example, if an A-T pair was replaced by a G-C pair in the original mutation, another mutation to restore the A-T pair will exactly regenerate the original sequence. The exact removal of a transposable element following its insertion is another example of a true reversion.

second-site reversion A second mutation suppressing the effect of a first mutation.

The second type of back mutation, **second-site reversion**, may occur elsewhere in the gene, and its effects compensate for the first mutation. For example, one amino acid change in a protein may abolish its function, but a second alteration may compensate for the first and restore protein activity.

A forward mutation results from any change that alters the function of a gene product, whereas a back mutation must restore the original function to the altered gene product. Therefore, the possibilities for back mutations are much more restricted than those for forward mutations. The rate of back mutations is correspondingly lower than that of forward mutations, typically by a factor of about 10.

suppression mutation A second event eliminates the effects of a mutation without reversing the original change in DNA.

Mutations in other genes can also occur to circumvent the effects of mutation in the original gene. This is called a **suppression mutation**. A locus in which a mutation suppresses the effect of a mutation in another locus is called a *suppressor*. For example, a point mutation may cause an amino acid substitution in a polypeptide, whereas a second mutation in a tRNA gene may cause it to recognize the mutated codon and, as a result, insert the original amino acid during translation. (Note that this suppresses the original mutation but causes errors during translation of other mRNAs.)

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Transposable elements were originally identified because the rate of reversion of mutations caused by them is much higher than the reversion rate for point mutations. Explain why the reversion rate is so high.

1.13 Mutations Are Concentrated at Hotspots

So far we have presented mutations in terms of individual changes in the sequence of DNA that influence the activity of the DNA in which they occur. When we consider mutations in terms of the alteration of function of the gene, most genes within a species show more or less similar rates of mutation relative to their size. This suggests that the gene can be regarded as a target for mutation and that damage to any part of it can alter its function. As a result, susceptibility to mutation is roughly proportional to the size of the gene. But are all base pairs in a gene equally susceptible, or are some more likely to be mutated than others?

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

The histogram of **FIGURE 1.26** shows the frequency with which mutations are found at each base pair in the *lacI* gene of *E. coli*. The statistical probability that more than one mutation occurs at a particular site is given by random-hit kinetics (as seen in the Poisson distribution). Some sites will gain one, two, or three mutations, whereas others will not gain any. Some sites gain far more than the number of mutations expected from a random distribution; they may have 10 or even 100 times more mutations than predicted by

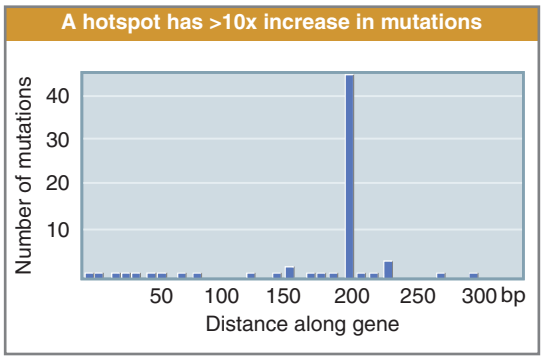


FIGURE 1.26 Spontaneous mutations occur throughout the *lacI* gene of *E. coli* but are concentrated at a hotspot.

hotspots A site in the genome at which the frequency of mutation (or recombination) is very much increased, usually by at least an order of magnitude relative to neighboring sites.

random hits. These sites are called **hotspots**. Spontaneous mutations may occur at hotspots, and different mutagens may have different hotspots.

A major cause of spontaneous mutation is the presence of an unusual base in the DNA. In addition to the four standard bases of DNA, *modified bases* are sometimes found. The name reflects their origin; they are produced by chemical modification of one of the four standard bases. The most common modified base is 5-methylcytosine, which is generated when a methylase enzyme adds a methyl group to cytosine residues at specific sites in the DNA. Sites containing 5-methylcytosine are hotspots for spontaneous point mutation in *E. coli*. In each case, the mutation is a G-C to A-T transition. The hotspots are not found in mutant strains of *E. coli* that cannot methylate cytosine.

The reason for the existence of these hotspots is that cytosine bases suffer a higher frequency of spontaneous deamination. In this reaction, the amino group is replaced by a keto group. Recall that deamination of cytosine generates uracil (see Figure 1.23). **FIGURE 1.27** compares this reaction with the deamination of 5-methylcytosine where deamination generates thymine. The effect is to generate the mismatched base pairs G-U and G-T, respectively.

FIGURE 1.28 shows that the consequences of deamination are different for 5-methylcytosine and cytosine. Deaminating the (rare) 5-methylcytosine causes a mutation, whereas deaminating cytosine does not have this effect. This happens because the DNA repair systems are much more effective in recognizing G-U than G-T and always correct the U (which normally should be present only in RNA, not in DNA).

E. coli contains an enzyme, uracil-DNA-glycosidase, that removes uracil residues from DNA (see the section titled "Base Excision Repair Systems Require Glycosylases" in the "Repair Systems" chapter). This action leaves an unpaired G residue, and a repair system then inserts a complementary C base. The net result of these reactions is to restore the original sequence of the DNA. Thus, this system protects DNA against the consequences of spontaneous deamination of cytosine. (This system is not, however, efficient enough to prevent the effects of the increased deamination caused by nitrous acid; see Figure 1.23.)

Note that the deamination of 5-methylcytosine creates thymine and results in a mismatched base pair, G-T. If the mismatch is not corrected before the next replication cycle, a mutation results; the bases in the mispaired G-T separate, and then they pair with the correct complements to produce the original G-C and the mutant A-T.

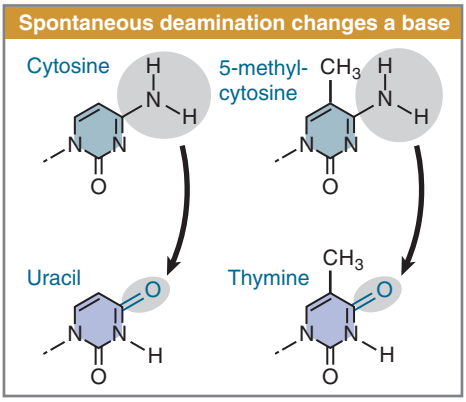


FIGURE 1.27 Deamination of cytosine produces uracil, whereas deamination of 5-methylcytosine produces thymine.

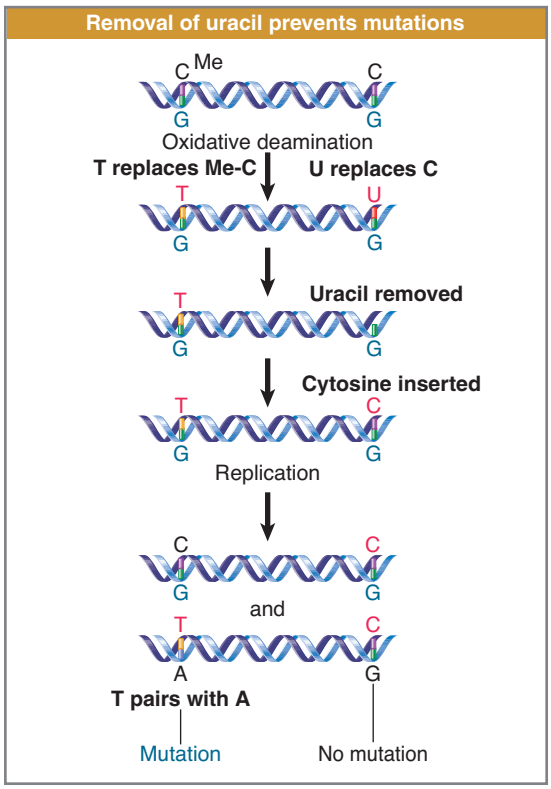


FIGURE 1.28 The deamination of 5-methylcytosine produces thymine (leading to C-G to T-A transitions), whereas the deamination of cytosine produces uracil (which usually is removed and then replaced by cytosine).

Deamination of 5-methylcytosine is the most common cause of mismatched G-T pairs in DNA. Repair systems that act on G-T mismatches have a bias toward replacing the T with a C (rather than the alternative of replacing the G with an A), which helps to reduce the rate of mutation (see the section titled "Controlling the Direction of Mismatch Repair" in the "Repair Systems" chapter). However, these systems are not as effective as those that remove U from G-U mismatches. As a result, deamination of 5-methylcytosine leads to mutation much more often than does deamination of cytosine.

5-methylcytosine also creates hotspots in eukaryotic DNA. It is common at CpG dinucleotides that are concentrated in regions called *CpG islands* (see the section titled "CpG Islands Are Subject to Methylation" in the "Epigenetics I" chapter). Although 5-methylcytosine accounts for only 1% of the bases in human DNA, sites containing the modified base account for about 30% of all point mutations.

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

Another type of hotspot, though not often found in coding regions, is the "slippery sequence"—a homopolymer run, or region where a very short sequence (one or a few nucleotides) is repeated many times in tandem. During replication, a DNA polymerase may skip one repeat or replicate the same repeat twice, leading to a decrease or increase in repeat number.

KEY CONCEPTS

- The frequency of mutation at any particular base pair is statistically equivalent, except for hotspots, where the frequency is increased by at least an order of magnitude.
- A common cause of hotspots is the modified base 5-methylcytosine, which is spontaneously deaminated to thymine.
- A hotspot can result from the high frequency of change in copy number of a short, tandemly repeated sequence.

CONCEPT AND REASONING CHECK

Suggest several possible reasons that a particular base pair can be a mutational hotspot.

1.14 Some Hereditary Agents Are Extremely Small

Viroid A small infectious nucleic acid that does not have a protein coat.

Viroids (or subviral pathogens) are infectious agents that cause diseases in higher plants. They are very small circular molecules of RNA. Unlike viruses—for which the infectious agent consists of a *virion*, a genome encapsulated in a protein coat—the viroid RNA is itself the infectious agent. The viroid consists solely of the RNA molecule, which is extensively folded by imperfect base pairing, forming a characteristic rod, as shown in **FIGURE 1.29**. Mutations that interfere with the structure of this rod reduce the infectivity of the viroid.

A viroid RNA consists of a single molecule that is replicated autonomously and accurately in infected cells. Viroids are categorized into several groups. A given viroid is assigned to a group according to sequence similarity with other members of the group. For example, four viroids in the potato spindle tuber viroid (PSTV) group have 70–83% sequence similarity with PSTV. Different isolates of a particular viroid strain vary from one another in sequence, which may result in phenotypic differences among infected cells. For example, the "mild" and "severe" strains of PSTV differ by three nucleotide substitutions.

