

# Nucleic Acids, Gene Expression, and Recombinant DNA Technology

## CHAPTER 5

### 1 Nucleotides and Nucleic Acids

- A. Nucleotides, Nucleosides, and Bases
- B. The Chemical Structures of DNA and RNA

### 2 DNA Is the Carrier of Genetic Information

- A. Transforming Principle Is DNA
- B. The Hereditary Molecule of Many Bacteriophages Is DNA

### 3 Double Helical DNA

- A. The Watson–Crick Structure: B-DNA
- B. DNA Is Semiconservatively Replicated
- C. Denaturation and Renaturation
- D. The Size of DNA

### 4 Gene Expression and Replication: An Overview

- A. RNA Synthesis: Transcription
- B. Protein Synthesis: Translation
- C. DNA Replication

### 5 Molecular Cloning

- A. Restriction Endonucleases
- B. Cloning Vectors
- C. Gene Manipulation
- D. The Identification of Specific DNA Sequences:
  - Southern Blotting
- E. Genomic Libraries
- F. The Polymerase Chain Reaction
- G. Production of Proteins
- H. Transgenic Organisms and Gene Therapy
- I. Social, Ethical, and Legal Considerations

Knowledge of how genes are expressed and how they can be manipulated is becoming increasingly important for understanding nearly every aspect of biochemistry. Consequently, although we do not undertake a detailed discussion of these processes until Part V of this textbook, we outline their general principles in this chapter. We do so by describing the chemical structures of nucleic acids, how we have come to know that DNA is the carrier of genetic information, the structure of the major form of DNA, and the general principles of how the information in genes directs the synthesis of RNA and proteins (how genes are expressed) and how DNA is replicated. The chapter ends with a discussion of how DNA is experimentally manipulated and expressed, processes that are collectively referred to as genetic engineering. These processes have revolutionized the practice of biochemistry.

### 1 NUCLEOTIDES AND NUCLEIC ACIDS

**Nucleotides** and their derivatives are biologically ubiquitous substances that participate in nearly all biochemical processes:

1. They form the monomeric units of nucleic acids and thereby play central roles in both the storage and the expression of genetic information.

2. **Nucleoside triphosphates**, most conspicuously ATP (Section 1-3B), are the “energy-rich” end products of the majority of energy-releasing pathways and the substances whose utilization drives most energy-requiring processes.

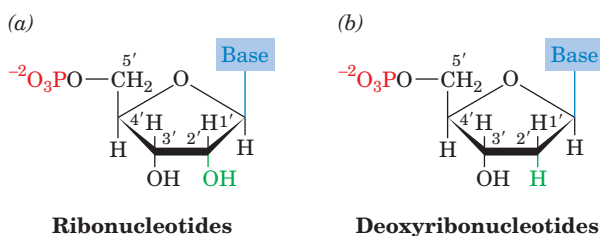
3. Most metabolic pathways are regulated, at least in part, by the levels of nucleotides such as ATP and ADP. Moreover, certain nucleotides, as we shall see, function as intracellular signals that regulate the activities of numerous metabolic processes.

4. Nucleotide derivatives, such as **nicotinamide adenine dinucleotide** (Section 13-2A), **flavin adenine dinucleotide** (Section 16-2C), and **coenzyme A** (Section 21-2), are required participants in many enzymatic reactions.

5. As components of the enzymelike nucleic acids known as **ribozymes**, nucleotides have important catalytic activities themselves.

#### A. Nucleotides, Nucleosides, and Bases

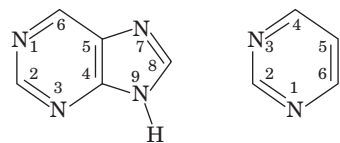
*Nucleotides are phosphate esters of a five-carbon sugar (which is therefore known as a **pentose**; Section 11-1A) in which a nitrogenous base is covalently linked to C1' of the sugar residue. In **ribonucleotides** (Fig. 5-1a), the monomeric units of RNA, the pentose is **D-ribose**, whereas in **deoxyribonucleotides** (or just **deoxynucleotides**; Fig. 5-1b),*



**Figure 5-1** Chemical structures of (a) ribonucleotides and (b) deoxyribonucleotides.

the monomeric units of DNA, the pentose is **2'-deoxy-D-ribose** (note that the “primed” numbers refer to the atoms of the ribose residue; “unprimed” numbers refer to atoms of the nitrogenous base). The phosphate group may be bonded to C5' of the pentose to form a **5'-nucleotide** (Fig. 5-1) or to its C3' to form a **3'-nucleotide**. If the phosphate group is absent, the compound is known as a **nucleoside**. A 5'-nucleotide, for example, may therefore be referred to as a **nucleoside-5'-phosphate**. In all naturally occurring nucleotides and nucleosides, the bond linking the nitrogenous base to the pentose C1' atom (which is called a glycosidic bond; Section 11-1Ca) extends from the same side of the ribose ring as does the C4'—C5' bond (the so-called  $\beta$  configuration; Section 11-1Ba) rather than from the opposite side (the  $\alpha$  configuration). Note that nucleotide phosphate groups are doubly ionized at physiological pH's; that is, *nucleotides are moderately strong acids*.

The nitrogenous bases are planar, aromatic, heterocyclic molecules which, for the most part, are derivatives of either **purine** or **pyrimidine**.



Purine

Pyrimidine

The structures, names, and abbreviations of the common bases, nucleosides, and nucleotides are given in Table 5-1. The major purine components of nucleic acids are **adenine** and **guanine** residues; the major pyrimidine residues are those of **cytosine**, **uracil** (which occurs mainly in RNA), and **thymine** (5-methyluracil, which occurs mainly in DNA). The purines form glycosidic bonds to ribose via

**Table 5-1** Names and Abbreviations of Nucleic Acid Bases, Nucleosides, and Nucleotides

Base Formula	Base (X = H)	Nucleoside (X = ribose <sup>a</sup> )	Nucleotide <sup>b</sup> (X = ribose phosphate <sup>a</sup> )
	Adenine Ade A	Adenosine Ado A	Adenylic acid Adenosine monophosphate AMP
	Guanine Gua G	Guanosine Guo G	Guanylic acid Guanosine monophosphate GMP
	Cytosine Cyt C	Cytidine Cyd C	Cytidylic acid Cytidine monophosphate CMP
	Uracil Ura U	Uridine Urd U	Uridylic acid Uridine monophosphate UMP
	Thymine Thy T	Deoxythymidine dThd dT	Deoxythymidylic acid Deoxythymidine monophosphate dTMP

<sup>a</sup>The presence of a 2'-deoxyribose unit in place of ribose, as occurs in DNA, is implied by the prefixes “deoxy” or “d.” For example, the deoxynucleoside of adenine is deoxyadenosine or dA. However, for thymine-containing residues, which rarely occur in RNA, the prefix is redundant and may be dropped. The presence of a ribose unit may be explicitly implied by the prefixes “ribo” or “r.” Thus the ribonucleotide of thymine is ribothymidine or rT.

<sup>b</sup>The position of the phosphate group in a nucleotide may be explicitly specified as in, for example, 3'-AMP and 5'-GMP.

their N9 atoms, whereas pyrimidines do so through their N1 atoms (note that purines and pyrimidines have dissimilar atom numbering schemes).

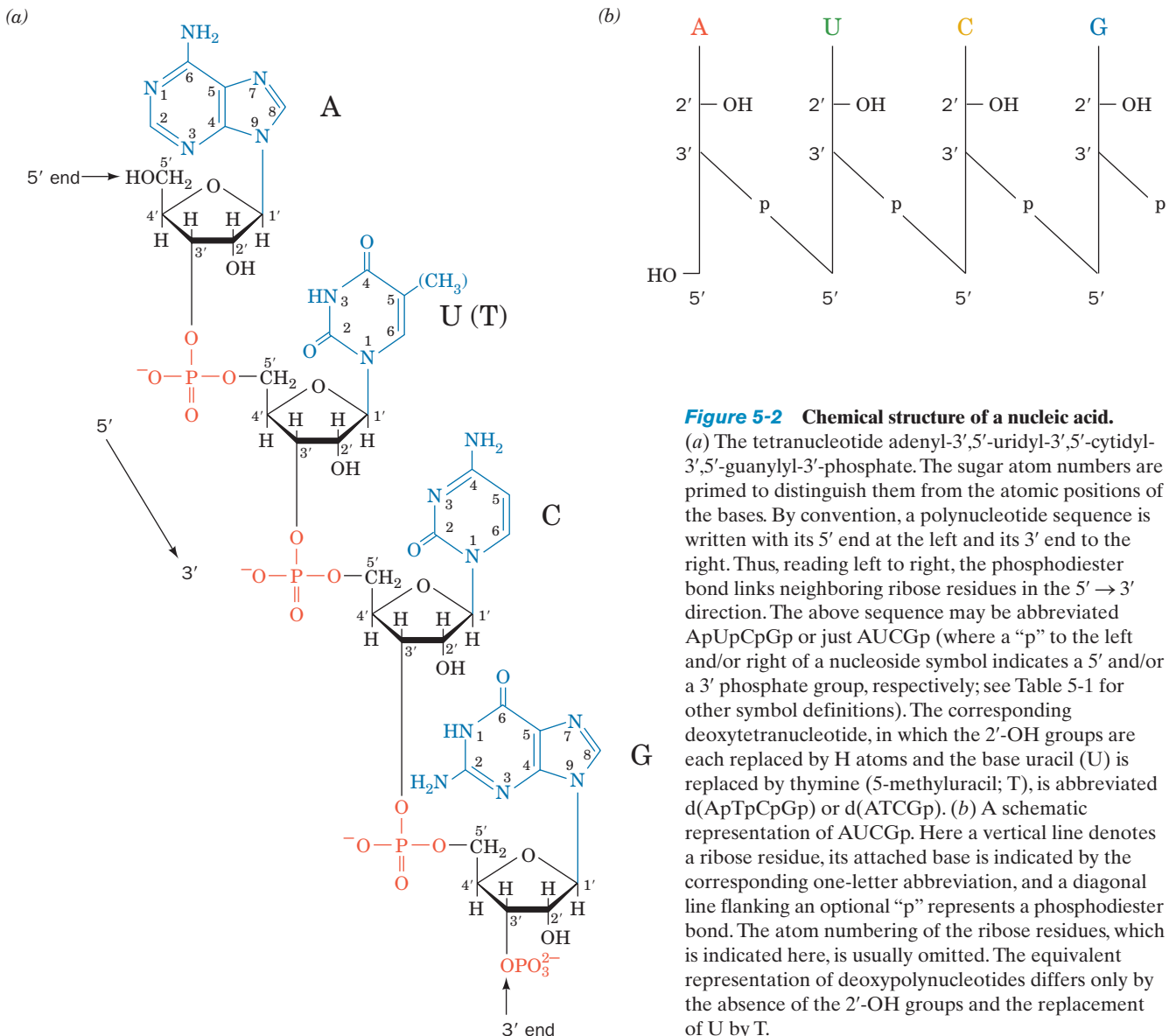
### B. The Chemical Structures of DNA and RNA

The chemical structures of the nucleic acids were elucidated by the early 1950s largely through the efforts of Phoebus Levene, followed by the work of Alexander Todd. *Nucleic acids are, with few exceptions, linear polymers of nucleotides whose phosphate groups bridge the 3' and 5' positions of successive sugar residues (e.g., Fig. 5-2).* The phosphates of these **polynucleotides**, the **phosphodiester** groups, are acidic, so that, *at physiological pH's, nucleic acids are polyanions.* Polynucleotides have directionality, that is, each has a **3' end** (the end whose C3' atom is not linked to a neighboring nucleotide) and a **5' end** (the end whose C5' atom is not linked to a neighboring nucleotide).

#### a. DNA's Base Composition Is Governed by Chargaff's Rules

*DNA has equal numbers of adenine and thymine residues ( $A = T$ ) and equal numbers of guanine and cytosine residues ( $G = C$ ).* These relationships, known as **Chargaff's rules**, were discovered in the late 1940s by Erwin Chargaff, who first devised reliable quantitative methods for the separation and analysis of DNA hydrolysates. Chargaff also found that the base composition of DNA from a given organism is characteristic of that organism; that is, it is independent of the tissue from which the DNA is taken as well as the organism's age, its nutritional state, or any other environmental factor. The structural basis for Chargaff's rules is that in double-stranded DNA, G is always hydrogen bonded (forms a **base pair**) with C, whereas A always forms a base pair with T (Fig. 1-16).

DNA's base composition varies widely among different organisms. It ranges from ~25% to 75% G + C in different

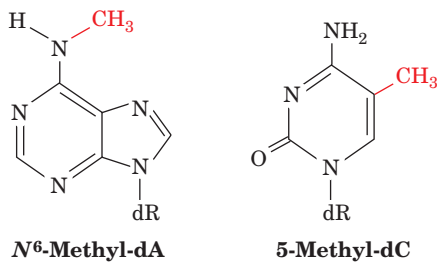


species of bacteria. It is, however, more or less constant among related species; for example, in mammals G + C ranges from 39% to 46%.

RNA, which usually occurs as single-stranded molecules, has no apparent constraints on its base composition. However, double-stranded RNA, which comprises the genetic material of certain viruses, also obeys Chargaff's rules (here A base pairs with U in the same way it does with T in DNA; Fig. 1-16). Conversely, single-stranded DNA, which occurs in certain viruses, does not obey Chargaff's rules. On entering its host organism, however, such DNA is replicated to form a double-stranded molecule, which then obeys Chargaff's rules.

### b. Nucleic Acid Bases May Be Modified

Some DNAs contain bases that are chemical derivatives of the standard set. For example, dA and dC in the DNAs of many organisms are partially replaced by *N*<sup>6</sup>-methyl-dA and 5-methyl-dC, respectively.



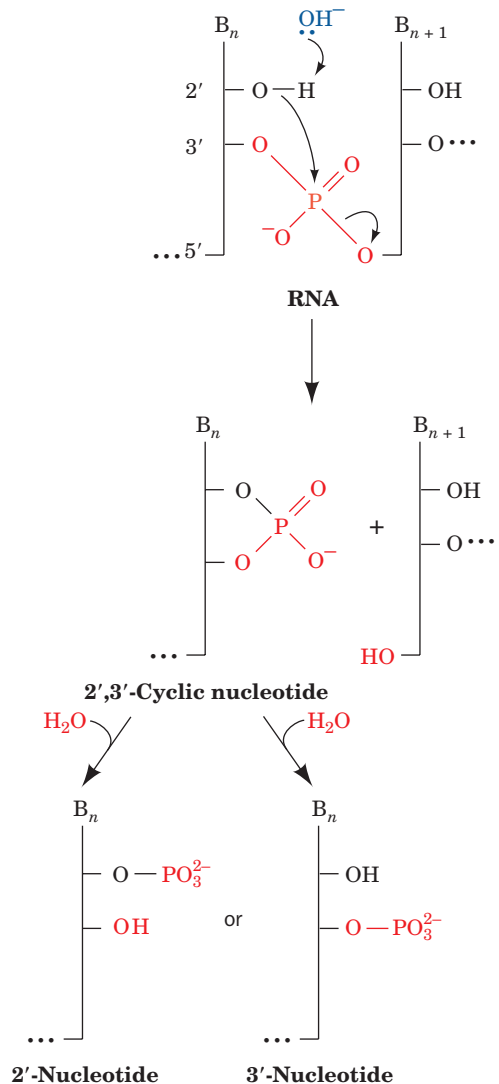
The altered bases are generated by the sequence-specific enzymatic modification of normal DNA (Sections 5-5A and 30-7). The modified DNAs obey Chargaff's rules if the derivatized bases are taken as equivalent to their parent bases. Likewise, many bases in RNAs and, in particular, those in transfer RNAs (tRNAs; Section 32-2Aa) are derivatized.

### c. RNA but Not DNA Is Susceptible to Base-Catalyzed Hydrolysis

RNA is highly susceptible to base-catalyzed hydrolysis by the reaction mechanism diagrammed in Fig. 5-3 so as to yield a mixture of 2' and 3' nucleotides. In contrast, DNA, which lacks 2'-OH groups, is resistant to base-catalyzed hydrolysis and is therefore much more chemically stable than RNA. This is probably why DNA rather than RNA evolved to be the cellular genetic archive.

## 2 DNA IS THE CARRIER OF GENETIC INFORMATION

Nucleic acids were first isolated in 1869 by Friedrich Miescher and so named because he found them in the nuclei of **leukocytes** (pus cells) from discarded surgical bandages. The presence of nucleic acids in other cells was demonstrated within a few years, but it was not until some 75 years after their discovery that their biological function was elucidated. Indeed, in the 1930s and 1940s it was widely held, in what was termed the **tetranucleotide hypothesis**, that nucleic acids have a monotonously repeating sequence of all four bases, so that they were not suspected of having a

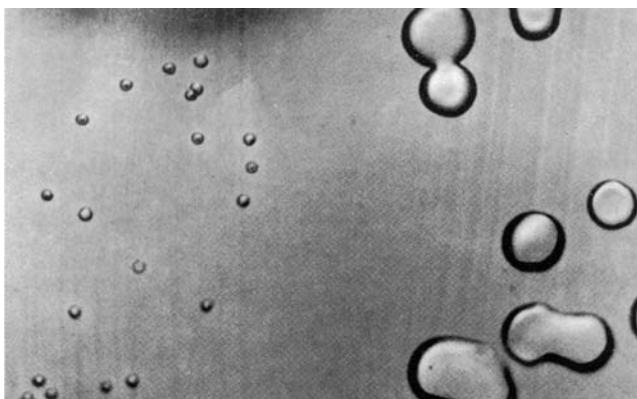


**Figure 5-3** Mechanism of base-catalyzed RNA hydrolysis. The base-induced deprotonation of the 2'-OH group facilitates its nucleophilic attack on the adjacent phosphorus atom, thereby cleaving the RNA backbone. The resultant 2',3'-cyclic phosphate group subsequently hydrolyzes to either the 2' or the 3' phosphate.

genetic function. Rather, it was generally assumed that genes were proteins since proteins were the only biochemical entities that, at that time, seemed capable of the required specificity. In this section, we outline the experiments that established DNA's genetic role.

### A. Transforming Principle Is DNA

The virulent (capable of causing disease) form of pneumococcus (*Diplococcus pneumoniae*), a bacterium that causes pneumonia, is encapsulated by a gelatinous polysaccharide coating that contains the binding sites (known as **O-antigens**; Section 11-3Bc) through which it recognizes the cells it infects. Mutant pneumococci that lack this coating, because of a defect in an enzyme involved in its formation, are not pathogenic (capable of causing disease). The virulent and



**Figure 5-4 Pneumococci.** The large glistening colonies are virulent S-type pneumococci that resulted from the transformation of nonpathogenic R-type pneumococci (smaller colonies) by DNA from heat-killed S pneumococci. [From Avery, O.T., MacLeod, C.M., and McCarty, M., *J. Exp. Med.* **79**, 153 (1944). Copyright © 1944 by Rockefeller University Press.]

nonpathogenic pneumococci are known as the S and R forms, respectively, because of the smooth and rough appearances of their colonies in culture (Fig. 5-4).

In 1928, Frederick Griffith made a startling discovery. He injected mice with a mixture of live R and heat-killed S pneumococci. This experiment resulted in the death of most of the mice. More surprising yet was that the blood of the dead mice contained live S pneumococci. The dead S pneumococci initially injected into the mice had somehow **transformed** the otherwise innocuous R pneumococci to the virulent S form. Furthermore, the progeny of the transformed pneumococci were also S; the transformation was permanent. Eventually, it was shown that the transformation could also be made *in vitro* (outside a living organism; literally “in glass”) by mixing R cells with a cell-free extract of S cells. The question remained: What is the nature of the **transforming principle**?

In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty, after a 10-year investigation, reported that *transforming principle is DNA*. The conclusion was based on the observations that the laboriously purified (few modern fractionation techniques were then available) transforming principle had all the physical and chemical properties of DNA, contained no detectable protein, was unaffected by enzymes that catalyze the hydrolysis of proteins and RNA, and was totally inactivated by treatment with an enzyme that catalyzes the hydrolysis of DNA. *DNA must therefore be the carrier of genetic information.*

Avery’s discovery was another idea whose time had not yet come. This seminal advance was initially greeted with skepticism and then largely ignored. Indeed, even Avery did not directly state that DNA is the hereditary material but merely that it has “biological specificity.” His work, however, influenced several biochemists, including Erwin Chargaff, whose subsequent accurate determination of DNA base ratios refuted the tetranucleotide hypothesis and thereby indicated that DNA could be a complex molecule.

It was eventually demonstrated that eukaryotes are also subject to transformation by DNA. Thus DNA, which cyto-

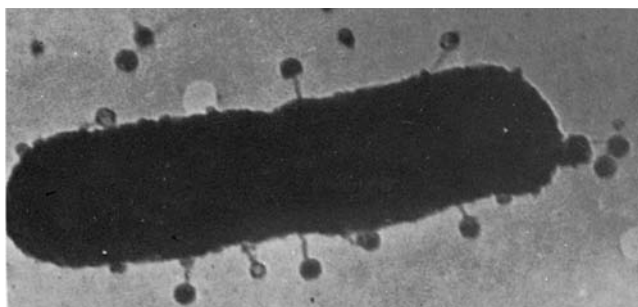


**Figure 5-5 Transgenic mouse.** The gigantic mouse (*left*) grew from a fertilized ovum that had been microinjected with DNA bearing the rat growth hormone gene. His normal-sized littermate (*right*) is shown for comparison. [Courtesy of Ralph Brinster, University of Pennsylvania.]

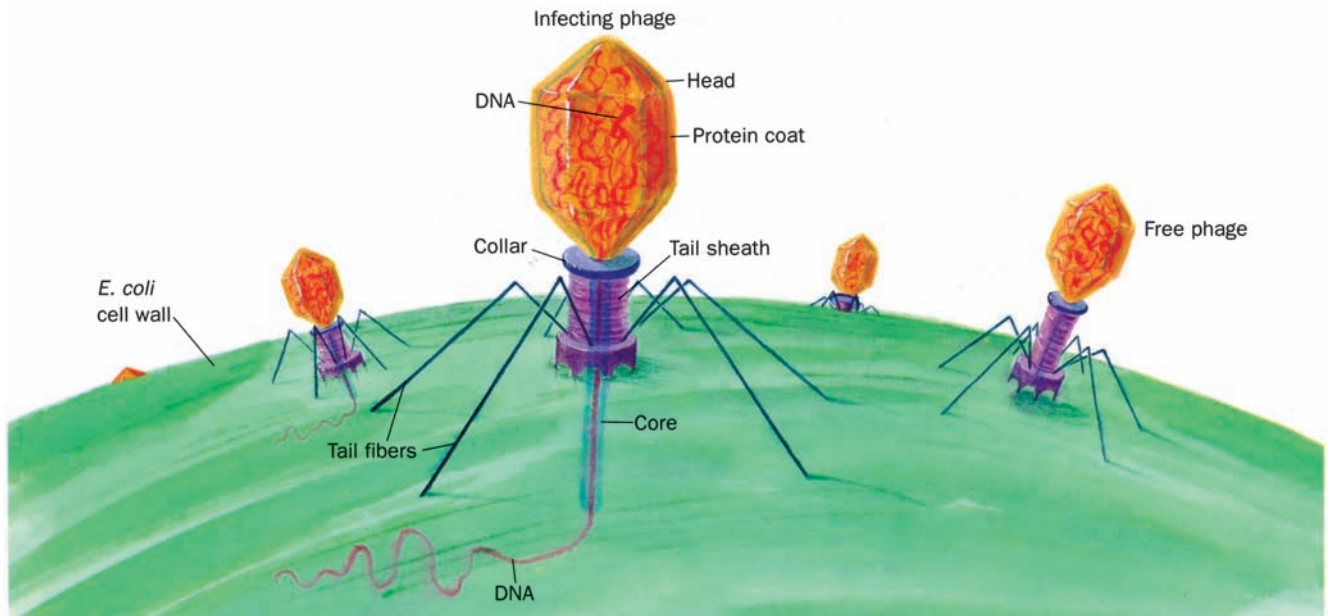
logical studies had shown resides in the chromosomes, must also be the hereditary material of eukaryotes. In a spectacular demonstration of eukaryotic transformation, Ralph Brinster, in 1982, microinjected DNA bearing the gene for rat **growth hormone** (a polypeptide) into the nuclei of fertilized mouse eggs (a technique discussed in Section 5-5H) and implanted these eggs into the uteri of foster mothers. The resulting “supermice” (Fig. 5-5), which had high levels of rat growth hormone in their serum, grew to nearly twice the weight of their normal littermates. Such genetically altered animals are said to be **transgenic**.

### B. The Hereditary Molecule of Many Bacteriophages Is DNA

Electron micrographs of bacteria infected with bacteriophages show empty-headed phage “ghosts” attached to the bacterial surface (Fig. 5-6). This observation led Roger Herriott to suggest that “the virus may act like a little hypodermic needle full of transforming principle,” which it in-



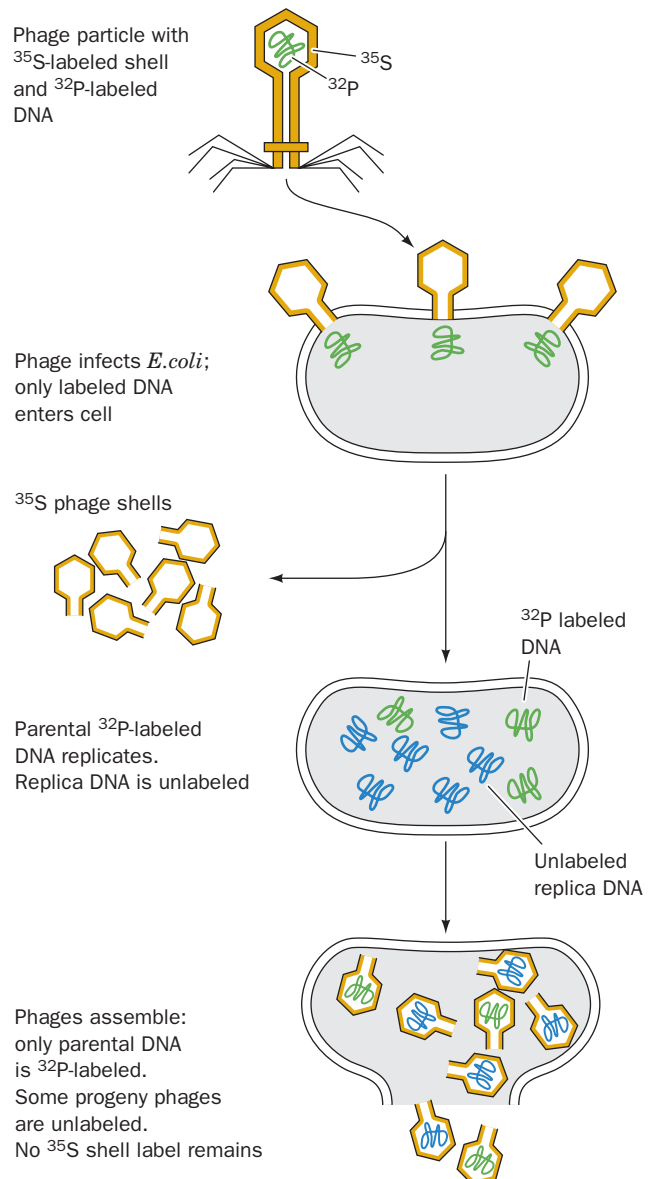
**Figure 5-6 Bacteriophages attached to the surface of a bacterium.** This early electron micrograph shows an *E. coli* cell to which **bacteriophage T5** are adsorbed by their tails. [Courtesy of Thomas F. Anderson, Fox Chase Cancer Center.]



**Figure 5-7** Diagram of T2 bacteriophage injecting its DNA into an *E. coli* cell.

jects into the bacterial host (Fig. 5-7). This proposal was tested in 1952 by Alfred Hershey and Martha Chase as is diagrammed in Fig. 5-8. **Bacteriophage T2** was grown on *E. coli* in a medium containing the radioactive isotopes  $^{32}\text{P}$  and  $^{35}\text{S}$ . This labeled the phage capsid, which contains no P, with  $^{35}\text{S}$ , and its DNA, which contains no S, with  $^{32}\text{P}$ . These phages were added to an unlabeled culture of *E. coli* and, after sufficient time was allowed for the phages to infect the bacterial cells, the culture was agitated in a kitchen blender so as to shear the phage ghosts from the bacterial cells. This rough treatment neither injured the bacteria nor altered the course of the phage infection. When the phage ghosts were separated from the bacteria (by centrifugation; Section 6-5), the ghosts were found to contain most of the  $^{35}\text{S}$ , whereas the bacteria contained most of the  $^{32}\text{P}$ . Furthermore, 30% of the  $^{32}\text{P}$  appeared in the progeny phages but only 1% of the  $^{35}\text{S}$  did so. Hershey and Chase therefore concluded that only the phage DNA was essential for the production of progeny. *DNA therefore must be the hereditary material.* In later years it was shown that, in a process known as **transfection**, purified phage DNA can, by itself, induce a normal phage infection of a properly treated bacterial host (transfection differs from transformation in that the latter results from the recombination of the bacterial chromosome with a fragment of homologous DNA).

In 1952, the state of knowledge of biochemistry was such that Hershey's discovery was much more readily accepted than Avery's identification of the transforming principle had been some 8 years earlier. Within a few months, the first speculations arose as to the nature of the **genetic**



**Figure 5-8** The Hershey–Chase experiment. This experiment demonstrated that only the nucleic acid component of bacteriophages enters the bacterial host during phage infection.

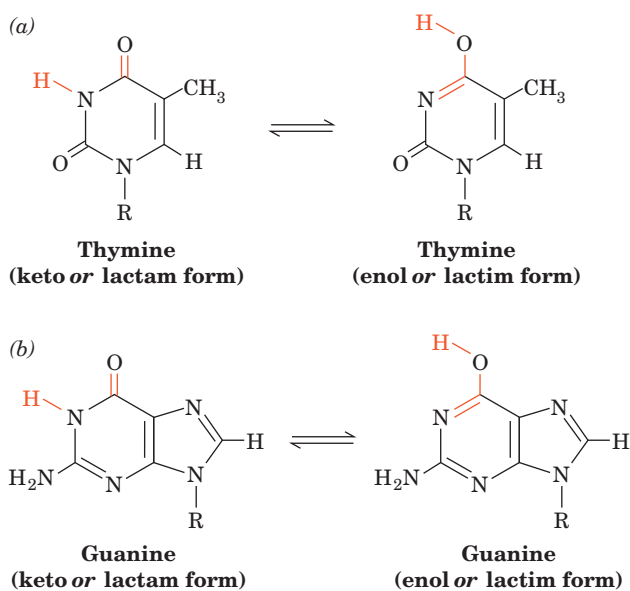
**code** (the correspondence between the base sequence of a gene and the amino acid sequence of a protein, Section 5-4Bb), and James Watson and Francis Crick were inspired to investigate the structure of DNA. In 1955, it was shown that the somatic cells of eukaryotes have twice the DNA of the corresponding germ cells. When this observation was proposed to be a further indicator of DNA's genetic role, there was little comment even though the same could be said of any other chromosomal component.

