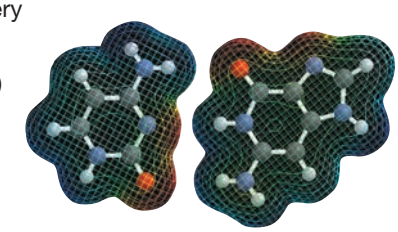
CHAPTER

25

Nucleic Acids and Protein Synthesis

Chemistry is the central science because it is involved in every aspect of life. Much of what we have learned about organic chemistry thus far is related to how things work chemically, how diseases can be treated at the molecular level with small molecules, and how we can create new compounds and materials that improve our daily lives. One of the most interesting of the many applications of organic chemistry is its ability to solve critical challenges in identification through DNA matching. Studying the structure of genes and DNA, scientists can determine genetic relationships between different species (and hence the course of evolution) or between people. They can also identify the remains of individuals through DNA matching, a valuable tool if there are no other physical means to make such an identification. In fact, DNA, the genetic material, is the key to all this work. DNA is the chemical fingerprint in every tissue of every individual. With the use of chemistry involving fluorescent dyes, radioactive isotopes, enzymes, gel electrophoresis, and a process called the polymerase chain reaction (PCR) that earned its inventor the 1993 Nobel Prize in Chemistry, it is now easy to synthesize millions of copies of DNA from a single molecule of DNA, as well as to sequence it rapidly and conveniently. To understand just how this amazing process works, we need to understand this final class of biomolecules in much more detail.



A guanine-cytosine base pair

PHOTO CREDIT: PASIEKA/Science Photo Library/Getty Images, Inc.

IN THIS CHAPTER WE WILL CONSIDER:

- the structures of nucleic acids and methods for their laboratory synthesis
- the primary and secondary structure of DNA
- RNA and its roles in protein synthesis
- methods of DNA sequencing

[WHY DO THESE TOPICS MATTER?] Not only will we show you how the PCR reaction works, but at the end of this chapter we will also show how chemists have developed and designed small molecules that, with hydrogen bonding, have the ability to selectively bind any specific DNA sequence desired. Through this technique, chemists can potentially target a drug molecule selectively to any DNA portion that might be critical to treating disease.

25.1 INTRODUCTION

Deoxyribonucleic acid (DNA) and **ribonucleic acid (RNA)** are molecules that carry genetic information in cells. DNA is the molecular archive of instructions for protein synthesis. RNA molecules transcribe and translate the information from DNA for the mechanics of protein synthesis. The storage of genetic information, its passage from generation to generation, and the use of genetic information to create the working parts of the cell all depend on the molecular structures of DNA and RNA. For these reasons, we shall focus our attention on the structures and properties of these **nucleic acids** and of their components, nucleotides and nucleosides.

DNA is a biological polymer composed of two molecular strands held together by hydrogen bonds. Its overall structure is that of a twisted ladder with a backbone of alternating sugar and phosphate units and rungs made of hydrogen-bonded pairs of heterocyclic amine bases (Fig. 25.1). DNA molecules are very long polymers. If the DNA from a

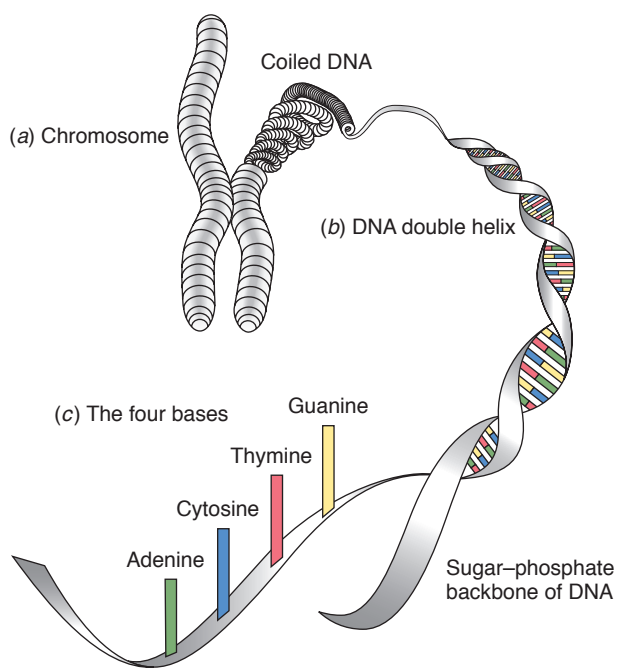


FIGURE 25.1 The basics of genetics. Each cell in the human body (except red blood cells) contains 23 pairs of chromosomes. Chromosomes are inherited: each parent contributes one chromosome per pair to their children. (a) Each chromosome is made up of a tightly coiled strand of DNA. The structure of DNA in its uncoiled state reveals (b) the familiar double-helix shape. If we picture DNA as a twisted ladder, the sides, made of sugar and phosphate molecules, are connected by (c) rungs made of heterocyclic amine bases. DNA has four, and only four, bases—adenine (A), thymine (T), guanine (G), and cytosine (C)—that form interlocking pairs. The order of the bases along the length of the ladder is called the DNA sequence. Within the overall sequence are genes, which encode the structure of proteins. (*Science and Technology Review*, November 1996, "The Human Genome Project," https://www.llnl.gov/str/Ashworth.html. Credit must be given to Linda Ashworth, the University of California, Lawrence Livermore National Laboratory, and the Department of Energy under whose auspices the work was performed, when this information or a reproduction of it is used.)

single human cell were extracted and laid straight end-to-end, it would be roughly a meter long. To package DNA into the microscopic container of a cell's nucleus, however, it is supercoiled and bundled into the 23 pairs of chromosomes with which we are familiar from electron micrographs.

Four types of heterocyclic bases are involved in the rungs of the DNA ladder, and it is the sequence of these bases that carries the information for protein synthesis. Human DNA consists of approximately 3 billion base pairs. In an effort that marks a milestone in the history of science, a working draft of the sequence of the 3 billion base pairs in the human genome was announced in 2000. A final version was announced in 2003, the 50th anniversary of the structure determination of DNA by Watson and Crick.

- Each section of DNA that codes for a given protein is called a **gene**.
- The set of all genetic information coded by DNA in an organism is its **genome**.

There are approximately 30,000–35,000 genes in the human genome. The set of all proteins encoded within the genome of an organism and expressed at any given time is called its **proteome** (Section 24.14). Some scientists estimate there could be up to one million different proteins in the cells of our various tissues—a number much greater than the number of genes in the genome due to gene splicing during protein expression and post-translational protein modification.

Hopes are very high that, having sequenced the human genome, knowledge of it will bring increased identification of genes related to disease states (Fig. 25.2) and study of these genes and the proteins encoded by them will yield a myriad of benefits for human health and longevity. Determining the structure of all of the proteins encoded in the genome, learning their functions, and creating molecular therapeutics based on this rapidly expanding store of knowledge are some of the key research challenges that lie ahead.

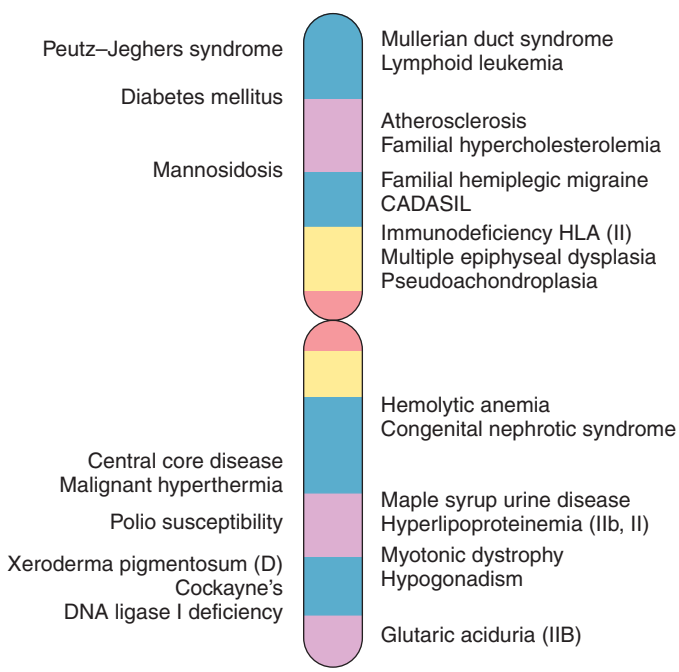


FIGURE 25.2 Schematic map of the location of genes for diseases on chromosome 19. *(From Dept. of Energy Joint Genome Institute Website. (http://www.jgi.doe.gov/whoweare/). Credit to the University of California, Lawrence Livermore National Laboratory, and the Department of Energy under whose auspices the work was performed.)*

Let us begin with a study of the structures of nucleic acids. Each of their monomer units contains a cyclic amine base, a carbohydrate group, and a phosphate ester.

25.2 NUCLEOTIDES AND NUCLEOSIDES

Mild degradations of nucleic acids yield monomeric units called **nucleotides**. A general formula for a nucleotide and the specific structure of one called adenylic acid are shown in Fig. 25.3.

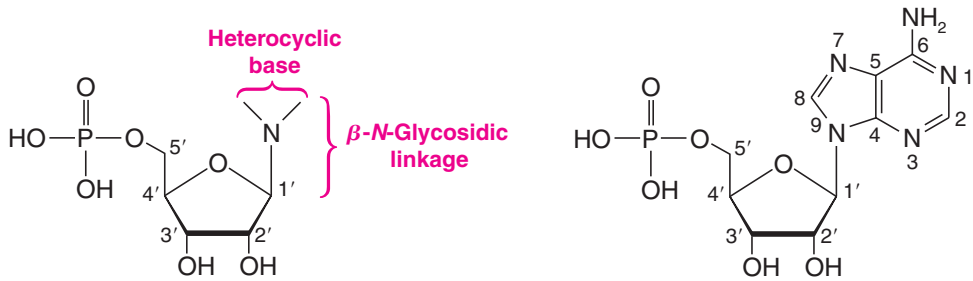


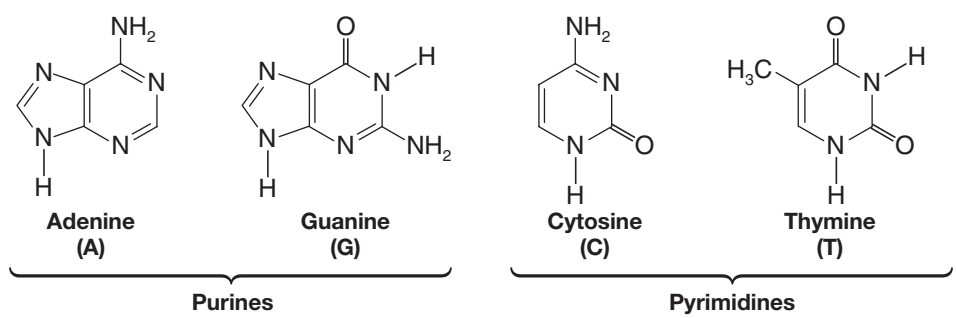
FIGURE 25.3 (*a*) General structure of a nucleotide obtained from RNA. The heterocyclic base is a purine or pyrimidine. *In nucleotides obtained from DNA, the sugar component is 2′-deoxy-D-ribose; that is, the —OH at position 2′ is replaced by —H.* The phosphate group of the nucleotide is shown attached at C5′; it may instead be attached at C3′. In DNA and RNA a phosphodiester linkage joins C5′ of one nucleotide to C3′ of another. The heterocyclic base is always attached through a β-*N*-glycosidic linkage at C1′. (*b*) Adenylic acid, a typical nucleotide.

Complete hydrolysis of a nucleotide furnishes:

1. A heterocyclic base from either the purine or pyrimidine family.
2. A five-carbon monosaccharide that is either D-ribose or 2-deoxy-D-ribose.
3. A phosphate ion.

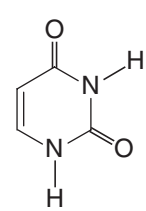
The central portion of the nucleotide is the monosaccharide, and it is always present as a five-membered ring, that is, as a furanoside. The heterocyclic base of a nucleotide is attached through an *N*-glycosidic linkage to C1′ of the ribose or deoxyribose unit, and this linkage is always β. The phosphate group of a nucleotide is present as a phosphate ester and may be attached at C5′ or C3′. (In nucleotides, the carbon atoms of the monosaccharide portion are designated with primed numbers, i.e., 1′, 2′, 3′, etc.)

Removal of the phosphate group of a nucleotide converts it to a compound known as a **nucleoside** (Section 22.15A). The nucleosides that can be obtained from DNA all contain 2-deoxy-D-ribose as their sugar component and one of four heterocyclic bases: adenine, guanine, cytosine, or thymine:



The nucleosides obtained from RNA contain D-ribose as their sugar component and adenine, guanine, cytosine, or uracil as their heterocyclic base.

Uracil replaces thymine in an RNA nucleoside (or nucleotide). Some nucleosides obtained from specialized forms of RNA may also contain other, but similar, purines and pyrimidines.



The heterocyclic bases obtained from nucleosides are capable of existing in more than one tautomeric form. The forms that we have shown are the predominant forms that the bases assume when they are present in nucleic acids.

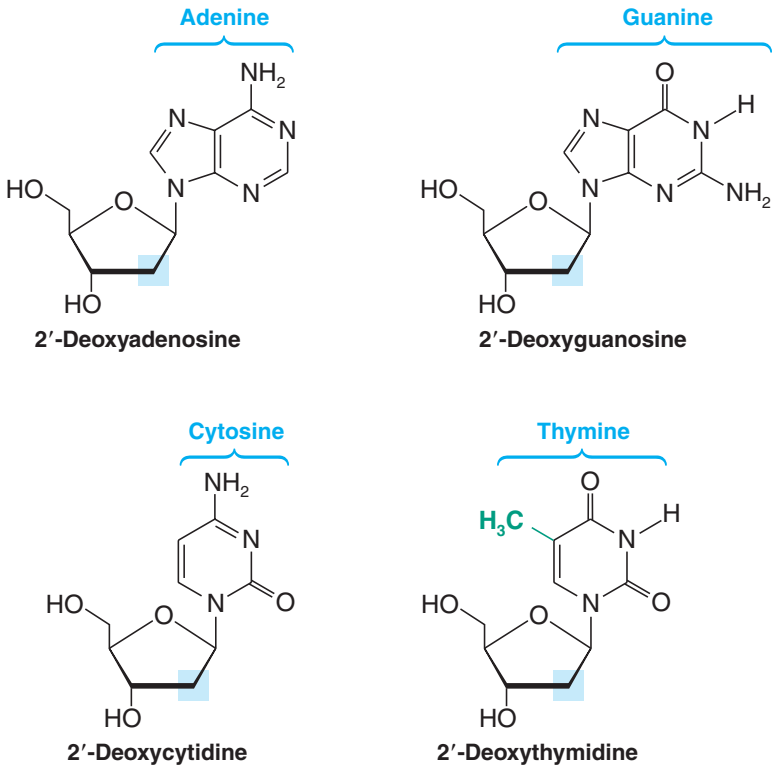


FIGURE 25.4 Nucleosides that can be obtained from DNA. DNA is 2′-deoxy at the position where the blue shaded box is shown. RNA (see Fig 25.5) has hydroxyl groups at that location. RNA has a hydrogen where there is a methyl group in thymine, which in RNA makes the base uracil (and the nucleoside uridine).

The names and structures of the nucleosides found in DNA are shown in Fig. 25.4; those found in RNA are given in Fig. 25.5.

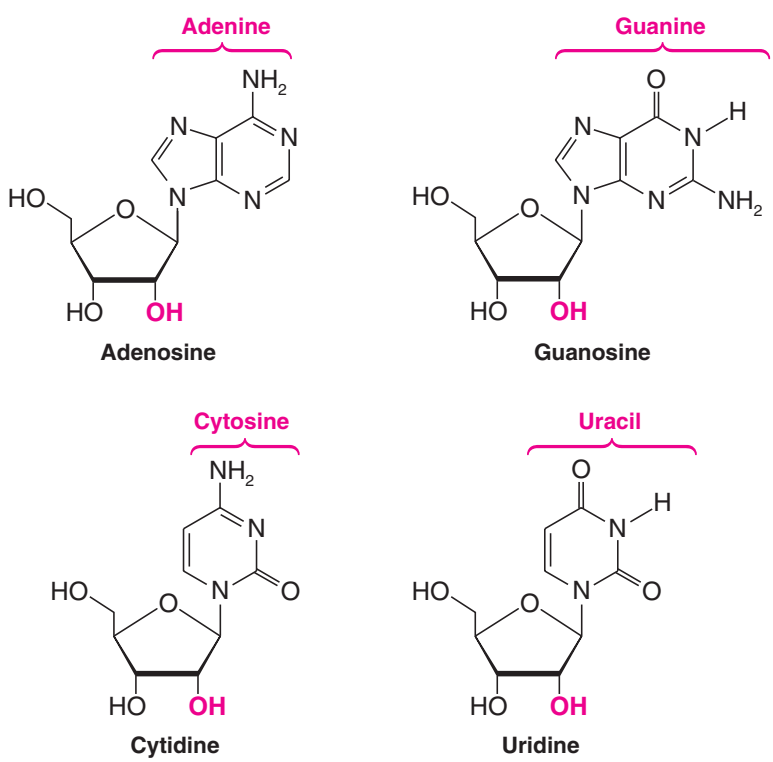


FIGURE 25.5 Nucleosides that can be obtained from RNA. DNA (see Fig 25.4) has hydrogen atoms where the red hydroxyl groups of ribose are shown (DNA is 2′-deoxy with respect to its ribose moiety).