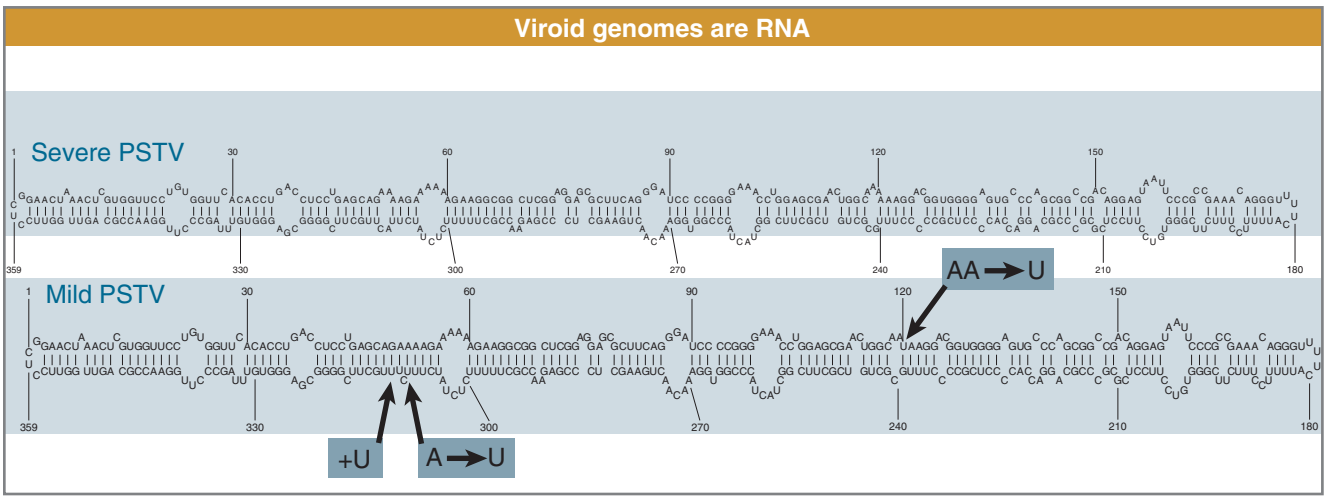


FIGURE 1.29 PSTV RNA is a circular molecule that forms an extensive double-stranded structure, interrupted by many interior loops. The severe and mild forms of PSTV have RNAs that differ at three sites.

Viroids are similar to viruses in having heritable nucleic acid genomes but differ from viruses in both structure and function. Viroid RNA does not appear to be translated into polypeptide, so it cannot itself encode the functions needed for its survival. This situation poses two as yet unanswered questions: How does viroid RNA replicate, and how does it affect the phenotype of the infected plant cell?

Replication must be carried out by enzymes of the host cell. The heritability of the viroid sequence indicates that viroid RNA is the template for replication. Thus, the modern version of the central dogma, as presented in **FIGURE 1.30**, includes replication of RNA as a component in some systems.

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

An even more unusual agent is the cause of scrapie, a degenerative neurological disease of sheep and goats. The disease is similar to the human diseases of kuru and Creutzfeldt–Jakob syndrome, which affect brain function. The infectious agent of scrapie does not contain nucleic acid. This extraordinary agent is called a **prion** (proteinaceous infectious agent). It is a 28-kD hydrophobic glycoprotein, PrP. PrP is encoded by a cellular gene (conserved among the mammals) that is expressed in normal brain cells. The protein exists in two forms: the version found in normal brain cells is called PrP^{c} and is entirely degraded by proteases during normal protein turnover. The version found in infected brains is called PrP^{sc} and is extremely resistant to degradation by proteases. PrP^{c} is converted to PrP^{sc} by a conformational change that confers protease resistance.

Prion A proteinaceous infectious agent that behaves as an inheritable trait, although it contains no nucleic acid. One example is PrP^{sc}, the agent of scrapie in sheep and bovine spongiform encephalopathy.

As the infectious agent of scrapie, PrP^{sc} must in some way modify the synthesis of its normal cellular counterpart so that it becomes infectious instead of harmless (see the section titled "Prions Cause Diseases in Mammals" in the "Epigenetics II" chapter). Mice that lack a PrP gene cannot develop scrapie, which demonstrates that PrP is essential for development of the disease.

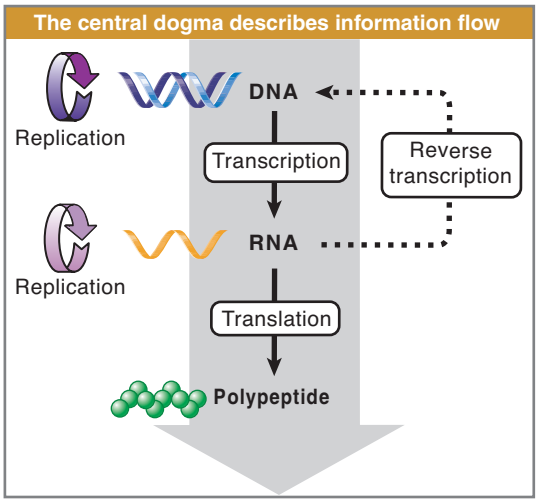


FIGURE 1.30 The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into a polypeptide (protein) is irreversible.

KEY CONCEPT

- Some very small hereditary agents are not polypeptide-encoding nucleic acids, but consist of RNA or protein with heritable properties.

CONCEPT AND REASONING CHECK

How would you distinguish whether a newly discovered infectious agent is an organism, a virus, a viroid, or a prion?

1.15 Most Genes Encode Polypeptides

Sir Archibald Garrod was the first person to make a connection between a metabolic defect and a heritable factor, suggesting that gene action has chemical effects in cells. In studying a patient with alkaptonuria (a disease characterized by brownish-black urine) in 1902, he determined both that the disorder resulted from an enzymatic deficiency and that its pattern of inheritance in the patient's family was consistent with an autosomal recessive allele. Garrod later identified several other "inborn errors of metabolism" with autosomal recessive inheritance.

The work of George Beadle and Edward Tatum in the 1940s was the first systematic attempt to associate genes with enzymes. Beadle and Tatum showed that each step in a metabolic pathway is catalyzed by a single enzyme and can be blocked by mutation in a single gene. This led to the **one gene–one enzyme hypothesis**. A mutation in a gene alters the activity of the protein enzyme it encodes.

one gene–one enzyme hypothesis Beadle and Tatum's hypothesis that a gene is responsible for the production of a single enzyme.

A modification in the hypothesis is needed to apply to proteins that consist of more than one polypeptide subunit. If the subunits are all the same, the protein is a **homomultimer**, and is encoded by a single gene. If the subunits are different, the protein is a **heteromultimer**, and each different subunit is encoded by a different gene. Stated as a more general rule applicable to any heteromultimeric protein, the one gene–one enzyme hypothesis becomes more precisely expressed as the **one gene–one polypeptide hypothesis**. (Even this modification is not completely descriptive of the relationship between genes and proteins, as many genes encode alternative versions of a polypeptide; see the section titled "Alternative Splicing Is a Rule, Rather Than an Exception, in Multicellular Eukaryotes" in the "RNA Splicing and Processing" chapter.)

homomultimer Molecular complex (such as a protein) in which the subunits are identical.

heteromultimer Molecular complex (such as a protein) composed of different subunits.

one gene–one polypeptide hypothesis Modified version of the not generally correct one gene–one enzyme hypothesis; the hypothesis that a gene is responsible for the production of a single polypeptide.

Identifying the biochemical effects of a particular mutation can be a protracted task. The mutation responsible for creating Mendel's wrinkled-pea phenotype was identified only in 1990 as an alteration that inactivates the gene for a starch debranching enzyme!

It is important to remember that a gene does not directly generate a polypeptide. As shown in Figure 1.2, a gene encodes an RNA, which *may* in turn encode a polypeptide. Most genes do encode polypeptides, but some genes encode RNAs that are not translated to polypeptides. These RNAs may be structural components of the protein synthesis machinery or may have roles in regulating gene expression. The basic principle is that *the gene is*

a sequence of DNA that specifies the sequence of an independent product. The process of gene expression may terminate in a product that is either RNA or polypeptide.

A mutation in a coding region is generally a random event with regard to the structure and function of the gene (but see the section earlier in this chapter titled "Mutations Are Concentrated at Hotspots"); mutations can have little or no effect (as in the case of neutral mutations), or they can damage or even abolish gene function. Most mutations that affect gene function are recessive; they result in an absence of function because the mutant allele does not produce its usual polypeptide. **FIGURE 1.31** illustrates the relationship between recessive and wild-type alleles. When a heterozygote contains one wild-type allele and one mutant allele, the wild-type allele is able to direct production of the enzyme and is therefore dominant. (This assumes that an adequate amount of protein is made by the single wild-type allele. When this is not true, the smaller amount made by one allele as compared to two alleles results in the intermediate phenotype of a partially dominant allele in a heterozygote.)

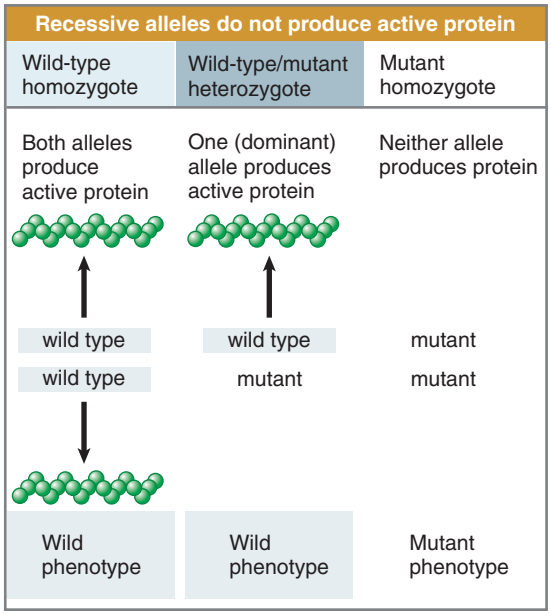


FIGURE 1.31 Genes encode proteins; dominance is explained by the properties of mutant proteins. A recessive allele does not contribute to the phenotype because it produces no protein (or protein that is nonfunctional).

KEY CONCEPTS

- The one gene–one polypeptide hypothesis summarizes the basis of modern genetics: that a typical gene is a stretch of DNA encoding a single polypeptide chain.
- Genes can also encode RNA products that are not translated into polypeptides.
- Mutations can damage gene function.