### 3 DOUBLE HELICAL DNA

The determination of the structure of DNA by Watson and Crick in 1953 is often said to mark the birth of modern molecular biology. The **Watson–Crick structure** of DNA is of such importance because, in addition to providing the structure of what is arguably the central molecule of life, it suggested the molecular mechanism of heredity. Watson and Crick's accomplishment, which is ranked as one of science's major intellectual achievements, tied together the less than universally accepted results of several diverse studies:

1. **Chargaff's rules.** At the time, the relationships  $A = T$  and  $G = C$  were quite obscure because their significance was not apparent. In fact, even Chargaff did not emphasize them.

2. **Correct tautomeric forms of the bases.** X-ray, nuclear magnetic resonance (NMR), and spectroscopic investigations have firmly established that the nucleic acid bases are overwhelmingly in the keto tautomeric forms shown in Table 5-1. In 1953, however, this was not generally appreciated. Indeed, guanine and thymine were widely believed to be in their enol forms (Fig. 5-9) because it was thought that the resonance stability of these aromatic molecules would thereby be maximized. Knowledge of the dominant tautomeric forms, which



**Figure 5-9** Some possible tautomeric conversions for bases. (a) Thymine and (b) guanine residues. Cytosine and adenine residues can undergo similar proton shifts.

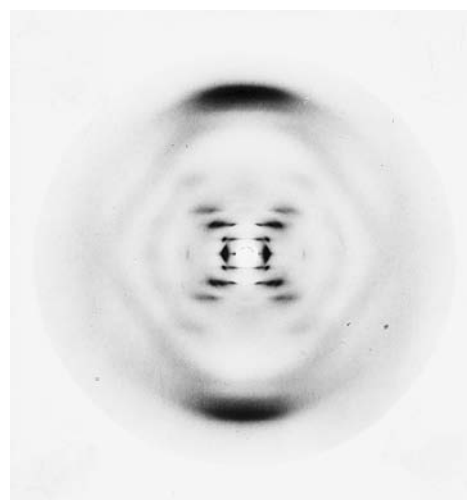
was prerequisite for the prediction of the correct hydrogen bonding associations of the bases, was provided by Jerry Donohue, an office mate of Watson and Crick and an expert on the X-ray structures of small organic molecules.

3. **Information that DNA is a helical molecule.** This was provided by an X-ray diffraction photograph of a DNA fiber taken by Rosalind Franklin (Fig. 5-10; DNA, being a thread-like molecule, does not crystallize but, rather, can be drawn out in fibers consisting of parallel bundles of molecules). This photograph enabled Crick, an X-ray crystallographer by training who had earlier derived the equations describing diffraction by helical molecules, to deduce (a) that DNA is a helical molecule and (b) that its planar aromatic bases form a stack of parallel rings which is parallel to the fiber axis.

This information only provided a few crude landmarks that guided the elucidation of the DNA structure. It mostly sprang from Watson and Crick's imaginations through model building studies. Once the Watson–Crick model had been published, however, its basic simplicity combined with its obvious biological relevance led to its rapid acceptance. Later investigations have confirmed the essential correctness of the Watson–Crick model, although its details have been modified.

#### A. The Watson–Crick Structure: B-DNA

Fibers of DNA assume the so-called B conformation, as indicated by their X-ray diffraction patterns, when the counterion is an alkali metal such as  $\text{Na}^+$  and the relative humidity is  $>92\%$ . **B-DNA** is regarded as the *native* (biologically functional) form of DNA because, for example, its X-ray pattern resembles that of the DNA in intact sperm heads.



**Figure 5-10** X-ray diffraction photograph of a vertically oriented  $\text{Na}^+$  DNA fiber in the B conformation taken by Rosalind Franklin. This is the photograph that provided key information for the elucidation of the Watson–Crick structure. The central X-shaped pattern of spots is indicative of a helix, whereas the heavy black arcs on the top and bottom of the diffraction pattern correspond to a distance of  $3.4 \text{ \AA}$  and indicate that the DNA structure largely repeats every  $3.4 \text{ \AA}$  along the fiber axis. [Courtesy of Maurice Wilkins, King's College, London.]

The Watson–Crick structure of B-DNA has the following major features:

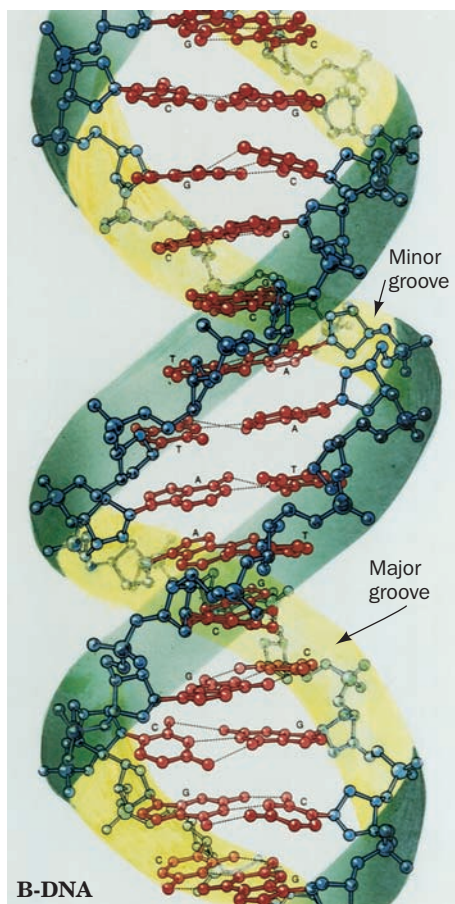
1. It consists of two polynucleotide strands that wind about a common axis with a right-handed twist to form an  $\sim 20\text{-\AA}$ -diameter double helix (Fig. 5-11). The two strands are antiparallel (run in opposite directions) and wrap around each other such that they cannot be separated without unwinding the helix. The bases occupy the core of the helix and the sugar–phosphate chains are coiled about its periphery, thereby minimizing the repulsions between charged phosphate groups.

2. The planes of the bases are nearly perpendicular to the helix axis. Each base is hydrogen bonded to a base on the opposite strand to form a planar base pair (Fig. 5-11). It is these hydrogen bonding interactions, a phenomenon

known as **complementary base pairing**, that result in the specific association of the two chains of the double helix.

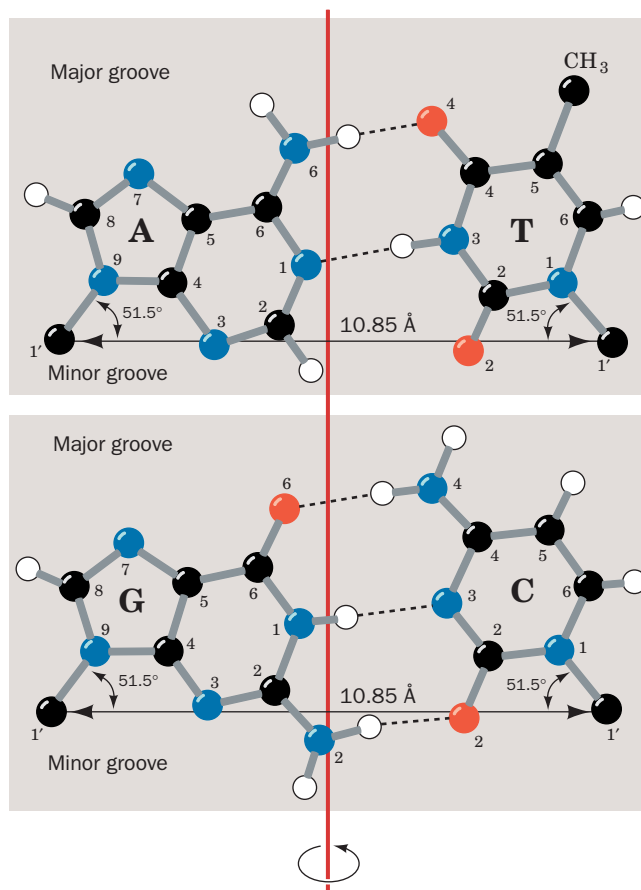
3. The “ideal” B-DNA helix has 10 base pairs (**bp**) per turn (a helical twist of  $36^\circ$  per bp) and, since the aromatic bases have van der Waals thicknesses of  $3.4\text{ \AA}$  and are partially stacked on each other (**base stacking**, Fig. 5-11), the helix has a **pitch** (rise per turn) of  $34\text{ \AA}$ .

The most remarkable feature of the Watson–Crick structure is that it can accommodate only two types of base pairs: Each adenine residue must pair with a thymine residue and vice versa, and each guanine residue must pair with a cytosine residue and vice versa. The geometries of these A · T and G · C base pairs, the so-called **Watson–Crick base pairs**, are shown in Fig. 5-12. It can be seen that both of these base pairs are interchangeable in that they can replace each other in the double helix without altering the positions of the sugar–phosphate backbone’s



**Figure 5-11** Three-dimensional structure of B-DNA. The repeating helix in this ball-and-stick drawing is based on the X-ray structure of the self-complementary dodecamer d(CGCGAATTCGCG) determined by Richard Dickerson and Horace Drew. The view is perpendicular to the helix axis. The sugar–phosphate backbones (blue with blue-green ribbon outlines) wind about the periphery of the molecule in opposite directions. The bases (red), which occupy its core, form hydrogen bonded base pairs. H atoms have been omitted for clarity. [Illustration, Irving Geis. Image from the Irving Geis Collection, Howard Hughes Medical Institute. Reprinted with permission.]

See Interactive Exercise 1 and Kinemage 2-1



**Figure 5-12** Watson–Crick base pairs. The line joining the C1' atoms is the same length in both base pairs and makes equal angles with the glycosidic bonds to the bases. This gives DNA a series of pseudo-twofold symmetry axes (often referred to as **dyad axes**) that pass through the center of each base pair (red line) and are perpendicular to the helix axis. Note that A · T base pairs associate via two hydrogen bonds, whereas C · G base pairs are joined by three hydrogen bonds. [After Arnott, S., Dover, S.D., and Wonacott, A.J., *Acta Cryst.* **B25**, 2192 (1969).] See Kinemages 2-2 and 17-2



*C1'* atoms. Likewise, the double helix is undisturbed by exchanging the partners of a Watson–Crick base pair, that is, by changing a  $G \cdot C$  to a  $C \cdot G$  or an  $A \cdot T$  to a  $T \cdot A$ . In contrast, any other combination of bases (e.g.,  $A \cdot G$  or  $A \cdot C$ ) would significantly distort the double helix since the formation of a non-Watson–Crick base pair would require considerable reorientation of the sugar–phosphate chain.

B-DNA has two deep exterior grooves that wind between its sugar–phosphate chains as a consequence of the helix axis passing through the approximate center of each base pair. However, the grooves are of unequal size (Fig. 5-11) because (1) the top edge of each base pair, as drawn in Fig. 5-12, is structurally distinct from the bottom edge; and (2) the deoxyribose residues are asymmetric. The **minor groove** exposes that edge of a base pair from which its  $C1'$  atoms extend (opening toward the bottom in Fig. 5-12), whereas the **major groove** exposes the opposite edge of each base pair (the top of Fig. 5-12).

Although B-DNA is, by far, the most prevalent form of DNA in the cell, double helical DNAs and RNAs can assume several distinct structures. The structures of these other double helical nucleic acids are discussed in Section 29-1B.

### B. DNA Is Semiconservatively Replicated

The Watson–Crick structure can accommodate any sequence of bases on one polynucleotide strand if the opposite strand has the complementary base sequence. This immediately accounts for Chargaff's rules. More importantly, it suggests that hereditary information is encoded in the sequence of bases on either strand. Furthermore, each polynucleotide strand can act as a template for the formation of its complementary strand through base pairing interactions (Fig. 1-17). The two strands of the parent molecule must therefore separate so that a complementary daughter strand may be enzymatically synthesized on the surface of each parent strand. This results in two molecules of **duplex** (double-stranded) DNA, each consisting of one polynucleotide strand from the parent molecule and a newly synthesized complementary strand. Such a mode of replication is termed **semiconservative** in contrast with **conservative** replication, which, if it occurred, would result in a newly synthesized duplex copy of the original DNA molecule with the parent DNA molecule remaining intact. The mechanism of DNA replication is the main subject of Chapter 30.

The semiconservative nature of DNA replication was elegantly demonstrated in 1958 by Matthew Meselson and Franklin Stahl. The density of DNA was increased by labeling it with  $^{15}\text{N}$ , a heavy isotope of nitrogen ( $^{14}\text{N}$  is the naturally abundant isotope). This was accomplished by growing *E. coli* for 14 generations in a medium that contained  $^{15}\text{NH}_4\text{Cl}$  as the only nitrogen source. The labeled bacteria were then abruptly transferred to an  $^{14}\text{N}$ -containing medium, and the density of their DNA was monitored as a function of bacterial growth by **equilibrium density gradi-**

**ent ultracentrifugation** (a technique for separating macromolecules according to their densities, which Meselson, Stahl, and Jerome Vinograd had developed for the purpose of distinguishing  $^{15}\text{N}$ -labeled DNA from unlabeled DNA; Section 6-5Bb).

The results of the Meselson–Stahl experiment are displayed in Fig. 5-13. After one generation (doubling of the cell population), all of the DNA had a density exactly halfway between the densities of fully  $^{15}\text{N}$ -labeled DNA and unlabeled DNA. This DNA must therefore contain equal amounts of  $^{14}\text{N}$  and  $^{15}\text{N}$ , as is expected after one generation of semiconservative replication. Conservative DNA replication, in contrast, would result in the preservation of the parental DNA, so that it maintained its original density, and the generation of an equal amount of unlabeled DNA. After two generations, half of the DNA molecules were unlabeled and the remainder were  $^{14}\text{N}$ – $^{15}\text{N}$  hybrids. This is also in accord with the predictions of the semiconservative replication model and in disagreement with the conservative replication model. In succeeding generations, the amount of unlabeled DNA increased relative to the amount of hybrid DNA, although the hybrid never totally disappeared. This is again in harmony with semiconservative replication but at odds with conservative replication, which predicts that the fully labeled parental DNA will always be present and that hybrid DNA never forms.

### C. Denaturation and Renaturation

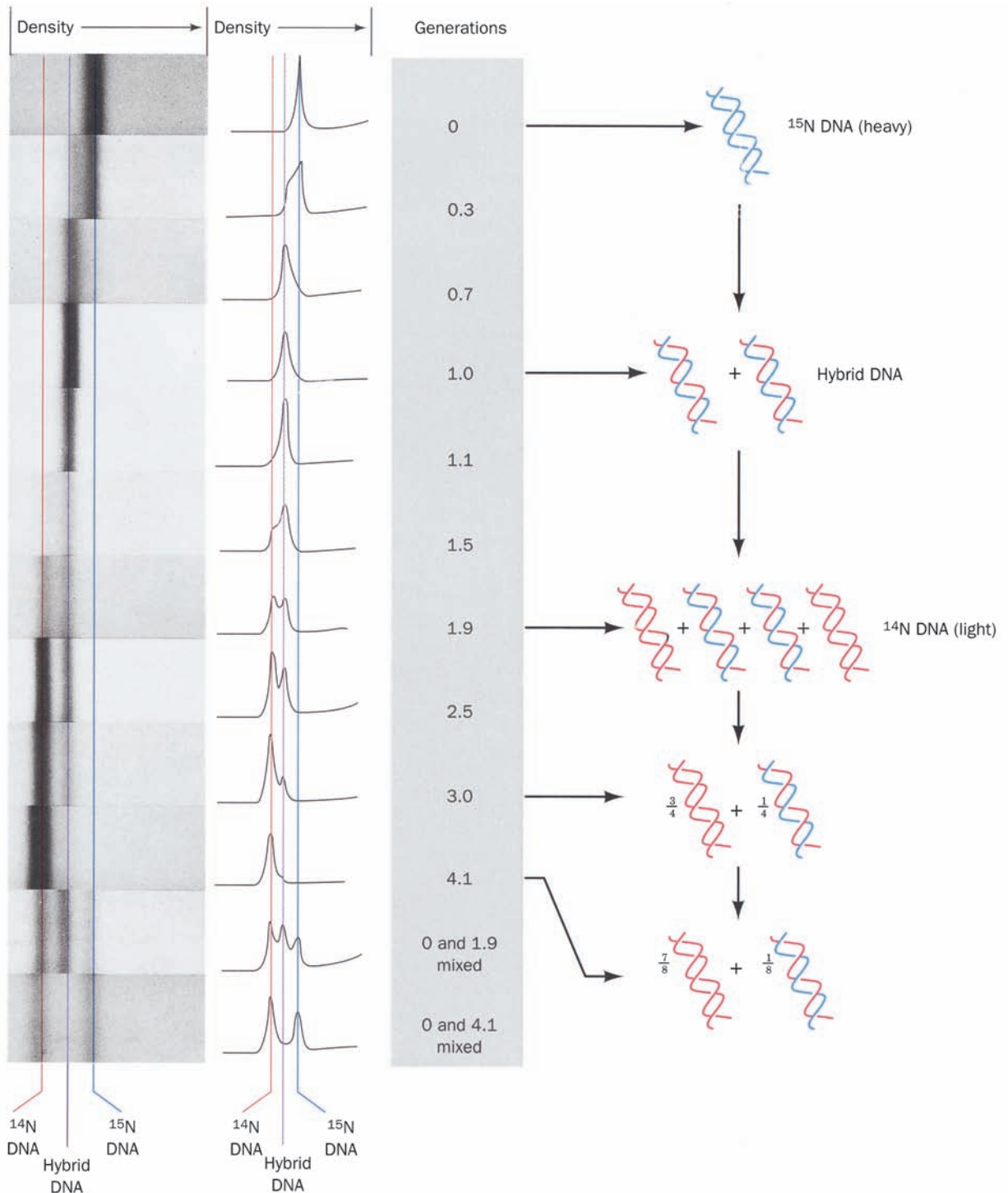
When a solution of duplex DNA is heated above a characteristic temperature, its native structure collapses and its two complementary strands separate and assume a flexible and rapidly fluctuating conformational state known as a **random coil** (Fig. 5-14). This **denaturation** process is accompanied by a qualitative change in the DNA's physical properties. For instance, the characteristic high viscosity of native DNA solutions, which arises from the resistance to deformation of its rigid and rodlike duplex molecules, drastically decreases when the duplex DNA decomposes (denatures) to two relatively freely jointed single strands.

#### a. DNA Denaturation Is a Cooperative Process


The most convenient way of monitoring the amount of nucleic acid present is by its ultraviolet (UV) absorbance spectrum. A solution containing a solute that absorbs light does so according to the **Beer–Lambert law**,

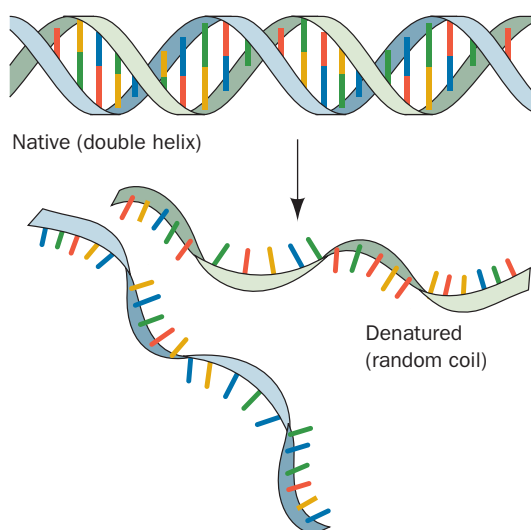
$$A = -\log\left(\frac{I}{I_0}\right) = \epsilon cl \quad [5.1]$$

where  $A$  is the solute's **absorbance** (alternatively, its **optical density**),  $I_0$  is the incident intensity of light at a given wavelength  $\lambda$ ,  $I$  is its transmitted intensity at  $\lambda$ ,  $\epsilon$  is the **molar extinction coefficient** of the solute at  $\lambda$ ,  $c$  is its molar concentration, and  $l$  is the length of the light path in centimeters. The value of  $\epsilon$  varies with  $\lambda$ ; a plot of  $\epsilon$  versus  $\lambda$  for the



**Figure 5-13** Demonstration of the semiconservative nature of DNA replication in *E. coli* by density gradient ultracentrifugation. The DNA was dissolved in an aqueous CsCl solution of density  $1.71 \text{ g} \cdot \text{cm}^{-3}$  and was subjected to an acceleration of 140,000 times that of gravity in an analytical ultracentrifuge (a device in which the rapidly spinning sample can be optically observed). This enormous acceleration induced the CsCl to redistribute in the solution such that its concentration increased with its radius in the ultracentrifuge. Consequently, the DNA migrated within the resulting density gradient to its position of buoyant density. The left panels are ultraviolet absorption photographs of ultracentrifuge cells (DNA strongly absorbs ultraviolet light) and are arranged such that regions of equal density have the same horizontal positions. The middle panels

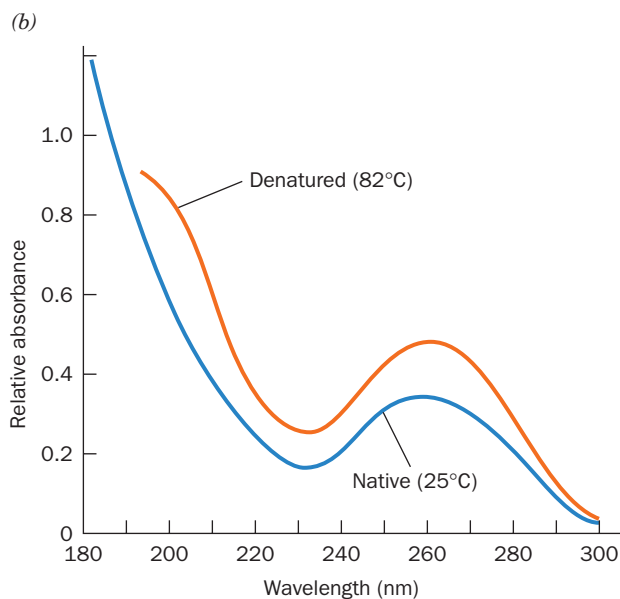
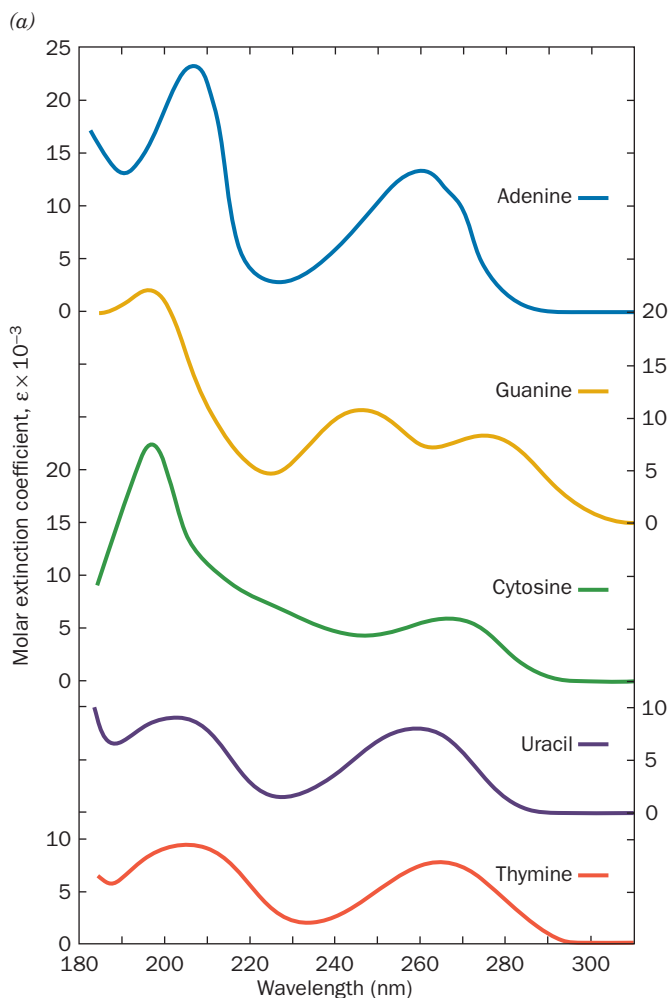
are microdensitometer traces of the corresponding photographs in which the vertical displacement is proportional to the DNA concentration. The buoyant density of DNA increases with its  $^{15}\text{N}$  content. The bands farthest to the right (of greatest radius and density) arise from DNA that is fully  $^{15}\text{N}$  labeled, whereas unlabeled DNA, which is  $0.014 \text{ g} \cdot \text{cm}^{-3}$  less dense, forms the leftmost bands. The bands in the intermediate position result from duplex DNA in which one strand is  $^{15}\text{N}$ -labeled and the other strand is unlabeled. The accompanying interpretive drawings (right) indicate the relative numbers of DNA strands at each generation donated by the original parents (blue,  $^{15}\text{N}$  labeled) and synthesized by succeeding generations (red, unlabeled). [From Meselson, M. and Stahl, F.W., *Proc. Natl. Acad. Sci.* **44**, 671 (1958).]  See the Animated Figures




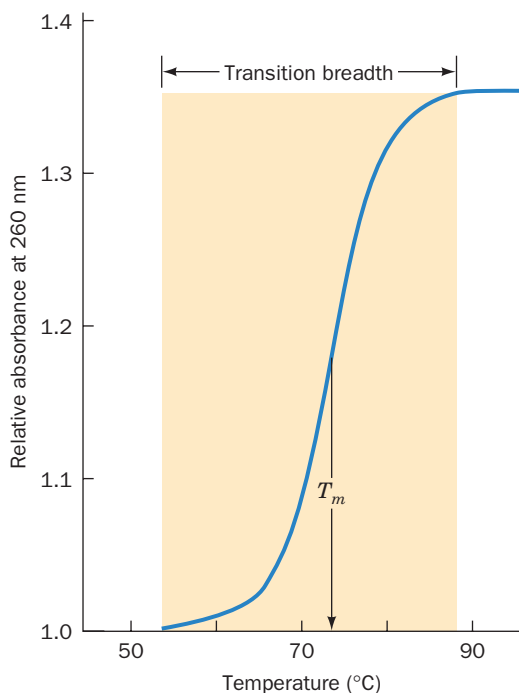
**Figure 5-14** Schematic representation of the strand separation in duplex DNA resulting from its heat denaturation.

solute is called its **absorbance spectrum**. The absorbance spectra of the five nucleic acid bases are shown in Fig. 5-15a. The spectra of the corresponding nucleosides and nucleotides are closely similar above 190 nm because, in this wavelength range, the molar extinction coefficients of ribose and phosphate groups are vanishingly small relative to those of the aromatic bases. As expected, the spectrum of native DNA (Fig. 5-15b) resembles that of its component bases in shape.

When DNA denatures, its UV absorbance increases by ~40% at all wavelengths (Fig. 5-15b). This phenomenon, which is known as the **hyperchromic effect** (Greek: *hyper*, above + *chroma*, color), results from the disruption of the electronic interactions among nearby bases. DNA's hyperchromic shift, as monitored at a particular wavelength (usually 260 nm), occurs over a narrow temperature range (Fig. 5-16). This indicates that the collapse of one part of the duplex DNA's structure destabilizes the remainder, a phenomenon known as a **cooperative process**. The denaturation of DNA may be described as the melting of a one-



**Figure 5-15** UV absorbance spectra of the nucleic acid bases and DNA. (a) Spectra of adenine, guanine, cytosine, thymine, and uracil near pH 7. (b) Spectra of native and heat-denatured *E. coli* DNA. Note that denaturation does not change the general shape of the absorbance spectrum but increases its absorbance at all wavelengths. [After Voet, D., Gratzer, W.B., Cox, R.A., and Doty, P., *Biopolymers* **1**, 193 (1963).]  See the Animated Figures



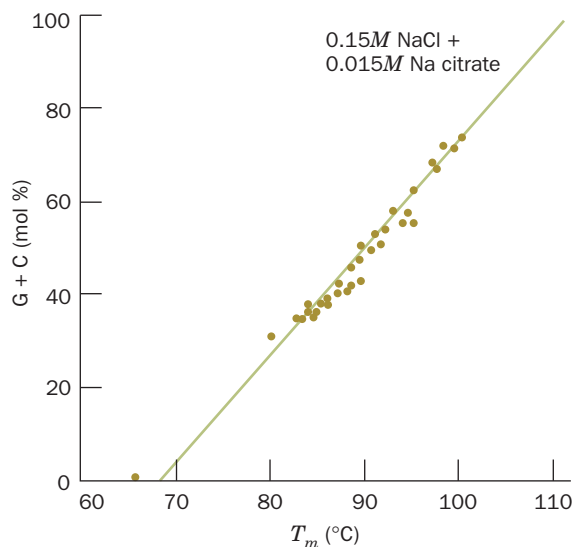
**Figure 5-16** Example of a DNA melting curve. The relative absorbance is the ratio of the absorbance (customarily measured at 260 nm) at the indicated temperature to that at 25°C. The melting temperature,  $T_m$ , is defined as the temperature at which half of the maximum absorbance increase is attained. See the Animated Figures

dimensional solid, so Fig. 5-16 is referred to as a **melting curve** and the temperature at its midpoint is known as the **melting temperature,  $T_m$** .

The stability of the DNA double helix, and hence its  $T_m$ , depends on several factors, including the nature of the solvent, the identities and concentrations of the ions in solution, and the pH. For example, duplex DNA denatures (its  $T_m$  decreases) under alkaline conditions that cause some of the bases to ionize and thereby disrupt their base pairing interactions. The  $T_m$  increases linearly with the mole fraction of G · C base pairs (Fig. 5-17), which indicates that triply hydrogen bonded G · C base pairs are more stable than doubly hydrogen bonded A · T base pairs.

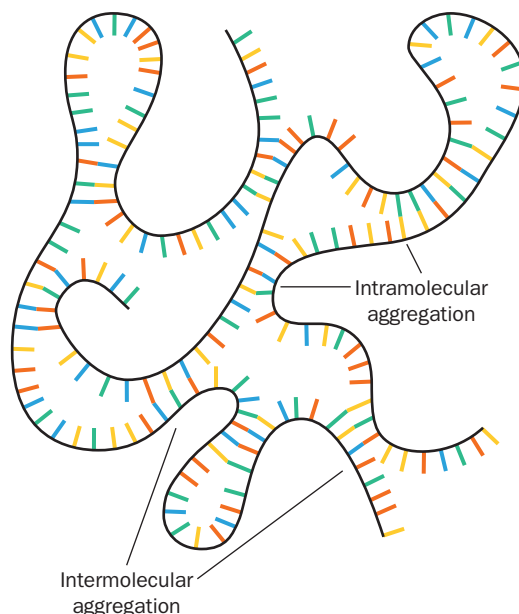
#### b. Denatured DNA Can Be Renatured

If a solution of denatured DNA is rapidly cooled to well below its  $T_m$ , the resulting DNA will be only partially base paired (Fig. 5-18) because its complementary strands will not have had sufficient time to find each other before the partially base paired structures become effectively “frozen in.” If, however, the temperature is maintained  $\sim 25^\circ\text{C}$  below the  $T_m$ , enough thermal energy is available for short base paired regions to rearrange by melting and reforming but not enough to melt out long complementary stretches. Under such **annealing conditions**, as Julius Marmur discovered in 1960, denatured DNA eventually



**Figure 5-17** Variation of the melting temperatures,  $T_m$ , of DNA with its G + C content. The DNAs were dissolved in a solution containing 0.15M NaCl and 0.015M sodium citrate. [After Marmur, J. and Doty, P., *J. Mol. Biol.* **5**, 113 (1962).]

completely renatures. Likewise, complementary strands of RNA and DNA, in a process known as **hybridization**, form RNA–DNA hybrid double helices that are only slightly less stable than the corresponding DNA double helices.