PRACTICE PROBLEM 25.1

Write the structures of other tautomeric forms of adenine, guanine, cytosine, thymine, and uracil.

PRACTICE PROBLEM 25.2 The nucleosides shown in Figs. 25.4 and 25.5 are stable in dilute base. In dilute acid, however, they undergo rapid hydrolysis yielding a sugar (deoxyribose or ribose) and a heterocyclic base.

(a) What structural feature of the nucleoside accounts for this behavior?

(b) Propose a reasonable mechanism for the hydrolysis.

Nucleotides are named in several ways. Adenylic acid (Fig. 25.3), for example, is usually called AMP, for adenosine monophosphate. The position of the phosphate group is sometimes explicitly noted by use of the names adenosine 5′-monophosphate or 5′-adenylic acid. Uridylic acid is usually called UMP, for uridine monophosphate, although it can also be called uridine 5′-monophosphate or 5′-uridylic acid. If a nucleotide is present as a diphosphate or triphosphate, the names are adjusted accordingly, such as ADP for adenosine diphosphate or GTP for guanosine triphosphate.

Nucleosides and nucleotides are found in places other than as part of the structure of DNA and RNA. We have seen, for example, that adenosine units are part of the structures of two important coenzymes, NADH and coenzyme A. The 5′-triphosphate of adenosine is, of course, the important energy source, ATP (Section 22.1B). The compound called 3′,5′-cyclic adenylic acid (or cyclic AMP) (Fig. 25.6) is an important regulator of hormone activity. Cells synthesize this compound from ATP through the action of an enzyme, *adenylate cyclase*. In the laboratory, 3′,5′-cyclic adenylic acid can be prepared through dehydration of 5′-adenylic acid with dicyclohexylcarbodiimide.

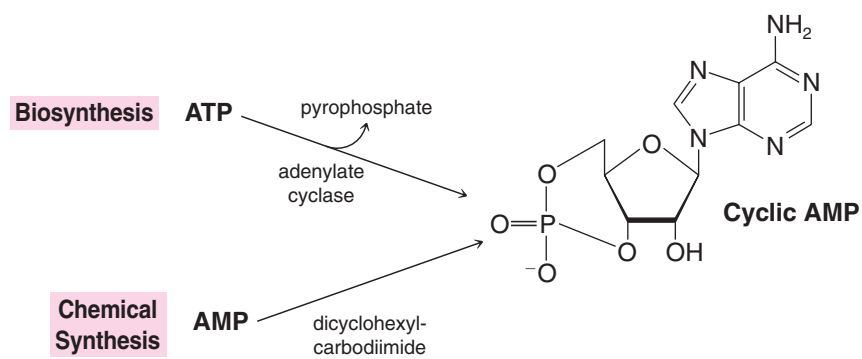


FIGURE 25.6 3′,5′-Cyclic adenylic acid (cyclic AMP) and its biosynthesis and laboratory synthesis.

SOLVED PROBLEM 25.1

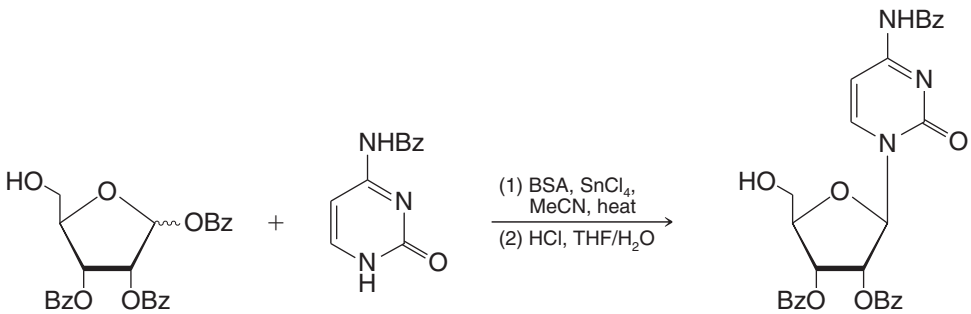
When 3′,5′-cyclic adenylic acid is treated with aqueous sodium hydroxide, the major product that is obtained is 3′-adenylic acid (adenosine 3′-phosphate) rather than 5′-adenylic acid. Suggest a mechanism that explains the course of this reaction.

STRATEGY AND ANSWER: The reaction appears to take place through an S_N2 mechanism. Attack occurs preferentially at the primary 5′-carbon atom rather than at the secondary 3′-carbon atom due to the difference in steric hindrance.

25.3 LABORATORY SYNTHESIS OF NUCLEOSIDES AND NUCLEOTIDES

A variety of methods have been developed for the chemical synthesis of nucleosides from the constituent sugars and bases or their precursors. The following is an example of a *silyl–Hilbert–Johnson nucleosidation*, where a benzoyl protected sugar (D-ribose) reacts in the presence of tin(IV) chloride with an *N*-benzoyl protected base (cytidine)

that is protected further by *in situ* silylation.* The trimethylsilyl protecting groups for the base are introduced using *N,O*-bis(trimethylsilyl)acetamide (BSA) and they are removed with aqueous acid in the second step. The result is a protected form of the nucleoside cytosine, from which the benzoyl groups can be removed with ease using a base:

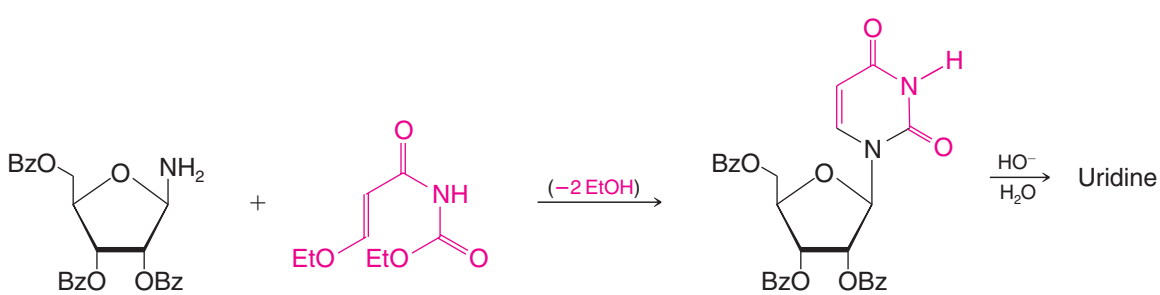


D-Ribose (Bz protected) **N^4-Benzoylcytosine** **Cytidine (Bz protected)**

Bz = C_6H_5CO (benzoyl)

BSA = $CH_3C[{=}NSi(CH_3)_3]OSi(CH_3)_3$, [*N,O*-bis(trimethylsilyl)acetamide, a reagent for trimethylsilylation of nucleophilic oxygen and nitrogen atoms]

Another technique involves formation of the heterocyclic base on a protected ribosylamine derivative:



2,3,5-Tri-*O*-benzoyl-β-D-ribofuranosylamine **β-Ethoxy-*N*-ethoxycarbonylacrylamide**

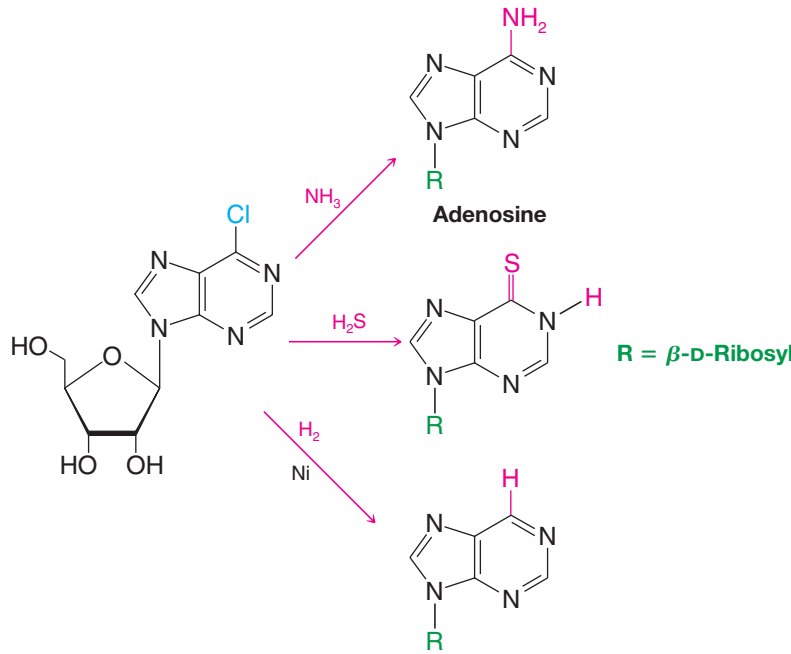
PRACTICE PROBLEM 25.3

Basing your answer on reactions that you have seen before, propose a likely mechanism for the condensation reaction in the first step of the preceding uridine synthesis.

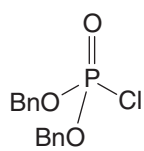
A third technique involves the synthesis of a nucleoside with a substituent in the heterocyclic ring that can be replaced with other groups. This method has been used extensively to synthesize unusual nucleosides that do not necessarily occur naturally. The

*These conditions were applied using L-ribose in a synthesis of the unnatural enantiomer of RNA (Pitsch, S. An efficient synthesis of enantiomeric ribonucleic acids from D-glucose. *Helv. Chim. Acta* **1997**, *80*, 2286–2314). The protected enantiomeric cytidine was produced in 94% yield by the above reaction. After adjusting protecting groups, solid-phase oligonucleotide synthesis methods (Section 25.7) were used with this compound and the other three nucleotide monomers (also derived from L-ribose) for preparation of the unnatural RNA enantiomer. See also Vorbrüggen, H.; Ruh-Pohlenz, C., *Handbook of Nucleoside Synthesis*; Wiley: Hoboken, NJ, 2001.

following example makes use of a 6-chloropurine derivative obtained from the appropriate ribofuranosyl chloride and chloromercuripurine:

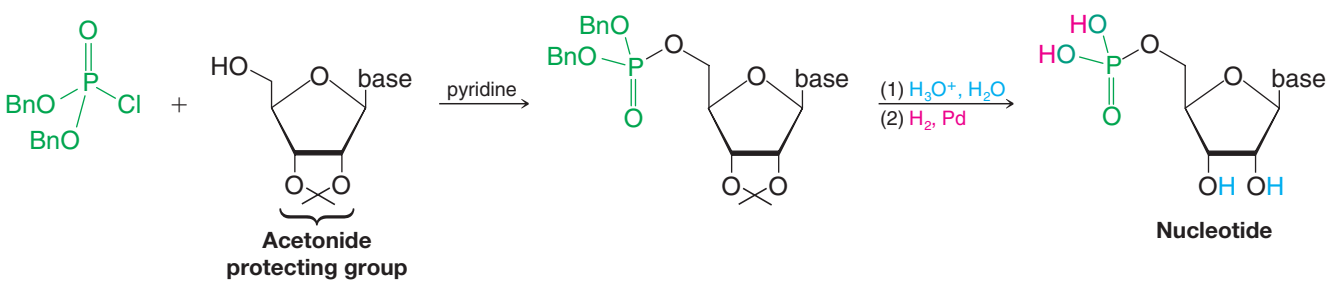


Numerous phosphorylating agents have been used to convert nucleosides to nucleotides. One of the most useful is dibenzyl phosphochloridate:



Bn = $C_6H_5CH_2$ **(benzyl)**

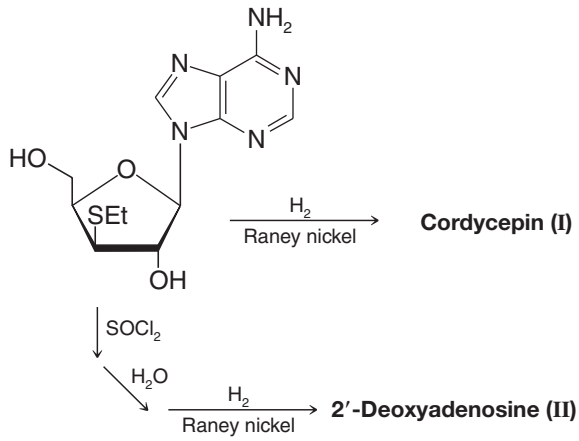
Specific phosphorylation of the 5′-OH can be achieved if the 2′- and 3′-OH groups of the nucleoside are protected by an acetonide group (see the following):



Mild acid-catalyzed hydrolysis removes the acetonide group, and hydrogenolysis cleaves the benzyl phosphate bonds.

PRACTICE PROBLEM 25.4 **(a)** What kind of linkage is involved in the acetonide group of the protected nucleoside, and why is it susceptible to mild acid-catalyzed hydrolysis? **(b)** How might such a protecting group be installed?

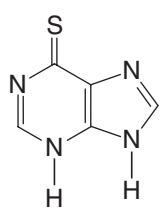
PRACTICE PROBLEM 25.5 The following reaction scheme is from a synthesis of cordycepin (a nucleoside antibiotic) and the first synthesis of 2′-deoxyadenosine (reported in 1958 by C. D. Anderson, L. Goodman, and B. R. Baker, Stanford Research Institute):



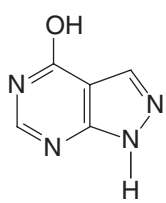
(a) What is the structure of cordycepin? (**I** and **II** are isomers.)
(b) Propose a mechanism that explains the formation of **II**.

25.3A Medical Applications

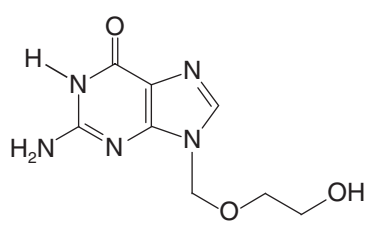
In the early 1950s, Gertrude Elion and George Hitchings (of the Wellcome Research Laboratories) discovered that 6-mercaptopurine had antitumor and antileukemic properties. This discovery led to the development of other purine derivatives and related compounds, including nucleosides, of considerable medical importance. Three examples are the following:



6-Mercaptopurine



Allopurinol



Acyclovir

ELION AND HITCHINGS shared the 1988 Nobel Prize in Physiology or Medicine for their work in the development of chemotherapeutic agents derived from purines.

6-Mercaptopurine is used in combination with other chemotherapeutic agents to treat acute leukemia in children, and almost 80% of the children treated are now cured. Allopurinol, another purine derivative, is a standard therapy for the treatment of gout. Acyclovir, a nucleoside that lacks two carbon atoms of its ribose ring, is highly effective in treating diseases caused by certain herpes viruses, including *herpes simplex* type 1 (fever blisters), type 2 (genital herpes), and varicella-zoster (shingles).

25.4 DEOXYRIBONUCLEIC ACID: DNA

25.4A Primary Structure

Nucleotides bear the same relation to a nucleic acid that amino acids do to a protein: they are its monomeric units. The connecting links in proteins are amide groups; in nucleic acids they are phosphate ester linkages. Phosphate esters link the 3′-OH of one ribose (or deoxyribose) with the 5′-OH of another. This makes the nucleic acid a long unbranched chain with a "backbone" of sugar and phosphate units with heterocyclic bases protruding

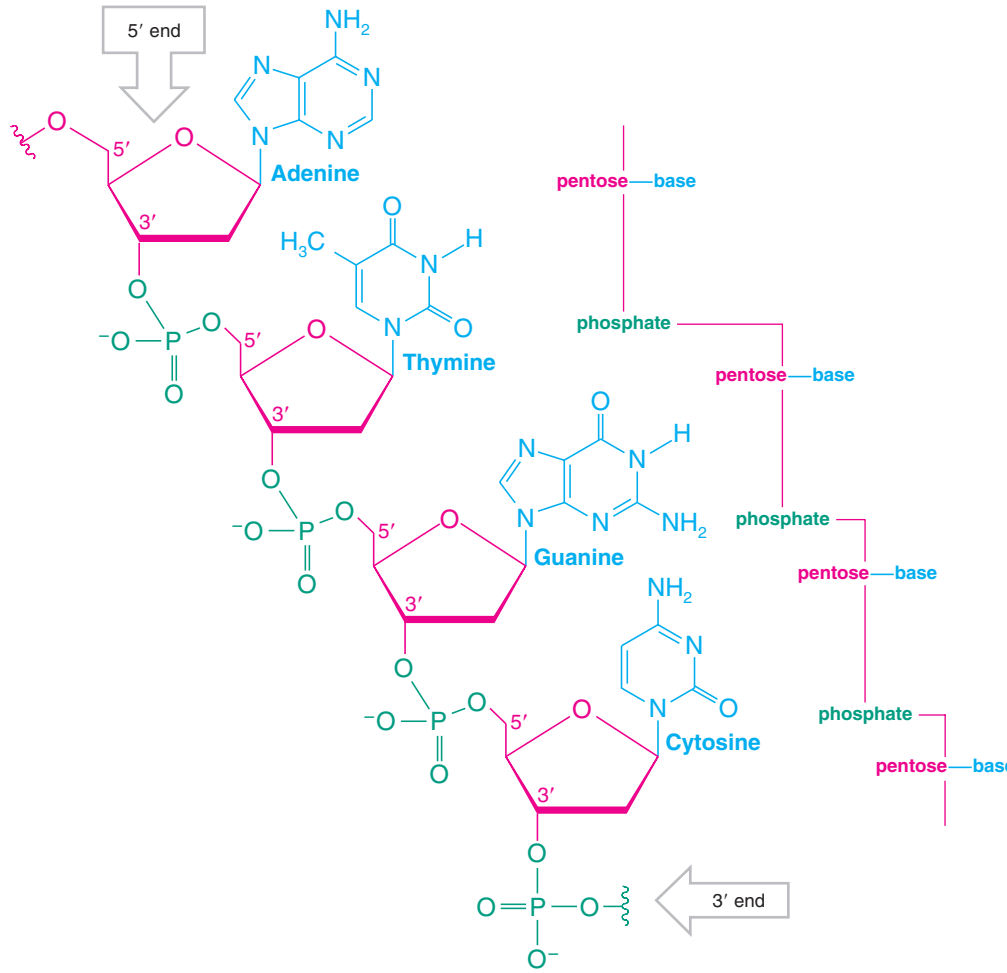


FIGURE 25.7 A segment of one DNA chain showing how phosphate ester groups link the 3′- and 5′-OH groups of deoxyribose units. RNA has a similar structure with two exceptions: a hydroxyl replaces a hydrogen atom at the 2′ position of each ribose unit and uracil replaces thymine.

from the chain at regular intervals (Fig. 25.7). We would indicate the direction of the bases in Fig. 25.7 in the following way:

$$5' \leftarrow A—T—G—C \rightarrow 3'$$

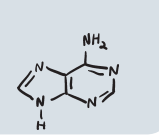
It is, as we shall see, the **base sequence** along the chain of DNA that contains the encoded genetic information. The sequence of bases can be determined using enzymatic methods and chromatography (Section 25.6).