CONCEPT AND REASONING CHECK

Propose a situation in which a mutant allele is dominant in a heterozygote with a wild-type allele.

HISTORICAL PERSPECTIVES

One Gene–One Enzyme—George W. Beadle and Edward L. Tatum, 1941

Genetic Control of Biochemical Reactions in *Neurospora*

How do genes control metabolic processes? The suggestion that genes are responsible for the production of enzymes was made very early in the history of genetics, most notably by the British physician Archibald Garrod in his 1908 book *Inborn Errors of Metabolism*. But the precise relationship between genes and enzymes was still uncertain. Perhaps each enzyme was controlled by more than one gene, or perhaps each gene contributed to the control of several enzymes. The classic experiments of George Beadle and Edward Tatum showed that the relationship is often remarkably simple: one gene encodes one enzyme. The pioneering experiments united genetics and biochemistry, and for the "one gene, one enzyme" concept, Beadle and Tatum were awarded a Nobel Prize in 1958 (Joshua Lederberg shared the prize for his contributions to microbial genetics). Because we now know that some enzymes contain polypeptide chains encoded by two (or occasionally more) different genes, a more accurate statement of the principle is "one gene, one polypeptide." Beadle and Tatum's experiments also demonstrate the importance of choosing the right organism. *Neurospora* had been introduced as a genetic model organism only a few years earlier, and Beadle and Tatum realized that they could take advantage of the ability of this organism to grow on a simple medium composed of known substances.

> From the standpoint of physiological genetics the development and functioning of an organism consist essentially of an integrated system of chemical reactions controlled in some manner by genes. . . .
>
> In investigating the roles of genes, the physiological geneticist usually attempts to determine the physiological and biochemical bases of already known hereditary traits. . . . There are, however, a number of limitations inherent in this approach. Perhaps the most serious of these is that the investigator must in general confine himself to the study of nonlethal heritable characters. Such characters are likely to involve more or less non-essential so-called "terminal" reactions. . . . A second difficulty . . . is that the standard approach to the problem implies the use of characters with visible manifestations. Many such characters involve morphological variations, and these are likely to be based on systems of biochemical reactions so complex as to make analysis exceedingly difficult. Considerations such as those just outlined have led us to investigate the general problem of the genetic control of development and metabolic reactions by reversing the ordinary procedure and, instead of attempting to work out the chemical bases of known genetic characters, to set out to determine if and how genes control known biochemical reactions. The ascomycete *Neurospora* offers many advantages for such an approach and is well suited to genetic studies. Accordingly, our program has been built around this organism. The procedure is based on the assumption that X-ray treatment will induce mutations in genes concerned with the control of known specific chemical reactions. If the organism must be able to carry out a certain chemical reaction to survive on a given medium, a mutant unable to do this will obviously be lethal on this medium. Such a mutant can be maintained and studied, however, if it will grow on a medium to which has been added the essential product of the genetically blocked reaction. . . .
>
> Among approximately 2000 . . . strains [derived from single cells after X-ray treatment], three mutants have been found that grow essentially normally on the complete medium and scarcely at all on the minimal medium. . . . One of these strains . . . proved to be unable to synthesize vitamin B6 (pyridoxine). A second strain . . . turned out to be unable to synthesize vitamin B1 (thiamine). . . . A third strain . . . has been found to be unable to synthesize para-aminobenzoic acid. . . . [These] preliminary results . . . appear to us to indicate that [this] approach . . . may offer considerable promise as a method of learning more about how genes regulate development and function. For example, it should be possible, by finding a number of mutants unable to carry out a particular step in a given synthesis, to determine whether only one gene is ordinarily concerned with the immediate regulation of a given specific chemical reaction.

Beadle, G. W., and Tatum, E. L. (1941). *Proc. Natl. Acad. Sci. USA* **27**, 499–506.

1.16 Mutations in the Same Gene Cannot Complement

How do we determine whether two mutations that cause a similar phenotype have occurred in the same gene? If they map to positions that are very close together (i.e., they recombine very rarely), they may be alleles. However, in the absence of information about their relative positions, they

could also represent mutations in two different genes whose proteins are involved in the same function. The **complementation test** is used to determine whether two recessive mutations are alleles of the same gene or in different genes. The test consists of generating a heterozygote for the two mutations (by mating parents homozygous for each mutation) and observing its phenotype.

complementation test A test that determines whether two mutations are alleles of the same gene. It is accomplished by crossing two different recessive mutations that have the same phenotype and determining whether the wild-type phenotype can be produced. If so, the mutations are said to complement each other and are probably not mutations in the same gene.

If the mutations are alleles of the same gene, the parental genotypes can be represented as:

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

The first parent provides an m_1 mutant allele, and the second parent provides an m_2 allele, so that the heterozygote progeny have the genotype:

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

No wild-type allele is present, so the heterozygotes have mutant phenotypes.If the mutations lie in different linked genes, the parental genotypes can be represented as:

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

Each chromosome has one wild-type allele at one locus (represented by the plus sign, +) and one mutant allele at the other locus. Then the heterozygote progeny have the genotype:

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

in which the two parents between them have provided a wild-type allele from each gene. The heterozygotes have wild-type phenotypes, and the two genes are said to complement.

The complementation test is shown in more detail in **FIGURE 1.32**. The basic test consists of the comparison shown in the top part of the figure. If two mutations are alleles of the same gene, we see a difference in the phenotypes of the *trans* configuration (both mutations are not in the same allele) and the *cis* configuration (both mutations are in the same allele). The *trans* configuration is mutant, because each allele has a different mutation. However, the *cis* configuration is wild-type, because one allele has two mutations and the other allele has no mutations. The lower part of the figure shows that if the two mutations are in different genes, we always see a wild-type phenotype. There is always one wild type and one mutant allele of each gene in both the *cis* and *trans* configurations. "Failure to complement" means that the two mutations occurred in the same gene. The term **cistron** is used to describe the gene as defined by the complementation test and (like "gene") describes a stretch of DNA that functions as a unit to produce an RNA or polypeptide product. The properties of the gene with regard to complementation are explained by the fact that its product is a single molecule that behaves as a functional unit.

cistron The genetic unit defined by the complementation test; it is equivalent to a gene and includes all noncomplementing alleles.

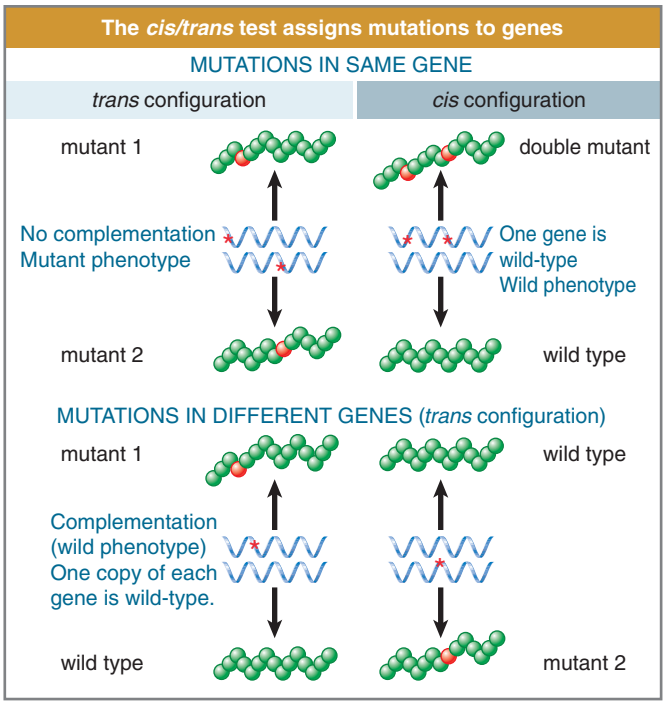


FIGURE 1.32 The cistron is defined by the complementation test. Genes are represented by DNA helices; red stars identify sites of mutation.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Would the complementation test work for mutations that are dominant? Why or why not?

1.17 Mutations May Cause Loss or Gain of Function

Null mutation A mutation that completely eliminates the function of a gene.

The various possible effects of mutation in a gene are summarized in **FIGURE 1.33**.

In principle, when a gene has been identified, insight into its function can be gained by generating a mutant organism that entirely lacks the gene. A mutation that completely eliminates gene function—usually because the gene has been deleted—is called a **null mutation**. If the gene is essential to the organism's survival, a null mutation is lethal, and the gene is referred to as an *essential gene*.

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

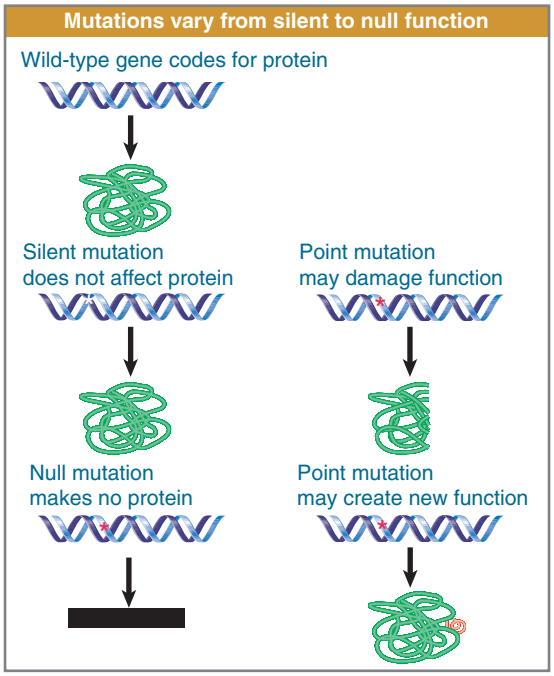


FIGURE 1.33 Mutations that do not affect protein sequence or function are silent. Mutations that abolish all protein activity are null. Mutations that cause loss of function are generally recessive; those that cause gain of function are dominant.

Null mutations, or other mutations that impede gene function (but do not necessarily abolish it entirely), are called **loss-of-function mutations**. A loss-of-function mutation is generally recessive (as in the example of Figure 1.32). Sometimes a mutation has the opposite effect and causes a protein to acquire a new function; such a change is called a **gain-of-function mutation**. A gain-of-function mutation is generally dominant.