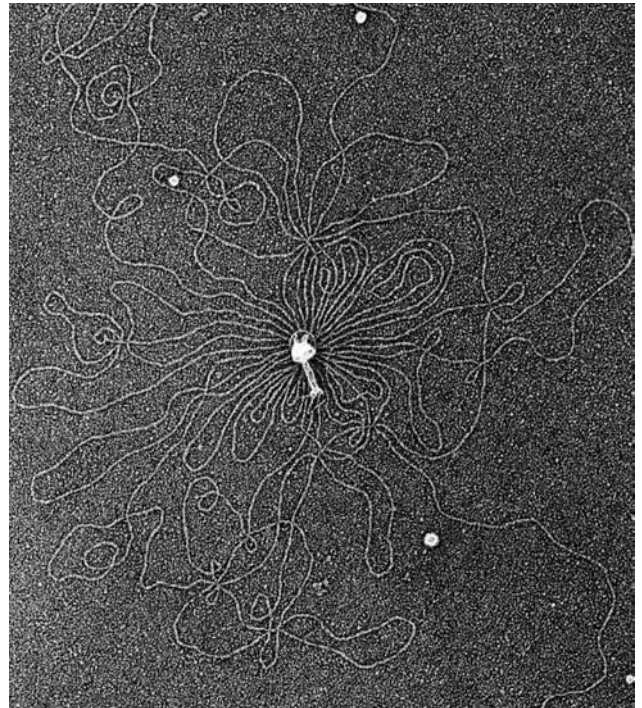
**Figure 5-18** Partially renatured DNA. A schematic representation showing the imperfectly base paired structures assumed by DNA that has been heat denatured and then rapidly cooled to well below its  $T_m$ . Note that both intramolecular and intermolecular aggregation may occur.

### D. The Size of DNA

DNA molecules are generally enormous (Fig. 5-19). The molecular mass of DNA has been determined by a variety of techniques including ultracentrifugation (Section 6-5A) and through length measurements by electron microscopy [a base pair of Na<sup>+</sup> B-DNA has an average molecular mass of 660 D and a length (thickness) of 3.4 Å] and **autoradiography** (a technique in which the position of a radioactive substance in a sample is recorded by the blackening of a photographic emulsion that the sample is laid over or embedded in; Fig. 5-20). The number of base pairs and the **contour lengths** (the end-to-end lengths of the stretched-out native molecules) of the DNAs from a selection of organisms of increasing complexity are listed in Table 5-2. Not surprisingly, an organism's haploid quantity (unique amount) of DNA varies more or less with its complexity (although there are notable exceptions to this generalization, such as the last entry in Table 5-2).

The visualization of DNAs from prokaryotes has demonstrated that their entire **genome** (complement of genetic information) is contained on a single, often circular, length of DNA. Similarly, Bruno Zimm demonstrated that the *largest chromosome of the fruit fly *Drosophila melanogaster* contains a single molecule of DNA* by comparing the molecular mass of this DNA with the cytologically measured length of DNA contained in the chromosome. Likewise, other eukaryotic chromosomes contain only single molecules of DNA.

The highly elongated shape of duplex DNA (recall B-DNA is only 20 Å in diameter), together with its stiffness, make it extremely susceptible to mechanical damage outside the cell's protective environment (for instance, if the *Drosophila* DNA of Fig. 5-20 were enlarged by a factor of 500,000, it would have the shape and some of the mechanical properties of a 6-km-long strand of uncooked spaghetti). The hydrodynamic shearing forces generated by such ordinary



**Figure 5-19** Electron micrograph of a T2 bacteriophage and its DNA. The phage has been osmotically lysed (broken open) in distilled water so that its DNA spilled out. Without special treatment, duplex DNA, which is only 20 Å in diameter, is difficult to visualize in the electron microscope. In the **Kleinschmidt procedure** used here, DNA is fattened to ~200 Å in diameter by coating it with a denatured basic protein. [From Kleinschmidt, A.K., Lang, D., Jacherts, D., and Zahn, R.K., *Biochim. Biophys. Acta* **61**, 857 (1962).]

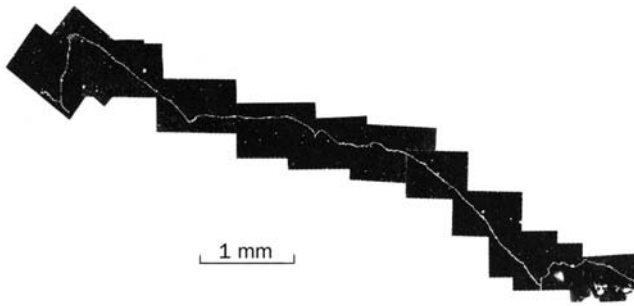
laboratory manipulations as stirring, shaking, and pipetting break DNA into relatively small pieces so that the isolation of an intact molecule of DNA requires extremely gentle

**Table 5-2** Sizes of Some DNA Molecules

Organism	Number of Base Pairs (kb) <sup>a</sup>	Contour Length (μm)
<b>Viruses</b>		
Polyoma, SV40	5.2	1.7
λ Bacteriophage	48.6	17
T2, T4, T6 bacteriophage	166	55
Fowlpox	280	193
<b>Bacteria</b>		
<i>Mycoplasma hominis</i>	760	260
<i>Escherichia coli</i>	4,600	1,600
<b>Eukaryotes</b>		
Yeast (in 17 haploid chromosomes)	12,000	4,100
<i>Drosophila</i> (in 4 haploid chromosomes)	180,000	61,000
Human (in 23 haploid chromosomes)	3,000,000	1,000,000
Lungfish (in 19 haploid chromosomes)	102,000,000	35,000,000

<sup>a</sup>kb = kilobase pair = 1000 base pairs (bp).


Source: Mainly Kornberg, A. and Baker, T.A., *DNA Replication* (2nd ed.), p. 20, Freeman (1992).

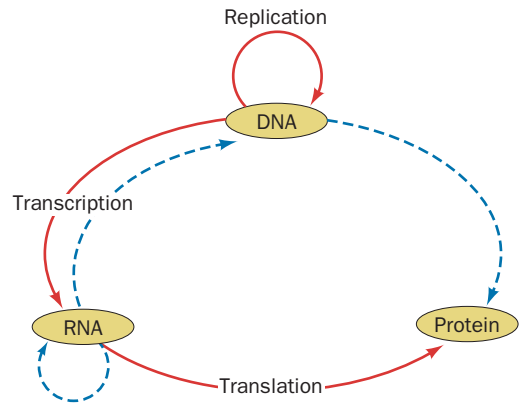
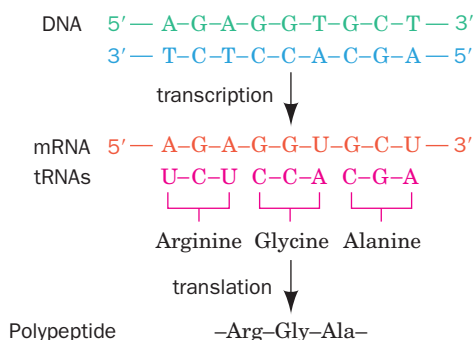


**Figure 5-20** **Autoradiograph of *Drosophila melanogaster* DNA.** Lysates of *D. melanogaster* cells that had been cultured with  $^3\text{H}$ -labeled thymidine were spread on a glass slide and covered with a photographic emulsion that was developed after a 5-month exposure. The white curve, which resulted from the radioactive decay of the  $^3\text{H}$ , traces the path of the DNA in this photographic positive. The DNA's measured contour length is 1.2 cm. [From Kavenoff, R., Klotz, L.C., and Zimm, B.H., *Cold Spring Harbor Symp. Quant. Biol.* **38**, 4 (1973). Copyright © 1974 by Cold Spring Harbor Laboratory Press.]

handling. Before 1960, when this was first realized, the measured molecular masses of DNA were no higher than  $\sim 10$  million D ( $\sim 15$  kb, where 1 kb = 1 kilobase pair = 1000 bp). DNA fragments of uniform molecular mass and as small as a few hundred base pairs may be generated by **shear degrading** DNA in a controlled manner; for instance, by forcing the DNA solution through a small orifice or by **sonication** (exposure to intense high-frequency sound waves).

## 4 GENE EXPRESSION AND REPLICATION: AN OVERVIEW

 **See Guided Exploration 1: Overview of transcription and translation**  
How do genes function, that is, how do they direct the synthesis of RNA and proteins, and how are they replicated? The answers to these questions form the multifaceted discipline known as **molecular biology**. In 1958, Crick neatly encapsulated the broad outlines of this process in a flow scheme he called the **central dogma of molecular biology**: *DNA directs its own replication and its transcription to yield RNA which, in turn, directs its translation to form proteins* (Fig. 5-21). Here the term “transcription” indicates that in transferring



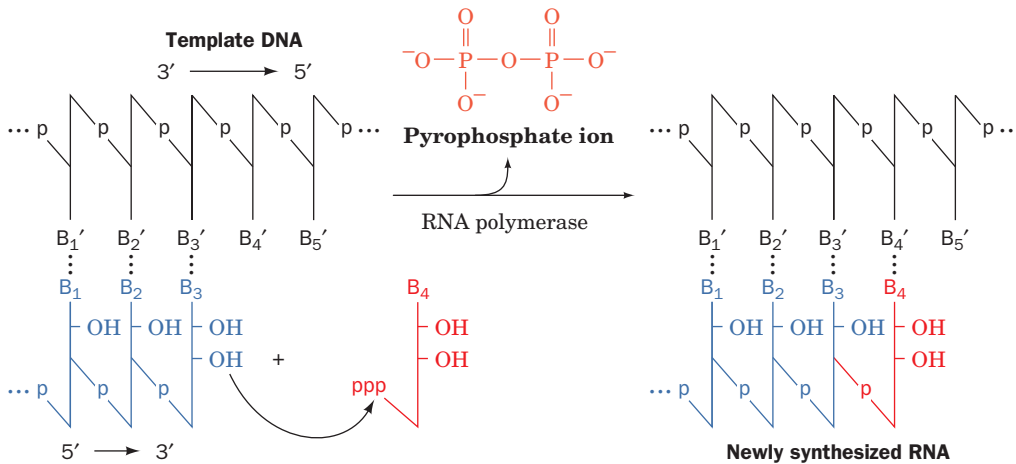
**Figure 5-21** **The central dogma of molecular biology.** Solid arrows indicate the types of genetic information transfers that occur in all cells. Special transfers are indicated by the dashed arrows: **RNA-directed RNA polymerase** is expressed both by certain RNA viruses and by some plants; **RNA-directed DNA polymerase (reverse transcriptase)** is expressed by other RNA viruses; and DNA directly specifying a protein is unknown but does not seem beyond the realm of possibility. However, the missing arrows are information transfers the central dogma postulates never occur: protein specifying either DNA, RNA, or protein. In other words, *proteins can only be the recipients of genetic information*. [After Crick, F., *Nature* **227**, 561 (1970).]

information from DNA to RNA, the “language” encoding the information remains the same, that of base sequences, whereas the term “translation” indicates that in transferring information from RNA to proteins, the “language” changes from that of base sequences to that of amino acid sequences (Fig. 5-22). The machinery required to carry out the complex tasks of gene expression and DNA replication in an organized manner and with high fidelity occupies a major portion of every cell. In this section we summarize how gene expression and replication occur to provide the background for understanding the techniques of recombinant DNA technology (Section 5-5). This subject matter is explored in considerably greater detail in Chapters 29 to 34.

### A. RNA Synthesis: Transcription

The enzyme that synthesizes RNA is named **RNA polymerase**. It catalyzes the DNA-directed coupling of the **nucleoside triphosphates (NTPs) adenosine triphosphate**

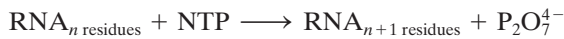
**Figure 5-22** **Gene expression.** One strand of DNA directs the synthesis of RNA, a process known as transcription. The base sequence of the transcribed RNA is complementary to that of the DNA strand. The RNAs known as **messenger RNAs (mRNAs)** are translated when molecules of **transfer RNA (tRNA)** align with the mRNA via complementary base pairing between segments of three consecutive nucleotides known as codons. Each type of tRNA carries a specific amino acid. These amino acids are covalently joined by the ribosome to form a polypeptide. Thus, the sequence of bases in DNA specifies the sequence of amino acids in a protein.



**Figure 5-23 Action of RNA polymerases.**

These enzymes assemble incoming ribonucleoside triphosphates on templates consisting of single-stranded segments of DNA such that the growing strand is elongated in the 5' to 3' direction.

(ATP), cytidine triphosphate (CTP), guanosine triphosphate (GTP), and uridine triphosphate (UTP) in a reaction that releases pyrophosphate ion ( $P_2O_4^{4-}$ ):



RNA synthesis proceeds in a stepwise manner in the 5' → 3' direction, that is, the incoming nucleotide is appended to the free 3'—OH group of the growing RNA chain (Fig. 5-23). RNA polymerase selects the nucleotide it incorporates into the nascent (growing) RNA chain through the requirement that it form a Watson–Crick base pair with the DNA strand that is being transcribed, the **template strand** (only one of duplex DNA's two strands is transcribed at a time). This is possible because, as the RNA polymerase moves along the duplex DNA it is transcribing, it separates a short (~14 bp) segment of its two strands to form a so-called **transcription bubble**, thereby permitting this portion of the template strand to transiently form a short DNA–RNA hybrid helix with the newly synthesized RNA (Fig. 5-24). Like duplex DNA, a DNA–RNA hybrid helix consists of antiparallel strands, and hence the DNA's template strand is read in its 3' → 5' direction.

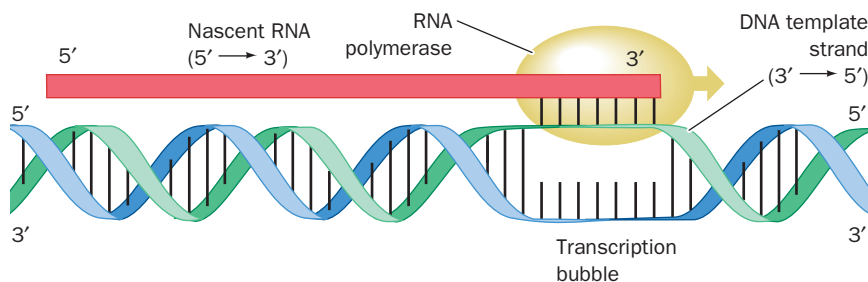
All cells contain RNA polymerase. In bacteria, one species of this enzyme synthesizes nearly all of the cell's RNA. Certain viruses generate RNA polymerases that syn-

thesize only virus-specific RNAs. Eukaryotic cells contain four or five different types of RNA polymerases that each synthesize a different class of RNA.

#### a. Transcriptional Initiation Is a Precisely Controlled Process

The DNA template strand contains control sites consisting of specific base sequences that specify both the site at which RNA polymerase initiates transcription (the site on the DNA at which the RNA's first two nucleotides are joined) and the rate at which RNA polymerase initiates transcription at this site. Specific proteins known in prokaryotes as **activators** and **repressors** and in eukaryotes as **transcription factors** bind to these control sites or to other such proteins that do so and thereby stimulate or inhibit transcriptional initiation by RNA polymerase. For the RNAs that encode proteins, which are named **messenger RNAs (mRNAs)**, these control sites precede the initiation site (that is, they are “upstream” of the initiation site relative to the RNA polymerase's direction of travel).

*The rate at which a cell synthesizes a given protein, or even whether the protein is synthesized at all, is mainly governed by the rate at which the synthesis of the corresponding mRNA is initiated.* The way that prokaryotes regulate the rate at which many genes undergo transcriptional initiation

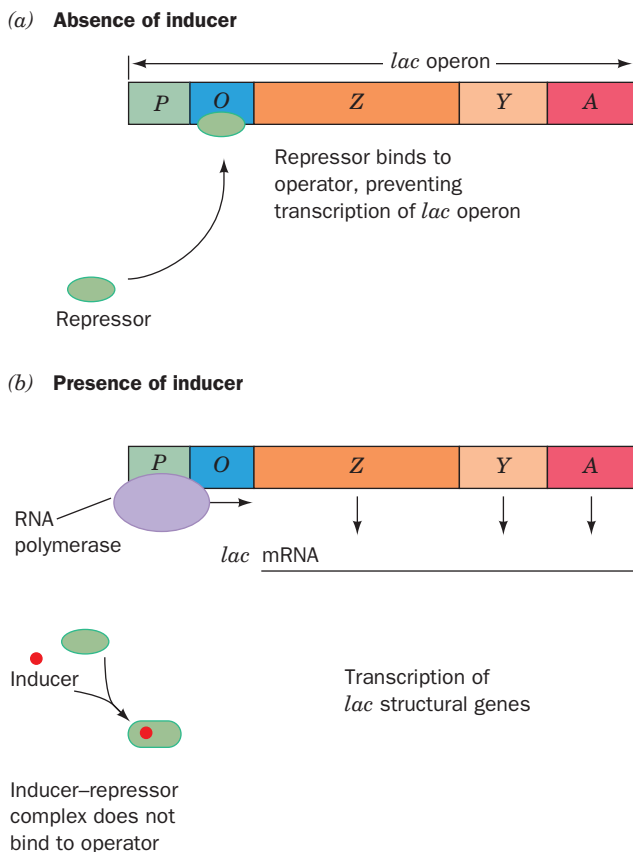


**Figure 5-24 Function of the transcription bubble.** RNA polymerase unwinds the DNA double helix by about a turn in the region being transcribed, thereby permitting the DNA's template strand to form a short segment of DNA–RNA hybrid double helix with the RNA's newly synthesized 3' end. As the

RNA polymerase advances along the DNA template (here to the right), the DNA unwinds ahead of the RNA's growing 3' end and rewinds behind it, thereby stripping the newly synthesized RNA from the template strand.

can be relatively simple. For example, the transcriptional initiation of numerous prokaryotic genes requires only that RNA polymerase bind to a control sequence, known as a **promoter**, that precedes the transcriptional initiation site. However, not all promoters are created equal: RNA polymerase initiates transcription more often at so-called efficient promoters than at those with even slightly different sequences. Thus the rate at which a gene is transcribed is governed by the sequence of its associated promoter.

A more complex way in which prokaryotes control the rate of transcriptional initiation is exemplified by the *E. coli lac operon*, a cluster of three consecutive genes (*Z*, *Y*, and *A*) encoding proteins that the bacterium requires to metabolize the sugar **lactose** (Section 11-2B). In the absence of lactose, a protein named the **lac repressor** specifically binds to a control site in the *lac* operon known as an **operator** (Section 31-3B). This prevents RNA polymerase from initiating the transcription of *lac* operon genes (Fig. 5-25a), thereby halting the synthesis of unneeded proteins. However, when lactose is available, the bacterium meta-



**Figure 5-25** Control of transcription of the *E. coli lac* operon.

(a) In the absence of an inducer such as allolactose, the *lac* repressor binds to the operator (*O*), thereby preventing RNA polymerase from transcribing the *Z*, *Y*, and *A* genes of the *lac* operon. (b) On binding inducer, the *lac* repressor dissociates from the operator, which permits RNA polymerase to bind to the promoter (*P*) and transcribe the *Z*, *Y*, and *A* genes. **See Guided Exploration 2: Regulation of gene expression by the *lac* repressor system**

bologically modifies a small amount of it to form the related sugar **allolactose**. This so-called **inducer** specifically binds to the *lac* repressor, thereby causing it to dissociate from the operator DNA so that RNA polymerase can initiate the transcription of the *lac* operon genes (Fig. 5-25b).

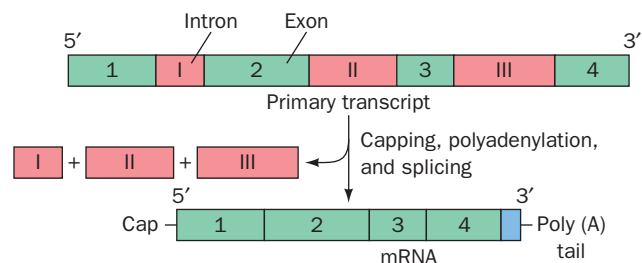
In eukaryotes, the control sites regulating transcriptional initiation can be quite extensive and surprisingly distant from the transcriptional initiation site (by as much as several tens of thousands of base pairs; Section 34-3). Moreover, the eukaryotic transcriptional machinery that binds to these sites and thereby induces RNA polymerase to commence transcription can be enormously complex (consisting of up to 50 or more proteins; Section 34-3).

### b. Transcriptional Termination Is a Relatively Simple Process

The site on the template strand at which RNA polymerase terminates transcription and releases the completed RNA is governed by the base sequence in this region. However, the control of transcriptional termination is rarely involved in the regulation of gene expression. In keeping with this, the cellular machinery that mediates transcriptional termination is relatively simple compared with that involved in transcriptional initiation (Section 31-2D).

### c. Eukaryotic RNA Undergoes Post-Transcriptional Modifications

Most prokaryotic mRNA transcripts participate in translation without further alteration. However, most primary transcripts in eukaryotes require extensive **post-transcriptional modifications** to become functional. For mRNAs, these modifications include the addition of a 7-methylguanosine-containing “cap” that is enzymatically appended to the transcript’s 5’ end and ~250-nucleotide **polyadenylic acid [poly(A)]** “tail” that is enzymatically appended to its 3’ end. However, the most striking modification that most eukaryotic transcripts undergo is a process called **gene splicing** in which one or more often lengthy RNA segments known as **introns** (for “intervening sequences”) are precisely excised from the RNA and the remaining **exons** (for “expressed sequences”) are rejoined in their original order to form the mature mRNA (Fig. 5-26; Section 31-4A). Different mRNAs can be generated from



**Figure 5-26** Post-transcriptional processing of eukaryotic mRNAs. Most primary transcripts require further covalent modification to become functional, including the addition of a 5’ cap and a 3’ poly(A) tail, and splicing to excise its introns from between its exons.



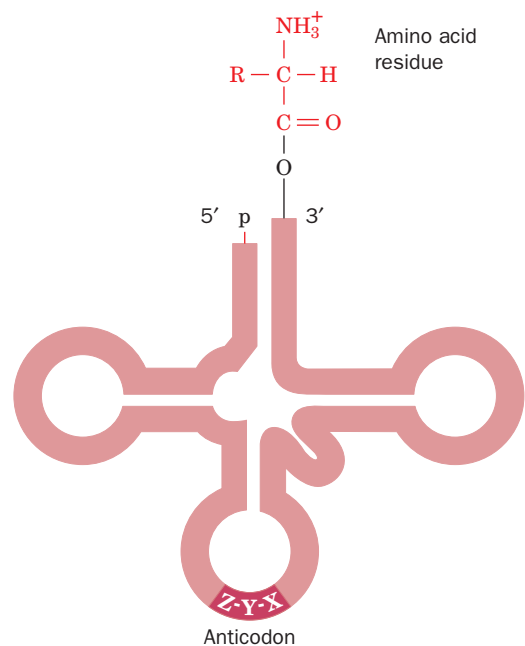
the same gene through the selection of alternate transcriptional initiation sites and/or alternative splice sites, leading to the production of somewhat different proteins, usually in a tissue-specific manner (Section 34-3C).

### B. Protein Synthesis: Translation

Polypeptides are synthesized under the direction of the corresponding mRNA by **ribosomes**, numerous cytosolic organelles that consist of about two-thirds RNA and one-third protein and have molecular masses of  $\sim 2500$  kD in prokaryotes and  $\sim 4200$  kD in eukaryotes. Ribosomal RNAs (**rRNAs**), of which there are several kinds, are transcribed from DNA templates, as are all other kinds of RNA.

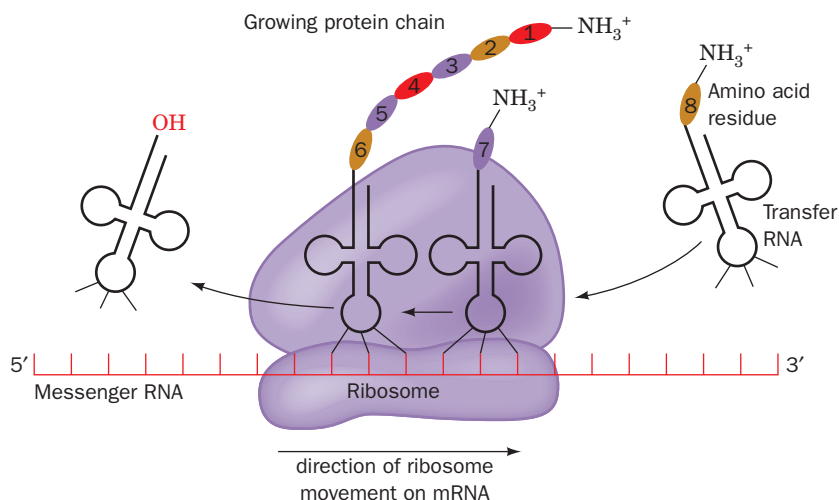
#### a. Transfer RNAs Deliver Amino Acids to the Ribosome

mRNAs are essentially a series of consecutive 3-nucleotide segments known as **codons**, each of which specifies a particular amino acid. However, codons do not bind amino acids. Rather, on the ribosome, they specifically bind molecules of **transfer RNA (tRNA)** that are each covalently linked to the corresponding amino acid (Fig. 5-27). A tRNA typically consists of  $\sim 76$  nucleotides (which makes it comparable in mass and structural complexity to a medium-sized protein) and contains a trinucleotide sequence, its **anticodon**, which is complementary to the codon(s) specifying its appended amino acid (see below). An amino acid is covalently linked to the 3' end of its corresponding tRNA to form an **aminoacyl-tRNA** (a process called “charging”) through the action of an enzyme that specifically recognizes both the tRNA and the amino acid (see below). During



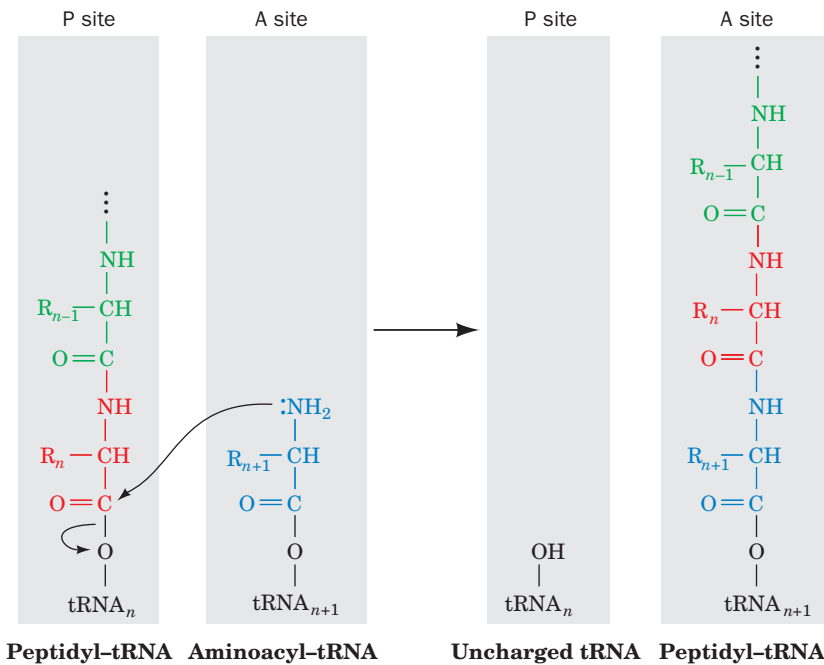
**Figure 5-27** Transfer RNA (tRNA) drawn in its “cloverleaf” form. Its covalently linked amino acid residue forms an aminoacyl-tRNA (top), and its anticodon (bottom), a trinucleotide segment, base pairs with the complementary codon on mRNA during translation.

translation, the mRNA is passed through the ribosome such that each codon, in turn, binds its corresponding aminoacyl-tRNA (Fig. 5-28). As this occurs, the ribosome transfers the amino acid residue on the tRNA to the C-terminal end of



**Figure 5-28** Schematic diagram of translation. The ribosome binds an mRNA and two tRNAs and facilitates their specific association through consecutive codon–anticodon interactions. The ribosomal binding site closer to the 5' end of the mRNA binds a **peptidyl-tRNA** (left, a tRNA to which the growing polypeptide chain is covalently linked) and is therefore known as the **P site**, whereas the ribosomal site closer to the 3' end of the mRNA binds an aminoacyl-tRNA (right) and is hence called the **A site**.

The ribosome catalyzes the transfer of the polypeptide from the peptidyl-tRNA to the aminoacyl-tRNA, thereby forming a new peptidyl-tRNA whose polypeptide chain has increased in length by one residue at its C-terminus. The discharged tRNA in the P site is then ejected, and the peptidyl-tRNA, together with its bound mRNA, is shifted from the A site to the P site, thereby permitting the next codon to bind its corresponding aminoacyl-tRNA in the ribosomal A site.



**Figure 5-29** The ribosomal reaction forming a peptide bond. The amino group of the aminoacyl-tRNA in the ribosomal A site nucleophilically displaces the tRNA of the peptidyl-tRNA ester

the growing polypeptide chain (Fig. 5-29). Hence, *the polypeptide grows from its N-terminus to its C-terminus*.

### b. The Genetic Code

The correspondence between the sequence of bases in a codon and the amino acid residue it specifies is known as the **genetic code** (Table 5-3). Its near universality among all forms of life is compelling evidence that life on Earth arose from a common ancestor and makes it possible, for example, to express human genes in *E. coli* (Section 5-5Ga). There are four possible bases (U, C, A, and G) that can occupy each of the three positions in a codon, and hence there are  $4^3 = 64$  possible codons. Of these codons, 61 specify amino acids (of which there are only 20) and the remaining three, UAA, UAG, and UGA, are **Stop codons** that instruct the ribosome to cease polypeptide synthesis and release the resulting transcript. All but two amino acids (Met and Trp) are specified by more than one codon and three (Leu, Ser, and Arg) are specified by six codons. Consequently, in a term borrowed from mathematics, the genetic code is said to be **degenerate** (taking on several discrete values).

Note that the arrangement of the genetic code is non-random: Most codons that specify a given amino acid, which are known as **synonyms**, occupy the same box in Table 5-3, that is, they differ in sequence only in their third (3') nucleotide. Moreover, most codons specifying non-polar amino acid residues have a G in their first position and/or a U in their second position (Table 5-3).

A tRNA may recognize as many as three synonymous codons because the 5' base of a codon and the 3' base of a

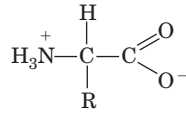
in the ribosomal P site, thereby forming a new peptide bond and transferring the growing polypeptide to the A site tRNA.