25.4B Secondary Structure

"I cannot help wondering whether some day an enthusiastic scientist will christen his newborn twins Adenine and Thymine."
F. H. C. Crick*

It was the now-classic proposal of James Watson and Francis Crick (made in 1953 and verified shortly thereafter through the X-ray analysis by Maurice Wilkins) that gave a model for the secondary structure of DNA. This work earned Crick, Watson, and Wilkins the 1962 Nobel Prize in Physiology or Medicine. Many believe that Rosalind Franklin, whose X-ray data was also key to solving the structure of DNA, should have shared the Nobel Prize, but her death from cancer in 1958 precluded it. The secondary structure of DNA is especially important because it enables us to understand how genetic information is preserved, how it can be passed on during the process of cell division, and how it can be transcribed to provide a template for protein synthesis.

*Taken from Crick, F. H. C., The structure of the hereditary material. *Sci. Am.* **1954**, *191*(10), 20, 54–61.



Of prime importance to Watson and Crick's proposal was an earlier observation (made in the late 1940s) by Erwin Chargaff that certain regularities can be seen in the percentages of heterocyclic bases obtained from the DNA of a variety of species. Table 25.1 gives results that are typical of those that can be obtained.

TABLE 25.1 DNA COMPOSITION OF VARIOUS SPECIES

Species	Base Proportions (mol %)							
	G	A	C	T	$\frac{G+A}{C+T}$	$\frac{A+T}{G+C}$	$\frac{A}{T}$	$\frac{G}{C}$
Sarcina lutea	37.1	13.4	37.1	12.4	1.02	0.35	1.08	1.00
Escherichia coli K12	24.9	26.0	25.2	23.9	1.08	1.00	1.09	0.99
Wheat germ	22.7	27.3	22.8[a]	27.1	1.00	1.19	1.01	1.00
Bovine thymus	21.5	28.2	22.5[a]	27.8	0.96	1.27	1.01	0.96
Staphylococcus aureus	21.0	30.8	19.0	29.2	1.11	1.50	1.05	1.11
Human thymus	19.9	30.9	19.8	29.4	1.01	1.52	1.05	1.01
Human liver	19.5	30.3	19.9	30.3	0.98	1.54	1.00	0.98

[a]Cytosine + methylcytosine.

Source: Reproduced with permission of The McGraw-Hill Companies, Inc. from Smith, E.L.; Hill, R.L.; Jehman, I.R.; Lefkowitz, R.J.; Handler, P.; and White, A., *Principles of Biochemistry: General Aspects*, 7th edition, ©1982.

Chargaff pointed out that for all species examined:

1. The total mole percentage of purines is approximately equal to that of the pyrimidines, that is, (%G + %A)/(%C + %T) $\cong$ 1.
2. The mole percentage of adenine is nearly equal to that of thymine (i.e., %A/%T $\cong$ 1), and the mole percentage of guanine is nearly equal to that of cytosine (i.e., %G/%C $\cong$ 1).

Chargaff also noted that the ratio which varies from species to species is the ratio (%A + %T)/(%G + %C). He noted, moreover, that whereas this ratio is characteristic of the DNA of a given species, it is the same for DNA obtained from different tissues of the same animal and does not vary appreciably with the age or conditions of growth of individual organisms within the same species.

Watson and Crick also had X-ray data that gave them the bond lengths and angles of the purine and pyrimidine rings of model compounds. In addition, they had data from Franklin and Wilkins that indicated a repeat distance of 34 Å in DNA.

Reasoning from these data, Watson and Crick proposed a double helix as a model for the secondary structure of DNA. According to this model, two nucleic acid chains are held together by hydrogen bonds between base pairs on opposite strands. This double chain is wound into a helix with both chains sharing the same axis. The base pairs are on the inside of the helix, and the sugar–phosphate backbone is on the outside (Fig. 25.8). The pitch of the helix is such that 10 successive nucleotide pairs give rise to one complete turn in 34 Å (the repeat distance). The exterior width of the spiral is about 20 Å, and the internal distance between 1′ positions of ribose units on opposite chains is about 11 Å.

Using molecular-scale models, Watson and Crick observed that the internal distance of the double helix is such that it allows only a purine–pyrimidine type of hydrogen bonding between base pairs. Purine–purine base pairs do not occur because they would be too large to fit, and pyrimidine–pyrimidine base pairs do not occur because they would be too far apart to form effective hydrogen bonds.

Helpful Hint

The use of models was critical to Watson and Crick in their Nobel prize-winning work on the three-dimensional structure of DNA.

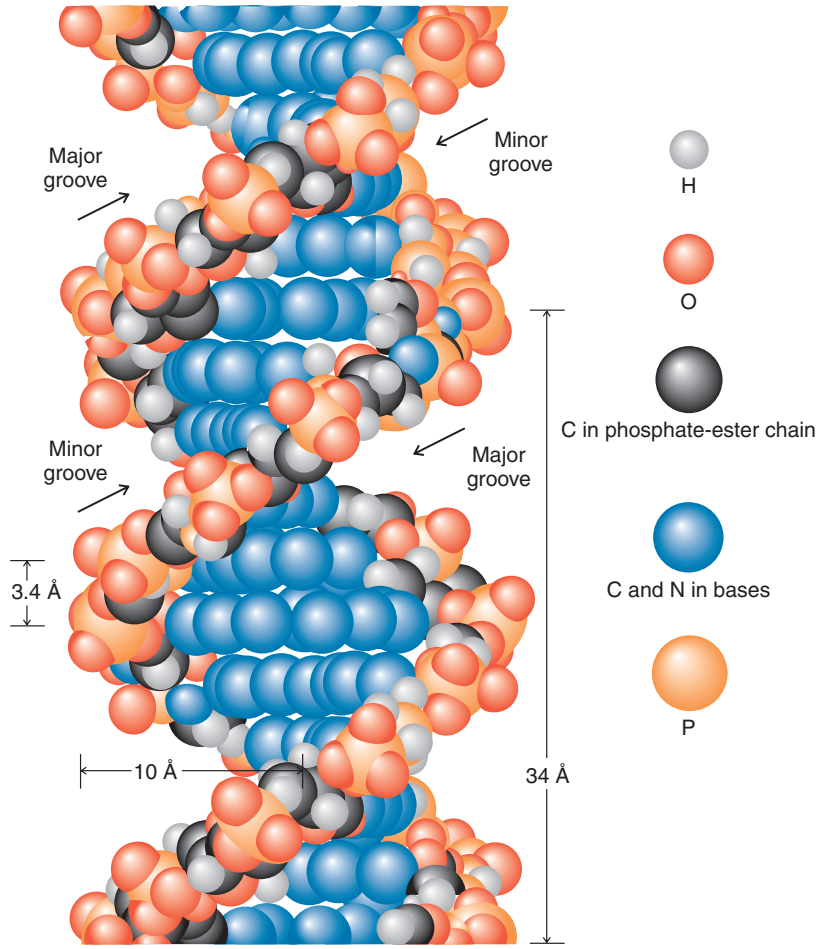
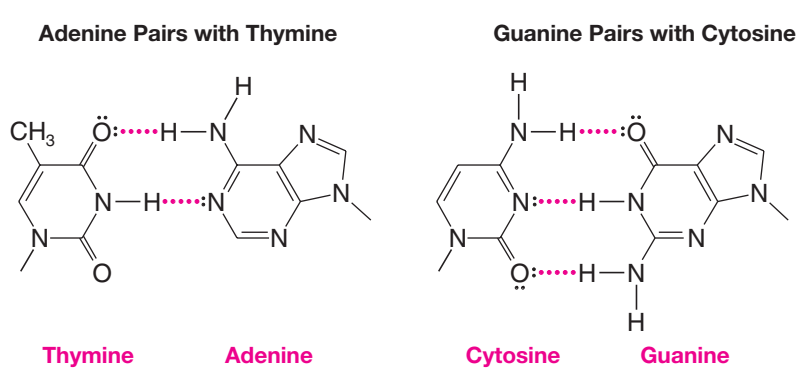


FIGURE 25.8 A molecular model of a portion of the DNA double helix. (Reprinted with permission of The McGraw-Hill Companies from Neal, L., *Chemistry and Biochemistry: A Comprehensive Introduction*, © 1971.)

Watson and Crick went one crucial step further in their proposal. Assuming that the oxygen-containing heterocyclic bases existed in keto forms, they argued that base pairing through hydrogen bonds can occur in only a specific way: adenine (A) with thymine (T) with cytosine (C) with guanine (G). Dimensions of the pairs and electrostatic potential maps for the individual bases are shown in Fig. 25.9.



Specific base pairing of this kind is consistent with Chargaff's finding that %A/%T ≅ 1 and %G/%C ≅ 1.

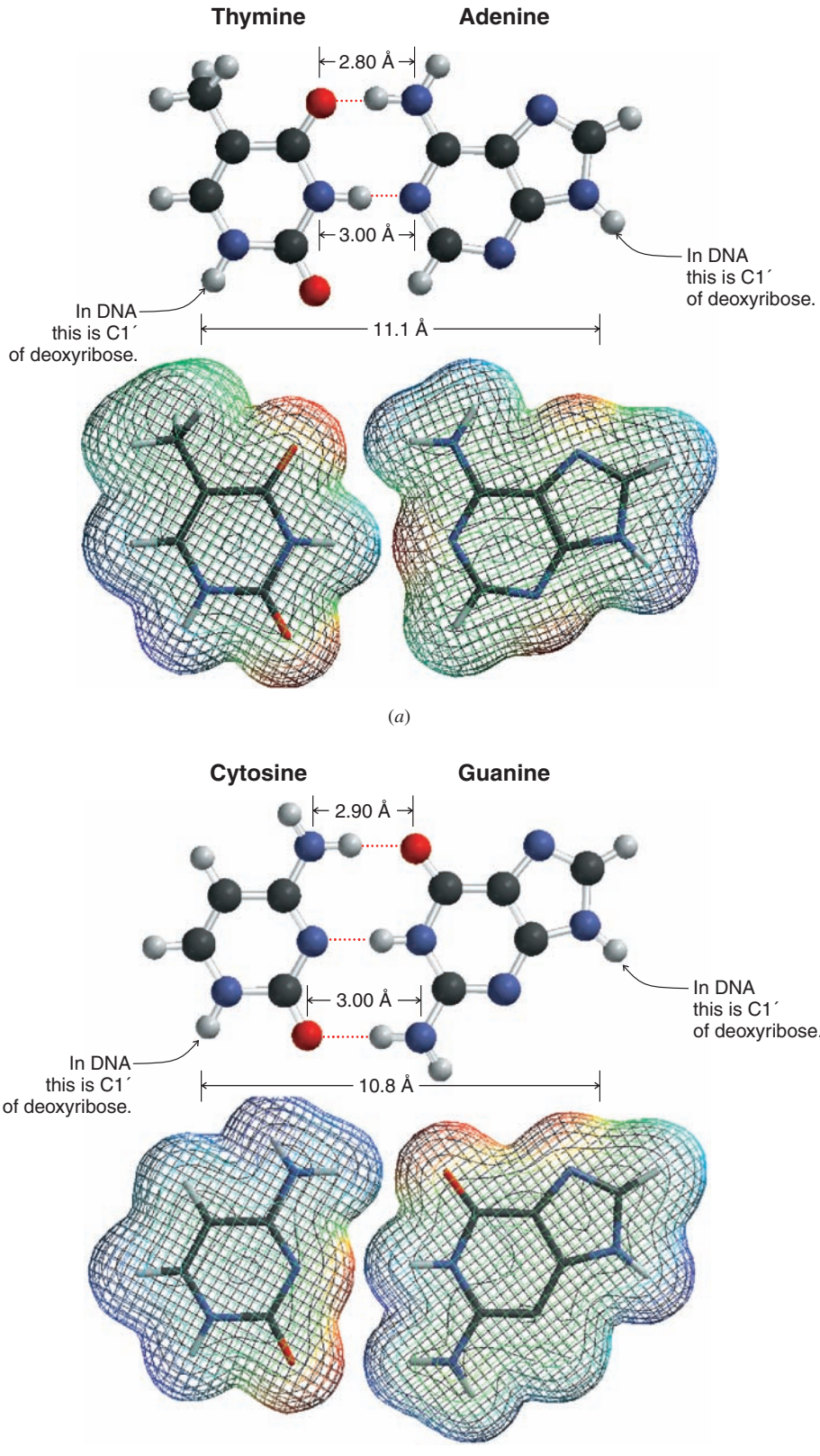


FIGURE 25.9 Base pairing of adenine with thymine (*a*) and cytosine with guanine (*b*). The dimensions of the thymine–adenine and cytosine–guanine hydrogen-bonded pairs are such that they allow the formation of strong hydrogen bonds and also allow the base pairs to fit inside the two phosphate–ribose chains of the double helix. Electrostatic potential maps calculated for the individual bases show the complementary distribution of charges that leads to hydrogen bonding. (Ball and stick models reprinted from Archives of *Biochemistry and Biophysics*, **65**, Pauling, I., Corey, R., p. 164–181, 1956. Copyright 1956, with permission from Elsevier.)

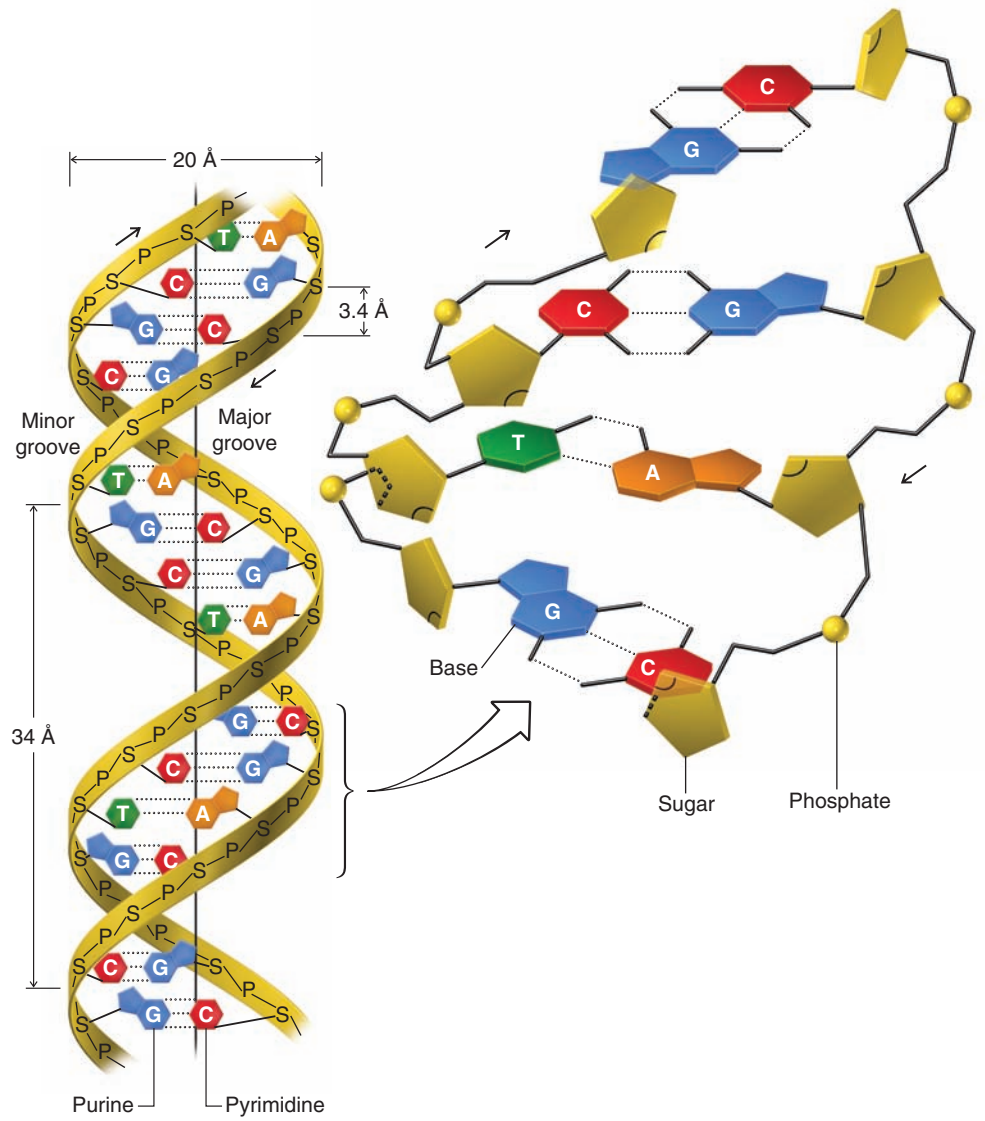


FIGURE 25.10 Diagram of the DNA double helix showing complementary base pairing. The arrows indicate the 3′ → 5′ direction.