Not all mutations in genes lead to a detectable change in the phenotype. Mutations without apparent phenotypic effect are called **silent mutations**. They fall into two categories: (1) base changes in a gene that do not cause any change in the amino acid in the resulting polypeptide and (2) base changes in a gene that change the amino acid, but the replacement in the polypeptide does not affect its activity. Silent mutations in the second category are called **neutral substitutions**.

loss-of-function mutation A mutation that eliminates or reduces the activity of a gene. It is often, but not always, recessive.

gain-of-function mutation A mutation that causes an increase in the normal gene activity or the acquisition of new abnormal properties. It is often, but not always, dominant.

silent mutation A mutation that does not change the sequence of a polypeptide because it produces synonymous codons.

neutral substitutions Mutations that cause changes in amino acids of the protein product but that do not affect the protein's activity.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Explain why loss-of-function mutations are generally recessive and gain-of-function mutations are generally dominant.

1.18 A Locus May Have Many Alleles

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

Different variants of the same gene are called *multiple alleles*, and their existence makes it possible to generate heterozygotes with two mutant alleles. The relationships between these multiple alleles can take various forms.

In the simplest case, a wild-type allele encodes a polypeptide product that is functional, whereas mutant allele(s) encode polypeptides that are nonfunctional. However, there are often cases in which a series of mutant alleles have different phenotypes. For example, wild-type function of the *white* locus of *Drosophila melanogaster* is required for development of the normal red color of the eye. The locus is named for the effect of null mutations, which, in homozygotes, causes the fly to have white eyes.

In *Drosophila*, the name of the wild-type allele is indicated by a "plus" superscript after the name of the locus, so, for example, w^+ is the wild-type *white* allele for red eye color in *D. melanogaster*. Sometimes "+" is used by itself to describe the wild-type allele, and only the mutant alleles are indicated by the name of the locus.

An entirely defective form of the gene (or absence of phenotype) may be indicated by a "minus" superscript. To distinguish among a variety of mutant alleles with different effects, other superscripts may be introduced, such as w^i (ivory eye color) or w^a (apricot eye color).The w^+ allele is dominant over any other allele in heterozygotes, and there are many different mutant alleles for this locus. **FIGURE 1.34** shows a small sample. Although some alleles have no eye color (i.e., a white eye), many alleles produce some color. Each of these mutant alleles must therefore represent a different mutation of the gene, many of which do not eliminate its function entirely but leave a residual activity that produces a characteristic phenotype. These alleles are named for the color of the eye in a homozygote. Most *w* mutations affect the quantity of pigment in the eye. The examples in the figure are arranged in roughly declining amount of color, but others, such as w^{sp}, affect the pattern in which pigment is deposited.

When multiple alleles exist, an organism may be a heterozygote that carries two different mutant alleles. The phenotype of such a heterozygote depends on the nature of the residual activity of each allele. The relationship between two mutant alleles is, in principle, no different from that between

Each allele has a different phenotype	
Allele	Phenotype of homozygote
w^+	red eye (wild type)
w^{bl}	blood
w^{ch}	cherry
w^{bf}	buff
w^h	honey
w^a	apricot
w^e	eosin
w^i	ivory
w^z	zeste (lemon-yellow)
w^{sp}	mottled, color varies
w^1	white (no color)

FIGURE 1.34 The *w* locus in *D. melanogaster* has an extensive series of alleles whose phenotypes extend from wild-type (red) color to complete lack of pigment.

wild-type and mutant alleles: one allele may be dominant, there may be partial dominance, or there may be codominance.

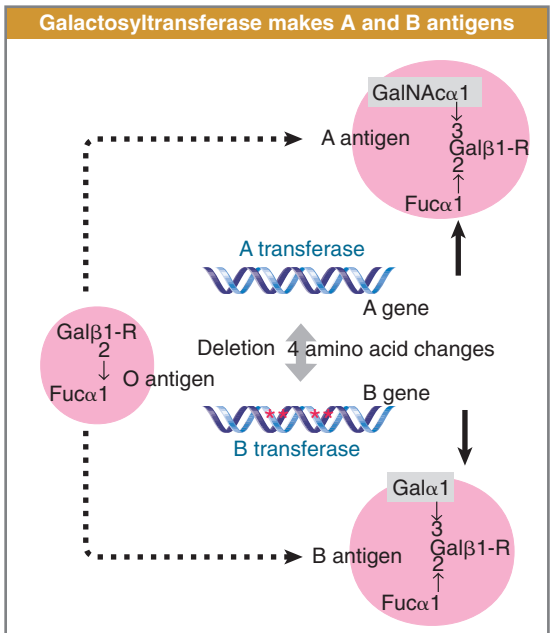
There is not necessarily a unique wild-type allele for any particular locus. Control of the *ABO* human blood group system provides an example. Lack of function is represented by the null, or *O*, allele. However, the functional alleles *A* and *B* are codominant with one another and dominant to the *O* allele. The basis for this relationship is illustrated in **FIGURE 1.35**.

The O antigen is generated in all individuals and consists of a particular carbohydrate group that is added to proteins. The *ABO* locus encodes a galactosyltransferase enzyme that puts an additional sugar group on the O antigen. The specificity of this enzyme determines the blood group. The *A* allele produces an enzyme that uses the cofactor UDP-N-acetylgalactose, forming the A antigen. The *B* allele produces an enzyme that uses the cofactor UDP-galactose, forming the B antigen. The A and B versions of the transferase enzyme differ in four amino acids that presumably affect its recognition of the type of cofactor. The *O* allele has a small deletion that eliminates the activity of the transferase, so no modification of the O antigen occurs.

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

Neither *A* nor *B* alleles can be regarded as uniquely wild-type because they represent alternative activities rather than loss or gain of function. A situation such as this, in which there are multiple functional alleles in a population, is described as a **polymorphism** (see the section titled "Individual Genomes Show Extensive Variation" in the chapter titled "The Content of the Genome").

polymorphism The simultaneous occurrence in the population of alleles showing variations at a given position.



Phenotype	Genotype	Activity
O	OO	None
A	AO or AA	N-Ac-gal transferase
B	BO or BB	Gal transferase
AB	AB	GalN-Ac-Gal-transferase

FIGURE 1.35 The *ABO* human blood group locus encodes a galactosyltransferase whose specificity determines the blood group.

KEY CONCEPTS

- The existence of multiple alleles allows the possibility of heterozygotes representing any pairwise combination of alleles.
- A locus may have a polymorphic distribution of alleles with no individual allele that can be considered to be the sole wild type.

CONCEPT AND REASONING CHECK

Explain why an individual fly with a new recessive mutation of the *white* locus will have a wild-type phenotype.

1.19 Recombination Occurs by Physical Exchange of DNA

genetic recombination New combinations of alleles at different loci resulting from a mechanism by which separate DNA molecules are joined into a single molecule due to such processes as crossing over or transposition.

The term **genetic recombination** describes the generation of new combinations of alleles at each generation in diploid organisms. Generally, each chromosome has a *homologous* partner with which it pairs during meiosis. The two homologous copies of each chromosome may have different alleles at some loci. By the exchange of corresponding segments between the homologs, called *crossing over*, recombinant chromosomes that are different from the parental chromosomes can be generated.

Recombination results from a physical exchange of chromosomal material. For example, recombination may result from the crossing over that occurs during meiosis (the specialized division that produces haploid germ cells), specifically during prophase I. Meiosis starts with a cell that has duplicated its chromosomes, so that it has four copies of each chromatid. (A *chromatid* is one of two identical copies of a chromosome following its duplication; because there are two homologs per diploid cell, there are four chromatids of each type of chromosome.) Early in meiosis, all four copies are closely associated (synapsed) in a structure called a *bivalent* and, later, a *tetrad*. At this point, pairwise exchanges of material between two nonsister (of the four total) chromatids may occur.

chiasma A site at which two homologous chromosomes synapse during prophase I of meiosis.

The point of synapsis between homologs is called a **chiasma** and is illustrated diagrammatically in **FIGURE 1.36**. A chiasma represents a site at which one strand in each of two nonsister chromatids in a tetrad has been broken and exchanged. If during the resolution of the chiasma the previously unbroken strands are also broken and exchanged, recombinant chromatids will be generated. Each recombinant chromatid consists of material derived from one chromatid on one side of the chiasma, with material from the other chromatid on the opposite side. The two recombinant chromatids have reciprocal structures. The event can be described as a "breakage and reunion." Because each individual crossing-over event involves only two of the four associated chromatids, a single recombination event can produce only 50% recombinants.

The complementarity of the two strands of DNA is essential for the recombination process. Each of the chromatids shown in Figure 1.36 consists of a very long duplex of DNA. For them to be broken and reconnected without any addition or loss of material requires a mechanism to recognize exactly corresponding positions; this mechanism is complementary base pairing.Recombination results from a process in which the single strands in the region of the crossover exchange their partners, resulting in a branch that may migrate for some

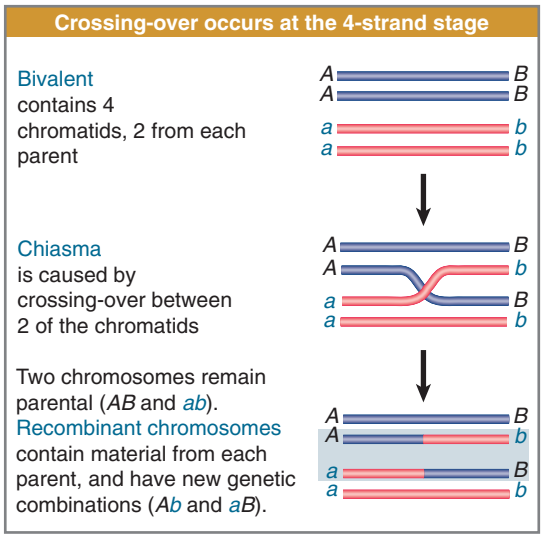


FIGURE 1.36 Chiasma formation at prophase I of meiosis is required for generating recombinant chromosomes.

distance in either direction. This creates a stretch of **heteroduplex DNA**, in which the single strand of one duplex is paired with its complement from the other duplex. The mechanism, of course, involves other stages in which strands must be broken and religated, which we discuss in more detail in the chapter titled "Homologous, Somatic and Site-Specific Recombination," but the crucial feature that makes precise recombination possible is the complementarity of DNA strands. A stretch of heteroduplex DNA forms in the recombination intermediate when a single strand crosses over from one duplex to the other. Each recombinant consists of one parental duplex DNA, which is connected by a stretch of heteroduplex DNA to the other parental duplex. Each duplex DNA corresponds to one of the chromatids involved in recombination in Figure 1.36.

heteroduplex DNA
DNA that is generated by base pairing between complementary single strands derived from the different parental duplex molecules; it occurs during crossing over.

