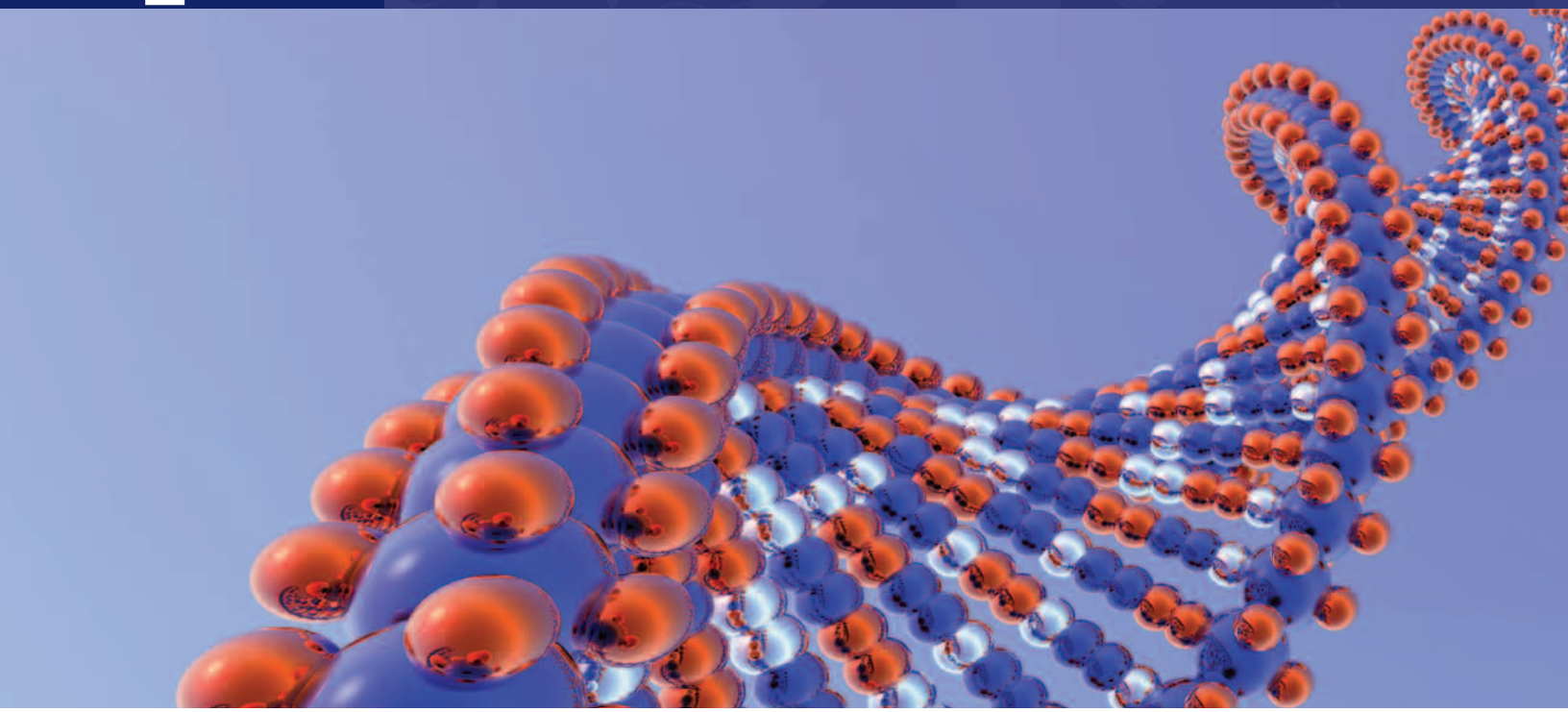
PART I

# Genes and Chromosomes

Photo courtesy of S. V. Flores, A. Mena, and B. F. McAllister. Used with permission of Bryant McAllister, Department of Biology, University of Iowa.

# 1



A strand of DNA in blue and red. DNA is the genetic material of eukaryotic cells, bacteria, and many viruses. © Artsilensecome/ ShutterStock, Inc.

# Genes Are DNA

## CHAPTER OUTLINE

## 1.1 Introduction

The hereditary basis of every living organism is its **genome**, a long sequence of deoxyribonucleic acid (DNA) that provides the complete set of hereditary information carried by the organism as well as its individual cells. The genome includes chromosomal DNA as well as DNA in plasmids and (in eukaryotes) organellar DNA, as found 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. The products of expression of nucleotide sequences within the genome determine development. By a complex series of interactions, the DNA sequence produces all of the proteins of the organism at the appropriate time and within the appropriate cells. Proteins serve a diverse series of roles in the development and functioning of an organism; they can form part of the structure of the organism, have the capacity to build the structures, perform the metabolic reactions necessary for life, and participate in regulation as transcription factors, receptors, key players in signal transduction pathways, and other molecules.

Physically, the genome may be divided into a number of different DNA molecules, or **chromosomes**. The ultimate definition of a genome is the sequence of the DNA of each chromosome. Functionally, the genome is divided into genes. Each gene is a sequence of DNA that encodes a single type of RNA and in many cases, ultimately a polypeptide. Each of the discrete chromosomes comprising the genome may contain a large number of genes. Genomes for living organisms may contain as few as ~500 genes (for a mycoplasma, a type of bacterium), ~20,000 to 25,000 for a human being, or as many as ~50,000 to 60,000 for rice.

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.

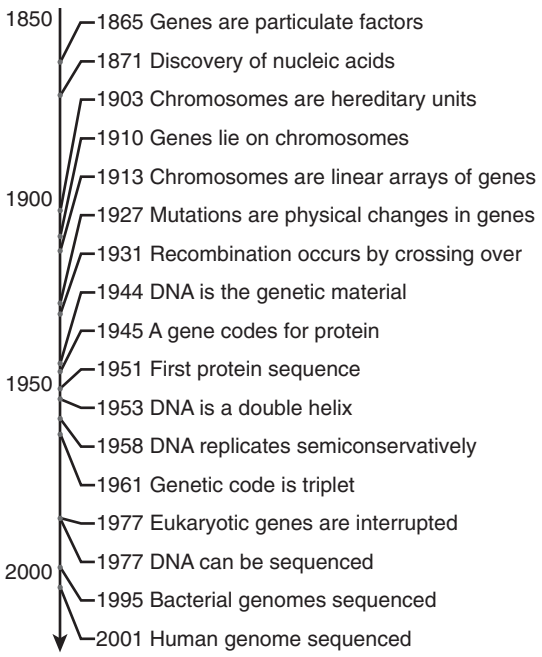



**FIGURE 1.1** A brief history of genetics.

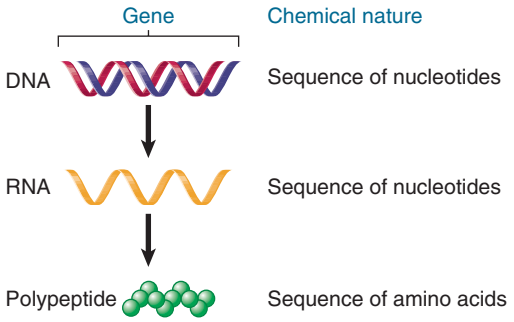



**FIGURE 1.2** A gene encodes an RNA, which may encode a polypeptide.

The first definition of the gene as a functional unit followed from the discovery that individual genes are responsible for the production of specific proteins. Later, the chemical differences between the DNA of the gene and its protein product led to the suggestion that a gene encodes a protein. This in turn led to the discovery of the complex apparatus by which the DNA sequence of a gene determines the amino acid sequence of 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 book. A gene is a sequence of DNA that directly produces a single strand of another nucleic acid, RNA, with a sequence that is 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. In other cases, such as rRNA and tRNA 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, or **structural**, genes, the RNA in turn encodes a polypeptide.