Specific base pairing also means that the two chains of DNA are complementary. Wherever adenine appears in one chain, thymine must appear opposite it in the other; wherever cytosine appears in one chain, guanine must appear in the other (Fig. 25.10).

Notice that while the sugar–phosphate backbone of DNA is completely regular, the sequence of heterocyclic base pairs along the backbone can assume many different permutations. This is important because it is the precise sequence of base pairs that carries the genetic information. Notice, too, that one chain of the double strand is the complement of the other. If one knows the sequence of bases along one chain, one can write down the sequence along the other, because A always pairs with T and G always pairs with C. It is this complementarity of the two strands that explains how a DNA molecule replicates itself at the time of cell division and thereby passes on the genetic information to each of the two daughter cells.

25.4C Replication of DNA

Just prior to cell division the double strand of DNA begins to unwind. Complementary strands are formed along each chain (Fig. 25.11). Each chain acts, in effect, as a template

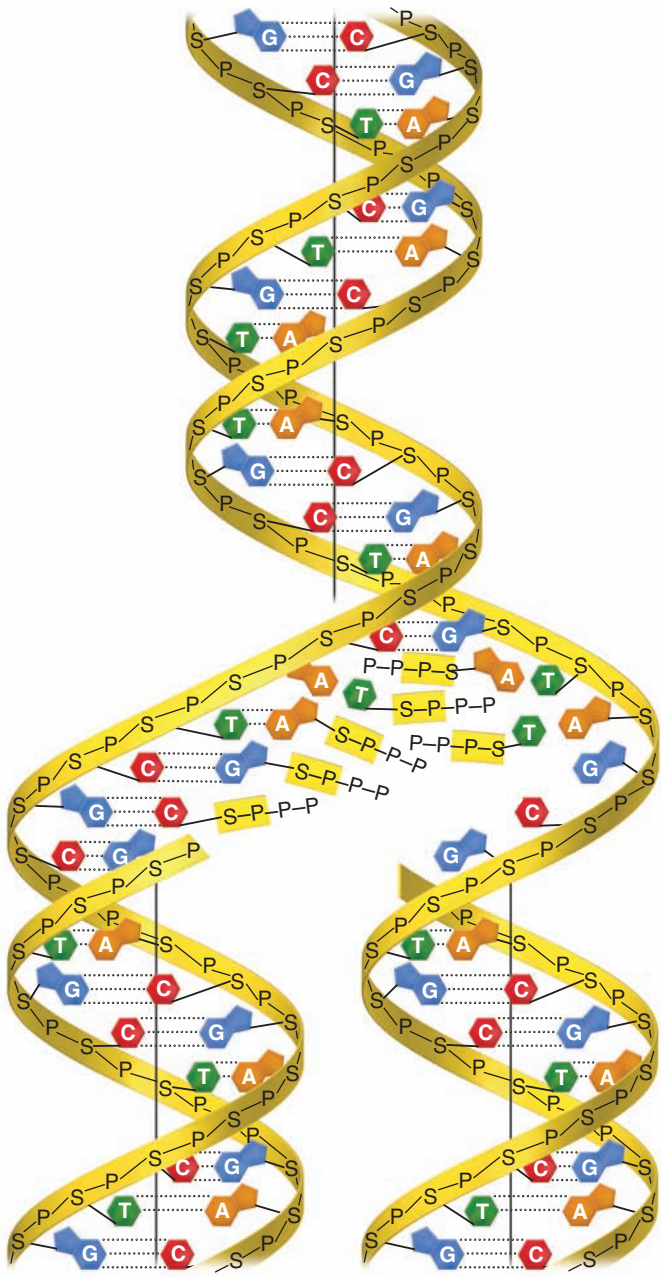


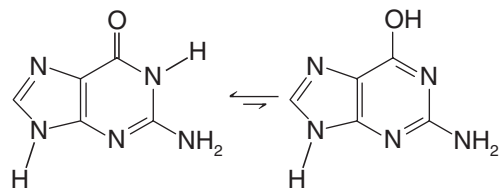
FIGURE 25.11 Replication of DNA. The double strand unwinds from one end and complementary strands are formed along each chain.

for the formation of its complement. When unwinding and **replication** are complete, there are two identical DNA molecules where only one had existed before. These two molecules can then be passed on, one to each daughter cell.

PRACTICE PROBLEM 25.6

(a) There are approximately 3 billion base pairs in the DNA of a single human cell. Assuming that this DNA exists as a double helix, calculate the length of all the DNA contained in a human cell. **(b)** The weight of DNA in a single human cell is 6×10^{-12} g. Assuming that Earth's population is about 6.5 billion, we can conclude that all of the genetic information that gave rise to all human beings now alive was once contained in the DNA of a corresponding number of fertilized ova. What is the total weight of DNA in this many ova? (The volume that this DNA would occupy is approximately that of a raindrop, yet if the individual molecules were laid end-to-end, they would stretch to the moon and back almost eight times.)

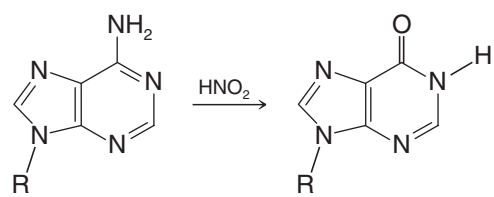
PRACTICE PROBLEM 25.7 **(a)** The most stable tautomeric form of guanine is the lactam form (or cyclic amide, see Section 17.8I). This is the form normally present in DNA, and, as we have seen, it pairs specifically with cytosine. If guanine tautomerizes (see Section 18.2) to the lactim form, it pairs with thymine instead. Write structural formulas showing the hydrogen bonds in this abnormal base pair.



Lactam form of guanine **Lactim form of guanine**

(b) Improper base pairings that result from tautomerizations occurring during the process of DNA replication have been suggested as a source of spontaneous mutations. We saw in part (a) that if a tautomerization of guanine occurred at the proper moment, it could lead to the introduction of thymine (instead of cytosine) into its complementary DNA chain. What error would this new DNA chain introduce into *its* complementary strand during the next replication even if no further tautomerizations take place?

PRACTICE PROBLEM 25.8 Mutations can also be caused chemically, and nitrous acid is one of the most potent chemical **mutagens**. One explanation that has been suggested for the mutagenic effect of nitrous acid is the deamination reactions that it causes with purines and pyrimidines bearing amino groups. When, for example, an adenine-containing nucleotide is treated with nitrous acid, it is converted to a hypoxanthine derivative:



Adenine nucleotide **Hypoxanthine nucleotide**

(a) Basing your answer on reactions you have seen before, what are likely intermediates in the adenine → hypoxanthine interconversion? **(b)** Adenine normally pairs with thymine in DNA, but hypoxanthine pairs with cytosine. Show the hydrogen bonds of a hypoxanthine–cytosine base pair. **(c)** Show what errors an adenine → hypoxanthine interconversion would generate in DNA through two replications.

25.5 RNA AND PROTEIN SYNTHESIS

Soon after the Watson–Crick hypothesis was published, scientists began to extend it to yield what Crick called "the central dogma of molecular genetics." This dogma stated that genetic information flows as follows:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

The synthesis of protein is, of course, all important to a cell's function because proteins (as enzymes) catalyze its reactions. Even the very primitive cells of bacteria require as many as 3000 different enzymes. This means that the DNA molecules of these cells must contain a corresponding number of genes to direct the synthesis of these proteins. A **gene** is that segment of the DNA molecule that contains the information necessary to direct the synthesis of one protein (or one polypeptide).

DNA is found primarily in the nucleus of eukaryotic cells. Protein synthesis takes place primarily in that part of the cell called the *cytoplasm*. Protein synthesis requires that two major processes take place; the first occurs in the cell nucleus, the second in the cytoplasm. The first is **transcription**, a process in which the genetic message is transcribed onto a form of RNA called messenger RNA (mRNA). The second process involves two other forms of RNA, called ribosomal RNA (rRNA) and transfer RNA (tRNA).

There are viruses, called retroviruses, in which information flows from RNA to DNA. The virus that causes AIDS is a retrovirus.

25.5A Messenger RNA Synthesis—Transcription

The events leading to protein synthesis begin in the cell nucleus with the synthesis of mRNA. Part of the DNA double helix unwinds sufficiently to expose on a single chain a portion corresponding to at least one gene. Ribonucleotides, present in the cell nucleus, assemble along the exposed DNA chain by pairing with the bases of DNA. The pairing patterns are the same as those in DNA with the exception that in RNA uracil replaces thymine. The ribonucleotide units of mRNA are joined into a chain by an enzyme called *RNA polymerase*. This process is illustrated in Fig. 25.12.

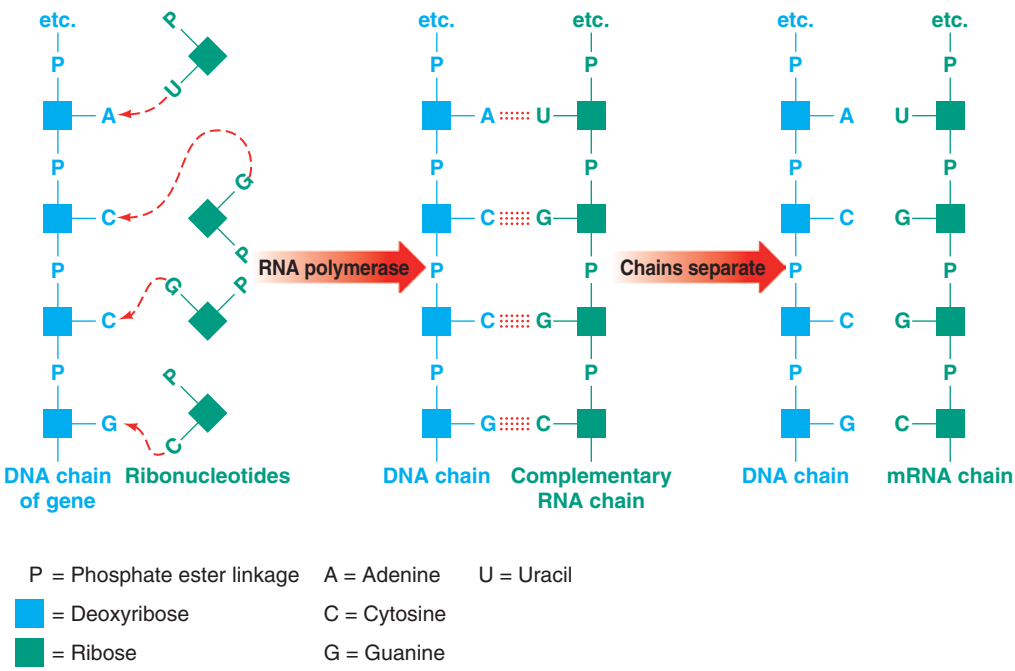


FIGURE 25.12 Transcription of the genetic code from DNA to mRNA.

PRACTICE PROBLEM 25.9

Write structural formulas showing how the keto form of uracil (Section 25.2) in mRNA can pair with adenine in DNA through hydrogen bond formation.

Most eukaryotic genes contain segments of DNA that are not actually used when a protein is expressed, even though they are transcribed into the initial mRNA. These segments are called **introns**, or intervening sequences. The segments of DNA within a gene that are expressed are called **exons**, or expressed sequences. Each gene usually contains a number of introns and exons. After the mRNA is transcribed from DNA, the introns in the mRNA are removed and the exons are spliced together.

After mRNA has been synthesized and processed in the cell nucleus to remove the introns, it migrates into the cytoplasm where, as we shall see, it acts as a template for protein synthesis.

25.5B Ribosomes—rRNA

Protein synthesis is catalyzed by ribosomes in the cytoplasm. Ribosomes (Fig. 25.13) are ribonucleoproteins, comprised of approximately two-thirds RNA and one-third protein. They have a very high molecular weight (about 2.6×10^6). The RNA component is present in two subunits, called the 50S and 30S subunits (classified according to their sedimentation behavior during ultracentrifugation*). The 50S subunit is roughly twice the molecular weight of the 30S subunit. Binding of RNA with mRNA is mediated by the 30S subunit. The 50S subunit carries the catalytic activity for translation that joins one amino acid by an amide bond to the next. In addition to the rRNA subunits there are approximately 30–35 proteins tightly bound to the ribosome, the entire structure resembling an exquisite three-dimensional jigsaw puzzle of RNA and protein. The mechanism for ribosome-catalyzed amide bond formation is discussed below.

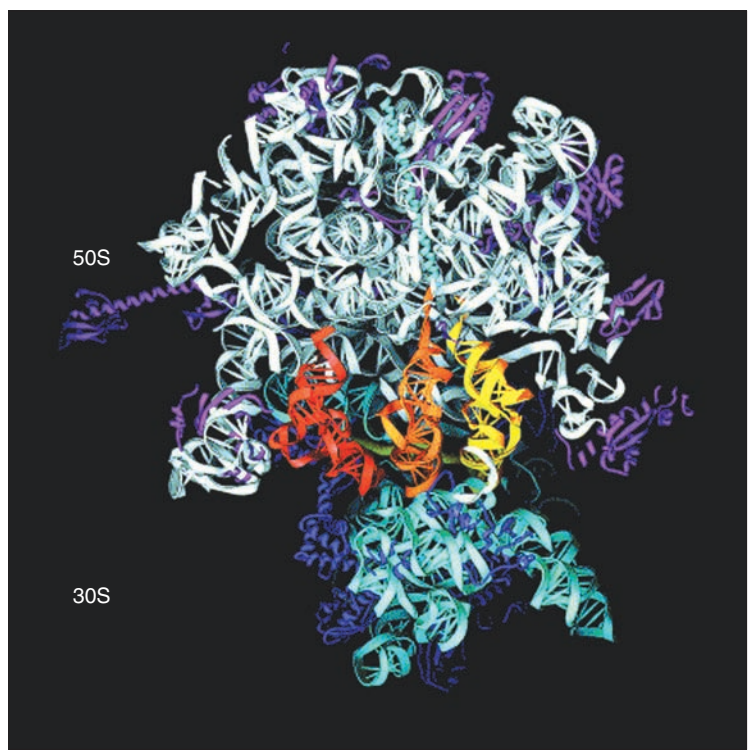


FIGURE 25.13 Structure of the *Thermus thermophilus* ribosome showing the 50S and 30S subunits and three bound transfer RNAs. The yellow tRNA is at the A site, which would bear the new amino acid to be added to the peptide. The light orange tRNA is at the P site, which would be the tRNA that bears the growing peptide. The red tRNA is at the E site, which is the "empty" tRNA after it has transferred the peptide chain to the new amino acid. *(Courtesy of Harry Noller, University of California, Santa Cruz.)*

Ribosomes, as reaction catalysts, are most appropriately classified as **ribozymes** rather than enzymes, because it is RNA that catalyzes the peptide bond formation during protein synthesis and not the protein subunits of the ribosome. The mechanism for peptide bond formation catalyzed by the 50S ribosome subunit (Fig. 25.14), proposed by Moore and co-workers based on X-ray crystal structures, suggests that attack by the α-amino group is facilitated by acid–base catalysis involving nucleotide residues along the 50S ribosome subunit chain, specifically a nearby adenine group. Full or partial removal of a proton from the α-amino group of the amino acid by N3 of the adenine group imparts greater nucleophilicity to the amino nitrogen, facilitating its attack on the acyl carbon of the adjacent peptide–tRNA moiety. A tetrahedral intermediate is formed, which collapses to form the new amide bond with release of the tRNA that had been joined to the peptide. Other moieties in the 50S ribosome

*S stands for svedberg unit; it is used in describing the behavior of proteins in an ultracentrifuge.