The formation of heteroduplex DNA requires the sequences of the two recombining duplexes to be close enough to allow pairing between the complementary strands. If there are no differences between the two parental genomes in this region, formation of heteroduplex DNA will be perfect. However, pairing can still occur even when there are small differences. In this case, the hybrid DNA has points of mismatch, at which a base in one strand is paired with a base in the other strand that is not complementary to it. The correction of such mismatches is another feature of genetic recombination (see the chapter titled "Repair Systems").

Over chromosomal distances, recombination events occur more or less at random, with a characteristic frequency. The probability that a crossover will occur within any specific region of the chromosome is more or less proportional to the length of the region, up to a saturation point. For example, a large human chromosome usually has three or four crossover events per meiosis, but a small chromosome may have only one on average.

FIGURE 1.37 compares three situations: two genes on different chromosomes, two genes that are far apart on the same chromosome, and two genes that are close together on the same chromosome. Genes on different chromosomes segregate independently according to Mendel's laws, resulting in the production of 50% "parental" types and 50% "recombinant" types during meiosis. When genes are sufficiently far apart on the same chromosome, the probability of at least one crossover in the region between them becomes so high that their association is the same as that of genes on different chromosomes, and they show 50% recombination.

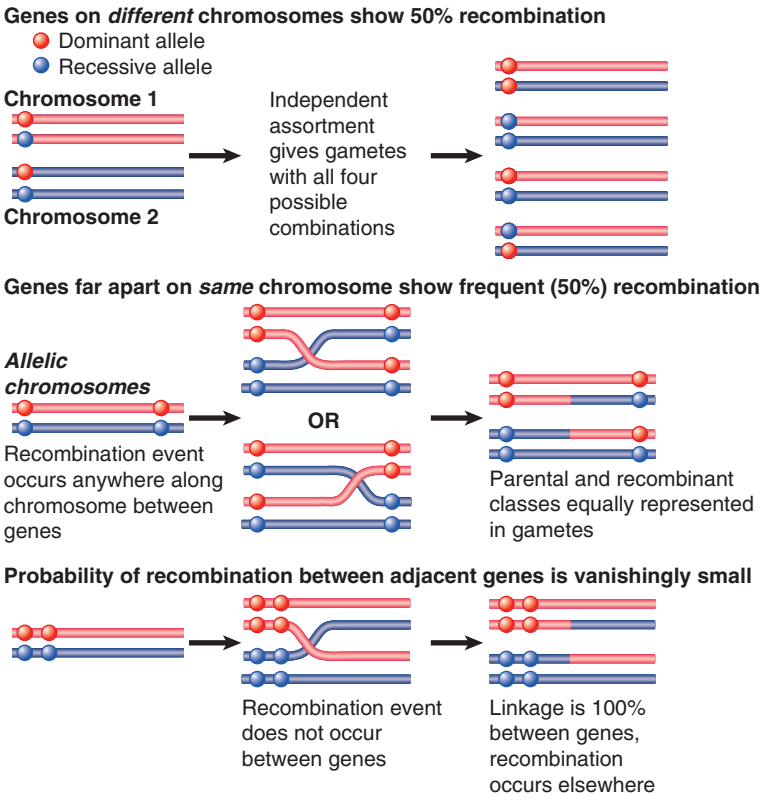


FIGURE 1.37 Genes on different chromosomes segregate independently so that all possible combinations of alleles are produced in equal proportions. Crossing over occurs so frequently between genes that are far apart on the same chromosome that they effectively segregate independently. But recombination is reduced when genes are closer together, and for adjacent genes it may hardly ever occur.

What if genes are close together on the same chromosome? The probability of a crossover between them is reduced, and recombination occurs only in some proportion of meioses. For example, if it occurs in one-quarter of the meioses, the overall rate of recombination is 12.5% (because a single recombination event produces 50% recombination, and this occurs in 25% of meioses). When genes are very close together, as shown in the bottom panel of Figure 1.37, recombination between them may never be observed in phenotypes of multicellular eukaryotes (because they produce few offspring).

This leads us to the concept that a chromosome is an array of many genes. Each protein-coding gene is an independent unit of expression and is represented in one or more polypeptide chains. The properties of a gene can be changed by mutation. The allelic combinations present on a chromosome can be changed by crossing over. We will next examine the relationship between the sequence of a gene and the sequence of the polypeptide chain it encodes.

KEY CONCEPTS

- Recombination is the result of crossing over that occurs at a chiasma during meiosis and involves two of the four chromatids.
- Recombination occurs by a breakage and reunion that proceeds via an intermediate of heteroduplex DNA.
- The distance between genes on the same chromosome is determined by the frequency of recombination between them.
- Genes that are close together are tightly linked because recombination between them is rare.
- Genes that are far apart may not show linkage because recombination is so frequent as to produce the same result as for genes on different chromosomes.

CONCEPT AND REASONING CHECK

Explain why the maximum frequency of recombination that can be measured between two loci is 50%.

1.20 The Genetic Code Is Triplet

Each protein-coding gene encodes a particular polypeptide chain. The concept that each polypeptide consists of a particular sequence of amino acids dates from Sanger's characterization of insulin in the 1950s. The discovery that a gene consists of DNA presents us with the issue of how a sequence of nucleotides in DNA is used to construct a sequence of amino acids in a polypeptide.

A crucial feature of the general structure of DNA, the double helix, is that it is independent of the particular sequence of its component nucleotides. The sequence of nucleotides in a protein-coding gene is important because it encodes the sequence of amino acids that constitutes the corresponding polypeptide. The relationship between a sequence of DNA and the sequence of the corresponding polypeptide is called the **genetic code**.

genetic code The correspondence between triplets in DNA (or RNA) and amino acids in polypeptide.

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

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

The coding region of a gene is deciphered by a complex apparatus that interprets the nucleic acid sequence. In any given region, it is usually the case that only one of the two strands of DNA encodes a functional RNA, so we write the genetic code as a sequence of bases (rather than base pairs). (There are cases where both strands are transcribed; examples are given in the chapter titled "Regulatory RNA.")

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

codon A triplet of nucleotides that codes for an amino acid, or a termination signal.

A coding sequence is read in nonoverlapping triplets from a fixed starting point:

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

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

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

then a nucleotide substitution will affect only one amino acid. For example, the substitution of an A by some other base (X) causes aa2 to be replaced by aa5:

UUU	AAX	GGG	CCC
aa1	aa5	aa3	aa4

because only the second codon has been changed.

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

frameshift A mutation caused by deletions or insertions that are not a multiple of three base pairs. They change the frame in which triplets are translated into polypeptides.

UUU	AAX	AGG	GCC	C
aa1	aa5	aa6	aa7	

Because the new sequence of triplets is completely different from the old one, the entire amino acid sequence of the polypeptide is altered downstream from the site of mutation, and may also be prematurely truncated, so the function of the protein is likely to be lost completely. Frameshift mutations are induced by **acridines**, compounds that bind to DNA and distort the structure of the double helix, causing additional bases to be incorporated or omitted during replication. Each mutagenic event in the presence of an acridine results in the addition or removal of a single base pair.

acridines Mutagens that act on DNA to cause the insertion or deletion of a single base pair. They were useful in defining the triplet nature of the genetic code.

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

FIGURE 1.38 illustrates the properties of frameshift mutations. An insertion or deletion changes the entire polypeptide sequence following the site of mutation. However, the combination of the insertion of a single nucleotide and the deletion of a single nucleotide causes the code to be read incorrectly only between the two sites of mutation; reading in the original frame resumes after the second site.

In 1961, genetic analysis of acridine mutations in the *rII* region of the phage T6 showed that all the mutations could be classified into one of two sets, described as (+) and (–). Either type of mutation by itself causes a frameshift: the (+) type by virtue of a base addition and the (–) type by virtue of a base deletion. Double mutant combinations of the types (++) and (– –) continue to show mutant behavior. However, combinations of the types (+–) or (–+) suppress one another, so that one mutation is described as a frameshift suppressor of the other. (In the context of this work, "suppressor" is used in an unusual sense because the second mutation is in the same gene as the first; in fact, these are second-site reversions.)

These results show that the genetic code must be read as a sequence that is fixed by the starting point. Therefore, a single base addition and deletion compensate for each other, whereas double additions or double deletions remain frameshift mutants. However, these observations do not suggest how many nucleotides make up each codon.

When triple mutants are constructed, only (+++) and (– – –) combinations show the wild-type phenotype, whereas other combinations remain mutant. If we take three single base additions or three deletions to

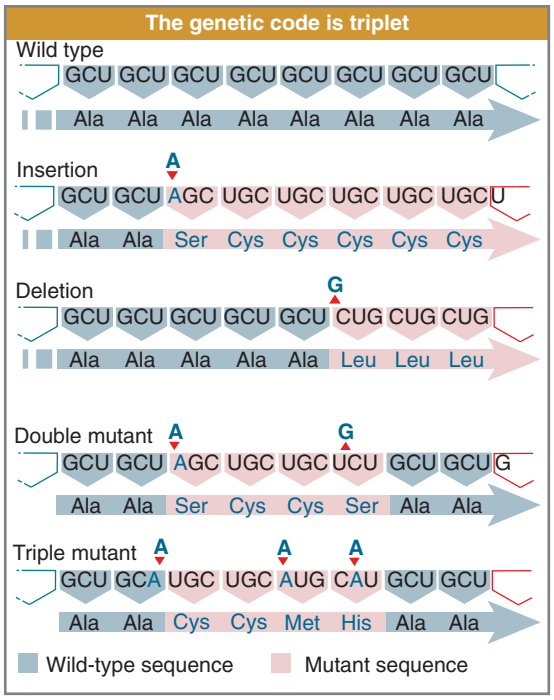


FIGURE 1.38 Frameshift mutations show that the genetic code is read in triplets from a fixed starting point.

correspond, respectively, to the addition or omission overall of a single amino acid, this implies that the code is read in triplets. An incorrect amino acid sequence is found between the two outside sites of mutation, and the sequence on either side remains wild type, as indicated in Figure 1.38.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECK

Consider mutagens that insert or delete two adjacent nucleotides. Using "+" to mean an insertion of two nucleotides and "–" to mean a deletion of two nucleotides, what combinations of insertions and deletions would suppress one another?

1.21 Every Coding Sequence Has Three Possible Reading Frames

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

Reading frame One of three possible ways of reading a nucleotide sequence. Each divides the sequence into a series of successive triplets.

open reading frame (ORF) A sequence of DNA consisting of triplets that can be translated into amino acids starting with an initiation codon and ending with a termination codon.

initiation codon A special codon (usually AUG) used to start synthesis of a polypeptide.

termination codon One of the three codons (UAA, UAG, UGA) that signal the termination of translation of a polypeptide. They are also known as stop codons.

closed (blocked) reading frame A reading frame that cannot be translated into a polypeptide because of the occurrence of termination codons.

unidentified reading frame (URF) An open reading frame with an as yet undetermined function.