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

## 1.2 DNA Is the Genetic Material of Bacteria and Viruses

**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.

The idea that the genetic material of organisms is DNA has its roots in the discovery of **transformation** by Frederick Griffith in 1928. The bacterium *Streptococcus* (formerly *Pneumococcus*) *pneumoniae* kills mice by causing pneumonia. The virulence of the bacterium is determined by its capsular polysaccharide, which allows the bacterium to escape destruction by its host. Several types of *S. pneumoniae* have different capsular polysaccharides, but they all have a smooth (S) appearance. Each of the S types can give rise to variants that fail to produce the capsular polysaccharide and therefore have a rough (R) surface (consisting of the material that was beneath the capsular polysaccharide). The R types are avirulent and do not kill the mice because the absence of the polysaccharide capsule allows the animal's immune system to destroy the bacteria.

When S bacteria are killed by heat treatment, they can no longer harm the animal. FIGURE 1.3, however, shows that when heat-killed S bacteria and avirulent R bacteria are jointly injected into a mouse, it dies as the result of a pneumonia infection. Virulent S bacteria can be recovered from the mouse's blood.

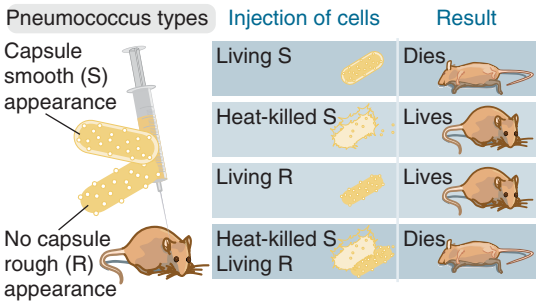



FIGURE 1.3 Neither heat-killed S-type nor live R-type bacteria can kill mice, but simultaneous injection of both can kill mice just as effectively as the live S type.

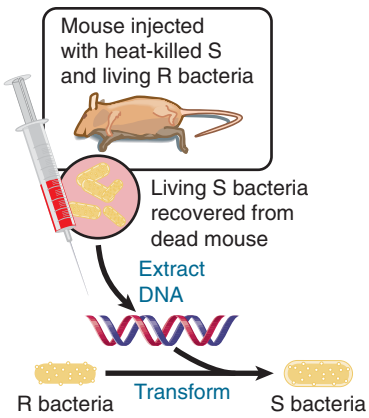



FIGURE 1.4 The DNA of S-type bacteria can transform R-type bacteria into the same S type.

In this experiment, the dead S bacteria were of type III. The live R bacteria had been derived from type II. The virulent bacteria recovered from the mixed infection had the smooth coat of type III. So, some property of the dead IIIS bacteria can transform the live IIR bacteria so that they make the capsular polysaccharide and become virulent. FIGURE 1.4 shows the identification of the component of the dead bacteria responsible for transformation. This was called the **transforming principle**. It was purified in a cell-free system in which extracts from the dead IIIS bacteria were added to the live IIR bacteria before being plated on agar and assayed for transformation (FIGURE 1.5). Purification of the transforming principle in 1944 by Avery, MacLeod, and McCarty showed that it is DNA.

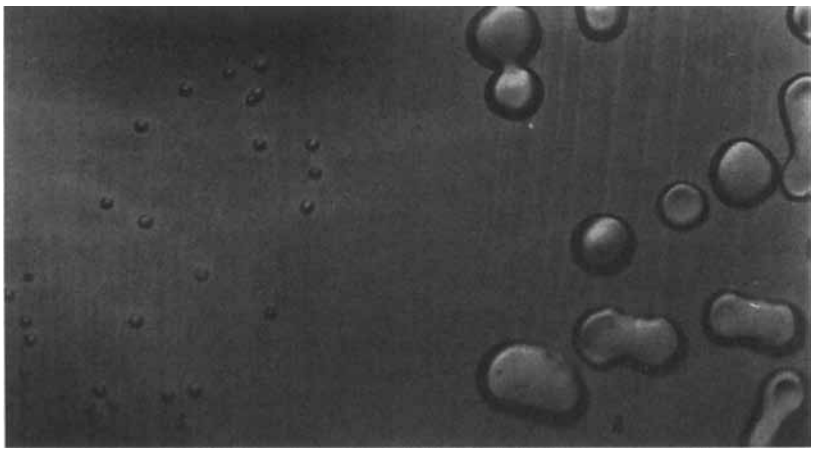


FIGURE 1.5 Rough (left) and smooth (right) colonies of *S. pneumoniae*. © Avery, et al., 1944. Originally published in *The Journal of Experimental Medicine,* 79: 137–158. Used with permission of The Rockefeller University Press.

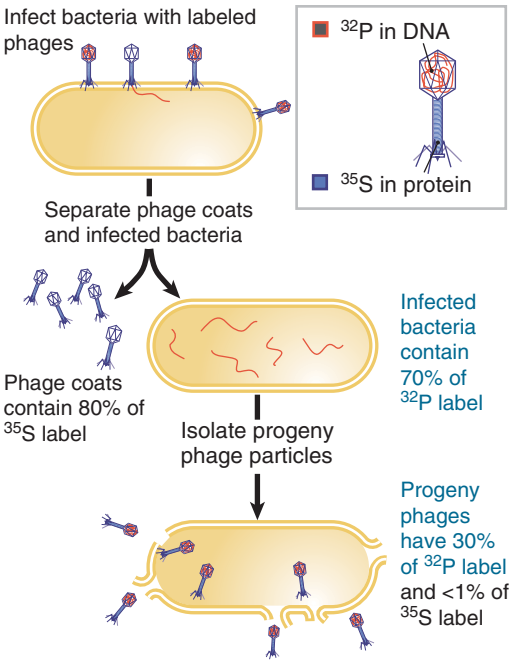



**FIGURE 1.6** The genetic material of phage T2 is DNA.

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 virus 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 ~20 minutes later each cell bursts open, or lyses, to release a large number of progeny phage.

**FIGURE 1.6** illustrates the results of an experiment conducted 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 phosphorus-32 [$^{32}$P]) or in their protein component (with sulfur-35 [$^{35}$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 ~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 consisted 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 expected behavior of genetic material.

A phage reproduces by commandeering the replication 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 the genetic material of the genome of a cell or a virus.

## 1.3 DNA Is the Genetic Material of Eukaryotic Cells

**Key concepts**

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

When DNA is added to eukaryotic cells growing in culture, it can enter 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 1.7**.

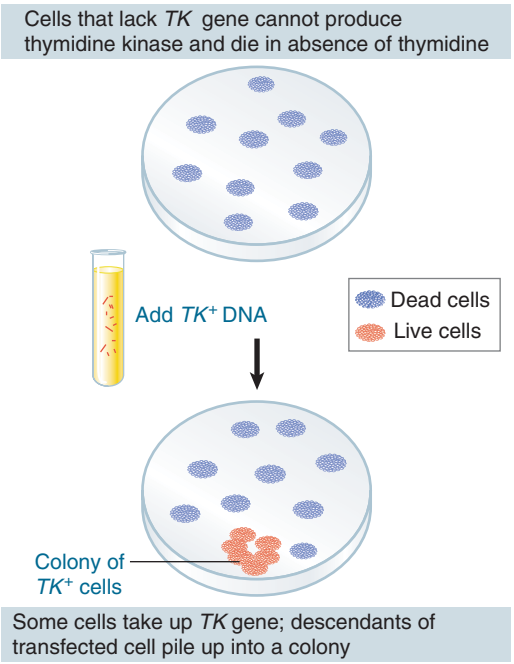



**FIGURE 1.7** Eukaryotic cells can acquire a new phenotype as the result of transfection by added DNA.

Although for historical reasons these experiments are described as **transfection** when performed with animal cells, they are analogous to bacterial transformation. The 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 upon the cells (synthesis of thymidine kinase in the example of Figure 1.7). 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.

