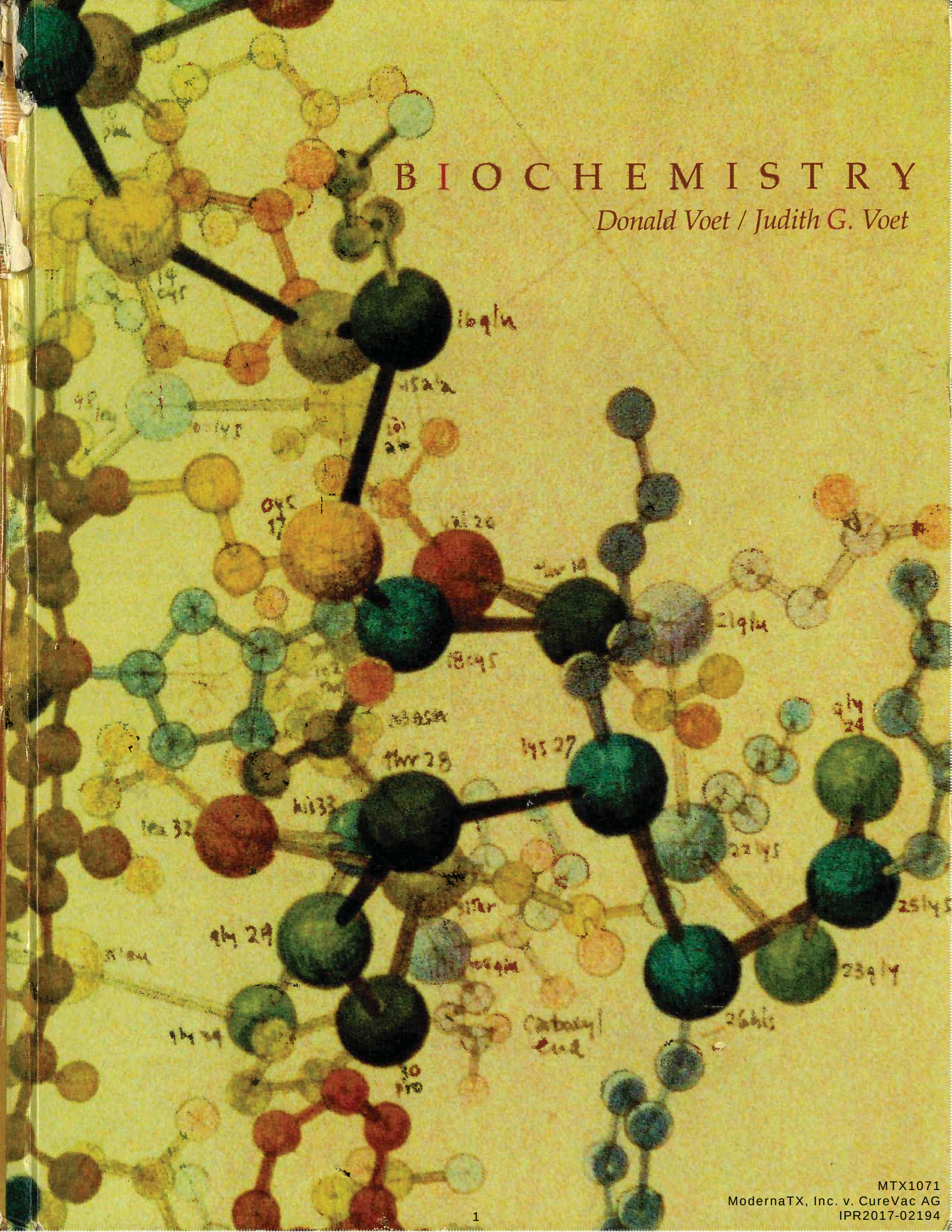


BIOCHEMISTRY

Donald Voet / Judith G. Voet

To:

Our parents, who encouraged us,
Our teachers, who enabled us, and
Our children, who put up with us.

Cover Art: One of a series of color studies of horse heart cytochrome *c* designed to show the influence of amino acid side chains on the protein's three-dimensional folding pattern. We have selected this study to symbolize the discipline of biochemistry: Both are beautiful but still in process and hence have numerous "rough edges." Drawing by Irving Geis in collaboration with Richard E. Dickerson.

Cover and part opening illustrations copyrighted by Irving Geis.

Cover Designer: Madelyn Lesure

Photo Research: John Schultz, Eloise Marion

Photo Research Manager: Stella Kupferberg

Illustration Coordinator: Edward Starr

Copy Editor: Jeannette Stiefel

Production Manager: Lucille Buonocore

Senior Production Supervisor: Linda Muriello

Library of Congress Cataloging in Publication Data:

Voet, Donald.
Biochemistry / by Donald Voet and Judith G. Voet.
p. cm.
Includes bibliographical references.
ISBN 0-471-61769-5
1. Biochemistry. I. Voet, Judith G. II. Title.
QP514.2.V64 1990 89-16727
574.19'2—dc20 CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Chapter 28

NUCLEIC ACID STRUCTURES AND MANIPULATION

There are two classes of nucleic acids, ***deoxyribonucleic acid (DNA)*** *and* ***ribonucleic acid (RNA).*** *DNA is the hereditary molecule in all cellular life forms, as well as in many viruses.* It has but two functions:

1. To direct its own replication during cell division.
2. To direct the **transcription** of complementary molecules of RNA.

RNA, in contrast, has more varied biological functions:

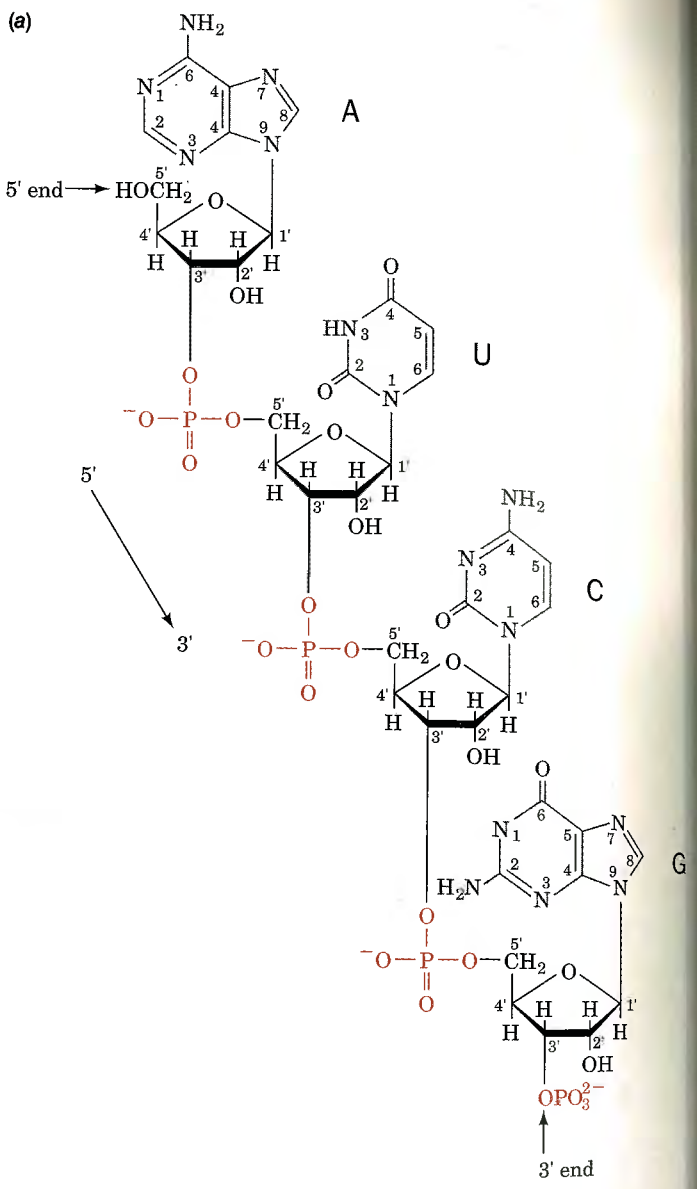
1. The RNA transcripts of DNA sequences that specify polypeptides, **messenger RNA (mRNA),** direct the ribosomal synthesis of these polypeptides in a process known as **translation.**

2. The RNAs of ribosomes, which are about two-thirds RNA and one-third protein, probably have functional as well as structural roles.
3. During protein synthesis, amino acids are delivered to the ribosome by molecules of **transfer RNA (tRNA).**
4. Certain RNAs are associated with specific proteins to form **ribonucleoproteins** that participate in the post-transcriptional processing of other RNAs.
5. In many viruses, RNA, not DNA, is the carrier of hereditary information.

In this chapter we examine the structures of nucleic acids with emphasis on DNA (the structure of RNA is detailed in Section 30-2A), and discuss methods of purifying, sequencing, and chemically synthesizing nucleic acids. We end by outlining how recombinant DNA technology, which has revolutionized the study of biochemistry, is used to manipulate, synthesize, and express DNA.

1. CHEMICAL STRUCTURE AND BASE COMPOSITION

The chemical structures of the nucleic acids were elucidated by the early 1950s largely through the efforts of Phoebus Levine followed by those of Alexander Todd. *Nucleic acids are, with few exceptions, linear polymers of nucleotides whose phosphates bridge the 3′ and 5′ positions of successive sugar residues (e.g., Fig. 28-1).* The phosphates of these **polynucleotides,** the **phosphodiester** groups, are acidic so that, *at physiological pH's, nucleic acids are polyanions.*



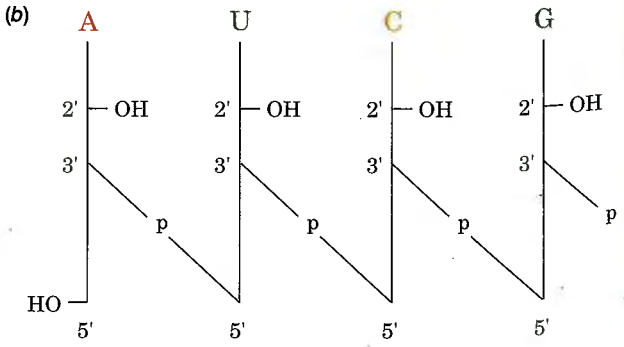


Figure 28-1
(*a*) The tetranucleotide adenyl-3′,5′-uridyl-3′,5′-cytidyl-3′,5′-guanylyl-3′-phosphate. The sugar atom numbers are primed to distinguish them from the atomic positions of the bases. By convention, polynucleotide sequences are written with their 5′ end at the left and their 3′ end to the right. Thus, reading left to right, the phosphodiester bond links neighboring ribose residues in the 5′ → 3′ direction. The above sequence may be abbreviated ApUpCpGp or just AUCGp (where a "p" to the left and/or right of a nucleoside symbol indicates a 5′ and/or a 3′ phosphoryl bond, respectively; see Table 26-1 for other symbol definitions). The corresponding deoxytetranucleotide is abbreviated d(ApUpCpGp) or d(AUCGp). (*b*) A schematic representation of AUCGp. Here a vertical line denotes a ribose residue, its attached base is indicated by the corresponding one letter abbreviation and a diagonal line flanking an optional "p" represents a phosphodiester bond. The atomic numbering of the ribose residues, which is indicated here, is usually omitted. The equivalent representation of deoxypolynucleotides differ only by the absence of the 2′-OH groups.

DNA's Base Composition Is Governed by Chargaff's Rules

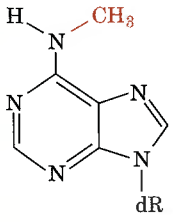
DNA has equal numbers of adenine and thymine residues (A = T) and equal numbers of guanine and cytosine residues (G = C). These relationships, known as **Chargaff's rules,** were discovered in the late 1940s by Erwin Chargaff who first devised reliable quantitative methods for the separation (by paper chromatography) 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 age of the organism, its nutritional state or any other environmental factor. The structural basis of Chargaff's rules derives from DNA's double-stranded character (Section 28-2A).

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%.

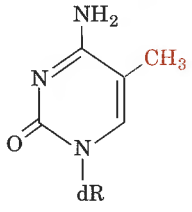
RNA, which usually occurs as a single-stranded molecule, has no apparent constraints on its base composition. However, double-stranded RNA, which comprises the genetic material of several viruses, obeys Chargaff's rules. Conversely, single-stranded DNA, which occurs in certain viruses, does not obey Chargaff's rules. Upon entering its host organism, however, such DNA is replicated to form a double-stranded molecule, which then obeys Chargaff's rules.

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^6-methyl-dA** and **5-methyl-dC,** respectively.



N^6-Methyl-dA



5-Methyl-dC

The altered bases are generated by the sequence specific enzymatic modification of normal DNA (Sections 28-6A and 31-7). The modified DNAs obey Chargaff's rules if the derivatized bases are taken as equivalent to their parent bases. Likewise, many bases in RNA and, in particular, in tRNA (Section 30-2), are derivatized.

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. 28-2 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.

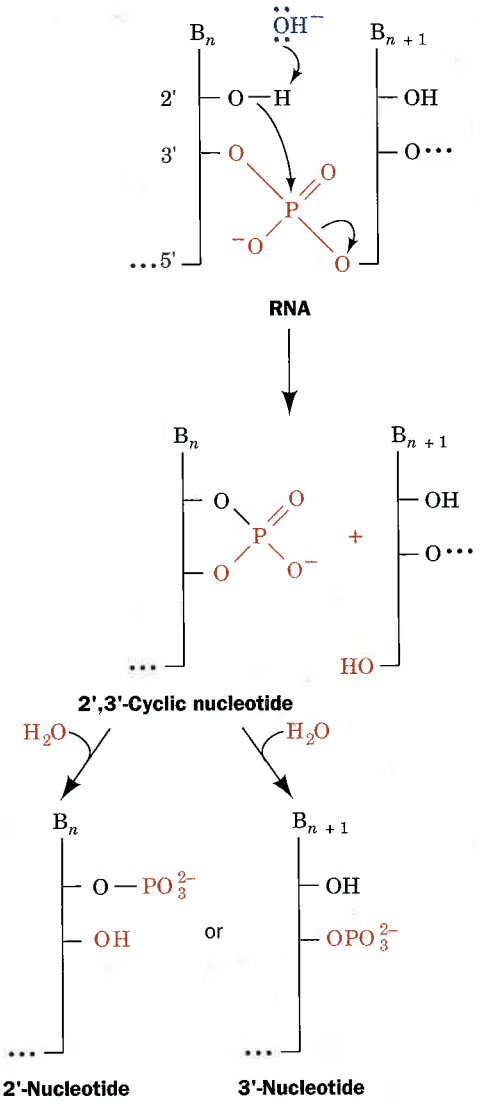


Figure 28-2
The mechanism of base-catalyzed RNA hydrolysis. The base-induced deprotonation of the 2′-OH group facilitates its nucleophilic attack on the adjacent phosphorus atom thereby cleaving the RNA backbone. The resultant 2′,3′-cyclic phosphate group subsequently hydrolyzes to either the 2′ or the 3′ phosphate. Note that the RNase-catalyzed hydrolysis of RNA follows a nearly identical reaction sequence (Section 14-1A).

2. DOUBLE HELICAL STRUCTURES

The determination of the structure of DNA by James Watson and Francis Crick in 1953 is often said to mark

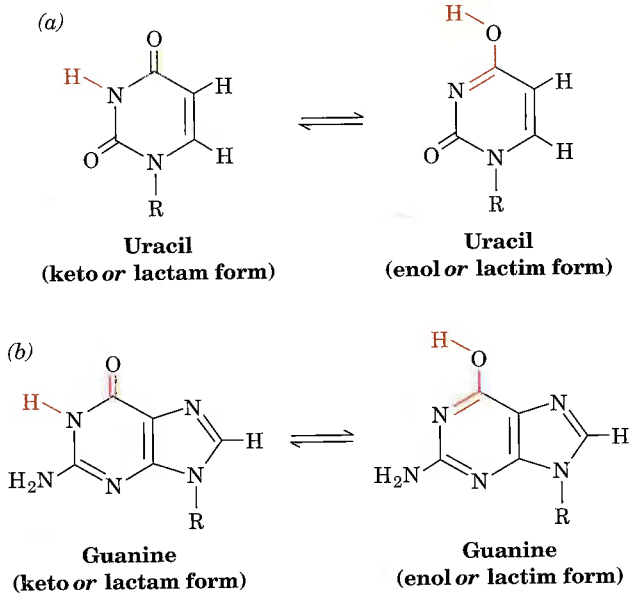


Figure 28-3
Some possible tautomeric conversions for (a) uracil and (b) guanine residues. Cytosine and adenine residues can undergo similar proton shifts.

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, these relationships were quite obscure because their significance was not apparent. In fact, even Chargaff did not emphasize them.

2. The correct tautomeric forms of the bases. X-ray, NMR, and spectroscopic investigations have firmly established that the nucleic acid bases are overwhelmingly in the keto tautomeric forms shown in Fig. 28-1. In 1953, however, this was not generally appreciated. Indeed, guanine and uracil were widely believed to be in their enol forms (Fig. 28-3) because it was thought that the resonance stability of these aromatic molecules would thereby be maximized. Knowledge of the dominant tautomeric forms, which 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. 28-4; DNA, being a threadlike molecule, does not crystallize but, rather, can be drawn out in fibers consisting of parallel bundles of molecules; Section 7-2). A description of the photograph enabled Crick, an X-ray crystallographer by training who had earlier derived the equations describing diffraction by helical molecules, to deduce that DNA is (a) a helical molecule, and (b) that its planar aromatic bases form a stack of parallel rings that 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.

It is now realized that double helical DNA and RNA can assume several distinct structures that vary with such factors as the humidity and the identities of the cations present, as well as with base sequence. In this section, we describe these various structures.

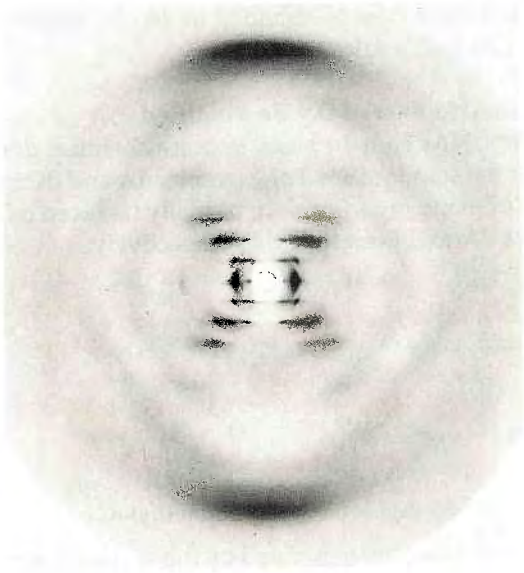


Figure 28-4
An X-ray diffraction photograph of a vertically oriented Na^+ DNA fiber in the B conformation. This is the photograph that provided key information for the elucidation of the Watson–Crick structure. The central X-shaped pattern of spots is indicative of a helix, whereas the heavy black arcs on the top and bottom of the diffraction pattern correspond to a distance of 3.4 Å and indicate that the DNA structure largely repeats every 3.4 Å along the fiber axis. [Courtesy of Maurice Wilkins, King's College, London.]

Table 28-1
Structural Features of Ideal A, B, and Z-DNA

	A	B	Z
Helical sense	Right handed	Right handed	Left handed
Diameter	~26 Å	~20 Å	~18 Å
Base pairs per helical turn	11	10	12 (6 dimers)
Helical twist per base pair	33°	36°	60° (per dimer)
Helix pitch (rise per turn)	28 Å	34 Å	45 Å
Helix rise per base pair	2.6 Å	3.4 Å	3.7 Å
Base tilt normal to the helix axis	20°	6°	7°
Major groove	Narrow and deep	Wide and Deep	Flat
Minor groove	Wide and shallow	Narrow and deep	Narrow and deep
Sugar pucker	C(3′)-*endo*	C(2′)-*endo*	C(2′)-*endo* for pyrimidines; C(3′)-*endo* for purines
Glycosidic bond	Anti	Anti	Anti for pyrimidines; syn for purines

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 Na^+ and the relative humidity is 92%. ***B-DNA is regarded as the native form because its X-ray pattern resembles that of the DNA in intact sperm heads.***

The Watson–Crick structure of B-DNA has the following major features (Table 28-1):

1. *It consists of two polynucleotide strands that wind about a common axis with a right-handed twist to form an ~20 Å in diameter double helix (Fig. 28-5). The two strands are antiparallel (run in opposite directions)* and wrap around each other such that they cannot be separated without unwinding the helix (a phenomenon known as **plectonemic coiling**). The bases occupy the core of the helix while its 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. 28-5).* 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° per bp) and, since the aromatic bases have van der Waals thicknesses of 3.4 Å and are partially stacked on each other (**base stacking;** Fig. 28-5*b*), the helix has a pitch (rise per turn) of 34 Å.

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. 28-6. 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 C(1′) atoms. Likewise, the double helix is undisturbed by exchanging the partners of a Watson–Crick base pair, that is, by changing a G·C to a C·G or a A·T to a T·A.* In contrast, any other combination of bases 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.

The two deep grooves that wind about the outside of B-DNA between the sugar–phosphate chains are of unequal size (Fig. 28-5*a*) because: (1) the top edge of each base pair, as drawn in Fig. 28-6, is structurally distinct from the bottom edge; and (2) the deoxyribose residues are asymmetric. The **minor groove** is that in which the

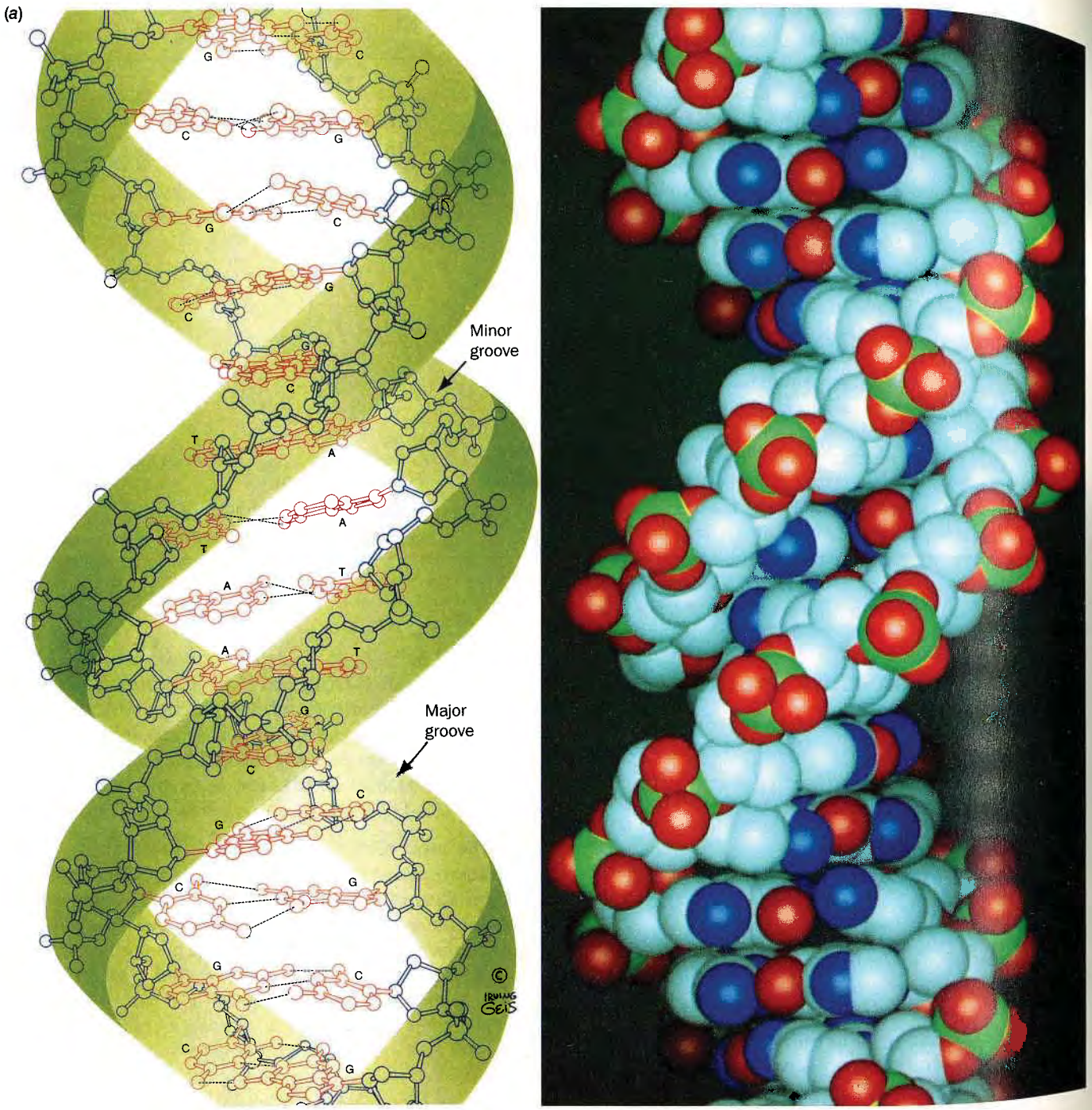


Figure 28-5
The structure of B-DNA as represented by ball-and-stick drawings and the corresponding computer-generated space-filling models. The repeating helix is based on the X-ray structure of the self-complementary dodecamer d(CGCGAATTCGCG) determined by Richard Dickerson and Horace Drew. (*a*) View perpendicular to the helix axis. In the drawing, the sugar–phosphate backbones, which wind about the periphery of the molecule, are blue, and the bases, which occupy its core, are red. In the space-filling model, C, N, O, and P atoms are white, blue, red, and green, respectively. H atoms have been omitted for clarity in both drawings. Note that the two sugar–phosphate chains run in opposite directions. (*b*) (*opposite*) View along the helix axis. In the drawing, the ribose ring O atoms are red and the nearest base pair is white. Note that the helix axis passes through the base pairs so that the helix has a solid core. [Drawings copyrighted © by Irving Geis. Computer graphics courtesy of Robert Stodola, Fox Chase Cancer Center.]

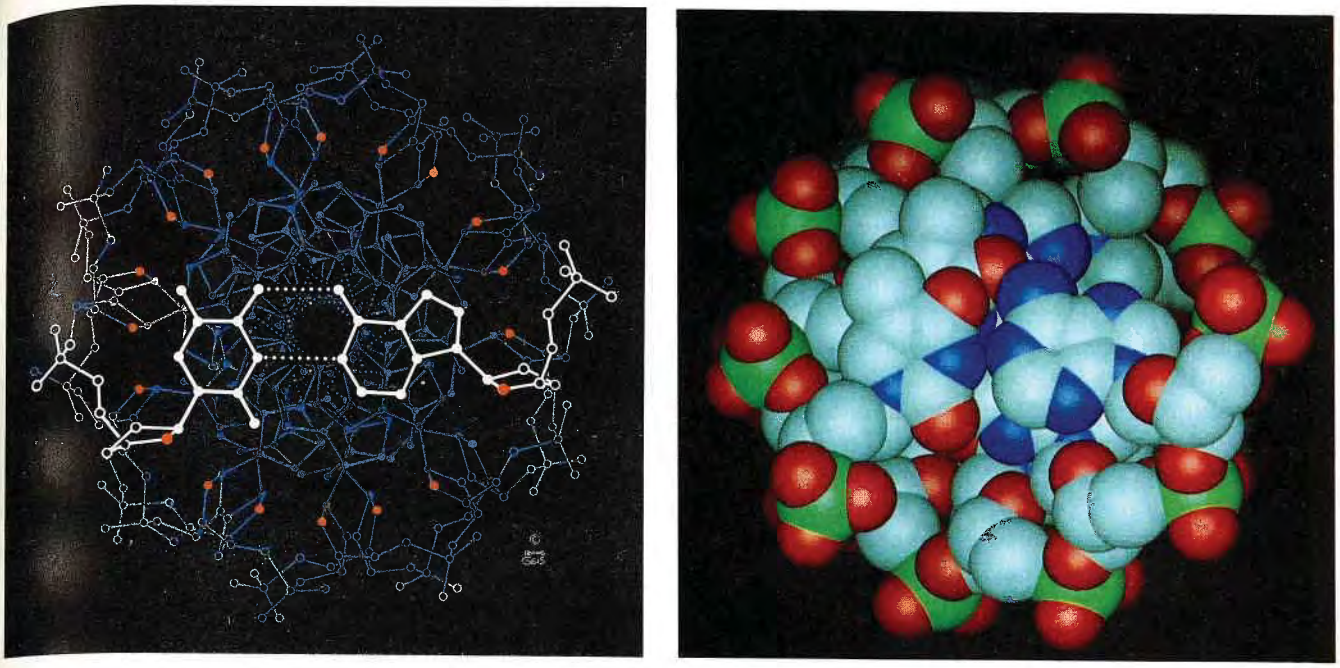


Figure 28-5 (*b*)

C(1′)-helix axis-C(1′) angle is $<180°$ (opening towards the bottom in Fig. 28-6; the helix axis passes through the middle of each base pair in B-DNA), whereas the **major groove** opens towards the opposite edge of each base pair (Fig. 28-6).

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.*

Real DNA Deviates from the Ideal Watson–Crick Structure

By the late 1970s, advances in nucleic acid chemistry permitted the synthesis and crystallization of ever longer oligonucleotides of defined sequences (Section 28-7). Consequently, some 25 years after the Watson–Crick structure had been formulated. the X-ray crystal structures of DNA fragments were clearly visualized for the first time (fiber diffraction studies provide only crude low resolution images in which the base pair elec-

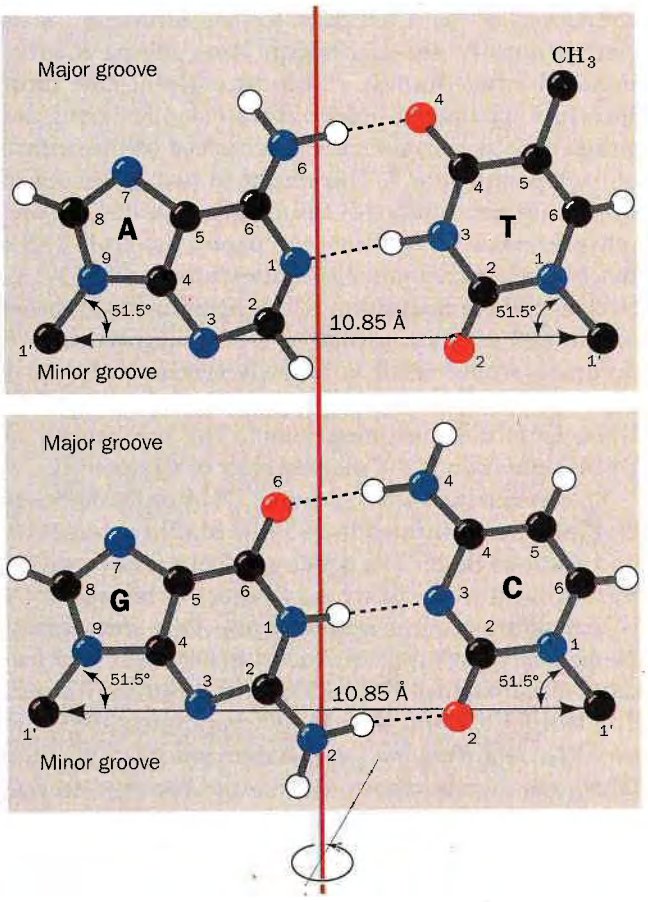


Figure 28-6
The Watson–Crick base pairs. The line joining the C(1′) atoms is the same length in both base pairs and makes equal angles with the glycosidic bonds to the bases. This gives DNA a series of pseudo-twofold symmetry axes (often referred to as **dyad axes**) that pass through the center of each 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,** 2196 (1969).]

tron density is the average electron density of all the base pairs in the fiber). Richard Dickerson and Horace Drew have shown that the self-complementary dodecamer d(CGCGAATTCGCG) crystallizes in the B-conformation. The molecule has an average rise per residue of 3.4 Å and has 10.1 bp per turn (a helical twist of 35.6°per bp), which is nearly equal to that of ideal B-DNA. Nevertheless, *individual residues significantly depart from this average conformation in a manner that appears to be sequence dependent (Fig. 28-5).* For example, the helical twist per base pair in this dodecamer ranges from 28 to 42°. Each base pair further deviates from its ideal conformation by such distortions as propeller twisting (the opposite rotation of paired bases about the base pair's long axis; in the above dodecamer these values range from 10 to 20°) and base pair roll (the tilting of a base pair as a whole about its long axis). Indeed, rapidly accumulating X-ray and NMR studies of other double helical DNA oligomers have amply demonstrated that *the structure of DNA is surprisingly irregular in a sequence-specific manner. This phenomenon, as we shall see (Sections 29-3C and E) is important for the sequence-specific binding to DNA of proteins that process genetic information.*

DNA Is Semiconservatively Replicated

The Watson–Crick structure also suggests how DNA can direct its own replication. Each polynucleotide strand can act as a template for the formation of its complementary strand through base pairing interactions. 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 (Fig. 1-16). 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 31.

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}N, a heavy isotope of nitrogen (^{14}N is the naturally abundant isotope). This was accomplished by growing *E. coli* for 14 generations in a medium that contained $^{15}NH_4Cl$ as its only nitrogen source. The labeled bacteria were then abruptly transferred to an ^{14}N-containing medium and the density of their DNA was monitored as a function of bacterial growth by equilibrium density gradient ultracentrifugation (Section 5-5B; a technique Meselson, Stahl, and Jerome Vinograd had developed for the purpose of distinguishing ^{15}N-labeled DNA from unlabeled DNA).

The results of the Meselson–Stahl experiment are displayed in Fig. 28-7. After one generation (doubling of the cell population), all of the DNA had a density exactly halfway between the densities of fully ^{15}N-labeled DNA and unlabeled DNA. This DNA must therefore contain equal amounts of ^{14}N and ^{15}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, one half of the DNA molecules were unlabeled and the remainder were $^{14}N—^{15}N$ 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.

Meselson and Stahl also demonstrated that DNA is double stranded. DNA from ^{15}N-labeled *E. coli* that were grown for one generation in an ^{14}N medium was heat denatured at 100°C (which causes strand separation; Section 28-3A) and then subjected to density gradient ultracentrifugation. Two bands were observed; one at the density of fully ^{15}N-labeled DNA and the other at the density of unlabeled DNA. Moreover the molecular masses of the DNA in these bands, as estimated from their peak shapes, was one half that of undenatured DNA (the peak width varies with molecular mass). Native DNA must therefore be composed of two equal-sized strands that separate upon heat denaturation.

B. Other Nucleic Acid Helices

Double-stranded DNA is a conformationally variable molecule. In the following subsections we discuss its major conformational states besides B-DNA and also those of double-stranded RNA.

A-DNA's Base Pairs Are Inclined to the Helix Axis

When the relative humidity is reduced to 75%, B-DNA undergoes a reversible conformational change to the so-called A form. Fiber X-ray studies indicate that *A-DNA forms a wider and flatter right-handed helix than does B-DNA* (Fig. 28-8; Table 28-1). A-DNA has 11 bp per turn and a pitch of 28 Å which gives A-DNA an axial hole (Fig. 28-8*b*). The most striking feature of A-DNA, however, is that the planes of its base pairs are tilted 20° with respect to the helix axis. A-DNA therefore has a deep major groove and a very shallow minor groove; it can be described as a flat ribbon wound around a 6 Å in diameter cylindrical hole. Most self-

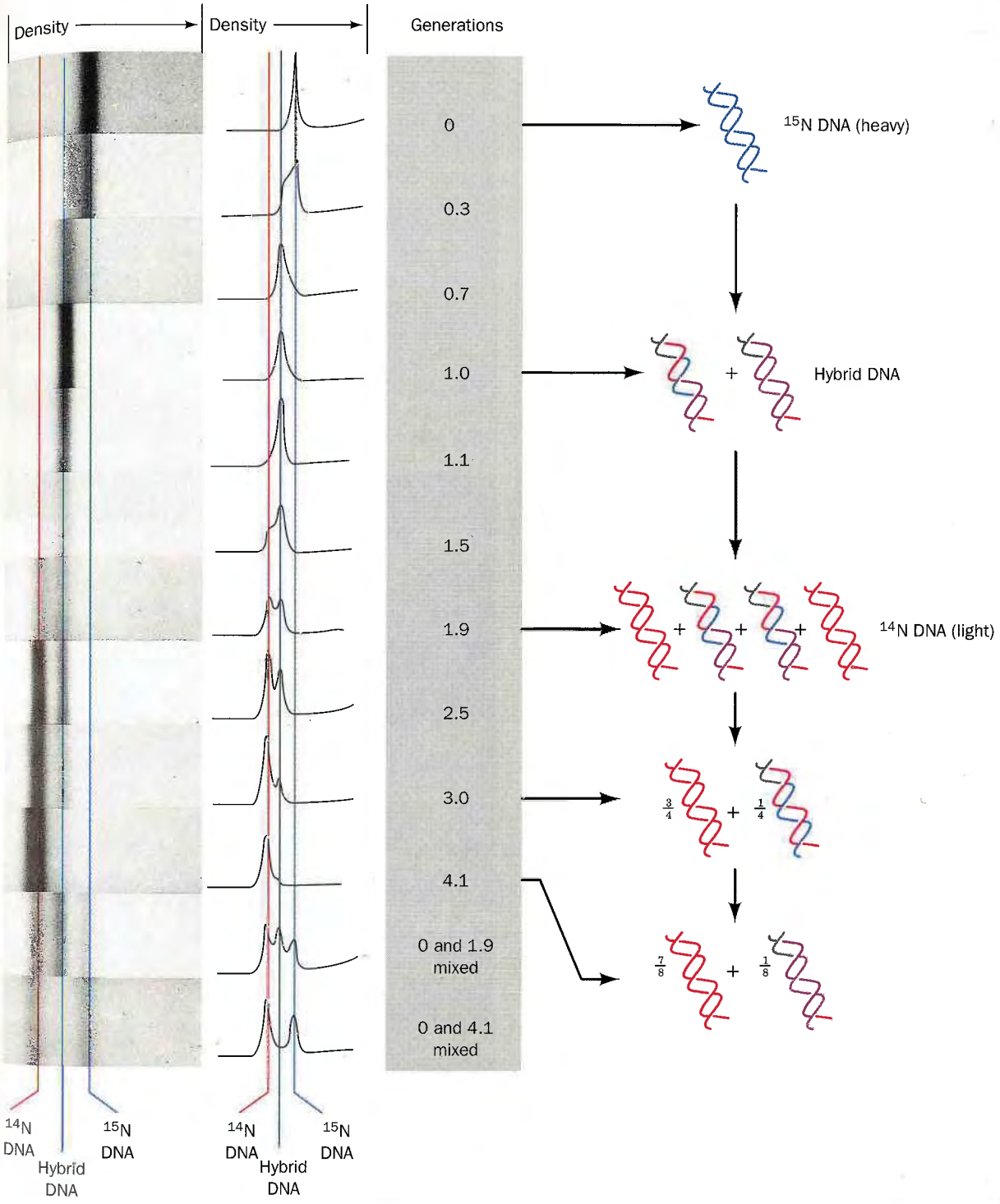


Figure 28-7
The demonstration of the semiconservative nature of DNA replication in *E. coli*. DNA in a CsCl solution of density 1.71 $g \cdot cm^{-3}$ was subjected to equilibrium density gradient ultracentrifugation at 140,000 *g* in an analytical ultracentrifuge (a device in which the spinning sample can be optically observed). The enormous centrifugal acceleration caused the CsCl to form a density gradient in which DNA migrated to its position of buoyant density. The left panels are UV absorption photographs of ultracentrifuge cells (DNA strongly absorbs UV 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}N content. The bands furthest to the right (greatest radius and density) arise from DNA that is fully ^{15}N labeled, whereas unlabeled DNA, which is 0.014 $g \cdot cm^{-3}$ less dense, forms the leftmost bands. The bands in the intermediate position result from duplex DNA in which one strand is ^{15}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}N labeled) and synthesized by succeeding generations (*red*, unlabeled). [From Meselson, M. and Stahl, F. W., *Proc. Natl. Acad. Sci.* **44**, 674 (1958).]

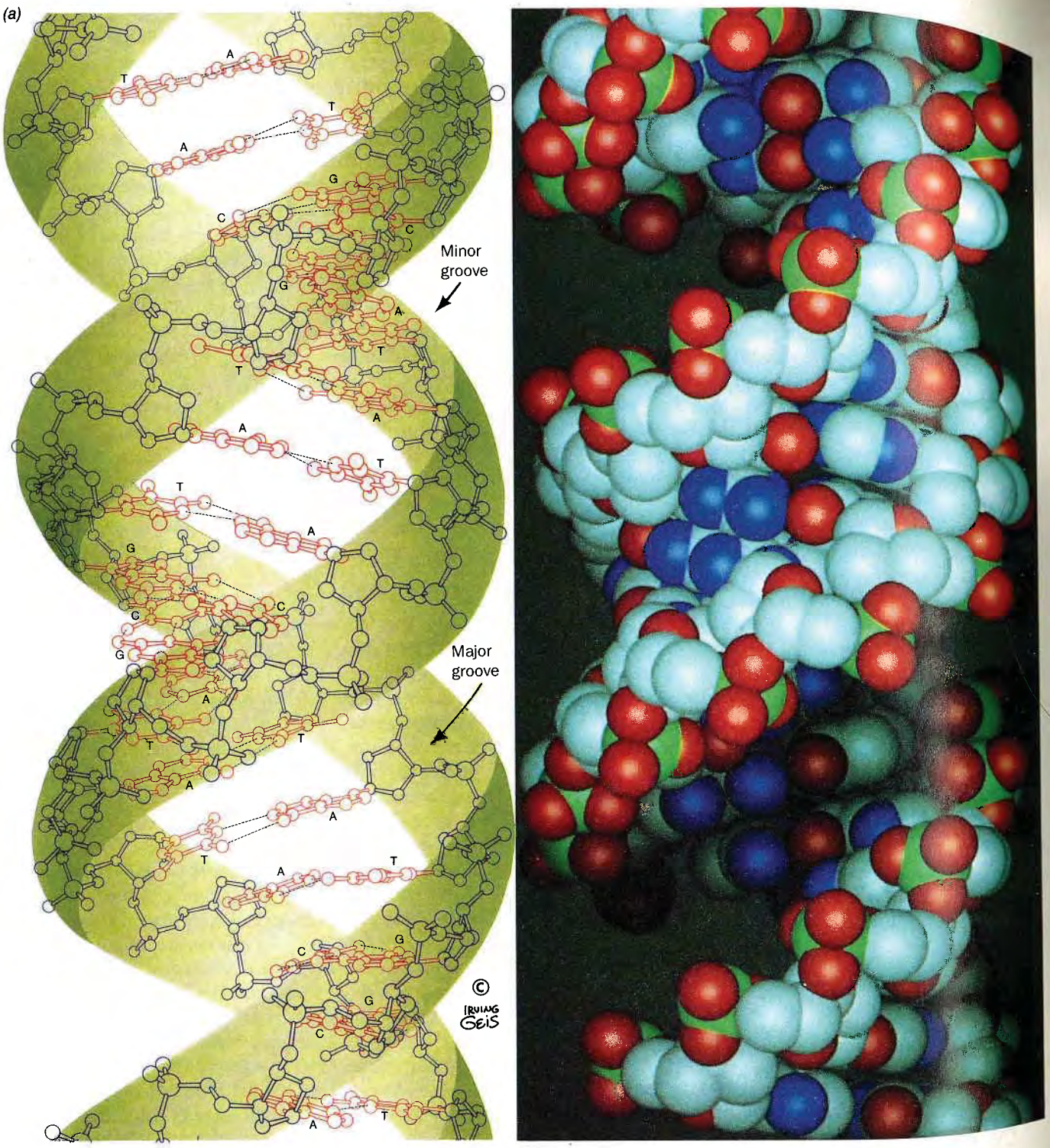


Figure 28-8
Ball-and-stick drawings and the corresponding space-filling models of A-DNA as viewed (*a*) perpendicular to the helix axis, and (*b*) (*opposite*) along the helix axis. The color codes are given in Fig. 28-5. The repeating helix was generated by Richard Dickerson based on the X-ray structure of the self-complementary octamer d(GGTATACC) determined by Olga Kennard, Dov Rabinovitch, Zippora Shakked, and Mysore Viswamitra. Note that the base pairs are inclined to the helix axis and that the helix has a hollow core. Compare this figure with Fig. 28-5. [Drawings copyrighted © by Irving Geis. Computer graphics courtesy of Robert Stodola, Fox Chase Cancer Center.]

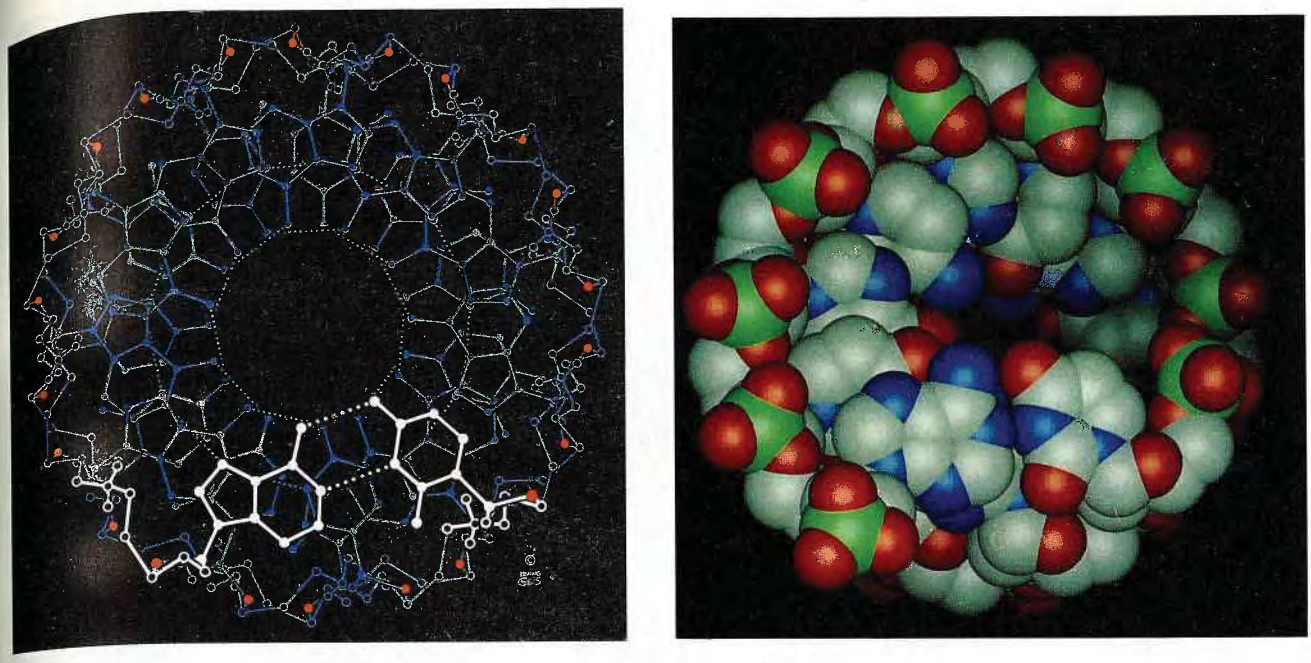


Figure 28.8 (*b*)

complementary oligonucleotides of <10 base pairs; for example, d(GGCCGGCC) and d(GGTATACC), crystallize in the A-DNA conformation. Like B-DNA, these molecules exhibit considerable sequence-specific conformational variation. It has not been established that A-DNA exists *in vivo* although a few experimental observations suggest that certain DNA segments normally assume the A conformation.

Z-DNA Forms a Left-Handed Helix

Occasionally, a seemingly well understood or at least familiar system exhibits quite unexpected properties. Over 25 years after the discovery of the Watson–Crick structure, the crystal structure determination of d(CGCGCG) by Andrew Wang and Alexander Rich revealed, quite surprisingly, *a left-handed double helix (Fig. 28-9; Table 28-1).* A similar helix is formed by d(CGCATGCG). *This helix, which has been dubbed Z-DNA, has 12 Watson–Crick base pairs per turn, a pitch of 45 Å and, in contrast to A-DNA, a deep minor groove and no discernable major groove.* Z-DNA therefore resembles a left-handed drill bit in appearance. The base pairs in Z-DNA are flipped 180° relative to those in B-DNA (Fig. 28-10) through conformational changes discussed in Section 28-3B. As a consequence, the repeating unit of Z-DNA is a dinucleotide, d(XpYp), rather than a single nucleotide as it is in the other DNA helices. Here, X is usually a pyrimidine residue and Y is usually a purine residue because the purine nucleotide assumes a conformation that would be sterically unfavorable in the pyrimidine nucleotide. The line joining successive phosphate groups on a polynucleotide strand of Z-DNA therefore follows a zigzag path around the helix (Fig. 28-9*a*; hence the name Z-DNA) rather than a smooth curve as it does in A- and B-DNAs (Figs. 28-5*a* and 28-8*a*).

Fiber diffraction and NMR studies have shown that complementary polynucleotides with alternating purines and pyrimidines, such as poly d(GC)·poly d(GC) or poly d(AC)·poly d(GT), take up the Z-DNA conformation at high salt concentrations. Evidently, *the Z-DNA conformation is most readily assumed by DNA segments with alternating purine–pyrimidine base sequences (for structural reasons explained in Section 28-3B).* A high salt concentration stabilizes Z-DNA relative to B-DNA by reducing the otherwise increased electrostatic repulsions between closest approaching phosphate groups on opposite strands (8 Å in Z-DNA vs 12 Å in B-DNA). The methylation of cytosine residues at C(5), a common biological modification (Section 31-7), also promotes Z-DNA formation since a hydrophobic methyl group in this position is less exposed to solvent in Z-DNA than it is in B-DNA.

Does Z-DNA have any biological significance? Rich has proposed that the reversible conversion of specific segments of B-DNA to Z-DNA under appropriate circumstances acts as a kind of switch in regulating genetic expression. Yet, the *in vivo* existence of Z-DNA has been difficult to prove. A major problem is demonstrating that a particular probe for detecting Z-DNA, a Z-DNA-specific antibody, for example, does not in itself cause what would otherwise be B-DNA to assume the Z conformation—a kind of biological uncertainty principle (the act of measurement inevitably disturbs the system being measured). Recently, however, Z-DNA has been shown to be present in *E. coli* by employing an *E. coli* enzyme that methylates a specific base sequence

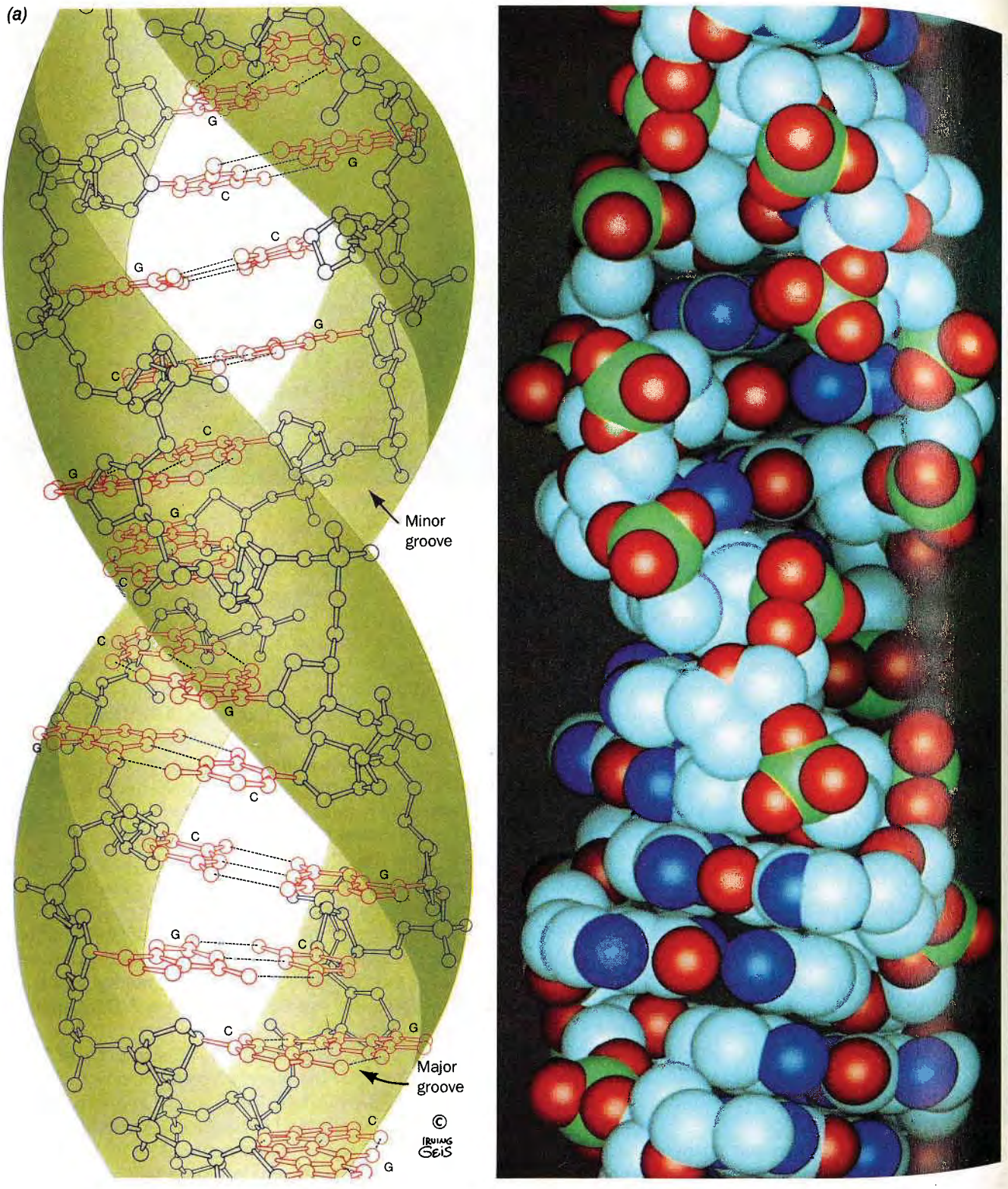


Figure 28-9
Ball-and-stick drawings and the corresponding space-filling models of Z-DNA as viewed (*a*) perpendicular to the helix axis and (*b*) (*opposite*) along the helix axis. The color codes are given in Fig. 28-5. The repeating helix was generated by Richard Dickerson based on the X-ray structure of the self-complementary hexamer d(CGCGCG) determined by Andrew Wang and Alexander Rich. Note that the helix is left handed and that the sugar–phosphate chains follow a zigzag course (alternate ribose residues lie at different radii in Part *b*) indicating that the Z-DNA's repeating motif is a dinucleotide. Compare this figure with Figs. 28-5 and 28-8. [Drawings copyrighted © by Irving Geis. Computer graphics courtesy of Robert Stodola, Fox Chase Cancer Center.]

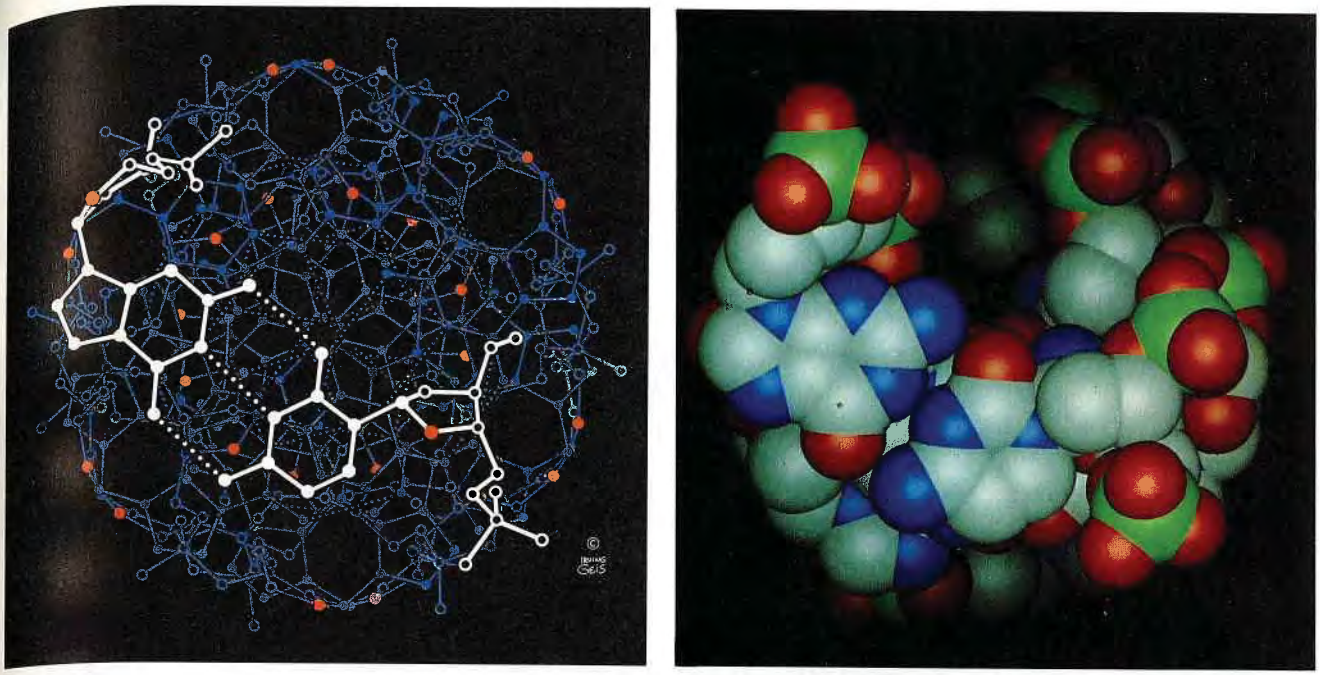


Figure 28-9 (*b*)

in vitro when the DNA is in the B form but not when it is in the Z form. The *in vivo* methylation of this base sequence is inhibited when it is cloned in *E. coli* (by techniques discussed in Section 28-8) within or adjacent to a DNA segment that can form Z-DNA. Moreover, there is a balance between the *in vivo* B and Z forms of these DNAs that is thought to be influenced by environmental factors such as salt concentration and protein binding. Nevertheless, the biological function of Z-DNA, if any, remains unknown.

RNA-11 and RNA–DNA Hybrids Have an A-DNA-Like Conformation

Double helical RNA is unable to assume a B-DNA-like conformation because of steric clashes involving its 2′-OH groups. Rather, it usually assumes a conformation resembling A-DNA (Fig. 28-8), known as **A-RNA** or **RNA-11,** which has 11 bp per helical turn, a pitch of 30 Å, and its base pairs inclined to the helix axis by ~14°. Many RNAs, for example, transfer and ribosomal RNAs (whose structures are detailed in Sections 30-2A

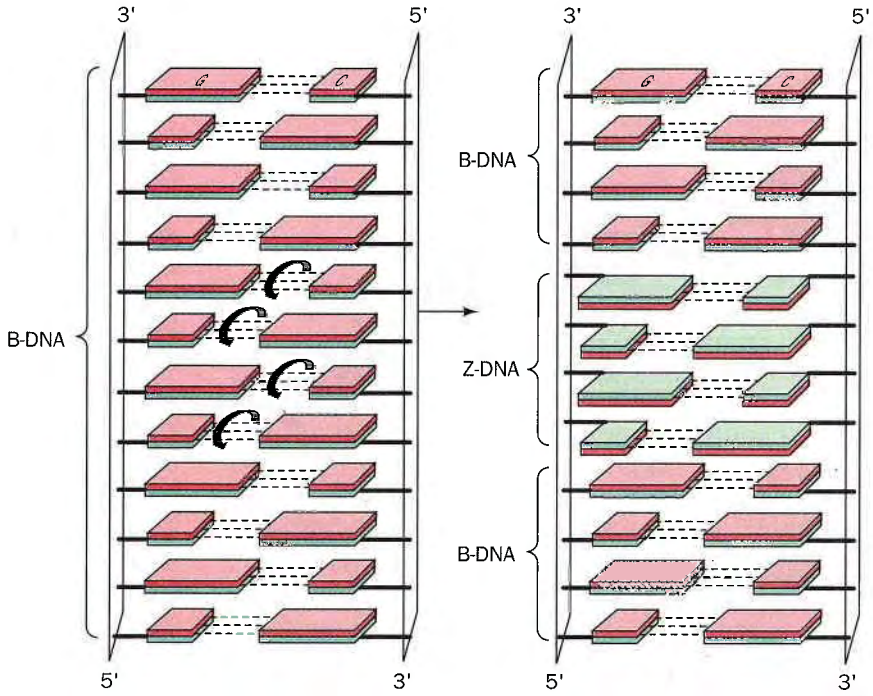


Figure 28-10
The conversion of B-DNA to Z-DNA, here represented by a 4 bp segment, involves a 180° flip of each base pair (*curved arrows*) relative to the sugar–phosphate chains. Here, the different faces of the base pairs are colored red and green. [After Rich, A., Nordheim, A., and Wang, A. H.-J., *Annu. Rev. Biochem.* **53,** 799 (1984).]