A C G A C G A C G A C G A C G A C G
the three possible reading frames (each starting in one of the bold nucleotides shown in the sequence above) are
ACG ACG ACG ACG ACG ACG
CGA CGA CGA CGA CGA CG
GAC GAC GAC GAC GAC G

A reading frame that consists exclusively of triplets coding for amino acids is called an **open reading frame,** or **ORF**. A sequence that is translated into a polypeptide has a reading frame that starts with a special **initiation codon** (AUG) and then extends through a series of triplets coding for amino acids until it ends at one of three **termination codons** (see the chapter titled "Using the Genetic Code"). ORFs can also be characterized by consensus promoter sequences at an appropriate distance upstream from the initiation codon and (in eukaryotes) the presence of intron splicing junctions.

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

When the sequence of a DNA region of unknown function is obtained, each possible reading frame can be analyzed to determine whether it is open or closed. Usually, no more than one of the three possible reading frames is open in any single stretch of DNA. **FIGURE 1.39** shows an example of a sequence that can be read in only one reading frame because the alternative reading frames are blocked by frequent termination codons. A long open reading frame is unlikely to exist by chance; if it had not been translated into polypeptide, there would have been no selective pressure to prevent the accumulation of termination codons. Therefore, the identification of a lengthy open reading frame is taken to be *prima facie* evidence that the sequence is (or until recently has been) translated into a polypeptide in that frame. An ORF for which no polypeptide product has been identified is sometimes called an **unidentified reading frame (URF)**.

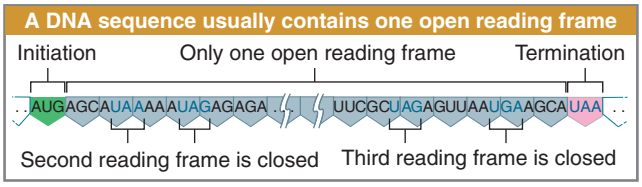


FIGURE 1.39 An open reading frame starts with AUG and continues in triplets to a termination codon. Closed reading frames may be interrupted frequently by termination codons.

KEY CONCEPT

- Usually, only one of the three possible reading frames is translated, and the other two are closed by frequent termination signals.

CONCEPT AND REASONING CHECK

Suppose you identify a stretch of DNA in which only one of the three possible reading frames is closed. What would you hypothesize from this observation?

1.22 Bacterial Genes Are Colinear with Their Products

By comparing the nucleotide sequence of a gene with the amino acid sequence of its polypeptide product, we can determine whether the gene and the polypeptide are **colinear**—that is, whether the sequence of nucleotides in the gene exactly corresponds to the sequence of amino acids in the polypeptide. In bacteria and their viruses, genes and their products are colinear. Each gene is a continuous stretch of DNA with a coding region that is three times the number of amino acids in the polypeptide that it encodes (due to the triplet nature of the genetic code). In other words, if a polypeptide contains *N* amino acids, the gene encoding that polypeptide contains 3*N* nucleotides.

colinearity The relationship that describes the 1:1 correspondence of a sequence of triplet nucleotides to a sequence of amino acids.

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

The colinearity of gene and polypeptide was originally investigated in the tryptophan synthetase gene of *E. coli*. Genetic distance was measured as the percentage of recombination between mutations; amino acid distance was measured as the number of amino acids separating sites of replacement. **FIGURE 1.40** compares the two maps. The order of seven sites of mutation is the same as the order of the corresponding sites of amino acid replacement, and the recombination distances are roughly similar to the actual distances in the protein. The recombination map expands the distances between some mutations, but otherwise there is little distortion of the recombination map relative to the physical map.

The recombination map leads to two further general points about the organization of the gene. Different mutations may cause a wild-type amino acid to be replaced with different alternatives. (See the codon table in Figure 22.1 to see how base changes in specific codons can alter the resulting amino acid.) If two such

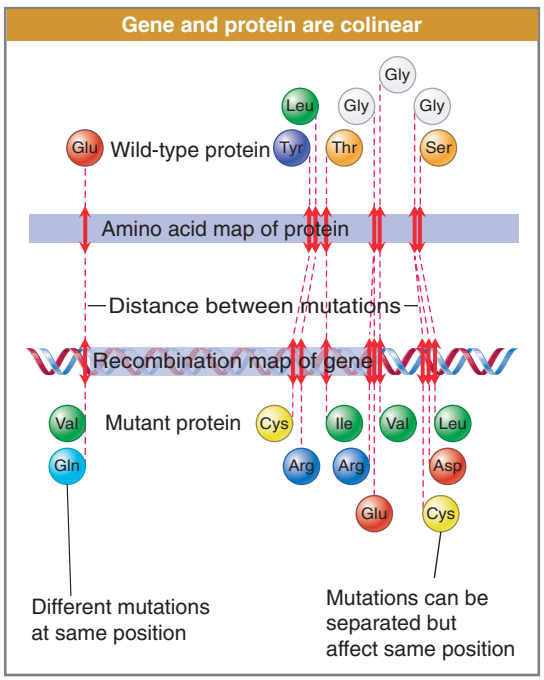


FIGURE 1.40 The recombination map of the tryptophan synthetase gene corresponds with the amino acid sequence of the polypeptide.

mutations cannot recombine, they must involve different point mutations at the same position in DNA. If the mutations can be separated on the genetic map but affect the same amino acid on the upper map (the connecting lines converge in the figure), they must involve point mutations at different positions in the same codon. This happens because the minimum size unit of genetic recombination (1 bp) is smaller than the unit coding for the amino acid (3 bp).

KEY CONCEPTS

- A bacterial gene consists of a continuous length of 3*N* nucleotides that encodes *N* amino acids.
- The gene is colinear with both its mRNA and polypeptide products.

CONCEPT AND REASONING CHECK

Although a recombination map and a physical map will show genetic markers in the same order, the distance between markers will usually be different when the two maps are compared. Why?

1.23 Several Processes Are Required to Express the Product of a Gene

messenger RNA (mRNA) The intermediate that represents one strand of a gene encoding a polypeptide. Its coding region is related to the polypeptide sequence by the triplet genetic code.

antisense (template) strand The DNA strand that is complementary to the sense strand and acts as the template for synthesis of mRNA.

coding (sense) strand The DNA strand that has the same sequence as the mRNA and is related by the genetic code to the polypeptide sequence that it represents.

gene expression The process by which the information in a sequence of DNA in a gene is used to produce an RNA or polypeptide, involving transcription and (for polypeptides) translation.

In comparing a gene and its polypeptide product, we are restricted to the sequence of DNA that lies between the points corresponding to the N-terminus and C-terminus of the polypeptide. However, a gene is not directly translated into polypeptide but is expressed via the production of a **messenger RNA** (abbreviated as **mRNA**), a nucleic acid intermediate actually used to synthesize a polypeptide (as we see in detail in the "Translation" chapter).

Messenger RNA is synthesized by the same process of complementary base pairing used to replicate DNA, with the important difference being that it corresponds to only one strand of the DNA double helix. **FIGURE 1.41** shows that the sequence of mRNA is complementary to the sequence of one strand of DNA (called the **antisense,** or **template, strand**) and is identical (apart from the replacement of T with U) to the other strand of DNA (called the **coding,** or **sense, strand**). The convention for writing DNA sequences is that the top strand is the coding strand and runs 5′ to 3′.

The process by which information from a gene is used to synthesize an RNA or polypeptide product is called **gene expression**. In bacteria,

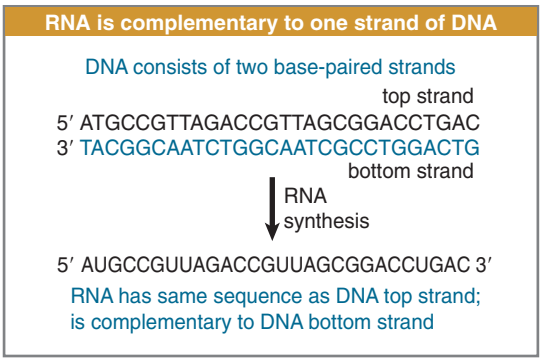


FIGURE 1.41 RNA is synthesized by using one strand of DNA as a template for complementary base pairing.

expression of a protein-coding gene consists of two stages. The first stage is **transcription**, when an mRNA copy of the template strand of the DNA is produced. The second stage is **translation** of the mRNA into polypeptide. This is the process by which the sequence of an mRNA is read in triplets to give the series of amino acids that make the corresponding polypeptide.

An mRNA includes a sequence of nucleotides that corresponds with the sequence of amino acids in the polypeptide. This part of the nucleic acid is called the **coding region**. However, the mRNA includes additional sequences on either end that do not encode amino acids. The 5′ untranslated region is called the **leader,** or **5′ UTR**, and the 3′ untranslated region is called the **trailer,** or **3′ UTR**.

The gene includes the entire sequence represented in mRNA. Sometimes mutations impeding gene function are found in the additional, noncoding regions, confirming the view that these sequences are a legitimate part of the genetic unit. **FIGURE 1.42** illustrates this situation, in which the gene is considered to comprise a continuous stretch of DNA needed to produce a particular polypeptide, including the leader, the coding region, and the trailer.

A bacterial cell has only a single compartment, so transcription and translation occur in the same place, as illustrated in **FIGURE 1.43**. In eukaryotes, transcription occurs in the nucleus, but the RNA product must be

transcription Synthesis of RNA from a DNA template.

translation Synthesis of polypeptide from an mRNA template.

coding region A part of a gene that encodes a polypeptide sequence.

leader (5′ UTR) In mRNA, the untranslated sequence at the 5′ end that precedes the initiation codon.

trailer (3′ UTR) The untranslated sequence at the 3′ end of an mRNA following the termination codon.

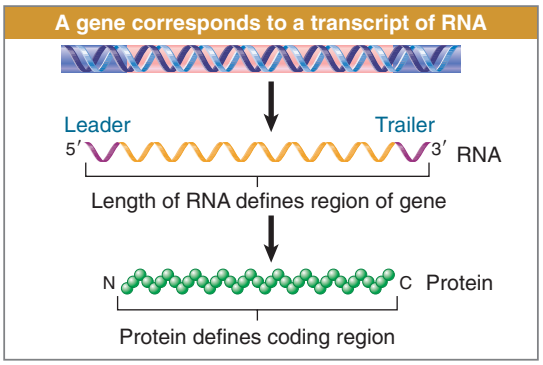


FIGURE 1.42 The gene is usually longer than the sequence encoding the polypeptide.