The genetic material of all known organisms and many viruses is DNA. Some viruses, though, use RNA as the genetic material. As a result, the general nature of the genetic material is that it is always nucleic acid; specifically, it is DNA, except in the RNA viruses.

## 1.4 Polynucleotide Chains Have Nitrogenous Bases Linked to a Sugar–Phosphate Backbone

**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 has uracil instead of thymine.

The basic building block of nucleic acids (DNA and RNA) is the nucleotide, which has three components:

1. A nitrogenous base
2. A sugar
3. 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**. To avoid ambiguity between the numbering systems of the heterocyclic rings and the sugar, positions on the pentose are given a prime (′).

Nucleic acids are named for the type of sugar: DNA has 2′–deoxyribose, whereas RNA has ribose. The difference is that the sugar in RNA has 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 1.8 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 5′ carbon of one pentose ring to the 3′ carbon of the next pentose ring via a phosphate group; thus the sugar–phosphate backbone is said to consist of 5′–3′ phosphodiester linkages. Specifically, the 3′ carbon of one pentose is bonded to one oxygen of the phosphate, while the 5′ carbon of the other pentose is bonded to the opposite oxygen of the phosphate. The nitrogenous bases "stick out" from the backbone.

Each nucleic acid contains four types of nitrogenous bases. 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

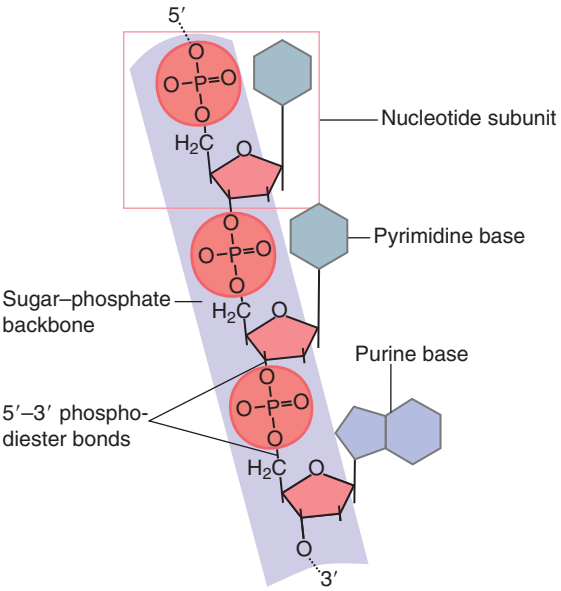



**FIGURE 1.8** A polynucleotide chain consists of a series of 5′–3′ sugar–phosphate links that form a backbone from which the bases protrude.

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.

## 1.5 Supercoiling Affects the Structure of DNA

**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.
- A closed DNA molecule has a linking number (L), which is the sum of twist (T) and writhe (W).
- The linking number can be changed only by breaking and reforming bonds in the DNA backbone.

The two strands of DNA are wound around each other to form a double helical structure (described in detail in the next section); 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 thus can only occur if the DNA has no free ends (otherwise the free ends can rotate to relieve the tension) or in linear DNA (FIGURE 1.9, top) 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 nonsupercoiled circular DNA lying flat in Figure 1.9 (center) with the supercoiled circular molecule that forms a twisted (and therefore more condensed) shape (Figure 1.9, bottom).

The consequences of supercoiling depend on whether the DNA is twisted around itself in the same direction as the two strands within the double helix (clockwise) or in the opposite direction. Twisting in the same direction produces *positive supercoiling*, which overwinds the DNA so that there are fewer base pairs per turn. Twisting in the opposite direction produces *negative supercoiling*, or underwinding, so there are more base pairs per turn. Both types of supercoiling of the double helix in space are

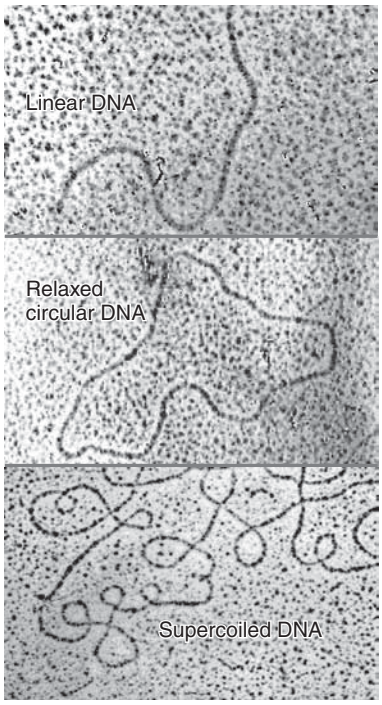



**FIGURE 1.9** Linear DNA is extended (top); a circular DNA remains extended if it is relaxed (nonsupercoiled; center); but a supercoiled DNA has a twisted and condensed form (bottom). Photos courtesy of Nirupam Roy Choudhury, International Centre for Genetic Engineering and Biotechnology (ICGEB).

tensions in the DNA (which is why DNA molecules with no supercoiling are called "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 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 1.10.

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. DNA in a typical chromosome, however, is not only extremely long, but is also coated with proteins that serve to anchor the DNA at numerous points. As a result, even a linear eukaryotic chromosome does not functionally possess free ends.

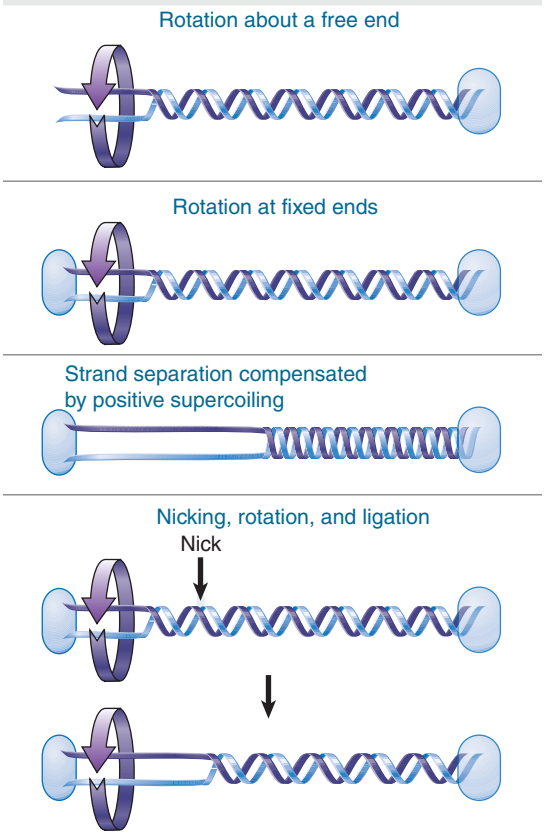



**FIGURE 1.10** Separation of the strands of a DNA double helix can be achieved in several ways.

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.

A closed molecule of DNA can be characterized by its **linking number** (**L**), which is the number of times one strand crosses over the other in space. Closed DNA molecules of identical sequence may have different linking numbers, reflecting different degrees of supercoiling. Molecules of DNA that are the same except for their linking numbers are called *topological isomers*.