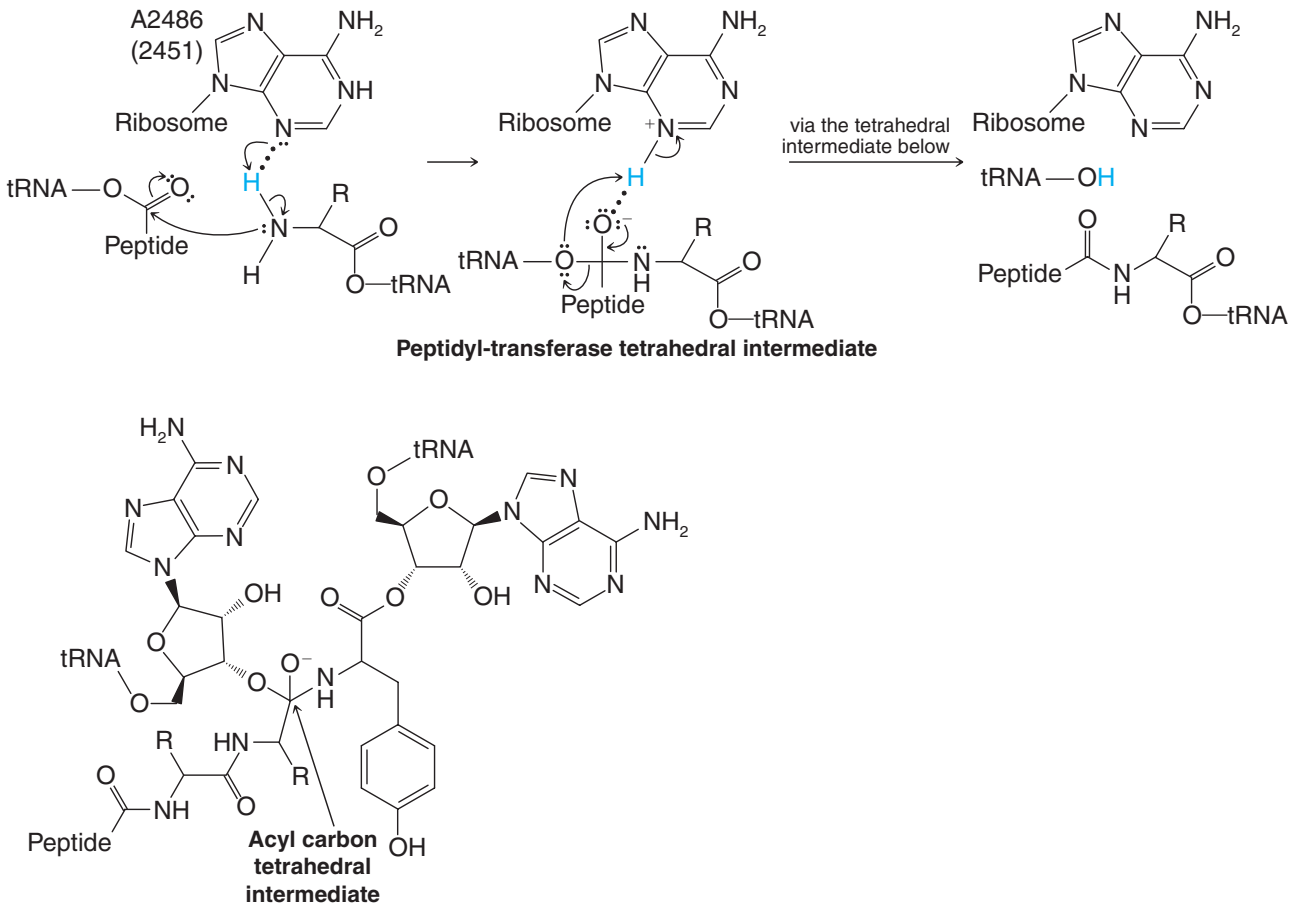


FIGURE 25.14 A mechanism for peptide bond formation catalyzed by the 50S subunit of the ribosome (as proposed by Moore and co-workers). The new amide bond in the growing peptide chain is formed by attack of the α-amino group in the new amino acid, brought to the A site of the ribosome by its tRNA, on the acyl carbon linkage of the peptide held at the P site by its tRNA. Acid–base catalysis by groups in the ribosome facilitate the reaction. (Reprinted with permission from Nissen et al., The structural basis of ribosome activity in peptide bond synthesis. SCIENCE 289:920–930 (2000). Reprinted with permission from AAAS. Also reprinted from Monro, R. E., and Marker, K. A., Ribosome-catalysed reaction of puromycin with a formylmethionine containing oligonucleotide, *J. Mol. Biol.* 25 pp. 347–350. Copyright 1967, with permission of Elsevier.)

subunit are believed to help stabilize the transfer of charge that occurs as N3 of the adenyl group accepts the proton from the α-amino group of the new amino acid (see Problem 25.16).

25.5C Transfer RNA

Transfer RNA has a very low molecular weight when compared to those of mRNA and rRNA. Transfer RNA, consequently, is much more soluble than mRNA or rRNA and is sometimes referred to as soluble RNA. The function of tRNA is to transport amino acids to specific areas on the mRNA bound to the ribosome. There are, therefore, many forms of tRNA, more than one for each of the 20 amino acids that is incorporated into proteins, including the redundancies in the **genetic code** (see Table 25.2).*

The structures of most tRNAs have been determined. They are composed of a relatively small number of nucleotide units (70–90 units) folded into several loops or

*Although proteins are composed of 22 different amino acids, protein synthesis requires only 20. Proline is converted to hydroxyproline and cysteine is converted to cystine after synthesis of the polypeptide chain has taken place.

TABLE 25.2 THE MESSENGER RNA GENETIC CODE

Amino Acid	mRNA Base Sequence 5′ → 3′	Amino Acid	mRNA Base Sequence 5′ → 3′	Amino Acid	mRNA Base Sequence 5′ → 3′
Ala	GCA	His	CAC	Ser	AGC
	GCC		CAU		AGU
	GCG	Ile	AUA		UCA
	GCU		AUC		UCG
Arg	AGA		AUU		UCC
	AGG	Leu	CUA		UCU
	CGA		CUC	Thr	ACA
	CGC		CUG		ACC
	CGG		CUU		ACG
	CGU		UUA		ACU
Asn	AAC		UUG	Trp	UGG
	AAU	Lys	AAA	Tyr	UAC
Asp	GAC		AAG		UAU
	GAU	Met	AUG	Val	GUA
Cys	UGC	Phe	UUU		GUG
	UGU		UUC		GUC
Gln	CAA	Pro	CCA		GUU
	CAG		CCC	Chain initiation	
Glu	GAA		CCG	fMet (*N*-formyl-methionine)	AUG
	GAG		CCU	Chain termination	UAA
Gly	GGA				UAG
	GGC				UGA
	GGG				
	GGU				

arms through base pairing along the chain (Fig. 25.15). One arm always terminates in the sequence cytosine–cytosine–adenine (CCA). It is to this arm that a specific amino acid becomes attached *through an ester* linkage to the 3′-OH of the terminal adenosine. This attachment reaction is catalyzed by an enzyme that is specific for the tRNA and for the amino acid. The specificity may grow out of the enzyme's ability to recognize base sequences along other arms of the tRNA.

At the loop of still another arm is a specific sequence of bases, called the **anticodon**. The anticodon is highly important because it allows the tRNA to bind with a specific site—called the **codon**—of mRNA. The order in which amino acids are brought by their tRNA units to the mRNA strand is determined by the sequence of codons. This sequence, therefore, constitutes a genetic message. Individual units of that message (the individual words, each corresponding to an amino acid) are triplets of nucleotides.

25.5D The Genetic Code

The triplets of nucleotides (the codons) on mRNA are the genetic code (see Table 25.2). The code must be in the form of three bases, not one or two, because there are 20 different amino acids used in protein synthesis but there are only four different bases in mRNA. If only two bases were used, there would be only 4^2, or 16, possible combinations, a number too small to accommodate all of the possible amino acids. However, with a three-base code, 4^3, or 64, different sequences are possible. This is far more than

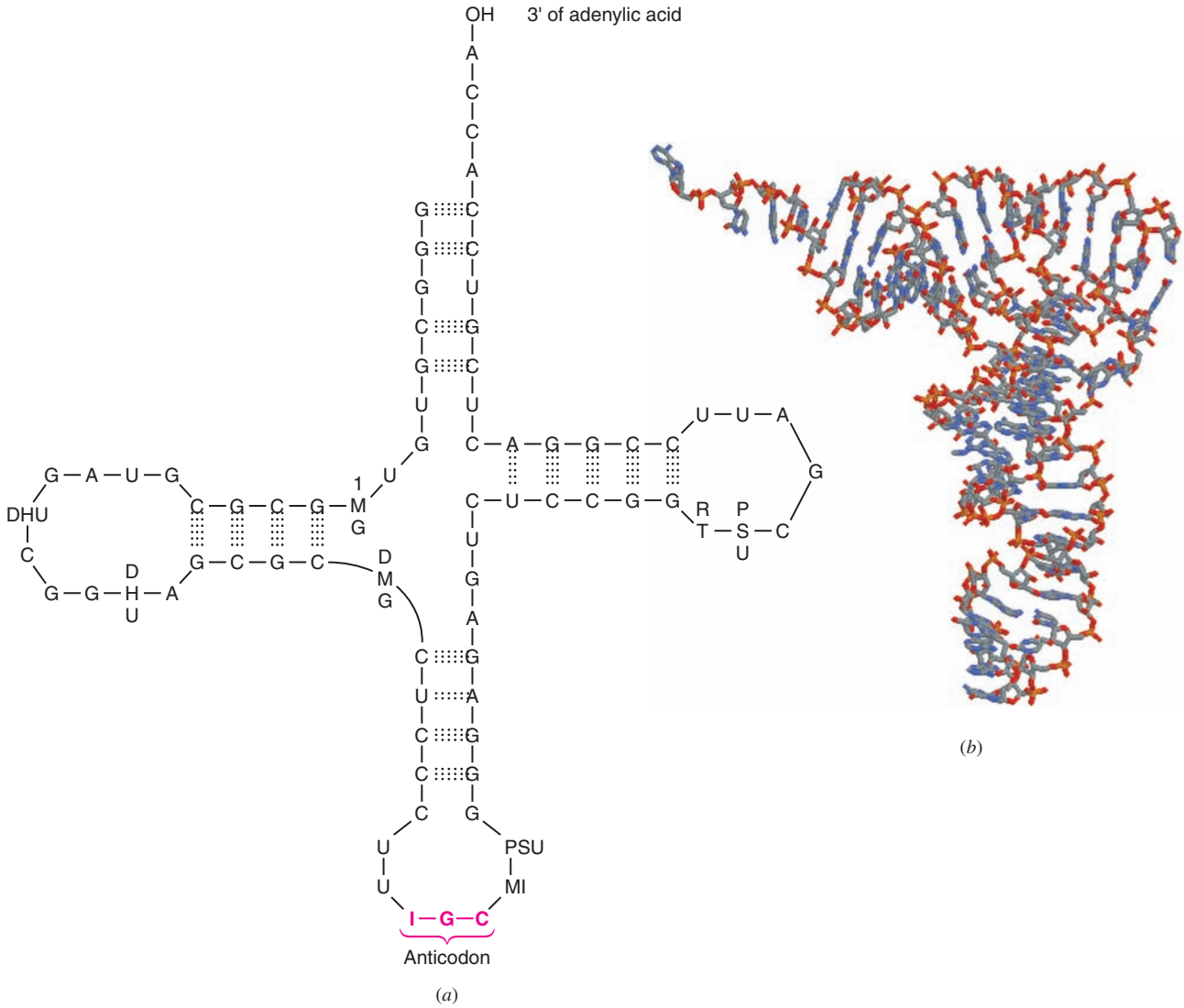
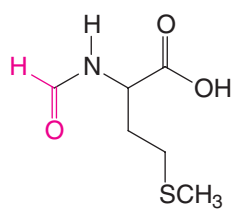


FIGURE 25.15 (*a*) Structure of a tRNA isolated from yeast that has the specific function of transferring alanine residues. Transfer RNAs often contain unusual nucleosides. PSU = pseudouridine, RT = ribothymidine, MI = 1-methylinosine, I = inosine, DMG = N^2-methylguanosine, DHU = 4,5-dihydrouridine, 1MG = 1-methylguanosine. (*b*) The X-ray crystal structure of a phenylalanine–tRNA from yeast. (For part b, Protein Data Bank PDB ID: 4TNA. *http://www.pdb.or* Reprinted from Hingerty, E., Brown, R.S., Jack, A., Further refinement of the structure of yeast $tRNA_{Phe}$, *J. Mol. Biol.* **124**, p. 523. Copyright 1978, with permission of Elsevier.)

are needed, and it allows for multiple ways of specifying an amino acid. It also allows for sequences that punctuate protein synthesis, sequences that say, in effect, "start here" and "end here."

Both methionine (Met) and *N*-formylmethionine (fMet) have the same mRNA code (AUG); however, *N*-formylmethionine is carried by a different tRNA from that which carries methionine. *N*-Formylmethionine appears to be the first amino acid incorporated into the chain of proteins in bacteria, and the tRNA that carries fMet appears to be the punctuation mark that says "start here." Before the polypeptide synthesis is complete, *N*-formylmethionine is removed from the protein chain by an enzymatic hydrolysis.



***N*-Formylmethionine (fMet)**

The genetic code can be expressed in mRNA codons (as we have shown in Table 25.2) or in DNA codons. We have chosen to show the mRNA codons because these are the codons that are actually read during the synthesis of polypeptides (the process called **translation** that we discuss next). However, each mRNA molecule (Section 25.5A) acquires its sequence of nucleotides by **transcription** from the corresponding gene of DNA. In transcription, RNA polymerase (along with other transcription factors) opens the DNA double helix and begins the process.

As RNA polymerase transcribes DNA to mRNA, it moves along the complementary strand of DNA reading it in the 3′ to 5′ direction (called the antisense direction), making an mRNA transcript that is the same as the sense strand (the 5′ to 3′ direction) of the DNA (except that uracil replaces thymine). For example:

Sense strand of DNA	5′... CAT	CGT	TTG	ACC	GAT ... 3′
Antisense strand of DNA	3′... GTA	GCA	AAC	TGG	CTA ... 5′
	⇓ Transcription of antisense strand				
mRNA	5′... CAU	CGU	UUG	ACC	GAU ... 3′
	⇓ Translation of mRNA				
Peptide	... His —	Arg —	Leu —	Thr —	Asp ...

Because the synthesis of mRNA proceeds in the 5′ to 3′ direction, the codons for the sense strand of DNA (with the exception of thymine replacing uracil) are the same as those for the mRNA. For example, one DNA codon for valine is GTA. The corresponding mRNA codon for valine is GUA.

25.5E Translation

We are now in a position to see how the synthesis of a hypothetical polypeptide might take place. This process is called **translation**. Let us imagine that a long strand of mRNA is in the cytoplasm of a cell and that it is in contact with ribosomes. Also in the cytoplasm are the 20 different amino acids, each acylated to its own specific tRNA.

As shown in Fig. 25.16, a tRNA bearing fMet uses its anticodon to associate with the proper codon (AUG) on that portion of mRNA that is in contact with a ribosome. The next triplet of bases on the mRNA chain in this figure is AAA; this is the codon that specifies lysine. A lysyl-tRNA with the matching anticodon UUU attaches itself to this site. The two amino acids, fMet and Lys, are now in the proper position for the 50S ribosome subunit to catalyze the formation of an amide bond between them, as shown in Fig. 25.16 (by the mechanism in Fig. 25.14). After this happens, the ribosome moves down the chain so that it is in contact with the next codon. This one, GUA, specifies valine. A tRNA bearing valine (and with the proper anticodon) binds itself to this site. Another peptide bond–forming reaction takes place attaching valine to the polypeptide chain. Then the whole process repeats itself again and again. The ribosome moves along the mRNA chain, other tRNAs move up with their amino acids, new peptide bonds are formed, and the polypeptide chain grows. At some point an enzymatic reaction removes fMet from the beginning of the chain. Finally, when the chain is the proper length, the ribosome reaches a punctuation mark, UAA, saying "stop here." The ribosome separates from the mRNA chain and so, too, does the protein.