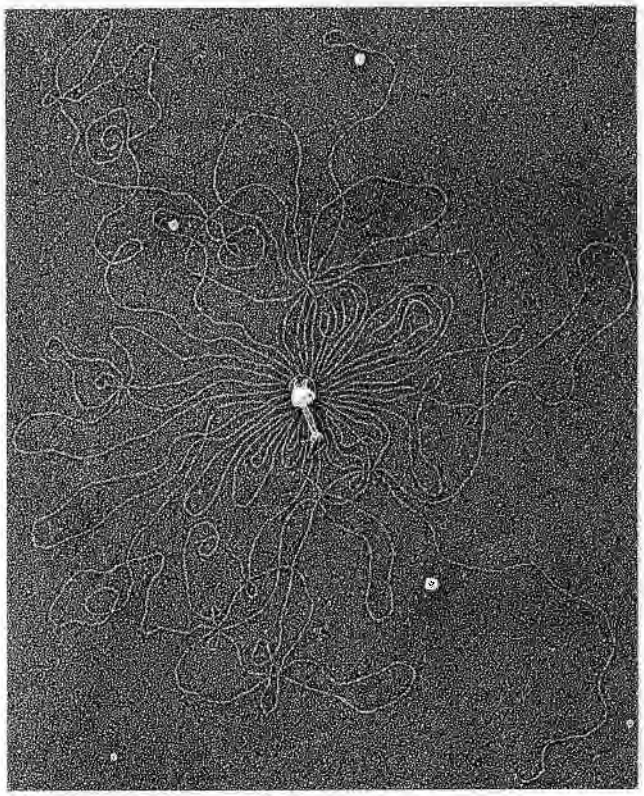


Figure 28-11
An electron micrograph of a T2 bacteriophage that had been osmotically lysed in distilled water so that its DNA spilled out. Without special treatment, duplex DNA, which is only 20 Å in diameter, is difficult to visualize in the electron microscope. In the **Kleinschmidt procedure,** DNA is fattened to ~200 Å in diameter by coating it with denatured cytochrome *c* or some other basic protein. The preparation is rendered visible in the electron microscope by shadowing it with platinum. [From Kleinschmidt, A. K., Lang, D., Jacherts, D., and Zahn, R. K., *Biochim. Biophys. Acta* **61,** 861 (1962).]

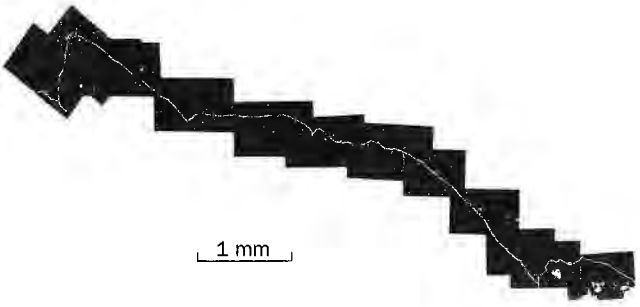


Figure 28-12
An autoradiograph of *Drosophila melanogaster* DNA. Lysates of *D. melanogaster* cells that had been cultured with [^{3}H]thymidine were spread on a glass slide and covered with a photographic emulsion that was developed after a 5-month exposure. The measured contour length of the DNA is 1.2 cm. [From Kavenoff, R., Klotz, L. C., and Zimm, B. H., *Cold Spring Harbor Symp. Quant. Biol.* **38,** 4 (1973). Copyright © 1973 by Cold Spring Harbor Laboratory.]

and 30-3A), contain complementary sequences that form double helical stems. Hybrid double helices, which consist of one strand each of RNA and DNA, also have an A-DNA-like conformation. Small segments of RNA·DNA hybrid helices must occur in both the transcription of RNA on DNA templates (Section 29-2D) and in the initiation of DNA replication by short lengths of RNA (Section 31-1D).

C. The Size of DNA

DNA molecules are generally enormous (Fig. 28-11). The molecular mass of DNA has been determined by a variety of techniques including hydrodynamic methods (Section 5-5), length measurements by electron microscopy, and autoradiography [Fig. 28-12; a base pair of Na^+ B-DNA has an average molecular mass of 660 D and a length (thickness) of 3.4 Å]. 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 presented in Table 28-2. 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 28-2).

The visualization of DNAs from prokaryotes has demonstrated that their entire **genome** (complement of genetic information) is contained on a single, usually 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 amount of DNA contained in the chromosome. Presumably other eukaryotic chromosomes also 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. 28-12 were expanded 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 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 handling. Before 1960, when this was first realized, the measured molecular masses of DNA were no higher than 10 million D. 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 pipetting, through the use of a high speed blender, or by **sonication** (exposure to high frequency sound waves).

Table 28-2
Sizes of Some DNA Molecules

Organism	Number of Base Pairs (kb)[a]	Contour Length (μm)
Viruses		
Polyoma, SV40	5.1	1.7
Bacteriophage λ	48.6	17
T2,T4,T6 bacteriophage	166	55
Fowlpox	280	193
Bacteria		
Mycoplasma hominis	760	260
Escherichia coli	4,000	1,360
Eukaryotes		
Yeast *(in 17 haploid chromosomes)*	13,500	4,600
Drosophila (in 4 haploid chromosomes)	165,000	56,000
Human *(in 23 haploid chromosomes)*	2,900,000	990,000
Lungfish *(in 19 haploid chromosomes)*	102,000,000	34,700,000

[a] kb = kilo base pair = 1000 base pairs (bp).

Source: Kornberg, A., *DNA Replication, p.* 20, Freeman (1980).

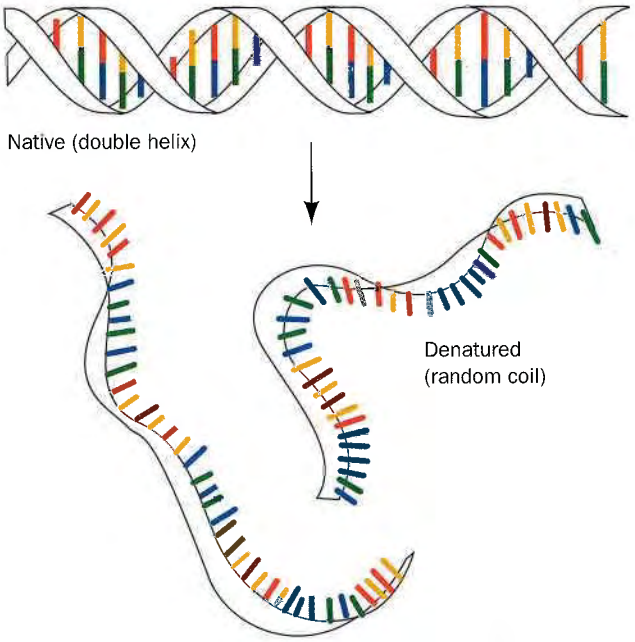


Figure 28-13
A schematic representation of DNA denaturation.

3. FORCES STABILIZING NUCLEIC ACID STRUCTURES

DNA does not exhibit the structural complexity of proteins because it has only a limited repertoire of secondary structures and no comparable tertiary or quaternary structures. This is perhaps to be expected since there is a far greater range of chemical and physical properties among the 20 amino acid residues of proteins than there is among the four DNA bases. As we discuss in Sections 30-2B and 3A, however, many RNAs have well-defined tertiary structures.

In this section we examine the forces that give rise to the structures of nucleic acids. These forces are, of course, much the same as those that are responsible for the structures of proteins (Section 7-4) but, as we shall see, the way they combine gives nucleic acids properties that are quite different from those of proteins.

A. 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 the random coil conformation (Fig. 28-13). 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 DNA decomposes to relatively freely jointed single strands.

DNA Denaturation Is a Cooperative Process

The most convenient way of monitoring the native state of DNA is by its ultraviolet (UV) absorbance spectrum. When DNA denatures, its UV absorbance, which is almost entirely due to its aromatic bases, increases by ~40% at all wavelengths (Fig. 28-14). 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. 28-15). This indicates that the denaturation of DNA is a cooperative phenomenon in which the collapse of one part of the structure destabilizes the remainder. The denaturation of DNA may be described as the melting of a one-dimensional solid so that Fig. 28-15 is referred to as a **melting curve** and the temperature at its midpoint is known as its **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. T_m also increases linearly with the mole fraction of G·C base pairs (Fig. 28-16), which indicates that triply hydrogen bonded G·C base pairs

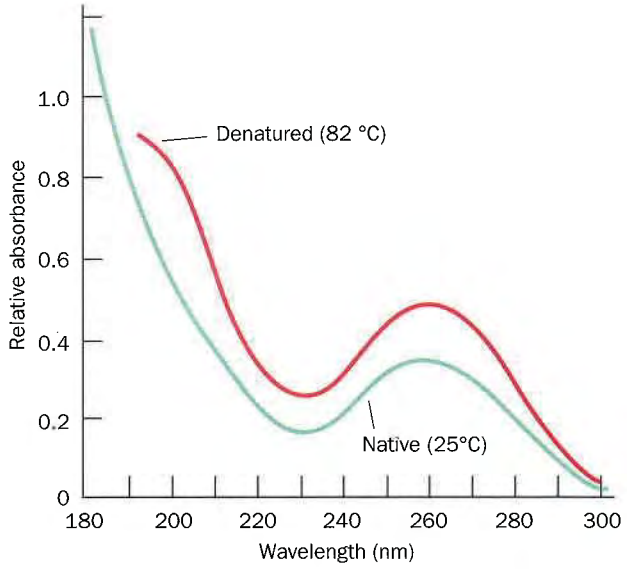


Figure 28-14
The UV absorbance spectra of native and heat denatured *E. coli* DNA. Note that denaturation does not change the general shape of the absorbance curve but only increases its intensity. [After Voet, D., Gratzer, W. B., Cox, R. A., and Doty, P., *Biopolymers* **1,** 205 (1963).]

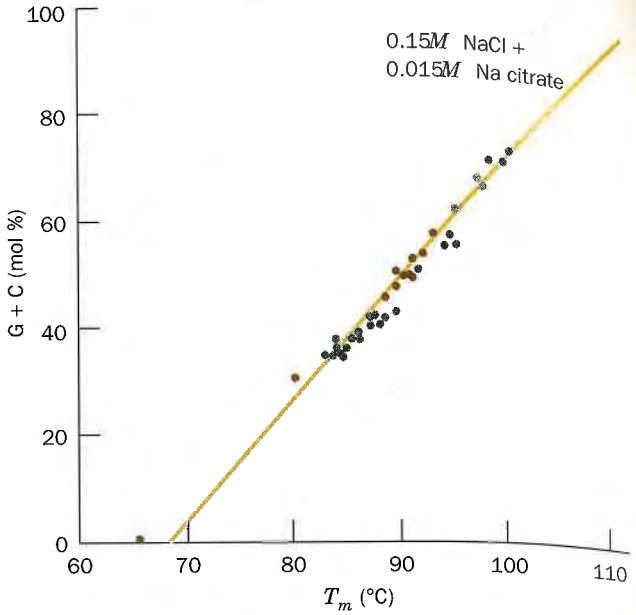


Figure 28-16
The variation of the melting temperatures, T_m, of various DNAs with their G + C content. The DNAs were dissolved in a solution containing 0.15*M* NaCl and 0.015*M* Na citrate. [After Marmur, J. and Doty, P., *J. Mol. Biol.* **5,** 113 (1962).]

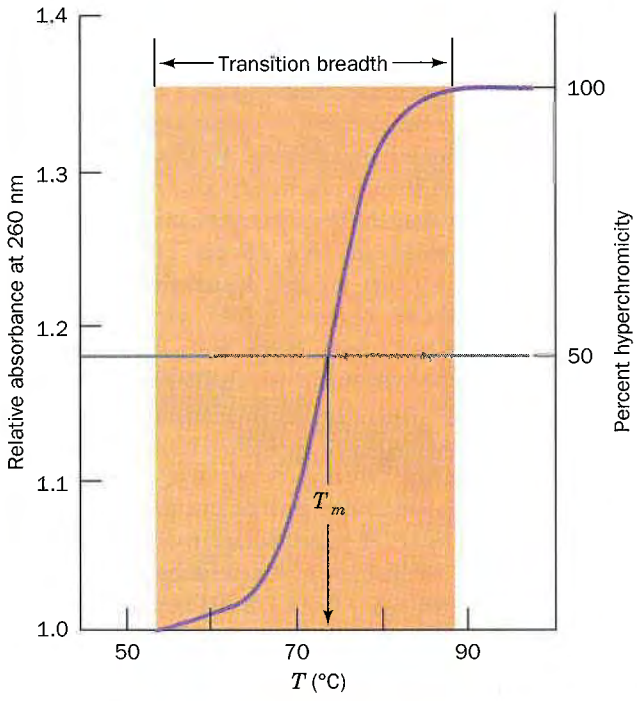


Figure 28-15
An example of a DNA melting curve. The relative absorbance is the ratio of the absorbance (customarily measured at 260 nm) at the indicated temperature to that at 25°C. The melting temperature, T_m, is the temperature at which one half of the maximum absorbance increase is attained.

are more stable than doubly hydrogen bonded A·T base pairs.

Denatured DNA Can Be Renatured

If a solution of denatured DNA is rapidly cooled below its T_m, the resulting DNA will be only partially base paired (Fig. 28-17) because the 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 ~25°C below the T_m, enough thermal energy is available for short base paired regions to rearrange by melting and reforming but not so much as to melt out long complementary stretches. Under such **annealing conditions,** as Julius Marmur discovered in 1960, denatured DNA eventually 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.

B. Sugar–Phosphate Chain Conformations

The conformation of a nucleotide unit, as Fig. 28-18 indicates, is specified by the six torsion angles of the sugar–phosphate backbone and the torsion angle describing the orientation of the base about the glycosidic bond [the bond joining C(1′) to the base]. It would seem that these seven degrees of freedom per nucleotide would render polynucleotides highly flexible. Yet, as we shall see, these torsion angles are subject to a variety of

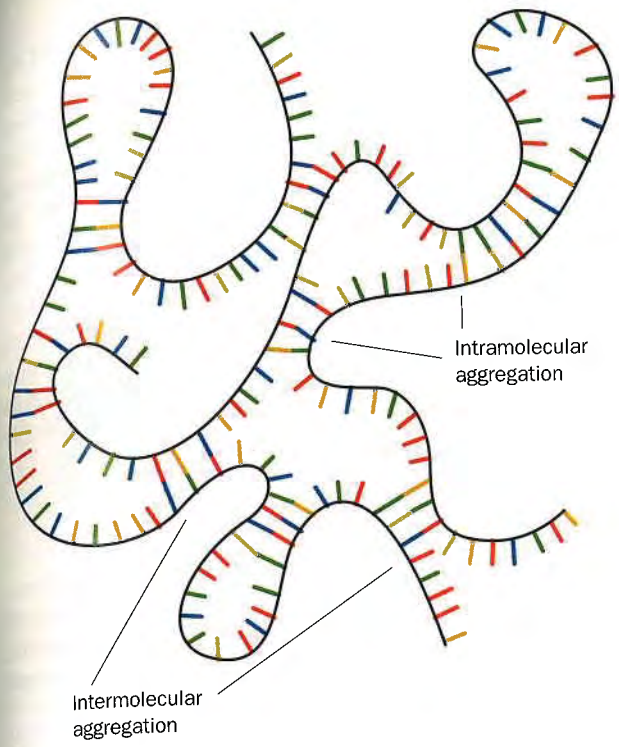


Figure 28-17
A schematic representation of the imperfectly base paired structures assumed by DNA that has been heat denatured and then rapidly cooled. Note that both intramolecular and intermolecular aggregation may occur.

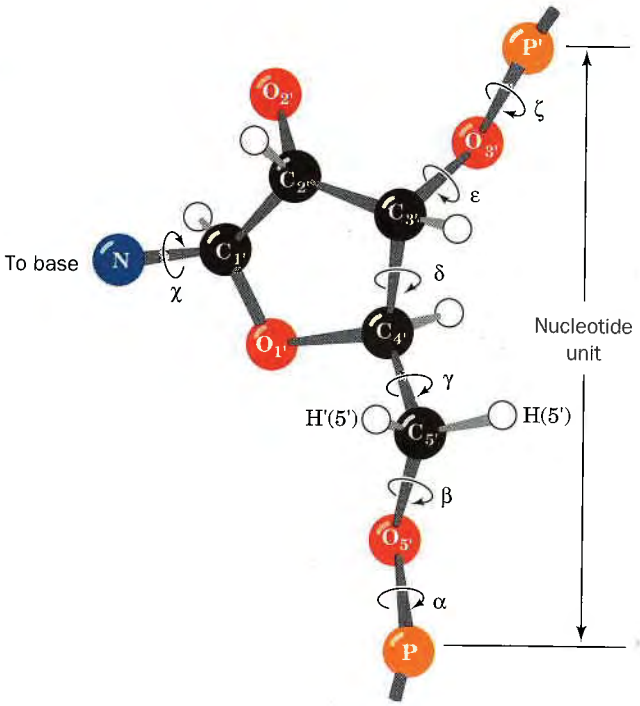


Figure 28-18
The conformation of a nucleotide unit is determined by the seven indicated torsional angles.

internal constraints that greatly restrict their conformational freedom.

Torsion Angles about Glycosidic Bonds Have One or Two Stable Positions

The rotation of a base about its glycosidic bond is greatly hindered, as is best seen by the manipulation of a space-filling molecular model. Purine residues have two sterically permissible orientations relative to the sugar known as the **syn** (Greek: with) and **anti** (Greek: against) conformations (Fig. 28-19). For pyrimidines, only the anti conformation is easily formed because, in the syn conformation, the sugar residue sterically interferes with the pyrimidine's C(2) substituent. In most double helical nucleic acids, all bases are in the anti conformation. The exception is Z-DNA (Section 28-2B), in which the alternating pyrimidine and purine residues are anti and syn, respectively. *This explains Z-DNA's pyrimidine–purine alternation.* Indeed, the base pair flips that convert B-DNA to Z-DNA (Fig. 28-10) are brought about by rotating each purine base about its glycosidic bond from the anti to syn conformations, whereas the sugars rotate in the pyrimidine nucleotides thereby maintaining their anti conformations.

Sugar Ring Pucker Is Limited to Only a Few of Its Possible Arrangements

The ribose ring has a certain amount of flexibility that significantly affects the conformation of the sugar–phosphate backbone. The vertex angles of a regular pentagon are 108°, a value quite close to the tetrahedral

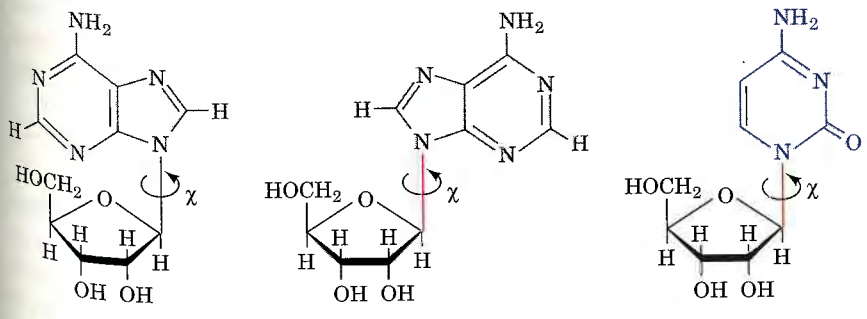


Figure 28-19
The sterically allowed orientations of purine and pyrimidine bases with respect to their attached ribose units.

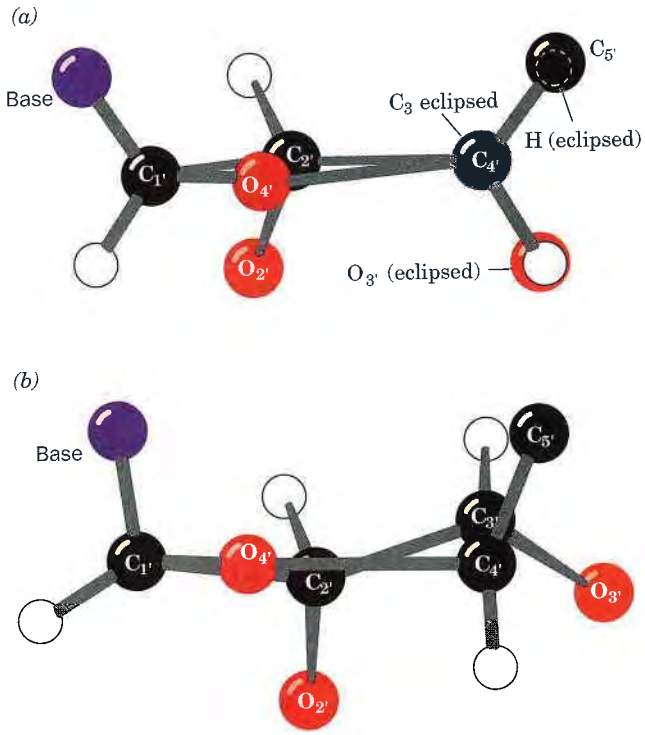


Figure 28-20
The substituents to (*a*) a planar ribose ring [here viewed down the C(3′)—C(4′) bond] are all eclipsed. The resulting steric strain is partially relieved by ring puckering such as in (*b*), a half-chair conformation in which C(3′) is the out-of-plane atom.

angle (109.5°), so that one might expect the ribofuranose ring to be nearly flat. However, the ring substituents are eclipsed when the ring is planar. To relieve the resultant crowding, which even occurs between hydrogen atoms, the ring **puckers;** that is, it becomes slightly nonplanar, so as to reorient the ring substituents (Fig. 28-20; this is readily observed by the manipulation of a skeletal molecular model).

One would, in general, expect only three of a ribose ring's five atoms to be coplanar since three points define a plane. Nevertheless, in the great majority of the >50 nucleoside and nucleotide crystal structures that have been reported, four of the ring atoms are coplanar to within a few hundreths of an Å and the remaining atom is out of this plane by several tenths of an Å (the **half-chair** conformation). If the out-of-plane atom is displaced to the same side of the ring as atom C(5′), it is said to have the **endo** conformation (Greek: *endon,* within), whereas displacement to the opposite side of the ring from C(5′) is known as the **exo** conformation (Greek: *exo,* out of). In the great majority of known nucleoside and nucleotide structures, the out-of-plane atom is either C(2′) or C(3′) (Fig. 28-21). C(2′)-*endo* is the most frequently occurring ribose pucker with C(3′)-*endo* and -*exo* also being common. Other ribose conformations are rare.

The ribose pucker is conformationally important in nucleic acids because it governs the relative orientations of the phosphate substituents to each ribose residue. For instance, it is difficult to build a model of a double helical nucleic acid unless the sugars are either C(2′)-*endo* or C(3′)-*endo.* In fact, B-DNA has the C(2′)-*endo* conformation whereas A-DNA and RNA-11 are C(3′)-*endo.* In Z-DNA, the purine nucleotides are all C(3′)-*endo* and the pyrimidine nucleotides are C(2′)-*endo,* which is another reason why the repeating unit of Z-DNA is a dinucleotide. Note that the most common sugar puckers of independent nucleosides and nucleotides, molecules that are subject to few of the conformational constraints of double helices, are the same as those of double helices.

The Sugar–Phosphate Backbone Is Conformationally Constrained

If the torsion angles of the sugar–phosphate chain (Fig. 28-18) were completely free to rotate, there could probably be no stable nucleic acid structure. However, the comparison, by Muttaiya Sundaralingam, of some

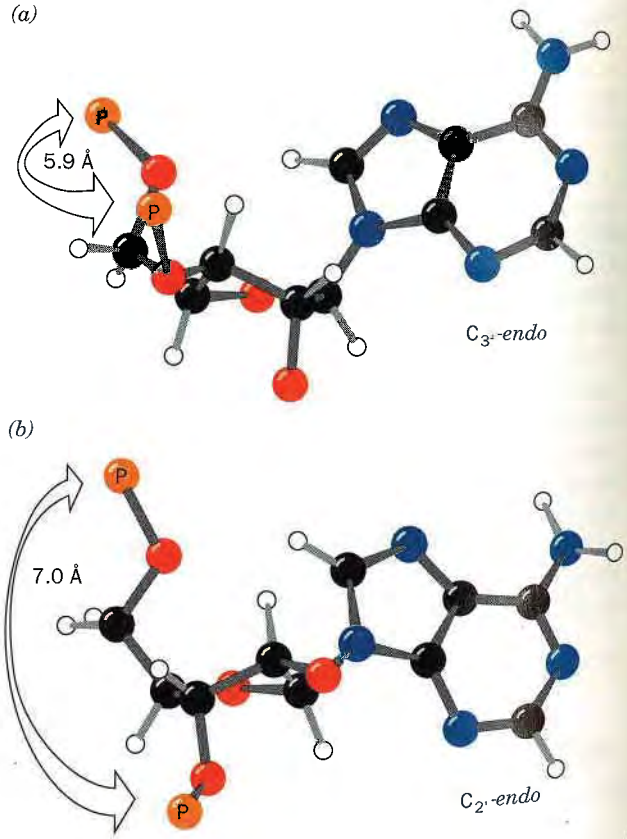


Figure 28-21
Nucleotides in (*a*) the C(3′)-*endo* conformation [on the same side of the sugar ring as C(5′)], and (*b*) the C(2′)-*endo* conformation which occur, respectively, in A-DNA and B-DNA. The distances between adjacent P atoms in the sugar–phosphate backbone are indicated. [After Saenger, W., *Principles of Nucleic Acid Structure, p.* 237, Springer–Verlag (1983).]

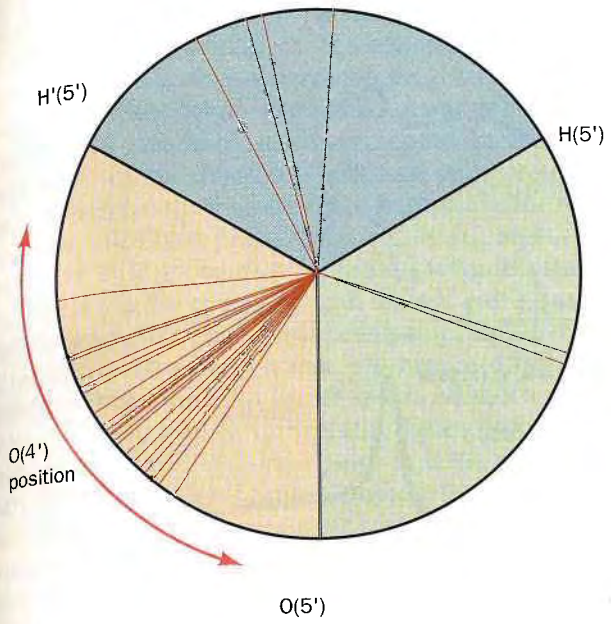


Figure 28-22
A conformational wheel showing the distribution of the torsion angle about the C(4′)—C(5′) bond (γ in Fig. 28-18) in 33 X-ray structures of nucleosides, nucleotides and polynucleotides. Each radial line represents the position of the C(4′)—O(4′) bond in a single structure relative to the substituents of C(5′) as viewed from C(5′) to C(4′). Note that most of the observed torsion angles fall within a relatively narrow range. [After Sundaralingam, M., *Biopolymers* **7**, 838 (1969).]

40 nucleoside and nucleotide crystal structures revealed that these angles are really quite restricted. For example, the torsion angle about the C(4′)—C(5′) bond (γ in Fig. 28-18) is rather narrowly distributed such that O(4′) usually has a gauche conformation with respect to O(5′) (Fig. 28-22). This is because the presence of the ribose ring together with certain noncovalent interactions of the phosphate group stiffens the sugar–phosphate chain by restricting its range of torsion angles. These restrictions are even greater in polynucleotides because of steric interference between residues.

The sugar–phosphate conformational angles of the various double helices are all reasonably strain free. *Double helices are therefore conformationally relaxed arrangements of the sugar–phosphate backbone.* Nevertheless, the sugar–phosphate backbone is by no means a rigid structure so that, upon strand separation, it assumes a random coil conformation.

C. Base Pairing

Base pairing is apparently a "glue" that holds together double-stranded nucleic acids. Only Watson–Crick pairs occur in the crystal structures of self-complementary oligonucleotides. It is therefore important to understand how Watson–Crick base pairs differ from other doubly hydrogen bonded arrangements of the bases that have reasonable geometries (e.g., Fig. 28-23).

Unconstrained A·T Base Pairs Assume Hoogsteen Geometry

When monomeric adenine and thymine derivatives are cocrystallized, the A·T base pairs that form invariably have adenine N(7) as the hydrogen bonding acceptor (**Hoogsteen geometry;** Fig. 28-23*b*) rather than N(1) (Watson–Crick geometry; Fig. 28-6). This suggests that Hoogsteen geometry is inherently more stable for A·T pairs than is Watson–Crick geometry. Apparently steric and other environmental influences make Watson–Crick geometry the preferred mode of base pairing in double helices. A·T pairs with Hoogsteen geometry are nevertheless of biological importance; for example, they help stabilize the tertiary structures tRNAs (Section

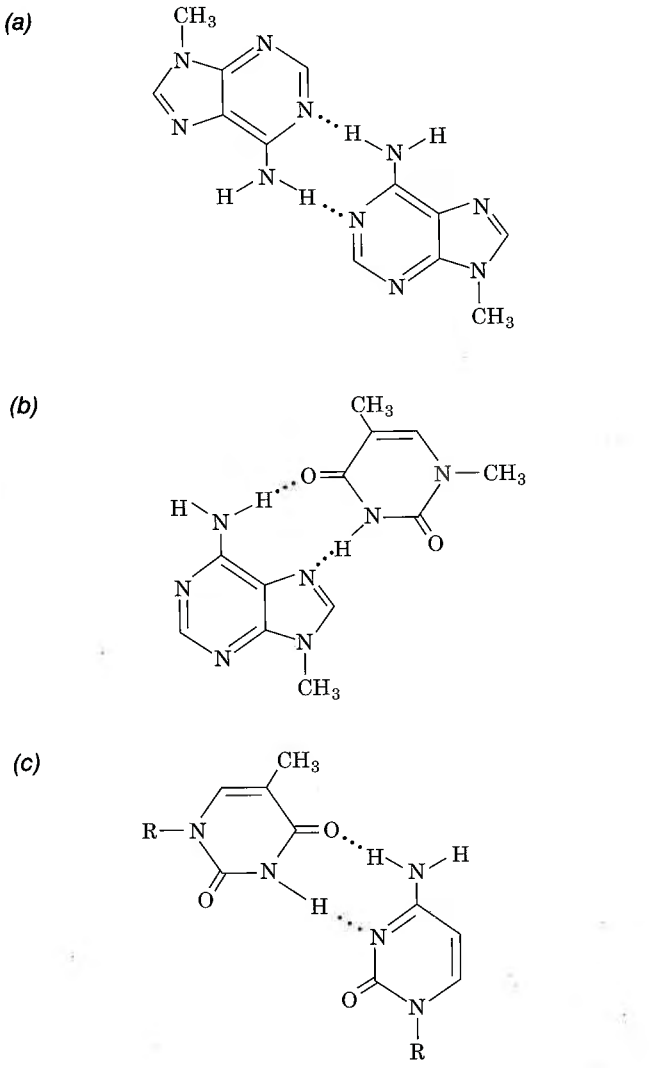


Figure 28-23
Some non-Watson–Crick base pairs. (*a*) the pairing of adenine residues in the crystal structure of 9-methyladenine. (*b*) The Hoogsteen pairing between adenine and thymine residues in the crystal structure of 9-methyladenine·1-methylthymine. (*c*) A hypothetical pairing between cytosine and thymine residues.

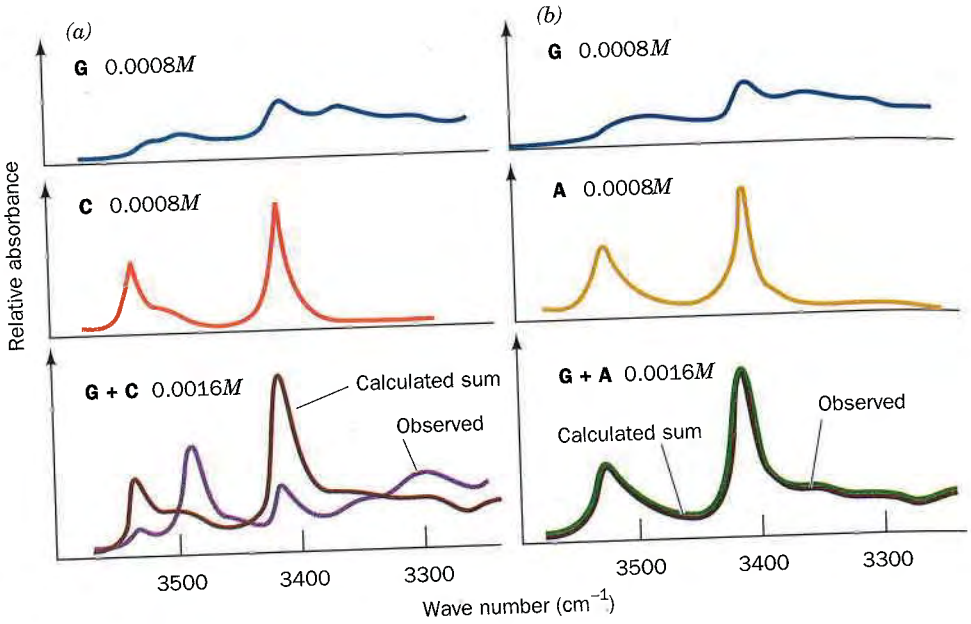


Figure 28-24
The IR spectra, in the N—H stretch region, of guanine, cytosine, and adenine derivatives, both separately and in the indicated mixtures. The solvent, $CDCl_3$, does not hydrogen bond with the bases and is relatively transparent in the frequency range of interest. (*a*) G + C. The line in the lower spectrum, which is the sum of the two upper spectra, is the calculated spectrum of G + C for noninteracting molecules. The band near 3500 cm^{-1} in the observed G + C spectrum is indicative of a specific hydrogen bonding association between G and C. (*b*) G + A. The close match between the calculated and observed spectra of the G + A mixture indicates that G and A do not significantly interact. [After Kyogoku, Y., Lord, R. C., and Rich, A., *Science* **154,** 5109 (1966).]

30-2B). In contrast, monomeric G·C pairs always co-crystallize with Watson–Crick geometry as a consequence of their triply hydrogen bonded structures.

Non-Watson–Crick Base Pairs Are of Low Stability

The bases of a double helix, as we have seen (Section 28-2A), associate such that any base pair position may interchangeably be A·T, T·A, G·C, or C·G without affecting the conformations of the sugar–phosphate chains. One might reasonably suppose that this requirement of **geometric complementarity** of the Watson–Crick base pairs, A with T and G with C, is the only reason that other base pairs do not occur in a double helical environment. In fact, this was precisely what was believed for many years after the DNA double helix was discovered.

Eventually, the failure to detect pairs of different bases in nonhelical environments other than A with T (or U) and G with C led Richard Lord and Rich to demonstrate, through spectroscopic studies, that *only the bases of Watson–Crick pairs have a high mutual affinity.* Figure 28-24*a* shows the infrared (IR) spectrum in the N—H stretch region of guanine and cytosine derivatives, both separately and in a mixture. The band in the spectrum of the G + C mixture that is not present in the spectra of either of its components is indicative of a specific hydrogen bonding interaction between G and C. Such an association, which can occur between like as well as unlike molecules, may be described by ordinary mass action equations.

$$B_1 + B_2 \rightleftharpoons B_1 \cdot B_2 \qquad K = \frac{[B_1 \cdot B_2]}{[B_1][B_2]} \qquad [28.1]$$

From the analyses of IR spectra such as Fig. 28-24, the values of K for the various base pairs have been determined. The self-association constants of the Watson–Crick bases are given in the top of Table 28-3 (the hydrogen bonded association of like molecules is indicated

Table 28-3
Association Constants for Base Pair Formation

Base Pair	$K(M^{-1})^a$
Self-Association	
A·A	3.1
U·U	6.1
C·C	28
G·G	10^3–10^4
Watson–Crick Base Pairs	
A·U	100
G·C	10^4–10^5

[a] Data measured in deuterochloroform at 25°C.

Source: Kyogoku, Y., Lord, R. C., and Rich, A., *Biochim. Biophys. Acta* **179,** 10 (1969).

by the appearance of new IR bands as the concentration of the molecule is increased). The bottom of Table 28-3 lists the association constants of the Watson–Crick pairs. Note that each of these latter quantities is larger than the self-association constants of both their component bases so that Watson–Crick base pairs preferentially form from their constituents. In contrast, the non-Watson–Crick base pairs, A·C, A·G, C·U, and G·U, whatever their geometries, have association constants that are negligible compared with the self-pairing association constants of their constituents (e.g., Fig. 28-24*b*). *Evidently, a second reason that non-Watson–Crick base pairs do not occur in DNA double helices is that they have relatively little stability.* Conversely, the exclusive presence of Watson–Crick base pairs in DNA results, in part, from an **electronic complementarity** matching A to T and G to C. The theoretical basis of this electronic complementarity, which is an experimental observation, is obscure. This is because the approximations inherent in present day theoretical treatments make them unable to accurately account for the few $kJ \cdot mol^{-1}$ energy differences between specific and nonspecific hydrogen bonding associations. The double helical segments of many RNAs, however, contain occasional non-Watson–Crick base pairs, most often G·U, which have functional as well as structural significance (e.g., Sections 30-2B and D).

Hydrogen Bonds Do Not Stabilize DNA

It is clear that hydrogen bonding is required for the specificity of base pairing in DNA that is ultimately responsible for the enormous fidelity required to replicate DNA with almost no error (Section 31-3D). Yet, as is also true for proteins (Section 7-4B), *hydrogen bonding contributes little to the stability of the double helix.* For instance, adding the relatively nonpolar ethanol to an aqueous DNA solution, which strengthens hydrogen bonds, destabilizes the double helix as is indicated by its decreased T_m. This is because hydrophobic forces, which are largely responsible for DNA's stability (see Section 28-3D), are disrupted by nonpolar solvents. In contrast, *the hydrogen bonds between the base pairs of native DNA are replaced in denatured DNA by energetically more or less equivalent hydrogen bonds between the bases and water.*

D. Base Stacking and Hydrophobic Interactions

Purines and pyrimidines tend to form extended stacks of planar parallel molecules. This has been observed in the structures of nucleic acids (Figs. 28-5, 8, and 9) and in the several hundred reported X-ray crystal structures that contain nucleic acid bases. The bases in these structures are usually partially overlapped (e.g., Fig. 28-25). In fact, crystal structures of chemically related bases often exhibit similar stacking patterns. Apparently

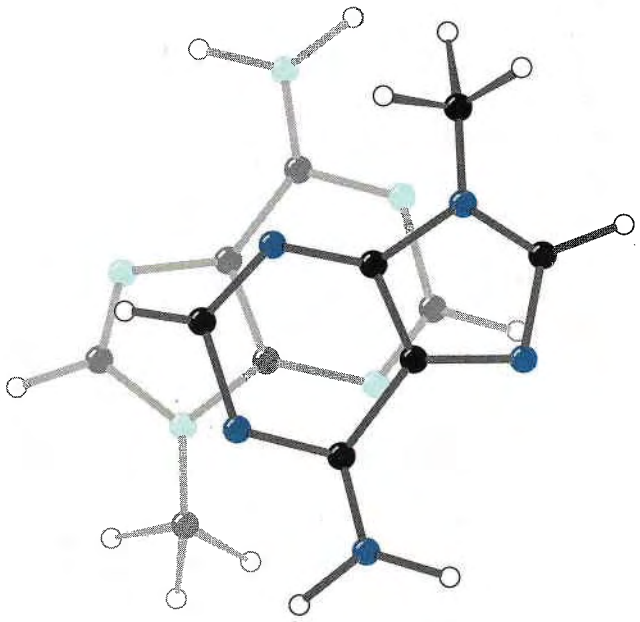


Figure 28-25
The stacking of adenine rings in the crystal structure of 9-methyladenine. The partial overlap of the rings is typical of the association between bases in crystal structures and in double helical nucleic acids. [After Stewart, R. F. and Jensen, L. H., *J. Chem. Phys.* **40,** 2071 (1964).]

stacking interactions, which in the solid state are a form of van der Waals interaction (Section 7-4A), have some specificity although certainly not as much as base pairing.

Nucleic Acid Bases Stack in Aqueous Solution

Bases aggregate in aqueous solution as has been demonstrated by the variation of osmotic pressure with concentration. The van't Hoff law of osmotic pressure is

$$\pi = RTm \qquad [28.2]$$

where π is the osmotic pressure, m is the molality of the solute (mol solute/kg solvent), R is the gas constant, and T is the temperature. The molecular mass, M, of an ideal solute can be determined from its osmotic pressure since $M = c/m$, where c = g solute/kg solvent.

If the species under investigation is of known molecular mass but aggregates in solution, Eq. [28.2] must be rewritten:

$$\pi = \phi RTm \qquad [28.3]$$

where ϕ, the **osmotic coefficient,** indicates the solute's degree of association. ϕ varies from 1 (no association) to 0 (infinite association). The variation of ϕ with m for nucleic acid bases in aqueous solution (e.g., Fig. 28-26) is consistent with a model in which the bases aggregate in successive steps:

$$A + A \rightleftharpoons A_2 + A \rightleftharpoons A_3 + A \rightleftharpoons \cdots \rightleftharpoons A_n$$

where n is at least 5 (if the reaction goes to completion,

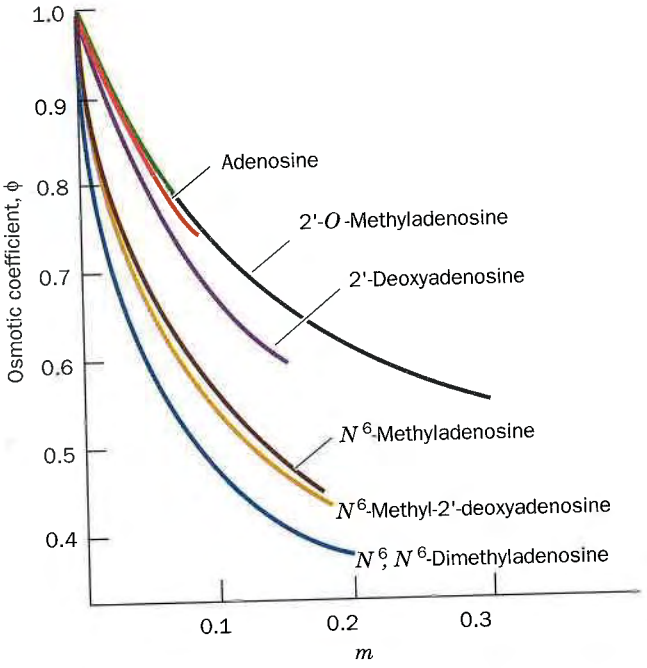
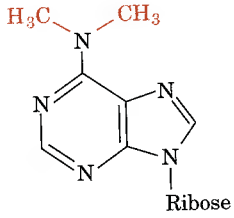


Figure 28-26
The variation of the osmotic coefficient ϕ with the molal concentrations m of adenosine derivatives in H_2O. The decrease of ϕ with increasing m indicates that these derivatives aggregate in solution. [After Broom, A. D., Schweizer, M. P., and Ts'o, P. O. P., *J. Am. Chem. Soc.* **89,** 3613 (1967).]

$\phi = 1/n$). This association cannot be a result of hydrogen bonding since **N^6,N^6-dimethyladenosine,**



N^6,N^6-Dimethyladenosine

which cannot form interbase hydrogen bonds, has a greater degree of association than does adenosine (Fig. 28-26). Apparently *the aggregation arises from the formation of stacks of planar molecules.* This model is corroborated by proton NMR studies: The directions of the aggregates' chemical shifts are compatible with a stacked but not a hydrogen bonded model. The stacking associations of monomeric bases are not observed in nonaqueous solutions.

Single-stranded polynucleotides also exhibit stacking interactions. For example, poly(A) shows a broad increase of UV absorbance with temperature (Fig. 28-27*a*). This hyperchromism is independent of poly(A) concentration so that it cannot be a consequence of intermolecular aggregation. Likewise, it is not due to intramolecular hydrogen bonding because poly(N^6,N^6-dimethyl A) has a greater degree of hyperchromism than does poly(A). The hyperchromism must therefore arise from some sort of stacking associations within a single strand that melt out with increasing temperature. This is not a very cooperative process as is indicated by the broadness of the melting curve and the observation that short polynucleotides, including dinucleoside phosphates such as ApA, exhibit similar melting curves (Fig. 28-27*b*).

Nucleic Acid Structures Are Stabilized by Hydrophobic Forces

Stacking associations in aqueous solutions are largely stabilized by hydrophobic forces. One might reasonably suppose that hydrophobic interactions in nucleic acids are similar in character to those that stabilize protein structures. However, closer examination reveals that these two types of interactions are qualitatively different in character. Thermodynamic analysis of dinucleoside phosphate melting curves in terms of the reaction

Dinucleoside phosphate *(unstacked)* $\rightleftharpoons$ dinucleoside phosphate *(stacked)*

(Table 28-4) indicates that *base stacking is enthalpically driven and entropically opposed. Thus the hydrophobic interactions responsible for the stability of base stacking associations in nucleic acids are diametrically opposite in character to those that stabilize protein structures* (which are enthalpically opposed and entropically driven; Section 7-4C). This is reflected in the differing structural properties of these interactions. For example, the aromatic side chains of proteins are almost never stacked and the crystal structures of aromatic hydrocarbons such as benzene, which resemble these side chains, are characteristically devoid of stacking interactions.

Hydrophobic forces in nucleic acids are but poorly understood. The observation that they are different in character from the hydrophobic forces that stabilize proteins

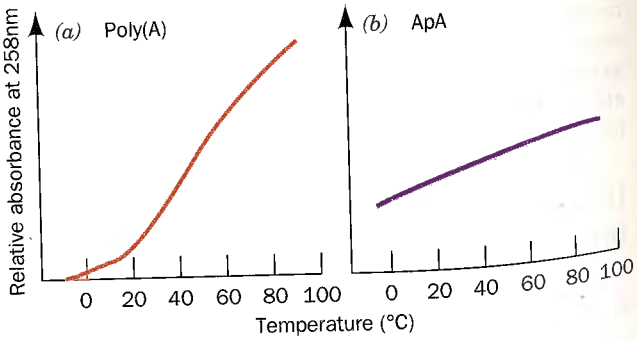


Figure 28-27
The broad temperature range of hyperchromic shifts at 258 nm of (*a*) poly(A) and (*b*) ApA is indicative of noncooperative conformational changes in these substances. Compare this figure with Fig. 28-15. [After Leng, M. and Felsenfeld, G., *J. Mol. Biol.* **15,** 457 (1966).]

Table 28-4
Thermodynamic Parameters for the Reaction

Dinucleoside phosphate *(unstacked)* $\rightleftharpoons$ *dinucleoside phosphate (stacked)*

Dinucleoside Phosphate	$\Delta H_{stacking}$ (kJ·mol^{-1})	$-T\Delta S_{stacking}$ (kJ·mol^{-1} at 25°C)
ApA	−22.2	24.9
ApU	−35.1	39.9
GpC	−32.6	34.9
CpG	−20.1	21.2
UpU	−32.6	36.2

Source: Davis, R. C. and Tinoco, I., Jr., *Biopolymers* **6**, 230 (1968).

is nevertheless not surprising because the nitrogenous bases are considerably more polar than the hydrocarbon residues of proteins that participate in hydrophobic bonding. There is, however, no theory available that adequately explains the nature of hydrophobic forces in nucleic acids (our understanding of hydrophobic forces in proteins, it will be recalled, is similarly incomplete). They are complex interactions of which base stacking is probably a significant component. Whatever their origins, hydrophobic forces are of central importance in determining nucleic acid structures.

E. Ionic Interactions

Any theory of the stability of nucleic acid structures must take into account the electrostatic interactions of their charged phosphate groups. Unfortunately, the theory of polyelectrolytes is, as yet, incapable of making reliable predictions of molecular conformations. We can, however, make experimental observations.

The melting temperature of duplex DNA increases with the cation concentration because these ions electrostatically shield the anionic phosphate groups from each other. The observed relationship for Na^+ is

$$T_m = 41.1\ X_{G+C} + 16.6 \log[Na^+] + 81.5 \quad [28.4]$$

where X_{G+C} is the mole fraction of G·C base pairs (recall that T_m increases with the G + C content); the equation is valid in the ranges $0.3 < X_{G+C} < 0.7$ and $10^{-3}M < [Na^+] < 1.0M$. Other monovalent cations such as Li^+ and K^+ have similar nonspecific interactions with phosphate groups. Divalent cations, such as Mg^{2+}, Mn^{2+}, and Co^{2+}, in contrast, specifically bind to phosphate groups so that *divalent cations are far more effective shielding agents for nucleic acids than are monovalent cations.* For example, an Mg^{2+} ion has an influence on the DNA double helix comparable to that of 100 to 1000 Na^+ ions. Indeed, enzymes that mediate reactions with nucleic acids or just nucleotides (e.g., ATP) usually require Mg^{2+} for activity.

4. NUCLEIC ACID FRACTIONATION

In Chapter 5 we considered the most commonly used procedures for isolating and, to some extent, characterizing proteins. Most of these methods, often with some modification, are also regularly used to fractionate nucleic acids according to size, composition, and sequence. There are also many techniques that are applicable only to nucleic acids. In this section we shall outline some of the most useful of the separation procedures that are specific for nucleic acids.

A. Solution Methods

Nucleic acids are invariably associated with proteins. Once cells have been broken open (Section 5-1B), the nucleic acids must be deproteinized. This may be accomplished by shaking (very gently if high molecular mass DNA is being isolated) the protein–nucleic acid mixture with a phenol solution and/or a $CHCl_3$–isoamyl alcohol mixture so that the protein precipitates and can be removed by centrifugation. Alternatively, the protein can be dissociated from the nucleic acids by detergents, guanidinium chloride, or high salt concentrations, or it can be enzymatically degraded by proteases such as pronase. In all cases, the nucleic acids, a mixture of RNA and DNA, can then be isolated by precipitation with ethanol. The RNA can be recovered from such precipitates by treating them with pancreatic DNase to eliminate the DNA. Conversely, the DNA can be freed of RNA by treatment with RNase. Alternatively, RNA and DNA may be separated by ultracentrifugation (Section 28-4D).

In all these and subsequent manipulations, the nucleic acids must be protected from degradation by nucleases that occur both in the experimental materials and on human hands. Nucleases may be inhibited by the presence of chelating agents such as EDTA, which sequester the divalent metal ions that nucleases require for activity. In cases where no nuclease activity can be tolerated, all glassware must be autoclaved to heat denature the nucleases and the experimenter should wear plastic gloves. Nevertheless, nucleic acids are generally easier to handle than proteins because their lack of a tertiary structure, in most cases, makes them relatively tolerant of extreme conditions.

B. Chromatography

Many of the chromatographic techniques that are used to separate proteins (Section 5-3) are also applicable to nucleic acids. Paper chromatography and thin layer chromatography are useful in fractionating oligonucleotides. They have been largely replaced, however, by the more powerful techniques of HPLC, partic-

ularly those using reverse-phase chromatography. Larger nucleic acids are often separated by procedures that include ion exchange chromatography and gel filtration chromatography.

Hydroxyapatite Binds Double-Stranded DNA More Tightly Than Single-Stranded DNA

Hydroxyapatite (a form of calcium phosphate; Section 5-3E) is particularly useful in the chromatographic purification and fractionation of DNA. Double-stranded DNA binds to hydroxyapatite more tightly than do most other molecules. Consequently DNA can be rapidly isolated by passing a cell lysate through a hydroxyapatite column, washing the column with a phosphate buffer of concentration low enough to release only the RNA and proteins, and then eluting the DNA with a concentrated phosphate solution.

Single-stranded DNA elutes from hydroxyapatite at a lower phosphate concentration than does double-stranded DNA (Fig. 28-28). This phenomenon forms the basis of a technique, known as **thermal chromatography,** for separating DNA according to its base composition. A hydroxyapatite column to which double-stranded DNA is bound is eluted with a phosphate buffer that releases only single-stranded DNA while the temperature of the column is gradually increased. As the DNA melts and is converted to the single-stranded form it is eluted from the column. Since the T_m of a duplex DNA varies with its G + C content (Eq. [28.4]), thermal chromatography permits the fractionation of double-stranded DNA according to its base composition.

Messenger RNAs Can Be Isolated by Affinity Chromatography

Affinity chromatography is useful in isolating specific nucleic acids. For example, most eukaryotic messenger RNAs (mRNAs) have a poly(A) sequence at their 3′ ends (Section 28-4A). They can be isolated on agarose or cellulose to which poly(U) is covalently attached. The poly(A) sequences specifically bind to the complementary poly(U) in high salt and at low temperatures and can later be released by altering these conditions. Moreover, if the (partial) sequence of an mRNA is known (e.g., as deduced from the corresponding protein's amino acid sequence), the complementary DNA strand may be synthesized (via methods discussed in Section 28-7) and used to isolate that particular mRNA.

C. Electrophoresis

Nucleic acids of a given type may be separated by polyacrylamide gel electrophoresis (Sections 5-4B and C) because their electrophoretic mobilities in such gels vary inversely with their molecular masses. However, DNAs of more than a few thousand base pairs cannot penetrate even a weakly cross-linked polyacrylamide gel. This difficulty is partially overcome through the use

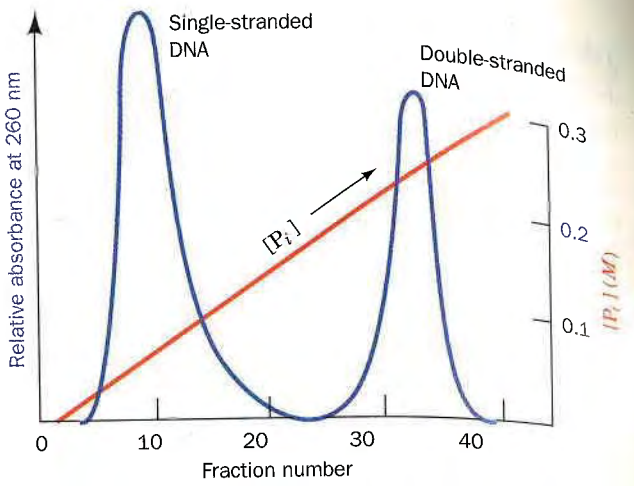


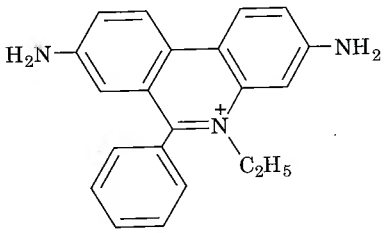
Figure 28-28
The chromatographic separation of single-stranded and duplex DNAs on hydroxyapatite by elution with a solution of increasing phosphate concentration.

of agarose gels. By using gels with an appropriately low agarose content, relatively large DNAs in various size ranges may be fractionated. In this manner, **plasmids** (small, autonomously replicating DNA molecules that occur in bacteria and yeast), for example, may be separated from the larger chromosomal DNA of bacteria.

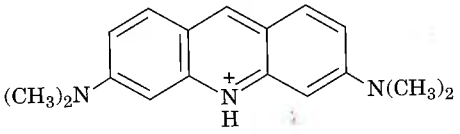
Very Large DNAs Are Separated by Pulsed-Field Gel Electrophoresis

The sizes of the DNAs that can be separated by conventional gel electrophoresis are limited to ~100,000 bp, even when gels containing as little as 0.1% agarose (which makes an extremely fragile gel) are used. However, the recent development of **pulsed-field gel electrophoresis (PFG)** by Charles Cantor and Cassandra Smith has extended this limit to DNAs with up to 10 million bp (6.6 million kD). The electrophoresis apparatus used in PFG has two or more pairs of electrodes arrayed around the periphery of an agarose slab gel. The different electrode pairs are sequentially pulsed for times varying from 0.1 to 1000 s depending on the sizes of the DNAs being separated. Gel electrophoresis of DNA requires that these elongated molecules worm their way through the gel's labyrinthine channels more or less in the direction from the cathode to the anode. If the direction of the electric field abruptly changes, these DNAs must reorient their long axes along the new direction of the field before they can continue their passage through the gel. The time required to reorient very long gel-embedded DNA molecules evidently increases with their size. Consequently, a judicious choice of electrode distribution and pulse lengths causes shorter DNAs to migrate through the gel faster than longer DNAs, thereby effecting their separation.

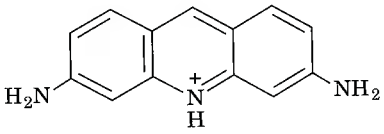
Figure 28-29
An agarose gel electrophoretogram of double helical DNA. After electrophoresis, the gel was soaked in a solution of ethidium bromide, washed, and photographed under UV light. The fluorescence of the ethidium cation is strongly enhanced by binding to DNA so that each fluorescent band marks a different sized DNA fragment. The three parallel lanes contain identical DNA samples so as to demonstrate the technique's reproducibility. [Photo by Elizabeth Levine. From Freifelder, D., *Biophysical Chemistry. Applications to Biochemistry and Molecular Biology* (2nd ed.), p. 294, W. H. Freeman (1982). Used by permission.]



Ethidium



Acridine orange



Proflavin

These dyes bind to duplex DNA by **intercalation** (slipping in between the stacked base pairs) where they exhibit a fluorescence under UV light that is far more intense than that of the free dye. As little as 50 ng of DNA may be detected in a gel by staining it with ethidium bromide (Fig. 28-29). Single-stranded DNA and RNA also stimulate the fluorescence of ethidium but to a lesser extent than does duplex DNA.

Duplex DNA Is Detected by Selectively Staining It with Intercalation Agents

The various DNA bands in a gel must be detected if they are to be isolated. Double-stranded DNA is readily stained by planar aromatic cations such as **ethidium ion, acridine orange,** or **proflavin.**

Southern Blotting Identifies DNAs with Specific Sequences

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. 28-30). This pro-

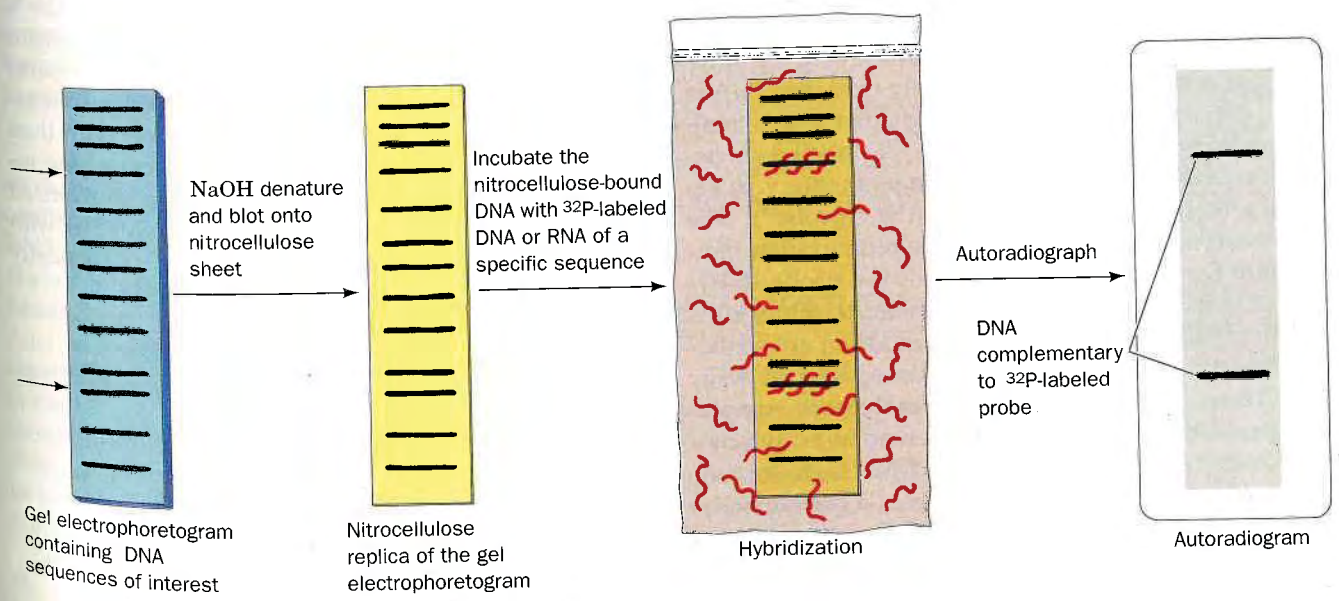


Figure 28-30
The detection of DNAs containing specific base sequences by the Southern transfer technique.