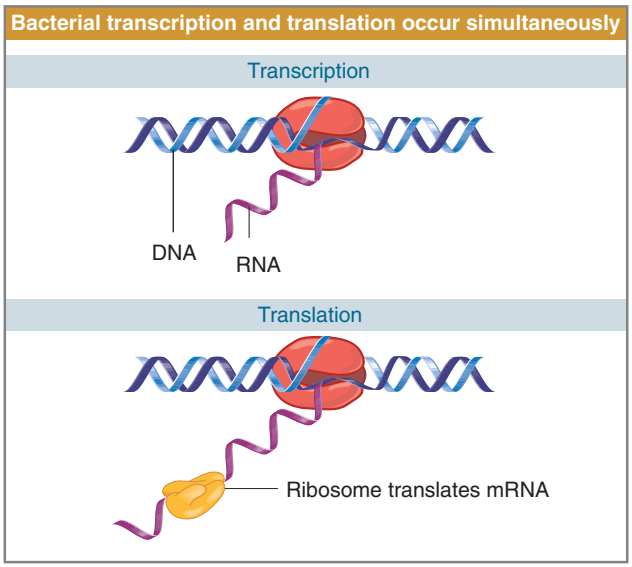


FIGURE 1.43 Transcription and translation take place in the same compartment in bacteria.

transported to the cytoplasm in order to be translated. This results in a spatial separation between transcription (in the nucleus) and translation (in the cytoplasm). The simplest eukaryotic genes are like bacterial genes; the transcript RNA is in fact the mRNA. However, for more complex genes, the primary transcript of the gene is a **pre-mRNA** that requires processing to generate the mature mRNA. The basic stages of gene expression in eukaryotes are outlined in **FIGURE 1.44**.

The most important stage in **RNA processing** is **splicing**. Many genes in eukaryotes (and the majority in multicellular eukaryotes) contain internal regions called **introns** that do not carry coding information for the polypeptide products encoded by those genes. The coding information is contained in **exons**. The process of splicing removes introns from the pre-mRNA to generate an RNA that has a continuous open reading frame (see Figure 3.1). Other processing events that occur at this stage involve the modification of the 5′ and 3′ ends of the pre-mRNA (see Figure 19.1).

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

pre-mRNA The nuclear transcript that is processed by modification and splicing to produce a mature mRNA.

RNA processing Modifications to RNA transcripts of eukaryotic genes. This may include alterations to the 3′ and 5′ ends and the removal of introns.

splicing The process of excising introns from RNA and connecting the exons into a continuous mRNA.

intron A segment of DNA that is transcribed but later removed from within the transcript by splicing together the sequences (exons) on either side of it.

exon Any segment of an interrupted gene that is represented in the mature RNA product.

ribosome A large assembly of RNA and proteins that synthesizes polypeptides under direction from an mRNA template.

ribosomal RNAs (rRNAs) A major component of the ribosome.

transfer RNA (tRNA) The intermediate in protein synthesis that interprets the genetic code. Each tRNA molecule can be linked to an amino acid and has an anticodon sequence that is complementary to a triplet codon representing the amino acid.

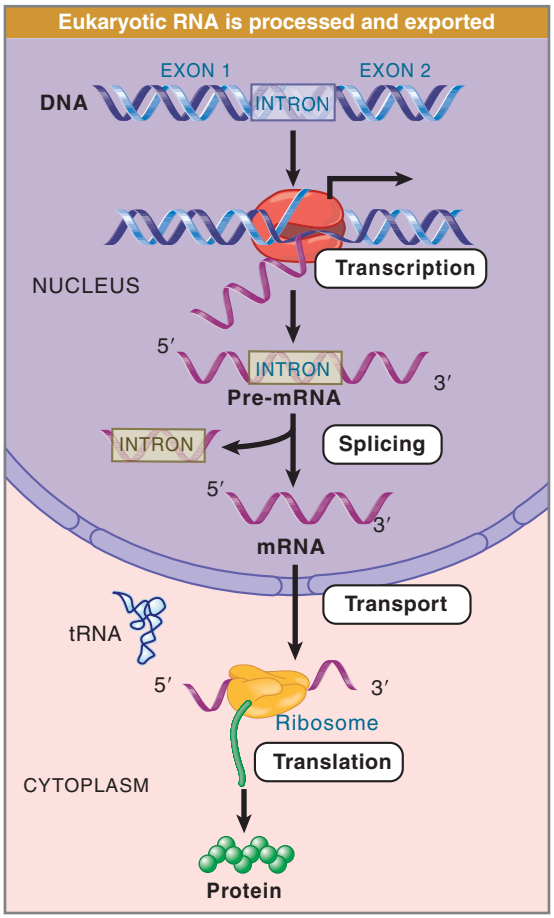


FIGURE 1.44 In eukaryotes, transcription occurs in the nucleus, and translation occurs in the cytoplasm.

It is an important point to note that the process of gene expression involves RNA not only as the essential substrate but also in providing components of the apparatus. The rRNA and tRNA components are encoded by genes and are generated by the process of transcription (like mRNA), but they are not translated to polypeptides.

KEY CONCEPTS

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

CONCEPT AND REASONING CHECKS

Why is bacterial gene expression generally faster than eukaryotic gene expression?
Why do the recombination maps of eukaryotic genes usually not correspond to the amino acid maps of their products?

1.24 Proteins Are *trans*-Acting; Sites on DNA Are *cis*-Acting

A crucial progression in the definition of the gene was the realization that all its parts must be present on one contiguous stretch of DNA. In genetic terminology, sites that are located on the same DNA are said to be in *cis*. Sites that are located on two different molecules of DNA are described as being in *trans*. So two mutations may be in *cis* (on the same DNA) or in *trans* (on different DNAs). The complementation test uses this concept to determine whether two mutations are in the same gene (see the section titled "Mutations in the Same Gene Cannot Complement" earlier in this chapter). We may now extend the concept of the difference between *cis* and *trans* effects from defining the coding region of a gene to describing the interaction between a gene and its regulatory elements.

Suppose that the ability of a gene to be expressed is controlled by a protein that binds to the DNA close to the coding region. In the example depicted in **FIGURE 1.45**, RNA can be synthesized only when the protein is bound to a control site on the DNA. Now suppose that a mutation occurs in the control site so that the protein can no longer bind to it. As a result, the gene can no longer be expressed.

So gene expression can be inactivated either by a mutation in a control site or by a mutation in a coding region. The mutations cannot be distinguished genetically because both have the property of acting only on the DNA sequence of the single allele in which they occur. They have identical properties in the complementation test, so a mutation in a control region is defined as comprising part of the gene in the same way as a mutation in the coding region. **FIGURE 1.46** shows that a change in the control site affects only the coding region to which it is connected; it does not affect the ability of the homologous allele to be expressed. A mutation that acts solely by affecting the properties of the contiguous sequence of DNA is called a ***cis*-acting sequence**.

***cis*-acting sequence** A site that affects the activity only of sequences on its own molecule of DNA (or RNA); this property usually implies that the site does not encode polypeptide.

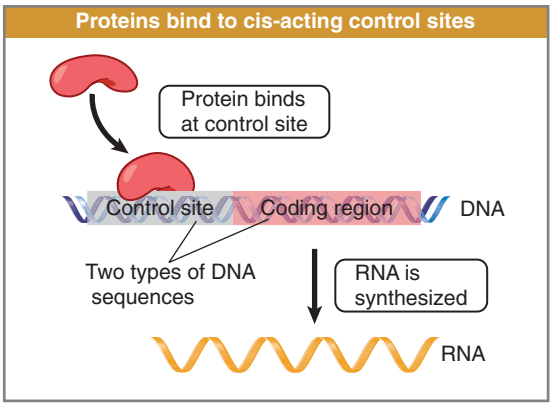


FIGURE 1.45 Control sites in DNA provide binding sites for proteins; coding regions are expressed via the synthesis of RNA.

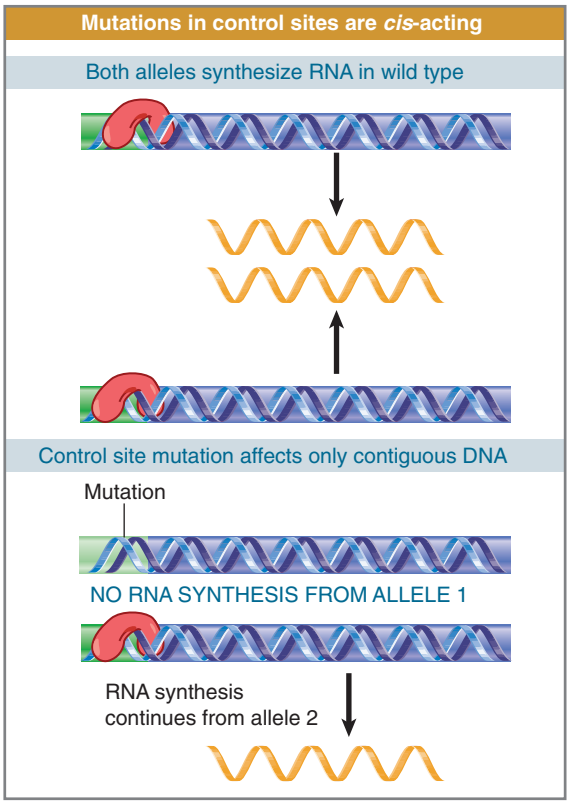


FIGURE 1.46 A *cis*-acting site controls the adjacent DNA but does not influence the other allele.

We may contrast the behavior of the *cis*-acting mutation shown in Figure 1.46 with the result of a mutation in the gene encoding the regulatory protein. **FIGURE 1.47** shows that the absence of regulatory protein would prevent both alleles from being expressed. A mutation of this sort is said to be a ***trans*-acting sequence**.

trans-**acting sequence** DNA sequence encoding a product that can function on any copy of its target DNA. This implies that it is a diffusible protein or RNA.