Even before the polypeptide chain is fully grown, it begins to form its own specific secondary and tertiary structure. This happens because its primary structure is correct—its amino acids are ordered in just the right way. Hydrogen bonds form, giving rise to specific segments of α helix, pleated sheet, and coil or loop. Then the whole chain folds and bends; enzymes install disulfide linkages, so that when the chain is fully grown, the whole protein has just the shape it needs to do its job. (Predicting 2° and 3° protein

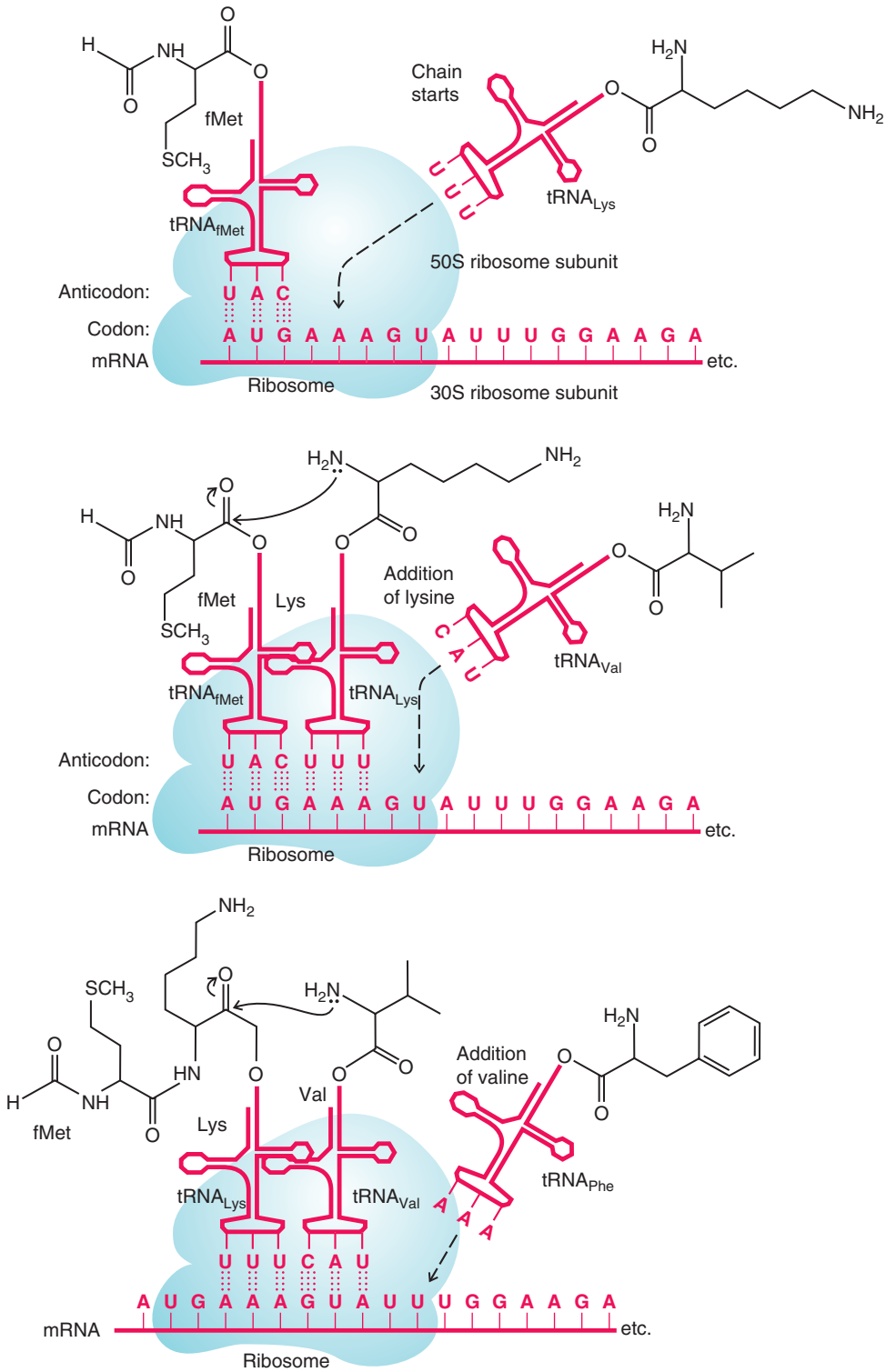


FIGURE 25.16 Step-by-step growth of a polypeptide chain with mRNA acting as a template. Transfer RNAs carry amino acid residues to the site of mRNA that is in contact with a ribosome. Codon–anticodon pairing occurs between mRNA and RNA at the ribosomal surface. An enzymatic reaction joins the amino acid residues through an amide linkage. After the first amide bond is formed, the ribosome moves to the next codon on mRNA. A new tRNA arrives, pairs, and transfers its amino acid residue to the growing peptide chain, and so on.

structure from amino acid sequence, however, remains a critical problem in structural biochemistry.)

In the meantime, other ribosomes nearer the beginning of the mRNA chain are already moving along, each one synthesizing another molecule of the polypeptide. The

time required to synthesize a protein depends, of course, on the number of amino acid residues it contains, but indications are that each ribosome can cause 150 peptide bonds to be formed each minute. Thus, a protein, such as lysozyme, with 129 amino acid residues requires less than a minute for its synthesis. However, if four ribosomes are working their way along a single mRNA chain, a protein molecule can be produced every 13 s.

But why, we might ask, is all this protein synthesis necessary—particularly in a fully grown organism? The answer is that proteins are not permanent; they are not synthesized once and then left intact in the cell for the lifetime of the organism. They are synthesized when and where they are needed. Then they are taken apart, back to amino acids; enzymes disassemble enzymes. Some amino acids are metabolized for energy; others—new ones—come in from the food that is eaten, and the whole process begins again.

PRACTICE PROBLEM 25.10 The sense strand of a segment of DNA has the following sequence of bases:

5′…T G G G G G T T T T A C A G C…3′

(a) What mRNA sequence would result from this segment?

(b) Assume that the first base in this mRNA is the beginning of a codon. What order of amino acids would be translated into a polypeptide synthesized along this segment?

(c) Give anticodons for each tRNA associated with the translation in part (b).

PRACTICE PROBLEM 25.11 **(a)** Using the first codon given for each amino acid in Table 25.2, write the base sequence of mRNA that would translate the synthesis of the following pentapeptide:

Arg · Ile · Cys · Tyr · Val

(b) What base sequence in the DNA sense strand would correspond with this mRNA?

(c) What anticodons would appear in the tRNAs involved in the pentapeptide synthesis?

SOLVED PROBLEM 25.2

Explain how an error of a single base in each strand of DNA could bring about the amino acid error that causes sickle-cell anemia (see "The Chemistry of…" box in Section 24.6B).

STRATEGY AND ANSWER: A change from GAA to GTA in DNA would lead to a change in mRNA from GAA to GUA (see Table 25.2). This change would result in the glutamic acid residue at position 6 in normal hemoglobin becoming valine (as it is in persons with sickle-cell anemia). Alternatively, a change from GAG to GTG in DNA would lead to a change in mRNA from GAG to GUG that would also result in valine replacing glutamic acid.

25.6 DETERMINING THE BASE SEQUENCE OF DNA: THE CHAIN-TERMINATING (DIDEOXYNUCLEOTIDE) METHOD

Certain aspects of the strategy used to sequence DNA resemble the methods used to sequence proteins. Both types of molecules require methods amenable to lengthy polymers, but in the case of DNA, a single DNA molecule is so long that it is absolutely

necessary to cleave it into smaller, manageable fragments. Another similarity between DNA and proteins is that small sets of molecular building blocks comprise the structures of each, but in the case of DNA, only four nucleotide monomer units are involved instead of the 20 amino acid building blocks used to synthesize proteins. Finally, both proteins and nucleic acids are charged molecules that can be separated on the basis of size and charge using chromatography.

The first part of the process is accomplished by using enzymes called **restriction endonucleases**. These enzymes cleave double-stranded DNA at specific base sequences. Several hundred restriction endonucleases are now known. One, for example, called *Alu*I, cleaves the sequence AGCT between G and C. Another, called *Eco*R1, cleaves GAATTC between G and A. Most of the sites recognized by restriction enzymes have sequences of base pairs with the same order in both strands when read from the 5′ direction to the 3′ direction. For example,

$$5' \leftarrow G—A—A—T—T—C \rightarrow 3'$$
$$3' \leftarrow C—T—T—A—A—G \rightarrow 5'$$

Such sequences are known as **palindromes**. (Palindromes are words or sentences that read the same forward or backward. Examples are "radar" and "Madam, I'm Adam.")

Sequencing of the fragments (often called restriction fragments) can be done chemically or with the aid of enzymes. The first chemical method was introduced by A. Maxam and W. Gilbert (both of Harvard University); the **chain-terminating (dideoxynucleotide) method** was introduced in the same year by F. Sanger (Cambridge University). Essentially all DNA sequencing is currently done using an automated version of the chain-terminating method, which involves enzymatic reactions and 2′,3′-dideoxynucleotides.

GILBERT AND SANGER shared the Nobel Prize in Chemistry in 1980 with Paul Berg for their work on nucleic acids. Sanger (Section 24.5B), who pioneered the sequencing of proteins, had won an earlier Nobel Prize in 1958 for the determination of the structure of insulin.

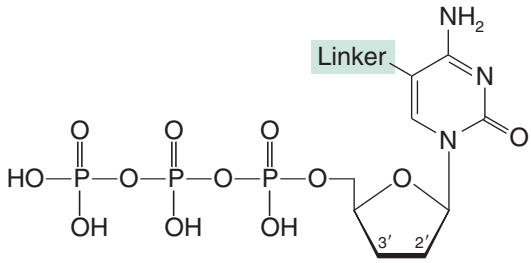
25.6A DNA Sequencing by the Chain-Terminating (Dideoxynucleotide) Method

The chain-terminating method for sequencing DNA involves replicating DNA in a way that generates a family of partial copies that differ in length by one base pair. These partial copies of the parent DNA are separated according to length, and the terminal base on each strand is detected by a covalently attached fluorescent marker.

The mixture of partial copies of the target DNA is made by "poisoning" a replication reaction with a low concentration of unnatural nucleotides. The unnatural terminating nucleotides are 2′,3′-dideoxy analogues of the four natural nucleotides. Lacking the 3′-hydroxyl, each 2′,3′-dideoxynucleotide incorporated is incapable of forming a phosphodiester bond between its 3′ carbon and the next nucleotide that would be needed to continue the polymerization, and hence the chain terminates. Because a low concentration of the dideoxynucleotides is used, only occasionally is a dideoxynucleotide incorporated at random into the growing chains, and thus DNA molecules of essentially all different lengths are synthesized from the parent DNA.

Each terminating dideoxynucleotide is labeled with a fluorescent dye that gives a specific color depending on the base carried by that terminating nucleotide. (An alternate method is to label the *primer*, a short oligonucleotide sequence used to initiate replication of the specific DNA, with specific fluorescent dyes, instead of the dideoxynucleotide terminators, but the general method is the same.) One of the dye systems in use (patented by ABI) consists of a donor chromophore that is initially excited by the laser and which then transfers its energy to an acceptor moiety which produces the observed fluorescence. The donor is tethered to the dideoxynucleotide by a short linker.

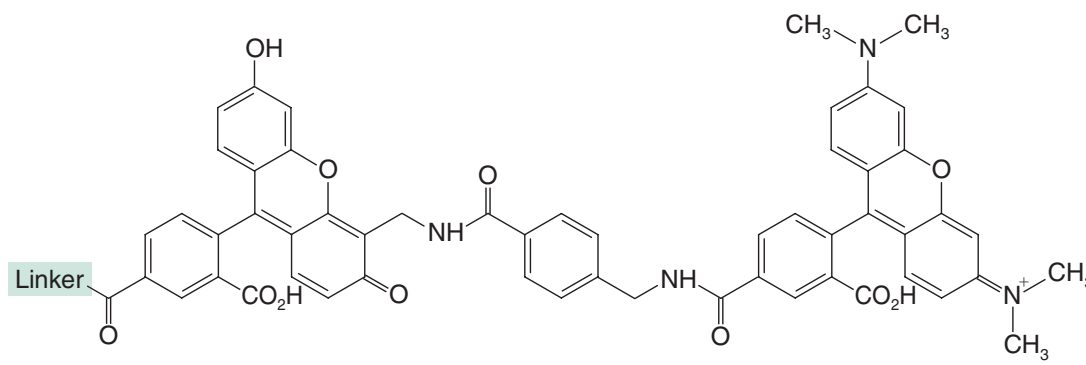
A 2′3′-Dideoxynucleotide, Linker, and Fluorescent Dye Moiety Like Those Used in Fluorescence-Tagged Dideoxynucleotide DNA Sequencing Reactions



Linker between nucleotide and dye

2′,3′-Dideoxycytosine triphosphate

Lack of 3′-hydroxyl causes chain termination.



Fluorescence donor moiety

Fluorescence acceptor moiety

Donor–acceptor fluorescent dye

The replication reaction used to generate the partial DNA copies is similar but not identical to the polymerase chain reaction (PCR) method (Section 25.8). In the dideoxy sequencing method only one primer sequence of DNA is used, and hence only one strand of the DNA is copied, whereas in the PCR, two primers are used and both strands are copied simultaneously. Furthermore, in sequencing reactions the chains are deliberately terminated by addition of the dideoxy nucleotides.

Capillary electrophoresis is the method most commonly used to separate the mixture of partial DNAs that results from a sequencing reaction. Capillary electrophoresis separates the DNAs on the basis of size and charge, allowing nucleotides that differ by only one base length to be resolved. Computerized acquisition of fluorescence data as the differently terminated DNAs pass the detector generates a four-color chromatogram, wherein each consecutive peak represents a DNA molecule one nucleotide longer than the previous one. The color of each peak represents the terminating nucleotide in that molecule. Since each of the four types of dideoxy terminating bases fluoresces a different color, the sequence of nucleotides in the DNA can be read directly. An example of sequence data from this kind of system is shown in Fig. 25.17.

Use of automated methods for DNA sequencing represents an exponential increase in speed over manual methods employing vertical slab polyacrylamide gel electrophoresis (Fig. 25.18). Only a few thousand bases per day (at most) could be sequenced by a person using the manual method. Now it is possible for a single machine running parallel and continuous analyses to sequence almost 3 million bases per day using automated capillary electrophoresis and laser fluorescence detection. As an added benefit, the ease of DNA sequencing often makes it easier to determine the sequence of a protein by the sequence of all or part of its corresponding gene, rather than by sequencing the protein itself (see Section 24.5).

The development of high-throughput methods for sequencing DNA is largely responsible for the remarkable success achieved in the Human Genome Project. Sequencing the 3 billion base pairs in the human genome could never have been completed before 2003 and the 50th anniversary of Watson and Crick's elucidation of the structure of DNA had high-throughput sequencing methods not come into existence.

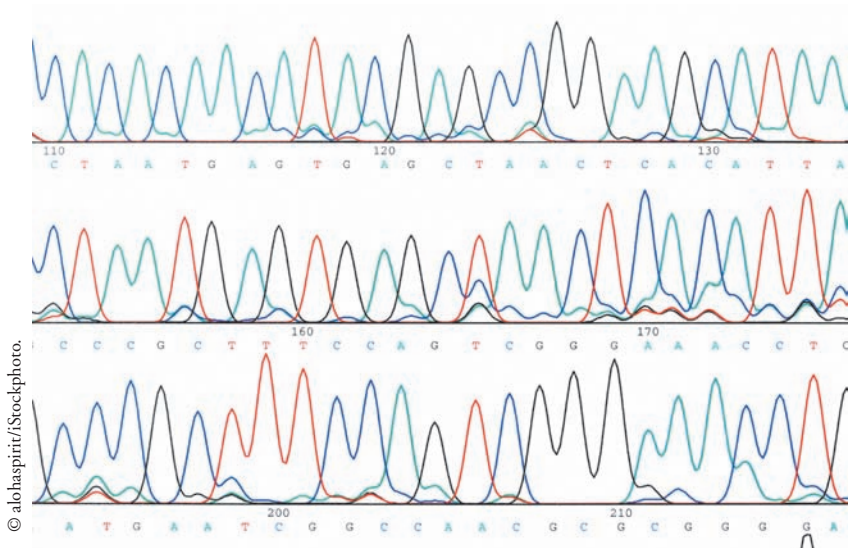


FIGURE 25.17 Example of data from an automated DNA sequencer.

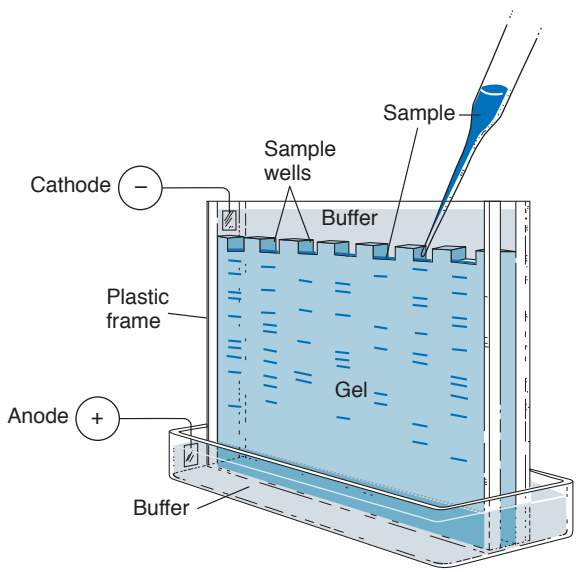


FIGURE 25.18 An apparatus for gel electrophoresis. Samples are applied in the slots at the top of the gel. Application of a voltage difference causes the samples to move. The samples move in parallel lanes. (Reprinted with permission of John Wiley & Sons, Inc., from Voet, D. and Voet, J. G., *Biochemistry*, Second Edition. © 1995 Voet, D. and Voet, J. G.)

25.7 LABORATORY SYNTHESIS OF OLIGONUCLEOTIDES

Synthetic oligonucleotides are needed for a variety of purposes. One of the most important and common uses of synthetic oligonucleotides is as primers for nucleic acid sequencing and for PCR (Section 25.8). Another important application is in the research and development of **antisense oligonucleotides**, which hold potential as therapies for a variety of diseases. An antisense oligonucleotide is one that has a sequence complementary to the coding sequence in a DNA or RNA molecule. Synthetic oligonucleotides that bind tightly to DNA or mRNA sequences from a virus, bacterium, or other disease condition may be able to shut down expression of the target protein associated with those conditions. For example, if the sense portion of DNA in a gene reads

A—G—A—C—C—G—T—G—G

the antisense oligonucleotide would read

T—C—T—G—G—C—A—C—C

The ability to deactivate specific genes in this way holds great medical promise. Many viruses and bacteria, during their life cycles, use a method like this to regulate some of their own genes. The hope, therefore, is to synthesize antisense oligonucleotides that will seek out and destroy viruses in a person's cells by binding with crucial sequences of the viral DNA or RNA. Synthesis of such oligonucleotides is an active area of research today and is directed at many viral diseases, including AIDS, as well as lung and other forms of cancer.