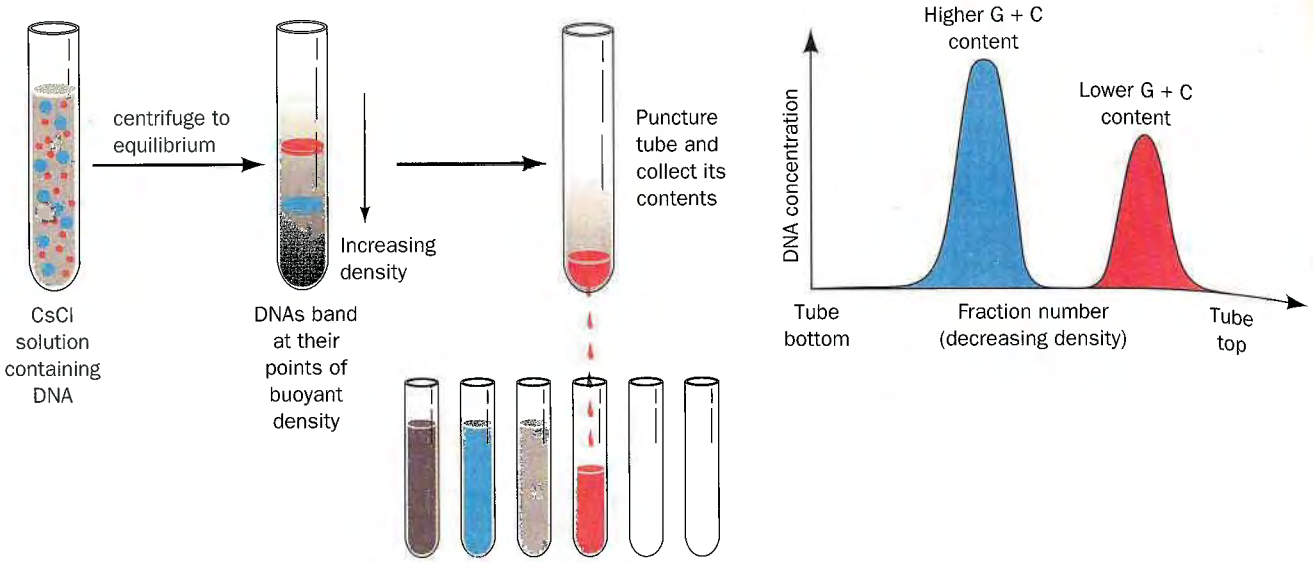
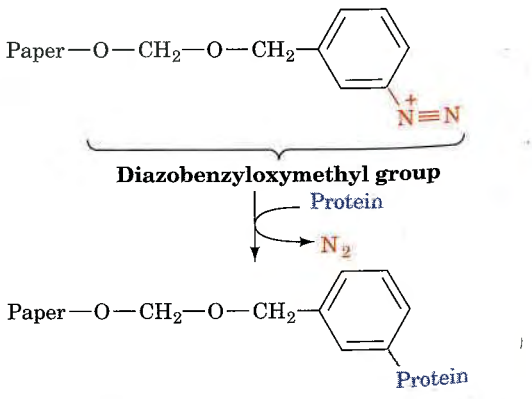


Figure 28-31
The separation of DNAs according to base composition by equilibrium density gradient ultracentrifugation in CsCl solution. An initially 8*M* CsCl solution forms a density gradient that varies linearly from ~1.80 g·cm^{-3} at the bottom of the centrifuge tube to ~1.55 g·cm^{-3} at the top. The amount of DNA in each fraction is estimated from its UV absorbance, usually at 260 nm.

cedure takes advantage of the valuable property of nitrocellulose that it tenaciously binds single-stranded but not duplex DNA. Following the gel electrophoresis of double-stranded DNA, the gel is soaked in 0.5*M* NaOH solution, which converts the DNA to the 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 named **electroblotting**). After vacuum drying the nitrocellulose at 80°C, which permanently fixes the DNA in place, the nitrocellulose sheet is moistened with a minimal quantity of solution containing ^{32}P-labeled single-stranded DNA or RNA that is complementary in sequence to the DNA of interest (the "probe"). The moistened filter is held at a suitable renaturation temperature for several hours to permit the probe to hybridize to its target sequence(s), washed to remove the unbound radioactive probe, dried, and then autoradiographed by placing it for a time 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. A DNA segment containing a particular base sequence (e.g., a gene specifying a certain protein) may, in this manner, be detected and isolated. Specific DNAs may likewise be detected by linking the probe to an enzyme that generates a colored or fluorescent deposit on the blot. Such nonradioactive detection techniques are desirable in a clinical setting because of the health hazards, disposal problems, and the more cumbersome nature of autoradiographic methods.

Northern and Western Blotting, Respectively, Detect RNAs and Proteins

Variations of Southern transfer, which are punningly called **northern transfer (northern blotting)** and **western transfer (western blotting),** respectively detect specific RNAs and proteins. In northern blotting, RNA is immobilized on nitrocellulose paper and detected through the use of complementary radiolabeled RNA or DNA probes. In western blotting, a protein mixture is bound to nitrocellulose paper and specific proteins identified by their binding of antibodies raised against them. Nitrocellulose, however, binds proteins so tenaciously that, in some cases, it may interfere with their immunochemical identification. In such cases, the nitrocellulose paper can be replaced by paper derivatized with diazobenzyloxymethyl groups, which react with the proteins' primary amino groups so as to covalently couple them to the paper.



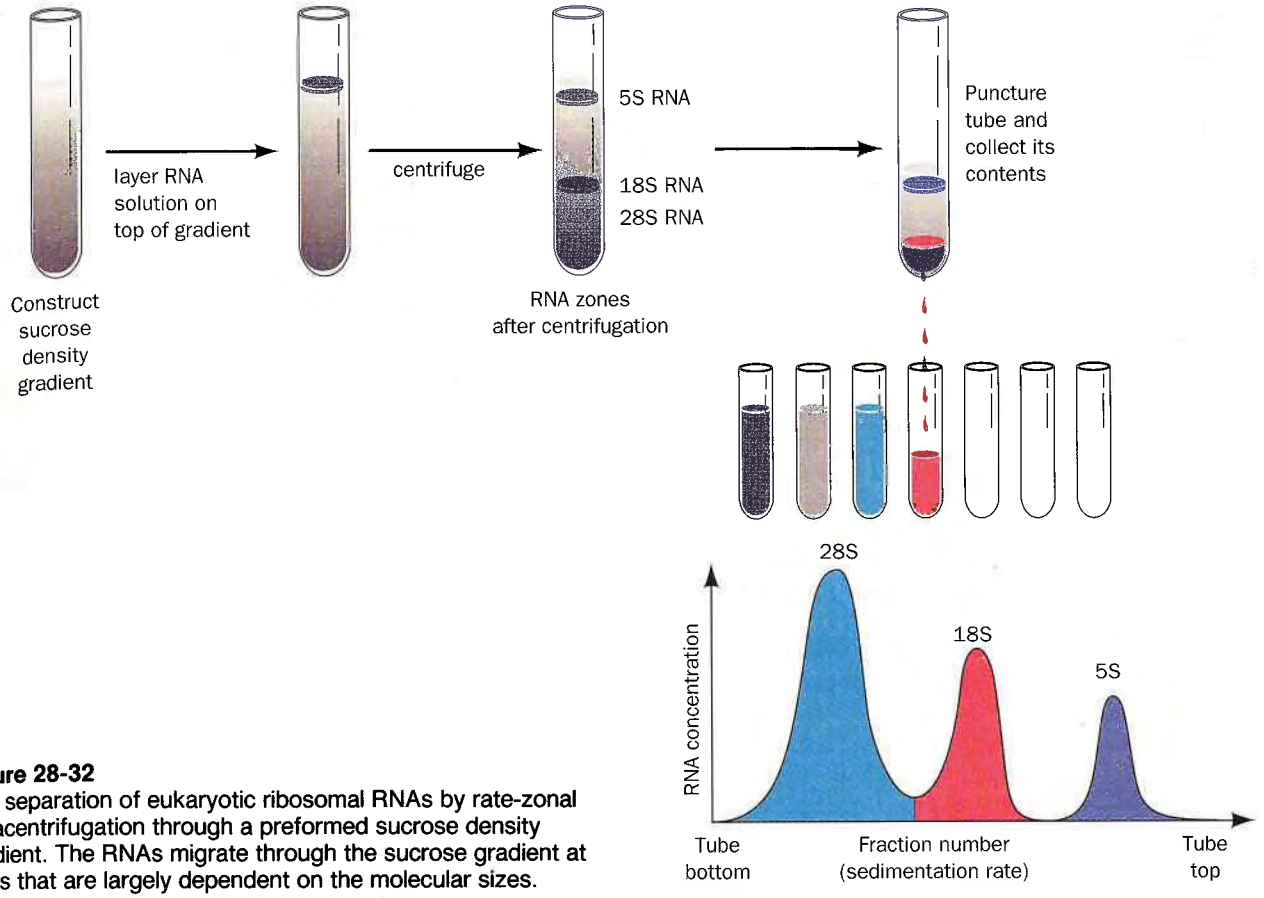


Figure 28-32
The separation of eukaryotic ribosomal RNAs by rate-zonal ultracentrifugation through a preformed sucrose density gradient. The RNAs migrate through the sucrose gradient at rates that are largely dependent on the molecular sizes.

D. Ultracentrifugation

Equilibrium density gradient ultracentrifugation (Fig. 28-31; Section 5-5B) in CsCl constitutes one of the most commonly used DNA separation procedures. The bouyant density, ρ, of double-stranded Cs^+ DNA depends on its base composition:

$$\rho = 1.660 + 0.098\ X_{G+C} \qquad [28.5]$$

so that a CsCl density gradient fractionates DNA according to its base composition. For example, eukaryotic DNAs often contain minor fractions that band separately from the major species. Some of these **satellite bands** consist of mitochondrial and chloroplast DNAs. Another important class of satellite DNA is composed of **repetitive sequences** that are short segments of DNA tandemly (one behind the other) repeated hundreds, thousands, and in some cases, millions of times in a chromosome (Section 33-2B). Likewise, plasmids may be separated from bacterial chromosomal DNA by equilibrium density gradient ultracentrifugation.

Single-stranded DNA is $\sim 0.015\ g \cdot cm^{-3}$ denser than the corresponding double-stranded DNA so that the two may be separated by equilibrium density gradient ultracentrifugation. RNA is too dense to band in CsCl but does so in Cs_2SO_4 solutions. RNA–DNA hybrids will band in CsCl but at a higher density than the corresponding duplex DNA.

RNA may be fractionated by rate-zonal ultracentrifugation through a sucrose gradient (Fig. 28-32; Section 5-5B). RNAs are separated by this technique largely on the basis of their size. In fact, ribosomal RNA, which constitutes the major portion of cellular RNA, is classified according to its sedimentation rate; for example, the RNA of the *E. coli* small ribosomal subunit is known as 16S RNA (Section 30-3A).

5. SUPERCOILED DNA

The circular genetic maps of viruses and bacteria implies that their chromosomes are likewise circular. This conclusion has been confirmed by electron micrographs in which circular DNAs are seen (Fig. 28-33). Some of these circular DNAs have a peculiar twisted appearance, a phenomenon that is known equivalently as **supercoiling, supertwisting,** or **superhelicity.** Supercoiling arises from a biologically important topological property of covalently closed circular duplex DNA that is the subject of this section. It is occasionally referred to as DNA's tertiary structure.

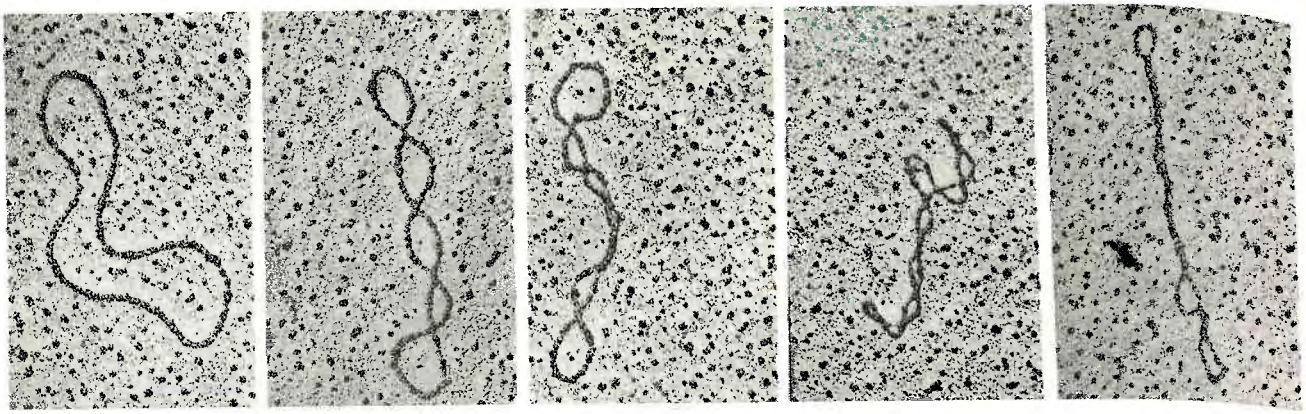


Figure 28-33
Electron micrographs of circular duplex DNAs that vary in their conformations from no supercoiling (*left*) to tightly supercoiled (*right*). [Electron micrographs by Laurien Polder. From Kornberg, A., *DNA Replication*, p. 29, W. H. Freeman (1980). Used by permission.]

A. Superhelix Topology

Consider a double helical DNA molecule in which both strands are covalently joined to form a circular duplex molecule as is diagrammed in Fig. 28-34 (each strand can only be joined to itself because the strands are antiparallel). *A geometric property of such an assembly is that its number of coils cannot be altered without first cleaving at least one of its polynucleotide strands.* You can easily demonstrate this to yourself with a buckled belt in which each edge of the belt represents a strand of DNA. The number of times the belt is twisted before it is buckled cannot be changed without unbuckling or cutting the belt (cutting a polynucleotide strand).

This phenomenon is mathematically expressed

$$L = T + W \qquad [28.6]$$

in which:

1. L, the **linking number,** is the number of times that one DNA strand winds about the other. This integer quantity is most easily counted when the molecule's duplex axis is constrained to lie in a plane (see below). However, *the linking number is invariant no matter how the circular molecule is twisted or distorted so long as both its polynucleotide strands remain covalently intact; the linking number is therefore a topological property of the molecule.*

2. T, the **twist,** is the number of complete revolutions that one polynucleotide strand makes about the duplex axis in the particular conformation under consideration. By convention, T is positive for right-handed duplex turns so that, for B-DNA in solution, the twist is normally the number of base pairs divided by 10.5 (the number of base pairs per turn of the B-DNA double helix under physiological conditions; see Section 28-5B).

3. W, the **writhing number,** is the number of turns that the duplex axis makes about the superhelix axis in the conformation of interest. *It is a measure of the DNA's superhelicity.* The difference between writhing and twisting is illustrated by the familiar example in Fig. 28-35. $W = 0$ when DNA's duplex axis is constrained to lie in a plane (e.g., Fig. 28-34); then $L = T$ so that L may be evaluated by counting the DNA's duplex turns.

The two DNA conformations diagrammed on the right of Fig. 28-36 are topologically equivalent; that is, they have the same linking number, L, but differ in their twists and writhing numbers. Note that T and W need not be integers, only L.

Since L is constant in an intact duplex DNA circle, for

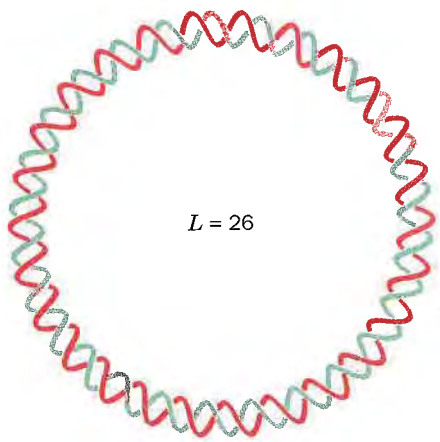


Figure 28-34
A schematic diagram of covalently closed circular duplex DNA that has 26 double helical turns. Its two polynucleotide strands are said to be **topologically bonded** to each other because, although they are not covalently linked, they cannot be separated without breaking covalent bonds.

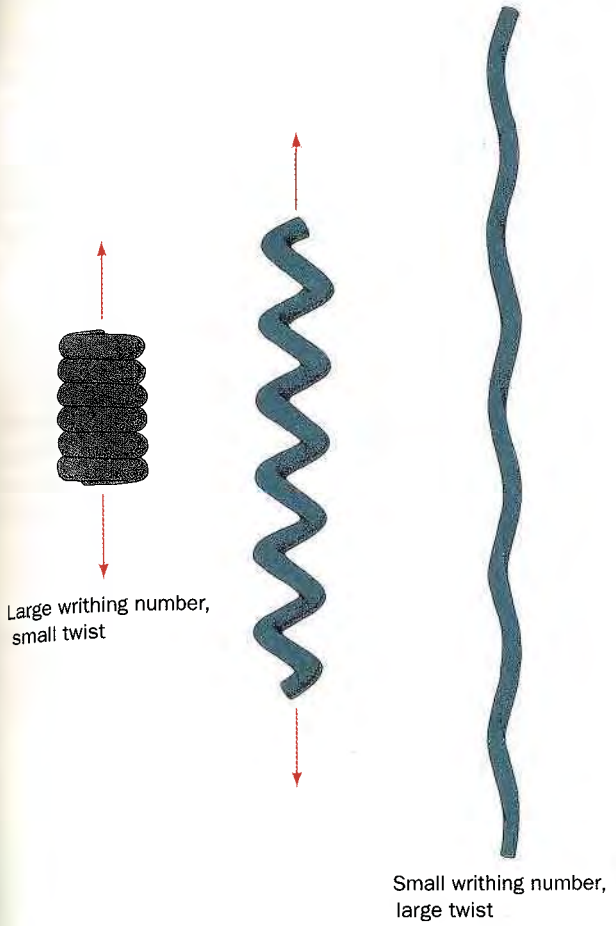


Figure 28-35
The difference between writhing and twist as demonstrated by a coiled telephone cord. In its relaxed state (*left*), the cord is in a helical form that has a large writhing number and a small twist. As the coil is pulled out (*middle*) until it is nearly straight (*right*), its writhing number becomes small as its twist becomes large.

every new double helical twist, ΔT, there must be an equal and opposite superhelical twist; that is, $\Delta W = -\Delta T$. For example, a closed circular DNA without supercoils (Fig. 28-36, *upper right*) can be converted to a negatively supercoiled conformation (Fig. 28-36, *lower right*) by winding the duplex helix the same number of positive (right handed) turns.

Supercoils May Be Toroidal or Interwound

A supercoiled duplex may assume two topologically equivalent forms:

1. **A toroidal helix** in which the duplex axis is wound as if about a cylinder (Fig. 28-37*a*).
2. An **interwound helix** in which the duplex axis is twisted around itself (Fig. 28-37*b*).

Note that these two interconvertible superhelical forms have opposite handedness. Since left-handed toroidal turns may be converted to left-handed duplex turns (see Fig. 28-35), left-handed toroidal turns and right-handed

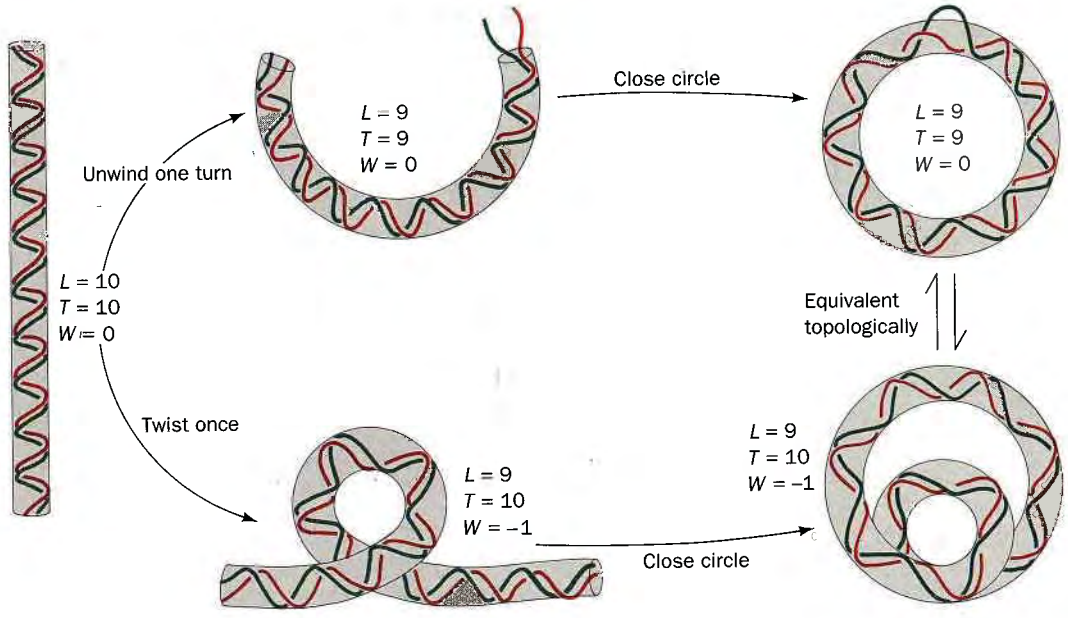


Figure 28-36
Two ways of introducing one supercoil into a DNA with 10 duplex turns. The two closed circular forms shown (*right*) are topologically equivalent; that is, they are interconvertible without breaking any covalent bonds. The linking number *L*, twist *T*, and writhing number *W* are indicated for each form. Strictly speaking, the linking number is only defined for a covalently closed circle.

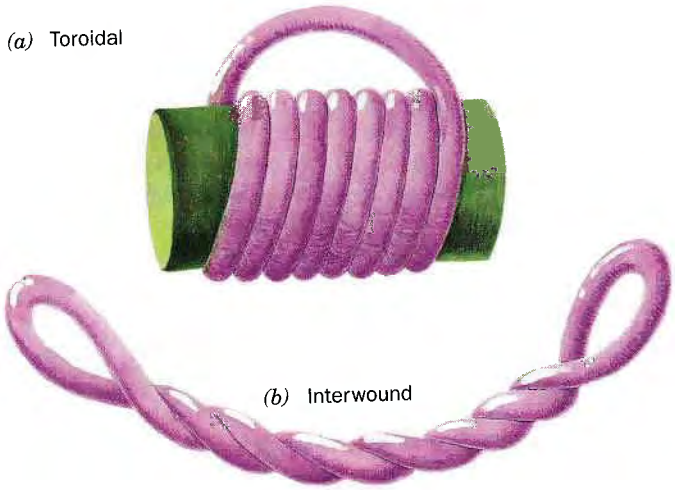


Figure 28-37
A rubber tube that has been (*a*) toroidally coiled around a cylinder with its ends joined such that it has no twist, jumps to (*b*) an interwound helix with the opposite handedness when the cylinder is removed. Neither the linking number, twist, nor writhing number are changed in this transformation.

interwound turns both have negative writhing numbers. Thus an underwound duplex ($T <$ number of bp/10.5), for example, will tend to develop right-handed interwound or left-handed toroidal superhelical turns when the constraints causing it to be underwound are released (the molecular forces in a DNA double helix promote its winding to its normal number of helical turns).

Supercoiled DNA Is Relaxed by Nicking One Strand

Supercoiled DNA may be converted to **relaxed circles** (as appears in the left-most panel of Fig. 28-33) by treatment with **pancreatic DNase I,** an **endonuclease** (an enzyme that cleaves phosphodiester bonds within a polynucleotide strand), which cleaves only one strand of a duplex DNA. *One single-strand nick is sufficient to relax a supercoiled DNA.* This is because the sugar–phosphate chain opposite the nick is free to swivel about its backbone bonds (Fig. 28-18) so as to change the molecule's linking number and thereby alter its superhelicity. Supercoiling builds up elastic strain in a DNA circle, much as it does in a rubber band. This is why the relaxed state of a DNA circle is not supercoiled.

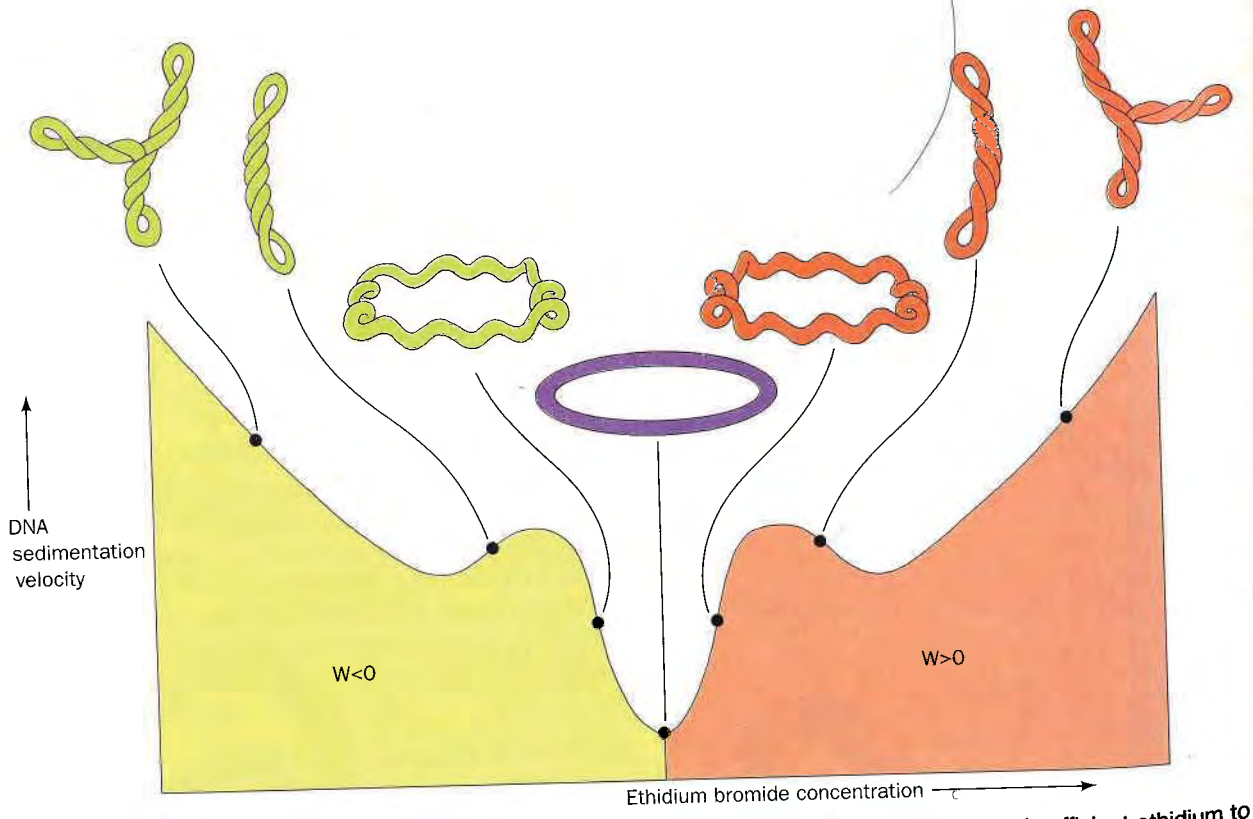


Figure 28-38
The sedimentation rate of closed circular duplex DNA as a function of ethidium bromide concentration. The intercalation of ethidium between the base pairs locally unwinds the double helix which, since the linking number of the circle is constant, is accompanied by an equivalent increase in the writhing number. As the superhelix unwinds, it becomes less compact and sediments more slowly. At the low point on the curve, the DNA circles have bound sufficient ethidium to become fully relaxed. As the ethidium concentration is further increased, the DNA supercoils in the opposite direction. The supertwisted appearances of the depicted DNAs have been verified by electron microscopy. [After Bauer, W. R., Crick, F. H. C., and White, J. H., *Sci. Am.* **243**(1): 129 (1980). Copyright © 1980 by Scientific American, Inc.]

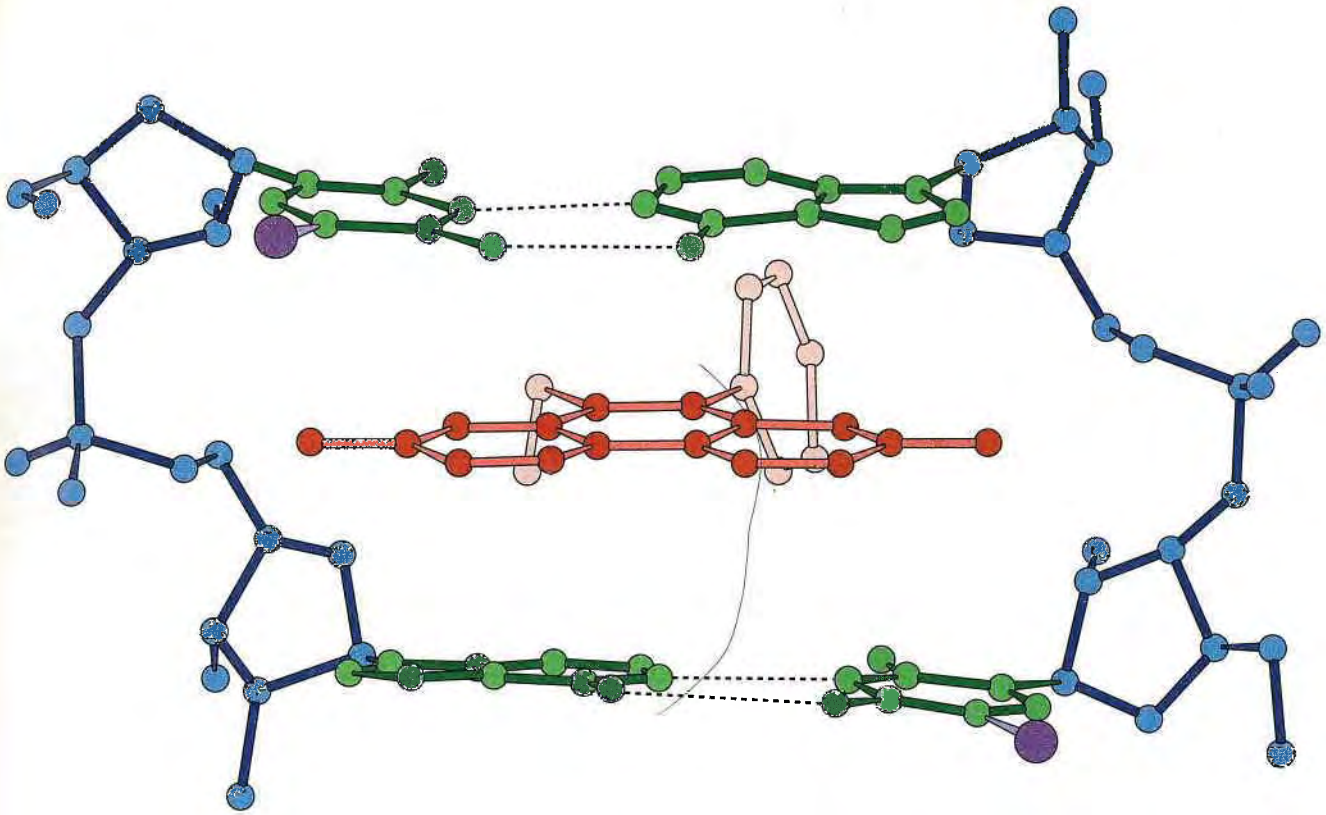


Figure 28-39
The X-ray structure of a complex of ethidium with 5-iodo UpA. Ethidium (*red*) intercalates between the base pairs of the double helically paired dinucleoside phosphate and thereby provides a model for the binding of ethidium to duplex DNA. [After Tsai, C.-C., Jain, S. C., and Sobell, H. M., *Proc. Natl. Acad. Sci.* **72,** 629 (1975).]

B. Measurements of Supercoiling

Supercoiled DNA, far from being just a mathematical curiosity, has been widely observed in nature. In fact, its discovery in polyoma virus DNA by Jerome Vinograd stimulated the elucidation of the topological properties of superhelices rather than *vice versa.*

Intercalating Agents Control Supercoiling by Unwinding DNA

All naturally occurring DNA circles are underwound; that is, their linking numbers are less than those of their corresponding relaxed circles. This phenomenon has been established by observing the effect of ethidium binding on the sedimentation rate of circular DNA (Fig. 28-38). Intercalating agents such as ethidium alter a circular DNA's degree of superhelicity because they cause the DNA double helix to unwind by $\sim 26°$ at the site of the intercalated molecule (Fig. 28-39). $W < 0$ in an unconstrained underwound circle because of the tendency of a duplex DNA to maintain its normal twist of 1 turn/10.5 bp. The titration of a DNA circle by ethidium unwinds the duplex (decreases T), which must be accompanied by a compensating increase in W. This, at first, lessens the superhelicity of an underwound circle. However, as the circle binds more and more ethidium, its value of W passes through zero (relaxed circles) and then becomes positive so that the circle again becomes superhelical. Thus the sedimentation rate of underwound DNAs, which is a measure of their compactness and therefore their superhelicity, passes through a minimum as the ethidium concentration increases. This is what is observed with native DNAs (Fig. 28-38). In contrast, the sedimentation rate of an overwound circle would only increase with increasing ethidium concentration.

DNAs Are Separated According to Their Linking Number by Gel Electrophoresis

Gel electrophoresis also separates similar molecules on the basis of their compactness so that the rate of migration of a circular duplex DNA increases with its degree of superhelicity. The agarose gel electrophoresis pattern of a population of chemically identical DNA molecules with different linking numbers therefore consists of a series of discrete bands (Fig. 28-40). The molecules in a given band all have the same linking number and differ from those in adjacent bands by $\Delta L = \pm 1$.

Comparison of the electrophoretic band patterns of **simian virus 40 (SV40)** DNA that had been enzymatic-

Figure 28-40
The agarose gel electrophoresis pattern of SV40 DNA. Lane 1 contains the negatively supercoiled native DNA (*lower band*). In lanes 2 and 3, the DNA has been exposed for 5 and 30 min, respectively, to an enzyme, known as a Type I topoisomerase (Section 28-5C), that relaxes the supercoils one at a time. Neighboring bands contain DNAs that differ by $\Delta L = \pm 1$. [From Keller, W., *Proc. Natl. Acad. Sci.* **72,** 2553 (1975).]

ally relaxed to varying degrees and then resealed (Fig. 28-40) reveals that 26 bands separate native from fully relaxed SV40 DNAs. Native SV40 DNA therefore has $W = -26$ (although it is somewhat heterogeneous in this quantity). Since SV40 DNA consists of 5243 bp, it has 1 superhelical turn per ~19 duplex turns. Such a **superhelix density** is typical of circular DNAs from various biological sources.

DNA in Physiological Solution Has 10.5 Base Pairs Per Turn

The insertion, using genetic engineering techniques (Section 28-8B), of an additional x base pairs into a superhelical DNA will increase its linking number by $x/h°$, where $h°$ is the number of base pairs per duplex turn. Such an insertion will shift the position of a band in the DNA's gel electrophoretic pattern by $x/h°$ of the distance to the preceding band. By measuring the effects of several such insertions James Wang established that $h° = 10.5 \pm 0.1$ bp for B-DNA in solution under physiological conditions.

C. Topoisomerases

The normal biological functioning of DNA occurs only if it is in the proper topological state. In such basic biological processes as RNA transcription, DNA replication and genetic recombination, the recognition of a base sequence requires the local separation of complementary polynucleotide strands. The negative supercoiling of naturally occurring DNAs results in a torsional strain that promotes such separations since it tends to unwind the duplex helix (an increase in W must be accompanied by a decrease in T). *If DNA lacks the proper superhelical tension, the above vital processes occur quite slowly, if at all.*

The supercoiling of DNA is controlled by a remarkable group of enzymes known as ***topoisomerases.*** They are so named because they alter the topological state (linking

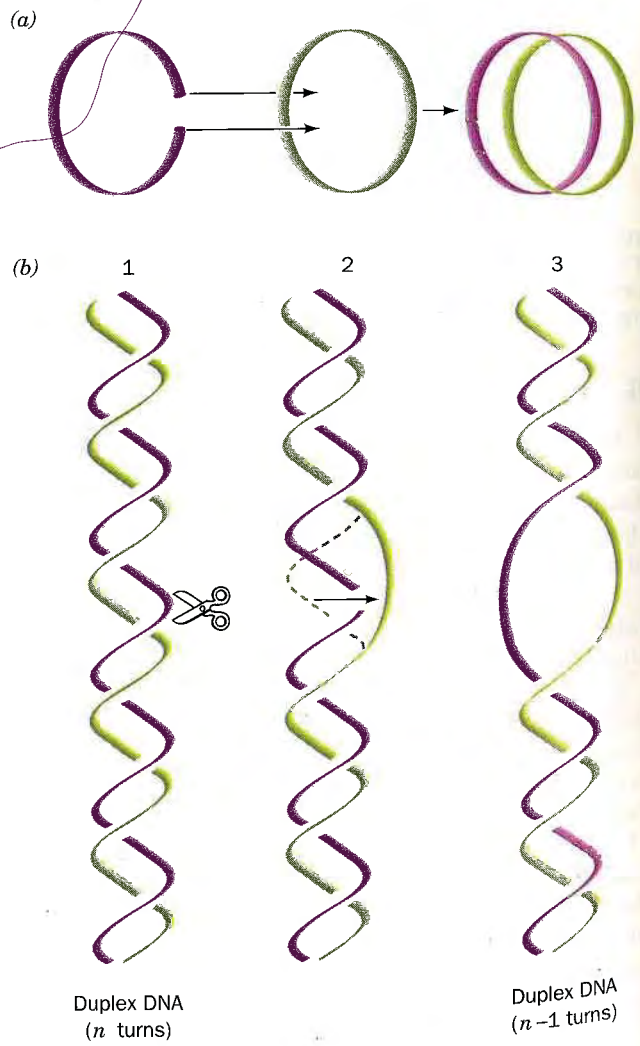


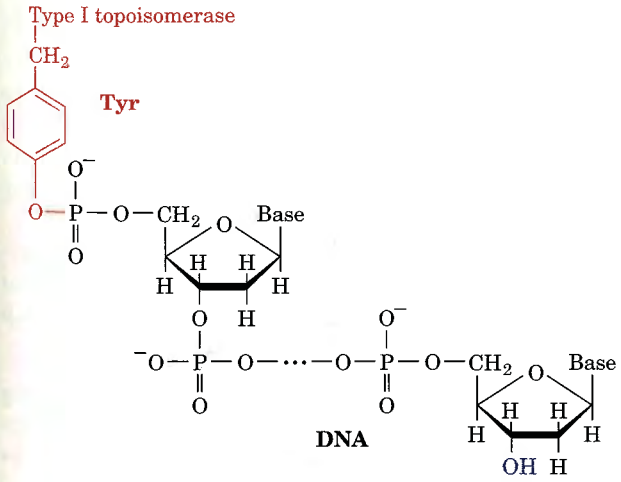
Figure 28-41
By cutting a single-stranded DNA, passing a loop of it through the break and then resealing the break, Type I topoisomerase can (*a*) catenate two single-stranded circles or (*b*) unwind duplex DNA by one turn.

number) of circular DNA but not its covalent structure. There are two classes of topoisomerases:

1. **Type I topoisomerases** act by creating transient single-strand breaks in DNA.
2. **Type II topoisomerases** act by making transient double-strand breaks in DNA.

Type I Topoisomerases Incrementally Relax Supercoiled DNA

Type I topoisomerases, which are also known as **nicking–closing enzymes,** are monomeric proteins of 100 to 120 kD that are widespread in both prokaryotes and eukaryotes. *They catalyze the relaxation of negative supercoils in DNA by increasing its linking number in increments of one turn.* The exposure of a negatively supercoiled DNA to nicking–closing enzyme sequentially increases its linking number until the supercoil is entirely relaxed. A clue to the mechanism of action of this enzyme was provided by the observation that it reversibly **catenates** (interlinks) single-stranded circles (Fig. 28-41*a*). Apparently the enzyme operates by cutting a single strand, passing a single-strand loop through the resulting gap, and then resealing the break (Fig. 28-41*b*) thereby twisting double helical DNA by one turn. In support of this hypothesis, the denaturation of prokaryotic nicking–closing enzyme that has been incubated with single-stranded circular DNA yields a linear DNA that has its 5′-terminal phosphoryl group linked to the enzyme via a phosphotyrosine diester linkage.

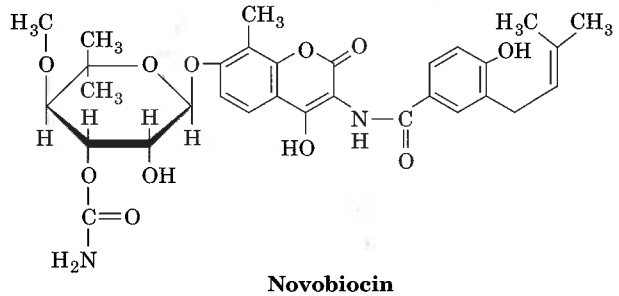


Denatured eukaryotic nicking–closing enzymes are instead linked to the 3′ end of DNA in a like manner. *By forming such covalent enzyme-DNA intermediates, the free energy of the cleaved phosphodiester bond is preserved so that no energy input is required to reseal the nick.*

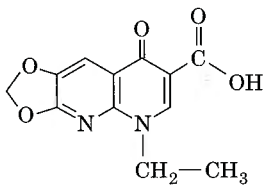
Type II Topoisomerases Supercoil DNA at the Expense of ATP Hydrolysis

Prokaryotic Type II topoisomerases, which are also known as **DNA gyrases,** are ~400-kD proteins that consist of two pairs of subunits designated *A* and *B*. *These enzymes catalyze the stepwise negative supercoiling of DNA with the concomitant hydrolysis of an ATP to ADP + P_i.* In the absence of ATP, DNA gyrase relaxes negatively supercoiled DNA but at a relatively slow rate. It can also tie knots in double-stranded circles as well as catenate them. Eukaryotic Type II topoisomerases only relax supercoils; they neither generate them nor hydrolyze ATP. DNA supercoiling in eukaryotes is generated somewhat differently (Section 33-1B).

Prokaryotic DNA gyrases are specifically inhibited by two classes of antibiotic. One of these classes includes the *Streptomyces* derived **novobiocin** and the other contains the clinically useful antibacterial agent **oxolinic acid.**



Novobiocin



Oxolinic acid

Both classes of antibiotic profoundly inhibit bacterial DNA replication and RNA transcription thereby demonstrating the importance of supercoiled DNA in these processes. Studies using antibiotic resistant *E. coli* mutants demonstrated that oxolinic acid associates with DNA gyrase's *A* subunit and novobiocin binds to its *B* subunit.

The gel electrophoretic pattern of duplex circles that have been exposed to DNA gyrase, with or without ATP, show a band pattern in which the linking numbers differ by increments of two rather than one as occurs with nicking–closing enzymes. This observation is strong evidence that *DNA gyrase acts by cutting both strands of a duplex, passing the duplex through the break and resealing it* (Fig. 28-42). This hypothesis is corroborated by the observation that when DNA gyrase is incubated with DNA and oxolinic acid, and subsequently denatured with guanidinium chloride, its *A* subunits remain covalently linked to the 5′ ends of both cut strands through phosphotyrosine linkages. Apparently oxolinic acid interferes with gyrase action by blocking the strand breaking–rejoining process. Novobiocin, on the other hand, prevents ATP from binding to the enzyme.

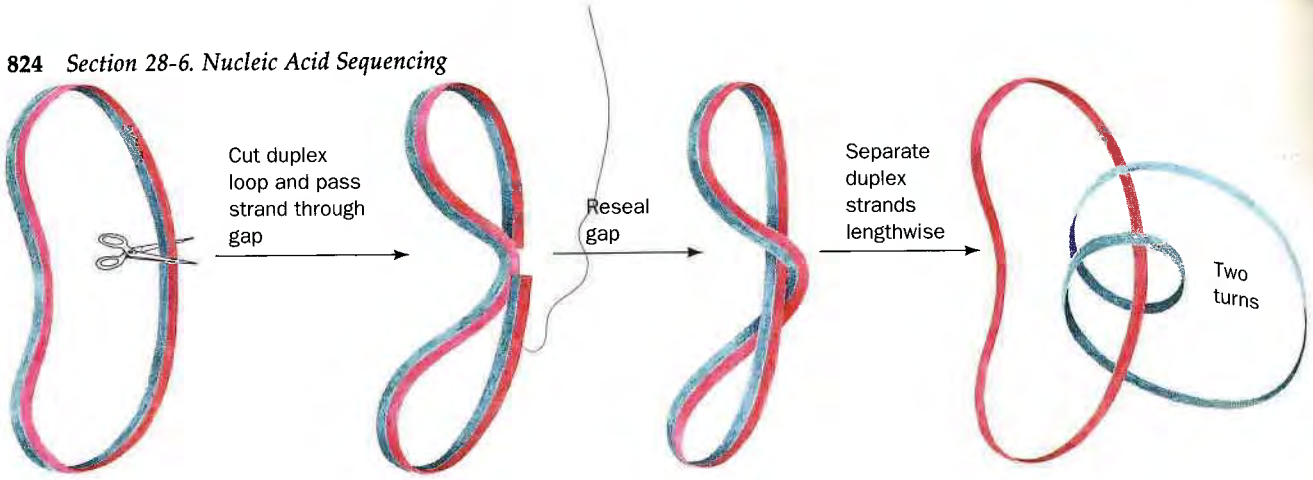
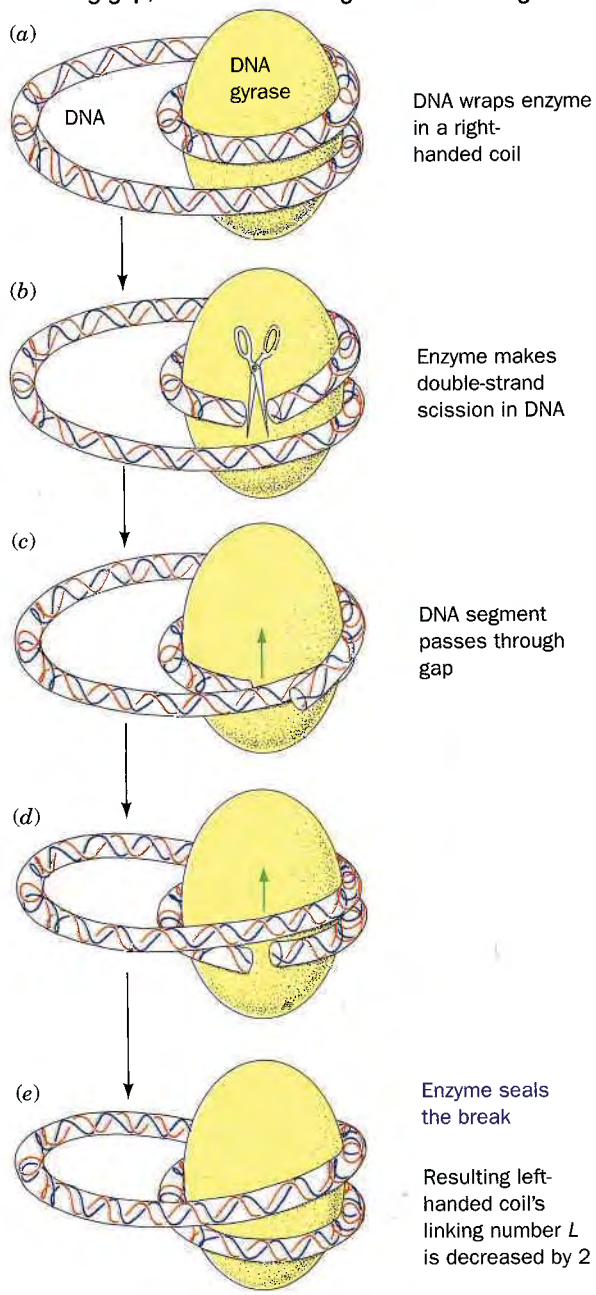


Figure 28-42
A demonstration, in which DNA is represented by a ribbon, that cutting a duplex circle, passing the strand through the resulting gap, and then resealing the break changes the linking number by two. Separating the resulting strands (slitting the ribbon along its length; *right*), indicates that one strand makes two complete revolutions about the other.



The exposure of a gyrase–DNA complex to *Staphylococcal* nuclease protects the DNA from nuclease degradation in a 140 bp fragment that is roughly centered on the gyrase cleavage site. The length of this protected fragment suggests that the DNA is wrapped around the enzyme. This observation led Nicholas Cozzarelli to propose the mechanism of gyrase–DNA action diagrammed in Fig. 28-43. It is named the **sign inversion** mechanism because it converts a right-handed toroidal supercoil to a left-handed toroidal supercoil.

6. NUCLEIC ACID SEQUENCING

The basic strategy of nucleic acid sequencing is identical to that of protein sequencing (Section 6-1). It involves:

1. The specific degradation and fractionation of the polynucleotide of interest to fragments small enough to be fully sequenced.
2. The sequencing of the individual fragments.
3. The ordering of the fragments by repeating the preceding steps using a degradation procedure that yields a set of polynucleotide fragments that overlap the cleavage points in the first such set.

Before about 1975, however, nucleic acid sequencing techniques lagged far behind those of protein sequencing largely because there were no available endonucle-

Figure 28-43
The sign inversion mechanism of DNA gyrase action. The duplex DNA is initially wrapped about the enzyme in a right-handed toroidal coil **(1).** The enzyme then makes a double-strand scission in the DNA **(2),** passes a DNA segment through the gap **(3, 4),** and reseals the break **(5).** This changes the handedness of the coil to the left-handed form so that the DNA's linking number *L* is decreased by 2.

ases that were specific for sequences greater than a nucleotide. Rather, nucleic acids were cleaved into relatively short fragments by partial digestion with enzymes such as **ribonuclease T1** (from *Aspergillus oryzae*), which cleaves RNA after guanine residues, or **pancreatic ribonuclease A,** which does so after pyrimidine residues. Moreover, there is no reliable polynucleotide reaction analogous to the Edman degradation for proteins (Section 6-1A). Consequently, the polynucleotide fragments were sequenced by their partial digestion with either of two **exonucleases** (enzymes that sequentially cleave nucleotides from the end of a polynucleotide strand): **snake venom phosphodiesterase,** which removes residues from the 3′ end of polynucleotides (Fig. 28-44), or **spleen phosphodiesterase,** which does so from the 5′ end. The resulting oligonucleotide fragments were identified from their chromatographic and electrophoretic mobilities. Sequencing RNA in this manner is a lengthy and painstaking procedure.

The first biologically significant nucleic acid to be sequenced was that of yeast alanine tRNA (Section 30-2A). The sequencing of this 76-nucleotide molecule by Robert Holley, a labor of 7 years, was completed in 1965, some 12 years after Frederick Sanger had determined the amino acid sequence of insulin. This was followed, at an accelerating pace, by the sequencing of >300 species of tRNAs and the 5S ribosomal RNAs (Section 30-3A) from several organisms. The art of RNA sequencing by these techniques reached its zenith in 1976 with the sequencing, by Walter Fiers, of the entire 3569 nucleotide genome of the **bacteriophage MS2.** In comparison, DNA sequencing was in a far more primitive state because of the lack of available DNA endonucleases with any sequence specificity.

Since 1975 there has been dramatic progress in nucleic acid sequencing technology. This has been made possible by three advances:

1. The discovery of **restriction endonucleases,** enzymes that cleave duplex DNA at specific sequences.
2. The development of DNA sequencing techniques.
3. The development of **molecular cloning** techniques (Section 28-8), which permit the acquisition of any identifiable DNA segment in the amounts required for sequencing. Their use is necessary because most specific DNA sequences are normally present in a genome in only a single copy.

These procedures are largely responsible for the present "revolution" in molecular biology that is discussed in succeeding chapters. The use of restriction endonucleases and DNA sequencing techniques are the subject of this section.

The pace of nucleic acid sequencing has become so rapid that directly determining a protein's amino acid sequence is far more difficult than determining the base

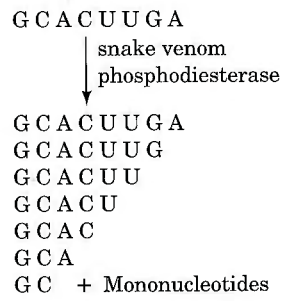


Figure 28-44
The sequence determination of an oligonucleotide by partial digestion with snake venom phosphodiesterase. This enzyme sequentially cleaves the nucleotides from the 3′ end of a polynucleotide that has a free 3′-OH group. Partial digestion of an oligonucleotide with snake venom phosphodiesterase yields a mixture of fragments of all lengths, as indicated, that may be chromatographically separated. Comparison of the base compositions of pairs of fragments that differ in length by one nucleotide establishes the identity of the 3′-terminal nucleotide of the larger fragment. In this way the base sequence of the oligonucleotide may be elucidated.

sequence of its corresponding gene (although amino acid and base sequences provide complementary information; Section 6-1K). There has been such a flood of new DNA sequences—so far ~40 million bases and increasing at the rate of 10 million bases per year—that only computers can keep track of them. A recent high point in the sequencer's art was the determination of the entire 172,282 bp sequence of **Epstein-Barr virus** (human **herpesvirus**) DNA. Indeed, preparations are under way to sequence the 2.9 billion bp human genome (although the magnitude of this project is such that if the DNA sequencing rate can be increased, as it is hoped, to 1 million bp/day, the project will still take nearly 10 years to complete).

A. Restriction Endonucleases

Bacteriophages that propagate efficiently on one bacterial strain, such as *E. coli* K12, have a very low rate of infection (~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. What is the molecular basis of this **host-specific modification** system? Werner Arber showed that it results from a **restriction-modification system** in the bacterial host that consists of a **restriction endonuclease** and a matched **modification methylase.** *The restriction endonuclease recognizes a specific base sequence of four to eight bases in double-stranded DNA and cleaves both strands of the duplex.* The modification methylase methylates a specific base (usually at the amino group of an adenine residue or the 5-position of a cytosine) in the same base sequence recognized by the restriction enzyme. The restriction enzyme does not cleave such a modified 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 modified before the next cycle of replication. *The restriction-modification system is therefore thought to protect 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 modified before being attacked by restriction enzymes. Once a viral genome becomes modified, however, it is able to reproduce in its new host. Its progeny, however, are no longer modified in the way that permits them to propagate in the original host.

There are three known types of restriction endonucleases. **Type I** and **Type III** restriction enzymes each carry both the endonuclease and the methylase 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, whereas Type III enzymes do so 24 to 26 bp distant from the recognition sequence. However, **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 modification methylases. *They cleave DNAs at specific sites within the recognition sequence, a property that makes Type II restriction enzymes indispensible biochemical tools for DNA manipulation.* In the remainder of this section we discuss only Type II restriction enzymes.

Over 500 species of Type II restriction enzymes with >100 differing specificities and from a variety of bacteria have been characterized. Several of the more widely used species are listed in Table 28-5. 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 contains more than one type of restriction enzyme. For example, *Eco*RI is produced by *E. coli* strain RY13.

Table 28-5
Recognition and Cleavage Sites of Some Type II Restriction Enzymes

Enzyme	Recognition Sequence[a]	Microorganism
*Alu*I	AG↓CT	*Arthrobacter luteus*
*Bam*HI	G↓GATC*C	*Bacillus amyloliquefaciens* H
*Bgl*I	GCCNNNN↓ NGCC	*Bacillus globigii*
*Bgl*II	A↓GATCT	*Bacillus globigii*
*Eco*RI	G↓AA*TTC	*Escherichia coli* RY13
*Eco*RII	↓CC*(A/T)GG	*Escherichia coli* R245
*Fnu*DI	GG↓CC	*Fusobacterium nucleatum* D
*Hae*II	PuGCGC↓Py	*Haemophilus aegyptius*
*Hae*III	GG↓C*C	*Haemophilus aegyptius*
*Hind*III	A*↓AGCTT	*Haemophilus influenzae* R_d
*Hpa*II	C↓C*GG	*Haemophilus parainfluenzae*
*Pst*I	CTGCA↓G	*Providencia stuartii* 164
*Sal*I	G↓TCGAC	*Streptomyces albus* G
*Taq*I	T↓CGA*	*Thermus aquaticus*
*Xho*I	C↓TCGAG	*Xanthomonas holcicola*

[a] 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 an asterisk (A* is N^6-methyladenine and C* is 5-methylcytosine). Pu, Py, and N represent purine nucleotide, pyrimidine nucleotide, and any nucleotide, respectively.

Source: Roberts, R. J., *Methods Enzymol.* **68**, 27–41 (1979).

Most Restriction Endonucleases Recognize Palindromic DNA Sequences

Most restriction enzyme recognition sites possess exact twofold rotational symmetry as is diagrammed in Fig. 28-45. Such sequences are known as **palindromes.**

> A palindrome is a word, verse, or sentence that reads the same backwards or forwards. Two examples are "Madam, I'm Adam" and "Sex at noon taxes."

Many restriction enzymes, such as *Eco*RI (Fig. 28-45*a*), catalyze the 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

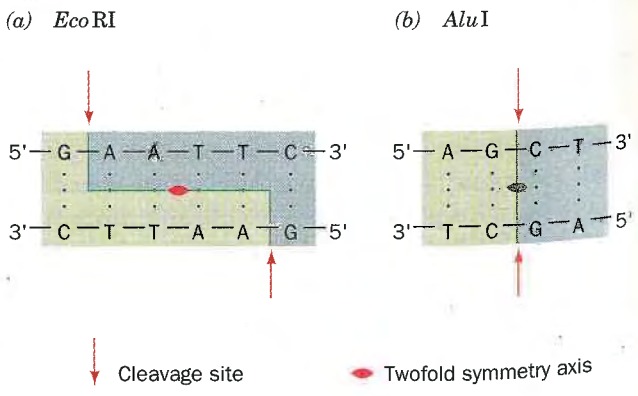


Figure 28-45
The recognition sequences of the restriction endonucleases (*a*) *Eco*RI and (*b*) *Alu*I showing their twofold (palindromic) symmetry and indicating their cleavage sites.

same restriction enzyme. Some restriction cuts, such as that of *Alu*I (Fig. 28-45*b*), pass through the twofold axis of the 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 *Alu*I (4 bp recognition sequence) and *Eco*RI (6 bp recognition sequence) restriction fragments should average $4^4 = 256$ and $4^6 = 4096$ bp in length, respectively.

The X-Ray Structure of the *Eco*RI·DNA Complex Reveals the Molecular Basis of Its Recognition Specificity

The X-ray structure of *Eco*RI endonuclease in complex with a segment of B-DNA containing the enzyme's recognition site was determined by John Rosenberg. The DNA binds in the twofold symmetric cleft between the two identical 276-residue subunits of the dimeric enzyme (Fig. 28-46) thereby accounting for the DNA's palindromic recognition sequence. The protein induces the DNA to kink in three places in a manner that partially unwinds the DNA so as to widen the major groove at the recognition site. Recognition specificity is provided by a tight complementary association of the protein with the major groove of the DNA involving 12 hydrogen bonds between the side chains of Glu 144, Arg 145, and Arg 200 on both protein subunits and the purine bases of the palindromic recognition site.

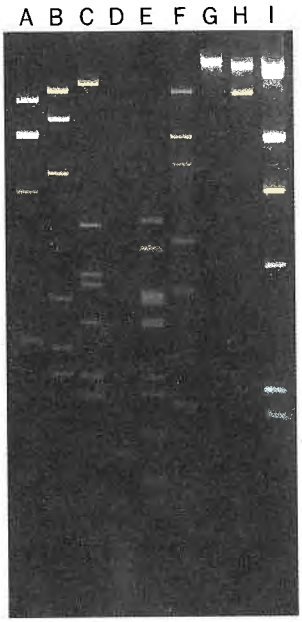


Figure 28-47
Agarose gel electrophoretograms of restriction digests of *Agrobacterium radiobacter* plasmid pAgK84 with (A) *Bam* HI, (B) *Pst* I, (C) *Bgl*II, (D) *Hae*III, (E) *Hin* cII, (F) *Sac*I, (G) *Xba*I, and (H) *Hpa*I. Lane (I) contains λ phage DNA digested with *Hin* dIII as a standard since these fragments have known sizes. [From Slota, J. E. and Farrand, S. F., *Plasmid* **8,** 180 (1982). Copyright © 1982 by Academic Press.]

Restriction Maps Provide a Means of Characterizing a DNA Molecule

The treatment of DNA with a restriction endonuclease produces a series of precisely defined fragments that can be separated according to size by gel electrophoresis (Fig. 28-47). Complementary single strands can be sepa-

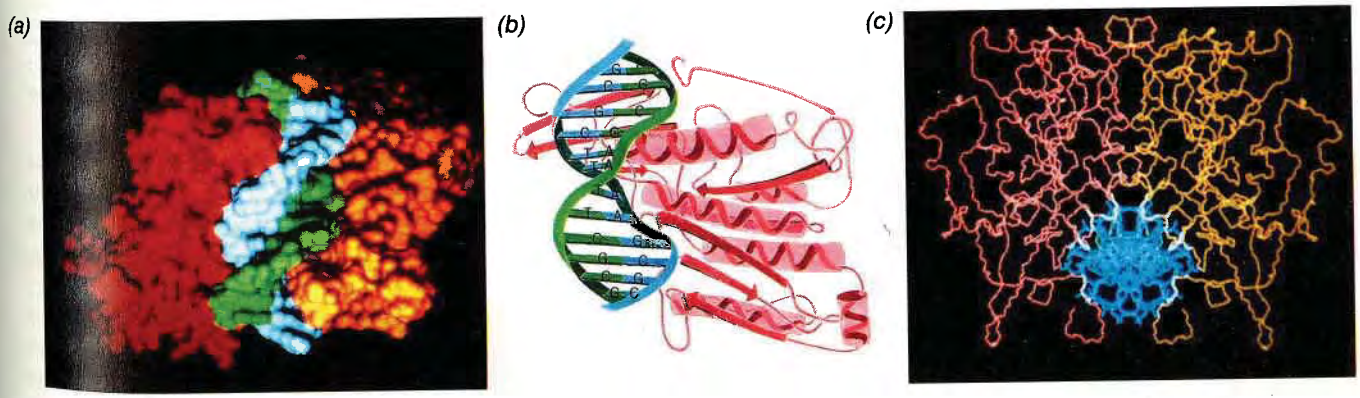


Figure 28-46
The X-ray structure of the *Eco*RI endonuclease·DNA complex. (*a*) Space-filling model showing the duplex DNA bound to the dimeric protein as viewed along the complex's twofold axis of symmetry. The two subunits of the dimeric protein are shown in yellow and orange while the DNA's identical strands are shown in green and blue. (*b*) Ribbon drawing of one *Eco*RI endonuclease subunit interacting with the DNA's major groove. The view is ~90° away from that in Part (*a*). (*c*) A skeletal model of the complex as viewed down the DNA's helical axis showing the protein's polypeptide backbone and the DNA's nonhydrogen atoms. The two protein subunits are drawn in yellow and pink while the DNA is drawn in blue. [Parts (*a*) and (*c*) Courtesy of John M. Rosenberg, University of Pittsburgh. Part (*b*) after Rosenberg, J. M., McClarin, J. A., Frederick, C. A., Wang, B.-C., Grable, J., Boyer, H. W., and Greene, P., *Trends Biochem. Sci.* **12,** 396 (1987).]

rated either by melting the DNA and subjecting it to gel electrophoresis, or by density gradient ultracentrifugation in alkaline CsCl. The single strands can be sequenced by one of the methods described below. If a DNA segment is too long to sequence, it may be further fragmented with a second, *etc.*, restriction enzyme before its strands are separated.