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

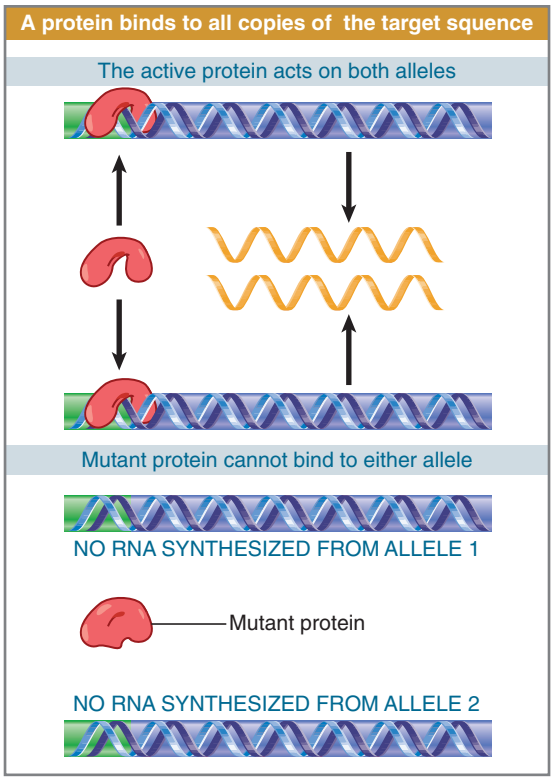


FIGURE 1.47 A *trans*-acting mutation in a gene for a regulatory protein affects both alleles of a gene that it controls.

KEY CONCEPTS

- All gene products (RNA or polypeptides) are *trans*-acting. They can act on any copy of a gene in the cell.
- *cis*-acting mutations identify sequences of DNA that are targets for recognition by *trans*-acting products. They are not expressed as RNA or polypeptide and affect only the contiguous stretch of DNA.

CONCEPT AND REASONING CHECK

Can a mutation in a control site alter the function of a gene's protein product without inactivating its expression? Explain.

1.25 Summary

Two classic experiments provided strong evidence that DNA is the genetic material of bacteria, eukaryotic cells, and many viruses. DNA isolated from one strain of *Pneumococcus* bacteria can confer properties of that strain upon another strain. In addition, DNA is the only component that is inherited by progeny phages from parental phages. DNA can be used to transfect new properties into eukaryotic cells.

DNA is a double helix consisting of antiparallel strands in which the nucleotide units are linked by 5′ to 3′ phosphodiester bonds. The backbone is on the exterior; purine and pyrimidine bases are stacked in the interior in pairs in which A is complementary to T and G is complementary to C. In semiconservative replication, the two strands separate, and daughter strands are assembled by complementary base pairing. Complementary base pairing is also used to transcribe an RNA from one strand of a DNA duplex.

A mutation consists of a change in the sequence of A-T and G-C base pairs in DNA. A mutation in a coding sequence may change the sequence of amino acids in the corresponding polypeptide. A point mutation changes only the amino acid represented by the codon in which the mutation occurs. Point mutations may be reverted by back mutation of the original mutation. Insertions may revert by loss of the inserted material, but deletions cannot revert. Mutations may also be suppressed indirectly when a mutation in a different gene counters the original defect.The natural incidence of mutations is increased by mutagens. Mutations may be concentrated at hotspots. A type of hotspot responsible for some point mutations is caused by deamination of the modified base 5-methylcytosine. Forward mutations occur at a rate of approximately 10^{-6} per locus per generation; back mutations are rarer.

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

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

When a gene encodes a polypeptide, the relationship between the sequence of DNA and sequence of the polypeptide is given by the genetic code. Only one of the two strands of DNA codes for polypeptide. A codon consists of three nucleotides that represent a single amino acid. A coding sequence of DNA consists of a series of codons, read from a fixed starting point. Usually, only one of the three possible reading frames can be translated into a polypeptide.

A gene may have multiple alleles. Recessive alleles are caused by loss-of-function mutations that interfere with the function of the protein. A null allele has total loss of function. Dominant alleles are caused by gain-of-function mutations that create a new property in the protein.

Chapter Questions

1. Which strain of *S. pneumoniae* was found to be nonvirulent (nonlethal) when mice were infected with it?
   A. the smooth strain
   B. the rough strain
   C. both strains
   D. none were virulent

2. What isotope can be used to specifically label protein?
   A. ^{14}C
   B. ^{3}H
   C. ^{32}P
   D. ^{35}S

3. The difference between RNA and DNA is the presence of a:
   A. 2′-PO_4 group on the ribose sugar in RNA.
   B. 3′-PO_4 group on the ribose sugar in RNA.

C. 2′-OH group on the ribose sugar in RNA.
D. 3′-OH group on the ribose sugar in RNA.

4. Which pair of scientists determined that DNA replication is semiconservative?
   A. Meselson and Stahl
   B. Watson and Crick
   C. Okazaki and Okazaki
   D. Griffith and Avery

5. In the density labeling experiment to study DNA replication, parental DNA in the cell was labeled with a high-density isotope and, after one or more generations of growth, subjected to density gradient centrifugation. Which of the following populations was not detected after the second generation?
   A. the light density population
   B. the hybrid density population
   C. the heavy density population
   D. both the light and heavy density populations

6. Which class of infectious agent is inserted into the host genome as a double-stranded DNA segment?
   A. retroviruses
   B. double-stranded RNA viruses
   C. double-stranded DNA viruses
   D. viroids

7. The mutation rate in bacteria is about:
   A. 10^{-6} per locus per generation.
   B. 10^{-7} per locus per generation.
   C. 10^{-8} per locus per generation.
   D. 10^{-9} per locus per generation.

8. About 30% of human point mutations are associated with which of the following modified bases?
   A. 5-methylguanine
   B. 5-methyladenine
   C. 5-methylthymine
   D. 5-methylcytosine

9. The presence of the modified base in the previous question often leads to:
   A. transitions.
   B. transversions.
   C. deletions.
   D. insertions.

10. The reversal of an original base pair that was changed from A-T to G-C and then back to A-T is an example of a:
    A. true reversion.
    B. second-site reversion.
    C. forward mutation.
    D. suppression.

11. Failure to complement means that two mutations are:
    A. part of the same genetic unit.
    B. in related genetic units on the same chromosome.
    C. in related genetic units on different chromosomes.
    D. in totally unrelated genetic units.

12. Most mutations that affect gene function are:
    A. dominant.
    B. recessive.
    C. codominant.
    D. corecessive.

13. A mutation in the DNA that changes an amino acid in a protein sequence without affecting the activity of the protein is called a:
    A. recessive mutation.
    B. neutral substitution.
    C. leaky mutation.
    D. null mutation.

14. The *w* gene for eye color in *Drosophila* is an example of:
    A. partial dominance.
    B. multiple alleles.
    C. codominance.
    D. a leaky mutation.

15. Which blood group is the null phenotype?
    A. A
    B. B
    C. AB
    D. O

16. Any given segment of a genome could have_____ possible reading frame(s) in a single strand of the DNA.
    A. 1
    B. 2
    C. 3
    D. 4

17. What types of mutations do acridines cause?
    A. frameshift mutations
    B. point mutations
    C. long deletions
    D. long insertions

18. The change in base sequence from AAA AGC TTC GAC to AAA GCT TCG ACC is an example of a(n):
    A. point mutation.
    B. insertion.
    C. frameshift mutation.
    D. deamination.

19. The convention for writing one strand of the DNA sequence of a protein-coding gene is:
    A. 5′ to 3′, with the sequence being the same as for the mRNA (except for T instead of U).
    B. 3′ to 5′, with the sequence being the same as for the mRNA (except for T instead of U).
    C. 5′ to 3′, with the sequence being opposite that for the mRNA (except for T instead of U).
    D. 3′ to 5′, with the sequence being opposite that for the mRNA (except for T instead of U).

20. What type of mutation would prevent both alleles of a gene from being expressed?
    A. dominant
    B. recessive
    C. *trans*-acting
    D. *cis*-acting

Key Terms

acridines
allele
annealing
antiparallel
antisense (template) strand
back mutation
central dogma
chiasma
chromosome
***cis*-acting sequence**
cistron
closed (blocked) reading frame
coding (sense) strand
coding region
codon
colinearity
complementary
complementation test
denaturation
DNA polymerase
DNase
endonuclease
exon
exonuclease
forward mutation
frameshift
gain-of-function mutation
gene expression
genetic code
genetic recombination
genome
heteroduplex DNA
heteromultimer
homomultimer
hotspots
hybridization
induced mutations
initiation codon
intron
leader (5′ UTR)
linkage
locus
loss-of-function mutation
major groove
melting temperature
messenger RNA (mRNA)
minor groove
mutagens
neutral substitutions
nucleoside
nucleotide
null mutation
one gene–one enzyme hypothesis
one gene–one polypeptide hypothesis
open reading frame (ORF)
overwound
point mutation
polymorphism
polynucleotide
pre-mRNA
prion
purine
pyrimidine
reading frame
renaturation
replication fork
reverse transcription
revertants
ribosomal RNAs (rRNAs)
ribosome
RNA polymerase
RNA processing
RNase
second-site reversion
semiconservative replication
silent mutation
splicing
spontaneous mutations
supercoiling
suppression mutation
termination codon
trailer (3′ UTR)
***trans*-acting sequence**
transcription

transfection
transfer RNA (tRNA)
transformation
transition
translation
transversion
true reversion
underwound
unidentified reading frame (URF)
viroid

Further Reading

Carter, C. W., Jr. (2008). Whence the genetic code?: thawing the 'frozen accident.' *Heredity* **100**, 339–340. A brief review of current hypotheses for the origin of the genetic code.

Holmes, F. (2001). *Meselson, Stahl, and the Replication of DNA: A History of the Most Beautiful Experiment in Biology*. Yale University Press, New Haven, CT. An account of Meselson and Stahl's scientific partnership with a unique look into the daily business of "doing science."

Maki, H. (2002). Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annu. Rev. Genet.* **36**, 279–303. A review of causes and effects of spontaneous mutations and mutational hotspots.

Prusiner, S. B. (1998). Prions. *Proc. Natl. Acad. Sci. USA* **95**, 13363–13383. An edited version of Prusiner's Nobel lecture, including an overview of the biology of prions and an account of their discovery.

Ripley, L. S. (1990). Frameshift mutation: determinants of specificity. *Annu. Rev. Genet.* **24**, 189–213. A review of the mechanisms of frameshift mutations.

Watson, J. D. (1981). *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (Norton Critical Editions). W. W. Norton, New York, NY. Watson's 1968 best-selling personal account of the discovery of the double helix along with reprints of original publications and additional commentary.

Yanofsky, C. (2001). Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism. *Annu. Rev. Biochem.* **70**, 1–37. A personal account of Yanofsky's research, including establishing the colinearity of genes and their protein products and the regulation of *trp* expression via attenuation.