The linking number is made up of two components: the **writhing number** (**W**) and the **twisting number** (**T**). The twisting number, T, is a property of the double helical structure itself, representing the rotation of one strand about the other. It represents the total number of turns of the duplex and is determined by the number of base pairs per turn. For a relaxed closed circular DNA lying flat in a plane, the twist is the total number of base pairs divided by the number of base pairs per turn. The writhing number, W, represents the turning of the axis of the duplex in space. It corresponds to the intuitive concept of supercoiling, but does not have exactly the same quantitative definition or measurement. For a relaxed molecule, $W = 0$, and the linking number equals the twist.

We are often concerned with the change in linking number, $\Delta L$, given by the equation:

$$\Delta L = \Delta W + \Delta T$$

The equation states that any change in the total number of revolutions of one DNA strand about the other can be expressed as the sum of the changes of the coiling of the duplex axis in space ($\Delta W$) and changes in the helical repeat of the double helix itself ($\Delta T$). In the absence of protein binding or other constraints, the twist of DNA does not tend to vary—in other words, the 10.5 base pairs per turn (bp/turn) helical repeat is a very stable conformation for DNA in solution. Thus, any $\Delta L$ (change in linking number) is mostly likely to be expressed by a change in W; that is, by a change in supercoiling.

A decrease in linking number (that is, a change of $-\Delta L$) corresponds to the introduction of some combination of negative supercoiling ($\Delta W$) and/or underwinding ($\Delta T$). An increase in linking number, measured as a change of $+\Delta L$, corresponds to an increase in positive supercoiling and/or overwinding.

We can describe the change in state of any DNA by the specific linking difference, $\sigma = \Delta L/L0$, for which L0 is the linking number when the DNA is relaxed. If all of the change in linking number is due to change in W (that is, $\Delta T = 0$), the specific linking difference equals the supercoiling density. In effect, $\sigma$ as defined in terms of $\Delta L/L0$ can be assumed to correspond to supercoiling density so long as the structure of the double helix itself remains constant.

The critical feature about the use of the linking number is that this parameter is an invariant property of any individual *closed* DNA molecule. The linking number cannot be changed by any deformation short of one that involves the breaking and rejoining of strands. A circular molecule with a particular linking number can express the number in terms of different combinations of T and W, but it cannot change their

sum so long as the strands are unbroken. (In fact, the partitioning of L between T and W prevents the assignment of fixed values for the latter parameters for a DNA molecule in solution.)

The linking number is related to the actual enzymatic events by which changes are made in the topology of DNA. The linking number of a particular closed molecule can be changed only by breaking one or both strands, using the free end to rotate one strand about the other, and rejoining the broken ends. When an enzyme performs such an action, it must change the linking number by an integer; this value can be determined as a characteristic of the reaction. The reactions to control supercoiling in the cell are performed by topoisomerase enzymes (this is explored in more detail in the chapter titled *DNA Replication*).

## 1.6 DNA Is a Double Helix

**Key concepts**

- The B-form of DNA is a double helix consisting of two polynucleotide chains that run antiparallel.
- The nitrogenous bases of each chain are flat purine or pyrimidine rings that face 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 base pairs per turn in solution).
- The double helix has a major (wide) groove and a minor (narrow) groove.

By the 1950s, the observation by Erwin Chargaff that the bases are present in different amounts in the DNAs of different species led to the concept that the sequence of bases is the form in which genetic information is carried. 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 (which is more hydrated than the A-form) is a regular helix, making a complete turn every 34 Å (3.4 nm), with a diameter of ~20 Å (2 nm). The distance between adjacent nucleotides is 3.4 Å (0.34 nm); thus there must be 10 nucleotides per turn. (In aqueous solution, the structure averages 10.4 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 most stably with C, whereas A can bond most stably 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**. Complementary base pairing occurs because of complementary shapes of the complementary 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.

The Watson–Crick model has the two polynucleotide chains running in opposite directions, so they are said to be **antiparallel**, as illustrated in FIGURE 1.11. Looking in one direction along the helix, one strand runs in the 5′ to 3′ direction, whereas its complement runs 3′ to 5′

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 $Na^+$. In the cell, positively charged proteins provide some of the neutralizing force. These proteins play important roles in determining the organization of DNA in the cell.

The base pairs 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

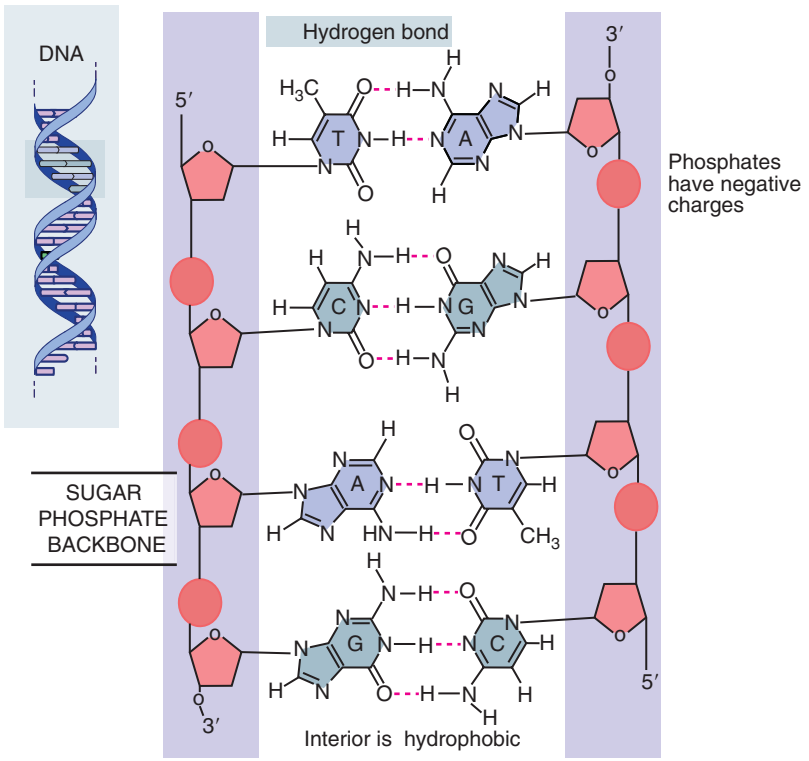



**FIGURE 1.11** The double helix maintains a constant width because purines always face pyrimidines in the complementary A-T and G-C base pairs. The sequence in the figure is T-A, C-G, A-T, G-C.

double helix as a spiral staircase, the base pairs form the steps, as illustrated schematically in **FIGURE 1.12**. Proceeding up the helix, bases are stacked above one another like a pile of plates.

Each base pair is rotated ~36° around the axis of the helix relative to the next base pair, so ~10 base pairs make a complete turn of 360°. The twisting of the two strands around one another forms a double helix with a **minor groove** that is ~12 Å (1.2 nm) across and a **major groove** that is ~22 Å (2.2 nm) across, as can be seen from the scale model of **FIGURE 1.13**. 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, observed when DNA is dehydrated, is also a right-handed helix and 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.)

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. If DNA has more base pairs per turn 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.

Another structural variant is *bent DNA*. A series 8 to 10 adenine residues on one strand can result in intrinsic bending of the double helix. This structure allows tighter packing with consequences for nucleosome assembly (see the chapter titled *Chromatin*) and gene regulation.