A diagram of a DNA molecule showing the relative positions of the cleavage sites of various restriction enzymes is known as its **restriction map.** Such a map is generated by subjecting the DNA to digestion with two or more restriction enzymes, both individually and in mixtures. By comparing the lengths of the fragments in the various digests, as determined, for instance, by their electrophoretic mobilities, a restriction map can be constructed. For example, consider the 4-kilobase pair **(kb)** linear DNA molecule that *Bam*HI, *Hin*dIII, and their mixture cut to fragments of the lengths indicated in Fig. 28-48*a*. This information is sufficient to deduce the positions of the restriction sites in the intact DNA and hence to construct the restriction map diagrammed in Fig. 28-48*b*. The restriction map of the SV40 chromosome is shown in Fig. 28-49. The restriction sites are physical reference points on a DNA molecule that are easily located. *Restriction maps therefore constitute a convenient framework for locating particular base sequences on*

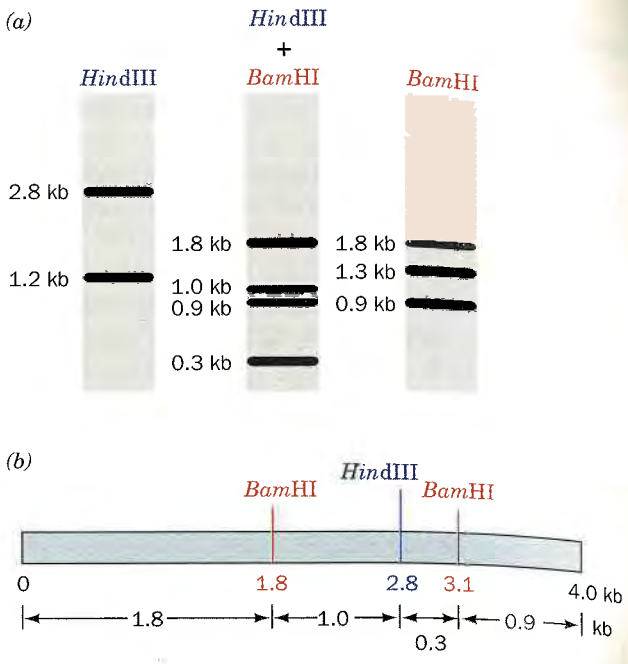


Figure 28-48
(*a*) The gel electrophoretic patterns of digests of a hypothetical DNA molecule with *Hin*dIII, *Bam*HI, and their mixtures. The lengths of the various fragments are indicated. (*b*) The restriction map of the DNA resulting from the information in Part (*a*). This map is equivalent to one that has been reversed, right to left.

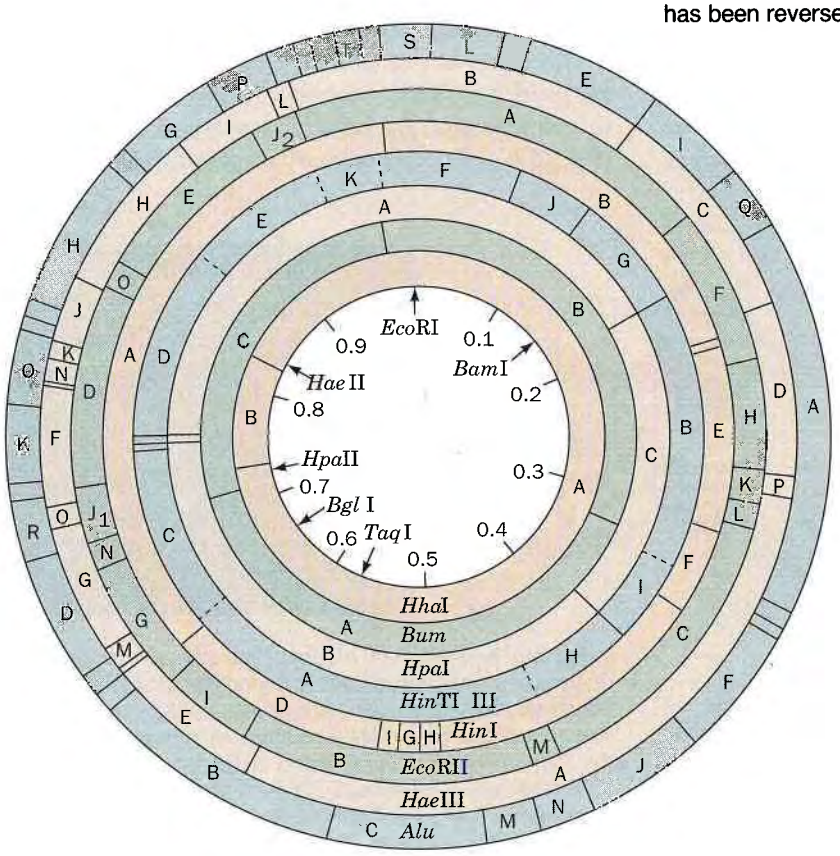


Figure 28-49
A restriction map for the 5243 bp circular DNA of SV40. The central circle indicates the fractional map coordinates with respect to the single *EcoR*I restriction site. The letters A,B,C, . . . in each ring represent the various restriction fragments of the corresponding restriction enzyme in order of decreasing length. [After Nathans, D., *Science* **206,** 905 (1979).]

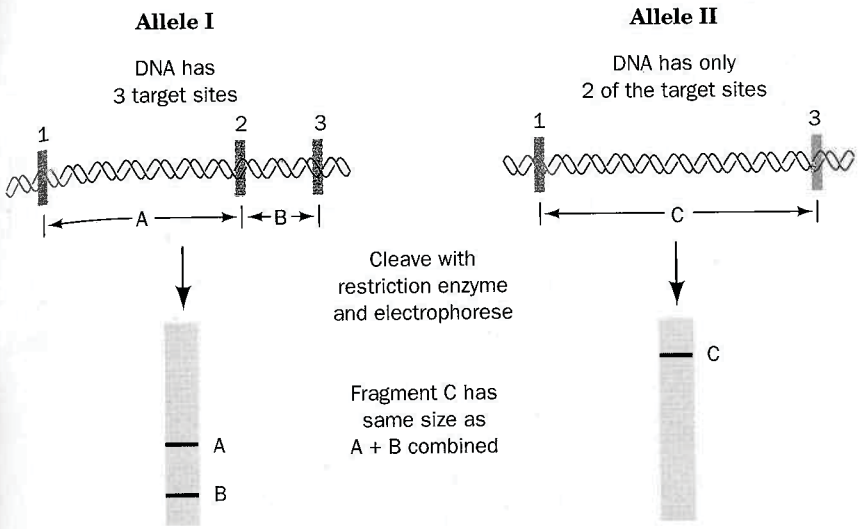


Figure 28-50
A mutational change that affects a restriction site in a DNA segment alters the number and sizes of its restriction fragments.

a chromosome and for estimating the degree of difference between related chromosomes.

Restriction-Fragment Length Polymorphisms Provide Markers for Characterizing Genes

Individuality in humans and other species derives from their high degree of genetic polymorphism; homologous human chromosomes differ in sequence, on average, every 200 to 500 bp. These genetic differences create or eliminate restriction sites. Restriction enzyme digests of the corresponding segments from homologous chromosomes therefore contain fragments with different lengths; that is, these DNAs exhibit **restriction-fragment length polymorphisms (RFLPs;** Fig. 28-50).

RFLPs are useful markers for identifying chromosomal differences (Fig. 28-51). They are particularly valuable for diagnosing inherited diseases for which the molecular defect is unknown. If a particular RFLP is so closely linked to a defective gene that there is little chance the two will recombine from generation to generation (recall that the probability of recombination between two genes increases with their physical separation on a chromosome; Section 27-1C), then the detection of that RFLP in an individual is indicative that the individual has also inherited the defective gene. For example, **Huntington's chorea,** a progressive and invariably fatal neurological deterioration, whose symptoms first appear around age 40, is caused by a dominant but unknown genetic defect. The identification of an RFLP that is closely linked to the defective Huntington's gene has permitted the children of Huntington's chorea victims (50% of whom inherit this devastating condition) to make informed decisions in ordering their lives.

By the same token, the identification of RFLPs associated with the genetic defects causing **cystic fibrosis** (a debilitating and often fatal autosomal recessive disease; heterozygotes, who comprise 5% of the Caucasian population, are asymptomatic), and **Duchenne muscular dystrophy** (an X-linked degenerative disease of muscle that is invariably fatal by around age 25) have permitted

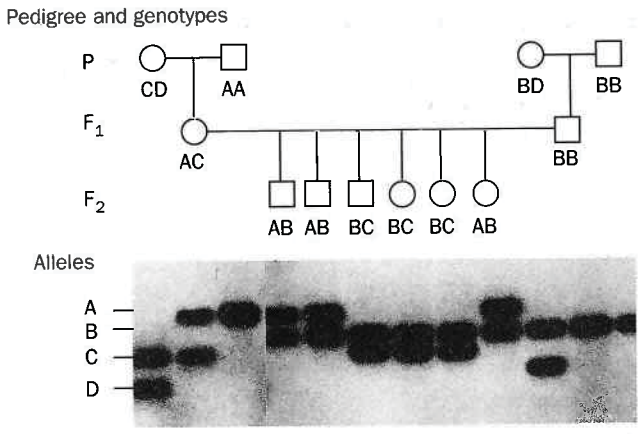


Figure 28-51
RFLPs are inherited according to the rules of Mendelian genetics. Four alleles of a particular gene, each characterized by different restriction markers, can occur in all possible pairwise combinations and segregate independently in each generation (circles represent females and squares represent males). In the P (parental) generation, two individuals are heterozygous (CD and BD) and the other two are homozygous (AA and BB) for the gene in question. Their children, the F_1 generation, are AC and 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. [Courtesy of Ray White, University of Utah Medical School.]

the *in utero* diagnoses of these diseases. (Note that the availability of fetal 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.)

RFLPs are also valuable markers for isolating and thus sequencing their closely linked but unknown genes. Indeed, the first phase in sequencing the human genome, which is already well underway, is to identify a series of ~100 equally spaced markers on each of the 23 human chromosomes.

B. Chemical Cleavage Method

After 1975, several methods were developed for the rapid sequencing of long stretches of DNA. Two of them, the **chemical cleavage** method of Allan Maxam and Walter Gilbert (Fig. 28-52), and the **chain-terminator** procedure of Frederick Sanger (the same individual who pioneered the amino acid sequencing of proteins), are widely used and are largely responsible for the vast number of DNA sequences that have been elucidated. In the remainder of this section, we discuss the chemical

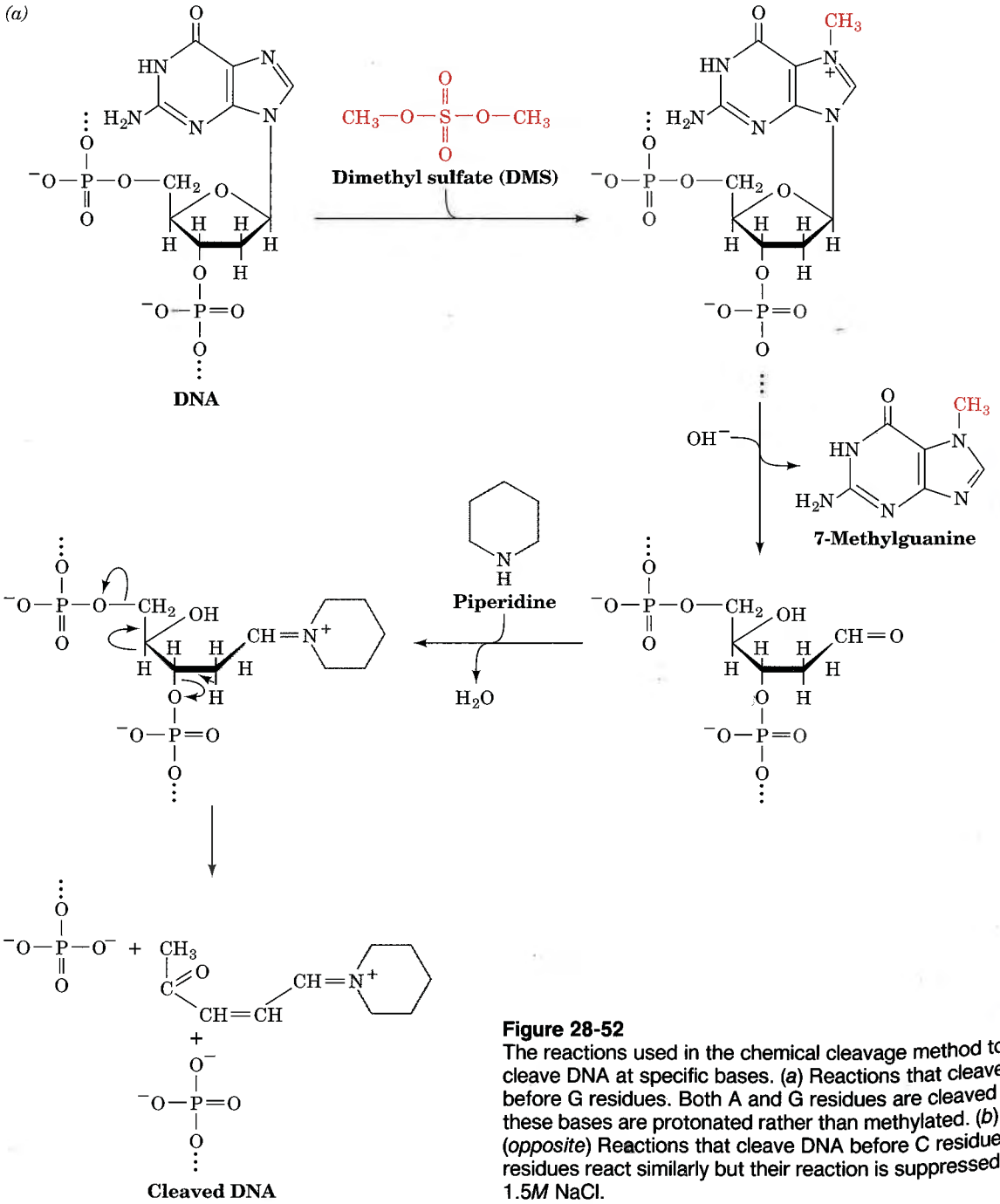
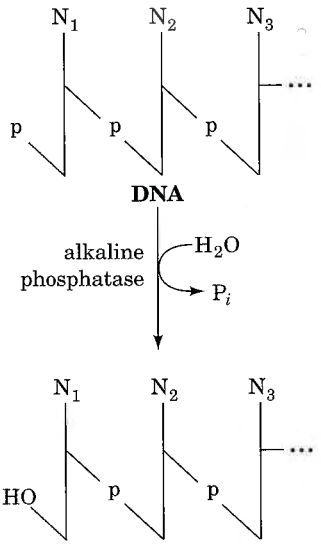


Figure 28-52
The reactions used in the chemical cleavage method to cleave DNA at specific bases. (*a*) Reactions that cleave DNA before G residues. Both A and G residues are cleaved if these bases are protonated rather than methylated. (*b*) (*opposite*) Reactions that cleave DNA before C residues. T residues react similarly but their reaction is suppressed in 1.5*M* NaCl.

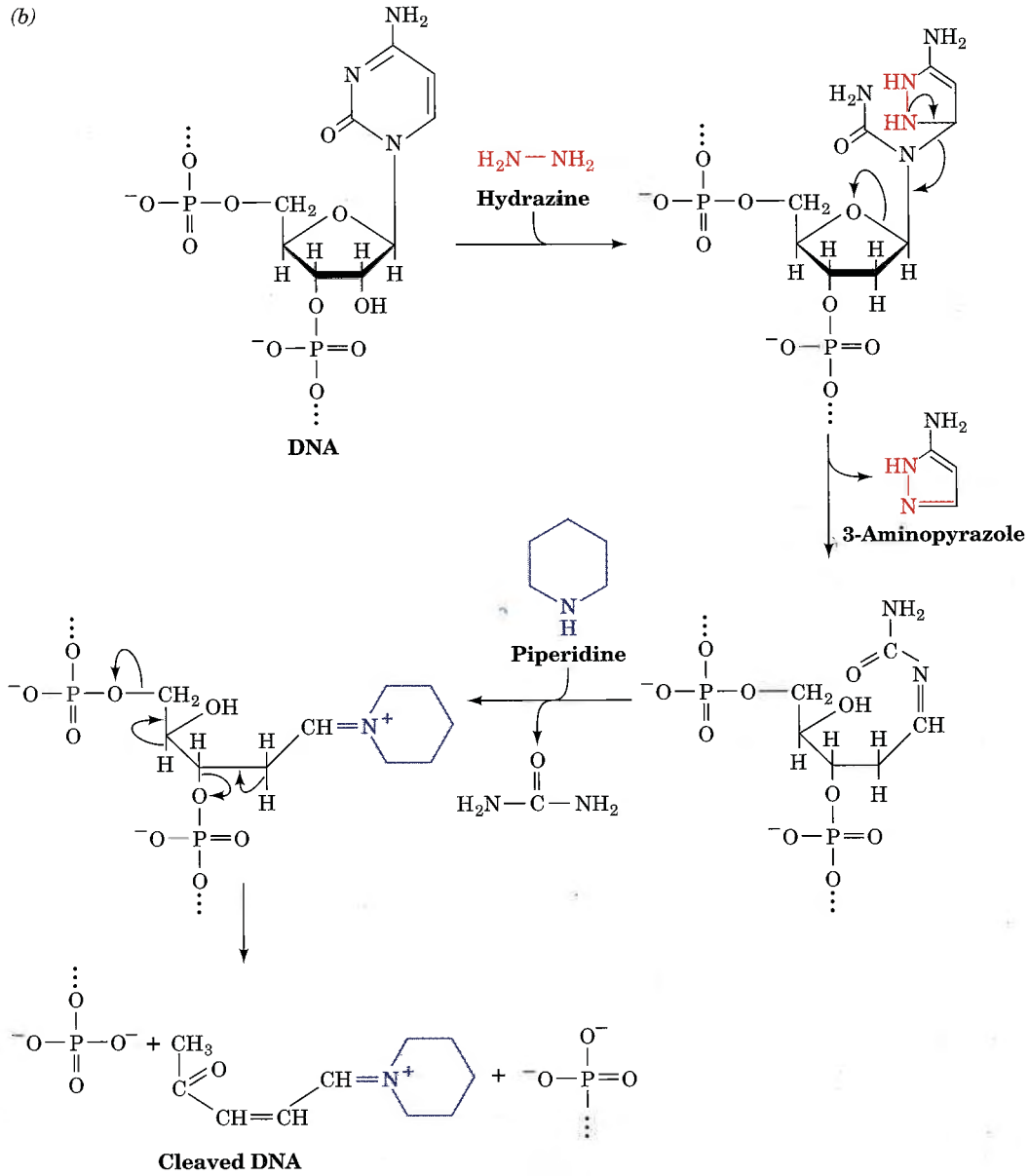
cleavage and chain-terminator methods as well as methods for sequencing RNA.

One End of the DNA Must Be Radioactively Labeled

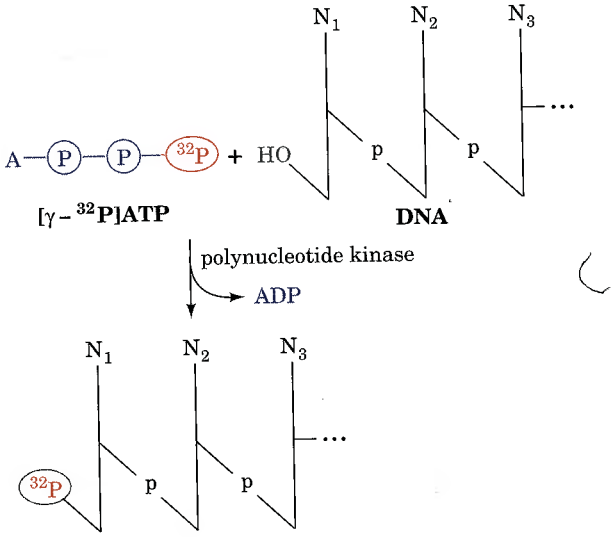
The first step in the chemical cleavage method is to radioactively label one end of the DNA, usually the 5′ end, with ^{32}P. If the DNA already has a 5′ phosphate group, this first must be removed by treatment with **alkaline phosphatase** from *E. coli*.



(b)



Then the 5′ terminus is labeled in a reaction with [γ-^{32}P]ATP as catalyzed by **polynucleotide kinase** from *E. coli* infected with bacteriophage T4.



The DNA Is Cleaved in a Base-Specific Manner

The basic strategy of the chemical cleavage method is to specifically cleave the end-labeled DNA at only one type of nucleotide under conditions such that each molecule is broken at an average of one randomly located susceptible bond. This produces a set of radioactive fragments whose members extend from the ^{32}P-labeled end to one of the positions occupied by the chosen base. For example, if the DNA to be sequenced is

^{32}P-TGTAGGAGCT

cleavage on the 5′ side of the G residues, for instance, would produce the following set of 5′-labeled fragments:

^{32}P-TGTAGGA
^{32}P-TGTAG
^{32}P-TGTA
^{32}P-T

Polyacrylamide gel electrophoresis separates these fragments according to size. Hence *the positions of the G residues in the DNA may be identified from the relative positions on the gel of their corresponding ^{32}P-labeled fragments as revealed by autoradiography.* (The unlabeled cleavage fragments are, of course, not observed in this procedure.) In order for this method to work, the gel must be of sufficient resolving power to unambiguously separate fragments that differ in length by only one nucleotide.

The DNA to be sequenced may be cleaved at specific bases by subjecting it, in separate aliquots, to four different treatments:

1. **G only**

   The DNA is reacted with **dimethyl sulfate (DMS)**, which methylates G residues at N(7), thereby rendering the glycosidic bond of the methylated residue susceptible to hydrolysis (Fig. 28-52*a*). Subsequent treatment by **piperidine** cleaves the polynucleotide chain before the depurinated residue.

2. **A + G**

   DMS preferentially methylates A residues at N(3) rather than N(7) and hence the above treatment cleaves DNA at A residues at only about one fifth the rate it does at G residues. If, instead, the DNA is treated with acid, both A and G are released at comparible rates to yield the same depurinated product indicated in Fig. 28-52*a*. Piperidine treatment then causes strand cleavage before both A and G residues. The A residues are identified by comparing the positions of the G and the A + G cleavages.

3. **C + T**

   The reaction of DNA with **hydrazine** ($NH_2—NH_2$) followed by piperidine treatment cleaves DNA before both its C and T residues (Fig. 28-52*b*).

4. **C only**

   If DNA is treated with the hydrazine in 1.5*M* NaCl, only its C residues react appreciably. Then, as with

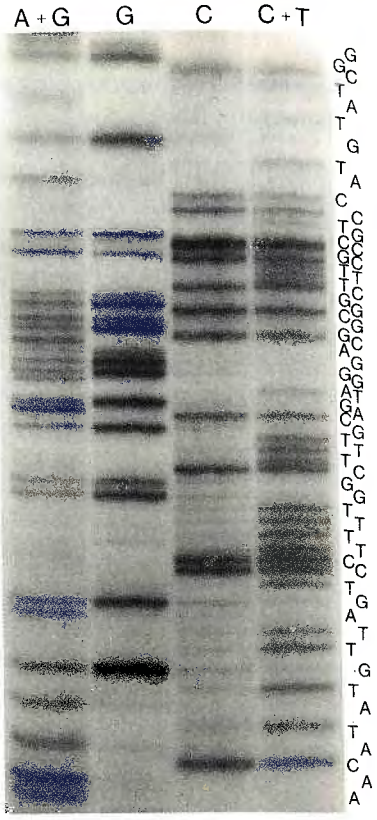


Figure 28-53
An autoradiograph of a sequencing gel containing fragments of a DNA segment that was treated according to the chemical cleavage method of sequence analysis. The DNA was ^{32}P labeled at its 5′ end. The DNA's deduced sequence is written beside the gel. Since the shorter fragments, which have the larger spacing, are at the bottom of the gel, the 5′ → 3′ direction in the sequence corresponds to the upward direction in the gel. [Courtesy of David Dressler, Harvard University Medical School.]

the purines, the comparison of the C and the C + T cleavage positions identifies the T residues.

In all four reactions, the conditions are adjusted so that the strands are cleaved at an average of one randomly located position each.

Cleavage Fragments Are Separated According to Size

The four differently fragmented samples of the DNA, A + G, G, C, and C + T, are simultaneously electrophoresed in parallel lanes on a **sequencing gel.** This is a long, thin (as little as 0.1 mm × up to 200 cm) polyacrylamide slab. It contains ~8*M* urea and is run at ~70°C so as to eliminate all hydrogen bonding associations. *These conditions ensure that the DNA fragments separate only according to their size. The sequence of the DNA can then be directly read off an autoradiogram of the sequencing gel* as is indicated in Fig. 28-53. Indeed, computerized devices are available to aid in doing so. However, a single gel is incapable of resolving much more than 100 consecutive fragments. This limitation is circumvented by electrophoresing three sets of the four differently cleaved samples for successively longer times so as to best resolve the shortest, intermediate length and longest fragments, respectively. In this manner, the base sequence of a 200 to 300 nucleotide DNA fragment can normally be determined from one set of sequencing reactions (although technical advances are steadily increasing this number).

Since the base-specific cleavages destroy the corresponding nucleotide, there is no fragment corresponding to the 5′-terminal nucleotide. Furthermore, the mononucleotide identifying the second base is usually not detected on a gel. The identities of these two nucleotides may be determined by sequencing the complementary strand which, just as importantly, verifies the sequence of the first strand.

C. Chain-Terminator Method

*The chain-terminator method utilizes the E. coli enzyme **DNA polymerase I** to make complementary copies of the single-stranded DNA being sequenced.* Under the direction of the strand being replicated (the **template strand**), DNA polymerase I assembles the four deoxy-

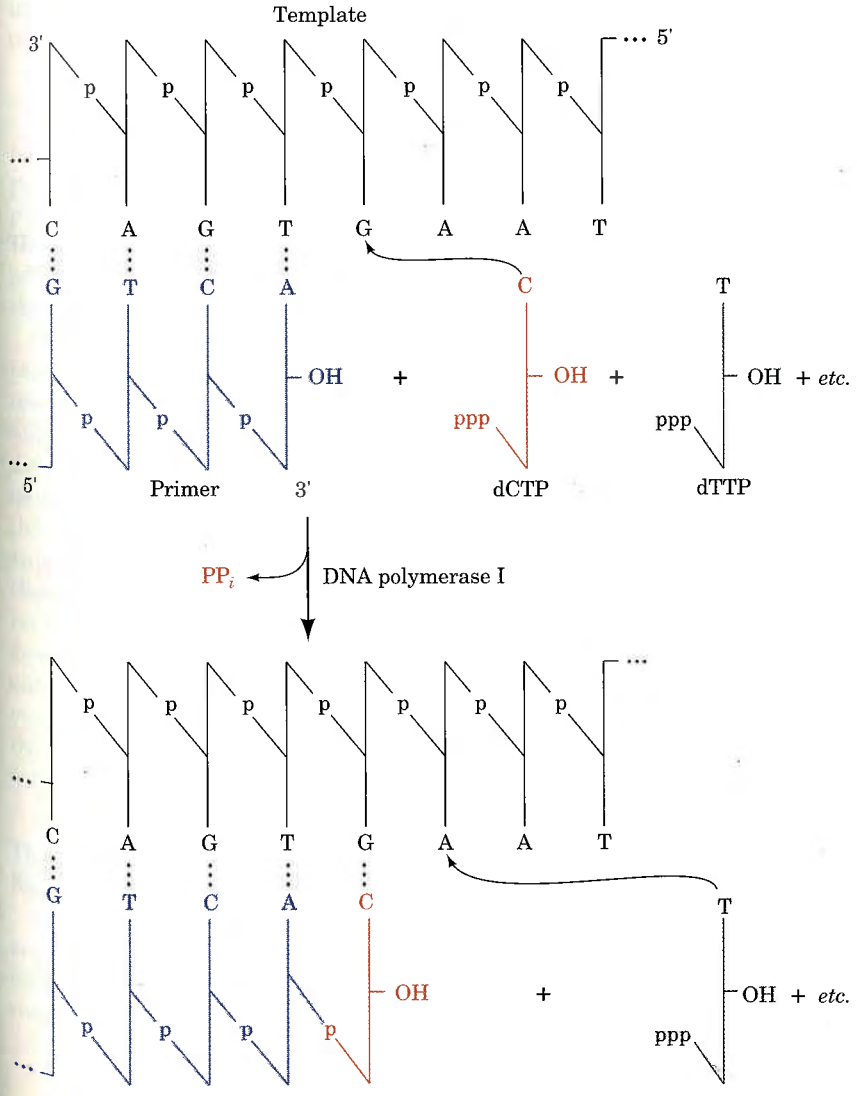


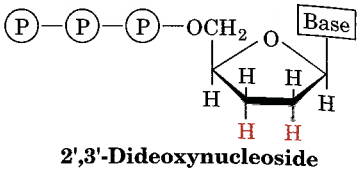
Figure 28-54
The replication of DNA as catalyzed by *E. coli* DNA polymerase I. Under the direction of the template strand, the primer is elongated by the stepwise addition of complementary nucleotides in the 5′ → 3′ direction on the growing polynucleotide.

nucleoside triphosphates, dATP, dCTP, dGTP, and dTTP, into a complementary polynucleotide chain that it elongates in the 5′ → 3′ direction (Fig. 28-54). DNA polymerase I can only sequentially add deoxyribonucleotides to the 3′ end of a polynucleotide. Hence, to initiate replication, it requires the presence of the 5′ end of the chain being synthesized (a **primer**) in a stable base paired complex with the template. If the DNA being sequenced is a restriction fragment, as it usually is, it begins and ends with a restriction site. The primer can therefore be a short DNA segment containing this restriction fragment annealed to the strand being replicated. The role of DNA polymerase I in DNA replication is examined in Section 31-2A.

DNA polymerase I has a 5′ → 3′ exonuclease activity (degrades DNA one nucleotide at a time from its 5′ end), which is catalyzed by a separate active site from that which mediates the polymerization reaction. This is demonstrated by the observation that upon proteolytic cleavage of the enzyme into two fragments, the larger fragment, which is known as the **Klenow fragment,** possesses the full polymerase activity of the enzyme whereas the smaller fragment has the 5′ → 3′ exonuclease activity. Only the Klenow fragment is used in DNA sequencing to ensure that all replicated chains have the same 5′ terminus.

The Synthesis of Labeled DNA by DNA Polymerase Is Terminated After Specific Bases

In the chain-terminator technique (Fig. 28-55), *the DNA to be sequenced is incubated with the Klenow fragment of DNA polymerase I, a suitable primer and the four deoxynucleoside triphosphates, of which at least one (usually dATP) is [α-^{32}P]-labeled. In addition, a small amount of the* ***2′,3′-dideoxynucleoside triphosphate***



2′,3′-Dideoxynucleoside triphosphate

of one of the bases is added to the reaction mixture. When the dideoxy analog is incorporated in the growing polynucleotide in place of the corresponding normal nucleotide, chain growth is terminated because of the absence of a 3′-OH group. By using only a small amount of the dideoxy analog, *a series of truncated chains are generated that are each terminated by the dideoxy analog at one of the positions occupied by the corresponding base. Sequence gel electrophoresis separates these chains according to their lengths and therefore indicates the positions at which that base occurs.*

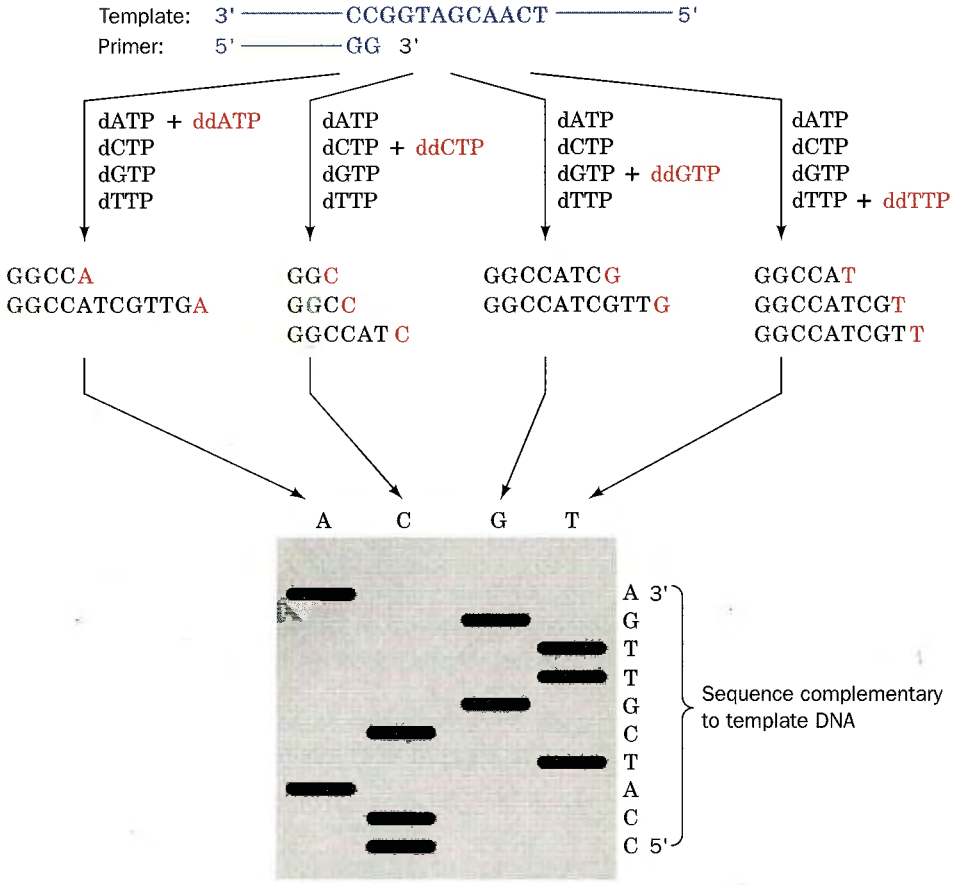


Figure 28-55
A flow diagram of the chain-terminator method of DNA sequencing. The symbol ddATP represents dideoxyadenosine triphosphate, *etc.*

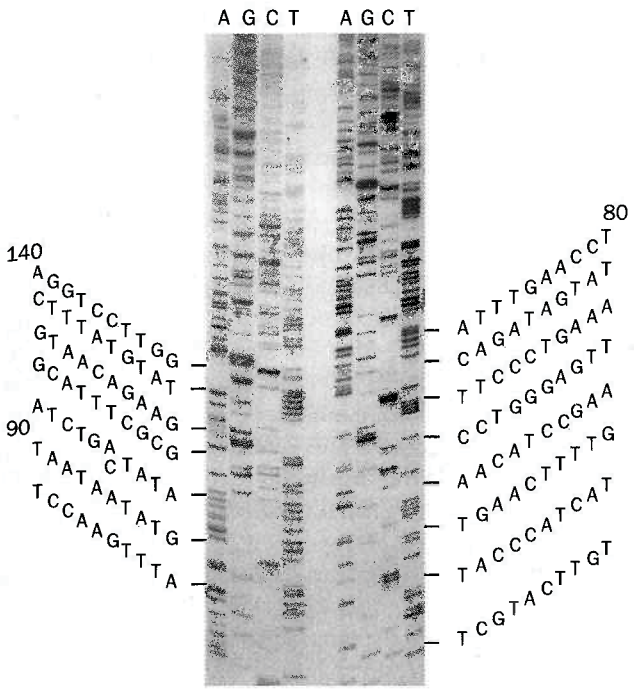


Figure 28-56
An autoradiograph of a sequencing gel containing DNA fragments produced by the chain-terminator method of DNA sequencing. A second loading of the gel (*right*) was made 90 min after the initial loading. The deduced sequence of 140 nucleotides is written along side. [From Hindley, J., DNA Sequencing, *In* Work, T. S. and Burdon, R. H. (Eds.), *Laboratory Techniques in Biochemistry and Molecular Biology*, Vol. 10, *p.* 82, Elsevier (1983). Used by permission.]

Each of the dideoxy analogs of the four bases are reacted in separate vessels and the resulting ^{32}P-labeled product mixtures are subjected to sequence gel electrophoresis in parallel lanes. The sequence of the replicated strand can then be directly read from an autoradiogram of the gel (Fig. 28-56), much like that in the chemical cleavage method.

Both the chain-terminator and the chemical cleavage procedures are widely used for DNA sequencing. With a few hours effort by a skilled operator, either method can sequence a DNA of several hundred nucleotides. Indeed, the major obstacle to sequencing a very long DNA molecule is ensuring that all of its fragments are cloned (by methods discussed in Section 28-8C) rather than sequencing them once they have been obtained. The chemical cleavage method is somewhat easier to set up for occasional use while the chain-terminator method is generally chosen for routine use. Note that the sequence obtained by the chain-terminator method is complementary to the DNA strand being sequenced, whereas the sequence obtained by the chemical cleavage method is that of the original DNA strand.

The Chain-Terminator Method Is Readily Automated

If large DNA segments such as entire chromosomes are to be sequenced, then existing sequencing methods must be greatly accelerated, that is, automated. The chain-terminator method has been adapted to computerized procedures. Rather than use radiolabeled nucleotides (with their inherent health hazards and storage problems), each dideoxynucleoside triphosphate is covalently linked to a differently fluorescing dye. The chain-extension reaction is carried out in a single vessel containing all four fluorescent dideoxy analogs and thus yielding a series of increasingly longer polynucleotides, each with a fluorescence spectrum characteristic of its 3′-terminal nucleotide. The reaction mixture is then subject to sequence gel electrophoresis in a single lane yielding a series of bands, each with the fluorescence spectrum indicative of a successive base in the DNA being sequenced (Fig. 28-57). The gel fluorescence detection system is computer-controlled and hence data acquisition is automated. This device can identify ~10,000 bases per day in contrast to the ~50,000 bases per year that a skilled operator can identify using the above-described manual methods (note that with the use of only one such device, it would still take nearly 1000 years to sequence the human genome).

D. RNA Sequencing

RNA may be rapidly sequenced by only a slight modification of DNA sequencing procedures. The RNA to be sequenced is transcribed into a complementary strand of DNA **(cDNA)** through the action of **RNA-directed DNA polymerase** (also known as **reverse transcriptase).** This enzyme, which is produced by certain RNA-containing viruses (Section 31-4C), uses an RNA tem-

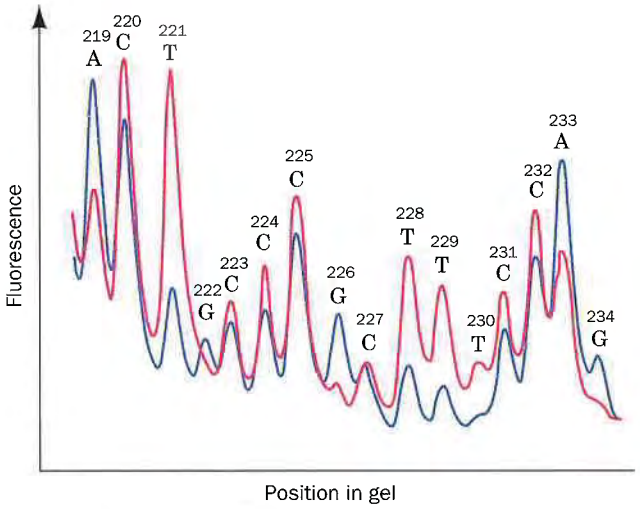


Figure 28-57
The detection, in a sequencing gel, of fluorescent terminator-labeled DNA fragments generated by the automated chain-terminator technique. The ratio of the intensities of the laser-excited dye fluorescence as separately measured in two wavelength bands (*blue*, short wavelength; *red*, long wavelength) unambiguously identifies the fluorescent 3′-terminal base in each gel band as A, T, G, or C. The number above each band indicates its position in the DNA segment being sequenced. [After Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A., and Baumeister, K., *Science* **238,** 340 (1987).]

plate but is otherwise similar in its action to DNA polymerase I. The resulting cDNA may then be sequenced by either the chemical cleavage or the chain-terminator methods. Alternatively, RNA may be directly sequenced by a chemical cleavage method similar to that of DNA sequencing, which employs reactions that cleave RNA after specific bases.

7. CHEMICAL SYNTHESIS OF OLIGONUCLEOTIDES

Molecular cloning techniques (Section 28-8) have permitted the genetic manipulation of organisms in order to investigate their cellular machinery, change their characteristics, and produce scarce or specifically altered proteins in large quantities. *The ability to chemically synthesize DNA oligonucleotides of specified base sequences is an indispensable part of this powerful technology.* For example, suppose we wished to obtain the gene specifying a protein whose amino acid sequence is at least partially known. Reference to the **genetic code** (the correspondence between an amino acid sequence and the base sequence of the gene specifying it; Section 30-1) permits the synthesis of a short (~15 nucleotide) ^{32}P-labeled oligonucleotide that is complementary to a segment of the gene of interest. The oligonucleotide is used as a probe in the Southern transfer procedure (Section 28-4C) on restriction enzyme-digested DNA from the organism that produced the protein. The probe specifically labels the required gene and thereby permits its isolation.

Synthetic oligonucleotides are also required to specifically alter genes through **site-directed mutagenesis.** An oligonucleotide containing a short gene segment with the desired altered base sequence is used as a primer in the DNA polymerase I replication of the gene of interest. Such a primer will hybridize to the corresponding wild-type sequence if there are only a few mismatched base pairs, and its extension, by DNA polymerase I (Section 28-6C), yields the desired altered gene (Fig. 28-58). The altered gene can then be inserted in a suitable organism via techniques discussed in Section 28-8 and grown (cloned) in quantity.

Oligonucleotides Are Valuable Diagnostic Tools

The use of synthetic oligonucleotides as probes in Southern transfer analysis has great promise 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

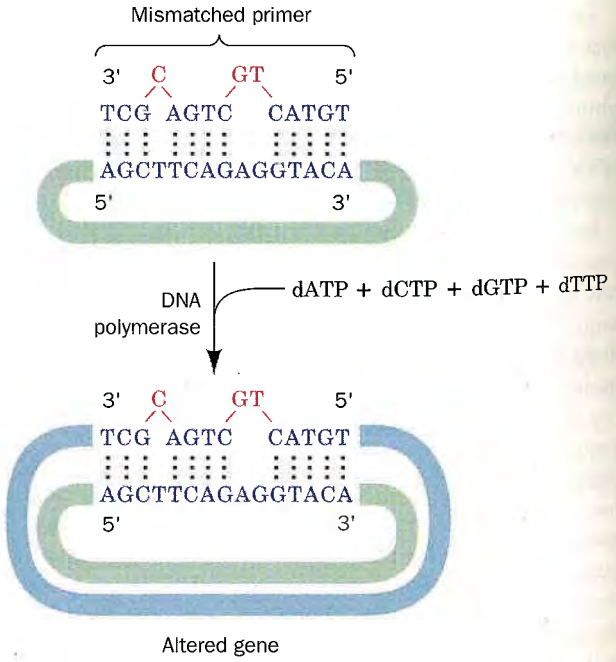


Figure 28-58
Site-directed mutagenesis. A chemically synthesized oligonucleotide incorporating the desired base changes is hybridized to the DNA encoding the gene to be altered. The mismatched primer is then extended by DNA polymerase I thereby generating the mutated gene. The mutated gene can subsequently be inserted into a suitable host organism so as to yield the mutant DNA, or the corresponding RNA, in quantity, produce a specifically altered protein, and/or generate a mutant organism.

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, sickle-cell anemia (Section 9-3B) arises from a single base change that causes the amino acid substitution Glu $\beta 6 \rightarrow$ Val in hemoglobin. 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 normal Hbβ gene gives opposite results. DNA from sickle-cell heterozygotes hybridizes to both probes but in reduced amounts relative to the DNAs from homozygotes. The oligonucleotides may consequently be used in the prenatal diagnosis of sickle-cell disease. DNA probes are also rapidly replacing the much slower and less accurate culturing techniques for the identification of pathogenic bacteria.

Oligonucleotides Are Synthesized in a Stepwise Manner

The basic strategy of oligonucleotide synthesis is analogous to that of polypeptide synthesis (Section 6-4): *A suitably protected nucleotide is coupled to the growing end of the oligonucleotide chain, the protecting group is removed, and the process is repeated until the desired oligonucleotide has been synthesized.* The first practical technique for DNA synthesis, the **phosphodiester method,** which was developed by H. Gobind Khorana in the 1960s, is a laborious process in which all reactions are carried out in solution and the products must be isolated at each stage of the multistep synthesis. Khorana, nevertheless, used this method, in combination with enzymatic techniques, to synthesize a 126-nucleotide tRNA gene, a project that required several years of intense effort by numerous skilled chemists.

The Phosphoramidite Method

By the early 1980s, these difficult and time consuming processes had been replaced by much faster solid phase methodologies that permitted oligonucleotide synthesis to be automated. The presently most widely used chemistry, which was formulated by Robert Letsinger and further developed by Marvin Caruthers, is known as the **phosphoramidite method.** This nonaqueous reaction sequence adds a single nucleotide to a growing oligonucleotide chain as follows (Fig. 28-59):

1. The **dimethoxytrityl (DMTr)** protecting group at the 5′ end of the growing oligonucleotide chain (which is anchored via a linking group at its 3′ end to a solid support, S) is removed by treatment with acid.
2. The newly liberated 5′ end of the oligonucleotide is coupled to the 3′-phosphoramidite derivative of the next deoxynucleoside to be added to the chain. The coupling agent in this reaction is **tetrazole.**
3. Any unreacted 5′ end (the coupling reaction has a yield of over 99%) is capped by acetylation so as to block its extension in subsequent coupling reactions. This prevents the extension of erroneous oligonucleotides.
4. The phosphite triester group resulting from the coupling step is oxidized to the phosphotriester thereby yielding a chain that has been lengthened by one nucleotide.

The above reaction sequence, in commercially available automated synthesizers, can be routinely repeated at least 50 times with a cycle time of 40 min or less. Once an oligonucleotide of desired sequence has been synthesized, it is released from its support and its various blocking groups, including those on the bases, are removed. The product can then be purified by HPLC and/or gel electrophoresis.

8. 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^{-1} 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 as much as one half of any protein originally present in an organism can be recovered in pure form. 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 in only 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 would be an all but impossible task. These difficulties have been largely eliminated in recent years 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 since the mid-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

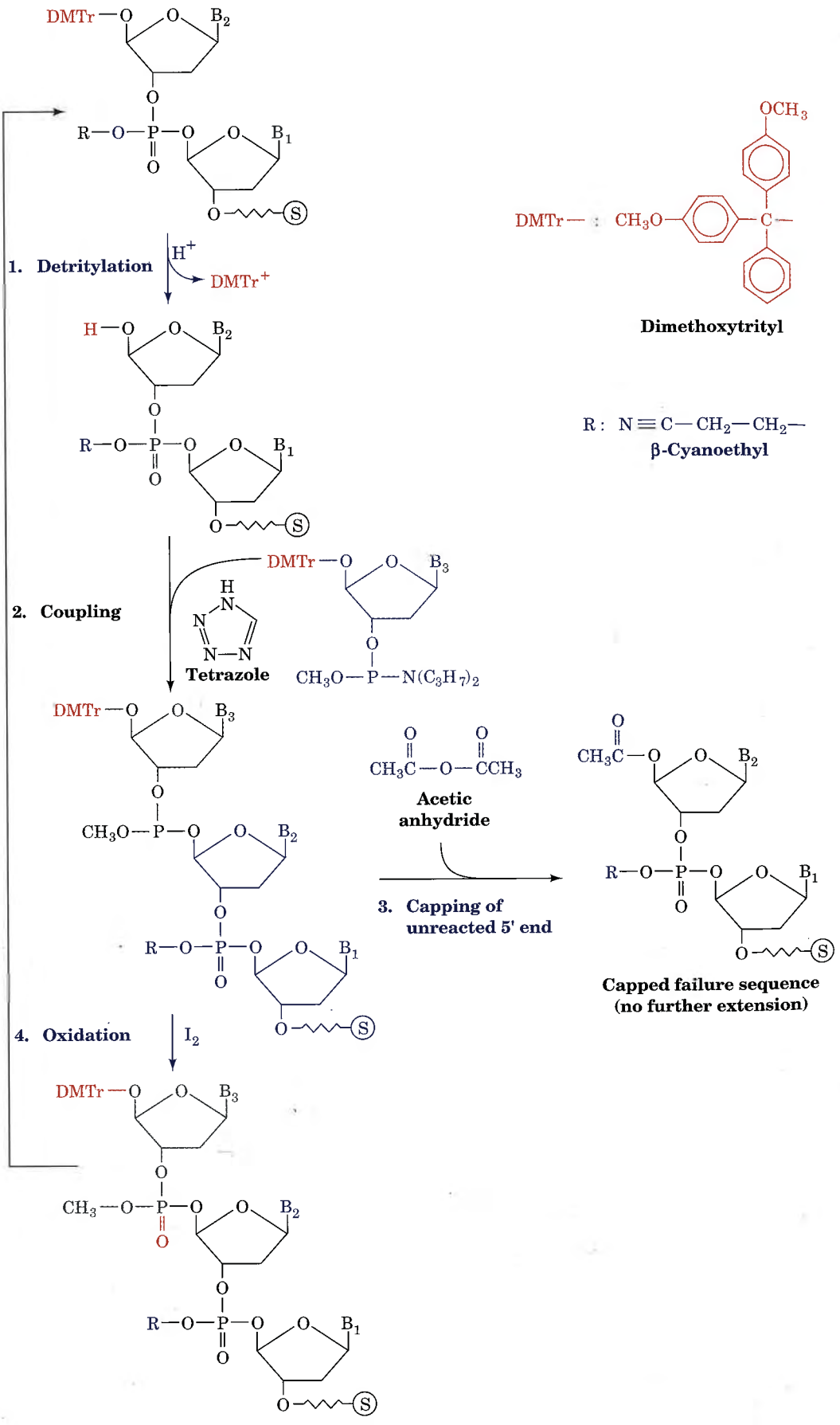

DMTr
B2
B1
B3
R
S
1. Detritylation
H+
DMTr+
2. Coupling
Tetrazole
3. Capping of unreacted 5' end
Acetic anhydride
4. Oxidation
I2
Capped failure sequence (no further extension)
OCH3
CH3O
Dimethoxytrityl
R: N≡C—CH2—CH2—
β-Cyanoethyl

Figure 28-59 (*opposite*)
The reaction cycle in the phosphite-triester method of oligonucleotide synthesis. Here B_1, B_2, and B_3 represent protected bases, and S represents an inert solid phase support such as controlled-pore glass.

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 RNA and protein synthesis (Chapters 29 and 30), the host may also produce large quantities of the RNA and protein specified by that gene. The techniques of genetic engineering are outlined in this section.

A. Cloning Vectors

Both plasmids, bacteriophages, and yeast artificial chromosomes are used as cloning vectors in genetic engineering.

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 31-2), to permit their autonomous propagation in a bacterial host or in yeast. Plasmids may be considered molecular parasites but in many instances they benefit their host by providing functions, such as resistance to an antibiotic, that the host lacks. Indeed, the widespread appearance, since antibiotics came into use, of antibiotic-resistant pathogens is a 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.** The plasmids used in molecular cloning, however, are under **relaxed control;** they are normally present in 10 to 200 copies per cell. Moreover, if protein synthesis in the bacterial host is inhibited, for example, by the antibiotic **chloramphenicol** (Section 30-3G), the copy number of these plasmids may increase to several thousand per cell (about one half of the cell's total DNA). The plasmids that have been constructed (by genetic engineering techniques) for use in molecular cloning are relatively small, carry genes specifying resistance to several antibiotics, and contain a number of conveniently located restriction endonuclease sites into which the DNA to be cloned may be inserted (via techniques described in Section 28-8B). The *E. coli* plasmid designated **pBR322** (Fig. 28-60) is among the most widely used cloning vectors.

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 Ca^{2+}, (which is thought to increase membrane permeability). An absorbed plasmid vector becomes permanently established in its bacterial host (transformation) with an efficiency of ~0.1%.

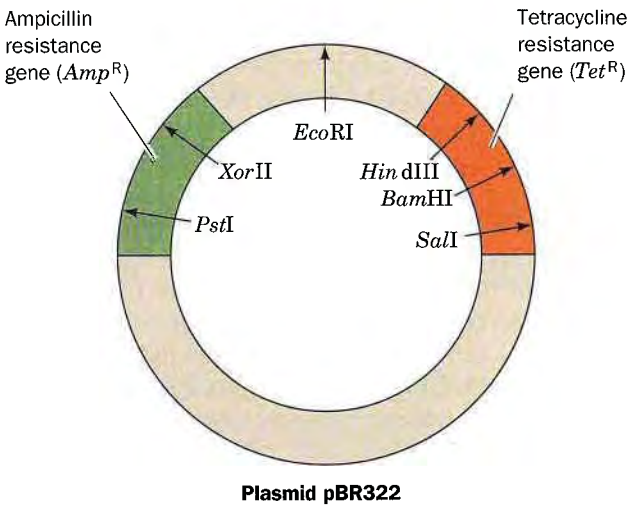


Figure 28-60
A restriction map of plasmid pBR322 indicating the positions of its antibiotic resistance genes.

Plasmid vectors cannot be used to clone DNAs of more than ~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.

Bacteriophage-Based Cloning Vectors

Bacteriophage λ (Fig. 28-61) is an alternative cloning

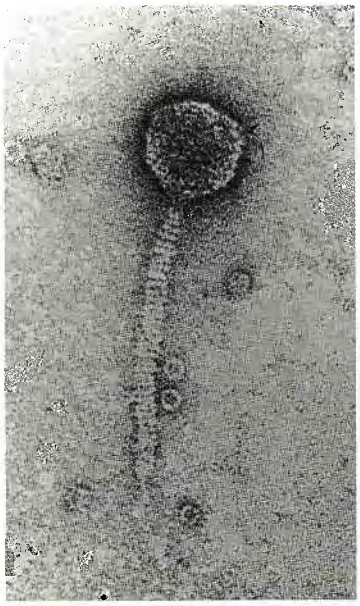


Figure 28-61
An electron micrograph of bacteriophage λ. [Courtesy of A. F. Howatson. From Lewin, B., *Gene Expression*, Vol. 3, Fig. 5.23, John Wiley & Sons Inc. (1977).]

vehicle that can be used to clone DNAs of up to 16 kb. The central third of this virus' 48.5-kb genome is not required for phage infection (Section 32-3A) and can therefore be replaced by foreign DNAs of up to slightly greater size using techniques discussed in Section 28-8B. 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 32-3B). 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.

λ 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 14 bp sequence known as a ***cos* site** located at both ends and that these ends be 36 to 51 kb apart (Section 32-3B). 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 ~49 kb. Cosmids have no phage genes and hence, upon introduction into a host cell via phage infection, reproduce as plasmids.

The **filamentous bacteriophage M13** (Fig. 28-62) is also a useful cloning vector. It has a single-stranded circular DNA that is contained in a protein tube composed of ~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 by the chain-terminator method (Section 28-6C) because these phages directly produce the ~300-nucleotide single-stranded DNA that this technique requires. Furthermore, since the DNA to be sequenced is always inserted at the same point in the viral chromosome (a restriction site; Section 28-8B), an ~15 base synthetic oligonucleotide (the so-called "universal primer") that is complementary to the viral DNA on the 3′ side of the cloning site may be used as the primer for any DNA segment sequenced by this method.

YAC Vectors

DNA segments larger than those that can be carried by cosmids may be cloned in **yeast artificial chromosomes (YACs).** YACs are linear DNA segments that contain all the molecular paraphernalia required for replication in yeast: a replication origin [known as an **autonomously replicating sequence (ARS)**], a centromere (the chromosomal segment attached to the spindle during mitosis and meiosis), and telomeres (the ends of linear chromosomes that permit their replication). DNAs of several hundred kb have been spliced into YACs and successfully cloned.

Figure 28-62
An 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, Emeritus and Harold Fisher, University of Rhode Island.]

B. Gene Splicing

A DNA to be cloned is, in many cases, obtained as a defined fragment through the application of restriction endonucleases (for M13 vectors, the restriction enzymes' requirement of duplex DNA necessitates the use of this phage DNA in double-stranded form). Recall that most restriction endonucleases cleave duplex DNA at specific palindromic sites so as to yield single-stranded ends that are complimentary to each other (Section 28-6A). 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. 28-63). The complimentary (cohesive) ends of the two DNAs specifically associate under annealing conditions and are covalently joined (spliced) through the action of an enzyme named* ***DNA ligase*** (Section 31-2C; DNA ligase produced by **bacteriophage T4** must be used for blunt-ended restriction cuts such as those generated by *Alu*I or *Hae*III; Table 28-5). *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 transferase (terminal transferase).** This mammalian enzyme, which has been implicated in the generation of antibody diversity (Section 34-2C), 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,

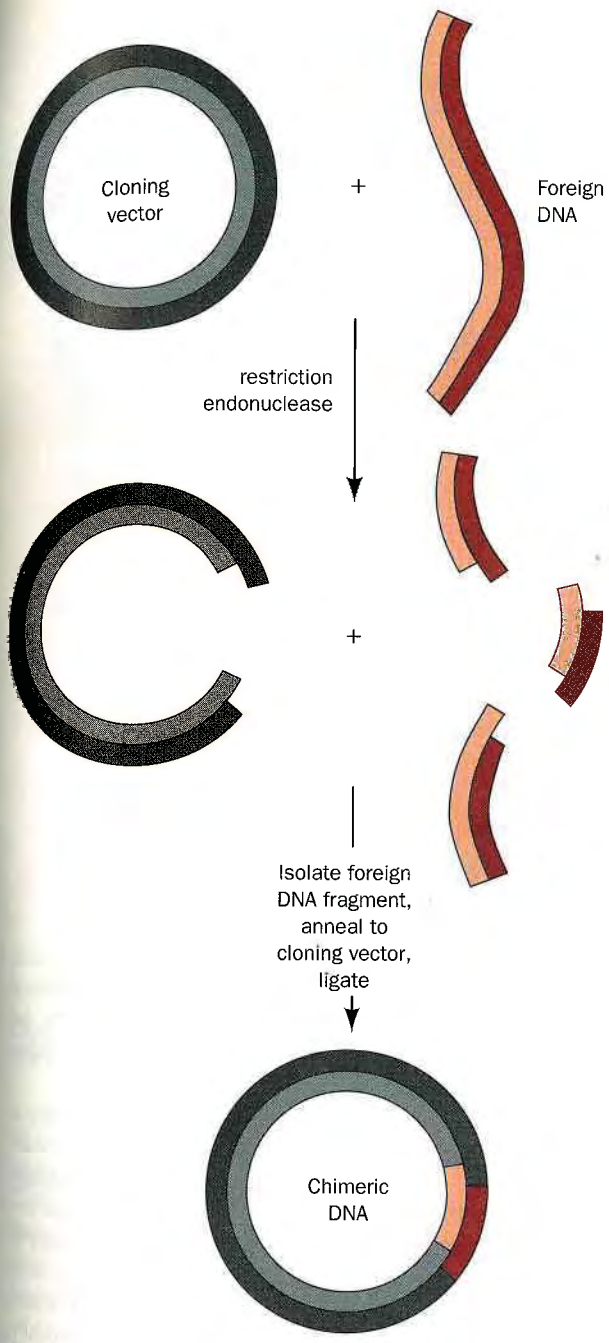


Figure 28-63
The construction of a recombinant DNA molecule by the insertion of a restriction fragment in a cloning vector's corresponding restriction cut.

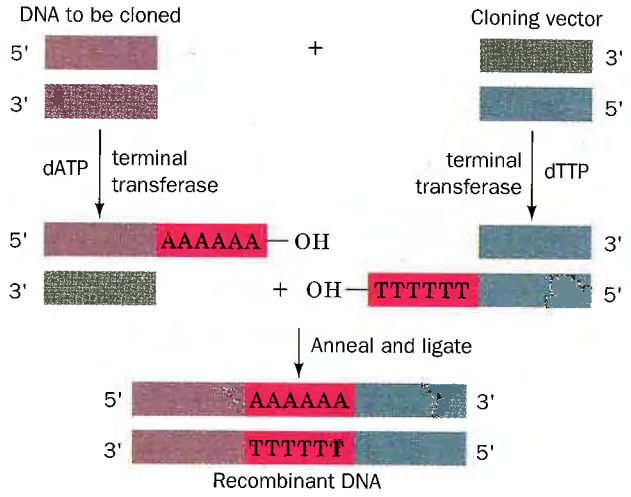


Figure 28-64
Two DNA fragments may be joined through the generation of complementary homopolymer tails. The poly(dA) and poly(dT) tails shown in this example may be replaced by poly(dC) and poly(dG) tails.

for example, can build up poly(dT) tails of ~100 residues on the 3′ ends of the DNA segment to be cloned (Fig. 28-64). 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 complimentary 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 can be circumvented by appending to both ends of the foreign DNA a chemically synthesized palindromic "linker" which has a restriction site matching that of the cloning vector. The linker is attached to the foreign DNA by blunt end ligation with T4 ligase and then cleaved with the appropriate restriction enzyme to yield the correct cohesive ends for ligation to the vector (Fig. 28-65).