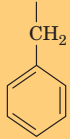
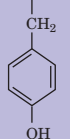


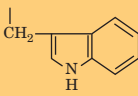
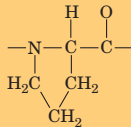
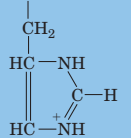
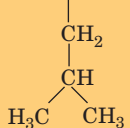
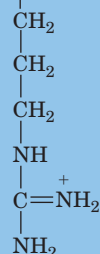
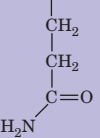
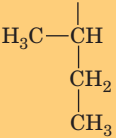
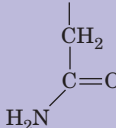

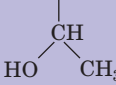
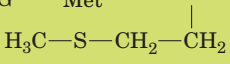
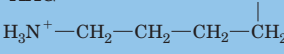
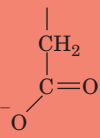
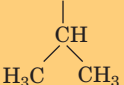


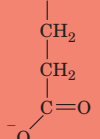
corresponding anticodon do not necessarily interact via a Watson-Crick base pair (Section 32-2D; keep in mind that the codon and the anticodon associate in an antiparallel fashion to form a short segment of an RNA double helix). Thus, cells can have far fewer than the 61 tRNAs that would be required for a 1:1 match with the 61 amino acid-specifying codons, although, in fact, some eukaryotic cells contain over 500 different tRNAs.

### c. tRNAs Acquire Amino Acids Through the Actions of Aminoacyl-tRNA Synthetases

In synthesizing a polypeptide, a ribosome does not recognize the amino acid appended to a tRNA but only whether its anticodon binds to the mRNA's codon (the anticodon and the amino acid on a charged tRNA are actually quite distant from one another, as Fig. 5-27 suggests). Thus, *the charging of a tRNA with the proper amino acid is as critical a step for accurate translation as is the proper recognition of a codon by its corresponding anticodon*. The enzymes that catalyze these additions are known as **aminoacyl-tRNA synthetases (aaRSs)**. Cells typically contain 20 aaRSs, one for each amino acid, and therefore a given aaRS will charge all the tRNAs that bear codons specifying its corresponding amino acid. Consequently, each aaRS must somehow differentiate its cognate (corresponding) tRNAs from among the many other types of structurally and physically quite similar tRNAs that each cell contains. Although many aaRSs recognize the anticodons of their cognate tRNAs, not all of them do so. Rather, they recognize other sites on their cognate tRNAs.

**Table 5-3** The “Standard” Genetic Code<sup>a</sup>



First Position (5' end)	Second Position				Third Position (3' end)
	U	C	A	G	
<b>U</b>	UUU Phe 	UCU	UAU Tyr 	UGU Cys 	<b>U</b>
	UUC	UCC	UAC	UGC	<b>C</b>
	UUA Leu	UCA Ser 	UAA <b>STOP</b>	UGA <b>STOP</b>	<b>A</b>
	UUG	UCG	UAG	UGG Trp 	<b>G</b>
<b>C</b>	CUU	CCU 	CAU His 	CGU	<b>U</b>
	CUC	CCC	CAC	CGC	<b>C</b>
	CUA Leu 	CCA Pro	CAA	CGA Arg 	<b>A</b>
	CUG	CCG	CAG Gln 	CGG	<b>G</b>
<b>A</b>	AUU	ACU	AAU	AGU	<b>U</b>
	AUC Ile 	ACC	AAC Asn 	AGC Ser 	<b>C</b>
	AUA	ACA Thr 	AAA	AGA	<b>A</b>
	AUG Met <sup>b</sup> 	ACG	AAG Lys 	AGG Arg	<b>G</b>
<b>G</b>	GUU	GCU	GAU Asp 	GGU	<b>U</b>
	GUC	GCC	GAC	GGC	<b>C</b>
	GUA Val 	GCA Ala 	GAA	GGA Gly 	<b>A</b>
	GUG	GCG	GAG Glu 	GGG	<b>G</b>

<sup>a</sup>Nonpolar residues are tan, basic residues are blue, acidic residues are red, and polar uncharged residues are purple.

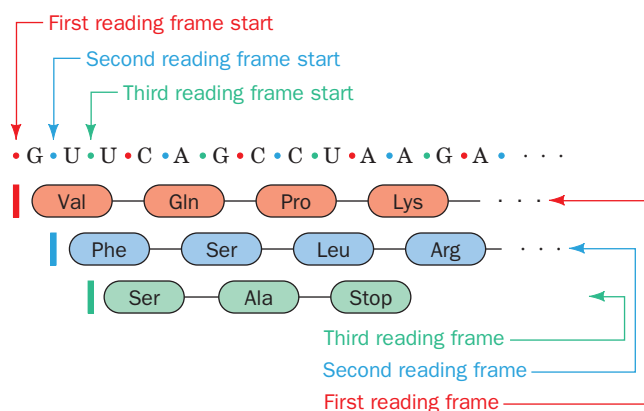
<sup>b</sup>AUG forms part of the initiation signal as well as coding for internal Met residues.

**d. Translation Is Initiated at Specific AUG Codons**

Ribosomes read mRNAs in their 5' to 3' direction (from “upstream” to “downstream”). The initiating codon is AUG, which specifies a Met residue. However, the tRNA that recognizes this initiation codon differs from the tRNA

that delivers a polypeptide’s internal Met residues to the ribosome, although both types of tRNA are charged by the same **methionyl-tRNA synthetase (MetRS)**.

If a polypeptide is to be synthesized with the correct amino acid sequence, it is essential that the ribosome main-



**Figure 5-30 Nucleotide reading frames.** An mRNA might be read in any of three different reading frames, each of which yields a different polypeptide.

tain the proper register between the mRNA and the incoming tRNAs, that is, that the ribosome maintain the correct **reading frame**. As is illustrated in Fig. 5-30, a shift of even one nucleotide along an mRNA will lead to the synthesis of an entirely different polypeptide from the point of the shift onward. Thus, the AUG codon that initiates polypeptide synthesis also sets the polypeptide's reading frame. Yet AUG also specifies a polypeptide's internal Met residues, and an mRNA is likely to contain numerous AUGs in different reading frames. How then does the ribosome select the initiation codon from among the many AUGs in an mRNA? In prokaryotes, the answer is that each mRNA contains a sequence on the upstream (5') side of the initiating codon (a region that does not encode polypeptide chain) through which the ribosome identifies this codon. In eukaryotes, the answer is simpler; the initiating codon is usually the first AUG that is downstream of the mRNA's 5' cap.

#### e. Prokaryotic mRNAs Have Short Lifetimes

In prokaryotes, transcription and translation both take place in the same cellular compartment, the cytosol (Figs. 1-2 and 1-13). Consequently ribosomes often attach to the 5' end of an mRNA before its synthesis is complete and commence synthesizing the corresponding polypeptide. This is essential because, since the mRNAs in prokaryotes have average lifetimes of only 1 to 3 minutes before being hydrolytically degraded by enzymes known as **nucleases**, the 5' end of an mRNA may be degraded before its 3' end is synthesized. This rapid turnover of its mRNAs permits a prokaryote to respond quickly to changes in its environment by synthesizing the proteins appropriate for its new situation within minutes of the change (recall that prokaryotes are adapted to live in environments in which there are rapid fluctuations in the available nutrients; Section 1-2).

Eukaryotic cells, in contrast, mostly lead a more sedentary existence. Their RNAs are transcribed and post-transcriptionally modified in the nucleus, whereas ribosomes occupy the cytosol where translation takes place (Fig. 1-5). Hence, mature mRNAs must be transported from the nucleus to the cytosol in order to participate in translation.

Eukaryotic mRNAs therefore tend to have lifetimes on the order of several days.

#### f. Proteins Are Subject to Post-Translational Modifications and Degradation

Newly synthesized polypeptides often require post-translational modifications to become functional. In many proteins, the leading (N-terminal) Met residue that was specified by its mRNA's initiating codon is excised by a specific **protease** (an enzyme that hydrolytically cleaves peptide bonds). Proteins are then subject to numerous other chemical modifications at specific residues, including specific proteolytic cleavages, acylation, hydroxylation, methylation, and phosphorylation (Section 4-3A). In addition, eukaryotic proteins, but not prokaryotic proteins, are subject to **glycosylation** (the addition of polysaccharides) at specific sites (Sections 11-3C and 23-3B). Indeed, **glycoproteins** (proteins that have been glycosylated) are the most common type of eukaryotic protein and can consist of up to 90% or more by mass of polysaccharide groups.

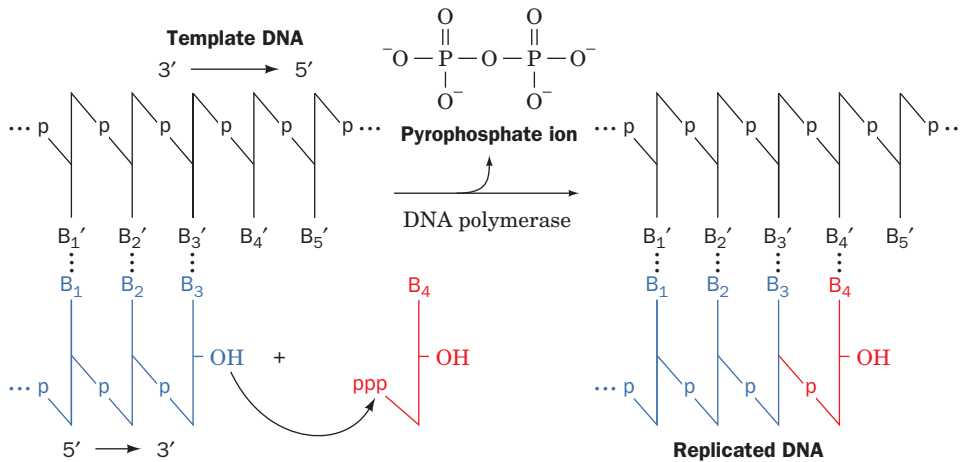
All cells have several mechanisms for degrading proteins to their component amino acids. This enables cells to eliminate damaged or abnormal proteins, destroy proteins that are no longer needed, and utilize proteins as nutrients. The lifetime of a protein in a cell can be surprisingly short, as little as a fraction of a minute, although many proteins in eukaryotes have lifetimes of days or weeks. Thus cells are dynamic entities that are constantly turning over most of their components, in particular their RNA and proteins.

### C. DNA Replication

The chemical reaction by which DNA is replicated (Fig. 5-31) is nearly identical to that synthesizing RNA (Fig. 5-23), but with two major differences: (1) deoxynucleoside triphosphates (**dNTPs**) rather than nucleoside triphosphates are the reactants and (2) the enzyme that catalyzes the reaction is **DNA polymerase** rather than RNA polymerase. The properties of DNA polymerase result in a third major difference between RNA and DNA synthesis: Whereas RNA polymerase can link together two nucleotides on a DNA template, *DNA polymerase can only extend (in the 5' to 3' direction) an existing polynucleotide that is base paired to the DNA's template strand*. Thus, whereas RNA polymerase can initiate RNA synthesis *de novo* (from the beginning), *DNA polymerase requires an oligonucleotide primer, which it lengthens*.

#### a. Primers Are RNA

If DNA polymerase cannot synthesize DNA *de novo*, where do primers come from? It turns out that they are not DNA, as might be expected, but rather RNA. In *E. coli*, these RNA primers are synthesized by both RNA polymerase (the same enzyme that synthesizes all other RNAs) and by a special RNA polymerase known as **primase**. DNA polymerase then extends this RNA primer, which is eventually excised and replaced by DNA, as is explained below. This extra complexity in DNA synthesis increases the fidelity of DNA replication. Whereas a cell makes many copies of an RNA and



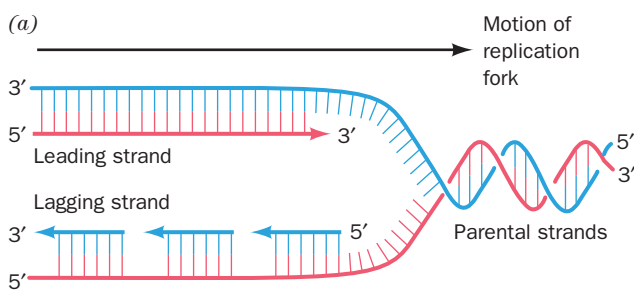
**Figure 5-31** Action of DNA polymerases. DNA polymerases assemble incoming deoxynucleoside triphosphates on single-

stranded DNA templates such that the growing strand is elongated in the 5' to 3' direction.

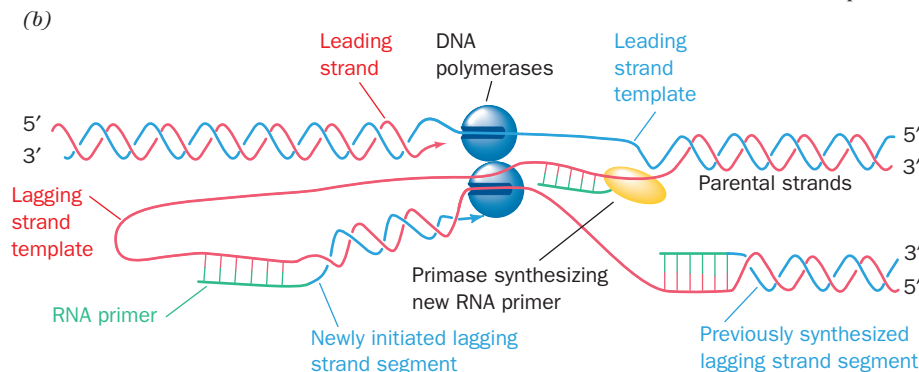
hence can tolerate an occasional mistake in its synthesis, a mistake (mutation) in the synthesis of DNA, the archive of genetic information, may be passed on to all of the cell's descendants. Since a Watson–Crick base pair is partially stabilized by its neighboring base pairs (a cooperative interaction), the first few base pairs that are formed in a newly synthesized polynucleotide will initially be less stable than the base pairs that are formed later. Consequently, these first few bases are more likely to be erroneously incorporated due to mispairing than those at the end of a longer chain. If a primer were DNA, there would be no way to differentiate it from other DNA so as to selectively replace it with more accurately synthesized DNA. Since the primer is RNA, however, it is readily identified and replaced.

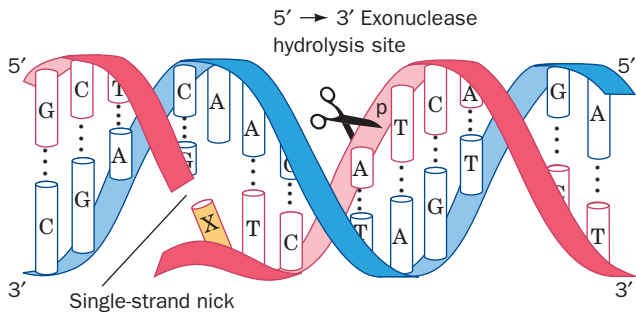
**b. DNA's Two Strands Are Replicated in Different Ways**

A fourth major difference between RNA and DNA synthesis is that, whereas only one DNA strand at a time is transcribed, in most cases both of its strands are simultaneously replicated. This takes place at a **replication fork**, the junction where the two strands of the parental DNA are pried apart and where the two daughter strands are synthesized (Fig. 1-17), each by a different molecule of DNA polymerase. One of these DNA polymerase molecules continuously copies the parental strand that extends in its 3' to 5' direction from the replication fork, thereby synthesizing the resulting daughter strand, which is known as the **leading strand**, in its 5' to 3' direction. However, since the second



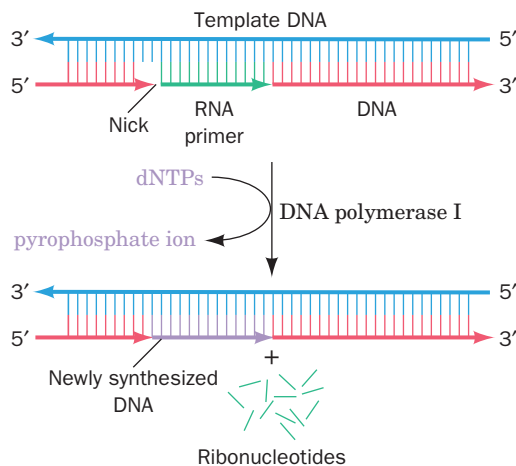
**Figure 5-32** Replication of duplex DNA in *E. coli*. (a) Since the two DNA polymerase molecules at the replication fork are linked together and DNA polymerase can only synthesize DNA in its 5' to 3' direction, the leading strand can be synthesized continuously but the lagging strand must be synthesized discontinuously, that is, in segments. (b) This is because the lagging strand template can only be copied if it loops around so as to feed through the DNA polymerase in its 3' to 5' direction. Consequently, when the DNA polymerase that is synthesizing the lagging strand encounters the previously synthesized lagging strand segment, it releases the lagging strand template and rebinds to it farther upstream so as to extend the next RNA primer to be synthesized.





**Figure 5-33** The  $5' \rightarrow 3'$  exonuclease function of DNA polymerase I. This enzymatic activity excises up to 10 nucleotides from the  $5'$  end of a single-strand nick. The nucleotide immediately past the nick (X) may or may not be base paired.

DNA polymerase at the replication fork also synthesizes DNA in the  $5'$  to  $3'$  direction and yet must travel with the replication fork, how does it copy the parental strand that extends from the replication fork in its  $5'$  to  $3'$  direction? The answer is that it synthesizes the so-called **lagging strand discontinuously**, that is, in pieces (Fig. 5-32a, opposite). It does so by binding the looped-around lagging strand template so as to extend its newly synthesized RNA primer in its  $5'$  to  $3'$  direction (Fig. 5-32b; in effect, reversing its direction of travel) until it encounters the previously synthesized primer. The DNA polymerase then disengages from the lagging strand template and rebinds to it upstream of its previous position, where it then extends the next RNA primer to be synthesized. Thus the lagging strand is synthesized discontinuously, whereas the leading strand is synthesized continuously. The synthesis of lagging strand primers in *E. coli* is catalyzed by primase, which accompanies the replication fork (Fig. 5-32b), whereas the synthesis of leading strand primers, a much rarer event, occurs most efficiently when both primase and RNA polymerase are present.



**Figure 5-34** Replacement of RNA primers by DNA in lagging strand synthesis. In *E. coli*, the RNA primer on the  $5'$  end of a newly synthesized DNA segment is excised through the action of DNA polymerase I's  $5' \rightarrow 3'$  exonuclease activity and is simultaneously replaced by DNA as catalyzed by the enzyme's DNA polymerase activity.

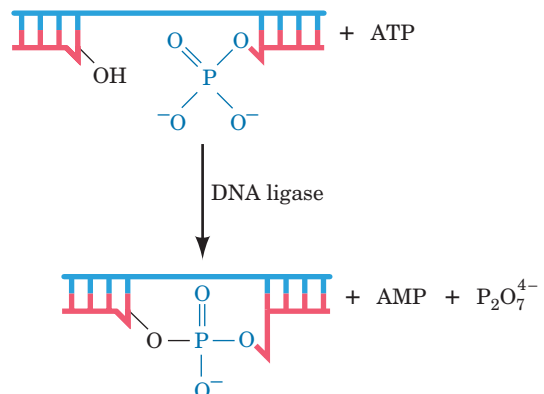
### c. Lagging Strand Synthesis Requires Several Enzymes

*Escherichia coli* contains two species of DNA polymerase that are essential for its survival. Of these, **DNA polymerase III (Pol III)** is the DNA replicase, that is, it synthesizes the leading strand and most of the lagging strand. **DNA polymerase I (Pol I)** has a different function, that of removing the RNA primers and replacing them with DNA. Pol I can do so because it has a second enzymatic activity besides that of a DNA polymerase; it is also a  **$5' \rightarrow 3'$  exonuclease** (an exonuclease hydrolytically removes one or more nucleotides from the end of a polynucleotide rather than cleaving it at an internal position). The  $5' \rightarrow 3'$  exonuclease function binds to single-strand nicks (places where successive nucleotides are not covalently linked such as on the  $5'$  side of an RNA primer after the succeeding lagging strand segment has been synthesized). It then excises a 1- to 10-nucleotide segment of the nicked strand in the  $5'$  to  $3'$  ( $5' \rightarrow 3'$ ) direction past the nick (Fig. 5-33). Pol I's  $5' \rightarrow 3'$  exonuclease and DNA polymerase activities work in concert, so as Pol I's  $5' \rightarrow 3'$  exonuclease removes the primer, its DNA polymerase activity replaces this RNA with DNA (Fig. 5-34).

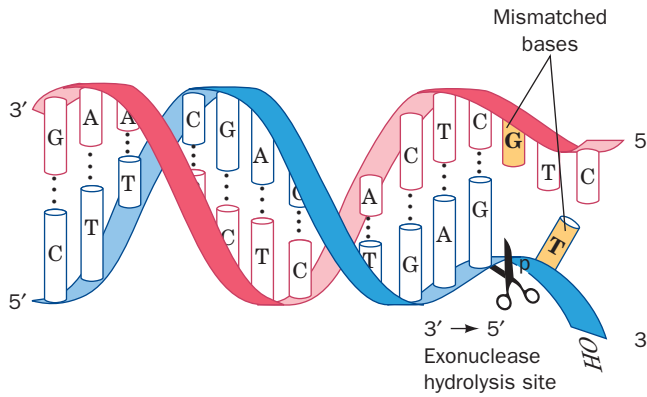
The synthesis of the leading strand is completed by the replacement of its single RNA primer with DNA. However, the completion of lagging strand synthesis requires that the nicks between its multiple discontinuously synthesized segments be sealed. This is the job of an independent enzyme named **DNA ligase** that covalently links adjacent  $3'$ -OH and  $5'$ -phosphate groups (Fig. 5-35).

### d. Errors in DNA Sequences Are Subject to Correction

In *E. coli*, RNA polymerase has an error rate of  $\sim 1$  wrong base for every  $10^4$  nucleotides it transcribes. In contrast, newly replicated DNA contains only  $\sim 1$  error per  $10^8$  to  $10^{10}$  base pairs. We have already seen that the use of RNA primers increases the fidelity of lagging strand synthesis. However, the main reason for the enormous fidelity of DNA replication is that both Pol I and Pol III have  **$3' \rightarrow 5'$  exonuclease** activities. The  $3' \rightarrow 5'$  exonuclease



**Figure 5-35** Function of DNA ligase. DNA ligase seals single-strand nicks in duplex DNA. It does so in a reaction that is powered by the hydrolysis of ATP or a similar compound.



**Figure 5-36** The 3' → 5' exonuclease function of DNA polymerase I and DNA polymerase III. In *E. coli*, this enzymatic activity excises mispaired nucleotides from the 3' end of a growing DNA strand.

degrades the newly synthesized 3' end of a daughter strand one nucleotide at a time (Fig. 5-36), thereby annulling the polymerase reaction. This enzymatic function is activated by non-Watson–Crick base pairing and consequently acts to edit out the occasional mistakes made by the polymerase function, thereby greatly increasing the fidelity of replication. However, in addition to this proofreading function on both Pol I and Pol III, all cells contain batteries of enzymes that detect and correct residual errors in replication as well as damage which DNA incurs through the action of such agents as UV radiation and **mutagens** (substances that damage DNA by chemically reacting with it) as well as by spontaneous hydrolysis (Section 30-5). In *E. coli*, Pol I also functions to replace the damaged DNA segments that these enzymes have excised.

## 5 MOLECULAR CLONING

A major problem in almost every area of biochemical research is obtaining sufficient quantities of the substance of interest. For example, a 10-L culture of *E. coli* grown to its maximum titer of  $\sim 10^{10}$  cells  $\cdot$  mL<sup>-1</sup> contains, at most, 7 mg of DNA polymerase I, and many of its proteins are present in far lesser amounts. Yet it is rare that even as much as half of any protein originally present in an organism can be recovered in pure form (Chapter 6). Eukaryotic proteins may be even more difficult to obtain because many eukaryotic tissues, whether acquired from an intact organism or grown in tissue culture, are available only in small quantities. As far as the amount of DNA is concerned, our 10-L *E. coli* culture would contain  $\sim 0.1$  mg of any 1000-bp length of chromosomal DNA (a length sufficient to contain most prokaryotic genes), but its purification in the presence of the rest of the chromosomal DNA (which consists of 4.6 million bp) would be an all but impossible task. These difficulties have been largely eliminated through the development of **molecular cloning** techniques (a **clone** is a collection of identical organisms that are derived from a single ancestor). These methods, which are also referred to as **genetic engineering** and **recombinant DNA** technology, deserve much of the credit for

the enormous progress in biochemistry and the dramatic rise of the biotechnology industry since the late 1970s.

The main idea of molecular cloning is to insert a DNA segment of interest into an autonomously replicating DNA molecule, a so-called **cloning vector** or **vehicle**, so that the DNA segment is replicated with the vector. Cloning such a **chimeric vector** (*chimera*: a monster in Greek mythology that has a lion's head, a goat's body, and a serpent's tail) in a suitable **host organism** such as *E. coli* or yeast results in the production of large amounts of the inserted DNA segment. If a cloned gene is flanked by the properly positioned control sequences for transcription and translation, the host may also produce large quantities of the RNA and protein specified by that gene. The techniques of genetic engineering, whose understanding is prerequisite to understanding many of the experiments discussed in this textbook, are outlined in this section.

### A. Restriction Endonucleases

In order to effectively carry out molecular cloning, it is necessary to be able to manipulate precisely sequence-defined DNA fragments. This is done through the use of enzymes known as **restriction endonucleases**.

Bacteriophages that propagate efficiently on one bacterial strain, such as *E. coli* K12, have a very low rate of infection ( $\sim 0.001\%$ ) in a related bacterial strain such as *E. coli* B. However, the few viral progeny of this latter infection propagate efficiently in the new host but only poorly in the original host. Evidently, the new host modifies these bacteriophages in some way. What is the molecular basis of this **host-specific modification**? Werner Arber showed that it results from a **restriction–modification system** in the bacterial host, which consists of a restriction endonuclease (alternatively, **restriction enzyme**; endonucleases are enzymes that hydrolytically cleave polynucleotides at internal sites) and a matched **DNA methyltransferase**. *Restriction endonucleases recognize a specific base sequence of four to eight bases in double-stranded DNA and cleave both strands of the duplex*. DNA methyltransferases methylate a specific base (at the amino group of an adenine or either the 5 position or the amino group of a cytosine) in the same base sequence recognized by the matched restriction enzyme.

A restriction enzyme does not cleave its corresponding methylated DNA. A newly replicated strand of bacterial DNA, which is protected from degradation by the methylated parent strand with which it forms a duplex, is methylated before the next cycle of replication. A restriction–modification system therefore protects the bacterium against invasion by foreign (usually viral) DNAs which, once they have been cleaved by a restriction endonuclease, are further degraded by bacterial exonucleases. Invading DNAs are only rarely methylated before being attacked by restriction enzymes. Yet if a viral DNA does become methylated, it is able to reproduce in its new host. Its progeny, however, are no longer methylated in the way that permits them to propagate in the original host (which has different restriction–modification systems).

There are four types of restriction endonucleases, Types I, II, III, and IV. **Type I** and **Type III** restriction enzymes

**Table 5-4** Recognition and Cleavage Sites of Some Type II Restriction Enzymes

Enzyme	Recognition Sequence <sup>a</sup>	Microorganism
AluI	AG↓C*T	<i>Arthrobacter luteus</i>
BamHI	G↓GATC*C	<i>Bacillus amyloliquefaciens</i> H
BglI	GCCNNNN↓NGCC	<i>Bacillus globigii</i>
BglII	A↓GATCT	<i>Bacillus globigii</i>
EcoRI	G↓AA*TTC	<i>Escherichia coli</i> RY13
EcoRII	↓CC*(A)GG	<i>Escherichia coli</i> R245
EcoRV	GA*T↓ATC	<i>Escherichia coli</i> J62 pLG74
HaeII	RGCGC↓Y	<i>Haemophilus aegyptius</i>
HaeIII	GG↓C*C	<i>Haemophilus aegyptius</i>
HindIII	A*↓AGCTT	<i>Haemophilus influenzae</i> R <sub>d</sub>
HpaII	C↓C*GG	<i>Haemophilus parainfluenzae</i>
MspI	C*↓CGG	<i>Moraxella</i> species
PstI	CTGCA*↓G	<i>Providencia stuartii</i> 164
PvuII	CAG↓C*TG	<i>Proteus vulgaris</i>
SalI	G↓TCGAC	<i>Streptomyces albus</i> G
TaqI	T↓CGA*	<i>Thermus aquaticus</i>
XhoI	C↓TCGAG	<i>Xanthomonas holcicola</i>

<sup>a</sup>The recognition sequence is abbreviated so that only one strand, reading 5' to 3', is given. The cleavage site is represented by an arrow (↓) and the modified base, where it is known, is indicated by a following asterisk (A\* is N<sup>6</sup>-methyladenine and C\* is 5-methylcytosine). R, Y, and N represent purine nucleotide, pyrimidine nucleotide, and any nucleotide, respectively.

Source: REBASE. The restriction enzyme database (<http://rebase.neb.com>).

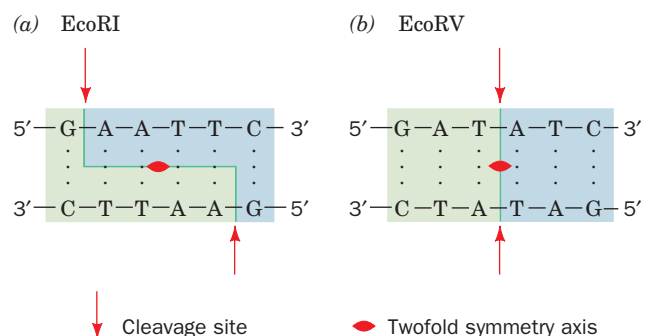
each carry both the endonuclease and the DNA methyltransferase activity on a single protein molecule. **Type I** restriction enzymes cleave the DNA at a possibly random site located at least 1000 bp from the recognition sequence, **Type III** enzymes do so 24 to 26 bp distant from the recognition sequence, and **Type IV** enzymes cleave methylated DNA. **Type II** restriction enzymes, which were discovered and characterized by Hamilton Smith and Daniel Nathans in the late 1960s, are separate entities from their corresponding DNA methyltransferases. They cleave DNAs at specific sites within or near the recognition sequence, a property that makes Type II restriction enzymes indispensable biochemical tools for DNA manipulation. In what follows, we discuss only Type II restriction enzymes.

Nearly 4000 species of Type II restriction enzymes from a variety of bacteria that have over 270 different sequence specificities have been characterized. Several of the more widely used species are listed in Table 5-4. A restriction endonuclease is named by the first letter of the genus of the bacterium that produced it and the first two letters of its species, followed by its serotype or strain designation, if any, and a roman numeral if the bacterium expresses more than one type of restriction enzyme. For example, **EcoRI** is produced by *E. coli* strain RY13.