Current methods for **oligonucleotide synthesis** are similar to those used to synthesize proteins, including the use of automated solid-phase techniques (Section 24.7D). A suitably protected nucleotide is attached to a solid phase called a "controlled pore glass," or CPG (Fig. 25.19), through a linkage that can ultimately be cleaved. The next protected nucleotide in the form of a **phosphoramidite** is added, and coupling is brought about by a coupling agent, usually 1,2,3,4-tetrazole. The phosphite triester that results from the

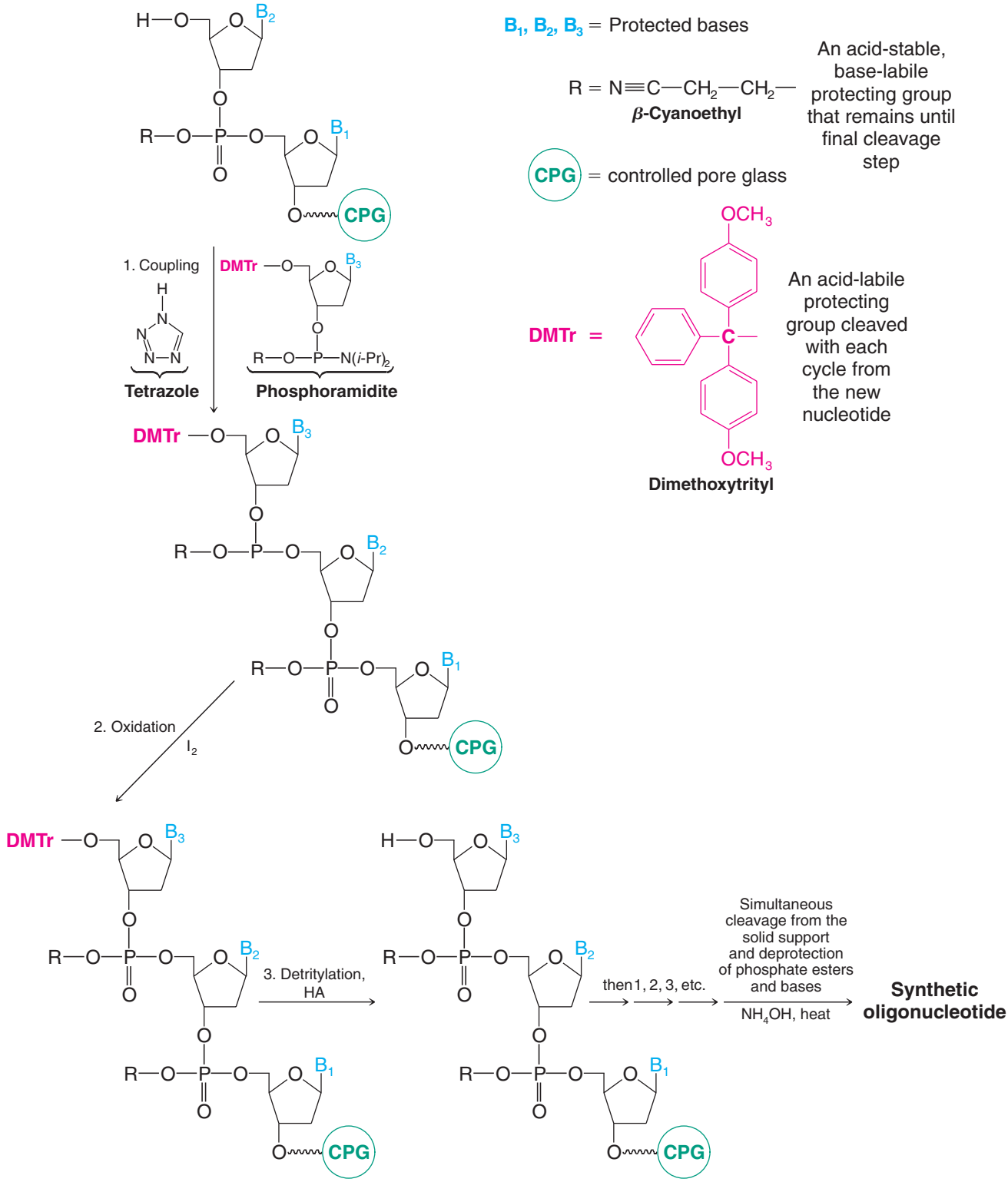


FIGURE 25.19 The steps involved in automated synthesis of oligonucleotides using the phosphoramidite coupling method.

coupling is oxidized to phosphate triester with iodine, producing a chain that has been lengthened by one nucleotide. The **dimethoxytrityl (DMTr)** group used to protect the 5′ end of the added nucleotide is removed by treatment with acid, and the steps **coupling, oxidation, detritylation**, as shown in Figure 25.19, are repeated. (All the steps are carried out in nonaqueous solvents.) With automatic synthesizers the process can be repeated at least 50 times and the time for a complete cycle is 40 min or less. The synthesis is monitored by spectrophotometric detection of the dimethoxytrityl cation as it is released in each cycle (much like the monitoring of Fmoc release in solid-phase peptide synthesis). After the desired oligonucleotide has been synthesized, it is released from the solid support and the various protecting groups, including those on the bases, are removed.

25.8 THE POLYMERASE CHAIN REACTION

The **polymerase chain reaction (PCR)** is an extraordinarily simple and effective method for exponentially multiplying (amplifying) the number of copies of a DNA molecule. Beginning with even just a single molecule of DNA, the PCR can generate 100 billion copies in a single afternoon. The reaction is easy to carry out: It requires only a miniscule sample of the target DNA (picogram quantities are sufficient), a supply of nucleotide triphosphate reagents and primers to build the new DNA, DNA polymerase to catalyze the reaction, and a device called a thermal cycler to control the reaction temperature and automatically repeat the reaction. The PCR has had a major effect on molecular biology. Perhaps its most important role has been in the sequencing of the human genome (Sections 25.6 and 25.9), but now virtually every aspect of research involving DNA involves the PCR at some point.

One of the original aims in developing the PCR was to use it in increasing the speed and effectiveness of prenatal diagnosis of sickle-cell anemia (Section 24.6B). It is now being applied to the prenatal diagnosis of a number of other genetic diseases, including muscular dystrophy and cystic fibrosis. Among infectious diseases, the PCR has been used to detect cytomegalovirus and the viruses that cause AIDS, certain cervical carcinomas, hepatitis, measles, and Epstein–Barr disease.

The PCR is a mainstay in forensic sciences as well, where it may be used to copy DNA from a trace sample of blood or semen or a hair left at the scene of a crime. It is also used in evolutionary biology and anthropology, where the DNA of interest may come from a 40,000-year-old woolly mammoth or the tissue of a mummy. It is also used to match families with lost relatives (see the chapter opening vignette). There is almost no area with biological significance that does not in some way have application for use of the PCR reaction.

The PCR was invented by Kary B. Mullis and developed by him and his co-workers at Cetus Corporation. It makes use of the enzyme DNA polymerase, discovered in 1955 by Arthur Kornberg and associates at Stanford University. In living cells, DNA polymerases help repair and replicate DNA. The PCR makes use of a particular property of DNA polymerases: their ability to attach additional nucleotides to a short oligonucleotide "primer" when the primer is bound to a complementary strand of DNA called a template. The nucleotides are attached at the 3′ end of the primer, and the nucleotide that the polymerase attaches will be the one that is complementary to the base in the adjacent position on the template strand. If the adjacent template nucleotide is G, the polymerase adds C to the primer; if the adjacent template nucleotide is A, then the polymerase adds T, and so on. Polymerase repeats this process again and again as long as the requisite nucleotides (as triphosphates) are present in the solution, until it reaches the 5′ end of the template.


Mullis was awarded the Nobel Prize in Chemistry for this work in 1993.

Figure 25.20 shows one PCR cycle. The target DNA, a supply of nucleotide triphosphate monomers, DNA polymerase, and the appropriate oligonucleotide primers (one primer sequence for each 5′ to 3′ direction of the target double-stranded DNA) are added to a small reaction vessel. The mixture is briefly heated to approximately 90 °C to separate the DNA strands (denaturation); it is cooled to 50–60 °C to allow the primer sequences and DNA polymerase to bind to each of the separated strands (annealing); and it is warmed to about 70 °C to extend each strand by polymerase-catalyzed condensation of nucleotide triphosphate monomers complementary to the parent DNA strand. Another cycle of the PCR begins by heating to separate the new collection of DNA molecules into single strands, cooling for the annealing step, and so on.

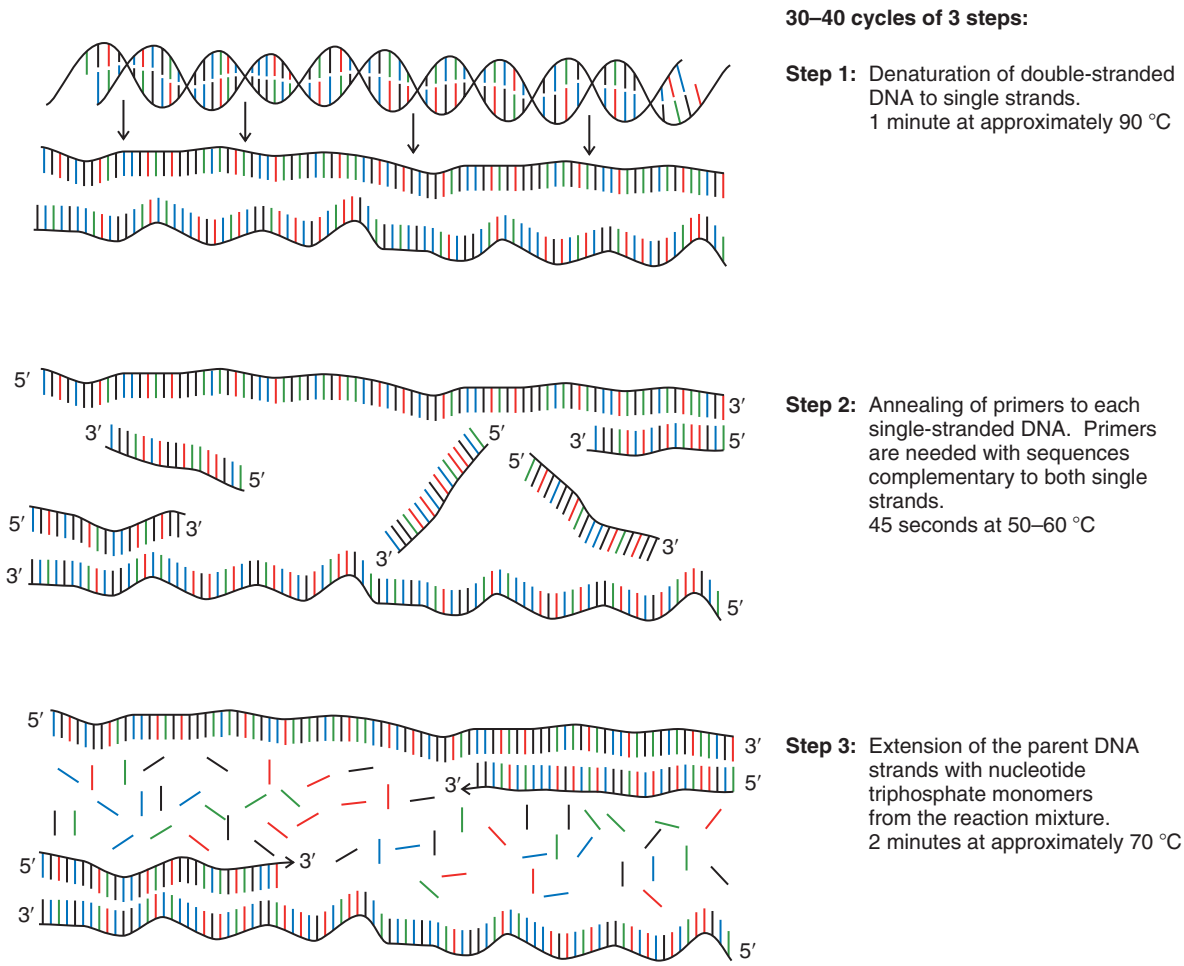


FIGURE 25.20 One cycle of the PCR. Heating separates the strands of DNA of the target to give two single-stranded templates. Primers, designated to complement the nucleotide sequences flanking the targets, anneal to each strand. DNA polymerase, in the presence of nucleotide triphosphates, catalyzes the synthesis of two pieces of DNA, each identical to the original target DNA. *(Used with permisson from Andy Vierstraete, University of Ghent.)*

Each cycle, taking only a few minutes, doubles the amount of target DNA that existed prior to that step (Fig. 25.21). The result is an exponential increase in the amount of DNA over time. After n cycles, the DNA will have been replicated 2^n times—after 10 cycles there is roughly 1000 times as much DNA; after 20 cycles roughly 1 million times as much; and so on. Thermal cycling machines can carry out approximately 20 PCR cycles per hour, resulting in billions of DNA copies over a single afternoon.

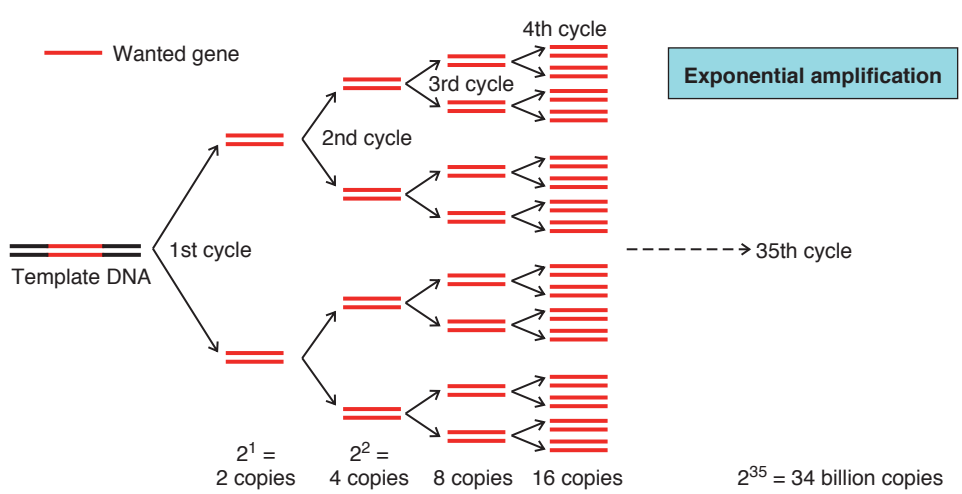


FIGURE 25.21 Each cycle of the PCR doubles the number of copies of the target area. *(Used with permisson from Andy Vierstraete, University of Ghent.)*

Each application of PCR requires primers that are 10–20 nucleotides in length and whose sequences are complementary to short, conveniently located sequences flanking the target DNA sequence. The primer sequence is also chosen so that it is near sites that are cleavable with restriction enzymes. Once a researcher determines what primer sequence is needed, the primers are usually purchased from commercial suppliers who synthesize them on request using solid-phase oligonucleotide synthesis methods like that described in Section 25.7.

As an intriguing adjunct to the PCR story, it turns out that cross-fertilization between disparate research fields greatly assisted development of current PCR methods. In particular, the discovery of extremozymes, which are enzymes from organisms that live in high-temperature environments, has been of great use. DNA polymerases now typically used in PCR are heat-stable forms derived from thermophilic bacteria. Polymerases such as Taq polymerase, from the bacterium *Thermus aquaticus*, found in places such as geyser hot springs, and $Vent_R^{TM}$, from bacteria living near deep-sea thermal vents, are used. Use of extremozyme polymerases facilitates PCR by allowing elevated temperatures to be used for the DNA melting step without having to worry about denaturing the polymerase enzyme at the same time. All materials can therefore be present in the reaction mixture throughout the entire process. Furthermore, use of a higher temperature during the chain extension also leads to faster reaction rates. (See "The Chemistry of…Stereoselective Reductions of Carbonyl Groups," Section 12.3C, for another example of the use of high-temperature enzymes.)

Simon Terry/Photo Researchers, Inc.

Thermophilic bacteria, growing in hot springs like these at Yellowstone National Park, produce heat-stable enzymes called extremozymes that have proved useful for a variety of chemical processes.

25.9 SEQUENCING OF THE HUMAN GENOME: AN INSTRUCTION BOOK FOR THE MOLECULES OF LIFE

The announcement by scientists from the public Human Genome Project and Celera Genomics Company in June 2000 that sequencing of the approximately 3 billion base pairs in the human genome was complete marked the achievement of one of the most important and ambitious scientific endeavors ever undertaken. To accomplish this feat, data were pooled from thousands of scientists working around the world using tools including PCR (Section 25.8), dideoxynucleotide sequencing reactions (Section 25.6), capillary electrophoresis, laser-induced fluorescence, and supercomputers. What was ultimately produced is a transcript of our chromosomes that could be called an instruction book for the molecules of life.

"This structure has novel features, which are of considerable biological importance." James Watson, one of the scientists who determined the structure of DNA.

But what do the instructions in the genome say? How can we best make use of the molecular instructions for life? Of the roughly 35,000 genes in our DNA, the function of only a small percentage of genes is understood. Discovering genes that can be used to benefit our human condition and the chemical means to turn them on or off presents some of the greatest opportunities and challenges for scientists of today and the future. Sequencing the genome was only the beginning of the story.