Properly Transformed Cells Must Be Selected

How can one select only those host organisms that contain a properly constructed vector? In the case of plasmid transformation, this is usually done through the use of antibiotics. For example, an *E. coli* transformed by a pBR322 plasmid (Fig. 28-60) containing a foreign DNA insert in its *Bam*HI site is **tetracycline-**sensitive (*tet*$^-$; tetracycline is an antibiotic that inhibits bacterial protein synthesis; Section 30-3G) because of the interruption of its *tet* gene by the insert, but is **ampicillin-**resistant (*amp*$^+$; ampicillin is a penicillin derivative) as conferred by the plasmid's intact *amp* gene. Bacterial colonies (clones) can therefore be grown on culture plates containing ampicillin to select for bacteria that have been transformed by this plasmid. Of these colonies, the ones with plasmids containing the foreign DNA can be detected, through replica plating (Section 27-1D), by their failure to grow on a tetracycline-containing medium.

Genetically engineered λ phage variants contain restriction sites that flank the dispensible central third of the phage genome (Section 28-8A). This segment may

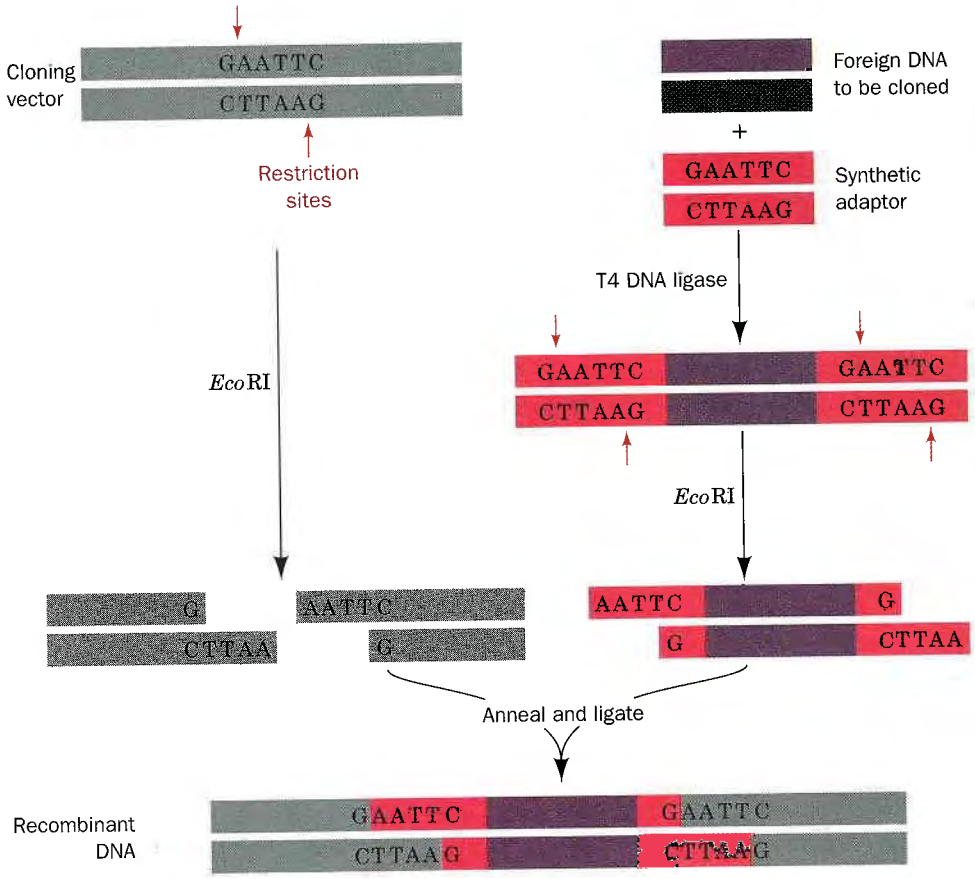


Figure 28-65
The construction of a recombinant DNA molecule through the use of synthetic oligonucleotide adaptors. In this example, the adaptor and the cloning vector have EcoRI restriction sites (*vertical arrows*).

therefore be replaced, as is described above, by a foreign DNA insert (Fig. 28-66). DNA is only packaged in λ phage heads if its length is from 75 to 105% of the 48.5-kb wild-type λ genome. Consequently, λ 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 other types of plasmids cannot.

C. 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 0.000035% of the 2.9 billion bp human genome. A DNA fragment might be identified by Southern blotting of a restriction digest of the genomic DNA under investigation (Section 28-4C). The radioactive probe used in this procedure could be the corresponding mRNA if it is produced in sufficient quantity to be isolated (e.g., reticulocytes, which produce little protein besides hemoglobin, are rich in globin mRNAs). Alternatively, in cases where the amino acid sequence of the protein encoded by the gene is known, the probe could be a mixture of the various synthetic oligonucleotides that might be complimentary to a segment of the gene's inferred base sequence (Section 30-1E).

In practice, 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 sequences(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 28-8B. The DNA is fragmented by partial rather than exhaustive restriction digestion 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 can also be used but requires further treatment of the fragments to insert them into cloning vectors. Genomic libraries have been established for a number of organisms including yeast, *Drosophila,* and humans.

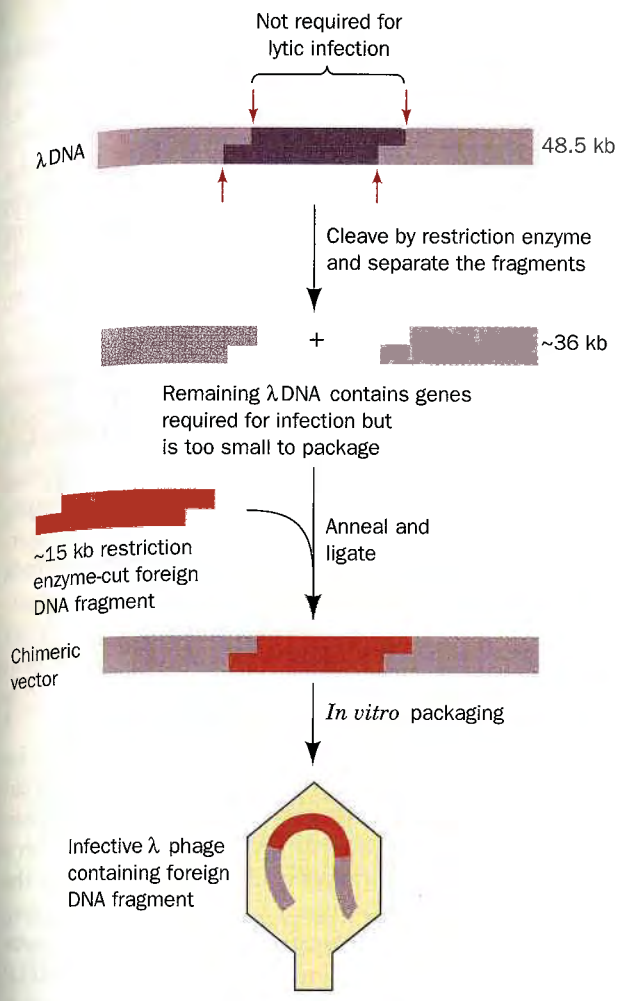


Figure 28-66
The cloning of foreign DNA in λ phages.

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

$$P = 1 - (1 - f)^N \qquad [28.7]$$

Consequently,

$$N = \ln(1 - P)/\ln(1 - f) \qquad [28.8]$$

Thus, in order for $P = 0.99$ for fragments averaging 10 kb in length, $N = 1840$ for the 4000-kb *E. coli* chromosome and 76,000 for the 165,000-kb *Drosophila* genome. The recent development of YAC-based genomic libraries therefore promises to greatly reduce the effort needed 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

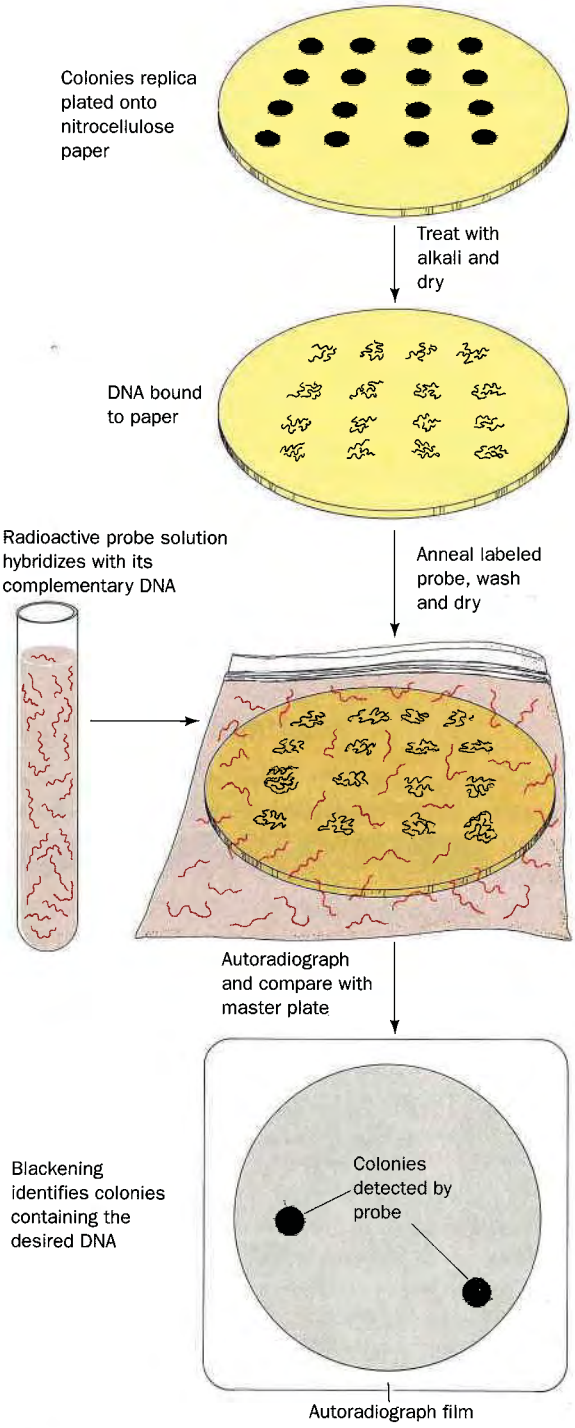


Figure 28-67
Colony (*in situ*) hybridization identifies the clones containing a DNA of interest.

by a process known as **colony** or ***in situ* hybridization** (Fig. 28-67; Latin: *in situ*, in position). The cloned yeast colonies, bacterial colonies, or phage plaques to be tested are transferred, by replica plating, from a master plate, to a nitrocellulose filter. The filter is treated with

NaOH, which lyses the cells/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/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. Using this technique, even an ~1 million clone human genomic library can be readily screened for the presence of a particular DNA segment.

Many eukaryotic genes and gene clusters span enormous tracts of DNA (Section 33-2); 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. 28-68): Each gene fragment 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, however, greatly reduces the need for this laborious and error-prone process.

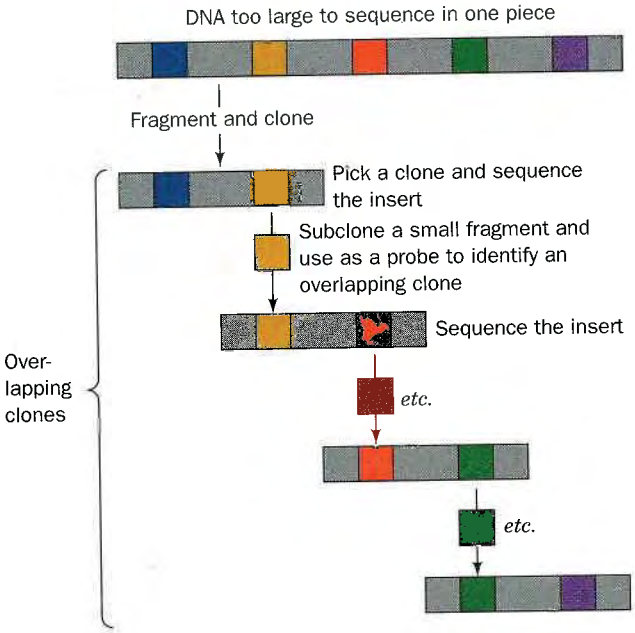


Figure 28-68
Chromosome walking. A DNA segment too large to sequence in one piece is fragmented and cloned. A clone is picked and the DNA insert it contains is sequenced. A small fragment of the insert near one end is subcloned (cloned from a clone) and used as a probe to select a clone containing an 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.

D. DNA Amplification by the Polymerase Chain Reaction

Although molecular cloning techniques are indispensible to modern biochemical research, the use of the **polymerase chain reaction (PCR)** offers a more convenient method of amplifying a specific DNA segment of up to 6 kb. In this technique, a denatured DNA sample is incubated with DNA polymerase and two oligonucleotide primers that direct the DNA polymerase to synthesize new complimentary strands. Multiple cycles of this process, each approximately doubling the amount of DNA present, exponentially 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, the primers are annealed to their complimentary segments on the DNA, and the DNA polymerase directs the synthesis of the complimentary strands (Section 28-6C). The use of a heat-stable DNA polymerase from the thermophilic bacterium *Thermus aquaticus* eliminates the need to add fresh enzyme after each heat denaturation step. Hence, each amplification cycle is controlled by simply varying the temperature.

Twenty-five cycles of PCR amplification increase the amount of the target sequence by around a millionfold with high specificity. 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. The amplified DNA can be characterized by the various techniques we have discussed: Southern blotting, RFLP analysis, and direct sequencing. PCR amplification is therefore a form of "cell-free molecular cloning" that can accomplish in an automated 3 to 4 h *in vitro* reaction what would otherwise take days or weeks via the cloning techniques discussed above.

The use of PCR amplification holds great promise for a variety of applications. Clinically, it can be used for the rapid diagnosis of infectious diseases and the detection of rare pathological events such as chromosomal translocations. Forensically, the DNA from a single hair or sperm can be used to unambiguously identify the donor. RNA may also be amplified by the PCR method by first converting it to cDNA through the use of reverse transcriptase (Section 28-6D).

E. Production of Proteins

One of the greatest potential uses of recombinant DNA technology is in the production of large quantities of scarce and/or novel proteins. This is a relatively straightforward procedure for bacterial proteins: A cloned structural gene must be accompanied by the properly positioned transcriptional and translational control sequences for its expression. With the use of a relaxed control plasmid and an efficient **promoter** (a type of transcriptional control element; Section 29-3A), the production of a protein

of interest may reach 30% of the host'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 inclusions. Protein extracted from these inclusions must therefore be renatured, usually by dissolving it in a guanidinium chloride or urea solution (Section 8-1A) and then dialyzing away the denaturant.

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. Most eukaryotic genes contain one or more internal unexpressed sequences called **introns,** which are specifically excised from the gene's RNA transcript to form the mature mRNA (Section 29-4A). Bacterial genes lack introns and hence, bacteria are unable to excise them.
3. Bacteria lack the enzyme systems to carry out the specific post-translational processing that many eukaryotic proteins require for biological activity (Section 30-5). 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 30-6A).

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 mRNA. Alternatively, genes encoding small proteins of known sequence can be chemically synthesized (Section 28-7). Neither of these strategies is universally applicable, however, because few mRNAs are sufficiently abundant to be isolated and many eukaryotic proteins are large (although the maximum available size of synthetic polynucleotides is increasing rapidly). Likewise, no general approach has been developed for the post-translational modification of eukaryotic proteins although polypeptide cleavage by treatment with trypsin or cyanogen bromide (Section 6-1E) has been successfully employed in the *in vitro* activation of some eukaryotic proenzymes. Lastly, the preferential bacterial proteolysis of certain eukaryotic proteins has been prevented by inserting the eukaryotic gene within a bacterial gene. The resulting hybrid protein has an N-terminal polypeptide of bacterial origin that, in some cases, prevents bacterial proteases from recognizing the eukaryotic segment as being foreign. However, 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 these problems (although post-translational processing may vary among different eukaryotes). Indeed, **shuttle vectors** are available that can propagate in both yeast and *E. coli* and thus transfer (shuttle) genes between these two types of cells.

The ability to synthesize a given protein in large quantities has enormous medical, agricultural, and industrial potential. Human insulin and human growth hormone, to mention but two, are already in widespread clinical use, and many others are under development. *Of equal importance is the ability to tailor proteins to specific applications through site-directed mutagenesis* (Section 28-7). For many purposes, however, it will be preferable to tailor an intact organism rather than just its proteins — true genetic engineering. For example, if nitrogen fixing bacteria can be persuaded to associate with agriculturally important plants besides legumes (a complicated process whose requirements are by no means understood), the need for nitrogenous fertilizers to grow these plants in high yield will perhaps be entirely eliminated.

F. Social 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 30-3G) 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 regulations 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, χ1776, the first approved "safe"

Figure 28-69
[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.]

strain of *E. coli,* has among its several defects the requirement for diaminopimelic acid, an intermediate in lysine biosynthesis (Section 24-5B), 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 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. 28-69). Bacterially produced human insulin is now routinely prescribed to treat diabetes and few would dispute the use of "gene therapy," if it can be developed, to cure such genetic defects as sickle-cell anemia (Section 9-3B) and Lesch-Nyhan syndrome (Section 26-2D). If, however, it becomes possible to alter complex 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? If 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? The U.S. Supreme Court has affirmed that novel life forms developed in the laboratory may be patented. But to what extent will such proprietory rights impede the free exchange of ideas and information that has heretofore permitted the rapid development of recombinant DNA technology?

Chapter Summary

Nucleic acids are linear polymers of nucleotides containing either ribose residues in RNA or deoxyribose residues in DNA that are linked by 3′ → 5′ phosphodiester bonds. In double helical DNAs and RNAs, the base compositions obey Chargaff's rules: A = T and G = C. RNA, but not DNA, is susceptible to base-catalyzed hydrolysis.

B-DNA consists of a right-handed double helix of antiparallel sugar–phosphate chains with ~10 bp per turn of 34 Å and with the bases all perpendicular to the helix axis. Bases on opposite strands hydrogen bond in a geometrically complementary manner to form A·T and G·C Watson–Crick base pairs. DNA replicates in a semiconservative manner as has been demonstrated by the Meselson–Stahl experiment. At low humidity, B-DNA undergoes a reversible transformation to a wider, flatter right-handed double helix known as A-DNA. Z-DNA, which is formed at high salt concentrations by polynucleotides of alternating purine and pyrimidine base sequences, is a left-handed double helix. Double-helical RNA and RNA·DNA hybrids have A-DNA-like structures. DNA occurs in nature as molecules of enormous lengths which, because they are also quite stiff, are easily mechanically cleaved by laboratory manipulations.

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. The orientations about the glycosidic bond and the various torsion angles in the sugar–phosphate chain are sterically constrained in nucleic acids. Likewise, only a few of the possible sugar pucker conformations are commonly observed. Watson–Crick base pairing is both geometrically and electronically complementary. Yet, hydrogen bonding interactions do not significantly stabilize nucleic acid structures. Rather, they are largely stabilized by hydrophobic interactions. Nevertheless, the hydrophobic forces in nucleic acids are qualitatively different in character from those that stabilize proteins. Electrostatic interactions between charged phosphate groups are also important structural determinants of nucleic acids.

Nucleic acids are fractionated by many of the techniques that are used to separate proteins. Hydroxyapatite chromatography separates single-stranded from double-stranded DNA. Polyacrylamide or agarose gel electrophoresis separates DNA largely on the basis of size. Very large DNAs can be separated by pulsed-field gel electrophoresis on agarose gels. Specific base sequences may be detected in DNA with the Southern transfer technique and in RNA by the similar northern transfer technique. DNA may be fractionated according to base composition by CsCl density gradient ultracentrifugation. Different species of RNA are separated by rate-zonal ultracentrifugation through a sucrose gradient.

The linking number of a covalently closed circular DNA is topologically invariant. Consequently, any change in the twist of a circular duplex must be balanced by an equal and opposite change in its writhing number, which indicates its degree of supercoiling. Supercoiling can be induced by intercalation agents. The gel electrophoretic mobility of DNA increases with its degree of superhelicity. Naturally occurring DNAs are all negatively supercoiled and must be so in order to participate in DNA replication, RNA transcription, and genetic recombination. Type I topoisomerases (nicking–closing enzymes) relax negatively supercoiled DNAs, one supertwist at a time, by creating a single-strand break, passing a single-strand loop through the gap, and resealing it. Type II topoisomerases (gyrases) generate negative supertwists at the expense of ATP hydrolysis. They do so, two supertwists at a time, by making a double-strand scisson in the DNA, passing the duplex through the break, and resealing it.

Nucleic acids may be sequenced by the same basic strategy used to sequence proteins. Defined DNA fragments are generated by Type II restriction endonucleases, which cleave DNA at specific and usually palindromic sequences of four to six bases. Restriction maps provide easily located physical reference points on a DNA molecule. In the chemical cleavage method of DNA sequencing, a defined fragment of DNA is ^{32}P-labeled at one end and subjected to a chemical cleavage process that randomly cleaves it after a particular type of base. The electrophoresis of the four differently cleaved DNA samples in parallel lanes of a sequencing gel resolves fragments that differ in size by one nucleotide. The base sequence of the DNA can be directly read from an autoradiogram of the gel. In the chain-terminator method, the DNA to be sequenced is replicated by DNA polymerase I in the presence of a [α-^{32}P]-labeled deoxynucleoside triphosphate and a small amount of the dideoxy analog of one of the nucleoside triphosphates. This results in a series of ^{32}P-labeled chains that are terminated after the various positions occupied by the corresponding base. An autoradiograph of the sequencing gel containing the four sets of fragments, each terminated after a different type of base, indicates the DNA's base sequence. RNA may be sequenced by determining the sequence of its corresponding cDNA or by directly sequencing it by a variation of the chemical cleavage method.

Oligonucleotides are indispensible to recombinant DNA technology; they are used to identify normal and mutated genes and to specifically alter genes through site-directed mutagenesis. Oligonucleotides of defined sequence are efficiently synthesized by the phosphite-triester method, a cyclic, nonaqueous, solid phase process that has been automated.

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, bacteriophages, cosmids, or yeast artificial chromosomes (YACs). 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. Introduction of a recombinant cloning vector into a suitable host organism permits the foreign DNA segment to be produced in nearly unlimited quantities. A particular gene may be isolated through the screening of a genomic library of the organism producing the gene. Genetic engineering techniques may also be used to produce otherwise scarce or specifically altered proteins in large quantities.

References

General

Bloomfield, V. A., Crothers, D. M., and Tinoco, I. Jr., *Physical Chemistry of Nucleic Acids,* Harper & Row (1974).

Cantor, C. R. and Schimmel, P. R., *Biophysical Chemistry,* Chapters 3, 5, 22–24, Freeman (1980).

Landegren, R., Kaiser, R., Caskey, C. T., and Hood, L., DNA diagnostics—molecular techniques and automation, *Science* **242,** 230–237 (1988).

Saenger, W., *Principles of Nucleic Acid Structure,* Springer–Verlag (1984). [A detailed and authoritative exposition.]

Structures of DNA, Cold Spring Harbor Symp. Quant. Biol. **47** (1983). [A cornucopia of structural information on DNA.]

Wood, W. B., Wilson, J. H., Benbow, R. M., and Hood, L. E., *Biochemistry, A Problems Approach* (2nd ed.), Chapters 17 and 19, Benjamin/Cummings (1980).

Structures and Stabilities of Nucleic Acids

Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V., and Kopka, M. L., The anatomy of A-, B- and Z-DNA, *Science* **216,** 475–485 (1982).

Dickerson, R. E., The DNA helix and how it is read, *Sci. Am.* **249**(6): 94–111 (1983).

Drew, H. R., McCall, M. J., and Calladine, C. R., Recent studies of DNA in the crystal, *Annu. Rev. Cell Biol.* **4,** 1–20 (1988).

Jaworski, A., Hsieh, W.-T., Blaho, J. A., Larson, J. E., and Wells, R. D., Left-handed DNA in vivo, *Science* **238,** 773–777 (1988).

Jurnak, F. A. and McPherson, A. (Eds.), *Biological Macromolecules and Assemblies,* Vol. 2: *Nucleic Acids and Interactive Proteins,* Wiley (1985). [Chapters 1–3 are authoritative reviews on the structures of A-, B-, and Z-DNAs, respectively.]

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.]

Rich, A., Nordheim, A., and Wang, A. H.-J., The chemistry and biology of left-handed Z-DNA, *Annu. Rev. Biochem.* **53,** 791–846 (1984).

Shakked, Z. and Rabinovich, D., The effect of the base sequence on the fine structure of the DNA double helix, *Prog. Biophy. Mol. Biol.* **47,** 159–195 (1986).

Sundaralingam, M., Stereochemistry of nucleic acids and their constituents. IV. Allowed and preferred conformations of nucleosides, nucleoside mono-, di-, tri-, and tetraphosphates, nucleic acids and polynucleotides, *Biopolymers* **7,** 821–860 (1969).

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. 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.]

Wells, R. D., Unusual DNA structures, *J. Biol. Chem.* **263,** 1095–1098 (1988).

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).

Zimmerman, S. B., The three-dimensional structures of DNA, *Annu. Rev. Biochem.* **51,** 395–427 (1982).

Fractionation of Nucleic Acids

Anand, R., Pulsed field gel electrophoresis: a technique for fractionating large DNA molecules, *Trends Genet.* **2,** 278–283 (1986).

Cantor, C. R., Smith, C. L., and Mathew, M. K., Pulsed-field gel electrophoresis of very large molecules, *Annu. Rev. Biophys. Biophys. Chem.* **17,** 287–304 (1988).

Freifelder, D., *Physical Biochemistry. Applications to Biochemistry and Molecular Biology* (2nd ed.), Freeman (1982).

Gould, H. and Matthews, H. R., Separation methods for nucleic acids and oligonucleotides, *in* Work, T. S. and Work E. (Eds.), *Laboratory Techniques in Biochemistry and Molecular Biology,* Vol. 4, Part III, North–Holland (1976).

Rickwood, D. and Hames, B. D. (Eds.), *Gel Electrophoresis of Nucleic Acids. A Practical Approach,* IRL Press (1982).

Schleif, R. F. and Wensink, P. C., *Practical Methods in Molecular Biology,* Chapter 5, Springer–Verlag (1981).

Walker, J. M. (Ed.), *Methods in Molecular Biology,* Vol. 2, *Nucleic Acids,* Humana Press (1984).

Supercoiled DNA

Bauer, W. R., Crick, F. H. C., and White, J. H., Supercoiled DNA, *Sci. Am.* **243**(1): 118–133 (1980). [The topology of supercoiling.]

Cozzarelli, N. R., DNA gyrase and the supercoiling of DNA, *Science* **207,** 953–960 (1980).

Gellert, M., DNA topoisomerases, *Annu. Rev. Biochem.* **50,** 879–910 (1981).

Maxwell, A. and Gellert, M., Mechanistic aspects of DNA topoisomerases, *Adv. Protein Chem.* **38,** 69–107 (1986).

Wang, J. C., DNA topoisomerases, *Sci. Am.* **247**(1): 94–109 (1982).

Wang, J. C., DNA topoisomerases, *Annu. Rev. Biochem.* **54,** 665–697 (1985).

Nucleic Acid Sequencing

Brownlee, G. G., Determination of sequences in RNA, *in* Work, T. S. and Work, E. (Eds.), *Laboratory Techniques in Biochemistry and Molecular Biology,* Vol. 3, Part I, North–Holland (1972). [A review of the sequencing procedures used before the advent of restriction enzymes and fast DNA sequencing techniques.]

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserantant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M., Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene, *Nature* **260,** 500–507 (1976).

Frederick, C. A., Wang, B.-C., Greene, P., Boyer, H. W., Grable, J., and Rosenberg, J. M., Kinked DNA in crystalline complex with *Eco*RI endonuclease, *Science* **234,** 1526–1551 (1987).

Ghosh, P. K., Reddy, V. B., Piatak, M., Lebowitz, P., and Weissman, S. M., Determination of RNA sequences by primer directed synthesis and sequencing of their cDNA transcripts, *Methods Enzymol.* **65,** 580–595 (1980).

Gusella, J. F., DNA polymorphism and human disease, *Annu. Rev. Biochem.* **55,** 831–854 (1986).

Hindley, J., DNA sequencing, *in* Work, T. S. and Burdon, R. S. (Eds.), *Laboratory Techniques in Molecular Biology,* Vol. 10, North–Holland (1983).

Maxam, A. M. and Gilbert, W., Sequencing end-labeled DNA with base-specific chemical cleavages, *Methods Enzymol.* **65,** 499–560 (1980).

Nathans, D. and Smith, M. O., Restriction endonucleases in the analysis and restructuring of DNA, *Annu. Rev. Biochem.* **44,** 273–293 (1975).

Peattie, D. M., Direct chemical method for sequencing RNA, *Proc. Natl. Acad. Sci.* **76,** 1760–1764 (1979).

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A., and Baumeister, K., A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides, *Science* **238,** 336–341 (1987).

Rosenberg, J. M., McClarin, J. A., Frederick, C. A., Wang, B.-C., Grable, J., Boyer, H. W., and Greene, P., Structure and recognition mechanism of *Eco*RI endonuclease, *Trends Biochem. Sci.* **12,** 395–398 (1987).

Sanger, F., Sequences, sequences, and sequences, *Annu. Rev. Biochem.* **57,** 1–28 (1988). [A scientific memoir.]

Smith, A. J. H., DNA sequence analysis by primed synthesis, *Methods Enzymol.* **65,** 560–580 (1980). [The chain-terminator method.]

Weissman, S. M., *Methods of DNA and RNA Sequencing,* Preager (1983).

Wells, R. D., Klein, R. D., and Singleton, C. K., Type II restriction enzymes, *in* Boyer, P. D. (Ed.), *The Enzymes* (3rd ed.), Vol. 14, *pp.* 137–156, Academic Press (1981).

White, R. and Lalouel, J.-M., Chromosome mapping with DNA markers, *Sci. Am.* **258**(2): 40–48 (1988). [Describes the use of RFLPs.]

Yuan, R., Structure and mechanism of multifunctional restriction endonucleases, *Annu. Rev. Biochem.* **50,** 285–315 (1981). [A discussion of Types I and III restriction endonucleases.]

Chemical Synthesis of Oligonucleotides

Caruthers, M. H., Barone, A. D., Beaucage, S. L., Dodds, D. R., Fisher, E. F., McBride, L. J., Matteucci, M., Stabinsky, Z., and Tang, J. Y., Chemical synthesis of deoxyoligonucleotides, *Methods Enzymol.* **154,** 287–313 (1987).

Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L., and Wallace, R. B., Detection of sickle cell β^s-globin allele by hybridization with synthetic oligonucleotides, *Proc. Natl. Acad. Sci.* **80,** 278–282 (1983).

Gait, M. J. (Ed.), *Oligonucleotide Synthesis. A Practical Approach,* IRL Press (1984).

Itakura, K., Rossi, J. J., and Wallace, B. R., Synthesis and use of synthetic oligonucleotides, *Annu. Rev. Biochem.* **53,** 323–356 (1984).

Khorana, H. G., Total synthesis of a gene, *Science* **203,** 614–625 (1979). [The use of the classical phosphodiester method of oligonucleotide synthesis.]

Molecular Cloning

Berger, S. L. and Kimmel, A. R. (Eds.), Guide to Molecular Cloning Techniques, *Methods Enzymol.* **152** (1987). [A "cookbook" describing the basic techniques of molecular biology.]

Burke, D. T., Carle, G. F., and Olso, M. V., Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors, *Science* **236,** 806–812 (1987).

Glover, D. M. (Eds.), *DNA Cloning. A Practical Approach,* Vols. 1 and 2, IRL Press (1985).

Johnson, I. S., Human insulin from recombinant DNA technology, *Science* **219,** 632–637 (1983).

Leatherbarrow, R. J. and Fersht, A. R., Protein engineering, *Protein Eng.* **1,** 7–16 (1986).

Maniatis, T., Fritsch, E. F., and Sambrook, J., *Molecular Cloning. A Laboratory Manual,* Cold Spring Harbor Laboratory (1982). [Contains extensive data and descriptions of techniques for preparing recombinant DNA.]

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A., Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science* **239,** 487–494 (1988).

Watson, J. D. and Tooze, J., *The DNA Story,* Freeman (1981). [A scrapbooklike account of the recombinant DNA debate of the 1970s.]

Watson, J. D., Tooze, J., and Kurtz, D. T., *Recombinant DNA. A Short Course,* Scientific American Books (1983).

Wu, R., Grossman, L., and Moldave, K., (Eds.), Recombinant DNA, Parts A–F, *Methods Enzymol.* **68, 100, 101,** and **153–155** (1979, 1983, and 1987).

Historical Aspects

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 narration of the discovery of the DNA double helix.]

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

Portugal, F. H. and Cohen, J. S., *A Century of DNA,* MIT Press (1977).

Sayre, A., *Rosalind Franklin and DNA,* Norton (1975). [A biographical work 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.]

Schlenk, F., Early nucleic acid chemistry, *Trends Biochem. Sci.* **13,** 67–68 (1988).

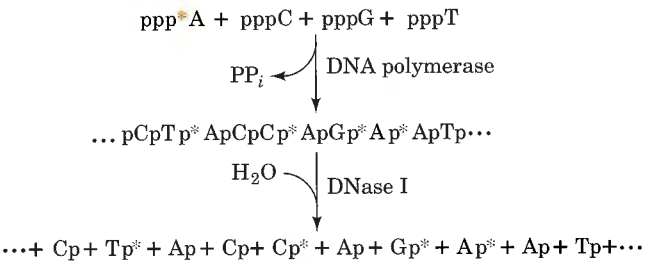
Watson, J. D., *The Double Helix,* Atheneum (1968). [A lively but controversial autobiographical account of the discovery of the DNA structure.]

Problems

1. 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 (the three consecutive nucleotides that base pair with mRNA). With what base on mRNA is hypoxanthine likely to pair? Draw the structure of this base pair. (b) tRNA often makes a G·U base pair with mRNA. 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 rare tautomeric form of adenine. What base sequence change would be caused by such mispairing?

2. What is the molecular mass and contour length of a section of B-DNA that specifies a 40-kD protein? Each amino acid is specified by three contiguous bases on a single strand of DNA (Section 30-1).

*3. 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 [α-^{32}P]-labeled and three unlabeled deoxynucleoside triphosphates. The resulting product is hydrolyzed by a DNase that cleaves 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. 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 equivalences in both cases?

4. What would be the effect of the following agents on the melting curve of an aqueous solution of duplex DNA? Explain. (a) Decreasing the ionic strength of the solution. (b) Squirting the DNA solution, at high pressure, through a very narrow orifice. (c) Bringing the solution to 0.1*M* adenine. (d) Heating the solution to 25°C above the DNA's melting point and then rapidly cooling it to 25°C below the DNA's melting point. (e) Adding a small amount of ethanol to the DNA solution.

5. What is the mechanism of alkaline denaturation of DNA?

*6. At Na^+ concentrations $> 10M$, the T_m of DNA decreases with increasing [Na^+]. Explain this behavior. (*Hint:* Consider the solvation requirements of Na^+.)

*7. Why are the most commonly observed conformations of the ribose ring those in which either atom C(2′) or C(3′) is out of the plane of the other four ring atoms. (*Hint:* In puckering a planar ring such that one atom is out of the plane of the other four, the substituents about the bond opposite the out-of-plane atom remain eclipsed. This is best observed with a ball-and-stick model.)

8. Polyoma virus DNA can be separated by sedimentation at neutral pH into three components that have sedimentation coefficients of 20, 16, and 14.5S and which are known as Types I, II, and III DNAs, respectively. These DNAs all have identical base compositions and molecular masses. In 0.15*M* NaCl, both Types II and III DNAs have melting curves of normal cooperativity and a T_m of 88°C. Type I DNA, however, exhibits a very broad melting curve and a T_m of 107°C. At pH 13, Types I and III DNAs have sedimentation coefficients of 53 and 16S, respectively, and Type II separates into two components with sedimentation coefficients of 16S and 18S. How do Types I, II, and III DNAs differ from one another? Explain their different physical properties.

9. A closed circular duplex DNA has a 100 bp segment of alternating C and G residues. Upon transfer to a solution containing a high salt concentration, this segment undergoes a transition from the B conformation to the Z conformation. What is the accompanying change in its linking number, writhing number, and twist?

10. You have discovered an enzyme secreted by a particularly virulent bacterium that cleaves the C(2′)—C(3′) bond in the deoxyribose residues of duplex DNA. What is the effect of this enzyme on supercoiled DNA ?

11. 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 *Taq*I, *Eco*RII, *Pst*I, and *Hae*II be expected to make in SV40 DNA? (Figure 28-49 indicates the number of restriction cuts that these enzymes actually make.)

12. A bacterial chromosome consists of a protein–DNA complex in which its single DNA molecule appears to be supercoiled as demonstrated by ethidium bromide titration. However, in contrast to the case with naked circular duplex DNA, the single-strand nicking of chromosomal DNA does not abolish this supercoiling. What does this indicate about the structure of the bacterial chromosome, that is, how do its proteins constrain its DNA?

13. Which of the restriction endonucleases listed in Table 28-5 produce blunt ends? Which sets of them are **isoschizomers** (enzymes that have the same recognition se-

quence 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)?

14. In investigating a newly discovered bacterial species that inhabits the sewers of Berkeley, you isolate a plasmid which you suspect carries genes that confer resistance to several antibiotics. To characterize this plasmid you decide to make its restriction map. The sizes of the plasmid's restriction fragments, as determined from their electrophoretic mobilities on agarose gels, are given in the following table. From the data, construct the restriction map of the plasmid.

Sizes of Restriction Fragments from a Plasmid DNA

Restriction Enzymes	Fragment Sizes (kb)
*Eco*RI	5.4
*Hin*dIII	2.1, 1.9, 1.4
*Sal*I	5.4
*Eco*RI + *Hin*dIII	2.1, 1.4, 1.3, 0.6
*Eco*RI + *Sal*I	3.2, 2.2
*Hin*dIII + *Sal*I	1.9, 1.4, 1.2, 0.9

15. Figure 28-70 pictures an autoradiograph of the sequencing gel of a *Hae*III restriction fragment from the *E. coli* K12 gene that codes for dihydrofolate reductase. The DNA was treated according to the chemical cleavage method of DNA sequencing after being ^{32}P labeled at its 3′ end. Read the sequence of the first 50 bases from the bottom of the gel.

16. 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 chromosome consists of 13,500 kb.

17. 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 λ 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?

Figure 28-70
[From Smith, D. R. and Calvo, J. M., *Nuclei Acids Res.* **8,** 2268 (1980).]

Chapter 29

TRANSCRIPTION

There are three major classes of RNA, all of which participate in protein synthesis: **ribosomal RNA (rRNA), transfer RNA (tRNA),** and **messenger RNA (mRNA).** All of these RNAs are synthesized under the direction of DNA templates, a process known as **transcription.**

RNA's involvement in protein synthesis became evident in the late 1930s through investigations by Torbjörn Caspersson and Jean Brachet. Caspersson, using microscopic techniques, found that DNA is confined almost exclusively to the eukaryotic cell nucleus, whereas RNA occurs largely in the cytosol. Brachet, who had devised methods for fractionating cellular organelles, came to similar conclusions based on direct chemical analyses. He found, in addition, that the cytosolic RNA-containing particles are also protein rich. Both investigators noted that the concentration of these RNA-protein particles (which were later named ribosomes) is correlated with the rate that a cell synthesizes protein, inferring a relationship between RNA and protein synthesis. Indeed, Brachet even suggested that *the RNA-protein particles are the site of protein synthesis.*

Brachet's suggestion was shown to be valid when radioactively labeled amino acids became available in the 1950s. A short time after injecting a rat with a labeled amino acid, most of the label that had been incorporated in proteins was associated with ribosomes. This experiment also established that *protein synthesis is not immediately directed by DNA because, at least in eukaryotes, DNA and ribosomes are never in contact.*

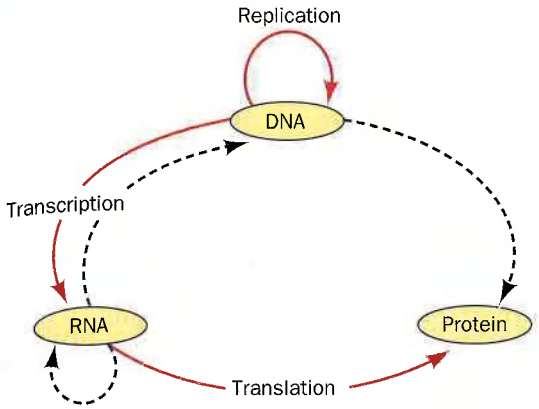


Figure 29-1
The central dogma of molecular biology. Solid arrows indicate the types of genetic information transfers that occur in all cells. Special transfers are indicated by the dashed arrows: RNA-directed RNA polymerase occurs both in certain RNA viruses and in some plants (where it is of unknown function); RNA-directed DNA polymerase (reverse transcriptase) occurs in 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 recipients of gentic information.* [After Crick, F., *Nature* **227,** 562 (1970).]

In 1958, Francis Crick summarized the then dimly perceived relationships among DNA, RNA, and protein in a flow scheme he described as the **central dogma** of molecular biology: *DNA directs its own replication and its transcription to RNA which, in turn, directs its translation to proteins (Fig. 29-1).*

> The peculiar use of the word "dogma," one definition of which is a religious doctrine that the true believer cannot doubt, stemmed from a misunderstanding. When Crick formulated the central dogma, he was under the impression that dogma meant "an idea for which there was no reasonable evidence."

We begin this chapter by discussing experiments that led to the elucidation of mRNA's central role in protein synthesis. We then study the mechanism of transcription and its control in prokaryotes. Finally, in the last section, we consider post-transcriptional processing of RNA in both prokaryotes and eukaryotes. Translation is the subject of Chapter 30.

1. THE ROLE OF RNA IN PROTEIN SYNTHESIS

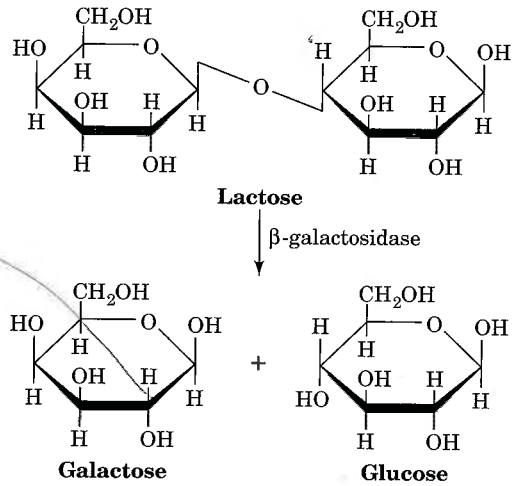
Proteins are specified by mRNA and synthesized on ribosomes. This idea arose from the study of **enzyme induction,** a phenomenon in which bacteria vary the synthesis rates of specific enzymes in response to environmental changes. We shall see below that *enzyme induction occurs as a consequence of the regulation of mRNA synthesis by proteins that specifically bind to the mRNA's DNA templates.*

A. Enzyme Induction

E. coli can synthesize an estimated 3000 different polypeptides (Section 27-1D). There is, however, enormous variation in the amounts of these different polypeptides that are produced. For instance, the various ribosomal proteins may each be present in over 10,000 copies per cell, whereas certain regulatory proteins (see below) normally occur in <10 copies per cell. Many enzymes, particularly those involved in basic cellular "housekeeping" functions, are synthesized at a more or less constant rate; they are called **constitutive enzymes.** Other enzymes, termed **adaptive** or **inducible enzymes,** are synthesized at rates that vary with the cell's circumstances.

Lactose-Metabolizing Enzymes Are Inducible

Bacteria, as has been recognized since 1900, adapt to their environments by producing enzymes that metabolize certain nutrients, for example, lactose, only when those substances are available. *E. coli* grown in the absence of lactose are initially unable to metabolize this disaccharide. To do so they require the presence of two proteins: **β-galactosidase,** which catalyzes the hydrolysis of lactose to its component monosaccharides;



and **galactoside permease** (also known as **lactose permease;** Section 18-4B), which transports lactose into the cell. *E. coli* grown in the absence of lactose contain only a few molecules of these proteins. Yet, a few minutes after lactose is introduced into their medium, *E. coli* increase the rate at which they synthesize these proteins by ~1000-fold and maintain this pace until lactose is no longer available. The synthesis rate then returns to its original miniscule level (Fig. 29-2). *This ability to produce a series of proteins only when the substances they metabo-*

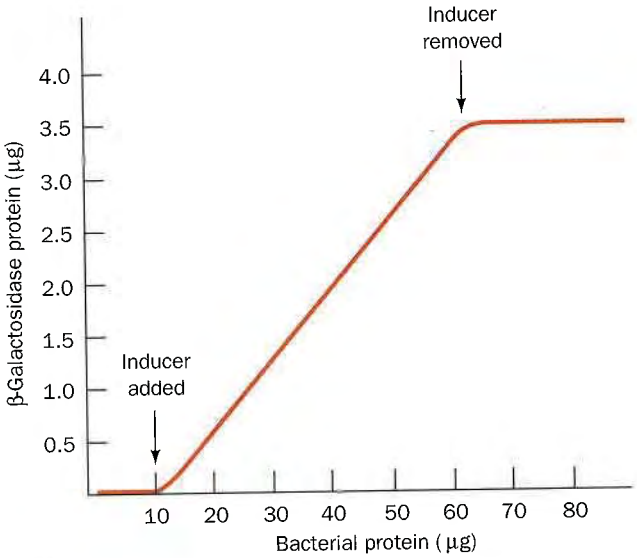
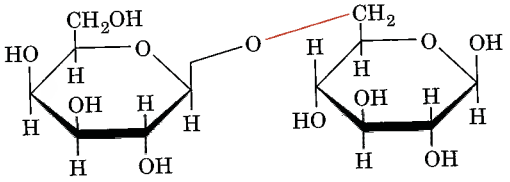


Figure 29-2
The induction kinetics of β-galactosidase in *E. coli.* [After Cohn, M., *Bacteriol. Rev.* **21,** 156 (1957).]

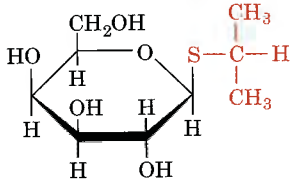
lize are present permits bacteria to adapt to their environment without the debilitating need to continuously synthesize large quantities of otherwise unnecessary substances.

Lactose or one of its metabolic products must somehow trigger the synthesis of the above proteins. Such a substance is known as an **inducer.** The physiological inducer of the lactose system, the lactose isomer **1,6-allolactose,**



1,6-Allolactose

arise's from lactose's occasional transglycosylation by β-galactosidase. Most studies of the lactose system use **isopropylthiogalactoside (IPTG),**



Isopropylthiogalactoside (IPTG)

a potent inducer that structurally resembles allolactose but which is not degraded by β-galactosidase.

Lactose system inducers also stimulate the synthesis of **thiogalactoside transacetylase,** an enzyme that, *in vitro,* transfers an acetyl group from acetyl-CoA to the C(6)-OH group of a β-thiogalactoside such as IPTG. Since lactose fermentation procedes normally in the absence of thiogalactoside transacetylase, however, this enzyme's physiological role is unknown.

lac System Genes Form an Operon

The genes specifying wild-type β-galactosidase, galactoside permease, and thiogalactoside transacetylase are designated Z^+, Y^+, and A^+, respectively. Genetic mapping of the defective mutants Z^-, Y^-, and A^- indicated that these *lac* **structural genes** (genes that specify polypeptides) are contiguously arranged on the *E. coli* chromosome (Fig. 29-3; genetic mapping is reviewed in Section 27-1). *These genes, together with the control elements P and O, form a genetic unit called an* ***operon,*** *specifically the* ***lac operon.*** The nature of the control elements is discussed below. The role of operons in prokaryotic gene expression is examined in Section 29-3.

lac Repressor Inhibits the Synthesis of *lac* Operon Proteins

An important clue as to how *E. coli* synthesizes protein was provided by a mutation that causes the proteins of the *lac* operon to be synthesized in large amounts in the absence of inducer. This so-called **constitutive mutation** occurs in a gene, designated *I,* that is distinct from although closely linked to the genes specifying the *lac* enzymes (Fig. 29-3). What is the nature of the *I* gene product? This riddle was solved through an ingeneous experiment performed by Arthur Pardee, Francois Jacob, and Jacques Monod. Hfr bacteria of genotype I^+Z^+ were mated to an F^- strain of genotype I^-Z^- in the absence of inducer while the β-galactosidase activity of the culture was monitored (Fig. 29-4; bacterial mating is described in Section 27-1D). At first, as expected, there was no β-galactosidase activity because the Hfr donors lacked inducer and the F^- recipients were unable to produce active enzyme (only DNA passes through the cytoplasmic bridge connecting mating bacteria). About 1 h after conjugation began, however, when the I^+Z^+ genes had just entered the F^- cells, β-galactosidase synthesis began and only ceased after about another hour. The explanation for these observations is that the donated Z^+ gene, upon entering the cytoplasm of the I^- cell, directs the synthesis of β-galactosidase in a constitutive manner. Only after the donated I^+ gene has had sufficient time to be expressed is it able to repress β-galactosidase synthesis. *The I^+ gene must therefore give rise*

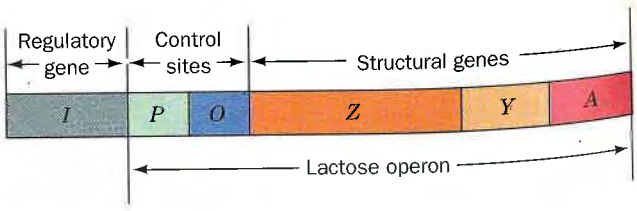


Figure 29-3
A genetic map of the *E. coli* genes encoding the proteins mediating lactose metabolism and the genetic sites that control their expression. The *Z*, *Y*, and *A* genes, respectively, specify β-galactosidase, galactoside permease, and thiogalactoside transacetylase.

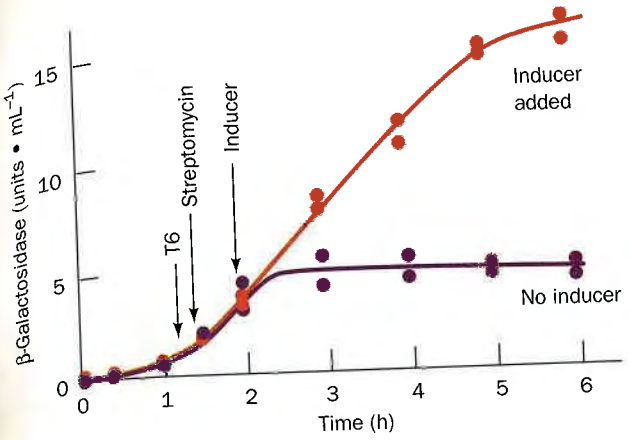


Figure 29-4
The appearance of β-galactosidase in the transient merozygotes (partial diploids) formed by mating I^+Z^+ Hfr donors with a I^-Z^- F^- recipients. The F^- strain was also resistant to both bacteriophage T6 and streptomycin, whereas the Hfr strain was sensitive to these agents. Both types of cells were grown and mated in the absence of inducer. After sufficient time had passed for the transfer of the *lac genes*, the Hfr cells were selectively killed by the addition of T6 phage and streptomycin. In the absence of inducer (*lower curve*), β-galactosidase synthesis commenced at around the time that the *lac* genes had entered the F^- cells and continued for ~1 h. If inducer was added shortly after the Hfr donors had been killed (*upper curve*), enzyme synthesis continued unabated. This demonstrates that the cessation of β-galactosidase synthesis in uninduced cells is not due to the intrinsic loss of the ability to synthesize this enzyme. [After Pardee, A. B., Jacob, F., and Monod, J., *J. Mol. Biol.* **1**, 173 (1959).]

to a diffusible product, the ***lac repressor****, which inhibits the synthesis of β-galactosidase (and the other lac proteins).* Inducers such as IPTG temporarily inactivate *lac* repressor, whereas I^- cells constitutively synthesize *lac* enzymes because they lack a functional repressor. *Lac* repressor, as we shall see in Section 29-3B, is a protein.

B. Messenger RNA

The nature of the *lac* repressor's target molecule was deduced in 1961 through a penetrating genetic analysis by Jacob and Monod. A second type of constitutive mutation in the lactose system, designated O^c (for **operator constitutive**), which complementation analysis indicated to be independent of the *I* gene, maps between the *I* and Z genes (Fig. 29-3). In the partially diploid F′ strain $O^cZ^-/F\ O^+Z^+$, β-galactosidase activity is inducible by IPTG whereas the strain $O^c\ Z^+/F\ O^+Z^-$ constitutively synthesizes this enzyme (in F′ bacteria, the F factor plasmid contains a segment of the bacterial chromosome, in this case a portion of the *lac* operon; Section 27-1D). *An O^+ gene can therefore only control the expression of a Z gene on the same chromosome.* The same is true with the Y^+ and A^+ genes.

Jacob and Monod's observations led them to conclude the proteins are synthesized in two-stage process:

1. The structural genes on DNA are transcribed onto complementary strands of **messenger RNA (mRNA).**
2. The mRNAs transiently associate with ribosomes, which they direct in polypeptide synthesis.

This hypothesis explains the behavior of the *lac* system *(Fig. 29-5). In the absence of inducer, the lac repressor specifically binds to the O gene (the* ***operator****) so as to physically block the enzymatic transcription of mRNA. Upon binding inducer, the repressor dissociates from the operator thereby permitting the transcription and subsequent translation of the lac enzymes.* The operator–repressor–inducer system thereby acts as a molecular switch so that the *lac* operator can only control the expression of *lac* enzymes on the same chromosome. The O^c mutants constitutively synthesize *lac* enzymes because they are unable to bind repressor. The **coordinate** (simultaneous) expression of all three *lac* enzymes under the control of a single operator site arises, as Jacob and Monod theorized, from the transcription of the *lac*

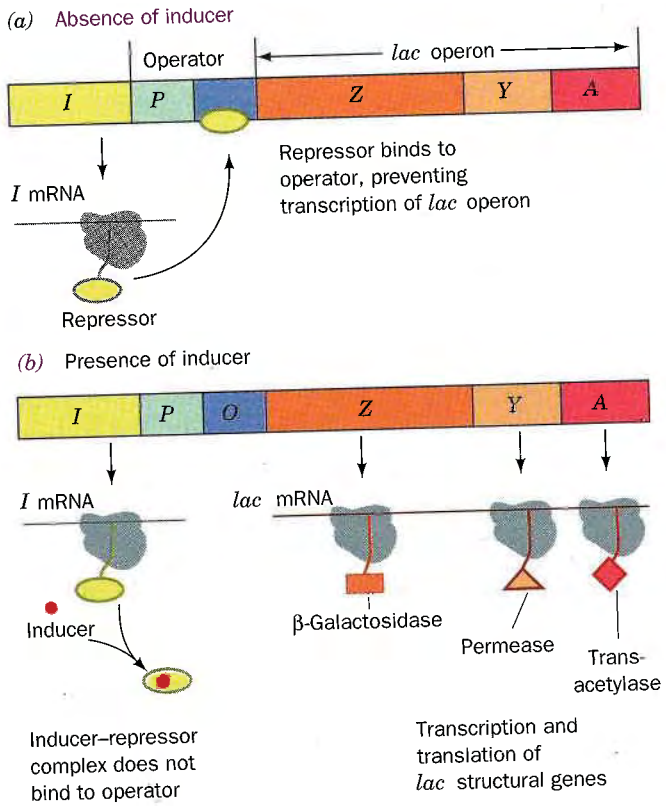


Figure 29-5
The expression in the *lac* operon. (*a*) In the absence of inducer, the repressor, the product of the *I* gene, binds to the operator thereby preventing transcription of the *lac* operon. (*b*) Upon binding inducer, the repressor dissociates from the operator, which permits the transcription and subsequent translation of the *lac* structural genes to proceed.

operon as a single **polycistronic mRNA** which directs the ribosomal synthesis of each of these proteins. This transcriptional control mechanism is further discussed in Section 29-3. [Pairs of DNA sequences, which are on the same DNA molecule, are said to be in cis (Latin: on this side) while those on different DNA molecules are said to be in trans (Latin: across). Control sequences such as the *O* gene, which are only active on the same DNA molecule as the genes they control, are called **cis-acting elements.** Those such as *lacI*, which specify the synthesis of diffusible products and can therefore be located on a different DNA molecule from the genes they control, are said to direct the synthesis of **trans-acting factors.**]

mRNAs Have Their Predicted Properties

The kinetics of enzyme induction, as indicated, for example, in Figs. 29-2 and 29-4, requires that the postulated mRNA be both rapidly synthesized and rapidly degraded. An RNA with such quick turnover had, in fact, been observed in T2-infected *E. coli.* Moreover, the base composition of this RNA fraction resembles that of the viral DNA rather than that of the bacterial RNA. Ribosomal RNA, which comprises up to 90% of a cell's RNA, turns over much more slowly than mRNA. Ribosomes are therefore not permanently committed to the synthesis of a particular protein (a once popular hypothesis). Rather, *ribosomes are nonspecific protein synthesizers that produce the polypeptide specified by the mRNA with which they are transiently associated.* A bacterium can therefore respond within a few minutes to changes in its environment.

Evidence favoring the Jacob and Monod model rapidly accumulated. Sydney Brenner, Jacob, and Matthew Meselson carried out experiments designed to characterize the RNA that *E. coli* synthesized after T4 phage infection. *E. coli* were grown in a medium containing ^{15}N and ^{13}C so as to label all cell constituents with these heavy isotopes. The cells were then infected with T4 phages and immediately transferred to an unlabeled medium (which contained only the light isotopes ^{14}N and ^{12}C) so that cell components synthesized before and after phage infection could be separated by equilibrium density gradient ultracentrifugation in CsCl solution. No "light" ribosomes were observed, which indicates, in agreement with the above mentioned T2 phage results, that no new ribosomes are synthesized after phage infection.

The growth medium also contained either ^{32}P or ^{35}S so as to radioactively label the newly synthesized and presumably phage-specific RNA and protein, respectively. Much of the ^{32}P-labeled RNA was associated, as was postulated for mRNA, with the preexisting "heavy" ribosomes (Fig. 29-6). Likewise, the ^{35}S-labeled proteins were transiently associated with, and therefore synthesized by, these ribosomes.

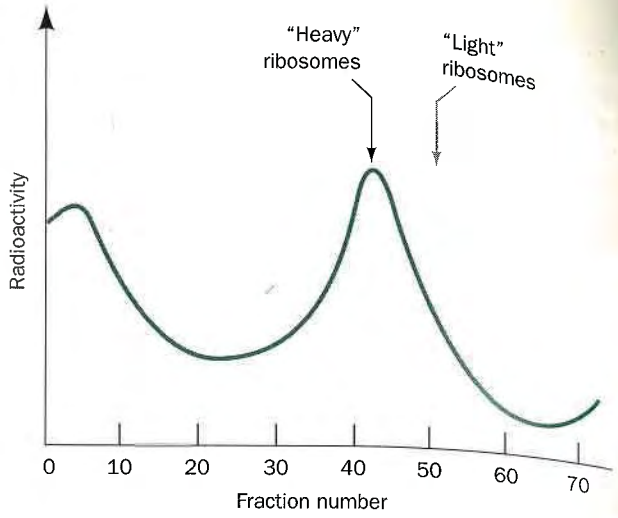


Figure 29-6
The distribution, in a CsCl density gradient, of ^{32}P-labeled RNA that had been synthesized by *E. coli* after T4 phage infection. Free RNA, being relatively dense, bands at the bottom of the centrifugation cell (*left*). Much of the RNA, however, is associated with the ^{15}N- and ^{13}C-labeled "heavy" ribosomes that had been synthesized before the phage infection. The predicted position of unlabeled "light" ribosomes, which are not synthesized by phage-infected cells, is also indicated. [After Brenner, S., Jacob, F., and Meselson, M., *Nature* **190,** 579 (1961).]