### a. Most Restriction Endonucleases Recognize Palindromic DNA Sequences

Most Type II restriction enzyme recognition sites possess exact twofold rotational symmetry, as is diagrammed in Fig. 5-37. Such sequences are known as **palindromes** (a palindrome is a word, verse, or sentence that reads the

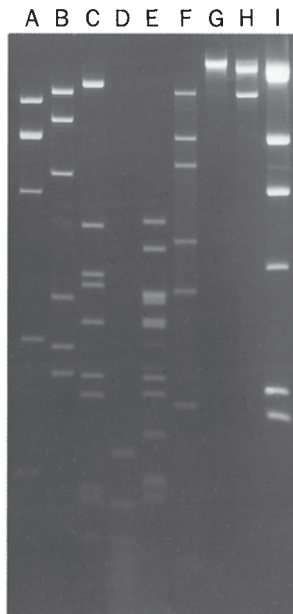
same backward and forward; two examples are “Madam, I’m Adam” and “Sex at noon taxes”). Many restriction enzymes, such as EcoRI (Fig. 5-37a), catalyze cleavage of the two DNA strands at positions that are symmetrically staggered about the center of the palindromic recognition sequence. This yields restriction fragments with complementary single-stranded ends that are from one to four nucleotides in length. Restriction fragments with such **cohesive** or **sticky ends** can associate by complementary base pairing with other restriction fragments generated by the same restriction enzyme. Some restriction cuts, such as that of EcoRV (Fig. 5-37b), pass through the twofold axis of the



**Figure 5-37** Restriction sites. The recognition sequences of the restriction endonucleases (a) EcoRI and (b) EcoRV have twofold (palindromic) symmetry (red symbol). The cleavage sites are indicated (arrows). Note that EcoRI generates DNA fragments with sticky ends, whereas EcoRV generates blunt-ended fragments.



palindrome to yield restriction fragments with fully base paired **blunt ends**. Since a given base has a one-fourth probability of occurring at any nucleotide position (assuming the DNA has equal proportions of all bases), a restriction enzyme with an  $n$ -base pair recognition site produces restriction fragments that are, on average,  $4^n$  base pairs long. Thus AluI (4-bp recognition sequence) and EcoRI (6-bp recognition sequence) restriction fragments should average  $4^4 = 256$  and  $4^6 = 4096$  bp in length, respectively.



**Figure 5-38** Agarose gel electrophoretogram of restriction digests. The *Agrobacterium radiobacter* plasmid pAgK84 was digested with (A) BamHI, (B) PstI, (C) BglII, (D) HaeIII, (E) HincII, (F) SacI, (G) XbaI, and (H) HpaI. Lane I contains  $\lambda$  phage DNA digested with HindIII as standards since these fragments have known sizes. The DNA fragments in the electrophoretogram are made visible by fluorescence against a black background. [From Slota, J.E. and Farrand, S.F., *Plasmid* 8, 180 (1982). Copyright © 1982 by Academic Press.]

### b. Restriction-Fragment Length Polymorphisms Provide Markers for Characterizing Genes

The treatment of a DNA molecule with a restriction endonuclease produces a series of precisely defined fragments that can be separated according to size by **gel electrophoresis** (Fig. 5-38). (In gel electrophoresis, charged molecules are applied to one end of a thin slab of polyacrylamide or agarose gel and are separated through the application of an electric field. Under the conditions used to separate DNA fragments, the molecules move according to size, with the smallest fragments moving fastest. Gel electrophoresis is further discussed in Section 6-4B.) Complementary single strands can be separated either by melting the DNA and subjecting it to gel electrophoresis, or by using density gradient ultracentrifugation in alkaline CsCl solution (recall that DNA is denatured under alkaline conditions).

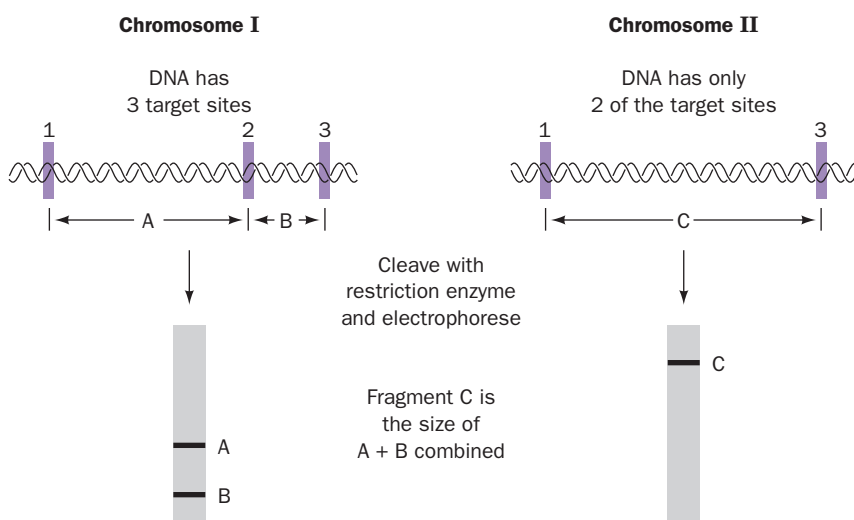
Individuality in humans and other species derives from their genetic polymorphism; homologous human chromosomes differ in sequence, on average, every  $\sim 1250$  bp. These genetic differences create and eliminate restriction sites (Fig. 5-39). Restriction enzyme digests of the corresponding segments from homologous chromosomes therefore contain fragments with different lengths; that is, these DNAs have **restriction-fragment length polymorphisms (RFLPs; Fig. 5-40)**. Since, with the exception of identical twins, each individual has a unique set of RFLPs (its **haplotype**), RFLPs can be used for purposes of identification.

### B. Cloning Vectors

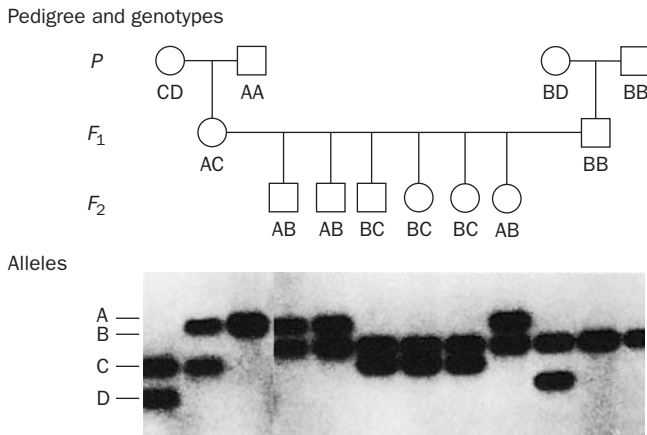
Plasmids, viruses, and artificial chromosomes are used as cloning vectors in genetic engineering.

#### a. Plasmid-Based Cloning Vectors

**Plasmids** are circular DNA duplexes of 1 to 200 kb that contain the requisite genetic machinery, such as a **replication origin** (a site at which DNA replication is initiated; Section 30-3Ca), to permit their autonomous propagation in a bacterial host or in yeast. Plasmids may be considered molecular parasites but in many instances they benefit



**Figure 5-39** Restriction-fragment length polymorphisms. A mutational change that affects a restriction site in a DNA segment alters the number and sizes of its restriction fragments.



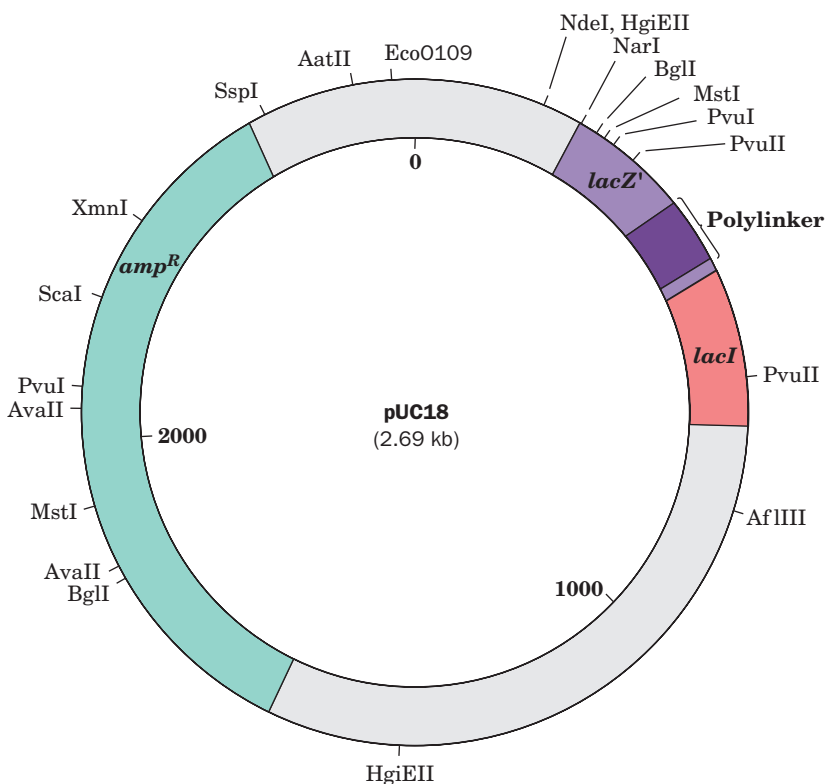
**Figure 5-40** Inheritance of RFLPs according to the rules of Mendelian genetics. Four alleles of a particular gene, each characterized by different restriction markers, can occur in all possible pairwise combinations and segregate independently in each generation (circles in the upper panel represent females and squares represent males). In the  $P$  (parental) generation, two individuals have heterozygous haplotypes (CD and BD) and the other two have homozygous haplotypes (AA and BB). Their children, the  $F_1$  generation, have the haplotypes AC or BB. Consequently, every individual in the  $F_2$  generation (grandchildren) inherited either an A or a C from their mother and a B from their father. The lower panel shows a gel electrophoretogram of these restriction fragments run in parallel lanes. [Courtesy of Ray White, University of Utah Medical School.]

their host by providing functions, such as resistance to antibiotics, that the host lacks. Indeed, the widespread and alarming appearance, since antibiotics came into use, of an-

tibiotic-resistant pathogens is partially the result of the rapid proliferation among these organisms of plasmids containing genes that confer resistance to antibiotics.

Some types of plasmids, which are present in one or a few copies per cell, replicate once per cell division as does the bacterial chromosome; their replication is said to be under **stringent control**. Most plasmids used in molecular cloning, however, are under **relaxed control**; they are normally present in 10 to as many as 700 copies per cell. Moreover, if protein synthesis in the bacterial host is inhibited, for example, by the antibiotic **chloramphenicol** (Section 32-3Gb), thereby preventing cell division, these plasmids continue to replicate until 2 or 3 thousand copies have accumulated per cell (which represents about half of the cell's total DNA). The plasmids that have been constructed (by genetic engineering techniques; Section 5-5C) for use in molecular cloning are relatively small, replicate under relaxed control, carry genes specifying resistance to one or more antibiotics, and contain a number of conveniently located restriction endonuclease sites into which the DNA to be cloned may be inserted. Indeed, many plasmid vectors contain a strategically located short (<100 bp) segment of DNA known as a **polylinker** that has been synthesized to contain a variety of restriction sites that are not present elsewhere in the plasmid. The *E. coli* plasmid designated **pUC18** (Fig. 5-41) is representative of the cloning vectors presently in use ("pUC" stands for "plasmid-Universal Cloning").

The expression of a chimeric plasmid in a bacterial host was first demonstrated in 1973 by Herbert Boyer and Stanley Cohen. The host bacterium takes up a plasmid when the two are mixed together in a process that is greatly enhanced by the presence of divalent cations such as  $\text{Ca}^{2+}$



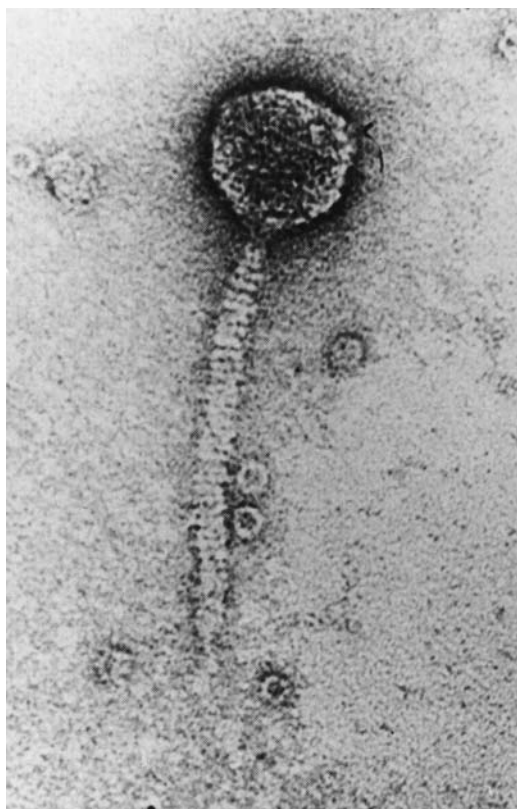
**Figure 5-41** The pUC18 cloning vector. A restriction map of the plasmid pUC18 indicates the positions of its  $amp^R$ ,  $lacZ'$ , and  $lacI$  genes. The  $amp^R$  gene confers resistance to the antibiotic **ampicillin** (a penicillin derivative; Section 11-3Bb);  $lacZ'$  is a modified form of the  $lacZ$  gene, which encodes the enzyme  **$\beta$ -galactosidase** (Section 11-2B); and  $lacI$  encodes the  $lac$  repressor, a protein that controls the transcription of  $lacZ$  (Section 5-4Aa). The polylinker, which encodes an 18-residue polypeptide segment inserted near the N-terminus of  $\beta$ -galactosidase, incorporates 13 different restriction sites that do not occur elsewhere in the plasmid.

and brief heating to  $\sim 42^{\circ}\text{C}$  (which increases cell membrane permeability to DNA; such cells are said to be **transformation competent**). Nevertheless, an absorbed plasmid vector becomes permanently established in its bacterial host (transformation) with an efficiency of only  $\sim 0.1\%$ .

Plasmid vectors cannot be used to clone DNAs of more than  $\sim 10$  kb. This is because the time required for plasmid replication increases with plasmid size. Hence intact plasmids with large, unessential (to them) inserts are lost through the faster proliferation of plasmids that have eliminated these inserts by random deletions.

### b. Virus-Based Cloning Vectors

**Bacteriophage  $\lambda$**  (Fig. 5-42) is an alternative cloning vehicle that can be used to clone DNAs of up to 16 kb. The central third of this virus's 48.5-kb genome is not required for phage infection (Section 33-3Aa) and can therefore be replaced by foreign DNAs of up to slightly greater size using techniques discussed in Section 5-5C. The chimeric phage DNA can then be introduced into the host cells by infecting them with phages formed from the DNA by an *in vitro* packaging system (Section 33-3Bc). The use of phages as cloning vectors has the additional advantage that the chimeric DNA is produced in large amounts and in easily purified form.



**Figure 5-42** Electron micrograph of bacteriophage  $\lambda$ . Bacteriophage  $\lambda$  reproduces in certain strains of *E. coli*. On binding to a susceptible *E. coli*, the DNA contained in the “head” of the phage particle is injected, through its “tail,” into the bacterial cell, where it is replicated  $\sim 100$  times and packaged to form progeny phage (Section 33-3). [Courtesy of A.F. Howatson. From Lewin, B., *Gene Expression*, Vol. 3, Fig. 5.23, Wiley (1977).]

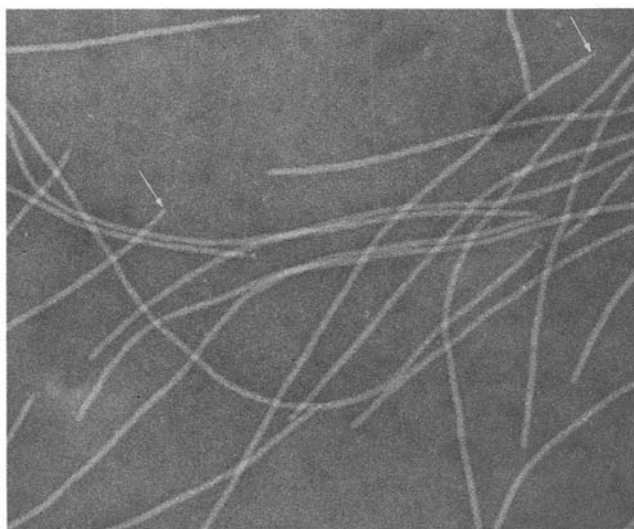
$\lambda$  Phages can be used to clone even longer DNA inserts. The viral apparatus that packages DNA into phage heads requires only that the DNA have a specific 16-bp sequence known as a **cos site** located at each end and that these ends be 36 to 51 kb apart (Section 33-3Bc). Placing two *cos* sites the proper distance apart on a plasmid vector yields, via an *in vitro* packaging system, a so-called **cosmid** vector, which can contain foreign DNA of up to  $\sim 49$  kb. Cosmids have no phage genes and hence, on introduction into a host cell via phage infection, reproduce as plasmids.

The **filamentous bacteriophage M13** (Fig. 5-43) is also a useful cloning vector. It has a single-stranded circular DNA that is contained in a protein tube composed of  $\sim 2700$  helically arranged identical protein subunits. This number is controlled, however, by the length of the phage DNA being coated; insertion of foreign DNA in a nonessential region of the M13 chromosome results in the production of longer phage particles. Although M13 cloning vectors cannot stably maintain DNA inserts of  $>1$  kb, they are widely used in the production of DNA for sequence analysis (Section 7-2Ba) because these phages directly produce the single-stranded DNA that the technique requires.

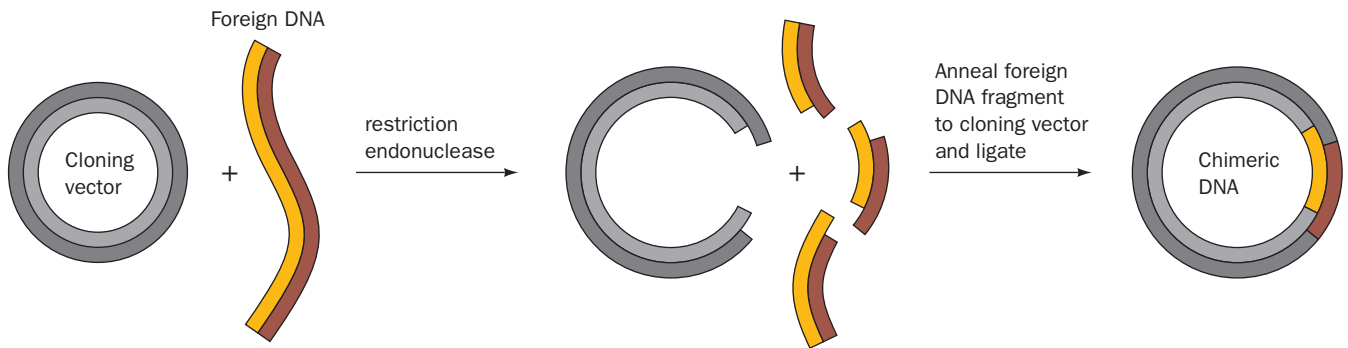
**Baculoviruses** are a large and diverse group of pathogenic viruses that infect mainly insects (but not vertebrates, so that they are safe for laboratory use) and hence can be grown in cultures of insect cells. A segment of the double-stranded DNA that forms the genome of some of these viruses is unnecessary for viral replication in tissue cultures of insect cells and hence can be replaced by a foreign DNA of up to 15 kb.

### c. YAC and BAC Vectors

DNA segments larger than those that can be carried by cosmids may be cloned in **yeast artificial chromosomes (YACs)** and in **bacterial artificial chromosomes (BACs)**. YACs are linear DNA segments that contain all the molecular paraphernalia required for replication in yeast: a replication



**Figure 5-43** Electron micrograph of the filamentous bacteriophage M13. Note that some filaments appear to be pointed at one end (arrows). [Courtesy of Robley Williams, Stanford University, and Harold Fisher, University of Rhode Island.]



**Figure 5-44 Construction of a recombinant DNA molecule.** A restriction fragment is inserted in a cloning vector's corresponding restriction cut. The sticky ends of the vector and the foreign

DNA anneal and are subsequently covalently joined by DNA ligase to yield a chimeric DNA.  See the Animated Figures

origin [known as an **autonomously replicating sequence (ARS)**], a **centromere** (the chromosomal segment that attaches to the spindle during mitosis and meiosis), and **telomeres** (the ends of linear chromosomes that permit their replication; Section 30-4D). BACs, which replicate in *E. coli*, are derived from circular plasmids that normally replicate long regions of DNA and are maintained at a level of approximately one copy per cell (properties similar to those of actual chromosomes). These vectors contain the minimal sequences required for autonomous replication, copy-number control, and proper partitioning of the plasmid during cell division. YACs and BACs containing inserts of several hundred kilobase pairs have been successfully cloned.

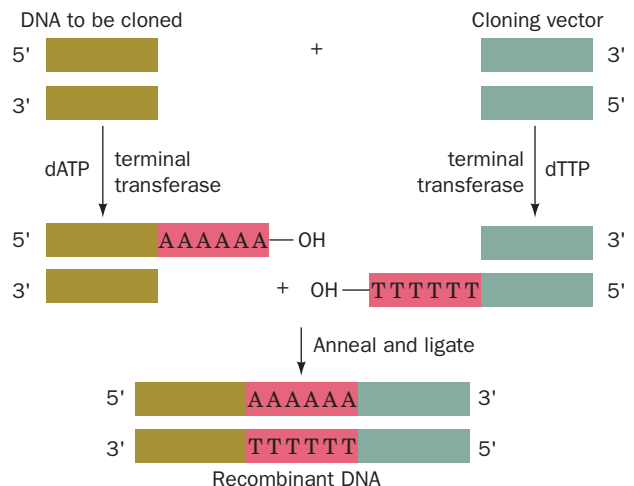
### C. Gene Manipulation

A DNA to be cloned is, in many cases, obtained as a sequence-defined fragment through the application of restriction endonucleases (for M13 vectors, the restriction enzymes' requirement of duplex DNA necessitates converting this phage DNA to its double-stranded form through the use of DNA polymerase I). Recall that most restriction endonucleases cleave duplex DNA at specific palindromic sites so as to yield single-stranded ends that are complementary to each other (cohesive or sticky ends; Section 5-5Aa). Therefore, as Janet Mertz and Ron Davis first demonstrated in 1972, a restriction fragment may be inserted into a cut made in a cloning vector by the same restriction enzyme (Fig. 5-44). The complementary (cohesive) ends of the two DNAs specifically associate under annealing conditions and are covalently joined (spliced) through the action of DNA ligase (Fig. 5-35; the DNA ligase produced by **bacteriophage T4** must be used for blunt-ended restriction cuts such as those generated by AluI, EcoRV, and HaeIII; Table 5-4). A great advantage of using a restriction enzyme to construct a chimeric vector is that the DNA insert can be precisely excised from the cloned vector by cleaving it with the same restriction enzyme.

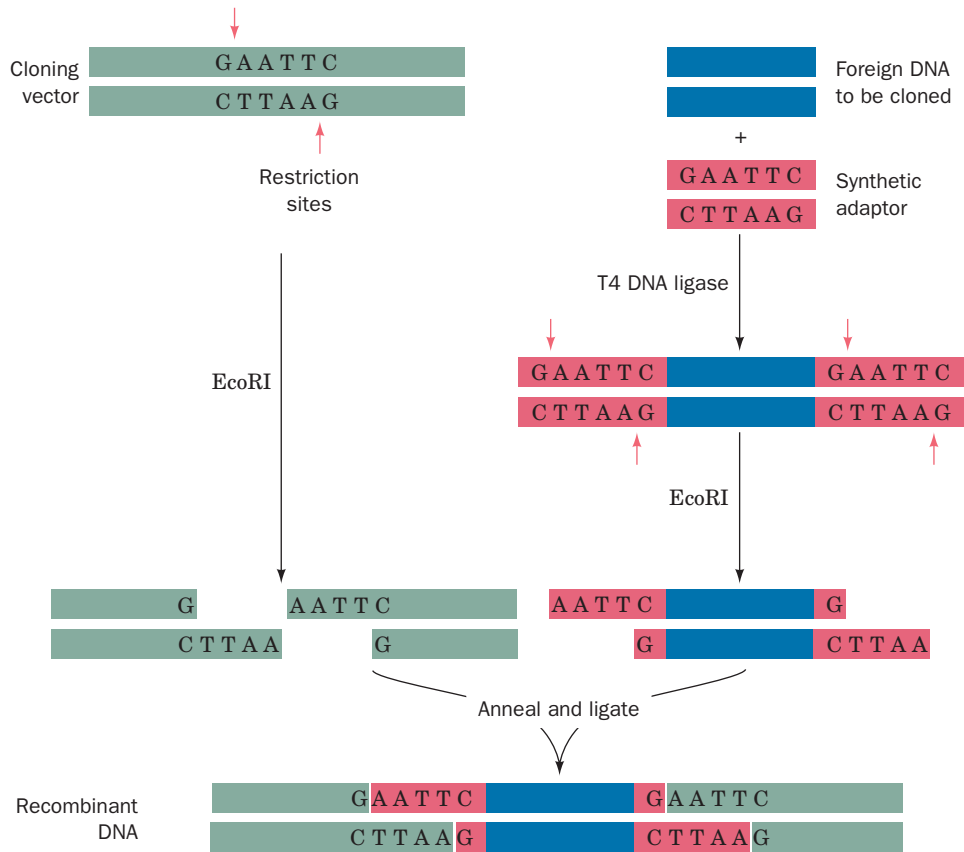
If the foreign DNA and cloning vector have no common restriction sites at innocuous positions, they may still be spliced, using a procedure pioneered by Dale Kaiser and Paul Berg, through the use of **terminal deoxynucleotidyl trans-**

**ferase (terminal transferase)**. This mammalian enzyme adds nucleotides to the 3'-terminal OH group of a DNA chain; it is the only known DNA polymerase that does not require a template. Terminal transferase and dTTP, for example, can build up poly(dT) tails of ~100 residues on the 3' ends of the DNA segment to be cloned (Fig. 5-45). The cloning vector is enzymatically cleaved at a specific site and the 3' ends of the cleavage site are similarly extended with poly(dA) tails. The complementary homopolymer tails are annealed, any gaps resulting from differences in their lengths filled in by DNA polymerase I, and the strands joined by DNA ligase.

A disadvantage of the above technique is that it eliminates the restriction sites that were used to generate the foreign DNA insert and to cleave the vector. It may therefore be difficult to recover the insert from the cloned vector. This difficulty is circumvented by a technique in which a chemically synthesized palindromic "linker" having a restriction site matching that of the cloning vector is appended to both ends of the foreign DNA (the chemical



**Figure 5-45 Splicing DNA using terminal transferase.** Two DNA fragments may be joined through the generation of complementary homopolymer tails via the action of the enzyme terminal transferase. The poly(dA) and poly(dT) tails shown in this example may be replaced by poly(dC) and poly(dG) tails.



**Figure 5-46** Construction of a recombinant DNA molecule through the use of synthetic oligonucleotide adaptors. In this ex-

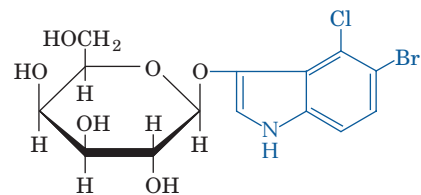
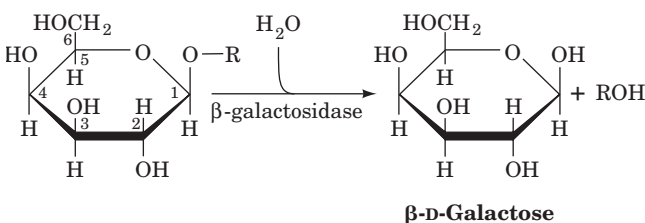
ample, the adaptor and the cloning vector have EcoRI restriction sites (red arrows).

synthesis of oligonucleotides is discussed in Section 7-6A). The linker is attached to the foreign DNA by blunt end ligation with T4 DNA ligase and then cleaved with the appropriate restriction enzyme to yield the correct cohesive ends for ligation to the vector (Fig. 5-46).

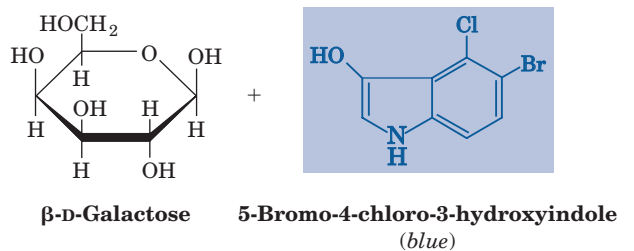
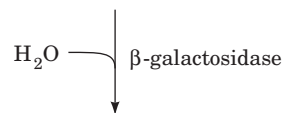
**a. Properly Transformed Cells Must Be Selected**

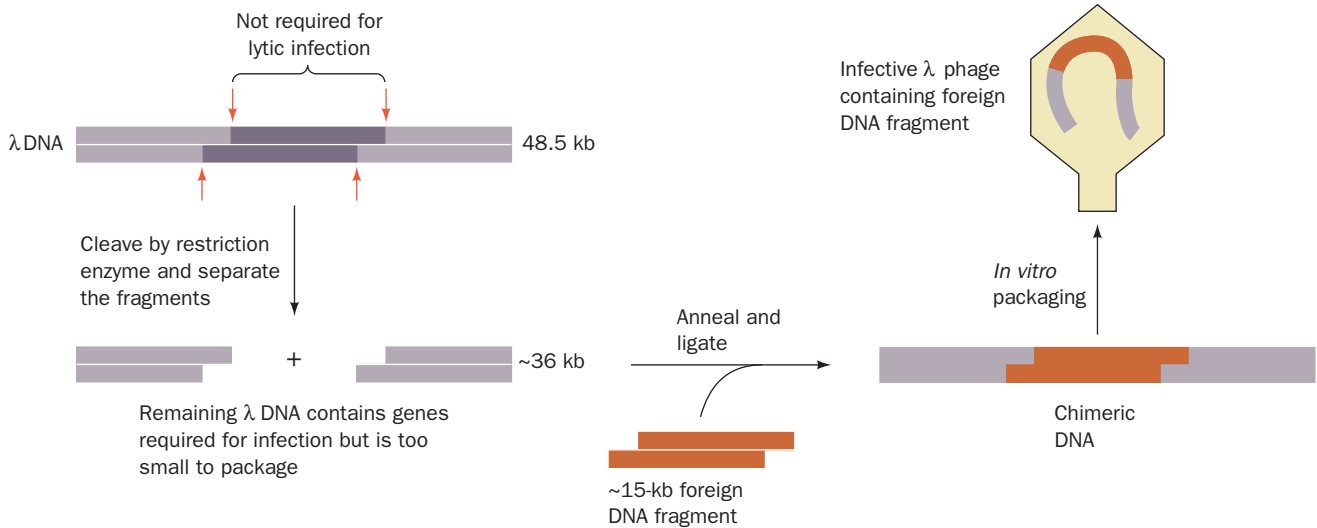
Both transformation and the proper construction of chimeric vectors occur with low efficiency. How can one select only those host organisms that have been transformed by the properly constructed vector? In the case of plasmid transformation, this is usually done through a double screen using antibiotics and/or **chromogenic** (color-producing) substrates. For example, the pUC18 plasmid contains the *lacZ'* gene (Fig. 5-41; a modified form of the *lac* operon's *Z* gene; Fig. 5-25). The *lacZ'* gene encodes the enzyme  $\beta$ -galactosidase, which catalyzes the hydrolysis of the bond from O1 of the sugar  $\beta$ -D-galactose to a substituent.