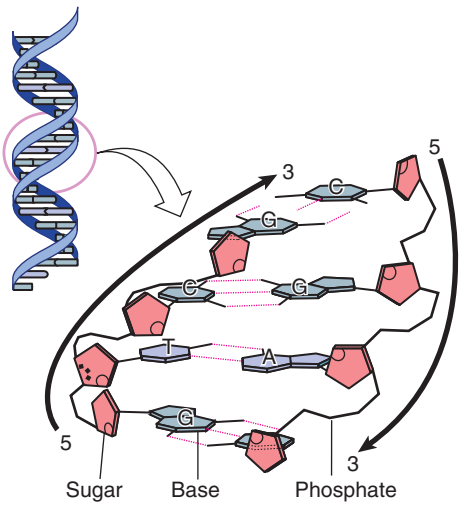



**FIGURE 1.12** Flat base pairs lie perpendicular to the sugar–phosphate backbone.

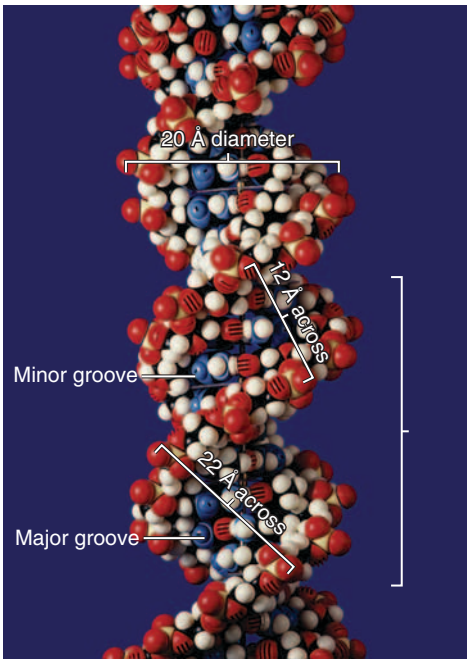



**FIGURE 1.13** The two strands of DNA form a double helix. © Photodisc.

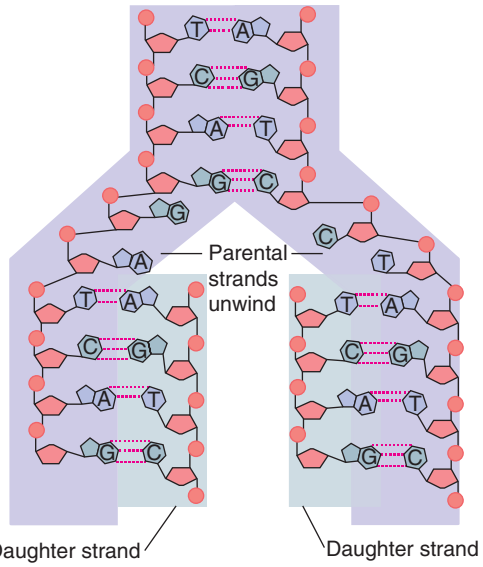



**FIGURE 1.14** Base pairing provides the mechanism for replicating DNA.

## 1.7 DNA Replication Is Semiconservative

**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.

It is crucial that DNA is reproduced accurately. The two polynucleotide strands are joined only by hydrogen bonds, so they are able to separate without the breakage of covalent bonds. The specificity of base pairing suggests that both of the separated parental strands could act as template strands for the synthesis of complementary daughter strands. **FIGURE 1.14** shows the principle that a new daughter strand is assembled from each 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 1.14 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 1.15**. The parental duplex is replicated to form two daughter duplexes, each of which consists of one parental strand and one newly synthesized daughter strand. The unit conserved from one generation to the next is one of the two individual strands comprising the parental duplex.

Figure 1.15 illustrates a prediction of this model. If the parental DNA carries a "heavy" density label because the organism has been

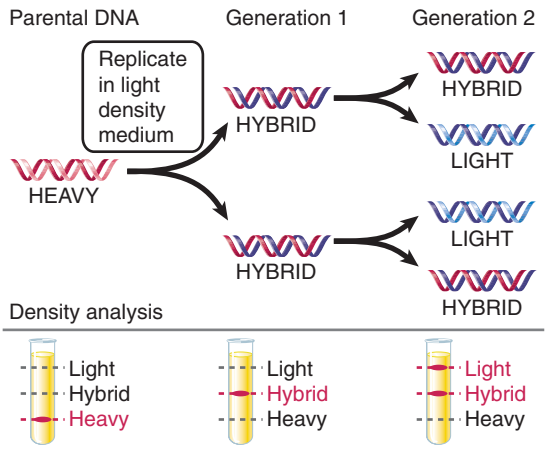



**FIGURE 1.15** Replication of DNA is semiconservative.

grown in a 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 a "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 and half light in the second generation.

## 1.8 Polymerases Act on Separated DNA Strands at the Replication Fork

**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.

Replication requires the two strands of the parental duplex to undergo separation, or **denaturation**. The disruption of the duplex, however, is 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. ("Denaturation" is also used to describe the loss of functional protein structure; it is a general term implying that the natural conformation of a macromolecule has been converted to some nonfunctional form.)

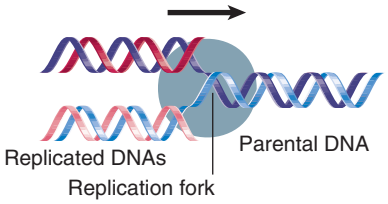



**FIGURE 1.16** The replication fork is the region of DNA in which there is a transition from the unwound parental duplex to the newly replicated daughter duplexes.

The helical structure of a molecule of DNA during replication is illustrated in **FIGURE 1.16**. 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 ligase that connects discontinuous DNA strands. Degradation of nucleic acids also requires specific enzymes: deoxyribonucleases (**DNases**) degrade DNA, and ribonucleases (**RNases**) degrade RNA. The nucleases 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 1.17**.
- Exonucleases remove nucleotide residues one at a time from the end of the molecule, generating mononucleotides. They always function on a single nucleic acid

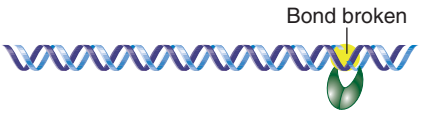



**FIGURE 1.17** An endonuclease cleaves a bond within a nucleic acid. This example shows an enzyme that attacks one strand of a DNA duplex.

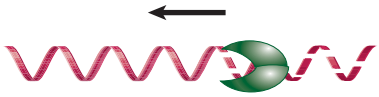


**FIGURE 1.18** An exonuclease removes bases one at a time by cleaving the last bond in a polynucleotide chain.

strand and each exonuclease proceeds in a specific direction, that is, starting either at a 5′ or a 3′ end and proceeding toward the other end. They are involved in trimming reactions, as shown in **FIGURE 1.18**.

## 1.9 Genetic Information Can Be Provided by DNA or RNA

**Key concepts**

- Cellular genes 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.

The **central dogma** is the dominant paradigm of molecular biology. Structural genes exist as sequences of nucleic acid but function by being expressed in the form of polypeptides. Replication makes possible the inheritance of genetic information, whereas transcription and translation are responsible for its expression to another form.

**FIGURE 1.19** illustrates the roles of replication, transcription, and translation in the context of the so-called *central dogma*:

- 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 (discussed shortly). Generally, polypeptides cannot be retrieved for use as genetic information; translation of RNA into polypeptide is always irreversible.

These mechanisms are equally effective for the cellular genetic information of prokaryotes or eukaryotes and for the information carried by viruses. The genomes of all living organisms consist of duplex DNA. Viruses have genomes that consist of DNA or RNA and there are examples of each type that are double-stranded (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 1.20**.

Cellular genomes reproduce DNA by the mechanism of semiconservative replication.