Sol Spiegelman developed the RNA–DNA hybridization technique (Section 28-3A) in 1961 to characterize the RNA synthesized by T2-infected *E. coli.* He found that this phage-derived RNA hybridizes with T2 DNA (Fig. 29-7) but neither does so with DNAs from unrelated phage nor with the DNA from uninfected *E. coli.* This RNA must therefore be complementary to T2 DNA in agreement with Jacob and Monod's prediction; that is, the phage-specific RNA is a messenger RNA. Hybridization studies have likewise shown that mRNAs from uninfected *E. coli* are complementary to portions of *E. coli* DNA. In fact, other RNAs, such as transfer RNA and ribosomal RNA, have corresponding complementary sequences on DNA from the same organism. Thus, *all cellular RNAs are transcribed from DNA templates.*

2. RNA POLYMERASE

RNA polymerase, the enzyme responsible for the DNA-directed synthesis of RNA, was discovered independently in 1960 by Samuel Weiss and Jerard Hurwitz. *The enzyme couples together the ribonucleoside triphosphates ATP, CTP, GTP, and UTP, on DNA templates in a reaction that is driven by the release and subsequent hydrolysis of* PP_i*:*

$$(\text{RNA})_{n\ \text{residues}} + \underset{\substack{\text{Nucleoside}\\ \text{triphosphate}}}{\text{NTP}} \rightleftharpoons (\text{RNA})_{n+1\ \text{residues}} + PP_i$$

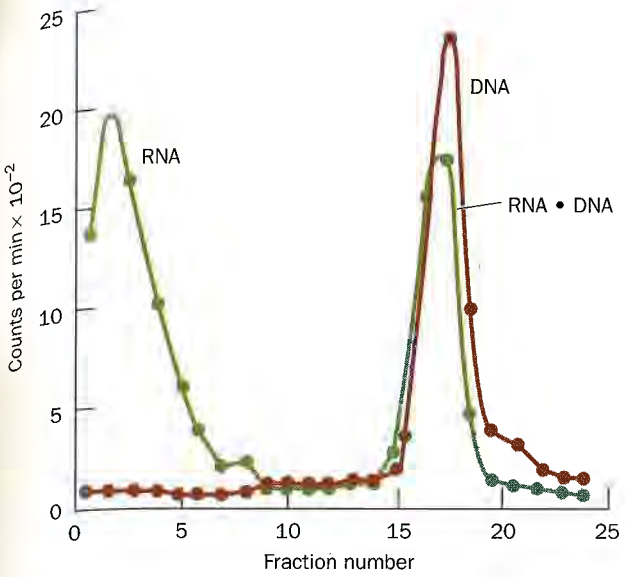


Figure 29-7
The hybridization of ^{32}P-labeled RNA produced by T2-infected *E. coli* with ^{3}H-labeled T2 DNA. Upon radioactive decay, ^{32}P and ^{3}H emit β particles with characteristically different energies so that these isotopes can be independently detected. Although free RNA (*left*) in a CsCl density gradient is denser than DNA, much of the RNA bands with the DNA (*right*). This indicates that the two polynucleotides have hybridized and are therefore complementary in sequence. [After Hall, B. D. and Spiegelman, S., *Proc. Natl. Acad. Sci.* **47,** 141 (1961).]

All cells contain RNA polymerase. In bacteria, one species of this enzyme synthesizes all of the cell's RNA except the short RNA primers employed in DNA replication (Section 31-1D). Various bacteriophages generate RNA polymerases that synthesize only phage-specific RNAs. Eukaryotic cells contain four or five RNA polymerases, that each synthesize a different class of RNA. In this section we first concentrate on the properties of the *E. coli* enzyme because it is the best characterized RNA polymerase; other bacterial RNA polymerases have similar properties. We then consider the eukaryotic enzymes.

A. Enzyme Structure

E. coli RNA polymerase's so-called **holoenzyme** is an ~480-kD protein with subunit composition $\alpha_2\beta\beta'\sigma$. Once RNA synthesis has been initiated, however, the σ subunit (also called **σ factor**) dissociates from the **core enzyme,** $\alpha_2\beta\beta'$, which carries out the actual polymerization process (see below). The β' subunit contains two atoms of Zn^{2+} which are thought to participate in the enzyme's catalytic function. The active enzyme also requires the presence of Mg^{2+}.

The holoenzyme, which is among the largest known soluble enzymes, is ~100 Å in diameter, which renders it visible in electron micrographs (Fig. 29-8); these clearly indicate that RNA polymerase binds to DNA as a protomer. The large size of the holoenzyme is presumably required by its several complex functions that include (1) template binding, (2) RNA chain initiation, (3) chain elongation, and (4) chain termination. We discuss these various functions below.

B. Template Binding

RNA synthesis is normally initiated only at specific sites on the DNA template. This was first demonstrated through hybridization studies of **bacteriophage ϕX174** DNA with the RNA produced by ϕX174-infected *E. coli.* Bacteriophage ϕX174 carries a single strand of DNA known as the "plus" strand. Upon its injection into *E. coli,* the plus strand directs the synthesis of the complementary "minus" strand with which it combines to form a circular duplex DNA known as the **replicative form** (Section 31-3B). The RNA produced by ϕX174-infected *E. coli* does not hybridize with DNA from intact phage but does so with the replicative form. Thus only the minus strand of ϕX174 DNA, the so-called **sense strand,** is transcribed, that is, acts as a template; the plus strand, the **antisense strand,** does not do so. Similar studies indicate that in larger phages, such as T4 and λ, the two viral DNA strands are the sense strands for different sets of genes. The same appears to be true of cellular organisms.

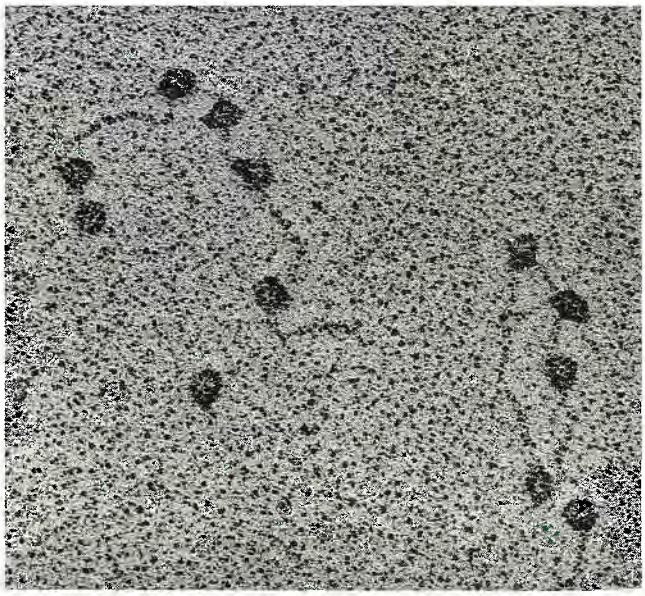


Figure 29-8
An electron micrograph of *E. coli* RNA polymerase holoenzyme attached to various promoter sites on bacteriophage T7 DNA. [From Williams, R. C., *Proc. Natl. Acad. Sci.* **74,** 2313 (1977).]

Holoenzyme Specifically Binds to Promoters

RNA polymerase binds to its initiation sites through base sequences known as ***promoters*** *that are recognized by the corresponding σ factor.* The existence of promoters was first recognized through mutations that enhance or diminish the transcription rates of certain genes including those of the *lac* operon. *Genetic mapping of such mutations indicated that the promoter consists of an ~40 bp sequence that is located on the 5′ side of the transcription start site.* [By convention, the sequence of template DNA is represented by its antisense (nontemplate) strand so that it will have the same directionality as the transcribed RNA. A base pair in a promoter region is assigned a positive or negative number that indicates its position, upstream or downstream in the direction of RNA polymerase travel, from the first nucleotide that is transcribed to RNA; this start site is +1 and there is no 0.] RNA, as we shall see, is synthesized in the 5′→ 3′ direction (Section 29-2D). Consequently, the promoter lies on the "upstream" side of the RNA's starting nucleotide. Sequencing studies indicate that the *lac* promoter *(lacP)* overlaps the *lac* operator (Fig. 29-3).

The holoenzyme forms tight complexes with promoters (dissociation constant $K \approx 10^{-14} M$) and thereby protects the bound DNA segments from digestion by DNase I. The region from about −20 to +20 is protected against exhaustive DNase I degradation. The region extending upstream to about −60 is also protected but to a lesser extent, presumably because it binds holoenzyme less tightly.

Sequence determinations of the protected regions from numerous *E. coli* and phage genes have revealed the "consensus" sequence of *E. coli* promoters (Fig. 29-9). *Their most conserved sequence is a hexamer centered at about the −10 position known as the* ***Pribnow box*** *(after David Pribnow who pointed out its existence in 1975).* It has a consensus sequence of TATAAT in which the leading TA and final T are highly conserved. *Upstream sequences around position −35 also have a region of sequence similarity,* TCTTGACAT, which is most evident in efficient promoters. The initiating (+1) nucleotide, which is nearly always A or G, is centered in a poorly conserved CAT or CGT sequence located 5 to 8 bp downstream from the Pribnow box. Most promoter sequences vary considerably from the consensus sequence (Fig. 29-9). Nevertheless, a mutation in one of the partially conserved regions can greatly increase or decrease a promoter's initiation efficiency. *The rates at which genes are transcribed, which span a range of at least 1000, varies directly with the rate that their promoters form stable initiation complexes with the holoenzyme.*

Initiation Requires the Formation of an Open Complex

The promoter regions in contact with the holoenzyme have been identified by determining where the enzyme

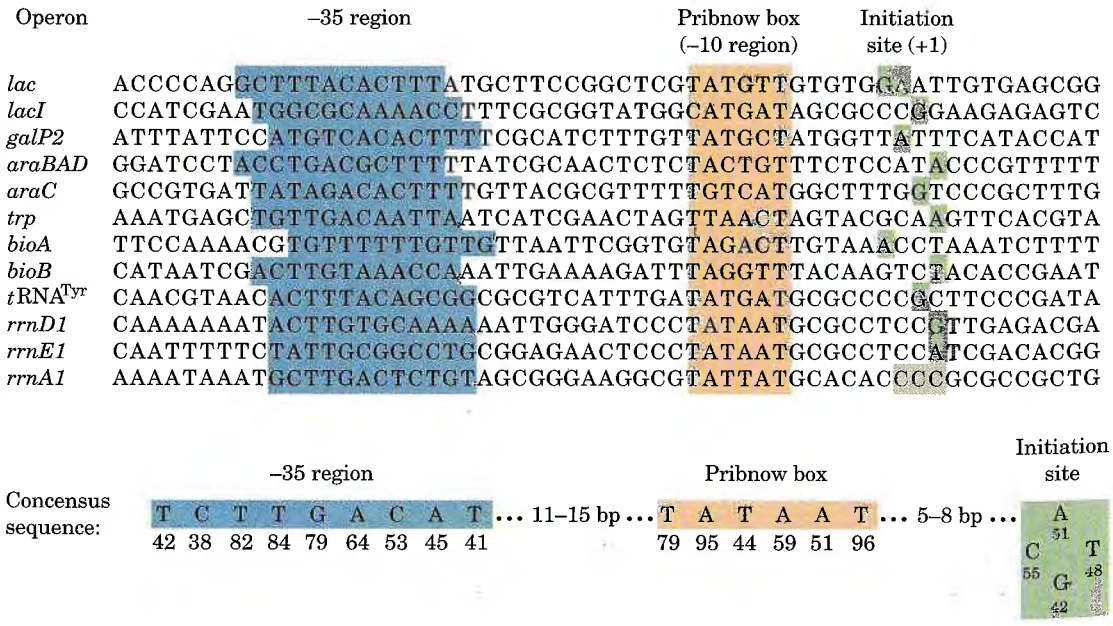


Figure 29-9
The noncoding strand nucleotide sequences of selected *E. coli* promoters. The Pribnow box (*red shading*), a 6 bp region centered around the −10 position, and an 8 to 12 bp sequence around the −35 region (*blue shading*) are both conserved. The transcription initiation sites (+1), which in most promoters occurs at a single purine nucleotide, are shaded in green. The bottom row shows the consensus sequence of 112 promoters with the number below each base indicating its percent occurrence. [After Rosenberg, M. and Court, D., *Annu. Rev. Genet.* **13,** 321–323 (1979). Consensus sequence from Hawley, D. K. and McClure, W. R., *Nucleic Acids Res.* **11,** 2244 (1983).]

alters the susceptibility of the DNA to alkylation by agents such as dimethyl sulfate (DMS), a procedure named **footprinting** (Section 33-3B). These experiments demonstrated that the holoenzyme contacts the promoter only around the Pribnow box and the −35 region. Model building indicates that these protected sites are both on the same side of the double helix which suggests that RNA polymerase binds to only one face of the double helical promoter.

DMS, in addition to methylating G residues at N(7) and A residues at N(3) (Section 28-6B), methylates N(1) of A and N(3) of C. Since these latter positions participate in base pairing interactions, however, they can only react with DMS in single-stranded DNA. This differential methylation of single- and double-standed DNAs provides a sensitive test for DNA strand separation or "melting." Footprinting studies indicate that the binding of holoenzyme "melts out" the promoter in an 11 bp region extending from the middle of the Pribnow box to just past the initiation site (−9 to +2). The need to form this "open complex" explains why promoter efficiency tends to decrease with the number of G·C base pairs in the Pribnow box; this presumably increases the difficulty in opening the double helix as is required for chain initiation (G·C pairs, it will be recalled, are stronger than A·T pairs).

Core enzyme, which does not specifically bind promoter, tightly binds duplex DNA (the complex's dissociation constant is $K \approx 5 \times 10^{-12} M$ and its half-life is ~60 min). Holoenzyme, in contrast, binds to nonpromoter DNA comparatively loosely ($K \approx 10^{-7} M$ and half-life >1 s). Apparently, the σ subunit allows holoenzyme to move rapidly along a DNA strand in search of the σ subunit's corresponding promoter. Once transcription has been initiated and the σ subunit jettisoned, the tight binding of core enzyme to DNA apparently stabilizes the ternary enzyme–DNA–RNA complex.

C. Chain Initiation

The 5′-terminal base of prokaryotic RNAs is almost always a purine with A occurring more often than G. The initiating reaction of transcription is the coupling of two nucleoside triphosphates in the reaction

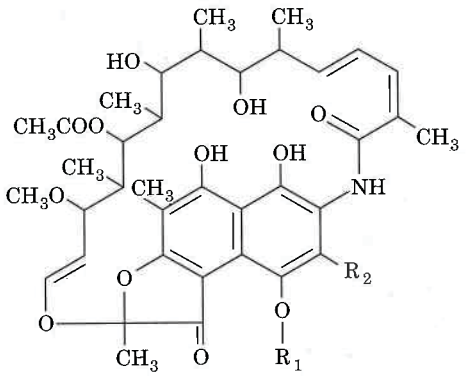
$$\text{pppA} + \text{pppN} \rightleftharpoons \text{pppApN} + \text{PP}_i$$

Bacterial RNAs therefore have 5′-triphosphate groups as was demonstrated by the incorporation of radioactive label into RNA when it was synthesized with [γ-^{32}P]ATP. Only the 5′ terminus of the RNA can retain the label because the internal phosphodiester groups of RNA are derived from the α-phosphate groups of nucleoside triphosphates.

Once holoenzyme has initiated RNA transcription, the σ factor dissociates from the core–DNA–RNA complex and can join with another core to form a new initiation complex. This is demonstrated by a burst of RNA synthesis upon the addition of core enzyme to a transcribing reaction mixture that initially contained holoenzyme.

Rifamycins Inhibit Prokaryotic Transcription Initiation

Two related antibiotics, **rifamycin B,** which is produced by *Streptomyces mediterranei,* and its semisynthetic derivative **rifampicin,**



Rifamycin B $R_1 = CH_2COO^-$; $R_2 = H$

Rifampicin $R_1 = H$; $R_2 = CH{=}N^+$(piperazine)$N{-}CH_3$

specifically inhibit transcription by prokaryotic, but not eukaryotic, RNA polymerases. This selectivity and their high potency (bacterial RNA polymerase is 50% inhibited by $2 \times 10^{-8} M$ rifampicin) has made them medically useful bacteriocidal agents against gram-positive bacteria and tuberculosis. The isolation of rifamycin resistant mutants whose β subunits have altered electrophoretic mobilities indicates that this subunit contains the rifamycin-binding site. Rifamycins neither inhibit the binding of RNA polymerase to the promoter nor the formation of the first phosphodiester bond, but they prevent further chain elongation. The inactivated RNA polymerase remains bound to the promoter thereby blocking its initiation by uninhibited enzyme. Once RNA chain initiation has occurred, however, rifamycins have no effect on the subsequent elongation process. The rifamycins are useful research tools because they permit the transcription process to be dissected into its initiation and its elongation phases.

D. Chain Elongation

What is the direction of RNA chain elongation; that is, does it occur by the addition of incoming nucleotides to the 3′ end of the nascent (growing) RNA chain (5′ → 3′ growth; Fig. 29-10*a*), or by their addition to its 5′ terminus (3′ → 5′ growth; Fig. 29-10*b*)? This question was answered by determining the rate that the radioactive label from [γ-^{32}P]GTP is incorporated into RNA. The

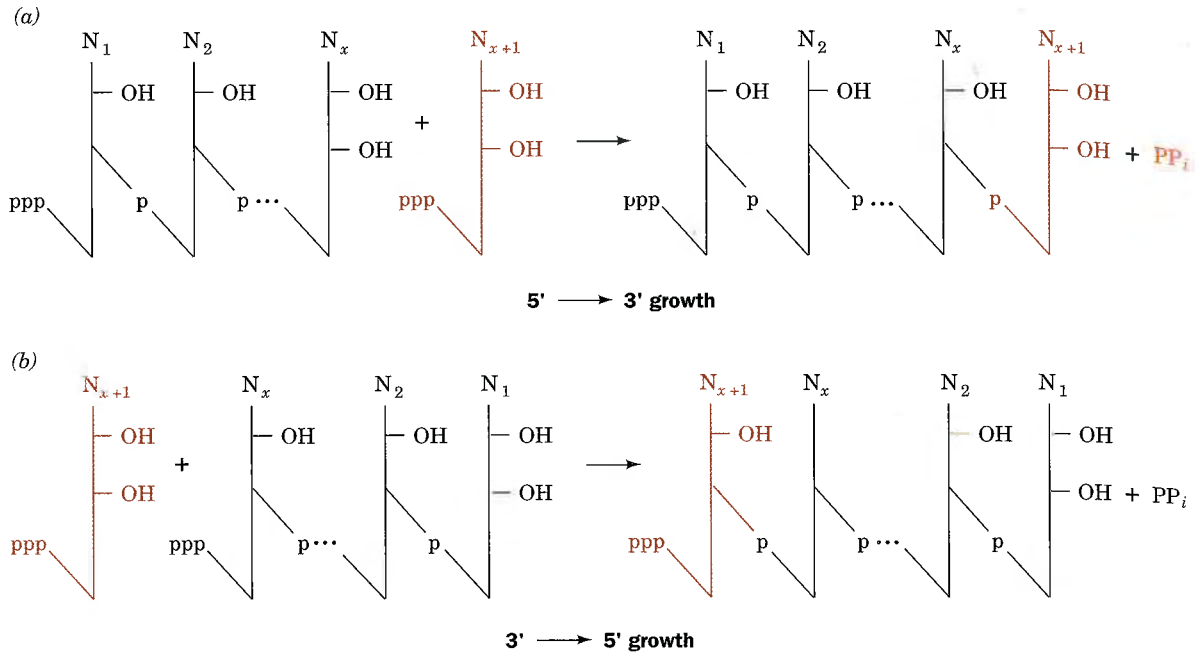
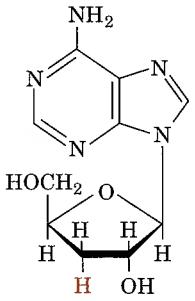


Figure 29-10
The two possible modes of RNA chain growth: (*a*) by the addition of nucleotides to the 3′ end, and (*b*) by the addition of nucleotides to the 5′ end. RNA polymerase catalyzes the former reaction.

ratio of ^{32}P to total nucleotide was highest just after chain initiation and decreased with time thereby indicating that the 5′-terminal pppG was incorporated into the RNA chain first rather than last. *Chain growth must therefore occur in the 5′ → 3′ direction (Fig. 29-10a).* This conclusion is corroborated by the observation that the antibiotic **cordycepin,**



Cordycepin
(3'-deoxyadenosine)

an adenosine analog that lacks a 3′-OH group, inhibits bacterial RNA synthesis. Its addition to the 3′ end of RNA, as is expected for 5′ → 3′ growth, prevents the RNA chain's further elongation. Cordycepin would not have this effect if chain growth occurred in the opposite direction because it cannot be appended to an RNA's 5′ end.

Transcription Probably Supercoils DNA

RNA chain elongation requires that the double-stranded DNA template be opened up at the point of RNA synthesis so that the sense strand can be transcribed onto its complementary RNA strand. In doing so, the RNA chain only transiently forms a short length of RNA–DNA hybrid duplex as is indicated by the observation that transcription leaves the template duplex intact and yields single-stranded RNA. The unpaired "bubble" of DNA in the open initiation complex apparently travels along the DNA with the RNA polymerase. There are two ways this might occur (Fig. 29-11):

1. If the RNA polymerase follows the template strand in its helical path around the DNA, the DNA would build up little supercoiling because the DNA duplex would never be unwound by more than about a turn. However, the RNA transcript would wrap around the DNA, once per duplex turn. This model is implausible since it is unlikely that its DNA and RNA could be readily untangled: The RNA would not spontaneously unwind from the long and often circular DNA in any reasonable time, and no topoisomerase is known to accelerate this process.

2. If the RNA polymerase moves in a straight line while the DNA rotates, the RNA and DNA would not become entangled. Rather, the DNA's helical turns would be pushed ahead of the advancing transcription bubble so as to more tightly wind the DNA ahead of the bubble (which promotes positive supercoiling) while the DNA behind the bubble would be equivalently unwound (which promotes negative supercoiling, although note that the linking number of the entire DNA remains unchanged). This model is supported by the observations that the transcription of plasmids in *E. coli* causes their positive supercoil-

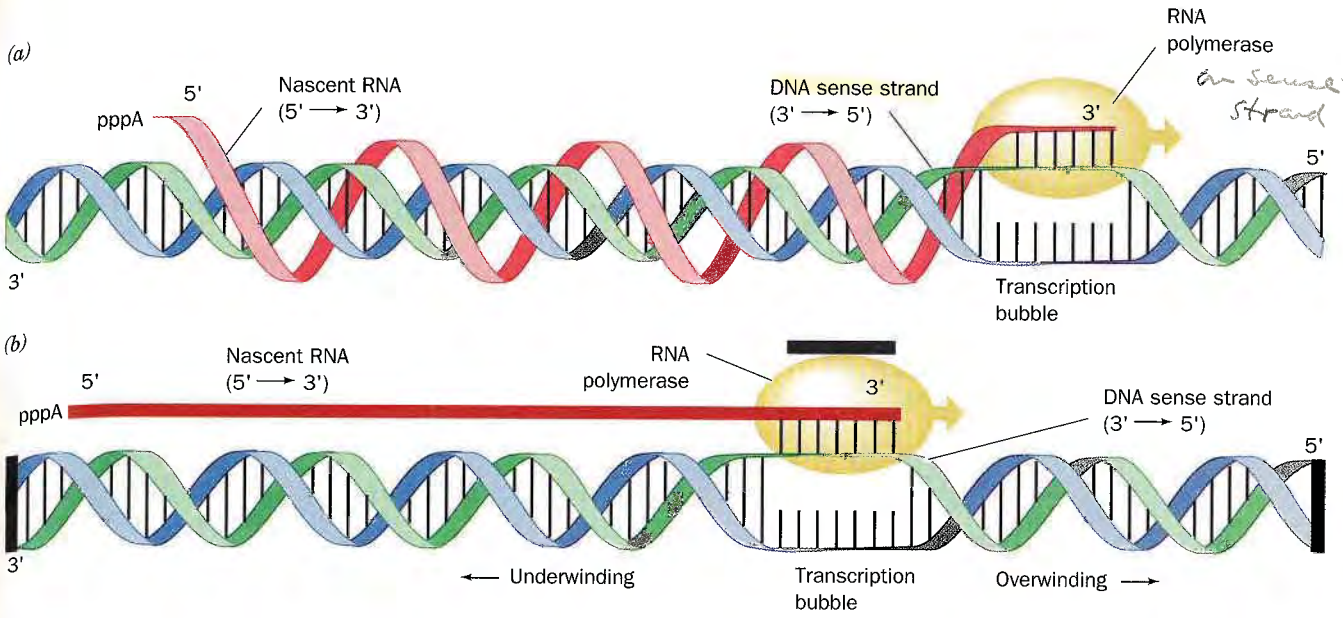


Figure 29-11
RNA chain elongation by RNA polymerase. In the region being transcribed, the DNA double helix is unwound by about a turn to permit the DNA's sense strand to form a short segment of DNA–RNA hybrid double helix with the RNA's 3′ end. As the RNA polymerase advances along the DNA template (here 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 sense strand. (*a*) One way this might occur is by the RNA polymerase following the path of the sense strand about the DNA double helix in which case the transcript becomes wrapped about the DNA once per duplex turn. (*b*) A second, and more plausible possibility, is that the RNA moves in a straight line while the DNA rotates beneath it. In this case the RNA would not wrap around the DNA but the DNA would become overwound ahead of the advancing transcription bubble and unwound behind it (consider the consequences of placing your finger between the twisted DNA strands in this model and pushing towards the right). The model presumes that the ends of the DNA as well as the RNA polymerase, are prevented from rotating by attachments within the cell (*black bars*). [After Futcher, B., *Trends Genet.* **4,** 271, 272 (1988).]

ing in gyrase mutants (which cannot relax positive supercoils; Section 28-5C) and their negative supercoiling in topoisomerase I mutants (which cannot relax negative supercoils).

Whatever the case, recall that inappropriate superhelicity halts transcription (Section 28-5C). Perhaps the torsional tension in the DNA generated by negative superhelicity behind the transcription bubble is required to help drive the transcriptional process, whereas too much such tension prevents the opening and maintenance of the transcription bubble.

Transcription Occurs Rapidly and Accurately

The *in vivo* rate of transcription is 20 to 50 nucleotides/s at 37°C as indicated by the rate that *E. coli* incorporate ^{3}H-labeled nucleosides into RNA (cells cannot take up nucleoside triphosphates from the medium). Once an RNA polymerase molecule has initiated transcription and moved away from the promoter, another RNA polymerase can follow suit. The synthesis of RNAs that are needed in large quantities, ribosomal RNAs, for example, are initiated as often as is sterically possible, about once per second (Fig. 29-12).

The error frequency in RNA synthesis, as estimated

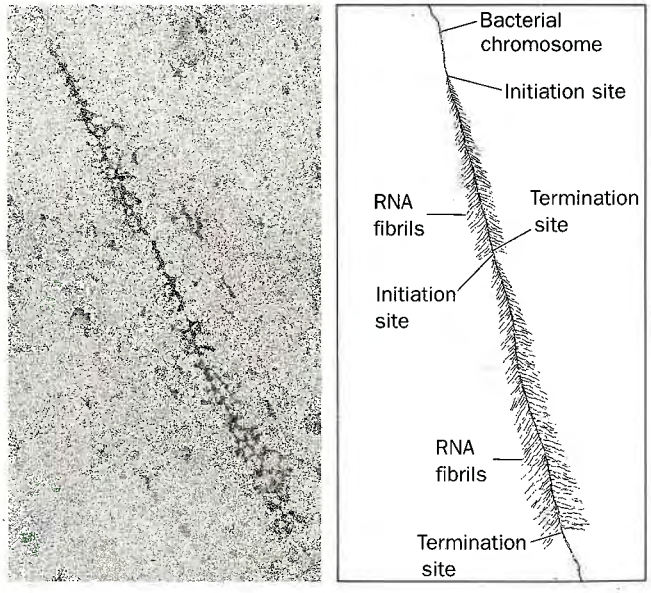
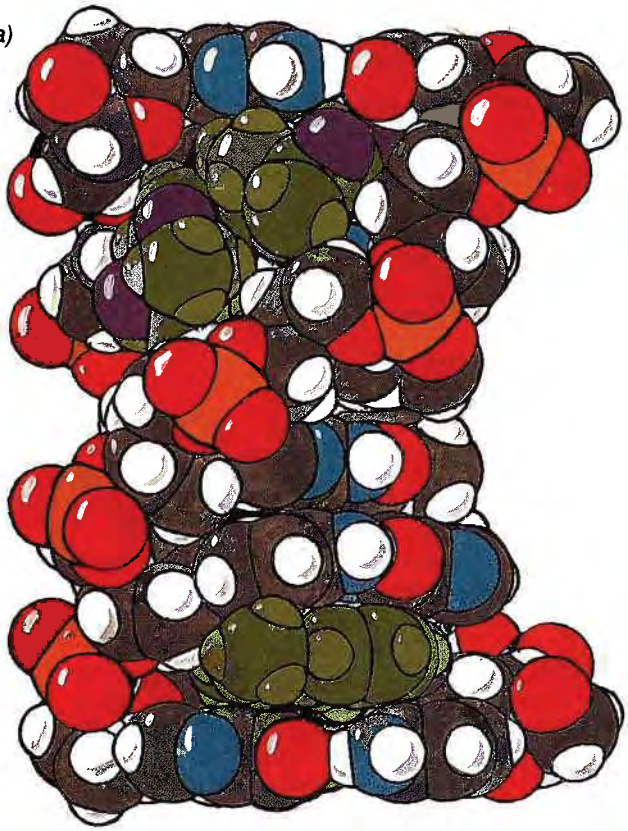


Figure 29-12
An electron micrograph and its interpretive drawing of two contiguous *E. coli* ribosomal genes undergoing transcription. The "arrowhead" structures result from the increasing lengths of the nascent RNA chains as the RNA polymerase molecules synthesizing them move from the initiation site on the DNA to the termination site. [Courtesy of Oscar L. Miller, Jr., University of Virginia.]

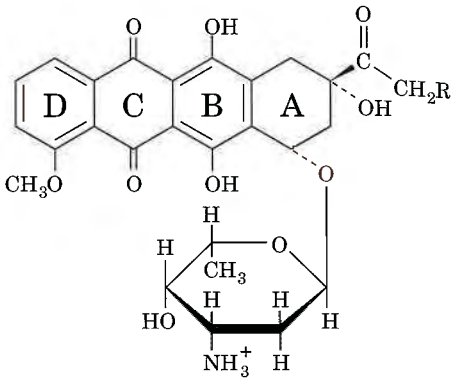
(a)



from the analysis of transcripts of simple templates such as poly[d(AT)]·poly[d(AT)], is one wrong base incorporated for every ~10^4 transcribed. This rate is tolerable because of the repeated transcription of most genes, because the genetic code contains numerous synonyms (Section 30-1E), and because amino acid substitutions in proteins are often functionally innocuous.

Intercalating Agents Inhibit Both RNA and DNA Polymerases

Daunomycin and the closely related **adriamycin,**



Daunomycin: R = H
Adriamycin: R = OH

(b)

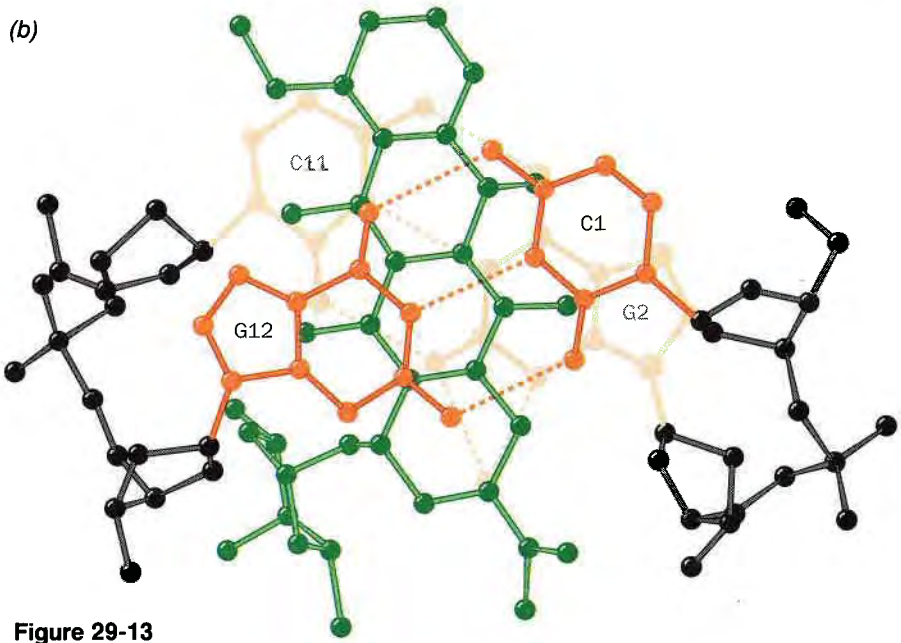


Figure 29-13
The structure of a complex of daunomycin with the self-complementary hexanucleotide d(CGTACG). Each double helical fragment binds two daunomycin molecules by intercalation between its G·C pairs to form a complex with twofold rotational symmetry. (*a*) A space-filling representation of the complex showing the upper daunomycin's amino sugar extending into the minor groove of the double helix and the edge of the lower daunomycin's intercalated ring. The daunomycin molecules are colored green with purple oxygen atoms. [After a drawing provided by Andrew Wang, University of Illinois.] (*b*) A view perpendicular to the bases indicating the stacking of the intercalated daunomycin ring system (*green*) with its surrounding G·C pairs. The C1·G12 base pair is closer to the viewer than the daunomycin ring system while the G2·C11 base pair is farther away. [After Wang, A. H.-J., Ughetto, G., Quigley, G. J., and Rich, A., *Biochemistry* **26,** 1155, 1157 (1987).]

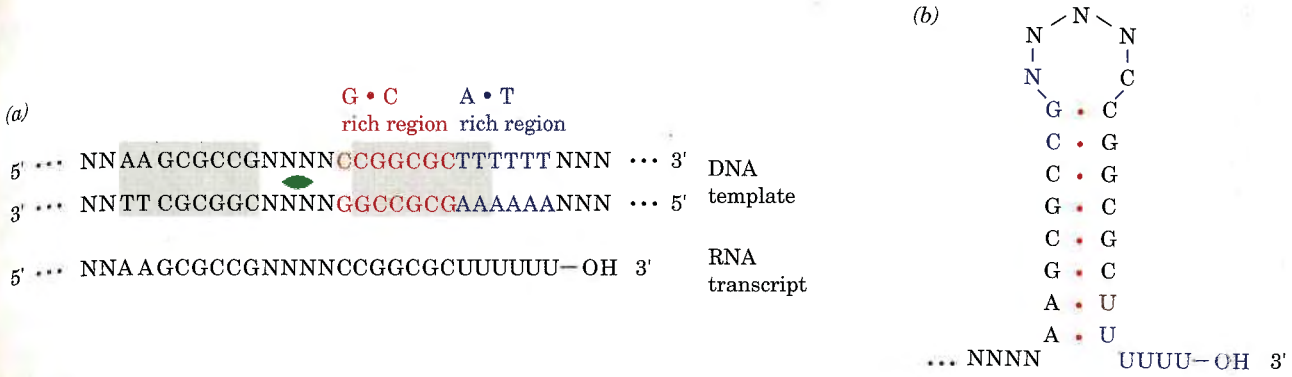
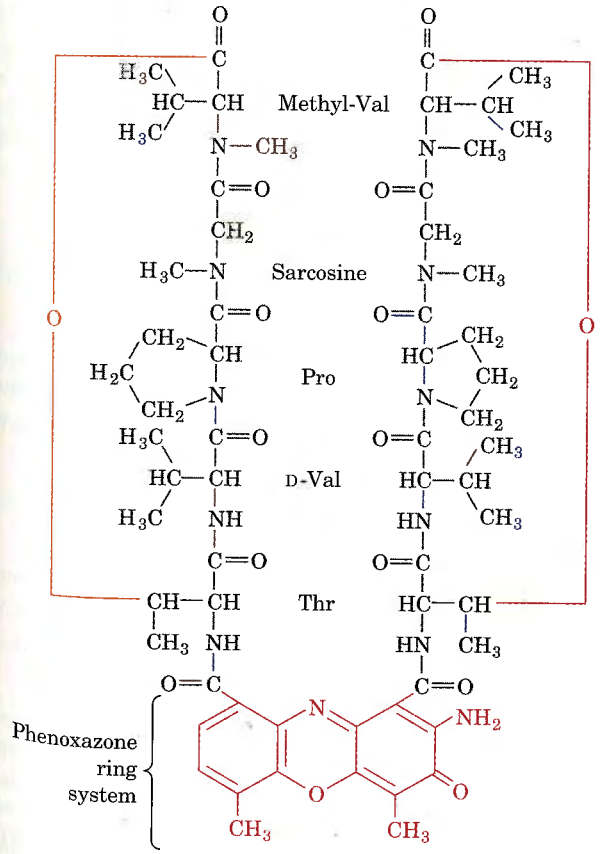


Figure 29-14
The base sequence of a hypothetical strong (efficient) terminator as deduced from the sequences of several transcripts. (*a*) The DNA sequence together with its corresponding RNA. The A·T-rich and G·C-rich sequences are shown in blue and red, respectively. The twofold symmetry axis (*lenticular symbol*) relates the flanking shaded segments that form an inverted repeat. (*b*) The RNA hairpin structure and poly(U) tail that triggers transcription termination. [After Pribnow, D., *in* Goldberger, R. F. (Ed.), *Biological Regulation and Development*, Vol. 1, *p.* 253, Plenum Press (1979).]

which are valuable chemotherapeutic agents in the treatment of certain human cancers, specifically bind to duplex DNA so as to inhibit both its transcription and its replication. These antibiotics presumably act by interfering with the passage of both RNA polymerase and DNA polymerase. The X-ray crystal structure of a complex of daunomycin with the self-complementary hexanucleotide d(CGTACG) reveals that daunomycin's planar aromatic ring system (rings B–D) is intercalated between the G·C pairs at both ends of the double helical fragment (Fig. 29-13). The nonplanar A ring extends into the minor groove where its side groups stabilize the complex through hydrogen bonding interactions with the DNA.

Actinomycin D,



Actinomycin D

an antibiotic produced by *Streptomyces antibioticus,* is also a potent inhibitor of nucleic acid synthesis. It acts by intercalating its **phenoxazone ring** between two successive G·C pairs of duplex DNA in a manner similar to daunomycin. Actinomycin's two identical cyclic pentapeptide groups, which have an unusual composition, stabilize this interaction through specific contacts with the double helix. Other intercalating agents, ethidium and proflavin (Section 28-4C), for example, also inhibit nucleic acid synthesis, presumably by similar mechanisms.

E. Chain Termination

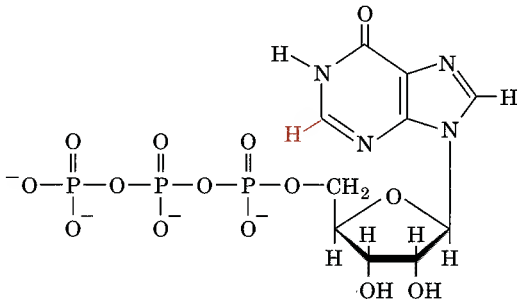
Electron micrographs such as Fig. 29-12 suggest that DNA contains specific sites at which transcription is terminated. The transcriptional termination sequences of several *E. coli* genes share two common features (Fig. 29-14*a*):

1. A series of 4 to 10 consecutive A·T's with the A's on the template strand. The transcribed RNA is terminated in or just past this sequence.
2. A G + C-rich region with a palindromic (twofold symmetric) sequence that immediately precedes the series of A·T's.

The RNA transcript of this region can therefore form a self-complementary "hairpin" structure that is terminated by several U residues (Fig. 29-14*b*).

The stability of a terminator's G + C-rich hairpin and the weak base pairing of its oligo(U) tail to template DNA appear to be important factors in ensuring proper chain

termination. Indeed, model studies have shown that oligo(dA·rU) forms a particularly unstable hybrid helix although oligo(dA·dT) forms a helix of normal stability. The formation of the G + C-rich hairpin causes RNA polymerase to pause for several seconds at the termination site. This, it has been proposed, induces a conformational change in the RNA polymerase, which permits the noncoding DNA strand to displace the weakly bound oligo(U) tail from the template strand thereby terminating transcription. Consistent with this notion is the observation that mutations that alter the strengths of these associations reduce the efficiency of chain termination and often eliminate it. Termination is similarly diminished when *in vitro* transcription is carried out with GTP replaced by **inosine triphosphate (ITP).**



Inosine triphosphate (ITP)

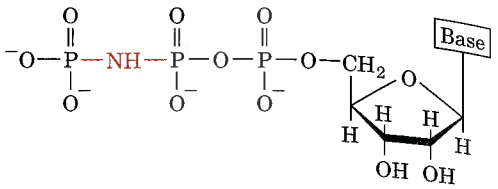
I·C pairs are weaker than those of G·C because the hypoxanthine base of I, which lacks the 2-amino group of G, can only make two hydrogen bonds to C thereby decreasing the hairpin's stability. UTP replacement by 5-bromo-UTP also diminishes chain termination because 5Br-U forms stronger base pairs with A than does U itself thus inhibiting the nascent RNA's displacement from the template DNA strand.

Termination Often Requires the Assistance of Rho Factor

The foregoing termination sequences induce the spontaneous termination of transcription. Other termination sites, however, lack any obvious similarities and are unable to form strong hairpins; *they require the participation of a protein known as* ***rho factor*** *to terminate transcription.* The existence of rho factor was suggested by the observation that *in vivo* transcripts are often shorter than the corresponding *in vitro* transcripts. Rho factor, a hexamer of identical 419-residue subunits, enhances the termination efficiency of spontaneously terminating transcripts as well as inducing the termination of nonspontaneously terminating transcripts.

Several key observations have led to a model of rho-dependent termination:

1. Rho factor is an enzyme that catalyzes the unwinding of RNA–DNA and RNA–RNA double helices. This process is powered by the hydrolysis of nucleoside triphosphates (NTPs) to nucleoside diphosphates + P_i with little preference for the identity of the base. NTPase activity is required for rho-dependent termination as is demonstrated by its *in vitro* inhibition when the NTPs are replaced by their β,γ-imido analogs,



β, γ-Imido nucleoside triphosphate

substances that are RNA polymerase substrates but cannot be hydrolyzed by rho factor.

2. Genetic manipulations indicate that rho-dependent termination requires the presence of a specific recognition sequence upstream of the termination site. The recognition sequence must be on the nascent RNA rather than the DNA as is demonstrated by rho's inability to terminate transcription in the presence of pancreatic RNase A. The essential features of this termination site have not been fully elucidated; the construction of synthetic termination sites indicate that it consists of 80 to 100 nucleotides which lack a stable secondary structure and probably contain multiple C-rich regions.

These observations suggest that rho factor attaches to nascent RNA at its recognition sequence and then migrates along the RNA in the 5′ → 3′ direction until it encounters an RNA polymerase paused at the termination site (without the pause, rho might not be able to overtake the RNA polymerase). There, rho unwinds the RNA–DNA duplex forming the transcription bubble thereby releasing the RNA transcript.

F. Eukaryotic RNA Polymerases

Eukaryotic nuclei contain three distinct types of RNA polymerases that differ in the RNAs they synthesize:

1. **RNA polymerase I,** which is located in the nucleoli (dense granular bodies in the nuclei that contain the ribosomal genes; Section 29-4B), synthesizes precursors of most ribosomal RNAs.
2. **RNA polymerase II,** which occurs in the nucleoplasm, synthesizes mRNA precursors.
3. **RNA polymerase III,** which also occurs in the nucleoplasm, synthesizes the precursors of 5S ribosomal RNA, the tRNAs, and a variety of other small nuclear and cytosolic RNAs.

In addition to these nuclear enzymes (which are also known as **RNA polymerases A, B,** and **C**), eukaryotic cells contain separate mitochondrial and chloroplast RNA polymerases.

Eukaryotic RNA polymerases, whose molecular masses vary between 500 and 700 kD, are characterized

by subunit compositions of Byzantine complexity. Each type of enzyme contains two nonidentical "large" (>100 kD) subunits and an array of up to 12 different "small" (<50 kD) subunits. Some of the small subunits occur in 2 or all 3 of the nuclear RNA polymerases. As yet, little is known about their functions or interactions although, intriguingly, the largest subunits of yeast RNA polymerases II and III exhibit extensive homology to each other and to the largest (β') subunit of *E. coli* RNA polymerase.

The RNA Polymerase I Promoter Consists of Nested Control Regions

The RNA polymerase I promoter has been identified by determining the transcription rates of a series of mutant rRNA genes from *Xenopus laevis* (an African clawed frog) with increasingly longer deletions from either their 5′ or their 3′ ends. (It is not possible to deduce the RNA polymerase I promoter from the sequence homologies common to the genes it transcribes because, as we shall see in Section 29-4B, there is only one type of rRNA gene.) Optimal rRNA expression requires the presence of the rRNA gene segment extending from −142 to +6. The minimal base sequence required for accurate initiation, however, extends between nucleotides −7 and +6. It therefore appears that this latter promoter element acts to guide RNA polymerase I to its proper initiation site, whereas the rest of the promoter functions to bind proteins known as **transcription factors** (see below).

RNA Polymerase II Promoters Are Complex and Diverse

The promoters recognized by RNA polymerase II, which are considerably longer and more diverse than those of prokaryotic genes, have, as yet, been only superficially described. The structural genes expressed in all tissues, the so-called "housekeeping" genes, which are thought to be constituitively transcribed, have one or more copies of the sequence GGGCGG or its complement (the **GC box**) located upstream from their transcription start sites. The analysis of deletion and point mutations in eukaryotic viruses such as SV40 indicates that GC boxes function analogously to prokaryotic promoters. On the other hand, structural genes that are selectively expressed in one or a few types of cells often lack these GC-rich sequences. Rather, *they contain a conserved AT-rich sequence located 25 to 30 bp upstream from their transcription start sites (Fig. 29-15).* Note that this so-called **TATA** or **Goldberg-Hogness box** (after Michael Goldberg and David Hogness who deduced its existence in 1978) resembles the prokaryotic Pribnow box (TATAAT) although they differ in their locations relative to the transcription start site (−27 vs −10). The functions of these two promoter elements are not strictly analogous, however, since the deletion of the TATA box does not necessarily eliminate transcription. Rather, TATA box deletion or mutation generates heterogeneities in the transcriptional start site thereby indicating that the TATA box participates in selecting this site.

The gene region extending between about −50 and −110 also contains promoter elements. For instance, many eukaryotic structural genes, including those encoding the various globins, have a conserved sequence of consensus CCAAT (the **CCAAT box**) located between about −70 and −80 whose alteration greatly reduces the gene's transcription rate. Globin genes have, in addition, a conserved **CACCC box** upstream from the CCAAT box that has also been implicated in transcriptional initiation. Evidently, the promoter sequences upstream of the TATA box form the initial DNA-binding sites for RNA polymerase II and the other proteins involved in transcriptional initiation (see below).

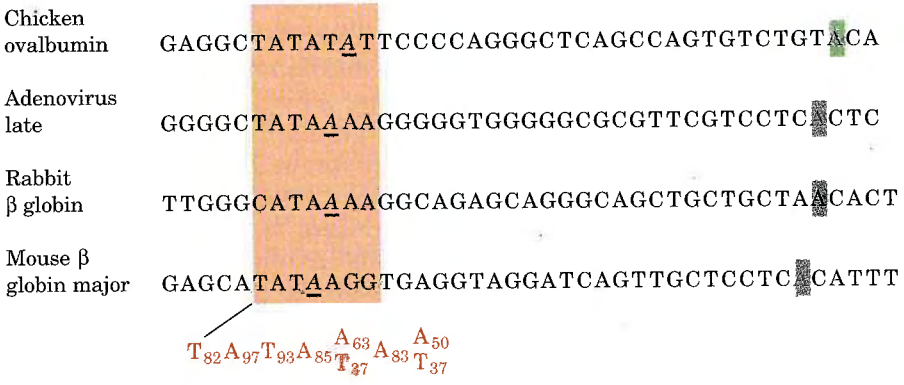


Figure 29-15
The promoter sequences of selected eukaryotic structural genes. The homologous segment, the TATA box, is shaded in red with the base at position −27 underlined and the initial nucleotide to be transcribed (+1) shaded in green. The bottom row indicates the consensus sequence of several such promoters with the subscripts indicating the percent occurrence of the corresponding base. [After Gannon, F., O'Hare, K., Perrin, F., Le Pennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., and Chambon, P., *Nature* **278,** 433 (1978).)

Enhancers Are Transcriptional Activators That Can Have Variable Positions and Orientations

Perhaps the most suprising aspect of eukaryotic transcriptional control elements is that some of them need not have fixed positions and orientations relative to their corresponding transcribed sequences. For example, the SV40 genome, in which such elements were first discovered, contains two repeated sequences of 72 bp each that are located upstream from the promoter for early gene expression. Transcription is unaffected if one of these repeats is deleted but is nearly eliminated when both are absent. The analysis of a series of SV40 mutants containing only one of these repeats demonstrated that its ability to stimulate transcription from its corresponding promoter is all but independent of its position and orientation. Indeed, transcription is unimpaired when this segment is several thousand base pairs upstream or downstream from the transcription start site. Gene segments with such properties are named **enhancers** to indicate that they differ from promoters, with which they must be associated in order to trigger site-specific and strand-specific transcription initiation (although the characterization of numerous promoters and enhancers indicates that their functional properties are similar). Enhancers occur both in eukaryotic viruses and cellular genes.

Enhancers are required for the full activities of their cognate promoters. But how do they act? Two not mutually exclusive possibilities are given the most credence:

1. Enhancers are "entry points" on DNA for RNA polymerase II, perhaps through a lack of binding affinity for the histones that normally coat eukaryotic DNA, so as to (as seems likely) block RNA polymerase II binding (Section 33-1A). Alternatively, enhancers may alter DNA's local conformation in a way that favors RNA polymerase II binding. In fact, some enhancers contain a segment of alternating purines and pyrimidines which, we have seen, is just the type of sequence most likely to form Z-DNA (Section 28-2B).
2. Enhancers are recognized by specific proteins called **transcription factors** that stimulate RNA polymerase II to bind to a nearby promoter.

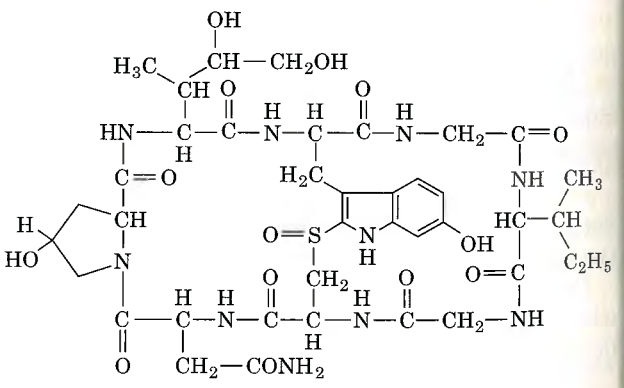
All cellular enhancers that have yet been identified are associated with genes that are selectively expressed in specific tissues. It therefore seems, as we discuss in Section 33-3B, that *enhancers mediate much of the selective gene expression in eukaryotes.*

RNA Polymerase III Promoters Can Be Located Downstream from Their Transcription Start Sites

The promoters of genes transcribed by RNA polymerase III can be located entirely within the genes' transcribed regions. Donald Brown established this through the construction of a series of deletion mutants of a *Xenopus borealis* 5S RNA gene. Deletions of base sequences that start from outside one or the other end of the transcribed portion of the 5S gene only prevent transcription if they extend into the segment between nucleotides +40 and +80. Indeed, a fragment of the 5S gene consisting of only nucleotides 41 to 87, when cloned in a bacterial plasmid, is sufficient to direct specific initiation by RNA polymerase III at an upstream site. This is because, as was subsequently demonstrated, the sequence contains the binding site for a transcription factor that stimulates the upstream binding of RNA polymerase III. Further studies have shown, however, that the promoters of other RNA polymerase III-transcribed genes may lie partially or even entirely upstream of their start sites.

Amatoxins Specifically Inhibit RNA Polymerases II and III

The poisonous mushroom *Amanita phalloides* (death cap), which is responsible for the majority of fatal mushroom poisonings, contains several types of toxic substances including a series of unusual bicyclic octapeptides known as **amatoxins. α-Amanitin,**



α-Amanitin

which is representative of the amatoxins, forms a tight 1 : 1 complex with RNA polymerase II ($K = 10^{-8}M$) and a looser one with RNA polymerase III ($K = 10^{-6}M$), so as to specifically block their elongation steps. α-Amanitin is therefore a useful tool for mechanistic studies of these enzymes. RNA polymerase I as well as mitochondrial, chloroplast, and prokaryotic RNA polymerases are insensitive to α-amanitin.

Despite the amatoxins' high toxicity (5–6 mg, which occur in ~40 g of fresh mushrooms, are sufficient to kill a human adult), they act slowly. Death, usually from liver dysfunction, occurs no earlier than several days after mushroom ingestion (and after recovery from the effects of other mushroom toxins). This, in part, reflects the slow turnover rate of eukaryotic mRNAs and proteins.

3. CONTROL OF TRANSCRIPTION IN PROKARYOTES

Prokaryotes respond to sudden environmental changes, such as the influx of nutrients, by inducing the synthesis of the appropriate proteins. This process takes only minutes because transcription and translation in prokaryotes are closely coupled: *Ribosomes commence translation near the 5′ end of a nascent mRNA soon after it is extruded from RNA polymerase (Fig. 29-16).* Moreover, *most prokaryotic mRNAs are enzymatically degraded within 1 to 3 min of their synthesis,* thereby eliminating the wasteful synthesis of unneeded proteins after a change in conditions (protein degradation is discussed in Section 30-6). In fact, the 5′ ends of some mRNAs are degraded before their 3′ ends have been synthesized.

In contrast, the induction of new proteins in eukaryotic cells frequently takes hours or days because transcription takes place in the nucleus and the resulting mRNAs must be transported to the cytoplasm where translation occurs. However, eukaryotic cells, particularly those of multicellular organisms, have relatively stable environments; changes in their transcriptional patterns usually occur only during cell differentiation.

In this section we examine some of the ways in which prokaryotic gene expression is regulated through transcriptional control. Eukaryotes, being vastly more complex creatures than are prokaryotes, have a correspondingly more complicated transcriptional control system whose general outlines are just coming into focus. We therefore defer discussion of eukaryotic transcriptional control until Section 33-3 where it can be considered in light of what we know about the structure and organization of the eukaryotic chromosome.

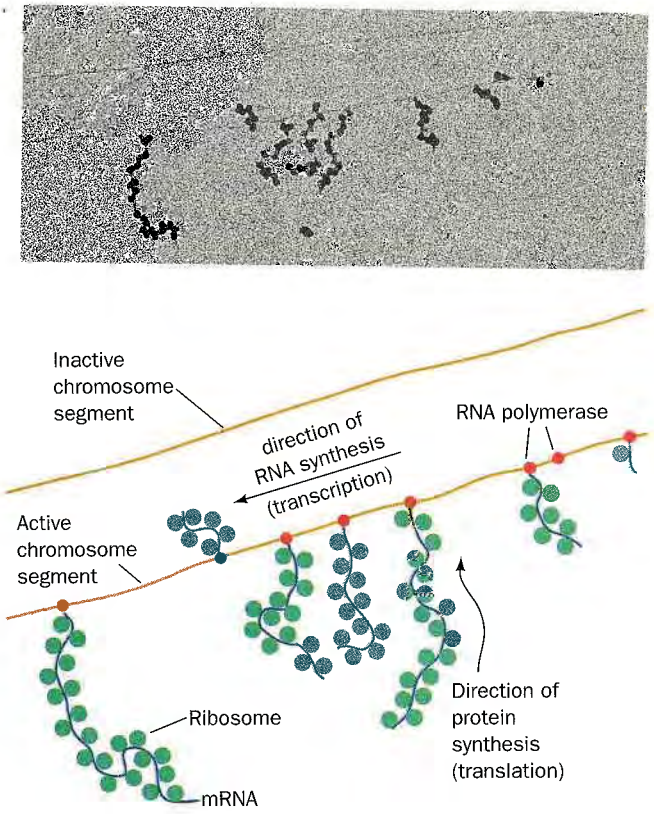


Figure 29-16
An electron micrograph and its interpretive drawing showing the simultaneous transcription and translation of an *E. coli* gene. RNA polymerase molecules are transcribing the DNA from right to left while ribosomes are translating the nascent RNAs (mostly from bottom to top). [Courtesy of Oscar L. Miller, Jr., University of Virginia.]

A. Promoters

In the presence of high concentrations of inducer, the *lac* operon is rapidly transcribed. In contrast, the *lacI* gene is transcribed at such a low rate that a typical *E. coli* cell contains <10 molecules of the *lac* repressor. Yet, the *I* gene has no repressor. Rather, it has such an inefficient promoter that it is transcribed an average of about once per bacterial generation. *Genes that are transcribed at high rates have efficient promoters.* In general, the more efficient a promoter, the more closely its sequence resembles that of the corresponding consensus sequence.

Gene Expression in Certain Phages Is Controlled by a Succession of σ Factors

The processes of development and differentiation involve the temporally ordered expression of sets of genes according to genetically specified programs. Phage infections are among the simplest examples of developmental processes. Typically, only a subset of the phage genome, often referred to as *early* genes, are expressed in the host immediately after phage infection. As time passes, *middle* genes start to be expressed and the *early* genes as well as the bacterial genes are turned off. In the final stages of phage infection, the *middle* genes give way to the *late* genes. Of course some phage types express more than three sets of genes and some genes may be expressed in more than one stage of an infection.

One way in which families of genes are sequentially expressed is through "cascades" of σ factors. In the infection of *Bacillus subtilus* by **bacteriophage SP01,** for example, the *early* gene promoters are recognized by the bacterial RNA polymerase holoenzyme. Among these *early* genes is gene 28 whose gene product is a new σ subunit, designated σ^{sp28}, that displaces the bacterial σ subunit from the core enzyme. This reconstituted holoenzyme recognizes only the phage *middle* gene promoters, which all have similar -35 and -10 (Pribnow box) regions, but bear little resemblance to the corresponding regions of bacterial and phage *early* genes.

The *early* genes therefore become inactive once their corresponding mRNAs have been degraded. The phage *middle* genes include genes 33 and 34, which together specify yet another σ factor, $\sigma^{gp33/34}$ which, in turn, permits the transcription of only *late* phage genes.

Several bacteria, including *E. coli* and *B. subtilus*, likewise have several different σ factors. These are not utilized in a sequential manner. Rather, those that differ from the predominant or primary σ factor control the transcription of coordinately expressed groups of special purpose genes whose promoters are quite different from those recognized by the primary σ factor.

B. *lac* Repressor

In 1966, Beno Müller-Hill and Walter Gilbert isolated *lac* repressor on the basis of its ability to bind ^{14}C-labeled IPTG and demonstrated that it is a protein. This was an exceedingly difficult task because *lac* repressor comprises only ~0.002% of the protein in wild-type *E. coli*. Now, however, *lac* repressor is available in quantity through the application of molecular cloning techniques (Section 28-8D).

lac Repressor Finds Its Operator by Sliding Along DNA

The *lac* repressor is a tetramer of identical 360-residue subunits arranged with three mutually perpendicular twofold axes (D_2 symmetry; Section 7-5B). Each subunit is capable of binding one IPTG molecule with a dissociation constant of $K = 10^{-6}M$. In the absence of inducer, the repressor tetramer nonspecifically binds duplex DNA with a dissociation constant of $K \approx 10^{-4}M$. However, it specifically binds to the *lac* operator with far greater affinity: $K \approx 10^{-13}M$. Limited proteolysis of *lac* repressor with trypsin splits a 58-residue N-terminal peptide from each subunit. The remaining "core" tetramer binds IPTG but is unable to bind DNA. Apparently the DNA and inducer binding regions of each subunit occupy separate domains.

The observed rate constant for the binding of *lac* repressor to *lac* operator is $k_f \approx 10^{10}M^{-1}\,s^{-1}$. This "on" rate is much greater than that calculated for the diffusion-controlled process in solution: $k_f = 10^{7}M^{-1}\,s^{-1}$ for molecules the size of *lac* repressor. Since it is impossible for a reaction to proceed faster than its diffusion-controlled rate, the *lac* repressor must not encounter operator from solution in a random three-dimensional search. Rather, *it appears that lac repressor finds operator by nonspecifically binding to DNA and diffusing along it in a far more efficient one-dimensional search.*

lac Operator Has a Nearly Palindromic Sequence

The availability of large quantities of *lac* repressor made it possible to characterize the *lac* operator. *E. coli* DNA that had been sonicated to small fragments was mixed with *lac* repressor and passed through a nitrocellulose filter. Protein, with or without bound DNA, sticks to nitrocellulose whereas duplex DNA, by itself, does not. The DNA was released from the filter-bound protein by washing it with IPTG solution, recombined with *lac* repressor, and the resulting complex treated with DNase I. The DNA fragment that *lac* repressor protects from nuclease degradation consists of a run of 26 bp that is embedded in a nearly twofold symmetric sequence of 35 bp (Fig. 29-17; *top*). *Such palindromic symmetry is a common feature of DNAs that are specifically bound by proteins;* recall that restriction endonuclease recognition sites are also palindromic (Section 28-6A).

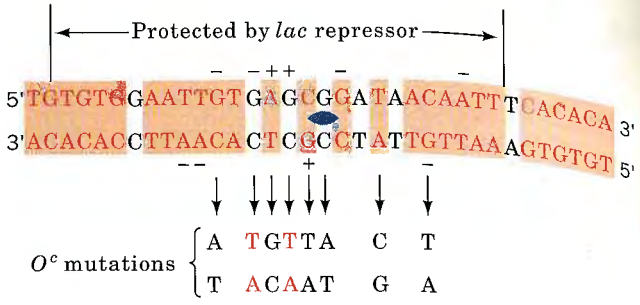


Figure 29-17
The base sequence of the *lac* operator. The symmetry related regions (*red*), comprise 28 of its 35 bp. A "+" denotes positions at which repressor binding enhances methylation by dimethyl sulfate [which methylates G at N(7) and A at N(3)] and a "−" indicates where this footprinting reaction is inhibited. The bottom row indicates the positions and identities of different point mutations that prevent *lac* repressor binding (O^c mutants). Those in color increase the operator's symmetry. [After Sobell, H.M., *in* Goldberger, R. F. (Ed.), *Biological Regulation and Development*, Vol. 1, *p.* 193, Plenum Press (1979).]