Thus, when grown in the presence of **5-bromo-4-chloro-3-indolyl- $\beta$ -D-galactoside** (commonly known as **X-gal**), a colorless substance which when hydrolyzed by  $\beta$ -galactosidase yields a blue product,



**5-Bromo-4-chloro-3-indolyl- $\beta$ -D-galactoside (X-gal)**  
(colorless)





**Figure 5-47 Cloning of foreign DNA in  $\lambda$  phages.** A nonessential portion of the phage genome can be replaced by a foreign DNA and packaged to form an infectious phage particle only if

the foreign DNA is approximately the same size as the DNA segment it replaced.  See the Animated Figures

*E. coli* transformed by an unmodified pUC18 plasmid form blue colonies. However, *E. coli* transformed by a pUC18 plasmid containing a foreign DNA insert in its polylinker region form colorless colonies because the insert interrupts the protein-encoding sequence of the *lacZ'* gene and hence they lack  $\beta$ -galactosidase activity. Bacteria that have failed to take up any plasmid, which would otherwise also form colorless colonies in the presence of X-gal, are excluded by adding the antibiotic **ampicillin** (Fig. 11-25) to the growth medium. Bacteria that do not contain the plasmid are sensitive to ampicillin, whereas bacteria containing the plasmid will grow, because the plasmid's intact *amp<sup>R</sup>* gene confers ampicillin resistance. Genes such as *amp<sup>R</sup>* are therefore known as **selectable markers**.

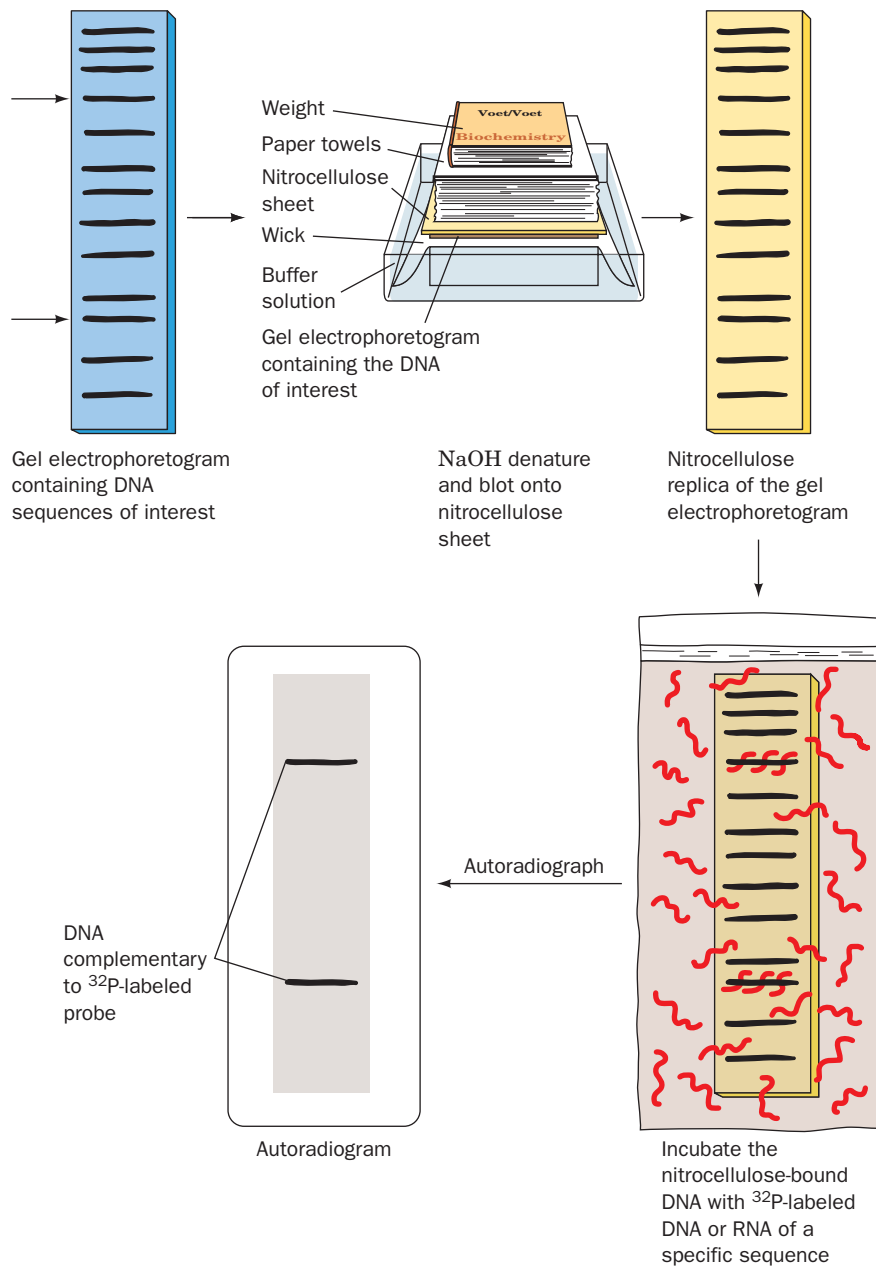
Genetically engineered  $\lambda$  phage variants contain restriction sites that flank the dispensable central third of the phage genome (Section 5-5Bb). This segment may therefore be replaced, as is described above, by a foreign DNA insert (Fig. 5-47). DNA is only packaged in  $\lambda$  phage heads if its length is from 75 to 105% of the 48.5-kb wild-type  $\lambda$  genome. Consequently,  $\lambda$  phage vectors that have failed to acquire a foreign DNA insert are unable to propagate because they are too short to form infectious phage particles. Cosmid vectors are subject to the same limitation. Moreover, cloned cosmids are harvested by repackaging them into phage particles. Hence, any cosmids that have lost sufficient DNA through random deletion to make them shorter than the above limit are not recovered. This is why cosmids can support the proliferation of large DNA inserts, whereas most other types of plasmids cannot.

#### **D. The Identification of Specific DNA Sequences: Southern Blotting**

DNA with a specific base sequence may be identified through a procedure developed by Edwin Southern known

as the **Southern transfer technique** or, more colloquially, as **Southern blotting** (Fig. 5-48). This procedure takes advantage of the valuable property of nitrocellulose that it tenaciously binds single-stranded (but not duplex) DNA [**nylon** and **polyvinylidene fluoride (PVDF)** membranes also have this property]. Following the gel electrophoresis of double-stranded DNA, the gel is soaked in 0.5M NaOH solution, which converts the DNA to its single-stranded form. The gel is then overlaid by a sheet of nitrocellulose paper which, in turn, is covered by a thick layer of paper towels, and the entire assembly is compressed by a heavy plate. The liquid in the gel is thereby forced (blotted) through the nitrocellulose so that the single-stranded DNA binds to it at the same position it had in the gel (the transfer to nitrocellulose can alternatively be accomplished by an electrophoretic process called **electroblotting**). After vacuum drying the nitrocellulose at  $\sim 80^\circ\text{C}$ , which permanently fixes the DNA in place, the nitrocellulose sheet is moistened with a minimal quantity of solution containing  $^{32}\text{P}$ -labeled single-stranded DNA or RNA (the “probe”) that is complementary in sequence to the DNA of interest. The moistened sheet is held at a suitable renaturation temperature for several hours to permit the probe to anneal to its target sequence(s), washed to remove the unbound radioactive probe, dried, and then autoradiographed by placing it for a time (hours to days) over a sheet of X-ray film. The positions of the molecules that are complementary to the radioactive sequences are indicated by a blackening of the developed film. Alternatively, a **phosphorimager**, essentially “electronic film” that detects radioactivity with tenfold greater sensitivity than X-ray film, can be used.

A DNA segment containing a particular base sequence (e.g., an RFLP) may, in this manner, be detected and isolated. The radioactive probe used in this procedure can be the corresponding mRNA if it is produced in sufficient quantity to be isolated [e.g., **reticulocytes** (immature red

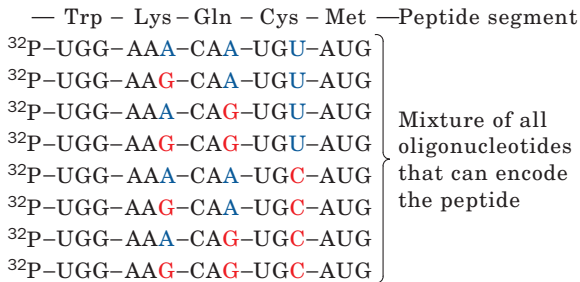


**Figure 5-48** Detection of DNAs containing specific base sequences by the Southern transfer technique.

blood cells), which produce little protein besides **hemoglobin** (the red protein that transports  $\text{O}_2$  in the blood), are rich in the mRNAs that specify hemoglobin]. Alternatively, the gene specifying a protein of known amino acid sequence may be found by synthesizing a probe that is a mixture of all oligonucleotides, which, according to the genetic code (Table 5-3), can specify a segment of the gene that has low degeneracy (Fig. 5-49).

*Southern blotting may be used for the diagnosis and prenatal detection of genetic diseases. These diseases often result from a specific change in a single gene such as a base substitution, deletion, or insertion.* The temperature at which probe hybridization is carried out may be adjusted so that only an oligonucleotide that is perfectly complementary to

a length of DNA will hybridize to it. Even a single base mismatch, under appropriate conditions, will result in a failure to hybridize. For example, the genetic disease **sickle-cell anemia** arises from a single  $\text{A} \rightarrow \text{T}$  base change in the gene specifying the  $\beta$  subunit of hemoglobin, which causes the amino acid substitution  $\text{Glu } \beta 6 \rightarrow \text{Val}$  (Section 7-3Aa). A 19-residue oligonucleotide that is complementary to the sickle-cell gene's mutated segment hybridizes, at the proper temperature, to DNA from homozygotes for the sickle-cell gene but not to DNA from normal individuals. An oligonucleotide that is complementary to the gene encoding the  $\beta$  subunit for normal hemoglobin yields opposite results. DNA from sickle-cell heterozygotes (who have one hemoglobin  $\beta$  gene bearing the sickle-cell anemia mu-



**Figure 5-49 A degenerate oligonucleotide probe.** Such a probe is a mixture of all oligonucleotides that can encode a polypeptide segment of known sequence. In practice, such a segment is chosen to contain a high proportion of residues specified by low-degeneracy codons. In the pentapeptide segment shown here, Trp and Met are each specified by only one codon and Lys, Gln, and Cys are each specified by two codons that differ in only their third positions (*blue and red*; Table 5-3) for a total of  $1 \times 2 \times 2 \times 2 \times 1 = 8$  oligonucleotides. The oligonucleotides are <sup>32</sup>P-labeled for use in Southern blotting.

tation and one that is normal) hybridizes to both probes but in reduced amounts relative to the DNAs from homozygotes. The oligonucleotide probes may consequently be used in the prenatal diagnosis of sickle-cell anemia. (Note that the availability of fetal genetic testing has actually increased the number of births because many couples who knew they had a high risk of conceiving a genetically defective child previously chose not to have children.) The use of DNA probes is also rapidly replacing the much slower and less accurate culturing techniques for the identification of pathogenic bacteria.

In a variation of the Southern blotting procedure, specific DNAs may be detected by linking the probe to an enzyme that generates a colored or fluorescent deposit on the blot when exposed to the proper reagents. Alternatively, a probe that is covalently linked to a dye that fluoresces when stimulated by a laser may be used. Such nonradioactive detection techniques are desirable in a clinical setting because of the health hazards, disposal problems, and more cumbersome nature of radiographic methods. Specific RNA sequences may be detected through a different variation of the Southern transfer, punningly named a **North-ern transfer (Northern blot)**, in which the RNA is immobilized on nitrocellulose paper and detected through the use of complementary RNA or DNA probes.

## E. Genomic Libraries

In order to clone a particular DNA fragment, it must first be obtained in relatively pure form. The magnitude of this task may be appreciated when it is realized that, for example, a 1-kb fragment of human DNA represents only 0.000033% of the 3 billion-bp human genome. A DNA fragment may be identified by Southern blotting of a restriction digest of the genomic DNA under investigation. In practice, however, it is usually more difficult to identify a particular gene from an organism and then clone it than

it is to clone the organism's entire genome as DNA fragments and then identify the clone(s) containing the sequence(s) of interest. Such a set of cloned fragments is known as a **genomic library**. A genomic library of a particular organism need only be made once since it can be perpetuated for use whenever a new probe becomes available.

Genomic libraries are generated according to a procedure known as **shotgun cloning**. The chromosomal DNA of the organism of interest is isolated, cleaved to fragments of clonable size, and inserted in a cloning vector by the methods described in Section 5-5B. The DNA is fragmented by partial rather than exhaustive restriction digestion (permitting the restriction enzyme to act for only a short time) so that the genomic library contains intact representatives of all the organism's genes, including those whose sequences contain restriction sites. Shear fragmentation by rapid stirring of a DNA solution or by sonication is also used but requires further treatment of the fragments to insert them into cloning vectors. Genomic libraries have been established for numerous organisms including yeast, *Drosophila*, mice, and humans.

### a. Many Clones Must Be Screened to Obtain a Gene of Interest

The number of random cleavage fragments that must be cloned to ensure a high probability that a given sequence is represented at least once in the genomic library is calculated as follows: The probability  $P$  that a set of  $N$  clones contains a fragment that constitutes a fraction  $f$ , in bp, of the organism's genome is

$$P = 1 - (1 - f)^N \quad [5.2]$$

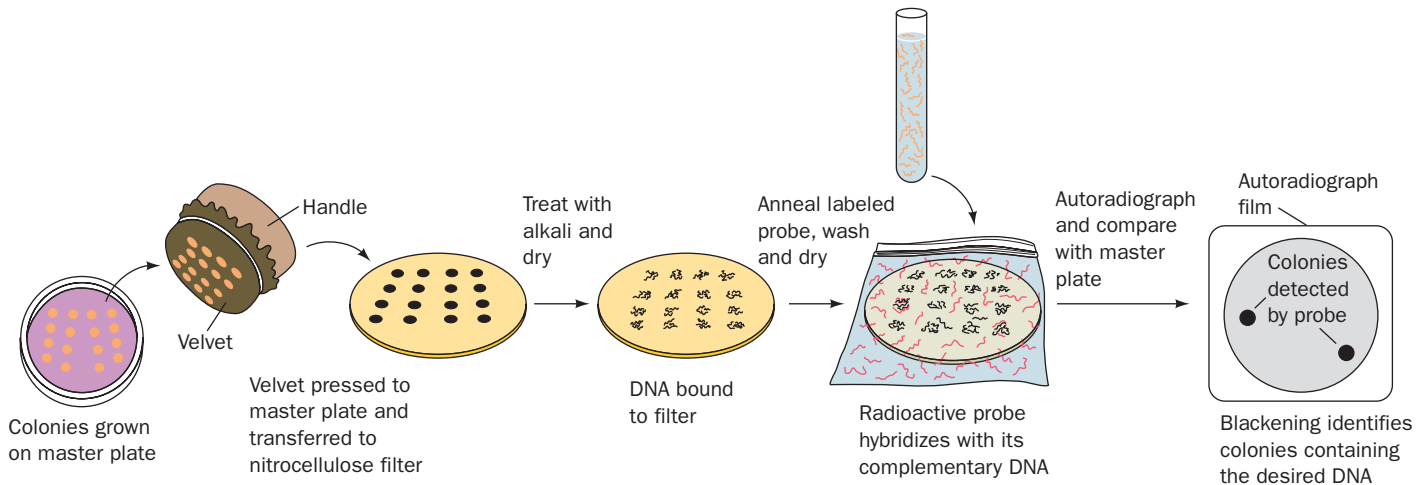
Consequently,

$$N = \log(1 - P)/\log(1 - f) \quad [5.3]$$

Thus, to have  $P = 0.99$  with fragments averaging 10 kb in length, for the 4639-kb *E. coli* chromosome ( $f = 0.00216$ ),  $N = 2134$  clones, whereas for the 180,000-kb *Drosophila* genome ( $f = 0.0000566$ ),  $N = 83,000$ . The use of YAC- or BAC-based genomic libraries therefore greatly reduces the effort necessary to obtain a given gene segment from a large genome.

Since a genomic library lacks an index, it must be screened for the presence of a particular gene. This is done by a process known as **colony** or **in situ hybridization** (Fig. 5-50; Latin: *in situ*, in position). The cloned yeast colonies, bacterial colonies, or phage plaques to be tested are transferred, by replica plating (Fig. 1-30), from a master plate to a nitrocellulose filter. The filter is treated with NaOH, which lyses the cells or phages and denatures the DNA so that it binds to the nitrocellulose (recall that single-stranded DNA is preferentially bound to nitrocellulose). The filter is then dried to fix the DNA in place, treated under annealing conditions with a radioactive probe for the gene of interest, washed, and autoradiographed. *Only those colonies or plaques containing the sought-after gene will bind the probe and thereby blacken the film.* The corresponding clones can then be retrieved from the master plate.






**Figure 5-50** Colony (*in situ*) hybridization. This technique identifies the clones containing a DNA of interest.

Using this technique, even an  $\sim 1$  million clone human genomic library can be readily screened for the presence of one particular DNA segment.

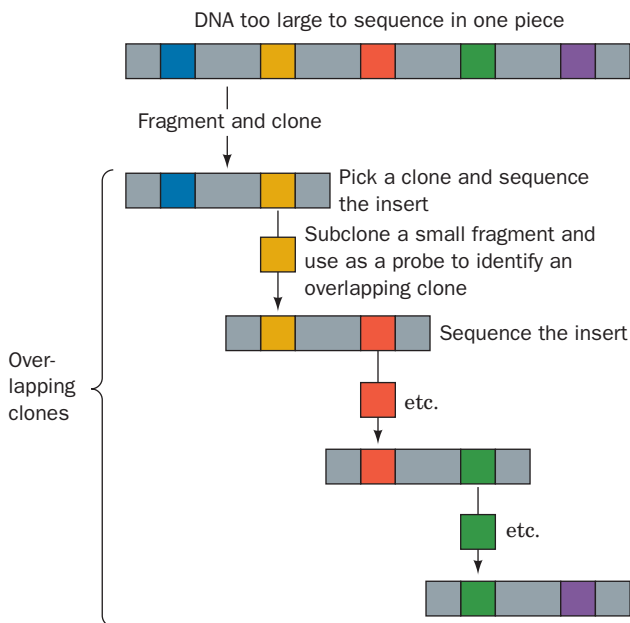
Many eukaryotic genes and gene clusters span enormous tracts of DNA (Section 34-2); some consist of  $>1000$  kb. With the use of plasmid-, phage-, or cosmid-based genomic libraries, such long DNAs can only be obtained as a series of overlapping fragments (Fig. 5-51). Each gene frag-

ment that has been isolated is, in turn, used as a probe to identify a successive but partially overlapping fragment of that gene, a process called **chromosome walking**. The use of YACs and BACs, however, greatly reduces the need for this laborious and error-prone process.

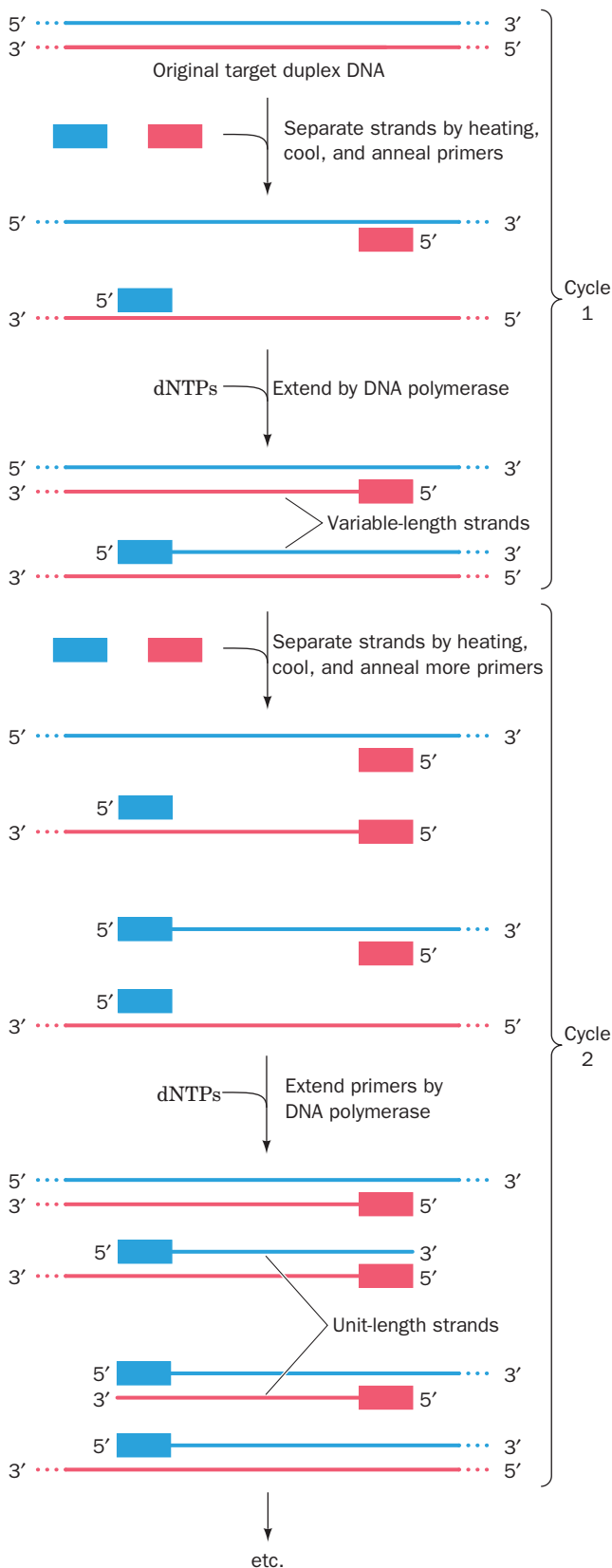
## F. The Polymerase Chain Reaction

 **See Guided Exploration 3: PCR and site-directed mutagenesis** Although molecular cloning techniques are indispensable to modern biochemical research, the use of the **polymerase chain reaction (PCR)** offers a faster and more convenient method of amplifying a specific DNA segment of up to 6 kb. In this technique (Fig. 5-52), which was formulated in 1985 by Kary Mullis, a heat denatured (strand-separated) DNA sample is incubated with DNA polymerase, dNTPs, and two oligonucleotide primers whose sequences flank the DNA segment of interest so that they direct the DNA polymerase to synthesize new complementary strands. Multiple cycles of this process, each doubling the amount of DNA present, geometrically amplify the DNA starting from as little as a single gene copy. In each cycle, the two strands of the duplex DNA are separated by heat denaturation at  $95^{\circ}\text{C}$ , the temperature is then lowered to permit the primers to anneal to their complementary segments on the DNA, and the DNA polymerase directs the synthesis of the complementary strands (Section 5-4C). The use of a heat-stable DNA polymerase, such as those from the thermophilic bacteria *Thermus aquaticus* (**Taq DNA polymerase**) or *Pyrococcus furiosus* (**Pfu DNA polymerase**), both of which are stable at  $95^{\circ}\text{C}$ , eliminates the need to add fresh enzyme after each heat denaturation step. Hence, in the presence of sufficient quantities of primers and dNTPs, the PCR is carried out simply by cyclically varying the temperature in an automatic device called a **thermocycler**.

Twenty cycles of PCR amplification theoretically increase the amount of the target sequence by  $2^{20} \approx 10^6$ -fold with high specificity (in practice, the number of copies of target sequence largely doubles with each PCR cycle until



**Figure 5-51** Chromosome walking. A DNA segment too large to sequence in one piece is fragmented and cloned. A clone is picked and the DNA insert it contains is sequenced. A small fragment of the insert near one end is subcloned (cloned from a clone) and used as a probe to select a clone containing an overlapping insert, which, in turn, is sequenced. The process is repeated so as to “walk” down the chromosome. Chromosome walking can, of course, extend in both directions.



**Figure 5-52 The polymerase chain reaction (PCR).** In each cycle of the reaction, the strands of the duplex DNA are separated by heat denaturation, the preparation is cooled such that synthetic DNA primers anneal to a complementary segment on each strand, and the primers are extended by DNA polymerase. The process is then repeated for numerous cycles. The number of “unit-length” strands doubles with every cycle after the second cycle.

more primer–template complex accumulates than the DNA polymerase can extend during a cycle, whereon the rate of increase of target DNA becomes linear rather than geometric; the actual yield is typically ~20% of the theoretical yield). Indeed, the method has been shown to amplify a target DNA present only once in a sample of  $10^5$  cells, thereby demonstrating that the method can be used without prior DNA purification (although, as a consequence of this enormous amplification, particular care must be taken that the DNA sample of interest is not contaminated by extraneous DNA that is similar in sequence to that under investigation). The amplified DNA can be characterized by a variety of techniques including RFLP analysis, Southern blotting, and direct sequencing (Section 7-2A). PCR amplification is therefore a form of “cell-free molecular cloning” that can accomplish in an automated *in vitro* procedure requiring as little as 30 minutes what would otherwise take days or weeks via the cloning techniques discussed above.

#### a. PCR Has Many Uses

*PCR amplification has become an indispensable tool in a great variety of applications.* Clinically, it is used for the rapid diagnosis of infectious diseases and the detection of rare pathological events such as mutations leading to cancer (Section 19-3Ba). Forensically, the DNA from a single hair, sperm, or drop of blood can be used to identify the donor. This is most commonly done through the analysis of **short tandem repeats (STRs)**, segments of DNA that contain repeating sequences of 2 to 7 bp such as  $(CA)_n$  and  $(ATGC)_n$  and that are scattered throughout the genome [e.g., the human genome contains ~100,000  $(CA)_n$  STRs]. The number of tandem repeats,  $n$ , in many STRs is genetically variable [ $n$  varies from 1 to 40 for particular  $(CA)_n$  STRs] and hence such repeats are markers of individuality (much as are RFLPs). The DNA of a particular STR can be PCR-amplified through the use of primers that are complementary to the unique (nonrepeating) sequences flanking the STR. The number of tandem repeats in that STR from a particular individual can then be determined, either by the measurement of its molecular mass through polyacrylamide gel electrophoresis (Section 6-6C) or by direct sequencing (Section 7-2A). The determination of this number for several well-characterized STRs (those whose numbers of repeats have been determined in numerous individuals of multiple ethnicities), that is, the DNA’s haplotype, can unambiguously identify the DNA’s donor.

STRs are also widely used to prove or disprove familial relationships. For example, oral tradition suggests that

Thomas Jefferson, the third American president, fathered a son, Eston Hemings (born in 1808), with his slave Sally Hemings (Eston Hemings was said to bear a striking physical resemblance to Jefferson). Only the tips of the Y chromosome undergo recombination (with the X chromosome) and the rest of it is passed unchanged from father to son (except for occasional mutations). The finding that the Y chromosomes of male-line descendants of both Eston Hemings and Jefferson's father's brother (Jefferson had no surviving legitimate sons) had identical STR-based haplotypes indicates that Thomas Jefferson was probably Eston Hemings' father (although this could also be true of any of Jefferson's contemporary male-line relatives).

RNA may also be amplified via PCR by first reverse-transcribing it into a complementary strand of DNA (cDNA) through the action of an enzyme named **RNA-directed DNA polymerase** (commonly known as **reverse transcriptase**). This enzyme, which is produced by certain RNA-containing viruses known as **retroviruses** (Section 30-4C), uses an RNA template but is otherwise similar in the reaction it catalyzes to DNA polymerase I.

Variations on the theme of PCR have found numerous applications. For instance, single-stranded DNA (which is required for DNA sequencing; Section 7-2A) can be rapidly generated via **asymmetric PCR**, in which such a small amount of one primer is used that it is exhausted after several PCR cycles. In subsequent cycles, only the strand extended from the other primer, which is present in excess, is synthesized (note that PCR amplification becomes linear rather than geometric after one primer is used up). In cases that primers may anneal to more than one site in the target DNA, **nested primers** can be used to ensure that only the target sequence is amplified. In this technique, PCR amplification is normally carried out using one pair of primers. The products of this amplification are then further amplified through the use of a second pair of primers that anneal to the target DNA within its amplified region. It is highly unlikely that both pairs of primers will incorrectly anneal in a nested fashion to a nontarget DNA, and hence only the target DNA will be amplified.

### b. Neanderthals Are Not Ancestors of Modern Humans

PCR is also largely responsible for the budding science of molecular archeology. For example, PCR-based techniques have been used by Svante Pääbo to determine whether or not Neanderthals form a different species from modern human beings. Neanderthals (*Homo neanderthalensis*; also called Neandertals) are extinct hominids that were about 30% larger than are modern humans, apparently had great muscular strength, and had low foreheads and protruding brows. According to the radiodated fossil record, they became extinct ~28,000 years ago after having inhabited Europe and western Asia for over 300,000 years. During the latter part of this period they coexisted with our direct ancestors (who might well have been responsible for their demise). Thus, an important anthropological issue is whether Neanderthals constituted an ancient race of *Homo sapiens* ancestral to modern humans or were a sep-

arate species. The morphological evidence has been cited as supporting both possibilities. A convincing way to settle this dispute would be by the comparison of the DNA sequences of modern humans with those of Neanderthals.

The DNA was extracted from a 0.4-g sample of a Neanderthal bone, and its mitochondrial DNA (**mtDNA**) was amplified by PCR (mtDNA rather than nuclear DNA was amplified because cells contain numerous mitochondria and hence an mtDNA sequence is 100- to 1000-fold more abundant than is any particular sequence of nuclear DNA). The sequence of the Neanderthal mtDNA was compared to those of 986 modern human lineages of a wide variety of ethnicities and 16 common lineages of chimpanzees (the closest living relatives of modern humans). A phylogenetic tree based on their sequence differences indicates that humans and chimpanzees diverged (had their last common ancestor) about 4 million years ago, humans and Neanderthals diverged around 660,000 years ago, and modern humans diverged from one another about 150,000 years ago. These sequence comparisons indicate that Neanderthals did not contribute significant genetic information to modern humans during their many thousand-year coexistence and hence that *Homo neanderthalensis* and *Homo sapiens* are separate species. This conclusion was confirmed by similar analyses of eleven Neanderthal samples from diverse locations throughout Europe.

### c. DNA Decays Quickly on the Geological Time Scale

There have been reports in the literature of DNAs that were PCR-amplified from fossils that were several million years old and from amber-entombed insects (amber is fossilized tree resin) that were as old as 135 million years (a phenomenon that formed the "basis" for the novel and movie *Jurassic Park*). Yet, over geological time spans, DNA decomposes, mostly through hydrolysis of the sugar-phosphate backbone and oxidative damage to the bases. How old can a fossil become before its DNA has decayed beyond recognition?

The amino acid residues in hydrated proteins racemize at a rate similar to the rate at which DNA decomposes. Since proteins in organisms are far more abundant than are specific DNA sequences, the enantiomeric (D/L) ratios of an amino acid residue can be determined directly (rather than requiring some sort of amplification, as in the case of DNA). The determination, in a variety of archeological specimens whose age could be authenticated, of the enantiomeric ratio of Asp (the fastest racemizing amino acid residue) revealed that DNA sequences can only be retrieved from samples in which the Asp D/L ratio is less than 0.08. These studies indicate that the survival of recognizable DNA sequences is limited to a few thousand years in warm regions such as Egypt and to as much as 100,000 years in cold regions such as Siberia. It therefore appears that the putatively very ancient DNAs, in reality, resulted from the artifactual amplification of contaminating modern DNAs, particularly those from the human operators carrying out the PCR amplifications. Indeed, the DNA in the above Neanderthal fossil had decomposed to the point that it appeared unlikely that its nuclear DNA could have

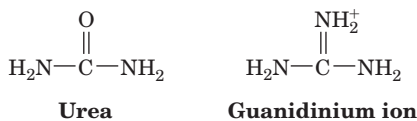
been successfully sequenced, which is why its mtDNA was amplified instead [although with contemporary methods of DNA sequencing (Section 7-2C) this has now become possible].