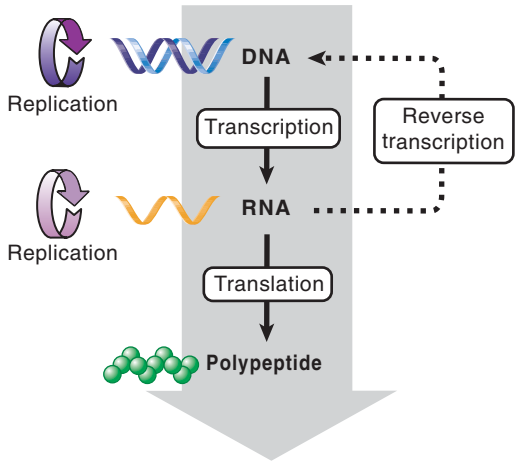



**FIGURE 1.19** The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into a polypeptide is irreversible.

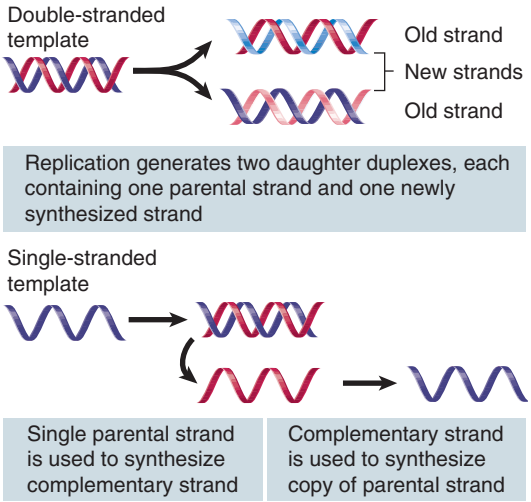



**FIGURE 1.20** Double-stranded and single-stranded nucleic acids both replicate by synthesis of complementary strands governed by the rules of base pairing.

Double-stranded viral genomes, whether DNA or RNA, also replicate by using the individual strands of the duplex as templates to synthesize complementary strands.

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

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 a ds DNA. This duplex DNA becomes part of the genome of the host cell and is inherited like any other gene. Thus reverse transcription allows a sequence of RNA to be retrieved and used as DNA in a cell.

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. On rare occasions though (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.

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 1.21** presents some examples that illustrate the range of genome types and sizes. The reasons for such variation in genome size and gene number will be explored in the chapters titled *The Content of the Genome* and *Genome Sequences and Gene Numbers*.

Among the various living organisms, with genomes varying in size over a 100,000-fold range, a common principle prevails: The DNA encodes all the proteins that the cell(s) of the organism must synthesize and the proteins in turn (directly or indirectly) provide the functions needed for survival. A similar principle describes the function of the genetic information of viruses, whether DNA or RNA: The nucleic acid encodes the protein(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.)

| 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 1.21** The amount of nucleic acid in the genome varies over an enormous range.

## 1.10 Nucleic Acids Hybridize by Base Pairing

**Key concepts**

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

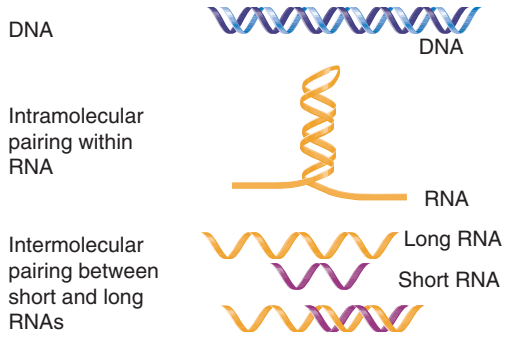



**FIGURE 1.22** Base pairing occurs in duplex DNA and also in intra- and intermolecular interactions in single-stranded RNA (or DNA).

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 1.22** 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 important for 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** ($\mathbf{T_m}$) and it depends on the G-C content of the duplex. Each G-C base pair has three hydrogen bonds; as a result 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 (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 1.23** 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. 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.

The purpose of the hybridization reaction is to combine two single-stranded nucleic

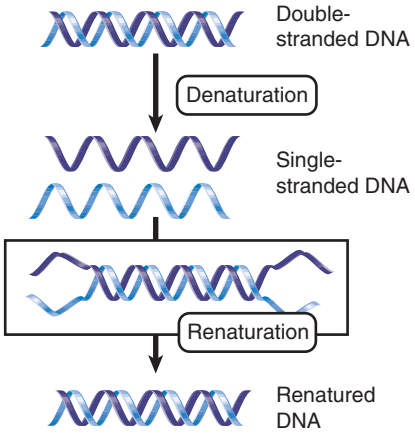



**FIGURE 1.23** Denatured single strands of DNA can renature to give the duplex form.

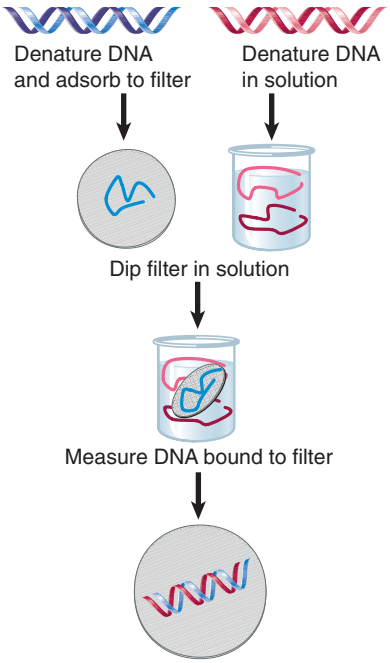



**FIGURE 1.24** Filter hybridization establishes whether a solution of denatured DNA (or RNA) contains sequences complementary to the strands immobilized on the filter.

acids in solution and then to measure the amount of double-stranded material that forms. **FIGURE 1.24** illustrates a procedure in which a DNA preparation is denatured and the single strands are attached to a filter. A second denatured DNA (or RNA) preparation is then 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 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.

The extent of hybridization between two single-stranded nucleic acids is determined by their complementarity. Two sequences need not be perfectly complementary to hybridize. If they are 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.

## 1.11 Mutations Change the Sequence of DNA

**Key concepts**

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

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

All organisms experience a certain number of mutations as the result of normal cellular operations or random interactions with the environment. 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 a smaller unit is observed.

Spontaneous mutations that inactivate gene function occur in bacteriophages and bacteria at a relatively constant rate of $3–4 \times 10^{-3}$ 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.

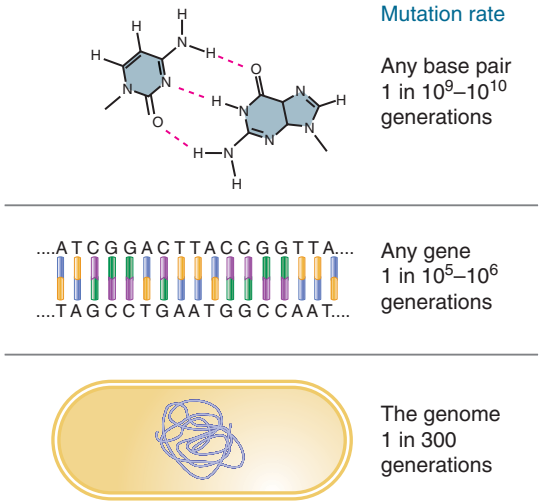



**FIGURE 1.25** A base pair is mutated at a rate of $10^{-9}$–$10^{-10}$ per generation, a gene of 1000 bp is mutated at $\sim 10^{-6}$ per generation, and a bacterial genome is mutated at $3 \times 10^{-3}$ per generation.

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. Such a 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 1.25** shows that in bacteria, the mutation rate corresponds to $\sim 10^{-6}$ events per locus per generation or to an average rate of change per base pair of $10^{-9}$–$10^{-10}$ per generation. The rate at individual base pairs varies very widely, over a 10,000-fold range. We have no accurate measurement of the rate of mutation in eukaryotes, although usually it is thought to be somewhat similar to that of bacteria on a per-locus, per-generation basis.