It has been suggested that the *lac* operator's symmetry matches that of its repressor; that is, operator binds to repressor in a twofold symmetric cleft between two subunits much like *Eco*RI restriction endonuclease binds to its recognition site (Section 28-6A). Methylation protection experiments, however, do not support this contention. There is an asymmetric pattern of differences between free and repressor-bound operator in the susceptibility of its bases to reaction with DMS (Fig. 29-17). Furthermore, point mutations in the operator that render it operator-constitutive (O^c), and which invariably weaken the binding of repressor to operator, may increase as well as decrease the operator's twofold symmetry (Fig. 29-17).

lac Repressor Prevents RNA Polymerase from Forming a Productive Initiation Complex

Operator occupies positions −7 through +28 of the *lac* operon relative to the transcription start site (Fig. 29-18). Nuclease protection studies, it will be recalled, indicate that, in the initiation complex, RNA polymerase tightly binds to the DNA between positions −20 and

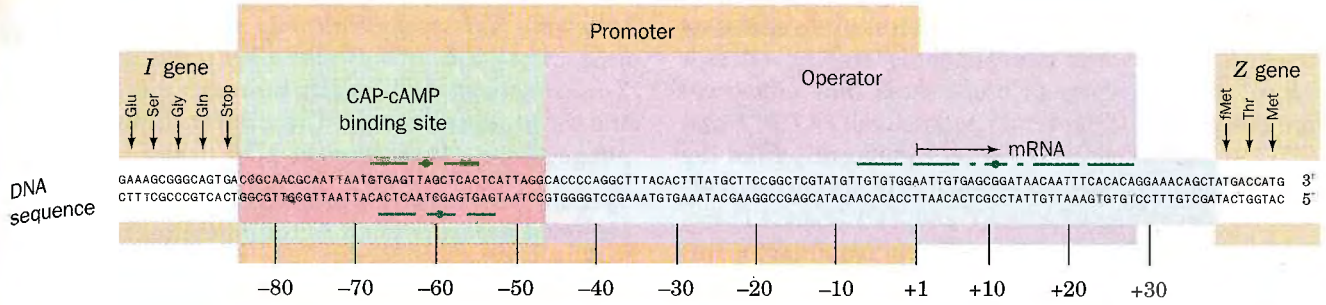


Figure 29-18
The nucleotide sequence of the *E. coli lac* promoter–operator region extending from the C-terminal region of *lacI* (*left*) to the N-terminal region of *lacZ* (*right*). The palindromic sequences of the operator and the CAP-binding site (Section 29-3C) are overscored or underscored. [After Dickson, R. C., Abelson, J., Barnes, W. M., and Reznikoff, W. A., *Science* **187,** 32 (1975).]

+20 (Section 29-2B). Thus, *the lac operator and promoter sites overlap.* This suggests that repressor binding and RNA polymerase binding are mutually exclusive. However, both proteins simultaneously bind to the *lac* operon, at least *in vitro,* to form a transcriptionally inactive complex. Evidently, operator-bound *lac* repressor prevents RNA polymerase from forming a productive initiation complex although how it does so is unknown.

C. Catabolite Repression: An Example of Gene Activation

Glucose is E. coli's metabolite of choice; the availability of adequate amounts of glucose prevents the full expression of genes specifying proteins involved in the fermentation of numerous other catabolites, including lactose (Fig. 29-19), arabinose and galactose, even when they are present in high concentrations. This phenomenon, which is known as **catabolite repression,** prevents the wasteful duplication of energy-producing enzyme systems.

cAMP Signals the Lack of Glucose

The first indication of the mechanism of catabolite repression was the observation that, in *E. coli,* the level of cAMP, which was known to be a second messenger in animal cells (Section 17-3E), is greatly diminished in the presence of glucose. This observation led to the finding that the addition of cAMP to *E. coli* cultures overcame catabolite repression by glucose. Recall that, in *E. coli,* adenylate cyclase is activated by a phosphorylated enzyme (E III_g), which is dephosphorylated upon the transport of glucose across the cell membrane (Section 18-3D). *The presence of glucose, therefore, normally lowers the cAMP level in E. coli.*

CAP–cAMP Complex Stimulates the Transcription of Catabolite Repressed Operons

Certain *E. coli* mutants, in which the absence of glucose does not relieve catabolite repression, are missing a cAMP-binding protein that is synonymously named **catabolite gene activator protein (CAP)** or **cAMP receptor protein (CRP).** CAP is a dimeric protein of identical 210-residue subunits that undergoes a large conformational change upon binding cAMP. Its function was elucidated by Ira Pastan who showed that *CAP–cAMP complex, but not CAP itself, binds to the lac operon (among others) and stimulates transcription from its otherwise low efficiency promoter in the absence of repressor.* CAP is therefore a **positive regulator** (turns on transcription), in contrast to *lac* repressor, which is a **negative regulator** (turns off transcription).

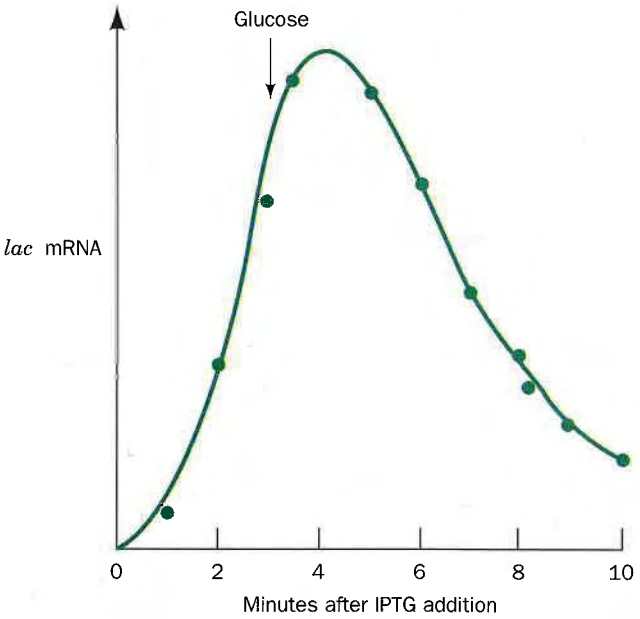


Figure 29-19
The kinetics of *lac* operon mRNA synthesis following its induction with IPTG, and of its degradation after glucose addition. *E. coli* were grown on a medium containing glycerol as their only carbon-energy source and ^{3}H-labeled uridine. IPTG was added to the medium at the beginning of the experiment to induce the synthesis of the *lac* enzymes. After 3 min, glucose was added to stop the synthesis. The amount of ^{3}H-labeled *lac* RNA was determined by hybridization with DNA containing the *lacZ* and *lacY* genes. [After Adesnik, M. and Levinthal, C., *Cold Spring Harbor Symp. Quant. Biol.* **35,** 457 (1970).]

Why is CAP–cAMP complex necessary to stimulate the transcription of its target operons? And how does it do so? The *lac* repressor has a weak (low efficiency) promoter; its −10 and −35 sequences (TATGTT and CTTTACACT; Fig. 29-18) differ significantly from the corresponding consensus sequences of strong (high efficiency) promoters (TATAAT and TCTTGACAT; Fig. 29-9). Such weak promoters evidently require some sort of help for efficient transcriptional initiation. There are two plausible (and not mutually exclusive) ways that CAP–cAMP could provide such help:

1. CAP–cAMP may stimulate transcriptional initiation through direct interaction with RNA polymerase. This hypothesis is supported by the observation that the *lac* operon fragment that CAP–cAMP complex protects from DNase I digestion contains two overlapping pseudopalindromic sequences that are located in the *lac* promoter's upstream segment (Fig. 29-18).

2. The binding of CAP–cAMP complex to promoter may conformationally alter this DNA. For example, it may induce the formation of the open RNA polymerase initiation complex (Section 29-2B). This idea is corroborated by the observation that negative supercoiling, which tends to unwind B-DNA, promotes the *in vitro* CAP stimulation of *lac* operon transcription. The binding of CAP to promoter, however, does not alter DNA's superhelicity so that CAP does not, by itself, unwind promoter. Another possibility is that CAP binding bends the DNA so as to store elastic energy for subsequent use in transcription. Indeed, the anomalously low polyacrylamide gel electrophoretic mobility of CAP in complex with its ~30 bp target sequence indicates that CAP binding induces at least a 90° bend in this DNA segment.

Many Prokaryotic Repressors and Activators Bind Their Operators in a Similar Fashion

Since genetic expression is controlled by proteins such as CAP and *lac* repressor, an important issue in the study of gene regulation is how do these proteins interact with DNA. The X-ray crystal structure of CAP, determined by Thomas Steitz, reveals that each monomer of this dimeric protein consists of two flexibly linked domains (Fig. 29-20*a*). The N-terminal domains bind cAMP and form the intersubunit contacts. The C-terminal domains form the DNA-binding site as is demonstrated by the observation that their excision by limited proteolysis results in a dimeric cAMP-binding protein that does not bind DNA.

The CAP dimer's two symmetrically disposed F helices protrude from the protein surface in such a way that, according to model building studies, they fit into successive major grooves of B-DNA (Fig. 29-20*b*). *CAP's E and F helices form a* ***helix–turn–helix*** *supersecondary structure that conformationally resembles analogous helix–turn–helix motifs in the other repressors of known X-ray structures:* the *E. coli* **trp repressor** (Section 29-3E) and the **cI repressors** and **Cro proteins** from **bacteriophages** λ and **434** (Section 32-3D).

Specific Protein–DNA Interactions Arise from Mutual Conformational Accommodations

Model building, such as that indicated in Fig. 29-20*b*, and, more importantly, the direct visualization of protein–DNA complexes (see below), indicates that *these DNA-binding proteins associate with their target base pairs mainly via the side chains extending from the second helix of the helix–turn–helix motif, the so-called* ***"recognition" helix*** (helix F in CAP, E in *trp* repressor, and $\alpha 3$ in the phage proteins). Indeed, replacing the outward-facing residues of the 434 repressor's "recognition" helix with the corresponding residues of the related **bacteriophage P22** (using the gentic engineering techniques described in Section 28-8) yields a hybrid repressor that binds to P22 operators but not to those of 434. Moreover, the ~20-residue helix–turn–helix motifs in all these proteins have amino acid sequences that are similar to each other and to polypeptide segments in numerous other prokaryotic DNA-binding proteins, including *lac* repressor. Evidently, *these proteins are evolutionarily related and bind their target DNAs in a similar manner* (but in a way that differs from that of *Eco*RI restriction endonuclease; Section 28-6A).

How does the "recognition" helix recognize its target sequence? Each base pair presents a different and presumably readily differentiated constellation of hydrogen bonding groups in DNA's major groove (see Fig. 28-6). It has therefore been proposed that there is a simple correspondence, analogous to Watson–Crick base pairing, between the amino acid residues of the "recognition" helix and the bases they contact in forming sequence-specific associations. The above X-ray structures, however, indicate this proposal to be incorrect. Rather, base sequence recognition arises from complex structural interactions. For instance:

1. The X-ray structures of 434 repressor and 434 Cro protein in complex with the identical 20 bp target DNA (434 phage expression is regulated through the differential binding of these proteins to the same DNA segments; Section 32-3D) were both determined by Stephen Harrison. Both dimeric proteins, as predicted for CAP (Fig. 29-20*b*), associate with the DNA in a twofold symmetric manner with their "recognition" helices bound in successive turns of the DNA's major groove (Figs. 29-21 and 29-22). In both complexes, the protein closely conforms to the DNA surface and interacts with its paired bases and sugar–phosphate chains through elaborate systems of hydrogen bonds, salt bridges, and van der Waals contacts. Nevertheless, the detailed geometries of

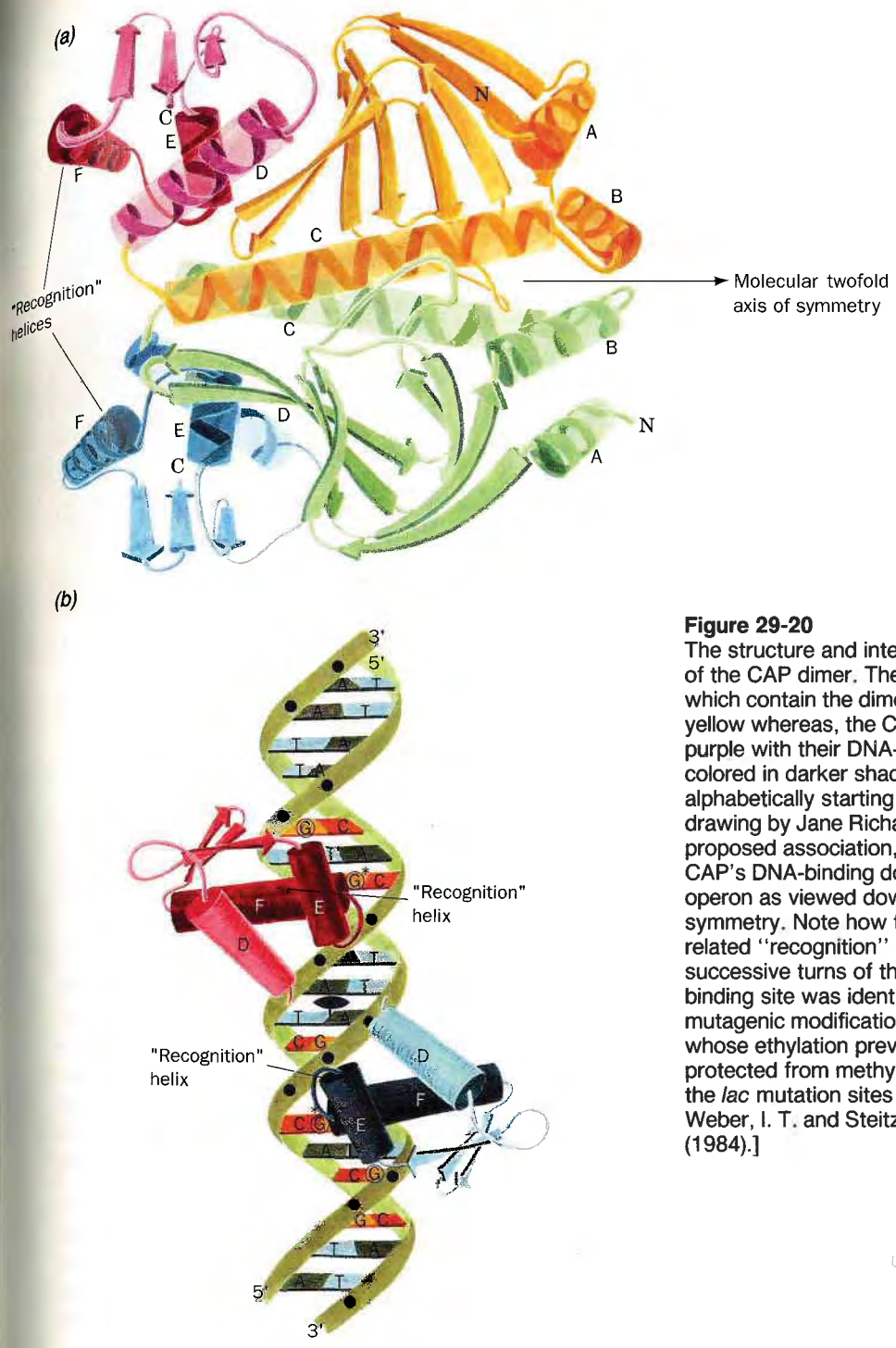


Figure 29-20
The structure and interactions of CAP. (*a*) A ribbon diagram of the CAP dimer. The cAMP-binding N-terminal domains, which contain the dimer contacts, are colored green and yellow whereas, the C-terminal domains are colored blue and purple with their DNA-binding helix–turn–helix domains colored in darker shades. The helices are labeled alphabetically starting from the N-terminus. [Based on a drawing by Jane Richardson, Duke University.] (*b*) The proposed association, based on model building, between CAP's DNA-binding domains and their binding site on the *lac* operon as viewed down the protein's twofold axis of symmetry. Note how the dimeric protein's two symmetry related "recognition" helices are spaced to fit into successive turns of the DNA's major groove. The DNA-binding site was identified through chemical, enzymatic, and mutagenic modification studies. Dots mark the phosphates whose ethylation prevents CAP binding, circled G's are protected from methylation when CAP binds, and * indicates the *lac* mutation sites that decrease CAP affinity. [After Weber, I. T. and Steitz, T. A., *Proc. Natl. Acad. Sci.* **81,** 3975 (1984).]

these associations are significantly different. In the repressor–DNA complex (Fig. 29-21), the DNA bends around the protein in an arc of radius ~65 Å so as to compress the minor groove by ~2.5 Å near its center (between the two protein monomers) and widen it by ~2.5 Å towards its ends [the phosphate–phosphate distance across the minor groove in canonical (ideal) B-DNA is 11.5 Å]. In contrast, the DNA in complex with Cro (Fig. 29-22), although also bent, is nearly straight at its center and has a less compressed minor groove (compare Figs. 29-21*a* and 29-22*a*). This explains why the simultaneous replacement of three residues in the repressor "recognition" helix with those occurring in Cro does not cause the resulting hybrid protein to bind DNA with Cro-like affinity: *The different conformations of the DNA in the*

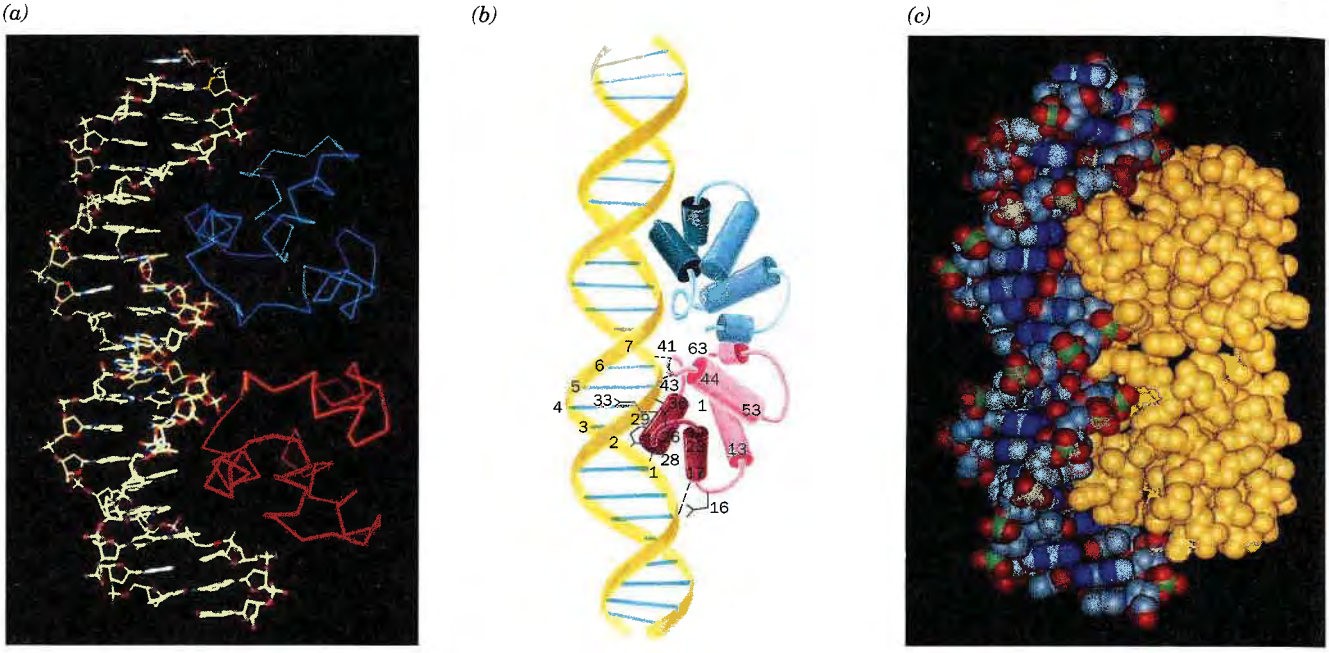


Figure 29-21
The X-ray structure of 434 phage repressor (actually only the repressor's 69-residue N-terminal domain) in complex with a 20 bp fragment of its target sequence [one strand of which has the sequence d(TATACAAGAAAGTTTGTACT)] as viewed perpendicularly to the complex's twofold axis of symmetry. (*a*) A skeletal model with the DNA on the left and with the protein's two identical subunits (C_α backbone only) shown in red and blue. Only the first 63 residues of the protein are visible. [Courtesy of Aneel Aggarwal, John Anderson, and Stephen Harrison, Harvard University.] (*b*) An interpretive drawing showing how the helix–turn–helix motif (*darker shading*) interacts with the DNA. In the lower protein monomer, the side chains important for interaction with the DNA are indicated and the numbers of the first and last residues of the helix are given. Note how the dimer's two "recognition" helices bind in successive major grooves of the DNA. [After Anderson, J. E., Ptashne, M., and Harrison, S. C., *Nature* **326,** 847 (1987).] (*c*) A space-filling model corresponding to Part (*a*). All of the protein's non-H atoms are drawn in yellow. [Courtesy of Aneel Aggarwal, John Anderson, and Stephen Harrison, Harvard University.]

repressor and Cro complexes prevents any particular side chain from interacting identically with the DNA in the two complexes.

2. Paul Sigler determined the X-ray structure of *E. coli trp* repressor in complex with an 18 bp palindromic DNA that closely resembles *trp* operator (Section 29-3E). The dimeric protein contacts the relatively straight DNA via 24 direct and 6 solvent-mediated (water bridged) hydrogen bonds to the DNA's phosphate groups (Fig. 29-23). Astoundingly, however, *there are no direct hydrogen bonds or nonpolar contacts that can explain the repressor's specificity for its operator* (the few such contacts in the structure are with bases that are tolerent to mutation). Evidently, *trp* repressor recognizes its operator via "indirect readout": The operator's sequence permits the DNA to assume a conformation that makes favorable contacts with the repressor. Model building indicates that canonical B-DNA can only make a small fraction of the contacts that operator makes to repressor. Other DNA sequences could conceivably assume repressor-bound operator's conformation but at too high an energy cost to form a stable complex with repressor (*trp* repressor's measured 10^4-fold preference for its operator over other DNAs implies an $\sim 23\ kJ \cdot mol^{-1}$ difference in their binding free energies). Thus, *specificity arises here from sequence-specific conformational variations in DNA rather than from sequence-specific hydrogen bonding interactions between DNA and protein.*

It therefore appears that *there are no simple rules governing how particular amino acid residues interact with bases. Rather, sequence specificity results from an ensemble of mutually favorable interactions between a protein and its target DNA.*

D. *araBAD* Operon: Positive and Negative Control by the Same Protein

Humans neither metabolize nor intestinally absorb the plant sugar L-arabinose. Hence, the *E. coli* that normally inhabit the human gut are periodically presented with a banquet of this pentose. Three of the five *E. coli*

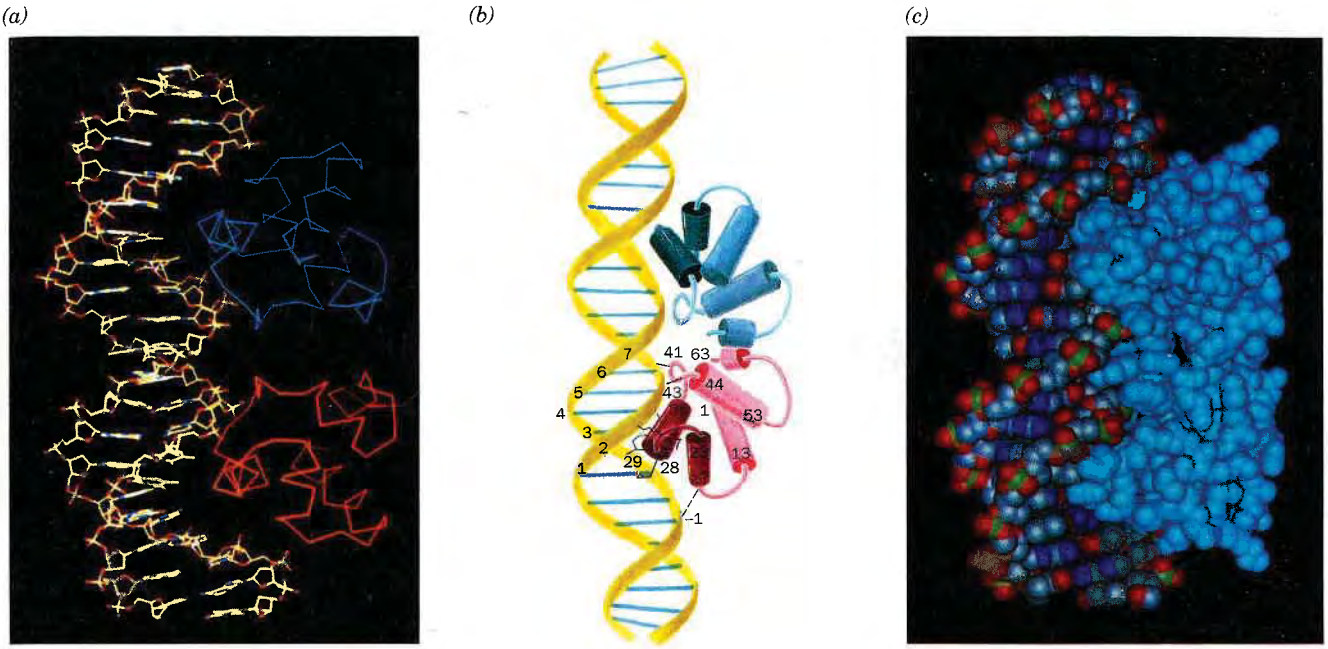


Figure 29-22
The X-ray structure of the 72-residue 434 Cro protein in complex with the same 20 bp DNA shown in Fig. 29-21 as viewed perpendicularly to the complex's twofold axis of symmetry. Only the first 64 residues of the protein are visible. Parts (*a*), (*b*), and (*c*) correspond to those in Fig. 29-21 with the protein in Part (*c*) shown in light blue. Note the close but not identical correspondence between the two structures. [Parts (*a*) and (*c*) courtesy of Alfonso Mondragon, Cynthia Wolberger, and Stephen Harrison, Harvard University. Part (*b*) after Wolberger, C., Dong, Y., Ptashne, M., and Harrison, S. C., *Nature* **335,** 791 (1988).]

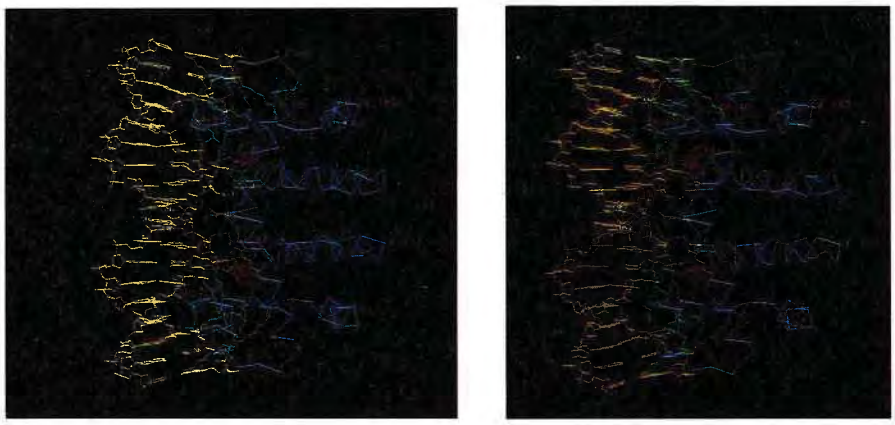


Figure 29-23
The X-ray structure of a *E. coli trp* repressor–operator complex as viewed, in stereo, perpendicular to the molecular twofold axis of symmetry. The protein's C_α backbone is shown (*blue*) together with the side chains (*green*) that make hydrogen bonds (*dashed lines*) to the 18 bp palindromic operator (*yellow*). The protein only binds its operator if L-tryptophan (*red*) is simultaneously bound to the protein. Note that the protein's "recognition" helices bind, as expected, in successive major grooves of the DNA but extend perpendicularly to the DNA duplex axis. In contrast, the "recognition" helices of 434 repressor and Cro proteins bind parallel to the major groove of their DNA (Figs. 29-21 and 29-22). Instructions for viewing stereo diagrams are given in the appendix to Chapter 7. [Courtesy of Paul Sigler, Yale University.]

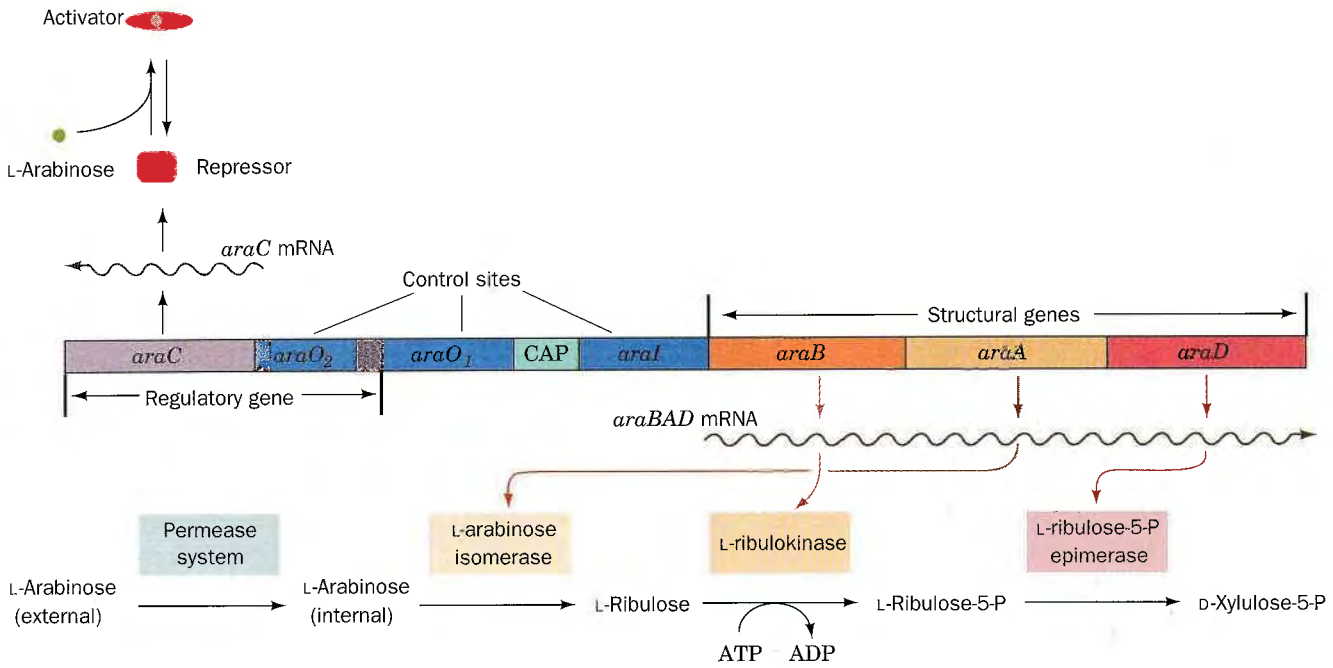


Figure 29-24
A genetic map of the *E. coli araC* and *araBAD* operons indicating the proteins they encode and the reactions in which these proteins participate. The permease system, which transports arabinose into the cell, is the product of the *araE* and *araF* genes, which occur in two independent operons. The pathway product, xylulose-5-phosphate, is converted, via the transketolase reaction, to the glycolytic intermediate fructose-6-phosphate (Section 21-4C). [After Lee, N., *in* Miller. J.H. and Rezinkoff, W. S. (Eds.), *The Operon, pp.* 390, Cold Spring Harbor Laboratory (1979).]

enzymes that metabolize L-arabinose are products of the catabolite repressible ***araBAD* operon** (Fig. 29-24).

The transcription of the araBAD operon is regulated by both CAP–cAMP and the L-arabinose-binding protein, **AraC** (the *araC* gene product; proteins may be assigned the name of the gene specifying them but in roman letters with the first letter capitalized; Fig. 29-25):

1. In the absence of AraC, RNA polymerase initiates transcription of the *araC* gene in the direction away from its upstream neighbor, *araBAD*. The *araBAD* operon remains repressed.

2. When AraC is present, with or without L-arabinose, but not CAP–cAMP (high glucose), AraC binds to three different gene sites: *araI*, which just precedes the *araBAD* promoter; $araO_1$, which overlaps the *araC* promoter; and $araO_2$, which, surprisingly, is located in a noncoding upstream region of the *araC* gene, around position −280 relative to the *araBAD* start site. $araO_1$ is the operator for the *araC* gene; its association with AraC blocks *araC* transcription so that this process is autoregulatory. A series of deletion mutations indicate that both $araO_2$ and *araI* must be present for *araBAD* to be repressed in the presence of AraC. The remarkably large separation between $araO_2$ and the *araBAD* promoter therefore suggests that the DNA is looped such that a single molecule or molecular complex of AraC protein simultaneously binds to both $araO_2$ and *araI*. This cooperative arrangement is required for the AraC-mediated repression of *araBAD* (negative control).

3. When the cAMP level is high (low glucose), CAP–cAMP binds to a site between $araO_1$ and *araI*. When L-arabinose is also present, it binds to AraC causing it to release $araO_2$ so as to open the DNA loop. This combined influence of CAP–cAMP and AraC–arabinose, which is probably mediated through a direct interaction between these two complexes, activates RNA polymerase to transcribe the *araBAD* operon (positive control). The observation that $araO_2$ deletion permits AraC–arabinose to activate *araBAD* in the absence of CAP–cAMP indicates that CAP–cAMP stimulates AraC–arabinose to release $araO_2$ and that this release is required to convert AraC–arabinose to an activator. *araC* remains repressed by AraC.

The function of DNA loop formation is obscure although it has been demonstrated to occur in numerous bacterial and eukaryotic systems. Perhaps it permits several regulatory proteins and/or regulatory sites on one protein to simultaneously influence transcription

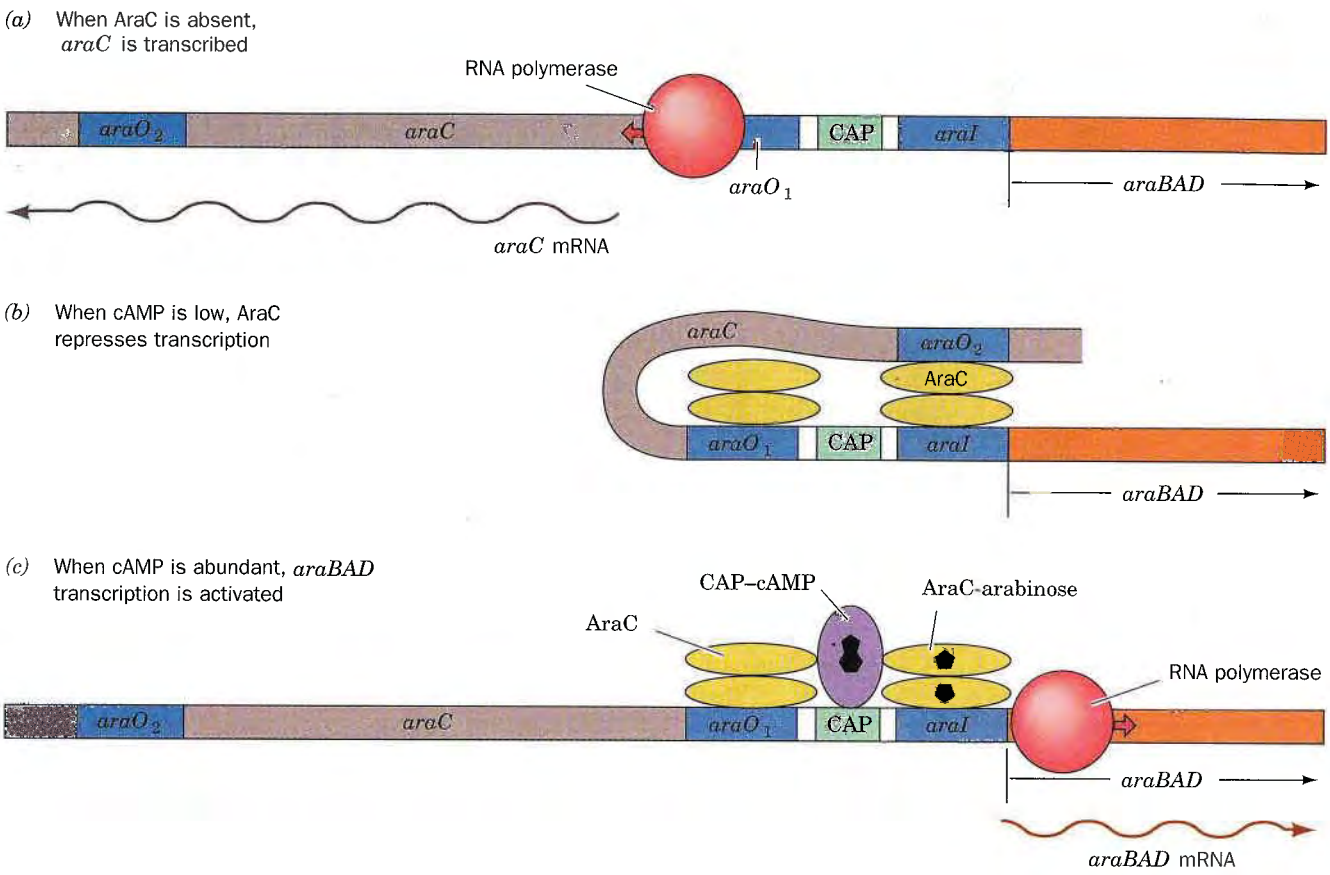


Figure 29-25
The proposed mechanism for *araBAD* regulaion: (*a*) In the absence of AraC, RNA polymerase initiates the transcription of *araC* but not *araBAD*. (*b*) When AraC is present, with or without L-arabinose, and the cAMP level is low, AraC binds to $araO_1$ and links together $araO_2$ and *araI* to form a DNA loop, thereby repressing both *araC* and *araBAD*. (*c*) When AraC and L-arabinose are both present and cAMP is abundant, the AraC–arabinose complex releases $araO_2$ but remains bound to *araI* where, in concert with CAP–cAMP, it activates *araBAD* transcription. *araC* remains repressed.

initiation by RNA polymerase. In fact, as recent studies have shown, the *lac* operon contains a second, relatively weak operator located 400 bp downstream from the transcription start site (within the *lacZ* gene). This secondary operator (O_2) cooperates with the primary operator (now called O_1) to form a repression complex that is stronger than with either operator alone. It is thought that during severe repression, both operators bind to a single *lac* repressor tetramer to form a DNA loop-containing complex.

E. *trp* Operon: Attenuation

In the following paragraphs we discuss a sophisticated transcriptional control mechanism named **attenuation** through which bacteria regulate the expression of certain operons involved in amino acid biosynthesis. This mechanism was discovered through the study of the ***E. coli trp operon*** (Fig. 29-26) which encodes five polypeptides comprising three enzymes that mediate the synthesis of tryptophan from chorismate (Section 24-5B). Charles Yanofsky established that the *trp* operon genes are coordinately expressed under the control of ***trp* repressor,** a dimeric protein of identical 107-residue subunits that is the product of the *trpR* gene (which forms an independent operon). *The trp repressor binds L-tryptophan, the pathway's end product, to form a complex that specifically binds to trp operator (trpO; Fig. 29-27), so as to reduce the rate of trp operon transcription by 70-fold.* The X-ray structure of the *trp* repressor–operator complex (Section 29-3C) indicates that tryptophan binding allosterically orients *trp* repressor's two symmetry related helix–turn–helix "DNA reading heads" so that they can simultaneously bind to *trpO* (Fig. 29-28; also see Fig. 29-23). Moreover, the bound tryptophan forms a hydrogen bond to a DNA phosphate group, thereby strengthening the repressor–operator association. Tryptophan therefore acts as a **corepressor;** its presence prevents what is then super-

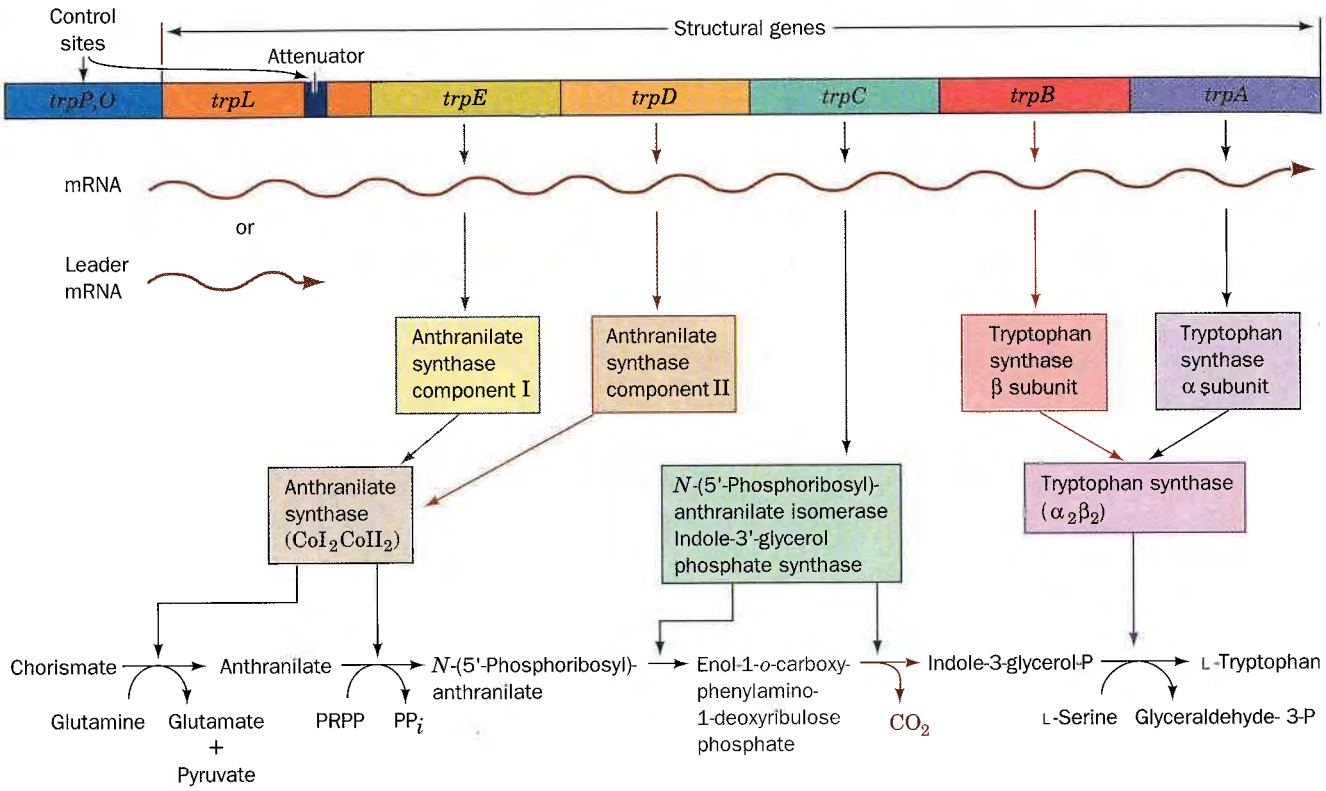


Figure 29-26
A genetic map of the *E. coli trp* operon indicating the enzymes it specifies and the reactions they catalyze. The gene product of *trpC* catalyzes two sequential reactions in the synthesis of tryptophan. [After Yanofsky, C., *J. Am. Med. Assoc.* **218,** 1027 (1971).]

fluous tryptophan biosynthesis. The *trp* repressor also controls the synthesis of at least two other operons: the ***trpR* operon** and the ***aroH* operon** (which encodes one of three isozymes that catalyze the initial reaction of aromatic amino acid biosynthesis: Section 24-5B).

Tryptophan Biosynthesis Is Also Regulated by Attenuation

The *trp* repressor–operator system was at first thought to fully account for the regulation of tryptophan biosynthesis in *E. coli.* However, the discovery of *trp* deletion mutants located downstream from *trpO* that increase *trp* operon expression sixfold indicated the existence of an additional transcriptional control element.

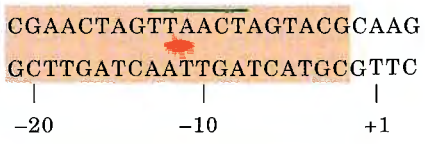


Figure 29-27
The base sequence of the *trp* operator. The nearly palindromic sequence is boxed and the Pribnow box is overscored.

Sequence analysis established that *trpE,* the *trp* operon's leading structural gene, is preceded by a 162-nucleotide **leader sequence** *(trpL).* Genetic analysis indicated that the new control element is located in *trpL,* ~30 to 60 nucleotides upstream of *trpE* (Fig. 29-26).

When tryptophan is scarce, the entire 6720-nucleotide polycistronic *trp* mRNA, including the *trpL* sequence, is synthesized. As the tryptophan concentration increases, the rate of *trp* transcription decreases as a result of the *trp* repressor–corepressor complex's consequent greater abundance. Of the *trp* mRNA that is transcribed, however, an increasing proportion consists of only a 140-nucleotide segment corresponding to the 5′ end of *trpL. The availability of tryptophan therefore results in the premature termination of trp operon transcription.* The control element responsible for this effect is consequently termed an **attenuator.**

The *trp* Attenuator's Transcription Terminator Is Masked When Tryptophan Is Scarce

What is the mechanism of attenuation? The attenuator transcript contains four complementary segments that can form one of two sets of mutually exclusive based paired hairpins (Fig. 29-29). *Segments 3 and 4 together with the succeeding residues comprise a normal transcription terminator (Section 29-2E):* a G + C-rich se-

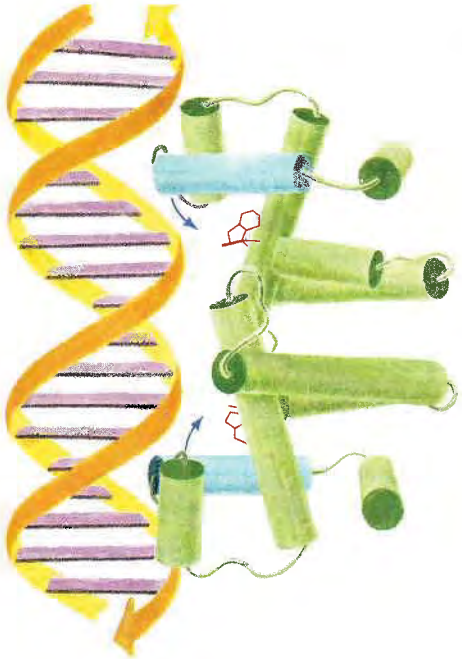


Figure 29-28
The structure of the *trp* repressor–tryptophan complex in association with its operator. The "recognition" helix (*blue*) of the dimeric protein's helix–turn–helix motif binds in the major groove of its operator DNA (see Fig. 29-23). Comparison of the X-ray structures of the *trp* repressor with and without bound tryptophan (*red*) indicates that, upon tryptophan dissociation, the "recognition" helices swing inwards (*arrows*) so that they can no longer simultaneously engage the DNA's major groove. [After Robertson, M., *Nature* **327,** 465 (1987).]

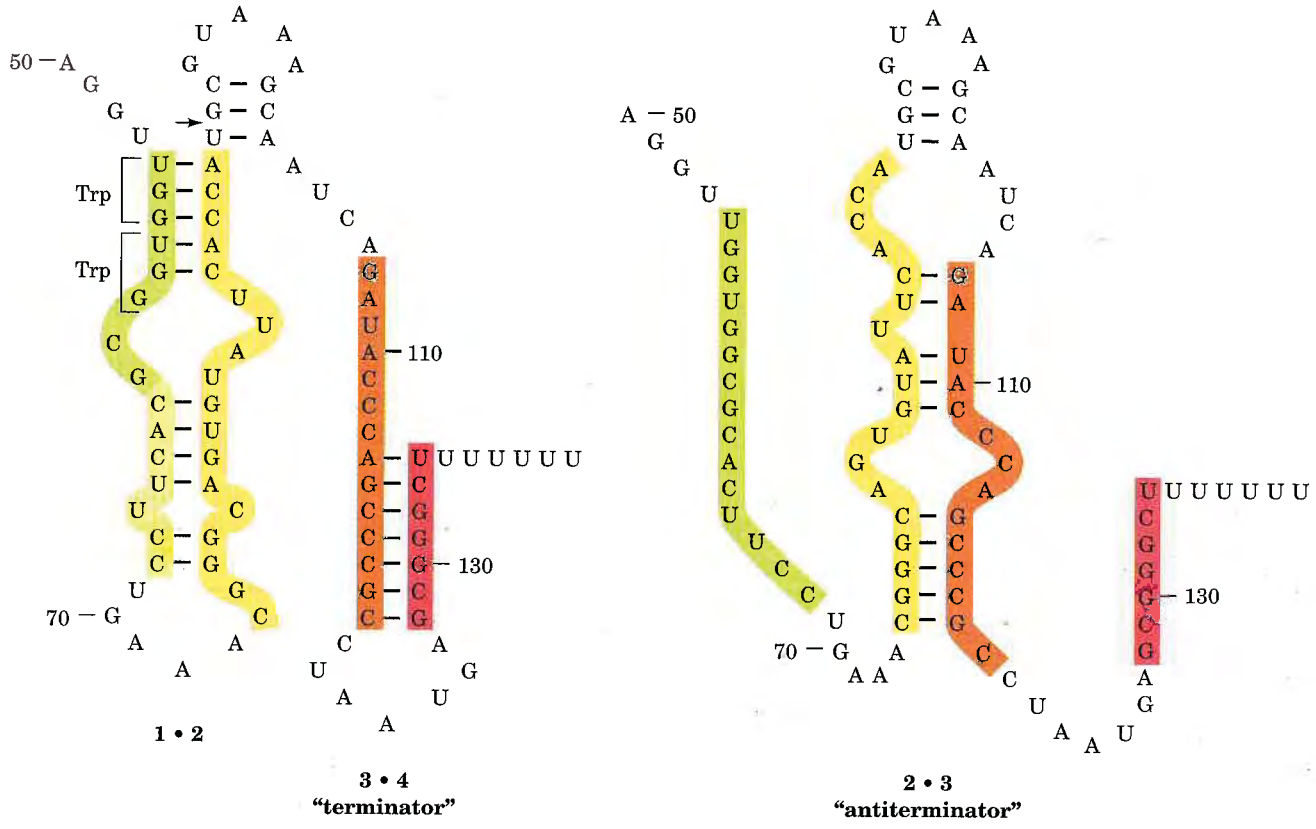


Figure 29-29
The alternative secondary structures of *trpL* mRNA. The formation of the base paired 2·3 (antiterminator) hairpin (*right*) precludes the formation of the 1·2 and 3·4 (terminator) hairpins (*left*) and *vice versa.* Attenuation results in the premature termination of transcription immediately after nucleotide 140 when the 3·4 hairpin is present. The arrow indicates the mRNA site past which RNA polymerase pauses until approached by an active ribosome. [After Fisher, R. F. and Yanofsky, C., *Proc. Natl. Acad. Sci.* **258,** 8147 (1983).]

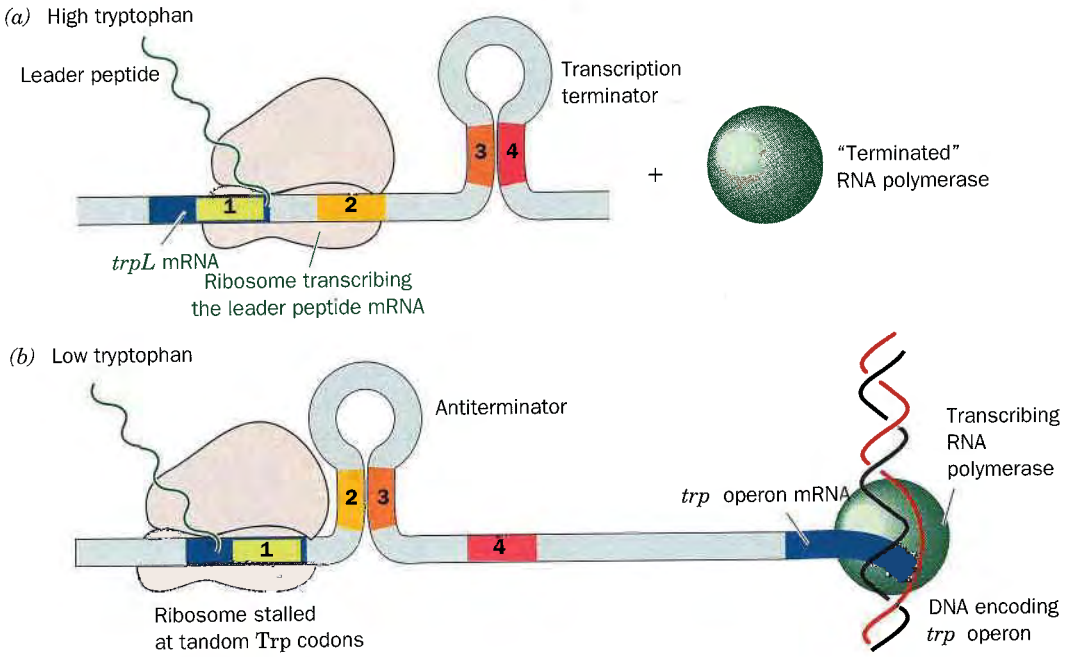


Figure 29-30
Attenuation in the *trp* operon. (*a*) When tryptophanyl–tRNA^{Trp} is abundant, the ribosome translates *trpL* mRNA. The presence of the ribosome on segment 2 prevents the formation of the base paired 2·3 hairpin. The 3·4 hairpin, an essential component of the transcriptional terminator, can thereby form thus aborting transcription. (*b*) When tryptophanyl–tRNA^{Trp} is scarce, the ribosome stalls on the tandem Trp codons of segment 1. This situation permits the formation of the 2·3 hairpin which, in turn, precludes the formation of the 3·4 hairpin. RNA polymerase therefore transcribes through this unformed terminator and continues *trp* operon transcription.

quence that can form a self-complementary hairpin structure followed by several sequential U's (compare with Fig. 29-14). *Transcription rarely proceeds beyond this termination site unless tryptophan is in short supply.*

A section of the leader sequence, which includes segment 1 of the attenuator, is translated to form a 14-residue polypeptide that contains two consecutive Trp residues (Fig. 29-29, *left*). The position of this particularly rare dipeptide segment (only ~1% of the residues in *E. coli* proteins are Trp) provided an important clue to the mechanism of attenuation. An additional essential aspect of this mechanism is that ribosomes commence the translation of a prokaryotic mRNA shortly after its 5′ end has been synthesized.

The above considerations led Yanofsky to propose the following model of attenuation (Fig. 29-30). An RNA polymerase that has escaped repression initiates *trp* operon transcription. Soon after the ribosomal initiation site of the *trpL* gene has been transcribed, a ribosome attaches to it and begins translation of the leader peptide. When tryptophan is abundant, so that there is a plentiful supply of tryptophanyl–tRNA^{Trp} (the transfer RNA specific for Trp with an attached Trp residue; Section 30-2C), the ribosome follows closely behind the transcribing RNA polymerase so as to sterically block the formation of the 2·3 hairpin. Indeed, RNA polymerase pauses past position 92 of the transcript and only continues transcription upon the approach of a ribosome, thereby ensuring the proximity of these two entities at this critical position. The prevention of 2·3 hairpin formation permits the formation of the 3·4 hairpin, the transcription terminator pause site, which results in the termination of transcription (Fig. 29-30*a*). When tryptophan is scarce, however, the ribosome stalls at the tandem UGG codons (the three sequential nucleotides specifying Trp; Section 30-1E) because of the lack of tryptophanyl–tRNA^{Trp}. As transcription continues, the newly synthesized segments 2 and 3 form a hairpin because the stalled ribosome prevents the otherwise competitive formation of the 1·2 hairpin (Fig. 29-30*b*). The formation of the transcriptional terminator's 3·4 hairpin is thereby preempted for sufficient time for RNA polymerase to transcribe through it and consequently through the remainder of the *trp* operon. The cell is thus provided with a regulatory mechanism that is responsive to tryptophanyl–tRNA^{Trp} level, which, in turn, depends on the protein synthesis rate as well as the tryptophan supply.

There is considerable evidence supporting the preceding model of attentuation. The *trpL* transcript is resistant to limited RNase T1 digestion indicating that it has extensive secondary structure. The significance of

the tandem Trp codons in the *trpL* transcript is corroborated by their presence in *trp* leader regions of several other bacterial species. Moreover, the leader peptides of the five other amino acid-biosynthesizing operons known to be regulated by attenuation (most exclusively so) are all rich in their corresponding amino acid residues (Table 29-1). For example, the *E. coli* ***his*** **operon,** which specifies enzymes synthesizing histidine, has seven tandem His residues in its leader peptide; similarly, the ***ilv*** **operon,** which specifies enzymes participating in isoleucine, leucine, and valine biosynthesis, has five Ile's, three Leu's, and six Val's in its leader peptide. Finally, the leader transcripts of these operons resemble that of the *trp* operon in their capacity to form two alternative secondary structures, one of which contains a trailing termination structure.

F. Regulation of Ribosomal RNA Synthesis: The Stringent Response

E. coli cells growing under optimal conditions divide every 20 min. Such cells contain nearly 10,000 ribosomes. Yet, RNA polymerase can initiate the transcription of an rRNA gene no faster than about once every second. If the *E. coli* genome contained only one copy of each of the three types of rRNA genes (those specifying the so-called 23S, 16S, and 5S rRNAs; Section 30-3A), there could be no more than ~1200 ribosomes/cell. However, *the E. coli chromosome contains seven separately located rRNA operons, all of which contain one nearly identical copy of each type of rRNA gene,* thereby accounting for the observed rRNA synthesis rate.

Cells have the remarkable ability to coordinate the rates at which their thousands of components are synthesized. For example, *E. coli* adjust their ribosome content to match the rate that they can synthesize proteins under the prevailing growth conditions. The rate of rRNA synthesis is therefore proportional to the rate of protein synthesis. One mechanism by which this occurs is known as the **stringent response:** *A shortage of any species of amino acid-charged tRNA (usually a result of "stringent" or poor growth conditions) that limits the rate of protein synthesis triggers a sweeping metabolic readjustment.* A major facet of this change is an abrupt 10-to 20-fold reduction in the rate of rRNA and tRNA synthesis. This **stringent control,** moreover, depresses numerous metabolic processes (including DNA replication and the biosynthesis of carbohydrates, lipids, nucleotides, proteoglycans, and glycolytic intermediates) while stimulating others (such as amino acid biosynthesis). The cell is thereby prepared to withstand nutritional deprivation.

ppGpp Mediates the Stringent Response

The stringent response is correlated with a rapid intracellular accumulation of the unusual nucleotide ***ppGpp*** *and its prompt decay when amino acids become available.* The observation that mutants, designated *relA*$^-$, which do not exhibit the stringent response (they are said to have **relaxed control**), lack ppGpp suggests that this substance mediates the stringent response. This idea was corroborated by *in vitro* studies demonstrating, for example, that ppGpp inhibits the transcription of rRNA genes but stimulates the transcription of the *trp* and *lac* operons as does the stringent response *in vivo.* It therefore seems that ppGpp acts by somehow altering RNA polymerase's promoter specificity at stringently controlled operons, an hypothesis that is supported by the isolation of RNA polymerase mutants that exhibit reduced responses to ppGpp.

Experiments with cell-free *E. coli* extracts have established that the protein encoded by wild-type *relA* gene, named **stringent factor,** catalyzes the reaction

$$\text{ATP} + \text{GDP} \rightleftharpoons \text{AMP} + \text{ppGpp}$$

Stringent factor is only active in association with a ribo-

Table 29-1
Amino Acid Sequences of Some Leader Peptides in Operons Subject to Attenuation

Operon	Amino Acid Sequence[a]
trp	Met-Lys-Ala-Ile-Phe-Val-Leu-Lys-Gly-TRP-TRP-Arg-Thr-Ser
pheA	Met-Lys-His-Ile-Pro-PHE-PHE-PHE-Ala-PHE-PHE-PHE-Thr-PHE-Pro
his	Met-Thr-Arg-Val-Gln-Phe-Lys-HIS-HIS-HIS-HIS-HIS-HIS-HIS-Pro-Asp
leu	Met-Ser-His-Ile-Val-Arg-Phe-Thr-Gly-LEU-LEU-LEU-LEU-Asn-Ala-Phe-Ile-Val-Arg-Gly-Arg-Pro-Val-Gly-Gly-Ile-Gln-His
thr	Met-Lys-Arg-ILE-Ser-THR-THE-ILE-THR-THR-THR-ILE-THR-ILE-THR-THR-Gln-Asn-Gly-Ala-Gly
ilv	Met-Thr-Ala-LEU-LEU-Arg-VAL-ILE-Ser-LEU-VAL-VAL-ILE-Ser-VAL-VAL-VAL-ILE-ILE-ILE-Pro-Pro-Cys-Gly-Ala-Ala-Leu-Gly-Arg-Gly-Lys-Ala

[a] Upper case residues are synthesized in the pathway catalyzed by the operon's gene products.