Despite the foregoing, there is credible evidence that certain bacterial spores can remain viable almost indefinitely. Bacterial spores, which several bacterial groups including bacilli form under adverse conditions, function to permit the bacterium's survival until conditions become favorable for growth. Bacterial spores have thick protective protein coats, their cytoplasm is partially dehydrated and mineralized, and their DNA is specifically stabilized by specialized proteins (Section 29-1Ba). Thus, a bacillus was cultured from a 25- to 40-million-year-old (Myr) amber-entombed bee after the surface of the amber had been chemically sterilized. Similarly, a halophilic (salt-loving) bacillus was cultured from a tiny ( $\sim 9 \mu\text{L}$ ) brine-filled inclusion in a surface-sterilized salt crystal from a 250-Myr salt deposit.

### G. Production of Proteins

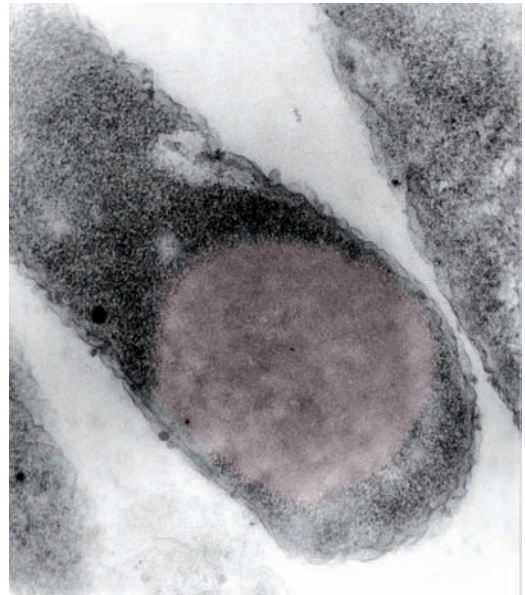
One of the most important uses of recombinant DNA technology is the production of large quantities of scarce and/or novel proteins. This is a relatively straightforward procedure for bacterial proteins: A cloned **structural gene** (a gene that encodes a protein) is inserted into an **expression vector**, a plasmid or virus that contains the properly positioned transcriptional and translational control sequences for the protein's expression. With the use of a relaxed control plasmid and an efficient promoter, the production of a protein of interest may reach 30% of the host bacterium's total cellular protein. Such genetically engineered organisms are called **overproducers**.

Bacterial cells often sequester such large amounts of useless (to the bacterium) protein as insoluble and denatured **inclusion bodies** (Fig. 5-53). A protein extracted from such inclusion bodies must therefore be renatured, usually by dissolving it in a solution of **urea** or **guanidinium ion** (substances that induce proteins to denature)



and then slowly removing the denaturant via a membrane through which the denaturant but not the protein can pass [**dialysis** or **ultrafiltration** (Section 6-3Bc); protein denaturation and renaturation are discussed in Section 9-1A].

A strategy for avoiding the foregoing difficulty is to engineer the gene for the protein of interest so that is preceded with a bacterial **signal sequence** that directs the protein synthesizing machinery of gram-negative bacteria such as *E. coli* to secrete the protein to their **periplasmic space** (the compartment between their plasma membrane and cell wall; signal sequences are discussed in Section 12-4Ba). The signal sequence is then removed by a specific bacterial protease. Secreted proteins, which are relatively few in number, can be released into the medium by the osmotic disruption (Section 6-1B) of the bacterial outer membrane



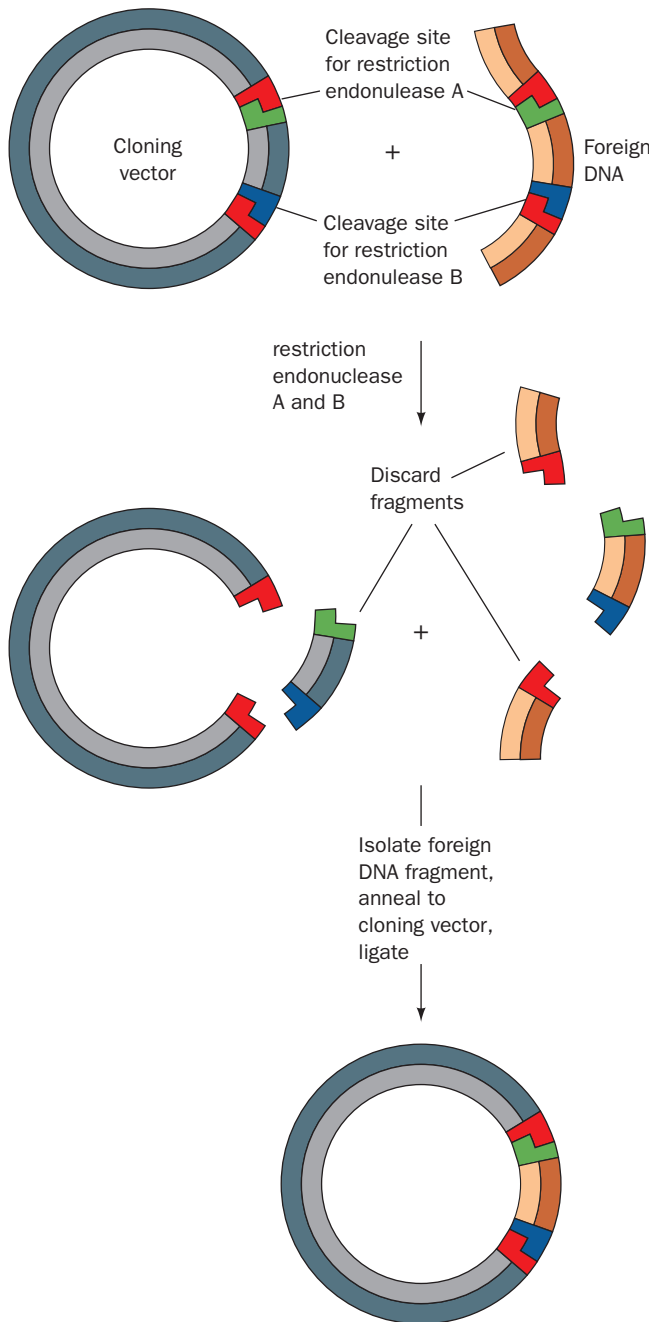
**Figure 5-53** Electron micrograph of an inclusion body of the protein prochymosin in an *E. coli* cell. [Courtesy of Teruhiko Beppu, Nikon University, Japan.]

(Section 1-1B; the bacterial cell wall is porous), so their purification is greatly simplified relative to that of intracellular proteins.

Another problem encountered when producing a foreign protein is that the protein may be toxic to the host cell (e.g., producing a protease may destroy the cell's proteins), thus killing the bacterial culture before sufficient amounts of the protein can be generated. One way to circumvent this problem is to place the gene encoding the toxic protein under the control of an inducible promoter, for example, the *lac* promoter in a plasmid that also includes the gene for the *lac* repressor (Section 5-4Aa). Then, the binding of the *lac* repressor to the *lac* promoter will prevent the expression of the foreign protein in the same way that it prevents the expression of the *lac* operon genes (Fig. 5-25a). However, after the cells have grown to a high concentration, an inducer is added that releases the repressor from the promoter and permits the expression of the foreign protein (Fig. 5-25b). The cells are thereby killed but not before they have produced large amounts of the foreign protein. For the *lac* repressor, the inducer of choice is **isopropylthiogalactoside (IPTG)**; (Section 31-1Aa), a synthetic, nonmetabolizable analog of the *lac* repressor's natural inducer, allolactose.

A problem associated with inserting a DNA segment into a vector, as is indicated in Fig. 5-44, is that any pair of sticky ends that have been made by a given restriction enzyme can be ligated together. Consequently, the products of a ligation reaction will include tandemly (one-after-the-other) linked vectors, inserts, and their various combinations in both linear and circular arrangements. Moreover, in the case of expression systems, 50% of the structural genes that are inserted into circular expression vectors will be installed backward with respect to the expression vector's

transcriptional and translational control sequences and hence will not be properly expressed. The efficiency of the ligation process can be greatly enhanced through the use of **directional cloning** (Fig. 5-54). In this process, two different restriction enzymes are employed to generate two different types of sticky ends on both the insert and the vector. In expression systems, these are arranged such that the structural gene can only be inserted into the expression vector in the correct orientation for expression.



**Figure 5-54 Construction of a recombinant DNA molecule by directional cloning.** Two restriction enzymes, which yield different sticky ends, are used so that the foreign DNA fragment can only be inserted into the cloning vector in one orientation.

### a. Eukaryotic Proteins Can Be Produced in Bacteria and in Eukaryotic Cells

The synthesis of a eukaryotic protein in a prokaryotic host presents several problems not encountered with prokaryotic proteins:

1. The eukaryotic control elements for RNA and protein synthesis are not recognized by bacterial hosts.
2. Bacteria lack the cellular machinery to excise the introns that are present in most eukaryotic transcripts, that is, bacteria cannot carry out gene splicing (Section 5-4Ac).
3. Bacteria lack the enzyme systems to carry out the specific post-translational processing that many eukaryotic proteins require for biological activity (Section 32-5). Most conspicuously, bacteria do not glycosylate proteins (although, in many cases, glycosylation does not seem to affect protein function).
4. Eukaryotic proteins may be preferentially degraded by bacterial proteases (Section 32-6A).

The problem of nonrecognition of eukaryotic control elements can be eliminated by inserting the protein-encoding portion of a eukaryotic gene into a vector containing correctly placed bacterial control elements. The need to excise introns can be circumvented by cloning the cDNA of the protein's mature mRNA. Alternatively, genes encoding small proteins of known sequence can be chemically synthesized (Section 7-6A). Neither of these strategies is universally applicable, however, because few mRNAs are sufficiently abundant to be isolated and the genes encoding many eukaryotic proteins are much larger than can presently be reliably synthesized. Likewise, no general approach has been developed for the post-translational modification of eukaryotic proteins.

The preferential bacterial proteolysis of certain eukaryotic proteins may be prevented by inserting the eukaryotic gene after a bacterial gene such that both have the same reading frame. The resulting **hybrid** or **fusion protein** has an N-terminal polypeptide of bacterial origin that, in some cases, prevents bacterial proteases from recognizing the eukaryotic segment as being foreign. The purification of a fusion protein may be greatly facilitated by the specific binding properties of its N-terminal portion via a process known as **affinity chromatography** (Section 6-3C). Moreover, the formation of a fusion protein may render soluble its otherwise insoluble C-terminal portion. The two polypeptide segments can later be separated by treatment with a protease that specifically cleaves a susceptible site that had been engineered into the boundary between the segments (see below).

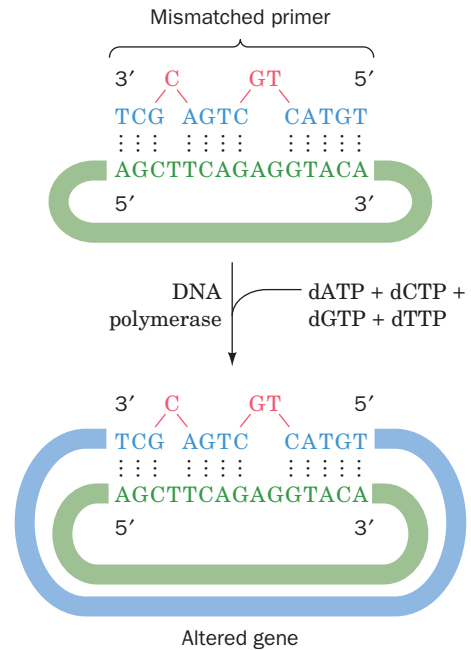
The development of cloning vectors that propagate in eukaryotic hosts, such as yeast or cultured animal cells, has led to the elimination of many of the above problems (although post-translational processing, and in particular glycosylation, may vary among different eukaryotes). Baculovirus-based vectors, which replicate in cultured insect cells, have been particularly successful in this regard. Moreover, **shuttle vectors** are available that can propagate in both yeast and *E. coli* and thus transfer (shuttle) genes between these two types of cells.


### b. Recombinant Protein Production Has Important Practical Consequences

The ability to synthesize a given protein in large quantities has had an enormous medical, agricultural, and industrial impact. Those that are in routine clinical use include **human insulin** (a polypeptide hormone that regulates fuel metabolism and whose administration is required for survival by certain types of diabetics; Section 27-3B), human growth hormone [**somatotropin**, which induces the proliferation of muscle, bone, and cartilage and is used to stimulate growth in children of abnormally short stature (Section 19-1J); before the advent of recombinant DNA techniques, somatotropin was only available in small quantities from the pituitaries of cadavers], **erythropoietin** (a protein growth factor secreted by the kidney that stimulates the production of red blood cells and is used in the treatment of anemia arising from kidney disease), several types of **colony-stimulating factors** (which stimulate the production and activation of white blood cells and are used clinically to counter the white cell-killing effects of chemotherapy and to facilitate bone marrow transplantation), and **tissue-type plasminogen activator (t-PA)**, which is used to promote the dissolution of the blood clots responsible for heart attacks and stroke; Section 35-1F). Synthetic vaccines consisting of harmless but immunogenic components of pathogens, for example, **hepatitis B** vaccine, are eliminating the risks attendant in using killed or attenuated viruses or bacteria in vaccines as well as making possible new strategies of vaccine development. The use of recombinant **blood clotting factors** in treating individuals with the inherited disease **hemophilia** (in which these factors are defective; Section 35-1Da) has replaced the need to extract these scarce proteins from large quantities of human blood and has thereby eliminated the high risk that hemophiliacs previously faced of contracting such blood-borne diseases as hepatitis and AIDS. Bovine somatotropin (**bST**) has long been known to stimulate milk production in dairy cows by ~15%. Its use has been made cost-effective, however, by the advent of recombinant DNA technology since bST could previously only be obtained in small quantities from cow pituitaries. Recombinant porcine somatotropin (**pST**), which is administered to growing pigs, induces ~15% greater growth on ~20% less feed while producing leaner meat.

### c. Site-Directed and Cassette Mutagenesis Generate Proteins with Specific Sequence Changes

Of equal importance to protein production is the ability to tailor proteins to specific applications by altering their amino acid sequences at specific sites. This is frequently done via a method pioneered by Michael Smith known as **site-directed mutagenesis**. In this technique, an oligonucleotide containing a short gene segment with the desired altered base sequence corresponding to the new amino acid sequence (and synthesized by techniques discussed in Section 7-6Aa) is used as a primer in the DNA polymerase I-mediated replication of the gene of interest. Such a primer can be made to hybridize to the corresponding wild-type sequence if there are only a few mismatched base



**Figure 5-55 Site-directed mutagenesis.** A chemically synthesized oligonucleotide incorporating the desired base changes is hybridized to the DNA encoding the gene to be altered (*green strand*). The mismatched primer is then extended by DNA polymerase I, thereby generating the mutated gene (*blue strand*). The mutated gene can subsequently be inserted into a suitable host organism so as to yield the mutant DNA, or its corresponding RNA, in quantity, produce a specifically altered protein, and/or generate a mutant organism.  See the Animated Figures

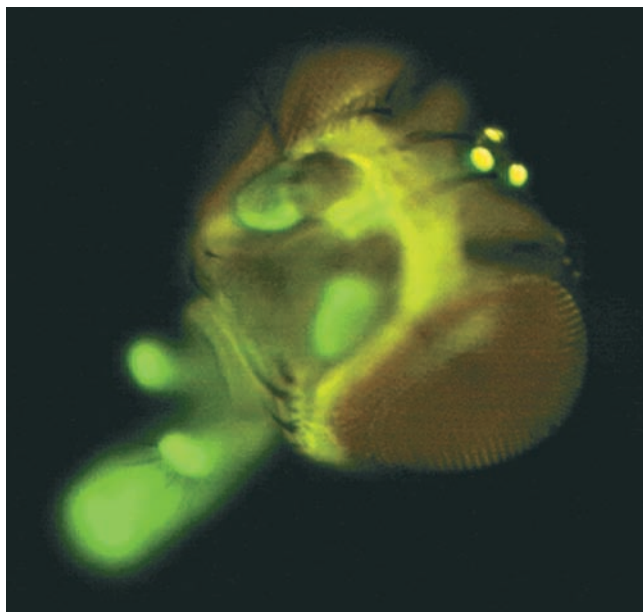
pairs, and its extension, by DNA polymerase I, yields the desired altered gene (Fig. 5-55). The altered gene can then be inserted in a suitable organism via techniques discussed in Section 5-5C and grown (cloned) in quantity. Similarly, PCR may be used as a vehicle for site-directed mutagenesis simply by using a mutagenized primer in amplifying a gene of interest so that the resulting DNA contains the altered sequence.

Using site-directed mutagenesis, the development of a variant form of the bacterial protease **subtilisin** (Section 15-3Bb) in which Met 222 has been changed to Ala (Met 222 → Ala or M222A) has permitted its use in laundry detergent that contains bleach (which largely inactivates wild-type subtilisin by oxidizing Met 222). **Monoclonal antibodies** (a single species of antibody produced by a clone of an antibody-producing cell; Sections 6-1Da and 35-2Bd) can be targeted against specific proteins and hence are used as antitumor agents. However, since monoclonal antibodies, as usually made, are mouse proteins, they are ineffective as therapeutic agents in humans because humans mount an immune response against mouse proteins. This difficulty has been rectified by “humanizing” monoclonal antibodies by replacing their mouse-specific sequences with those of humans (which the human immune system ignores) through site-directed mutagenesis. Thus the monoclonal antibody known as **trastuzumab** (trade name **Herceptin**), which binds specifically to the growth factor receptor **HER2** that is overexpressed in ~25% of breast

cancers, blocks HER2's growth-signaling activity, thereby causing the tumor to stop growing or even regress.

In an alternative mutagenesis technique called **cassette mutagenesis**, complementary oligonucleotides containing the mutation(s) of interest are chemically synthesized (Section 7-6Aa) and annealed to create a duplex "cassette." The cassette is then ligated into the target gene, which must therefore contain an appropriately placed unique restriction site (which can be introduced through site-directed mutagenesis; the cassette must, of course, have the corresponding sticky ends) or, if the cassette is to replace an existing segment of the target gene, two possibly different restriction sites flanking the replaceable segment. Cassette mutagenesis is particularly useful for the insertion of short peptide sequences into a protein of interest (e.g., for the introduction of a proteolytic target site for the cleavage of a fusion protein), when a specific region of the protein is to be subjected to extensive and/or repeated mutagenesis, and for the generation of proteins containing all possible sequences in a short segment (by synthesizing a mixture of cassettes containing all possible variants of the corresponding codons; Section 7-6C).

We will see numerous instances throughout this textbook of protein function being mutagenically character-

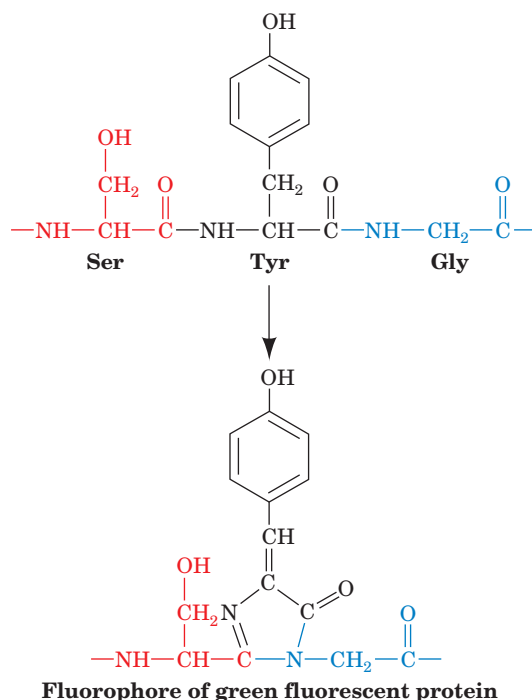


**Figure 5-56 Use of green fluorescent protein (GFP) as a reporter gene.** The gene for GFP was placed under the control of the *Drosophila per* gene promoter and transformed into *Drosophila*. The *per* gene encodes a so-called clock protein that is involved in controlling the fruit fly's circadian (daily) rhythm. The intensity of the green fluorescence of the isolated fly head seen here, which also occurs in other body parts, follows a daily oscillation that can be reset by light. These observations indicate that individual cells in *Drosophila* have photoreceptors and suggest that each of these cells possesses an independent clock. Evidently the head, which was previously thought to be the fly's master oscillator, does not coordinate all of its rhythms. [Courtesy of Steve A. Kay, The Scripps Research Institute, La Jolla, California.]

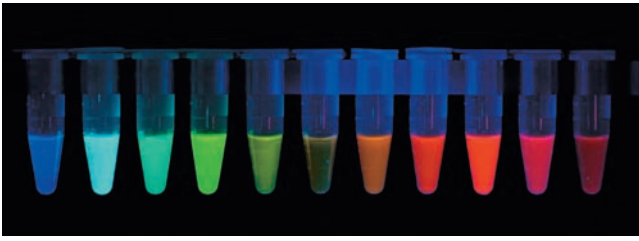
ized through the replacement of a specific residue(s) or a polypeptide segment suspected of having an important mechanistic or structural role. Indeed, mutagenesis has become an indispensable tool in the practice of enzymology.

#### d. Reporter Genes Can Be Used to Monitor Transcriptional Activity

The rate at which a structural gene is expressed depends on its upstream control sequences. Consequently, the rate of expression of a gene can be monitored by replacing its protein-encoding portion with or fusing it in frame to a gene expressing a protein whose presence can be easily determined. An already familiar example of such a **reporter gene** is the *lacZ* gene in the presence of X-gal (Section 5-5Ca) because its level of expression is readily quantitated by the intensity of the blue color that is generated. Although numerous reporter genes have been developed, that which has gained the greatest use encodes **green fluorescent protein (GFP)**. GFP, a product of the bioluminescent jellyfish *Aequorea victoria*, fluoresces with a peak wavelength of 508 nm (green light) when irradiated by UV or blue light (optimally 400 nm). This nontoxic protein, whose use was pioneered by Osamu Shimomura and Martin Chalfie, is intrinsically fluorescent; its light-emitting group is the product of the spontaneous cyclization and oxidation by O<sub>2</sub> of three consecutive residues, Ser-Tyr-Gly, to yield a conjugated system of double bonds that gives the protein its fluorescent properties.



Hence GFP requires no substrate or small molecule cofactor to fluoresce as do other highly fluorescent proteins. Its presence can therefore be monitored through the use of UV light or a fluorometer, and its cellular location can be determined through fluorescence microscopy (Fig. 5-56). Consequently, when the GFP gene is placed under control of the gene expressing a particular protein (GFP's fluorescence

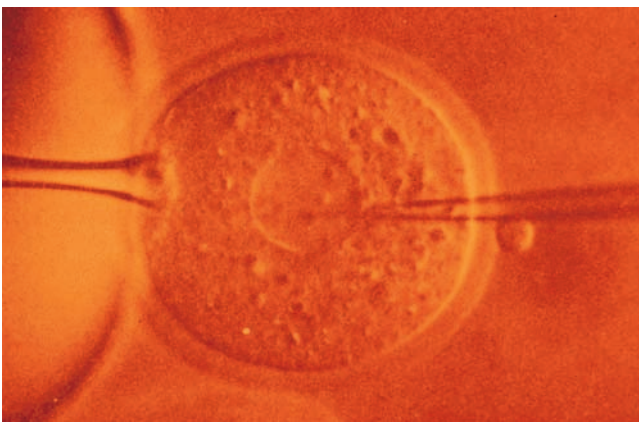


**Figure 5-57** Fluorescence of solutions of GFP and ten of its variants. Note that each of these proteins fluoresces with a different color. [Courtesy of Roger Tsien, University of California at San Diego.]

is unaffected by the formation of a fusion protein), the protein's expressional activity can be readily determined. In fact, a number of GFP variants with distinct sets of excitation and emission wavelengths were developed by Roger Tsien through genetic engineering (Fig. 5-57) and hence the expressional activities of several different genes can be monitored simultaneously. Moreover, the development of pH-sensitive GFP variants permits the monitoring of the pH in subcellular compartments.

#### H. Transgenic Organisms and Gene Therapy

For many purposes it is preferable to tailor an intact organism rather than just a protein—true genetic engineering. Multicellular organisms expressing a foreign (from another organism) gene are said to be **transgenic** and their transplanted foreign genes are often referred to as **transgenes**. For the change to be permanent, that is, heritable, a transgene must be stably integrated into the organism's germ cells. For mice, in techniques pioneered by Mario Capecchi, Martin Evans, and Oliver Smithies, this is accomplished by microinjecting (transfecting) DNA encoding the desired altered characteristics into a **pronucleus** of a fertilized ovum (Fig. 5-58; a fertilized ovum contains two pronuclei,



**Figure 5-58** Microinjection of DNA into the pronucleus of a fertilized mouse ovum. The ovum is being held in place by gentle suction from the pipette on the left. [Science Vu/Visuals Unlimited.]

one from the sperm and the other from the ovum, which eventually fuse to form the nucleus of the one-celled embryo) and implanting it into the uterus of a foster mother. The DNA integrates at a random position in the genome of the pronucleus through a poorly understood mechanism. Alternatively, an **embryonic stem cell** (an undifferentiated embryonic cell that can give rise, *in utero*, to an entire organism) may be transfected with an altered gene, which occasionally replaces its normal counterpart via recombination. A normal gene may, in this manner, be “knocked out” (permanently inactivated) by transfection with a defective version of the gene. With either method, mating heterozygotes for the altered gene yields progeny that are homozygotes for the altered gene. The use of transgenic mice, and in particular **knockout mice**, has greatly enhanced our understanding of vertebrate gene expression (Section 34-3).

#### a. Transgenic Organisms Have Many Uses

Procedures are being developed to generate transgenic farm animals such as cows, goats, pigs, and sheep. Animals may thus be induced to grow larger on lesser amounts of feed and/or to be resistant to particular diseases, although this will require a greater understanding of the genes involved than is presently available. An intriguing application of transgenic farm animals is for them to secrete pharmaceutically useful proteins, such as human growth hormone and blood clotting factors, into their milk. Such a transgenic cow, it is expected, will yield several grams of a foreign protein per liter of milk (tens of kilograms per year), which can thereby be produced far more economically than it can by bacteria. A small herd of such “pharm animals” could satisfy the world's need for a particular medicinally useful protein.

The transplantation between humans of organs such as hearts, lungs, livers, and kidneys (a process known as **allo-transplantation**; Greek: *allos*, other) has saved tens of thousands of lives since the late 1960s. However, the demand for transplantable organs has so outstripped the supply that as little as 5% of the organs needed in the United States become available. This organ shortage could be entirely eliminated if organs from human-sized animals such as pigs could be transplanted into humans (a process known as **xenotransplantation**; Greek: *xenos*, strange or foreign). However, the xenotransplantation of a pig organ into a human results in the destruction of that organ in as little as a six minutes through a series of **complement system**-mediated reactions that are triggered by the foreign antigens lining the blood vessels of the xenograft (the complement system constitutes the body's first line of immunological defenses; Section 35-2F). This **hyperacute rejection** occurs because the porcine tissue lacks the human proteins that inhibit the human complement system. However, when the organs from pigs that were made transgenic in these human inhibitory proteins were transplanted into baboons, the hyperacute rejection of these organs did not occur. Thus, although not all the problems of xenotransplantation have been eliminated (baboons with transplanted pig hearts have survived as long as six months), it now seems likely that genetic engineering techniques will eventually



make xenotransplantation a practical alternative to allotransplantation.

Transgenic plants are becoming increasingly available, promising a significant extension of the “green revolution” that has changed the face of agriculture throughout the world since the 1950s. For example, during sporulation, various strains of the soil microbe *Bacillus thuringiensis* (**Bt**) express proteins that specifically bind to the intestinal cells of certain insects in a manner that lyses these cells, thereby killing the insect through starvation and infection. These so-called  **$\delta$ -endotoxins** (also known as **crystal proteins** because Bt spores contain them in microcrystalline form) are innocuous to vertebrates and, hence, Bt spores have been used to control such pests as the **gypsy moth**. Unfortunately, Bt decays after a short time. However, the gene for a  $\delta$ -endotoxin has been cloned into corn, where, for example, it confers protection against the **European corn borer** (a commercially significant pest that, for much of its life cycle, lives inside the corn plant, where it is largely inaccessible to chemical insecticides). The use of such **Bt corn**, which is now widely planted in the United States, has greatly reduced the need for chemical insecticides.  $\delta$ -Endotoxin genes have likewise been successfully cloned into a variety of agriculturally significant plants including potatoes, soybeans, and cotton.

Among the properties of crop plants that have been generated through genetic engineering are increased herbicide resistance (which permits the more selective use of herbicides to control weeds); resistance to viruses, bacteria, and fungi; control of ripening (to permit a crop to be brought to market at the optimum time); altered plant architecture such as height (which can improve crop productivity); increased tolerance to environmental stresses such as cold, heat, drought and salinity; and modified or additional vitamins, starch, proteins, and oils (for improved nutritional properties and the production of sustainable supplies of raw materials). For example, a strain of rice has been developed that contains the foreign genes encoding the enzymes that synthesize  **$\beta$ -carotene** (Section 24-2Ad), an orange pigment that is a precursor of **vitamin A**. This genetically modified rice, which is named golden rice (Fig. 5-59), could help alleviate vitamin A deficiency, which afflicts ~400 million people and causes an estimated 1.5 million deaths and 400,000 cases of irreversible blindness per year, mainly in children. Unfortunately, the irrational fear of genetically modified organisms fostered by certain environmental and antiglobalization organizations has led to a thicket of regulations on transgenic crops that, as yet, has prevented golden rice from being made available to farmers in the poor countries where vitamin A deficiency is widespread.

#### b. Gene Therapy Has Enormous Medical Potential

**Gene therapy**, the transfer of new genetic material to the cells of an individual resulting in therapeutic benefit to that individual, has been under clinical investigation since 1990, when W. French Anderson and Michael Blaese employed this technology with two children (Section 28-4A) in an effort to alleviate their **severe combined immunodeficiency disease (SCID)**; any of several genetic diseases that



**Figure 5-59 Golden rice.** The rice grains on the left are the wild type. The grains on the right have been engineered to synthesize  $\beta$ -carotene, which gives them their yellow color. [Courtesy of Golden Rice Humanitarian Board, <http://www.goldenrice.org>.]

so impair the immune system that a victim must be kept in a sterile environment in order to survive). Around 4000 genetic diseases are presently known and are therefore potential targets of gene therapy.

Hundreds of gene-transfer protocols are currently under development for use in gene therapy. Ideally, these would deliver the gene(s) of interest to only their target cells such that the gene is permanently installed in these cells in a way that it undergoes sustained expression at the proper level and without causing any adverse side effects. Both nonviral and viral vectors are under development. Nonviral vectors deliver naked DNA to target cells, for example, by direct injection and via **liposomes** (membrane enveloped vesicles that are designed to be taken up by specific cells; Section 12-2B). A large variety of viral vectors are under investigation including those whose genetic material is DNA and those in which it is RNA. These viral vectors have been engineered so as to replace the genes encoding essential viral proteins with therapeutic genes. Hence, cells that have been “infected” by these “viruses” contain the therapeutic genes in their chromosomes but they lack the genetic information to replicate the virus.