## 1.12 Mutations May Affect Single Base Pairs or Longer Sequences

**Key concepts**

- A point mutation changes a single base pair.
- Point mutations can be caused by the chemical conversion of one base into another or by 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 and/or deletions can result from the movement of transposable elements.

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

- Chemical modification of DNA directly changes one base into a different base.
- 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**, which results from the substitution of one pyrimidine by the other, or of one purine by the other. This replaces a G-C pair 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 an A-T pair becomes a T-A or C-G pair.

As shown in **FIGURE 1.26**, 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, which occurs when noncomplementary bases pair instead of the usual Watson–Crick

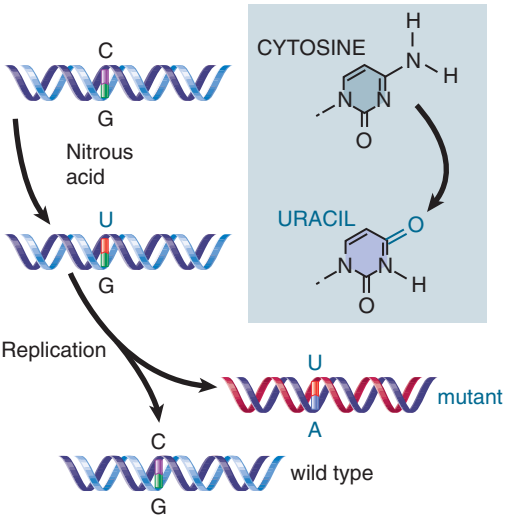



**FIGURE 1.26** Mutations can be induced by chemical modification of a base.

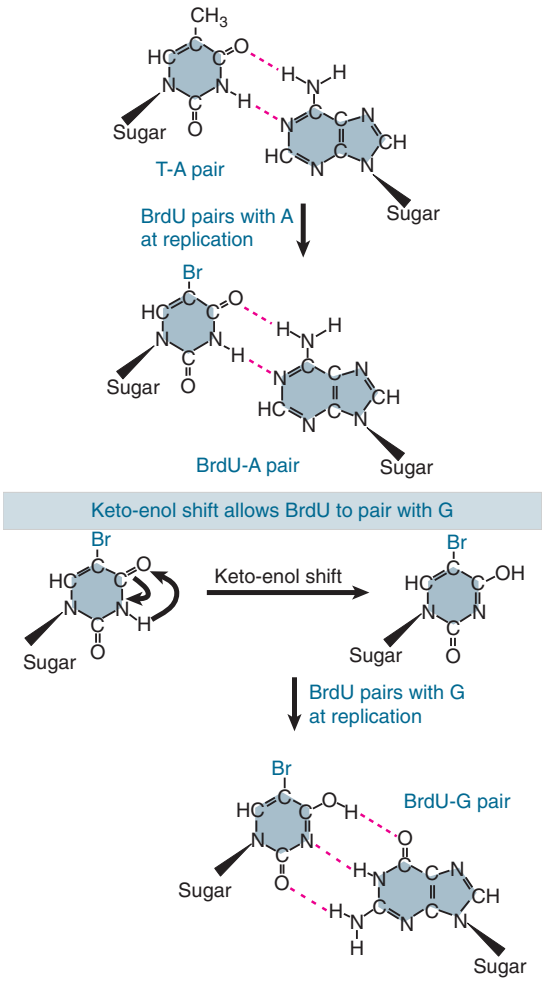



**FIGURE 1.27** Mutations can be induced by the incorporation of base analogs into DNA.

pairs. Base mispairing usually occurs as an aberration resulting from the incorporation into DNA of an abnormal base that has flexible pairing properties. **FIGURE 1.27** 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 can be incorporated into DNA in place of thymine. BrdU has flexible pairing properties, though, 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.

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

Point mutations were thought for a long time to be the principal means of change in individual genes. We now know, though, that insertions of short sequences are quite frequent. 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 because it may alter the reading frame; such an insertion is a *frameshift mutation*. (Similarly, a deletion within a coding region is usually a frameshift mutation.) Where such insertions have occurred, deletions of part or all of the inserted material, and sometimes of the adjacent regions, may subsequently occur.

A significant difference between point mutations and insertions is that mutagens can increase the frequency of point mutations, but do not affect the frequency of transposition. Both insertions and deletions of short sequences (often called *indels*) can occur by other mechanisms, however—for example, those involving errors during replication or recombination. In addition, a class of mutagens called the acridines introduces very small insertions and deletions.

## 1.13 The Effects of Mutations Can Be Reversed

**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.

**FIGURE 1.28** shows that the possibility of reversion mutations, or **revertants**, is an important characteristic that distinguishes point mutations and insertions from deletions:

- A point mutation can revert either by restoring the original sequence or by gaining a compensatory mutation elsewhere in the gene.
- 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.

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. An exact reversal of the original mutation is

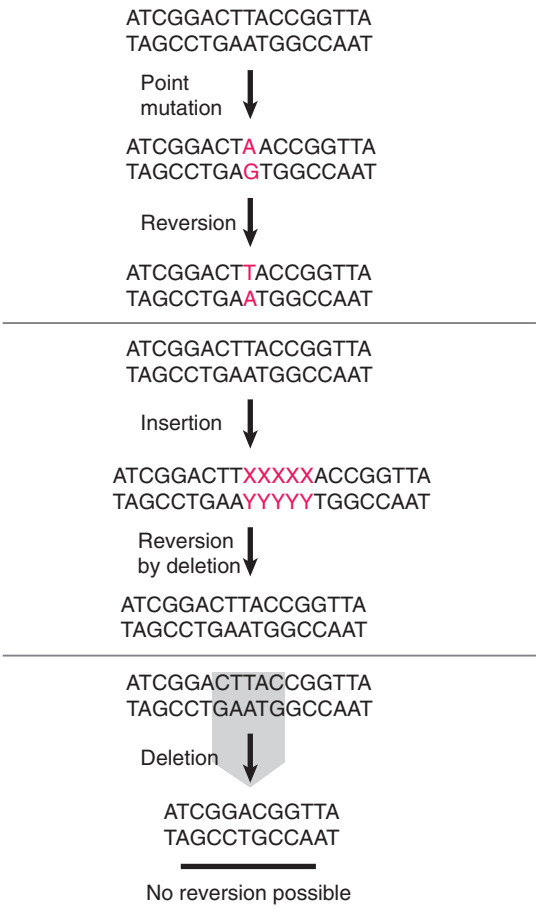



**FIGURE 1.28** Point mutations and insertions can revert, but deletions cannot revert.

called a **true reversion**. Consequently, if an A-T pair has been replaced by a G-C pair, 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. 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 gene 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. The possibilities for back mutations are thus 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 ~10.

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, while 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.)

## 1.14 Mutations Are Concentrated at Hotspots

**Key concept**

- 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.

So far we have dealt with 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. Are all base pairs in a gene equally susceptible, though, 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 in FIGURE 1.29 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

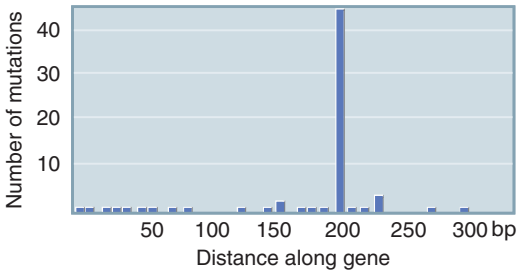



**FIGURE 1.29** Spontaneous mutations occur throughout the *lacI* gene of *E. coli*, but are concentrated at a hotspot.

expected from a random distribution; they may have 10× or even 100× more mutations than predicted by random hits. These sites are called **hotspots**. Spontaneous mutations may occur at hotspots, and different mutagens may have different hotspots.