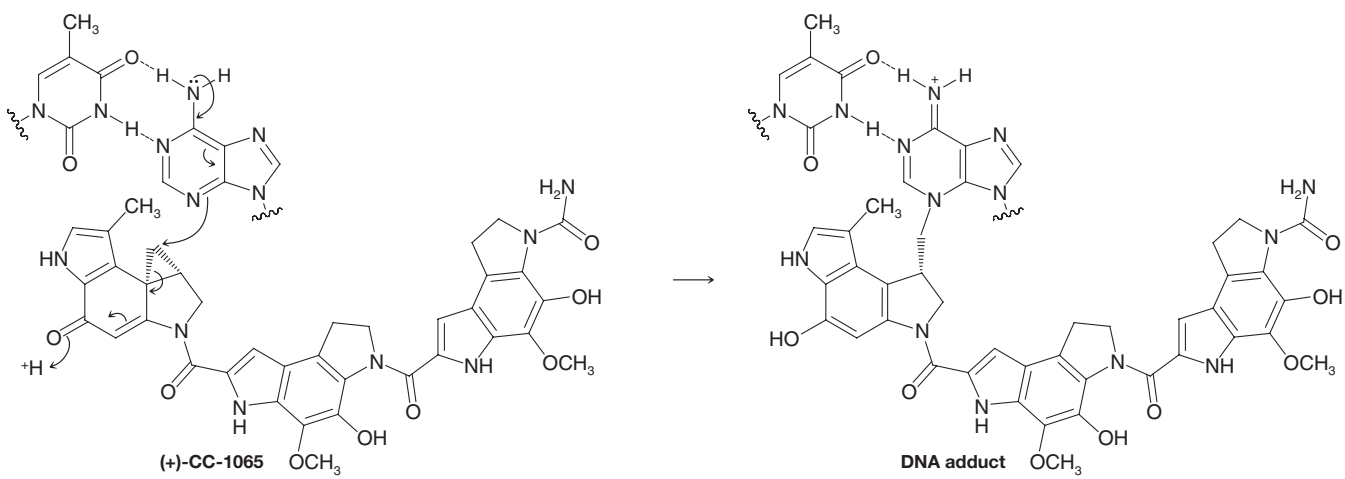
As the story unfolds, chemists will continue to add to the molecular archive of compounds used to probe our DNA. DNA microchips, with 10,000 or more short diagnostic sequences of DNA chemically bonded to their surface in predefined arrays, will be used to test DNA samples for thousands of possible genetic conditions in a single assay. With the map of our genome in hand, great libraries of potential drugs will be tested against genetic targets to discover more molecules that either promote or inhibit expression of key gene products. Sequencing of the genome will also accelerate development of molecules that interact with proteins, the products of gene expression. Knowledge of

the genome sequence will expedite identification of the genes coding for interesting proteins, thus allowing these proteins to be expressed in virtually limitless quantities. With an ample supply of target proteins available, the challenges of solving three-dimensional protein structures and understanding their functions will also be overcome more easily. Optimization of the structures of small organic molecules that interact with proteins will also occur more rapidly because the protein targets for these molecules will be available faster and in greater quantity. There is no doubt that the pace of research to develop new and useful organic molecules for interaction with gene and protein targets will increase dramatically now that the genome has been sequenced. The potential to use our chemical creativity in the fields of genomics and proteomics is immense.

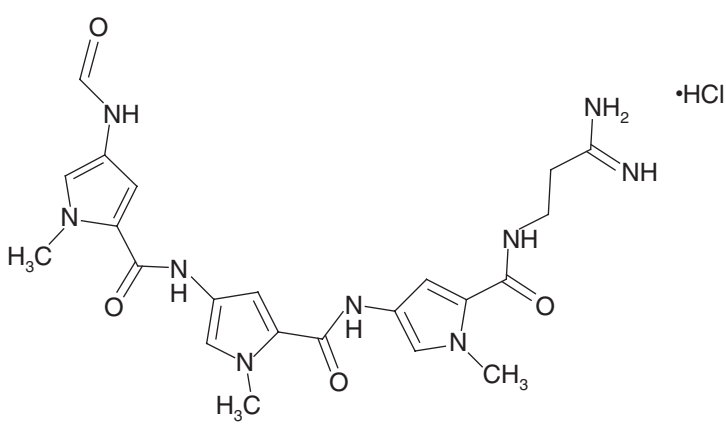
WHY Do These Topics Matter?]

SELECTIVELY TARGETING A DNA SEQUENCE

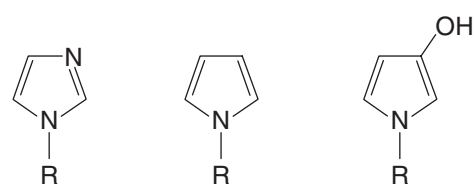
Just as specific hydrogen bonds are the basis for the base-pairing of A with T and C with G in a DNA molecule, hydrogen bonds can also be used to bind small molecules to specific DNA sequences. The result is often a significant biochemical outcome. For instance, the unique cyclopropane-containing natural product (+)-CC-1065 possesses helicity that allows it to line up with the edge of the DNA minor groove. Once it encounters a domain rich in AT sequences, it can serve as an electrophile in an S_N2 reaction that leads to alkylation of an adenine residue as shown below. As a result, the cell is eventually destroyed. Thus, (+)-CC-1065 can serve as an antitumor agent if the target is a cancerous cell.



This mechanism of action is not unique. In fact, there are several natural products that can similarly "read" the edge of the minor groove of DNA through hydrogen bond interactions. One such compound is distamycin A, which, as shown below, contains three peptide bond–linked pyrroles that can also target AT-rich regions of DNA.



Distamycin



Imidazole **Pyrrole** **Hydroxypyrrole**

This molecule, however, has proved highly significant in that it served as the main inspiration for chemists to go beyond natural products and develop a set of molecules with the power to literally read, or differentiate, not only AT-rich sequences, but any specific DNA target sequence desired. The leader of these efforts has been Peter Dervan of the California Institute of Technology. Over a period spanning two decades, his team developed a group of molecules reminiscent of distamycin that contain pyrrole, imidazole, and hydroxypyrrole rings in two separate domains linked by a flexible tether. These compounds can read any of the main DNA sequences as shown on the previous page based on the use of A, C, G, and T as the purines and pyrimidines comprising the backbone structure, and you can learn more about their specific structures in reference 2 cited below.

To give a sense of the significance of this finding, a DNA segment containing 8 base pairs has 32,896 different possible sequences. Rather than having to identify thousands of different and distinct solutions for selectively targeting such an array sequences, this solution provides a common, predictable system that can be easily tailored to target any of these possible sequences at will simply by changing the positioning of these three heterocyclic systems within the two arms of the molecules. Current work is directed toward determining whether drugs can be combined with such sequence-specific molecules to provide novel treatments. While time will tell if new medicines will result, for now it is satisfying to see how natural products, first chemical principles like hydrogen bonding, and thoughtful molecular design can be combined to do something that even nature does not appear to be able to accomplish!

To learn more about these topics, see:

1. Boger, D. L.; Johnson, D. S. "CC-1065 and the duocarmycins: Unraveling the keys to a new class of naturally derived DNA alkylating agents." *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 3642–3649.
2. Dervan, P. B. "Molecular recognition of DNA by small molecules." *Bioorg. Med. Chem*. **2001**, *9*, 2215–2135.

SUMMARY AND REVIEW TOOLS

The study aids for this chapter include key terms and concepts, which are highlighted in bold, blue text within the chapter, defined in the Glossary at the back of the book, and which have hyperlinked definitions in the accompanying *WileyPLUS* course (www.wileyplus.com).

PROBLEMS

Note to Instructors: Many of the homework problems are available for assignment via *WileyPLUS*, an online teaching and learning solution.

NUCLEIC ACID STRUCTURE

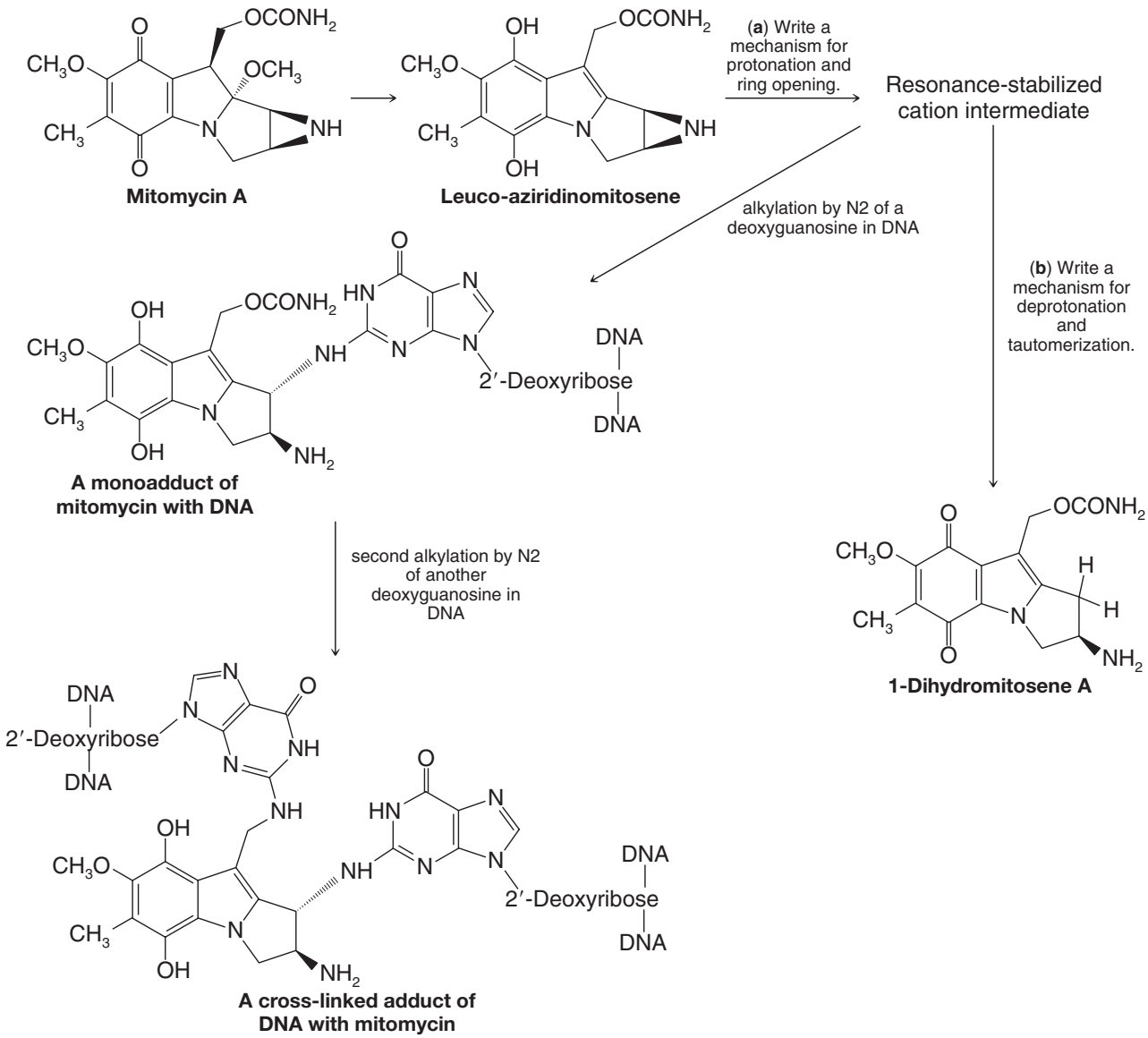
25.12 Write the structure of the RNA dinucleotide G–C in which G has a free 5′-hydroxyl group and C has a free 3′-hydroxyl group.

25.13 Write the structure of the DNA dinucleotide T–A in which T has a free 5′-hydroxyl group and A has a free 3′-hydroxyl group.

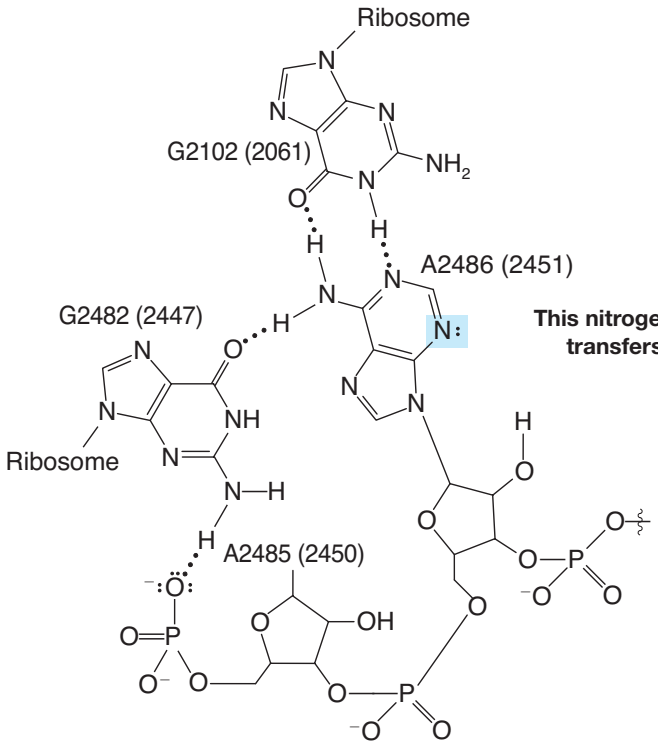
MECHANISMS

25.14 The example of a silyl–Hilbert–Johnson nucleosidation reaction in Section 25.3 is presumed to involve an intermediate ribosyl cation that is stabilized by intramolecular interactions involving the C2 benzoyl group. This intermediate blocks attack by the heterocyclic base from the α face of the ribose ring but allows attack on the β face, as required for formation of the desired product. Propose a structure for the ribosyl cation intermediate that explains the stereoselective bonding of the base.

25.15 **(a)** Mitomycin is a clinically used antitumor antibiotic that acts by disrupting DNA synthesis through covalent bond-forming reactions with deoxyguanosine in DNA. Maria Tomasz (Hunter College) and others have shown that alkylation of DNA by mitomycin occurs by a complex series of mechanistic steps. The process begins with reduction of the quinone ring in mitomycin to its hydroquinone form, followed by elimination of methanol from the adjacent ring to form an intermediate called leuco-aziridinomitosene. One of the paths by which leuco-aziridinomitosene alkylates DNA involves protonation and opening of the three-membered aziridine ring, resulting in an intermediate cation that is resonance stabilized by the hydroquinone group. Attack on the cation by N2 of a deoxyguanosine residue leads to a monoalkylated DNA product, as shown in the scheme. Write a detailed mechanism to show how the ring opening might occur, including resonance forms for the cation intermediate, followed by nucleophilic attack by DNA. (Intra- or interstrand cross-linking of DNA can further occur by reaction of another deoxyguanosine residue to displace the carbamoyl group of the initial mitosene–DNA monoadduct. A cross-linked adduct is also shown.) **(b)** 1-Dihydromitosene A is sometimes formed from the cation intermediate in part (a) by loss of a proton and tautomerization. Propose a detailed mechanism for the formation of 1-dihydromitosene A from the resonance-stabilized cation of part (a).



25.16 As described in Section 25.5B, acid–base catalysis is believed to be the mechanism by which ribosomes catalyze the formation of peptide bonds in the process of protein translation. Key to this proposal is assistance by the N3 nitrogen (highlighted in the scheme on the next page) of a nearby adenine in the ribosome for the removal of a proton from the α-amino group of the amino acid adding to the growing peptide chain (Fig. 25.14). The ability of this adenine group to remove the proton is, in turn, apparently facilitated by relay of charge made possible by other nearby groups in the ribosome. The constellation of these groups is shown in the scheme. Draw mechanism arrows to show formation of a resonance contributor wherein the adenine group could carry a formal negative charge, thereby facilitating its removal of the α-amino proton of the amino acid. (The true electronic structure of these groups is not accurately represented by any single resonance contributor, of course. A hybrid of the contributing resonance structures weighted according to stability would best reflect the true structure.)



LEARNING GROUP PROBLEM

Research suggests that expression of certain genes is controlled by conversion of some cytosine bases in the genome to 5-methylcytosine by an enzyme called DNA methyltransferase. Cytosine methylation may be a means by which some genes are turned off as cells differentiate during growth and development. It may also play a role in some cancer processes and in defending the genome from foreign DNA such as viral genes. Measuring the level of methylation in DNA is an important analytical process. One method for measuring cytosine methylation is known as methylation-specific PCR. This technique requires that all unmethylated cytosines in a sample of DNA be converted to uracil by deamination of the C4 amino group in the unmethylated cytosines. This is accomplished by treating the DNA with sodium bisulfite ($NaHSO_3$) to form a bisulfite addition product with its unmethylated cytosine residues. The cytosine sulfonates that result are then subjected to hydrolysis conditions that convert the C4 amino group to a carbonyl group, resulting in uracil sulfonate. Finally, treatment with base causes elimination of the sulfonate group to produce uracil. The modified DNA is then amplified by PCR using primers designed to distinguish DNA with methylated cytosine from cytosine-to-uracil converted bases.

Write detailed mechanisms for the reactions used to convert cytosine to uracil by the above sequence of steps.

This page intentionally left blank