There are three categories of gene therapy:

1. In the *ex vivo* (out of the body) approach, cells are removed from the body, incubated with a vector, and then returned to the body. This procedure is usually done with bone marrow cells, which are blood cell precursors.

2. In the *in situ* approach, the vector is applied directly to affected tissues. Such methods are being developed, for example, to treat tumors by injecting into the tumor a vector bearing the gene for a toxin or a gene that would make the tumor susceptible to a chemotherapeutic agent or to attack by the immune system; and to treat **cystic fibrosis**, by inhaling an aerosol containing a vector encoding the normal protein. (Cystic fibrosis, one of the most common genetic diseases, is caused by a defect in a protein involved in the secretion of chloride ion in the lungs and other tissues. This causes the secretion of abnormally thick mucus, which

results in recurrent and often damaging lung infections leading to early death.)

3. In the *in vivo* (in the body) approach, the vector would be injected directly into the bloodstream. There are, as yet, no clinical examples of this approach, although vectors to do so must ultimately be developed if gene therapy is to fulfill its promise.

In the first well-documented clinical success of gene therapy, Alain Fischer reported that several infants appear to have been cured, via the *ex vivo* treatment of their bone marrow cells, of a form of SCID called **SCID-X1** (which is caused by a mutation in the gene encoding the  $\gamma$  **cytokine receptor**, a receptor for certain protein growth factors, whose proper function is essential for the differentiation, growth, and survival of the white blood cells known as **T cells**; Section 35-2A). However, because the viral vector containing the  $\gamma$  cytokine receptor gene integrates into the genome at random, the location of the transgene may affect the expression of other genes, triggering cancer (Section 34-4Ca). In fact, two children have developed leukemia (a white blood cell cancer) as a result of gene therapy for SCID-X1.

Recently, several patients with the rare inherited blinding disease **Leber's congenital amaurosis 2 (LCA2)**; which results in degeneration of the retina) have shown improved vision without significant adverse side effects, after the subretinal injection of a viral vector bearing a normal copy of the defective gene responsible for LCA2. This gene, named **RPE65**, encodes an enzyme that helps convert vitamin A to **retinal** (Section 12-3Ab), the light-sensing pigment of the eye. Evidently, steady progress is being made in our understanding of the requirements for the construction and administration of effective gene therapy vectors. It therefore seems likely that, over the next few decades, gene therapy will revolutionize the practice of medicine.

### I. Social, Ethical, and Legal Considerations

In the early 1970s, when strategies for genetic engineering were first being discussed, it was realized that little was known about the safety of the proposed experiments. Certainly it would be foolhardy to attempt experiments such as introducing the gene for **diphtheria toxin** (Section 32-3Ge) into *E. coli* so as to convert this human symbiont into a deadly pathogen. But what biological hazards would result, for example, from cloning tumor virus genes in *E. coli* (a useful technique for analyzing these viruses)? Consequently, in 1975, molecular biologists declared a voluntary moratorium on molecular cloning experiments until these risks could be assessed. There ensued a spirited debate, at first among molecular biologists and later in the public arena, between two camps: those who thought that the enormous potential benefits of recombinant DNA research warranted its continuation once adequate safety precautions had been instituted, and those who felt that its potential dangers were so great that it should not be pursued under any circumstances.

The former viewpoint eventually prevailed with the promulgation, in 1976, of a set of U.S. government regula-

tions for recombinant DNA research. Experiments that are obviously dangerous were forbidden. In other experiments, the escape of laboratory organisms was to be prevented by both physical and biological containment. By biological containment it is meant that vectors will only be cloned in host organisms with biological defects that prevent their survival outside the laboratory. For example,  $\chi$ 1776, the first approved "safe" strain of *E. coli*, has among its several defects the requirement for diaminopimelic acid, an intermediate in lysine biosynthesis (Section 26-5Ba), which is neither present in human intestines nor commonly available in the environment.

As experience with recombinant DNA research accumulated, it became evident that the foregoing reservations were largely groundless. *No genetically altered organism yet reported has caused an unexpected health hazard*. Indeed, recombinant DNA techniques have, in many cases, eliminated the health hazards of studying dangerous pathogens such as the virus causing **acquired immune deficiency syndrome (AIDS)**. Consequently, since 1979, the regulations governing recombinant DNA research have been gradually relaxed.

There are other social, ethical, and legal considerations that will have to be faced as new genetic engineering techniques become available (Fig. 5-60). Recombinant erythropoietin is now routinely prescribed to treat the effects of certain types of kidney disease. However, should athletes be permitted to use this protein, as many reportedly have, to increase the number of red cells in their blood and hence its oxygen-carrying capacity (a dangerous procedure if uncontrolled since the increased number of cells in the blood can put a great strain on the heart)? Few would dispute the use of gene therapy, if it can be developed, to cure such devastating genetic defects as **sickle-cell anemia** (a painful and debilitating condition caused by deformed red blood cells that often results in early death; Section 10-3B) and **Tay-Sachs disease** (which is caused by the absence of the lysosomal enzyme **hexosaminidase A** and results in progressive neuronal dysfunction that is invariably fatal by around age 3; Section 25-8Ce). If, however, it becomes possible to alter complex (i.e., multigene) traits such as athletic ability or intelligence, which changes would be considered desirable, under what circumstances would they be made, and who would decide whether to make them? Should gene therapy be used on individuals with inherited diseases only to correct defects in their somatic cells or should it also be used to alter genes in their germ cells, which could then be transmitted to succeeding generations? Animals such as sheep, cows, dogs, and mice have already been cloned. Should humans with particularly desirable traits, either naturally occurring or generated through genetic engineering, be cloned? When it becomes easy to determine an individual's genetic makeup, should this information be used, for example, in evaluating applications for educational and employment opportunities, or in assessing a person's eligibility for health insurance (which has recently been made illegal in the United States)? These conundrums have led to the advent of a branch of philosophy named **bioethics** designed to deal with them.



**Figure 5-60** [Drawing by T.A. Bramley, in Andersen, K., Shanmugam, K.T., Lim, S.T., Csonka, L.N., Tait, R., Hennecke, H., Scott, D.B., Hom, S.S.M., Haury, J.F., Valentine, A., and Valentine,

R.C., *Trends Biochem. Sci.* 5, 35 (1980). Copyright © Elsevier Biomedical Press, 1980. Used by permission.]

## CHAPTER SUMMARY

**1 Nucleotides and Nucleic Acids** A nucleotide consists of either a ribose or a 2'-deoxyribose residue whose C1' atom forms a glycosidic bond with a nitrogenous base and whose 3' or 5' position is esterified to a phosphate group. Nucleosides lack the phosphate groups of nucleotides. The nitrogenous bases in the great majority of nucleotides are the purines adenine and guanine and the pyrimidines cytosine and either thymine in DNA or uracil in RNA. Nucleic acids are linear polymers of nucleotides containing either ribose residues in RNA or deoxyribose residues in DNA and whose 3' and 5' positions are linked by phosphate groups. In double helical DNAs and RNAs, the base compositions obey Chargaff's rules:  $A = T(U)$  and  $G = C$ . RNA, but not DNA, is susceptible to base-catalyzed hydrolysis.

**2 DNA Is the Carrier of Genetic Information** Extracts of virulent S-type pneumococci transform nonpathogenic R-type pneumococci to the S form. The transforming principle is DNA. Similarly, radioactive labeling has demonstrated that the genetically active substance of bacteriophage T2 is its DNA. The viral capsid serves only to protect its enclosed DNA and to inject it into the bacterial host. This establishes that DNA is the hereditary molecule.

**3 Double Helical DNA** B-DNA consists of a right-handed double helix of antiparallel sugar-phosphate chains with  $\sim 10$  bp per turn of  $34 \text{ \AA}$  and with its bases nearly perpendicular to the helix axis. Bases on opposite strands hydrogen bond in a geometrically complementary manner to form  $A \cdot T$  and  $G \cdot C$  Watson-Crick base pairs. DNA replicates in a semi-conservative manner, as has been demonstrated by the Meselson-Stahl experiment. When heated past its melting temperature,  $T_m$ , DNA denatures and undergoes strand separation. This process may be monitored by the hyperchromism of the DNA's UV spectrum. Denatured DNA can be renatured by maintaining it at  $\sim 25^\circ\text{C}$  below its  $T_m$ . DNA occurs in nature as molecules of enormous lengths which, because they are also quite stiff, are easily mechanically cleaved by laboratory manipulations.

**4 Gene Expression and Replication: An Overview** Genes are expressed according to the central dogma of molecular biology: DNA directs its own replication and its transcription to yield RNA, which, in turn, directs its translation to form proteins. RNA is synthesized from ribonucleoside triphosphates on DNA templates by RNA polymerase, a process in which the DNA template strand is read in its 3' to 5' direction and the

RNA is synthesized in its 5' to 3' direction. The rate at which a particular gene is transcribed is governed by control sites, which, for mRNAs, are located upstream of the transcriptional initiation site and can be quite extensive, particularly in eukaryotes. Eukaryotic mRNAs often require substantial post-transcriptional modifications, including gene splicing (the excision of introns and the rejoining of their flanking exons), to become functional.

mRNAs direct the ribosomal synthesis of polypeptides. In this process, ribosomes facilitate the binding of the mRNA's codons to the anticodons of tRNAs bearing their cognate amino acids, and the ribosomes then catalyze the formation of peptide bonds between successive amino acids. The correspondence between codons and the amino acid carried by the tRNAs that bind to them is called the genetic code. Enzymes known as aminoacyl-tRNA synthetases covalently link their corresponding tRNAs to their cognate amino acids. The selection of the correct initiation site on the mRNA also sets the reading frame for the polypeptide being synthesized. Newly synthesized proteins often require post-translational modifications to be functional, including specific proteolytic cleavages and, in eukaryotes only, glycosylation. The lifetime of a protein in a cell varies from fractions of a minute to days or weeks.

DNA is synthesized from deoxynucleoside triphosphates by DNA polymerase, an enzyme that can only extend existing polynucleotides bound to the template DNA and hence requires a primer. In cells, primers are RNA, which are synthesized on DNA templates by an RNA polymerase. The replication of both strands of duplex DNA takes place at a replication fork. In *E. coli*, duplex DNA replication is carried out by two molecules of DNA polymerase III, one of which synthesizes the leading strand and the other of which synthesizes the lagging strand. The leading strand is synthesized continuously. However, since all DNA polymerases can only extend DNA in its 5' to 3' direction, the lagging strand template must loop around to be read in its 3' to 5' direction, which requires that the lagging strand be synthesized discontinuously. The RNA primers for the lagging strand are synthesized by primase, and once a lagging strand segment has been synthesized, its primer is replaced through the combined actions of DNA polymerase I's 5' → 3' exonuclease and DNA polymerase activities. The single-strand nicks between successive

lagging strand segments are then sealed by DNA ligase. Both DNA polymerase I and DNA polymerase III also have 3' → 5' exonuclease activities that function to proofread the newly synthesized DNA for mispairing errors and excise the mispaired nucleotides.

**5 Molecular Cloning** Molecular cloning techniques have revolutionized the practice of biochemistry. Defined DNA fragments are generated through the use of Type II restriction endonucleases (restriction enzymes), which cleave DNA at specific and usually palindromic sequences of four to six bases. Restriction-fragment length polymorphisms (RFLPs) provide markers for identifying chromosomal differences and hence are useful in identity tests and in establishing familial relationships. A DNA fragment may be produced in large quantities by inserting it, using recombinant DNA techniques, into a suitable cloning vector. These may be genetically engineered plasmids, viruses, cosmids, yeast artificial chromosomes (YACs), or bacterial artificial chromosomes (BACs). The DNA to be cloned is usually obtained as a restriction fragment so that it can be specifically ligated into a corresponding restriction cut in the cloning vector. Gene splicing may also occur through the generation of complementary homopolymer tails on the DNA fragment and the cloning vector or through the use of synthetic palindromic linkers containing restriction sequences. The introduction of a recombinant cloning vector into a suitable host organism permits the foreign DNA segment to be produced in nearly unlimited quantities. Those cells that have been properly transformed by a vector can be chosen through the use of selectable markers and chromogenic substrates. Specific base sequences may be detected in DNA by Southern blotting and in RNA by the similar Northern blotting. A particular gene may be isolated through the screening of a genomic library of the organism producing the gene. The polymerase chain reaction (PCR) is a particularly fast and convenient method of identifying and obtaining specific sequences of DNA. Genetic engineering techniques may be used to produce otherwise scarce or specifically altered proteins in large quantities and to monitor gene expression through the use of reporter genes such as green fluorescent protein. They are also used to produce transgenic plants and animals and in gene therapy. The development of recombinant DNA techniques has generated numerous social, ethical, and legal issues whose resolution will govern how biotechnology is used.

## REFERENCES

### Useful Websites

REBASE. The restriction enzyme database. <http://rebase.neb.com>

### The Role of DNA

Avery, O.T., MacLeod, C.M., and McCarty, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types, *J. Exp. Med.* **79**, 137–158 (1944). [The milestone report identifying transforming principle as DNA.]

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

McCarty, M., *The Transforming Principle*, Norton (1985). [A chronicle of the discovery that genes are DNA.]

Palmiter, R.D., Brinster, R.L., Hammer, R.E., Trumbauer, M.E., Rosenfeld, M.G., Birnberg, N.C., and Evans, R.M., Dramatic

growth of mice that develop from eggs microinjected with metallothionein–growth hormone fusion genes, *Nature* **300**, 611–615 (1982).

Stent, G.S., Prematurity and uniqueness in scientific discovery, *Sci. Am.* **227**(6), 84–93 (1972). [A fascinating philosophical discourse on what it means for discoveries such as Avery's to be “ahead of their time” and on the nature of creativity in science.]

### Structure and Properties of B-DNA

Bloomfield, V.A., Crothers, D.M., and Tinoco, I., Jr., *Nucleic Acids. Structures, Properties, and Functions*, University Science Books (2000).

Crick, F., *What Mad Pursuit*, Basic Books (1988). [A scientific autobiography.]

- Judson, H.F., *The Eighth Day of Creation*, Part I, Simon & Schuster (1979). [A fascinating narrative on the discovery of the DNA double helix.]
- Manchester, K.L., Historical opinion: Erwin Chargaff and his 'rules' for the base composition of DNA: why did he fail to see the possibility of complementarity? *Trends Biochem. Sci.* **33**, 65–70 (2008).
- Meselson, M. and Stahl, F.W., The replication of DNA in *Escherichia coli*, *Proc. Natl. Acad. Sci.* **44**, 671–682 (1958). [The classic paper establishing the semiconservative nature of DNA replication.]
- Saenger, W., *Principles of Nucleic Acid Structure*, Springer-Verlag (1984).
- Sayre, A., *Rosalind Franklin and DNA*, Norton (1975) [A biography which argues that Rosalind Franklin, who died in 1958, deserves far more credit than is usually accorded her for the discovery of the structure of DNA.]; and Piper, A., Light on a dark lady, *Trends Biochem. Sci.* **23**, 151–154 (1998). [A biographical memoir on Rosalind Franklin.]
- Schlenk, F., Early nucleic acid chemistry, *Trends Biochem. Sci.* **13**, 67–68 (1988).
- Voet, D. and Rich, A., The crystal structures of purines, pyrimidines and their intermolecular structures, *Prog. Nucleic Acid Res. Mol. Biol.* **10**, 183–265 (1970).
- Watson, J.D., *The Double Helix. A Personal Account of the Discovery of the Structure of DNA*, Simon & Schuster (1968). [A provocative autobiographical narrative.]
- Watson, J.D. and Crick, F.H.C., Molecular structure of nucleic acids, *Nature* **171**, 737–738 (1953); and Genetical implications of the structure of deoxyribonucleic acid, *Nature* **171**, 964–967 (1953). [The seminal papers that are widely held to mark the origin of modern molecular biology.]
- Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R.E., Crystal structure analysis of a complete turn of B-DNA, *Nature* **287**, 755–758 (1980). [The first X-ray crystal structure of a segment of B-DNA, which largely confirmed the less reliable fiber diffraction-based structure of B-DNA proposed by Watson and Crick.]
- Zimm, B.H., One chromosome: one DNA molecule, *Trends Biochem. Sci.* **24**, 121–123 (1999). [A scientific reminiscence on how it was established that chromosomes each contain only one piece of DNA.]
- Molecular Cloning**
- Bainbridge, J.W.B., et al., Effect of gene therapy on visual function in Leber's congenital amaurosis; and Maguire, A.M., et al., Safety and efficacy of gene transfer for Leber's congenital amaurosis, *New Eng. J. Med.* **358**, 2231–2239; and 2240–2248 (2008).
- Brown, T.A., *Gene Cloning and DNA Analysis, An Introduction* (5th ed.), Blackwell Publishing (2006).
- Cavazzana-Calvo, M., et al., Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease, *Science* **288**, 669–672 (2000).
- Clark, D.P. and Pazdernik, N.J., *Biotechnology. Applying the Genetic Revolution*, Academic Press (2009).
- Cooper, A. and Wayne, R., New uses for old DNA, *Curr. Opin. Biotech.* **9**, 49–53 (1998). [Reviews the successes and pitfalls of studying ancient DNAs.]
- Cooper, D.K.C., Gollackner, B., and Sachs, D.H., Will the pig solve the transplantation backlog? *Annu. Rev. Med.* **53**, 133–147 (2002).
- Dale, J.W. and von Schantz, M., *From Genes to Genomes. Concepts and Application of DNA Technology* (2nd ed.), Wiley (2007).
- Foster, E.A., Jobling, M.A., Taylor, P.G., Donnelly, P., de Knijff, P., Mieremet, R., Zerjal, T., and Tyler-Smith, C., Jefferson fathered slave's last child, *Nature* **396**, 27–28 (1998).
- Gilbert, S.F., Tyler, A.L., and Zackin, E.J., *Bioethics and the New Embryology: Springboards for Debate*, Sinauer (2005). [A balanced discussion of many of the most contentious bioethical issues of our time.]
- Green, R.E., et al., Analysis of one million base pairs of Neanderthal DNA, *Nature* **444**, 330–336 (2006); Noonan, J.P., et al., Sequencing and analysis of Neanderthal genomic DNA, *Science* **314**, 1113–1118 (2006); and Green, R.E., et al., A complete Neanderthal mitochondrial genome sequence determined by high-throughput sequencing, *Cell* **134**, 416–426 (2008).
- Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H., Stoneking, M., and Pääbo, S., Neanderthal DNA sequences and the origin of modern humans; and Lindahl, T., Facts and artifacts of ancient DNA, *Cell* **90**, 19–30; and 1–3 (1997); and Ovchinnikov, I.V., Götherström, A., Romanova, G.P., Kharitonov, V.M., Lidén, K., and Goodwin, W., Molecular analysis of Neanderthal DNA from the northern Caucasus, *Nature* **404**, 490–493 (2000).
- Lodge, J., Lund, P., and Minchlin, S., *Gene Cloning: Principles and Applications*, Taylor and Francis (2007).
- Mullis, K.B., The unusual origin of the polymerase chain reaction. *Sci. Am.* **262**(4): 56–65 (1990).
- Nicholl, D.S.T., *An Introduction to Genetic Engineering* (2nd ed.), Cambridge University Press (2003).
- Primrose, S.B. and Twyman, R.M., *Principles of Gene Manipulation and Genomics* (7th ed.), Blackwell Publishing (2006).
- Sambrook, J. and Russel, D.W., *Molecular Cloning* (3rd ed.), Cold Spring Harbor Laboratory (2001). [A three-volume collection of laboratory protocols with accompanying background explanations.]
- Tsien, R.Y., The green fluorescent protein, *Annu. Rev. Biochem.* **67**, 509–544 (1998).
- Verma, I.M. and Weitzman, M.D., Gene therapy: Twenty-first century medicine, *Annu. Rev. Biochem.* **74**, 711–738 (2005).
- Vreeland, R.H., Rosenzweig, W.D., and Powers, D.W., Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal, *Nature* **407**, 897–900 (2000).
- Watson, J.D., Meyers, R.M., Caudy, A.A., and Witkowski, J.A., *Recombinant DNA. Genes and Genomes—A Short Course* (3rd ed.), Freeman (2007). [An exposition of the methods, findings, and results of recombinant DNA technology and research.]

## BIOINFORMATICS EXERCISES

Bioinformatics Exercises are available at [www.wiley.com/college/voet](http://www.wiley.com/college/voet).

### Chapter 5

#### Electrophoresis

- One-Dimensional Electrophoresis.** Perform an SDS-PAGE electrophoresis simulation with known and unknown proteins.
- Two-Dimensional Electrophoresis.** Explore the predicted and observed electrophoretic parameters ( $pI$ , molecular mass, and fragmentation pattern) for a known protein.

## PROBLEMS

1. The base sequence of one of the strands of a 20-bp duplex DNA is

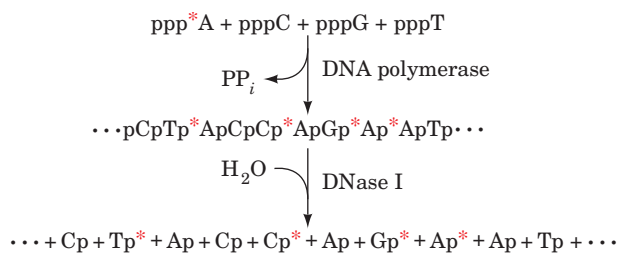


What is the base sequence of its complementary strand?

2. Non-Watson–Crick base pairs are of biological importance. For example: (a) **Hypoxanthine** (6-oxopurine) is often one of the bases of the anticodon of tRNA. With what base on mRNA is hypoxanthine likely to pair? Draw the structure of this base pair. (b) The third position of the codon–anticodon interaction between tRNA and mRNA is often a G · U base pair. Draw a plausible structure for such a base pair. (c) Many species of tRNA contain a hydrogen bonded U · A · U assembly. Draw two plausible structures for this assembly in which each U forms at least two hydrogen bonds with the A. (d) Mutations may arise during DNA replication when mispairing occurs as a result of the transient formation of a rare tautomeric form of a base. Draw the structure of a base pair with proper Watson–Crick geometry that contains a tautomeric form of adenine. What base sequence change would be caused by such mispairing?

3. (a) What is the molecular mass and contour length of a segment of B-DNA that specifies a 40-kD protein? (b) How many helical turns does this DNA have and what is its axial ratio (length to width ratio)?

\*4. The antiparallel orientation of complementary strands in duplex DNA was elegantly demonstrated in 1960 by Arthur Kornberg by **nearest-neighbor analysis**. In this technique, DNA is synthesized by DNA polymerase I from one deoxynucleoside triphosphate whose  $\alpha$ -phosphate group is radioactively labeled with  $^{32}\text{P}$  and three unlabeled deoxynucleoside triphosphates. The resulting product is treated with an enzyme, DNase I, that catalyzes the hydrolysis of the phosphodiester bonds on the 3' sides of all deoxynucleotides.



In this example, the relative frequencies of occurrence of ApA, CpA, GpA, and TpA in the DNA can be determined by measuring the relative amounts of Ap\*, Cp\*, Gp\*, and Tp\*, respectively, in the product (where p\* represents a  $^{32}\text{P}$ -labeled phosphate group). The relative frequencies with which the other 12 dinucleotides occur may likewise be determined by labeling, in turn, the other 3 nucleoside triphosphates in the above reactions. There are equivalencies between the amounts of certain pairs of dinucleotides. However, the identities of these equivalencies depend on whether the DNA consists of parallel or antiparallel strands. What are these equivalencies in both cases?

5. What would be the effect of the following treatments on the melting curve of B-DNA dissolved in 0.5M NaCl solution? Explain. (a) Decreasing the NaCl concentration. (b) Squirting the DNA solution, at high pressure, through a narrow orifice.

(c) Bringing the solution to 0.1M adenine. (d) Heating the solution to 25°C above the DNA's melting point and then rapidly cooling it to 35°C below the DNA's melting point.

6. What is the mechanism of alkaline denaturation of duplex DNA? [Hint: Some of the bases are relatively acidic.]

7. The following duplex DNA is transcribed from right to left as is printed here.



(a) Identify the template strand. (b) What is the amino acid sequence of the polypeptide that this DNA sequence encodes? Assume that translation starts at the first initiation codon. (c) Why doesn't the UGA sequence in the mRNA transcript cause transcription to terminate?

8. After undergoing splicing, a mature mRNA has the following sequence, where the vertical line indicates the position of the splice junction (the nucleotides from between which an intron had been removed).



(a) What is the sequence of the polypeptide specified by this mRNA? Assume that translation starts at the first initiation codon. (b) What would the polypeptide sequence be if the splicing system had erroneously deleted the GU on the 3' side of the splice junction? (c) What would the polypeptide sequence be if the splicing system had erroneously failed to excise a G at the splice junction? (d) Is there any relationship between the polypeptides specified in b and c and, if so, why?

9. Explain why the charging of a tRNA with the correct amino acid is equally as important for accurate translation as is the correct recognition of a codon by its corresponding aminoacyl-tRNA.

10. Describe how to select recombinant clones if a foreign DNA is inserted into the polylinker site of pUC18 and then introduced into *E. coli* cells.

11. Describe the possible outcome of a PCR experiment in which (a) one of the primers is inadvertently omitted from the reaction mixture; (b) one of the primers is complementary to several sites in the starting DNA sample; (c) there is a single-stranded break in the target DNA sequence, which is present in only one copy in the starting sample; (d) there is a double-stranded break in the target DNA sequence, which is present in only one copy in the starting sample.

12. In DNA replication, the leading strand and lagging strand are so named because any particular portion of the lagging strand is always synthesized after the corresponding portion of the leading strand has been synthesized. Explain why this must be the case.

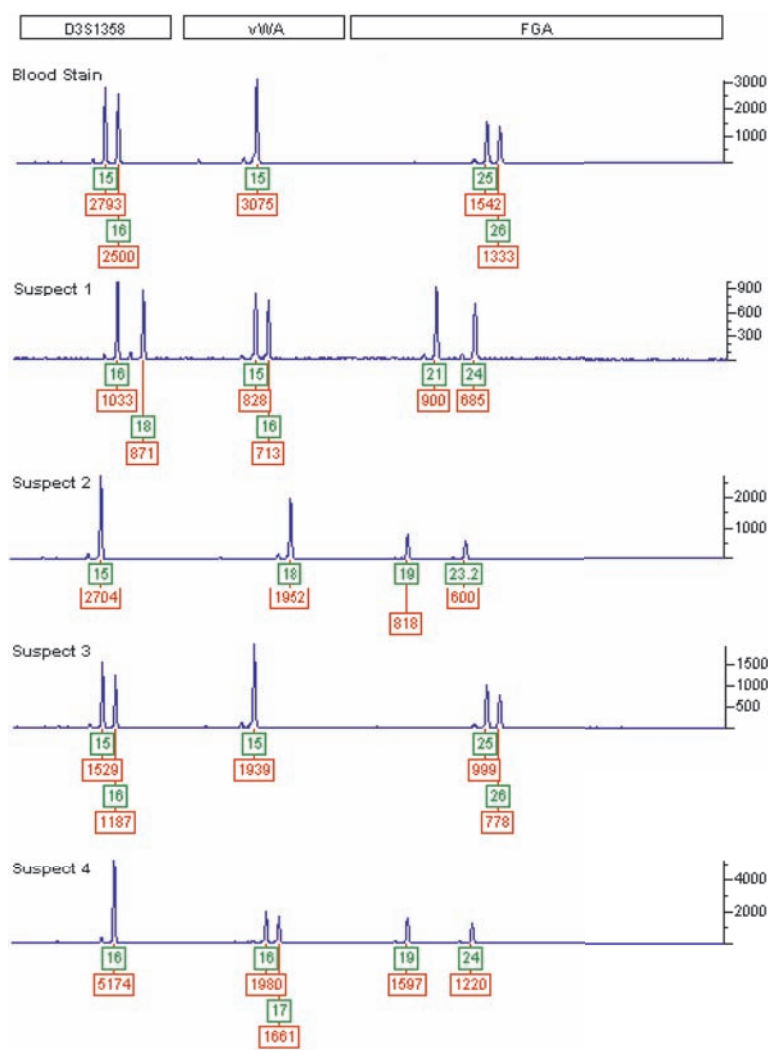
13. SV40 DNA is a circular molecule of 5243 bp that is 40% G + C. In the absence of sequence information, how many restriction cuts would TaqI, EcoRII, PstI, and HaeII be expected to make, on average, in SV40 DNA?

14. Which of the restriction endonucleases listed in Table 5-4 produce blunt ends? Which sets of them are **isoschizomers** (enzymes

that have the same recognition sequence but do not necessarily cleave at the same sites; Greek: *isos*, equal + *schizein*, to cut); which of them are **isocaudamers** (enzymes that produce identical sticky ends; Latin: *cauda*, tail)?

**15.** The plasmid pBR322 contains the *amp<sup>R</sup>* and *tet<sup>R</sup>* genes, which respectively confer resistance to the antibiotics ampicillin and **tetracycline** (Section 32-3Gd). The *tet<sup>R</sup>* gene contains a cleavage site for the restriction enzyme Sall, the only such site in the entire plasmid. Describe how one can select for *E. coli* that had been transformed by pBR322 that contains a foreign DNA insert in its Sall site.

**16.** A blood stain from a crime scene and blood samples from four suspects were analyzed by PCR using fluorescent primers associated with three STR loci (positions): D3S1358, vWA, and FGA. The resulting electrophoretograms are shown below. The numbers beneath each peak identify the allele (*green box*) and the height of the peak in relative fluorescence units (*red box*). (a) Since everyone has two copies of each chromosome and therefore two copies of each gene, what accounts for the appearance of only one allele at some loci? (b) Which suspect, if any, is a possible source of the blood? (c) Could the suspect be identified using just one of the three STR loci? (d) What can you conclude about the amount of DNA obtained from Suspect 1 compared to Suspect 4?



**17.** How many yeast DNA fragments of average length 5 kb must be cloned in order to be 90%, 99%, and 99.9% certain that a genomic library contains a particular segment? The yeast genome consists of 12,100 kb.

**18.** Many of the routine operations in genetic engineering are carried out using commercially available “kits.” Genbux Inc., a prospective manufacturer of such kits, has asked your advice on the feasibility of supplying a kit of intact  $\lambda$  phage cloning vectors with the nonessential central section of their DNA already removed. Presumably a “gene jockey” could then grow the required amount of phage, isolate its DNA, and restriction cleave it without having to go to the effort of separating out the central section. What advice would you give the company?

**19.** Indicate the sequences of the two 10-residue primers that could be used to amplify the central 40-nucleotide region of the following 50-nucleotide single-stranded DNA by PCR.

5'-AGCTGGACCACTGATCATTGACTGCTAGCGTCA  
GTCTAGTAGACTGACG-3'

**20.** A protein segment of sequence -Phe-Cys-Gly-Val-Leu-His-Lys-Met-Glu-Thr- is encoded by the following DNA segment:

5'-TTCTGCGGAGTCCTACACAAGATGGAGACA-3'

Design an 18-base oligonucleotide that could be used to change the protein's Leu-His segment to Ile-Pro via site-directed mutagenesis.