## 1.15 Many Hotspots Result from Modified Bases

**Key concepts**

- A common cause of hotspots is the modified base 5-methylcytosine, which is spontaneously deaminated to thymine.
- A hotspot can result from the high frequency of change in copy number of a short, tandemly repeated sequence.

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 1.26). **FIGURE 1.30** 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.

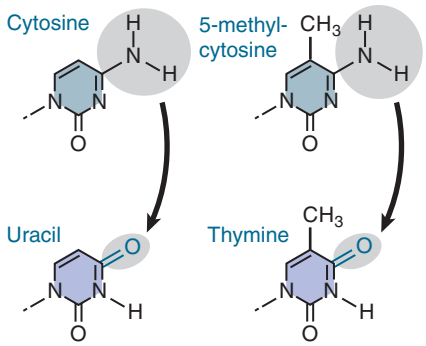



**FIGURE 1.30** Deamination of cytosine produces uracil, whereas deamination of 5-methylcytosine produces thymine.

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

**FIGURE 1.31** 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 accurately repairing G-U than G-T.

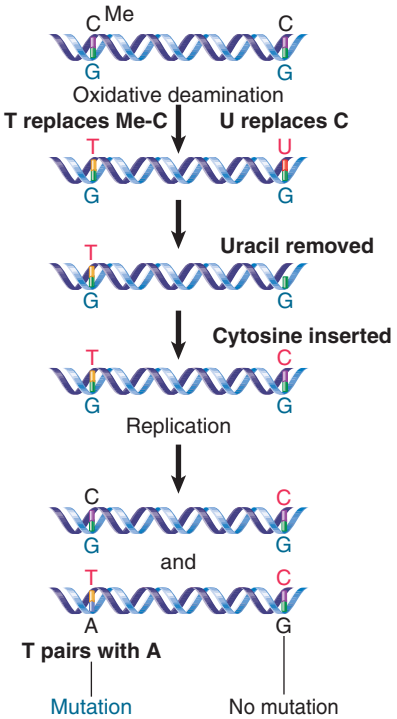



**FIGURE 1.31** The deamination of 5-methylcytosine produces thymine (by C-G to T-A transitions), whereas the deamination of cytosine produces uracil (which usually is removed and then replaced by cytosine).

*E. coli* contain an enzyme, uracil-DNA-glycosidase, that removes uracil residues from DNA (see the chapter titled *Repair Systems*). 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 1.26.)

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 first separate and then pair with the correct complements to produce the wild-type G-C in one daughter DNA and the mutant A-T in the other.

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 chapter titled *Repair Systems*). These systems are not, however, 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.

Additionally, 5-methylcytosine creates hotspots in eukaryotic DNA. It is common in CpG dinucleotides that are concentrated in regions called CpG islands (see the chapter titled *Epigenetic Effects Are Inherited*). Although 5-methylcytosine accounts for ~1% of the bases in human DNA, sites containing the modified base account for ~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; most likely the elimination of all the systems would increase the mutation rate much more.

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

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.

## 1.16 Some Hereditary Agents Are Extremely Small

**Key concept**

- Some very small hereditary agents do not encode polypeptide, but consist of RNA or protein with heritable properties.

**Viroids** (or subviral pathogens) are infectious agents that cause diseases in some 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 1.32**. 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 particular 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.

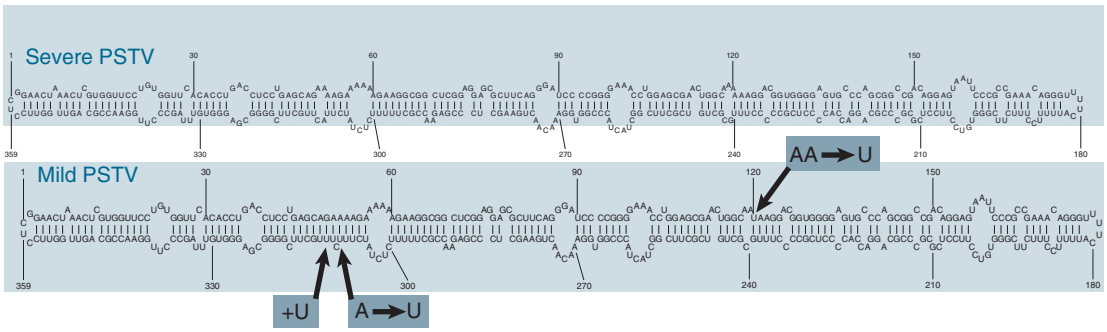



**FIGURE 1.32** PSTV RNA is a circular molecule that forms an extensive double-stranded structure, interrupted by many interior loops. The severe and mild forms of PSTV have RNAs that differ at three sites.

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.

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

An even more unusual agent is 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 disease, 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; 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, and which has yet to be fully defined.

As the infectious agent of scrapie, $PrP^{sc}$ must in some way modify the synthesis of its normal cellular counterpart so that it becomes infectious instead of harmless (see the chapter titled *Epigenetic Effects Are Inherited*). Mice that lack a PrP gene cannot develop scrapie, which demonstrates that PrP is essential for development of the disease.

## 1.17 Summary

Two classic experiments provided strong evidence that DNA is the genetic material of bacteria, viruses, and eukaryotic cells. 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 stretch of DNA may encode a polypeptide. The genetic code describes the relationship between the sequence of DNA and the sequence of the polypeptide. In general, only one of the two strands of DNA encodes a polypeptide. A codon consists of three nucleotides that encode a single amino acid. A coding sequence of DNA consists of a series of codons, which are read from a fixed starting point. In most cases only

one of the three possible reading frames can be translated into polypeptide.

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

The natural incidence of mutations is increased by mutagens. Mutations may be concentrated at hotspots. A type of hotspot responsible for some point mutations is caused by deamination of the modified base 5-methylcytosine. Forward mutations occur at a rate of $\sim 10^{-6}$ per locus per generation; back mutations are rarer. Not all mutations have an effect on the phenotype.

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

## References


### 1.1 Introduction

Review

Cairns, J., Stent, G., and Watson, J. D. (1966). *Phage and the Origins of Molecular Biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

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

### 1.2 DNA Is the Genetic Material of Bacteria and Viruses

Research

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

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

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

### 1.3 DNA Is the Genetic Material of Eukaryotic Cells

Research

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

### 1.6 DNA Is a Double Helix

Review

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.

Research

Franklin, R. E., and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature* 171, 740–741.

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

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

Wilkins, M. F. H., Stokes, A. R., and Wilson, H. R. (1953). Molecular structure of deoxypentose nucleic acids. *Nature* 171, 738–740.

### 1.7 DNA Replication Is Semiconservative

Review

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.

Research

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

### 1.11 Mutations Change the Sequence of DNA

Reviews

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

Drake, J. W., and Balz, R. H. (1976). The biochemistry of mutagenesis. *Annu. Rev. Biochem*. 45, 11–37.

### Research

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

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

## 1.12 Mutations May Affect Single Base Pairs or Longer Sequences

### Review

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

## 1.14 Mutations Are Concentrated at Hotspots

### Research

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

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

## 1.16 Some Hereditary Agents Are Extremely Small

### Reviews

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

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

Prusiner, S. B. (1998). Prions. *Proc. Natl. Acad. Sci. USA* 95, 13363–13383.

### Research

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

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