Source: Yanofsky, C., *Nature* **289,** 753 (1981).

some that is actively engaged in translation. ppGpp synthesis occurs at a maximal rate when a ribosome binds its mRNA-specified but uncharged (lacking an amino acid residue) tRNA. The binding of a specified and charged tRNA greatly reduces the rate of ppGpp synthesis. *The ribosome apparently signals the shortage of an amino acid by stimulating the synthesis of ppGpp which, acting as an intracellular messenger, influences the rates at which a great variety of operons are transcribed.*

ppGpp degradation is catalyzed by the *spoT* gene product. The *spoT*$^-$ mutants show a normal increase in ppGpp level upon amino acid starvation but an abnormally slow decay of ppGpp to basal levels when amino acids again become available. The *spoT*$^-$ mutants therefore exhibit a sluggish recovery from the stringent response. *The ppGpp level is apparently regulated by the countervailing activities of stringent factor and the spoT gene product.*

4. POST-TRANSCRIPTIONAL PROCESSING

The immediate products of transcription, the **primary transcripts,** are not necessarily functional entities. In order to acquire biological activity, many of them must be specifically altered in several ways: (1) by the exo and endonucleolytic removal of polynucleotide segments; (2) by appending nucleotide sequences to their 3′ and 5′ ends; and (3) by the modification of specific nucleosides. The three major classes of RNAs, mRNA, rRNA, and tRNA, are altered in different ways in prokaryotes and in eukaryotes. In this section we shall outline these **post-transcriptional modification** processes.

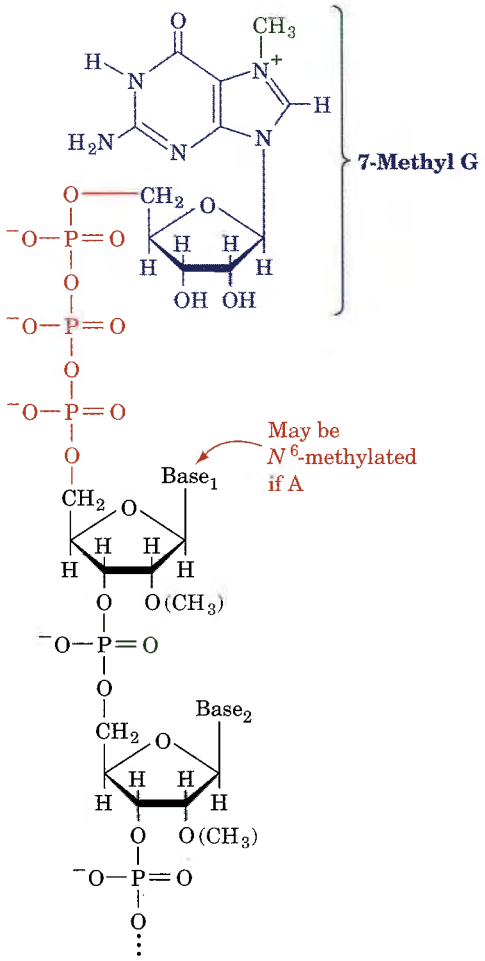


Figure 29-31
The structure of the 5′ cap of eukaryotic mRNAs. It is known as cap-0, cap-1, or cap-2, respectively, if it has no further modifications, if the leading nucleoside of the transcript is *O*(2′)-methylated, or if its first two nucleosides are 0(2′)-methylated.

A. Messenger RNA Processing

In prokaryotes, most primary mRNA transcripts function in translation without further modification. Indeed, as we have seen, ribosomes in prokaryotes usually commence translation on nascent mRNAs. In eukaryotes, however, mRNAs are synthesized in the cell nucleus while translation occurs in the cytosol. Eukaryotic mRNA transcripts can therefore undergo extensive post-transcriptional processing while still in the nucleus.

Eukaryotic mRNAs Are Capped

Eukaryotic mRNAs have a peculiar enzymatically appended ***cap structure*** *consisting of a 7-methylguanosine residue joined to the transcript's initial (5′) nucleoside via a 5′–5′ triphosphate bridge (Fig. 29-31).* The cap, which a specific guanylyltransferase adds to the growing transcript before it is > 20-nucleotides long, defines the eukaryotic translational start site (Section 30-3C). A cap may be *O* (2′)-methylated at the transcript's leading nucleoside (**cap-1,** the predominant cap in multicellular organisms), at its first two nucleosides **(cap-2),** or at neither of these positions (**cap-0,** the predominant cap in unicellular eukaryotes). If the leading nucleoside is adenosine (it is usually a purine), it may also be N^6-methylated.

Eukaryotic mRNAs Have Poly(A) Tails

Eukaryotic mRNAs, in contrast to those of prokaryotes, are invariably monocistronic. Yet, the sequences signaling transcriptional termination in eukaryotes have not been identified. This is largely because the termination process is imprecise; that is, the primary transcripts of a given structural gene have heterogeneous 3′ sequences. Nevertheless, mature eukaryotic mRNAs have well-defined 3′ ends; *almost all of them have 3′-poly(A) tails of 100*

to 200 nucleotides. The poly(A) tails are enzymatically appended to the primary transcripts in two reactions:

1. A transcript is cleaved 10 to 30 nucleotides past a highly conserved AAUAAA sequence, whose mutation abolishes cleavage and polyadenylation, and within 50 nucleotides before a less conserved U-rich or GU-rich sequence.
2. The poly(A) tail is subsequently generated from ATP through the stepwise action of **poly(A) polymerase.**

The precision of the cleavage reaction has apparently eliminated the need for accurate transcriptional termination; to put things another way, all's well that ends well.

In vitro studies indicate that a poly(A) tail is not required for mRNA translation. Rather, the observations that an mRNA's poly(A) tail shortens as it ages in the cytosol and that unadenylated mRNAs have abbreviated cytosolic lifetimes suggest that poly(A) tails have a protective role. In fact, the only mature mRNAs that generally lack poly(A) tails, those of histones (which, with few exceptions, lack the AAUAAA cleavage–polyadenylation signal), have lifetimes of <30 min in the cytosol, whereas most other mRNAs last hours or days.

Eukaryotic Genes Consist of Alternating Expressed and Unexpressed Sequences

The most striking difference between eukaryotic and prokaryotic structural genes is that the coding sequences of most eukaryotic genes are interspersed with unexpressed regions. Early investigations of eukaryotic structural gene transcription found, quite surprisingly, that primary transcripts are quite heterogeneous in length (from ~2000 to well over 20,000 nucleotides) and are much larger than is expected from the known sizes of eukaryotic proteins. Rapid labeling experiments demonstrated that little of this so-called **heterogeneous nuclear RNA (hnRNA)** is ever transported to the cytosol; most of it is quickly turned over in the nucleus. Yet, the hnRNA's 5′ caps and 3′ tails eventually appear in cytosolic mRNAs. *The straightforward explanation of these observations, that* ***pre-mRNAs*** *are processed by the excision of internal sequences, seemed so bizarre that it came as a great suprise in 1977 when it was independently demonstrated in several laboratories that this is actually the case.* In fact, the pre-mRNA's noncoding **intervening sequences (IVSs** or **introns)** are usually of greater length than their flanking **expressed sequences (exons).** This situation is graphically illustrated in Fig. 29-32, which is an electron micrograph of chicken **ovalbumin** mRNA hybridized to the sense strand of the ovalbumin gene (ovalbumin is the major protein component of egg white). The lengths of introns in vertebrate genes ranges from ~65 to over 100,000 nucleotides with no obvious periodicity. Indeed, the corresponding introns from genes in two vertebrate species can vary extensively in both length and sequence so as to bear little resemblance to one another.

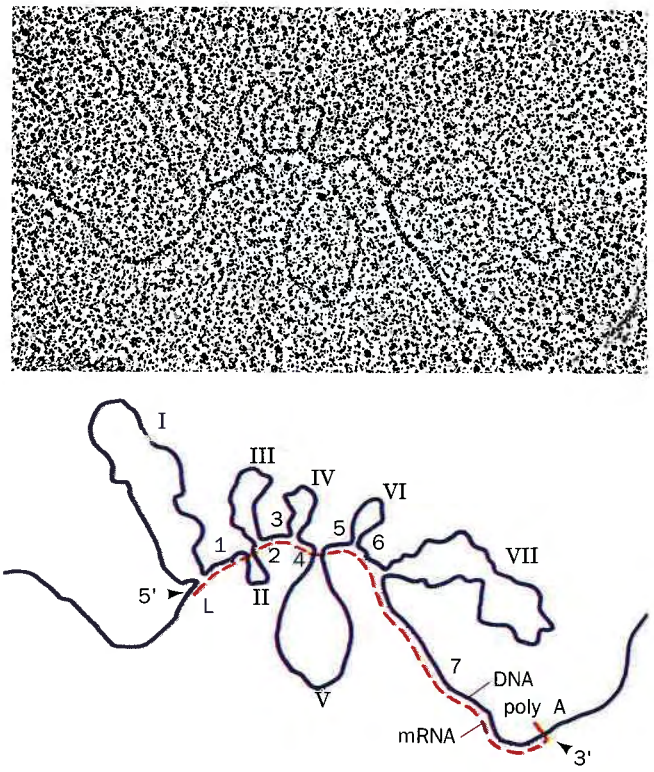


Figure 29-32
An electron micrograph and its interpretive drawing of a hybrid between the sense strand of the chicken ovalbumin gene (as obtained by molecular cloning methods; Section 31-5) and its corresponding mRNA. The complementary segments of the DNA (*purple line in drawing*) and mRNA (*red line*) have annealed to reveal the exon positions (*L*, 1–7). The looped-out segments (I–VII), which have no complementary sequences in the mRNA, are the introns. [From Chambon, P., *Sci. Am.* **244**(5): 61 (1981).

Further investigations established that the formation of eukaryotic mRNA begins with the transcription of an entire structural gene, including its introns, to form pre-mRNA (Fig. 29-33). Then, following capping and perhaps polyadenylation, the introns are excised and their flanking exons are connected, a process called **gene splicing,** to yield the mature mRNA. *The most striking aspect of gene splicing is its precision; if one base too few or too many were excised, the resulting mRNA could not be translated properly (Section 30-1B). Moreover, exons are never shuffled; their order in the mature mRNA is exactly the same as that in the gene from which it is derived.* In the following subsections we discuss the mechanism of this remarkable splicing process.

Exons Are Spliced in a Two-Stage Reaction

Sequence comparisons of exon–intron junctions from a diverse group of eukaryotes indicate that they

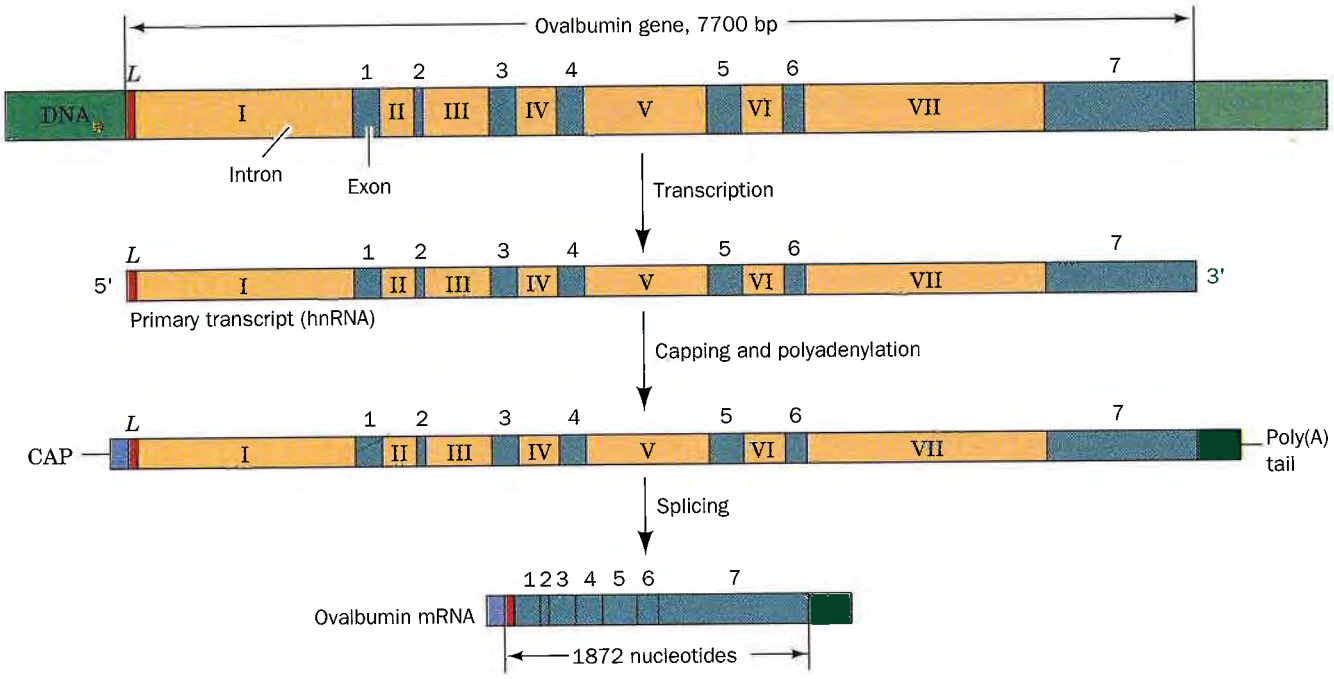


Figure 29-33
The sequence of steps in the production of mature eukaryotic mRNA as shown for the chicken ovalbumin gene. Following transcription, the primary transcript is capped and polyadenylated. The introns are then excised and the exons spliced together to form the mature mRNA.

have a high degree of homology (Fig. 29-34), including, as Richard Breathnach and Pierre Chambon first pointed out, *an invariant GU at the intron's 5′ boundary and an invariant AG at its 3′ boundary. These sequences are necessary and sufficient to define a splice junction:* Mutations that alter the sequences interfere with splicing, whereas mutations that change a nonjunction to a consensuslike sequence can generate a new splice junction.

Investigations of both cell free and *in vivo* splicing systems by Argiris Efstradiadis, Michael Rosbash, Phillip Sharp, and Tom Maniatis have established that intron excision occurs in two reactions (Fig. 29-35):

1. The formation of a 2′,5′ phosphodiester bond between an intron adenosine residue and its 5′-terminal phosphate group with the concommitant release of the 5′ exon. *The intron thereby assumes a novel* ***lariat structure.*** The adenosine residue at the lariat branch has been identified as the *A* in the sequence CURAY [where R represents purines (A or G) and Y represents pyrimidines (C or U)], which is highly conserved in vertebrate mRNAs and is typically located 20 to 50 residues upstream of the 3′ splice site (yeast have a similar UACUAAC sequence that occurs ~50-residues upstream from all its 3′ splice sites). Mutations that change this branch point A residue abolish splicing at that site.
2. The now free 3′-OH group of the 5′ exon forms a phosphodiester bond with the 5′-terminal phosphate of the 3′ exon yielding the spliced product. The intron is thereby eliminated in its lariat form and, *in vivo,* is rapidly degraded. Mutations that alter the conserved

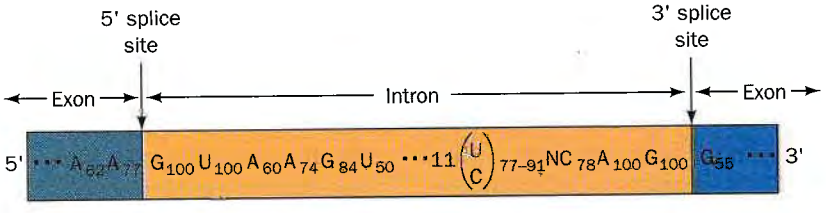


Figure 29-34
The consensus sequence at the exon–intron junctions of eukaryotic pre-mRNAs. The subscripts indicate the percent of pre-mRNAs in which specified base(s) occurs. Note that the 3′ splice site is preceded by a tract of 11 predominantly pyrimidine nucleotides. [Based on data from Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. S., and Sharp, P. A., *Annu. Rev. Biochem.* **55,** 1123 (1986).]

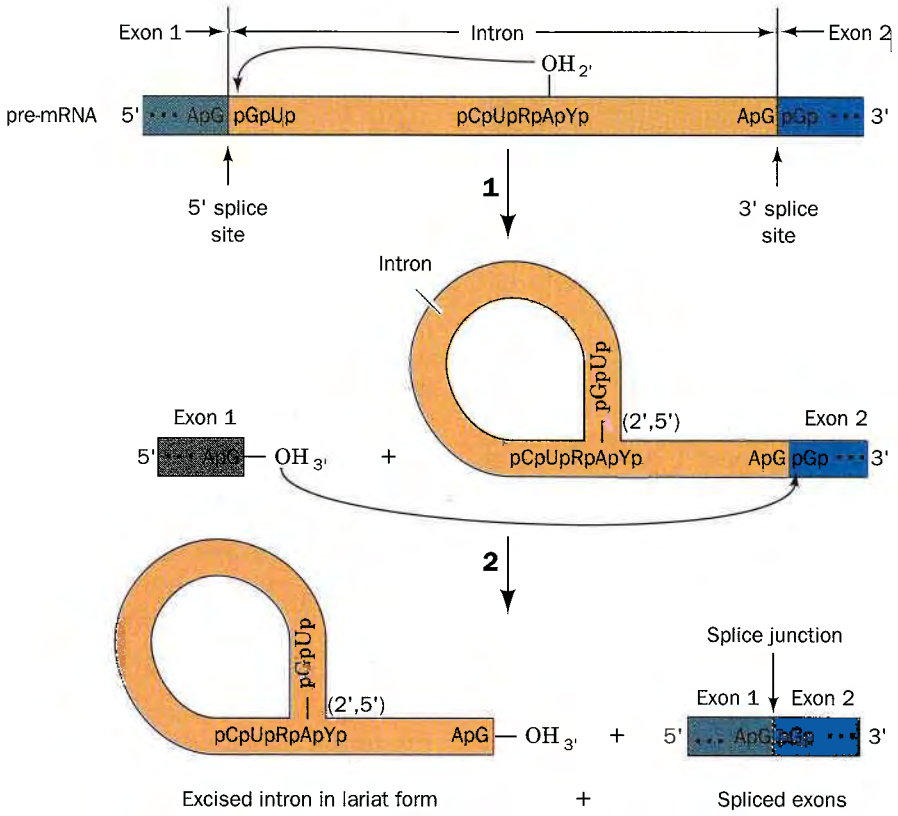


Figure 29-35
The sequence of transesterification reactions that splice together the exons of eukaryotic pre-mRNAs (the exons and introns are respectively drawn in black and red; R and Y, resectively, represent purine and pyrimidine residues): **(1)** The 2′-OH group of a specific intron A residue nucleophilically attacks the 5′ phosphate at the 5′ intron boundary to yield an unusual 2′,5′-phosphodiester bond and thus form a lariat structure. **(2)** The liberated 3′-OH group forms a 3′,5′-phosphodiester bond with the 5′ terminal residue of the 3′ exon, thereby splicing the two exons together and releasing the intron in lariat form.

AG at the 3′ splice junction block this second step although they do not interfere with lariat formation.

Note that the splicing process proceeds without free energy input; its transphosphorylation reactions preserve the free energy of each cleaved phosphodiester bond through the concomitant formation of a new one.

Splicing Is Mediated by snRNPs

How are splice junctions recognized and how are the two exons to be joined brought together in the splicing process? Part of the answer to this question was established by Joan Steitz going on the assumption that one nucleic acid is best recognized by another. The eukaryotic nucleus, as has been known since the 1960s, contains numerous copies of several highly conserved 60 to 300 nucleotide RNAs called **small nuclear RNAs (snRNAs),** which form protein complexes termed **small nuclear ribonucleoproteins (snRNPs;** pronounced "snurps"). Steitz recognized that the 5′ end of one of these snRNAs, **U1-snRNA** (so-called because it is a member of a U-rich subfamily of snRNAs), is partially complementary to the 5′ consensus sequence of mRNA splice junctions. The consequent hypothesis, that *U1-snRNP recognizes the 5′ splice junction,* was corroborated by the observations that splicing is inhibited by the selective destruction of the U1-snRNP sequences that are complementary to the 5′ splice junction and by the presence of anti-U1-snRNP antibodies (produced by patients suffering from **systemic lupus erythematosus,** an often fatal autoimmune disease). Similar studies have implicated **U2-snRNP** in recognizing the intron region that forms the lariat branch point and **U5-snRNP** in recognizing the 3′ splice junction. Altogether, ~65 pre-mRNA nucleotides participate in this recognition process, which rationalizes why introns are minimally 65 nucleotides in length.

Splicing takes place in an as yet poorly characterized 50S to 60S particle dubbed the ***splicosome*** *(Fig. 29-36).* The splicosome brings together a pre-mRNA, the foregoing snRNPs, **U4–U6-snRNP** (in which U4 and U6 snRNAs associate by base pairing and which binds to the other snRNPs rather than directly to pre-mRNA), and a vari-

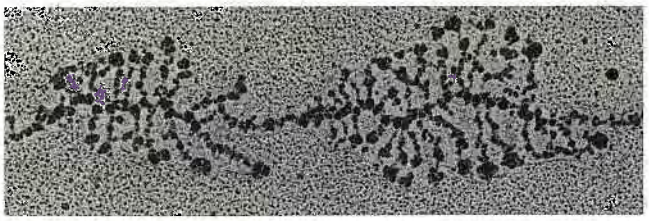


Figure 29-36
An electron micrograph of splicosomes in action. The splicosomes are the large beads on the pre-mRNAs extending above and below the horizontal DNA. [From Steitz, J. A., *Sci. Am.* **258**(6): 59 (1988). Electron micrograph by Yvonne N. Osheim.]

ety of pre-mRNA binding proteins (U4 – U6-snRNP has also been implicated in the previously described polyadenylation reaction). Note that the splicosome is a large particle; the similarly sized large ribosomal subunit of E. coli consists of 3004 nucleotides and 31 polypeptides and has a particle mass of 1.6 million daltons (Section 30-3A). The biochemical significance of splicing is discussed in Section 33-2F.

mRNA Is Methylated at Certain Adenylate Residues

During or shortly after the synthesis of vertebrate pre-mRNAs, ~0.1% of their A residues are methylated at N(6). These m^6A's tend to occur in the sequence RRm^6ACX, where X is rarely G. Although the functional significance of these methylated A's is unknown, it should be noted that a large fraction of them are components of the corresponding mature mRNAs.

B. Ribosomal RNA Processing

The seven *E. coli* rRNA operons all contain one (nearly identical) copy of each of the three types of rRNA genes (Section 29-3F). Their polycistronic primary transcripts, which are >5500 nucleotides in length, contain 16S rRNA at their 5′ ends followed by the transcripts for 1 or 2 tRNAs, 23S rRNA, 5S rRNA and, in some rRNA operons, 1 or 2 more tRNAs at the 3′ end (Fig. 29-37). The steps in processing these primary transcripts to mature rRNAs (Fig. 29-37) were elucidated with the aid of mutants defective in one or more of the processing enzymes.

The initial processing, which yields products known as **pre-rRNAs,** commences while the primary transcript is still being synthesized. It consists of specific endonucleolytic cleavages by **RNase III, RNase P, RNase E,** and **RNase F** at the sites indicated in Fig. 29-37. The base sequence of the primary transcript suggests the existence of several base paired stems. The RNase III cleavages occur in a stem consisting of complementary sequences flanking the 5′ and 3′ ends of the 23S segment (Fig. 29-38) as well as that of the 16S segment. Presumably certain features of these stems constitute the RNase III recognition site.

The 5′ and 3′ ends of the pre-rRNA's are trimmed away in secondary processing steps (Fig. 29-37) through the action of **RNAses M16, M23,** and **M5** to produce the mature rRNAs. These final cleavages only occur after the pre-rRNAs become associated with ribosomal proteins.

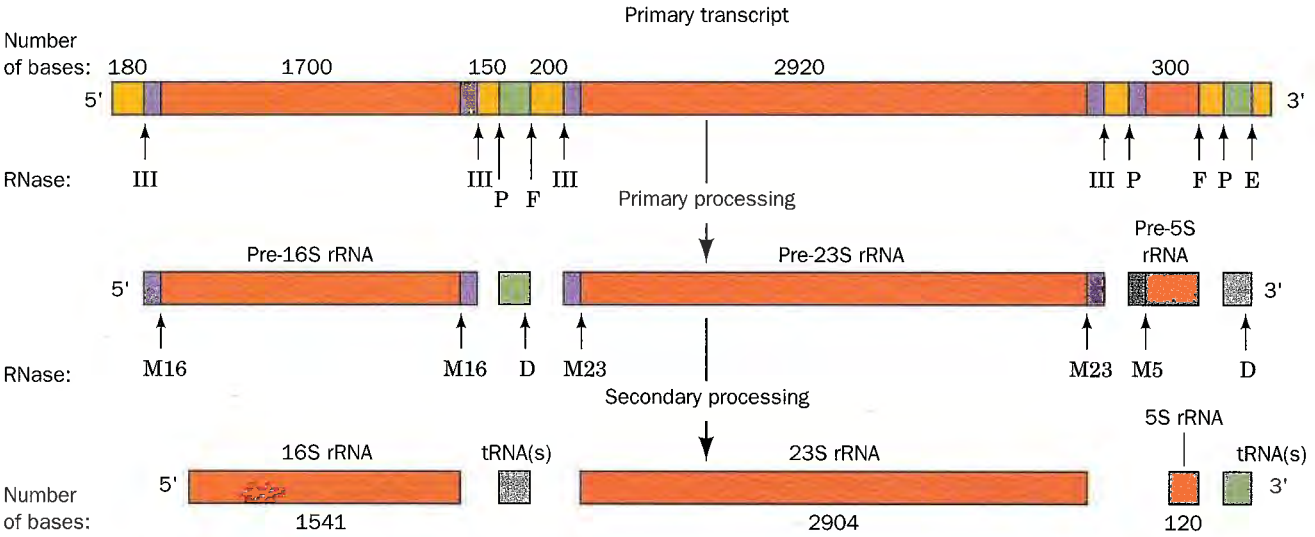


Figure 29-37
The post-transcriptional processing of *E. coli* rRNA. The transcriptional map is shown approximately to scale. The labeled arrows indicate the positions of the various nucleolytic cuts and the nucleases that generate them.

[After Apiron, D., Ghora, B. K., Plantz, G., Misra, T. K., and Gegenheimer, P., *in* Söll, D., Abelson, J. N., and Schimmel P. R. (Eds.), *Transfer RNA: Biological Aspects, p.* 148, Cold Spring Harbor Laboratory (1980).]

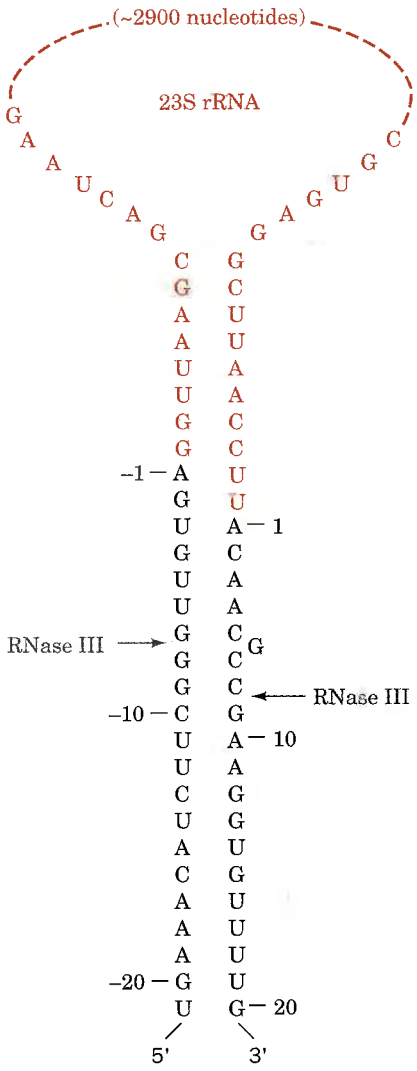


Figure 29-38
The proposed stem-and-giant-loop secondary structure in the 23S region of the *E. coli* primary rRNA transcript. The RNase III cleavage sites are indicated. [After Young. R. R., Bram, R. J., and Steitz, J. A., *in* Söll, D., Abelson, J. N., and Schimmel, P. R. (Eds.), *Transfer RNA: Biological Aspects*, p. 102, Cold Spring Harbor Laboratory (1980).]

Ribosomal RNAs Are Methylated

During ribosomal assembly, the 16S and 23S rRNAs are methylated at a total of 24 specific nucleosides. The methylation reactions, which employ *S*-adenosylmethionine (Section 24-3E) as a methyl donor, yield N^6,N^6-dimethyladenine and $O^{2'}$-methylribose residues. $O^{2'}$-methyl groups are thought to protect adjacent phosphodiester bonds from degradation by intracellular RNases (the mechanism of RNase hydrolysis involves utilization of the free 2′-OH group of ribose to eliminate the substituent on the 3′-phosphoryl group via the formation of a 2′,3′-cyclic phosphate intermediate; Section 28-1). However, the function of base methylation is unknown.

Eukaryotic rRNA Processing Resembles That of Prokaryotes

The eukaryotic genome typically has several hundred tandemly repeated copies of rRNA genes that are contained in small dark-staining nuclear bodies known as nucleoli (the site of rRNA transcription, processing, and ribosomal subunit assembly; Fig. 1-5). The primary rRNA transcript is an ~7500-nucleotide 45S RNA that contains, starting from the 5′ end, the 18S, 5.8S, and 28S rRNAs separated by spacer sequences (Fig. 29-39). In the first stage of its processing, 45S RNA is specifically methylated at ~110 sites that occur mostly in its rRNA sequences. About 80% of these modifications yield $O^{2'}$-methylribose residues and the remainder form methylated bases such as N^6,N^6-dimethyladenine and 2-methylguanine. The subsequent cleavage and trimming of the 45S RNA superfically resembles that of prokaryotic rRNAs. In fact, enzymes exhibiting RNAse III and RNase P-like activities occur in eukaryotes. The 5S eukaryotic rRNA is separately processed in a manner resembling that of tRNA (Section 29-4C).

Some Eukaryotic rRNA Genes Are Self-Splicing

Only a few eukaryotic rRNA genes contain introns. Nevertheless, Thomas Cech's study of how such genes are spliced in the ciliated protozoan *Tetrahymena thermophila* led to an astonishing discovery: *RNA can act as an enzyme. When the isolated pre-rRNA of this organism is incubated with guanosine or a free guanine nucleotide (GMP, GDP, or GTP), but in the absence of protein, its single 413-nucleotide intron excises itself and splices together its flanking exons; that is, this pre-rRNA is self-splicing.* The three-step reaction sequence of this process (Fig. 29-40) resembles that of mRNA splicing:

1. The 3′-OH group of the guanosine forms a phosphodiester bond with the intron's 5′ end.
2. The 3′-terminal OH group of the newly liberated 5′ exon forms a phosphodiester bond with the 5′-terminal phosphate of the 3′ exon thereby splicing together the two exons and releasing the intron.
3. The 3′-terminal OH group of the intron forms a phosphodiester bond with the phosphate of the nucleotide 15 residues from the intron's 5′ end, yielding the 5′-terminal fragment with the remainder of the intron in cyclic form.

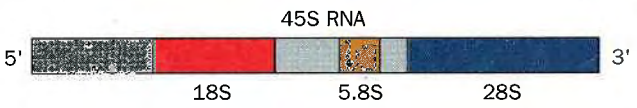


Figure 29-39
The organization of the 45S primary transcript of eukaryotic rRNA.

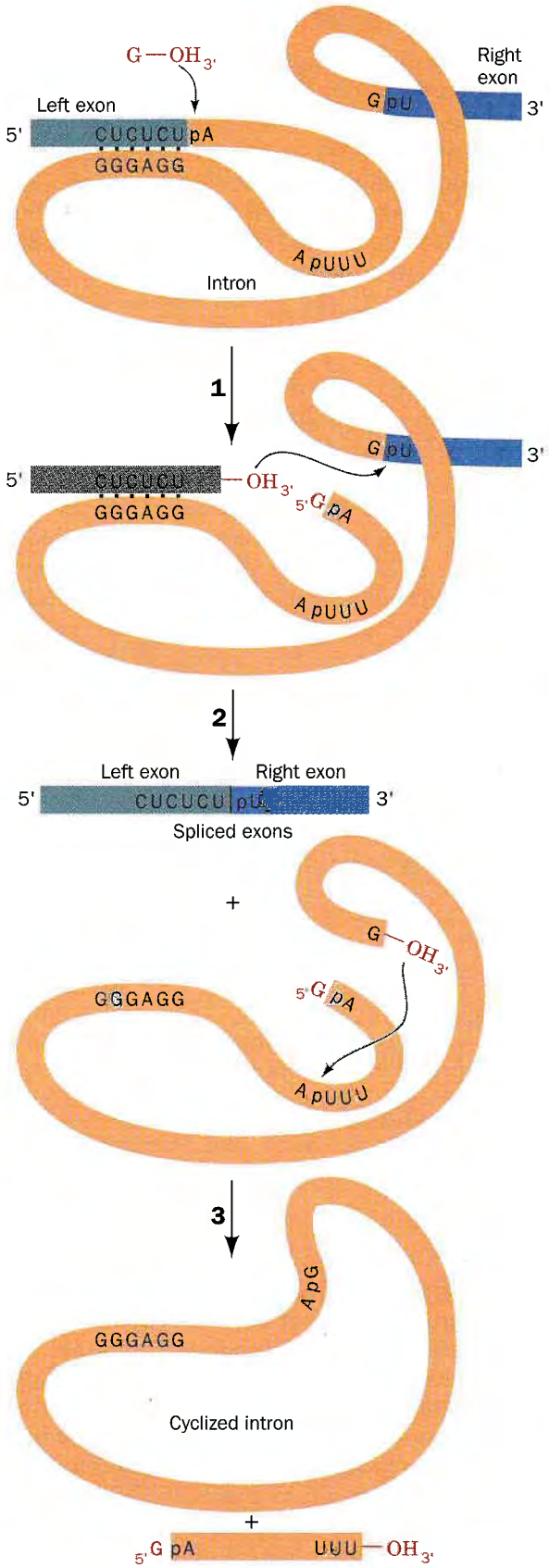


Figure 29-40
The sequence of reactions in the self-splicing of *Tetrahymena* pre-rRNA: **(1)** The 3′-OH group of a guanine nucleotide attacks the intron's 5′-terminal phosphate so as to form a phosphodiester bond and release the 5′ exon. **(2)** The newly generated 3′-OH group of the 5′ exon attacks the 5′-terminal phosphate of the 3′ exon thereby splicing the two exons and releasing the intron. **(3)** The 3′-OH group of the intron attacks the phosphate of the nucleotide that is 15 residues from the 5′ end so as to cyclize the intron and release its 5′-terminal fragment. Throughout this process, the RNA maintains a folded, internally hydrogen bonded conformation that permits the precise excision of the intron.

This self-splicing process, which similarly occurs in fungal mitochondrial rRNA, consists of a series of transesterifications and therefore does not require free energy input. Cech further established the enzymatic properties of the *Tetrahymena* intron, which presumably stem from its three-dimensional structure, by demonstrating that it catalyzes the *in vitro* cleavage of poly(C) with an enhancement factor of 10^{10} over the rate of spontaneous hydrolysis. Indeed, this RNA catalyst even exhibits Michaelis–Menton kinetics ($K_M = 42\ \mu M$ and $k_{cat} = 0.033\ s^{-1}$ for C_5). Such RNA enzymes have been named **ribozymes.**

Although the idea that an RNA can have enzymatic properties may be unorthodox, *there is no fundamental reason why an RNA, or any other macromolecule, cannot have catalytic activity.* Of course, in order to be an efficient catalyst, a macromolecule must be able to assume a stable structure but, as we shall see in Sections 30-2B and 3A, RNAs in the form of tRNA and most probably rRNA do just that. The chemical similarities of the mRNA and rRNA splicing reactions therefore suggest that splicosomes are ribozymal systems that evolved from primordial self-splicing RNAs and that their protein components merely serve to fine tune the ribozymes' structure and function. Similarly, the RNA components of ribosomes, which are more than one-half RNA and the rest protein, probably have catalytic functions in addition to the structural and recognition roles usually attributed to them (Section 30-3). Thus, the observations that cells contain batteries of enzymes for manipulating DNA but few for processing RNA, and that many coenzymes are ribonucleotides (e.g., ATP, NAD^+, and CoA), led to the hypothesis that *RNAs were the original biological catalysts in precellular times and that the chemically more versatile proteins were relative latecomers in macromolecular evolution (Section 1-4C).*

C. Transfer RNA Processing

tRNAs, as we discuss in Section 30-2A, consist of ~80 nucleotides that assume a secondary structure with four base paired stems known as the **cloverleaf structure**

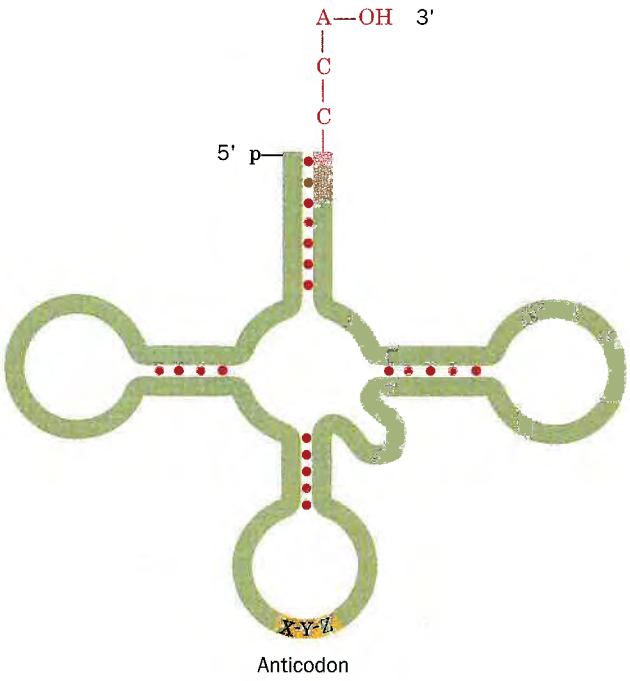


Figure 29-41
A schematic diagram of the tRNA cloverleaf secondary structure. Each dot indicates a base pair in the hydrogen bonded stems. The position of the anticodon triplet and the 3′-terminal —CCA are indicated.

(Fig. 29-41). All tRNAs have a large fraction of modified bases (whose structure, function, and synthesis is also considered in Section 30-2A) and each has the 3′-terminal sequence —CCA to which the corresponding amino acid is appended in the amino acid-charged tRNA. The **anticodon** (which is complementary to the codon specifying the tRNA's corresponding amino acid) occurs in the loop of the cloverleaf structure opposite the stem containing the terminal nucleotides.

The *E. coli* chromosome contains ~60 tRNA genes. Some of them are components of rRNA operons (Section 29-4A); the others are distributed, often in clusters, throughout the chromosome. The primary tRNA transcripts, which contain from one to as many as four or five identical tRNA species, have extra nucleotides at the 3′ and 5′ ends of each tRNA sequence. The excision and trimming of these tRNA sequences resembles that for *E. coli* rRNAs (Section 29-4B) in that both processes employ some of the same nucleases.

RNase P Is a Ribozyme

RNase P, which generates the 5′ ends of all *E. coli* tRNAs (Fig. 29-37), is a particularly interesting enzyme because it has a 377-nucleotide RNA component (~125 kD vs 14 kD for its protein subunit) that is essential for enzymatic activity. The enzyme's RNA was, quite understandably, first proposed to function in recognizing the substrate RNA through base pairing and to thereby guide the protein subunit, which was presumed to be the actual nuclease, to the cleavage site. However, Sidney Altman has shown that *the RNA component of RNase P is, in fact, the enzyme's catalytic subunit* by demonstrating that protein-free RNase P RNA catalyzes the cleavage of substrate RNA at high salt concentrations. RNase P protein, which is basic, evidently functions at physiological salt concentrations to electrostatically reduce the repulsions between the polyanionic ribozyme and substrate RNAs. The argument that trace quantities of RNase P protein are really responsible for the RNase P reaction was disposed of by showing that catalytic activity is exhibited by RNase P RNA that has been transcribed in a cell-free system. Thus we now have two independent examples of ribozymes.

Many Eukaryotic Pre-tRNAs Have Introns

Eukaryotic genomes contain from several hundred to several thousand tRNA genes. Many eukaryotic primary tRNA transcripts, for example, yeast $tRNA^{Tyr}$ (Fig. 29-42), contain a small intron adjacent to their anticodons as well as extra nucleotides at their 5′ and 3′ ends. Note that this intron is unlikely to disrupt the tRNA's cloverleaf structure. Eukaryotic tRNA transcripts lack the obligatory —CCA sequence at their 3′ end. This is appended to the immature tRNAs by the enzyme **tRNA nucleotidyltransferase,** which sequentially adds two C's and an A to tRNA using CTP and ATP as substrates. This enzyme also occurs in prokaryotes although, at least in *E. coli,* the tRNA genes all encode a —CCA terminus. The *E. coli* tRNA nucleotidyltransferase is therefore thought to function in the repair of degraded tRNAs.

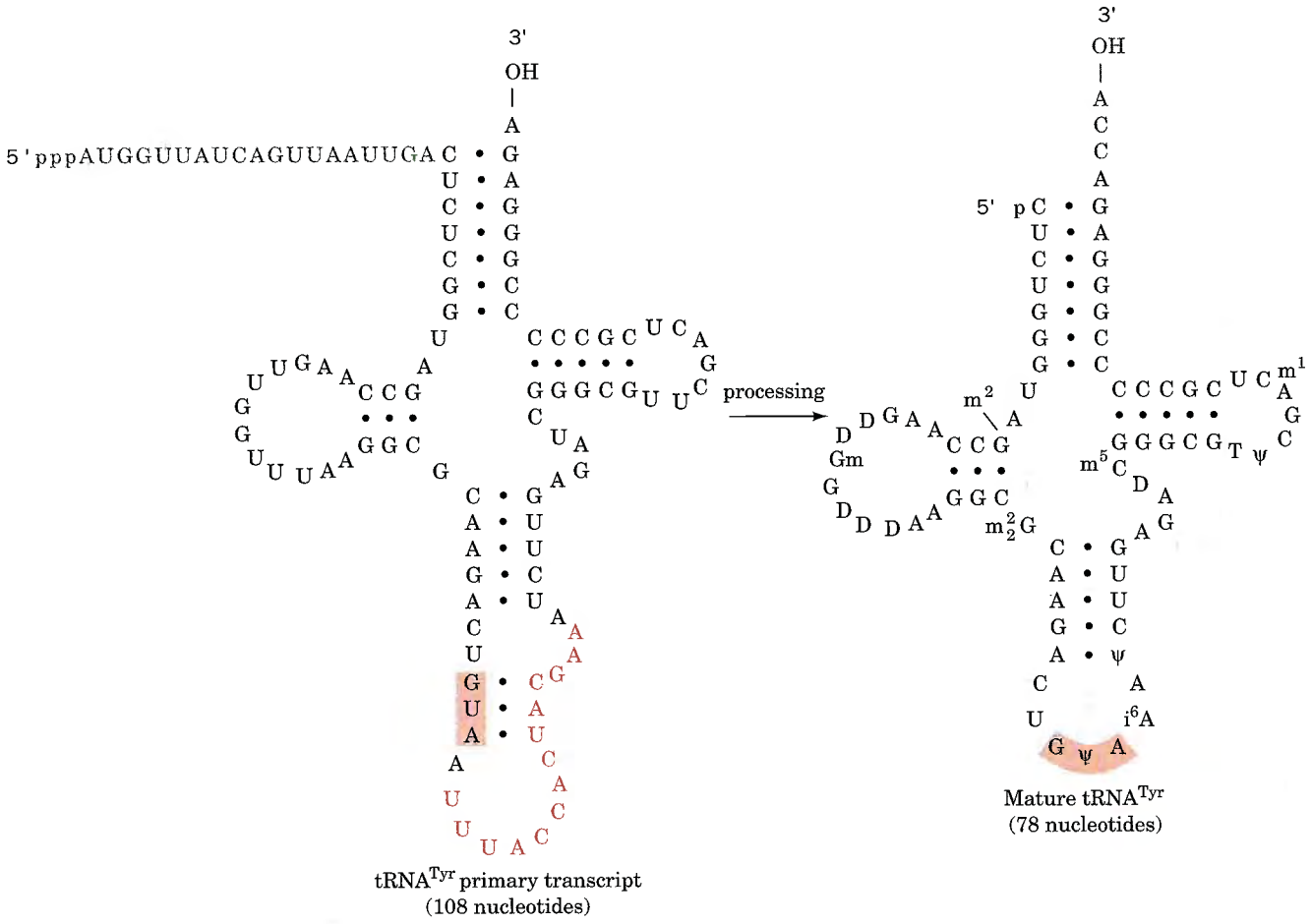


Figure 29-42
The post-transcriptional processing of yeast $tRNA^{Tyr}$. A 14-nucleotide intervening sequence and a 19-nucleotide 5′-terminal sequence are excised from the primary transcript, a —CCA is appended to the 3′ end and several of the bases are modified (their symbols are defined in Fig. 30-13) to form the mature tRNA. The anticodon is shaded. [After DeRobertis, E. M. and Olsen, M. V., *Nature* **278,** 142 (1989).]

Chapter Summary

The central dogma of molecular biology states that "DNA makes RNA makes protein" (although RNA can also "make" DNA). There is, however, enormous variation among the rates that the various proteins are made. Certain enzymes, such as those of the *lac* operon, are synthesized only when the substances they metabolize are present. The *lac* operon consists of the control sequences *lacP* and *lacO* followed by the tandemly arranged genes for β-galactosidase *(lacZ)*, galactoside permease *(lacY)*, and thiogalactoside transacetylase *(lacA)*. In the absence of inducer, physiologically allolactose, the *lac* repressor, the product of the *lacI* gene, binds to operator *(lacO)* so as to prevent the transcription of the *lac* operon by RNA polymerase. The binding of inducer causes the repressor to release the operator that allows the *lac* structural genes to be transcribed onto a single polycistronic mRNA. The mRNAs transiently associate with ribosomes so as to direct them to synthesize the encoded polypeptides.

The holoenzyme of *E. coli* RNA polymerase has the subunit structure $\alpha_2\beta\beta'\sigma$. It initiates transcription on the sense strand of a gene at a position designated by its promoter. The most conserved region of the promoter is the Pribnow box, which is centered at about the −10 position and has the consensus sequence TATAAT. The −35 region is also conserved in efficient promoters. Methylation protection studies indicate that holoenzyme forms an "open" initiation complex with the promoter. After the initiation of RNA synthesis, the σ subunit dissociates from the core enzyme, which then autonomously catalyzes chain elongation in the 5′ → 3′ direction. RNA synthesis is terminated by a segment of the transcript that forms a G + C-rich hairpin with an oligo(U) tail that spontaneously dissociates from the DNA. Termination sequences that lack these sequences require the assistance of rho factor for proper chain termination. In the nuclei of eukaryotic cells, RNA polymerases I, II, and III, respectively, synthesize rRNA precursors, hnRNA, and tRNAs + 5S RNA. The minimal RNA polymerase I promoter extends between nucleotides −7 and +6. Many RNA polymerase II promoters contain a conserved TATAAAA sequence, the TATA box, located around position −27. Enhancers are transcriptional activators that can have variable positions and orientations relative to the transcription start site. RNA polymerase III promoters are located within the transcribed regions of their gene between positions +40 and +80.

Prokaryotes can respond rapidly to environmental changes, in part, because the translation of mRNAs commences during their transcription and because most mRNAs are degraded within 1 to 3 min of their synthesis. The temporally ordered expression of sets of genes in some bacteriophages is controlled by cascades of σ factors. The *lac* repressor is a tetrameric protein of identical subunits that, in the absence of inducer, nonspecifically binds to duplex DNA but binds much more tightly to *lac* promoter. The promoter sequence that *lac* repressor protects from nuclease digestion has nearly palindromic symmetry. Yet, methylation protection and mutational studies indicate that repressor is not symmetrically bound to promoter. Repressor and RNA polymerase compete for the same promoter-binding sites.

The presence of glucose represses the transcription of operons specifying certain catabolic enzymes through the mediation of cAMP. Upon binding cAMP, which is only formed in the absence of glucose, catabolite gene activator protein (CAP) binds to the promoters of certain operons, such as the *lac* operon, thereby activating their transcription. CAP's two symmetry equivalent DNA-binding domains each bind in the major groove of their target DNA via a helix–turn–helix motif that occurs in numerous prokaryotic repressors. The binding between these repressors and their target DNAs is mediated by mutually favorable associations between these macromolecules rather than any specific interactions between base pairs and amino acid side chains analogous to Watson–Crick base pairing. *araBAD* transcription is controlled by CAP–cAMP and AraC through a remarkable complex of AraC to two-binding sites, $araO_2$ and *araI*, that forms a DNA loop. In this system, AraC also regulates its own synthesis by binding to the $araO_1$ site so as to repress the transcription of the *araC* gene. The expression of the *E. coli trp* operon is regulated both by attenuation as well as repression. Upon binding tryptophan, its corepressor, *trp* repressor binds to the *trp* operator thereby blocking *trp* operon transcription. When tryptophan is available, much of the *trp* transcript that has escaped repression is prematurely terminated in the *trpL* sequence because its transcript contains a segment that forms a normal terminator structure. When tryptophanyl–$tRNA^{Trp}$ is scarce, ribosomes stall at the transcript's two tandem Trp codons. This permits the newly synthesized RNA to form a base paired stem and loop that prevents the formation of the terminator structure. Several other operons are similarly regulated by attenuation. The stringent response is another mechanism by which *E. coli* match the rate of transcription to charged tRNA availability. When a specified charged tRNA is scarce, stringent factor on active ribosomes synthesizes ppGpp, which inhibits the transcription of rRNA and some mRNAs while stimulating the transcription of other mRNAs.

Prokaryotic mRNA transcripts require no additional processing. However, eukaryotic mRNAs have an enzymatically appended 5′ cap and, in most cases, an enzymatically generated poly(A) tail. Moreover, the introns of eukaryotic mRNA primary transcripts (hnRNAs) are precisely excised and their flanking exons are spliced together to form mature mRNAs in a snRNP-mediated process that takes place in splicosomes. The primary transcript of *E. coli* rRNAs contains all three rRNAs together with some tRNAs. These are excised and trimmed by specific endonucleases and exonucleases. The rRNAs are also modified by the methylation of specific nucleosides. The eukaryotic 18S, 5.8S, and 28S rRNAs are similarly transcribed as a 45S precursor which is processed in a manner resembling that of *E. coli* rRNAs. The intron of *Tetrahymena* pre-rRNA is removed in an RNA-catalyzed self-splicing reaction. Prokaryotic tRNAs are excised from their primary transcripts and trimmed in much the same manner as rRNAs. In RNase P, one of the enzymes mediating this process, the catalytic subunit is an RNA. Eukaryotic tRNA transcripts also require the excision of a short intron and the enzymatic addition of a 3′-terminal —CCA to form the mature tRNA.

References

General

Lewin, B., *Genes* (3rd ed.), Chapters 8–11 and 22–24, Wiley (1987).

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M., *Molecular Biology of the Gene* (4th ed.), Chapters 16 and 20, Benjamin/Cummings (1987).

The Genetic Role of RNA

Brenner, S., Jacob, F., and Meselson, M., An unstable intermediate carrying information from genes to ribosomes for protein synthesis, *Nature* **190,** 576–581 (1960). [The experimental verification of mRNA's existence.]

Brachet, J., Reminiscences about nucleic acid cytochemistry and biochemistry, *Trends Biochem. Sci.* **12,** 244–246 (1987).

Crick, F., Central dogma of molecular biology, *Nature* **227,** 561–563 (1970).

Hall, B. D. and Spiegelman, S., Sequence complementarity of T2-DNA and T2-specific RNA, *Proc. Natl. Acad. Sci.* **47,** 137–146 (1964). [The first use of RNA–DNA hybridization.]

Jacob, F. and Monod, J., Genetic regulatory mechanisms in the synthesis of proteins, *J. Mol. Biol.* **3,** 318–356 (1961). [The classic paper postulating the existence of mRNA and operons and explaining how the transcription of operons is regulated.]

Spiegelman, S., Hybrid nucleic acids, *Sci. Am.* **210**(5): 48–56 (1964).

RNA Polymerase and mRNA

Adhya, S. and Gottesman, M., Control of transcription termination, *Annu. Rev. Biochem.* **47,** 967–996 (1978).

Bear, D. G. and Peabody, D. S., The *E. coli* rho protein: an ATPase that terminates transcription, *Trends Biochem. Sci.* **13,** 343–347 (1988).

Berman, H. M. and Young, P. R., The interaction of intercalculating drugs with nucleic adids, *Annu. Rev. Biophys. Bioeng.* **10,** 87–114 (1981).

Chamberlain, M. J., Bacterial DNA-dependent RNA polymerases, *in* Boyer, P. D. (Ed.), *The Enzymes* (3rd ed.), Vol. 15, *pp.* 61–86, Academic Press (1982).

Futcher, B., Supercoiling and transcription, or vice versa? *Trends Genet.* **4,** 271–272 (1988).

Gannan, F., O'Hare, K., Perrin, F., LePennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., and Chambon, P., Organization and sequences of the 5′ end of a cloned complete ovalbumin gene, *Nature* **278,** 428–434 (1979).

Gale, E. F., Cundliffe, E., Reynolds, P. E., Richmond, M. H., and Waring, M. J., *The Molecular Basis of Antibiotic Action* (2nd ed.), Chapter 5, Wiley (1981).

Geiduschek, E. P. and Tocchini-Valentini, G. P., Transcription by RNA polymerase III, *Annu. Rev. Biochem.* **57,** 873–914 (1988).

Hansen, U. and Sharp, P. A., Transcription by RNA polymerase II, *in* Fraenkel-Conrat, H. and Wagner, R. R. (Eds.), *Comprehensive Virology,* Vol. 19, *pp.* 65–97, Plenum Press (1984).

Khoury, G. and Gruss, P., Enhancer elements, *Cell* **33,** 313–314 (1983).

Lewis, M. K. and Burgess, R. R., Eukaryotic RNA polymerases, *in* Boyer, P. D. (Ed.), *The Enzymes* (3rd ed.), Vol. 15, *pp.* 110–153, Academic Press (1982).

Losick, R. and Chamberlain, M. (Eds.), *RNA Polymerase,* Cold Spring Harbor Laboratory (1976). [A series of authorative articles on various aspects of RNA polymerase.]

McKnight, S. L. and Kingsbury, R., Transcriptional control signals of a eukaryotic protein-coding gene, *Science* **217,** 316–324 (1982).

Paule, M. R. Comparitive subunit composition of eukaryotic nuclear RNA polymerases, *Trends Biochem. Sci.* **6,** 128–131 (1981).

Platt, T., Transcription termination and the regulation of gene expression, *Annu. Rev. Biochem.* **55,** 339–372 (1986).

Pribnow, D., Genetic control signals in DNA, *in* Goldberger, R. F. (Ed.), *Biological Regulation and Development,* Vol. 1, *pp.* 217–277, Plenum Press (1979).

Rosenberg, M. and Court, D., Regulatory sequences involved in the promotion and termination of RNA transcription, *Annu. Rev. Genet.* **13,** 319–353 (1979).

Sakonju, S., Bogenhagen, D. F., and Brown, D. D., A control region in the center of the 5S RNA gene directs specific initiation of transcription: I. The 5′ border of the region; *and* II. The 3′ border of the region, *Cell* **19,** 13–25, 27–35 (1980).

Sentenac, A., Eukaryotic RNA polymerases, *CRC Crit. Rev. Biochem.* **18,** 31–90 (1985).

Siebenlist, U., RNA polymerase unwinds an 11-base pair segment of phage T7 promoter, *Nature* **279,** 651–652 (1979).

Sollner-Webb, B., Wilkinson, J. A. K., Roan, J., and Reeder, R. H., Nested control regions promote Xenopus ribosomal RNA synthesis by RNA polymerase I, *Cell* **35,** 199–206 (1983).

Wang, A. H.-J., Ughetto, G., Quigley, G. J., and Rich, A., Interactions between an anthracycline antibiotic and DNA: molecular structure of daunomycin complexed to d(CpGpTpApCpG) at 1.2 Å resolution, *Biochemistry* **26,** 1152–1163 (1987).

Control of Transcription

Aggarwal, A. K., Rogers, D. W., Drottar, M., Ptashne, M., and Harrison, S. C., Recognition of the DNA operator by the repressor of phage 434: a view at high resolution, *Science* **242,** 899–907 (1988); *and* Anderson, J. E., Ptashne, M., and Harrison, S. C., The structure of the repressor–operator complex of bacteriophage 434, *Nature* **326,** 846–852 (1987).

Brennan, R. G. and Matthews, B. W., The helix–turn–helix DNA binding motif, *J. Biol. Chem.* **264,** 1903–1906 (1989).

de Crombrugghe, B., Busby, S., and Buc, H., Cyclic AMP receptor protein: role in transcription activation, *Science* **244,** 831–838 (1984).

Dickson, R. C., Abelson, J., Barnes, W. M., and Reznikoff, W. S., Genetic regulation: The lac control region, *Science* **187,** 27–35 (1975).

Dunn, T. M., Hahn, S., Ogden, S., and Schleif, R. F., An operator at −280 base pairs that is required for repression of *araBAD*

operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression, *Proc. Natl. Acad. Sci.* **81,** 5017–5020 (1984).

Friedman, D. I., Imperiale, M. J., and Adhya, S. L., RNA 3′ end formation in the control of gene expression, *Annu. Rev. Genet.* **21,** 453–488 (1987).

Gallant, J. A., Stringent control in *E. coli, Annu. Rev. Genet.* **13,** 393–415 (1979).

Gilbert, W. and Müller-Hill, B., Isolation of the lac repressor, *Proc. Natl. Acad. Sci.* **56,** 1891–1898 (1966).

Gralla, J.D., Specific repression in the *lac* repressor—the 1988 version, *in* Gralla, J.D. (Ed.), *DNA–Protein Interactions in Transcription, pp.* 3–10, Liss (1989).

Helmann, J. D. and Chamberlin, M. J., Structure and function of bacterial sigma factors, *Annu. Rev. Biochem.* **57,** 839–872 (1988).

Kolter, R. and Yanofsky, C., Attenuation in amino acid biosynthetic operons, *Annu. Rev. Genet.* **16,** 113–134 (1982).

Lamond, A. I. and Travers, A. A., Stringent control of bacterial transcription, *Cell* **41,** 6–8 (1985).

Losick, R. and Pero. J., Cascades of sigma factors, *Cell* **25,** 582–584 (1981).

McClure, W. R., Mechanism and control of transcription initiation in prokaryotes, *Annu. Rev. Biochem.* **54,** 171–204 (1985).

Miller, J. H. and Reznikoff, W. S. (Eds.), *The Operon,* Cold Spring Harbor Laboratory (1978). [An informative collection of reviews on the *lac* operon as well as other operons.]

Otwinowski, Z., Schevitz, R. W., Zhang, R. -G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B., Crystal structure of *trp* repressor/operator complex at atomic resolution, *Nature* **335,** 321–329 (1988).

Pabo C. O. and Sauer, R. T., Protein–DNA recognition, *Annu. Rev. Biochem.* **53,** 293–321 (1984).

Pastan, I. and Adhya, S., Cyclic adenosine 5′-monophosphate in *Escherichia coli, Bacteriol. Rev.* **40,** 527–551 (1976).

Reznikoff, W. S., Siegele, D. A., Cowing, D. W., and Gross, C. A., The regulation of transcription initiation in bacteria, *Annu. Rev. Genet.* **19,** 355–387 (1985).

Raibaud, O. and Schwartz, O., Positive control of transcription in bacteria, *Annu. Rev. Genet.* **18,** 173–206 (1984).

Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F., and Matthews, B. W., Structural similarity in the DNA-binding domains of catabolite gene activator and *cro* repressor proteins, *Proc. Natl. Acad. Sci.* **79,** 3097–3100 (1982).

von Hippel, P. H., Bear, D. G., Morgan, W. D., and McSwiggen, J. A., Protein-nucleic acid interactions in transcription: a molecular analysis, *Annu. Rev. Biochem.* **53,** 389–346 (1984).

Weber, I. T. and Steitz, T. A., Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity, *Proc. Natl. Acad. Sci.* **81,** 3973–3977 (1984).

Wolberger, C., Dong, Y., Ptashne, M., and Harrison, S. C., Structure of phage 434 Cro/DNA complex, *Nature* **335,** 789–795 (1988).

Yanofsky, C., Transcription attenutation, *J. Biol. Chem.* **263,** 609–612 (1988).

Yanofsky, C., Attenuation in the control of expression of bacterial operons, *Nature* **289,** 751–758 (1981).

Post-Transcriptional Processing

Altman, S., Baer, M., Guerrier-Takada, C., and Vioque, A., Enzymatic cleavage of RNA by RNA, *Trends Biochem. Sci.* **11,** 515–518 (1986). [Discusses RNase P.]

Banerjee, A. K., 5′-Terminal cap structure in eukaryotic messenger ribonucleic acids, *Microbiol. Rev.* **44,** 175–205 (1980).

Birnstiel, M. L., Busslinger, M., and Strub, K., Transcription termination and 3′ processing: the end is in sight! *Cell* **41,** 349–359 (1985).

Boyer, P. D. (Ed.), *The Enzymes* (3rd ed.), Vol. 15, Academic Press (1982). [Contains articles on RNA processing enzymes.]

Breathnach, R. and Chambon, P., Organization and expression of eucaryotic split genes coding for proteins, *Annu. Rev. Biochem.* **50,** 349–383 (1981).

Cech, T. R., The chemistry of self-splicing RNA and RNA enzymes, *Science* **236,** 1532–1539 (1987).

Cech, T. R. and Bass, B. L., Biological catalysis by RNA, *Annu. Rev. Biochem.* **55,** 599–629 (1986).

Cech, T. R., RNA as an enzyme, *Sci. Am.* **255**(5): 64–75 (1986).

Chambon, P., Split genes, *Sci. Am.* **244**(5): 60–71 (1981).

Darnell, J. E., Jr., RNA, *Sci. Am.* **253**(4): 68–78 (1985).

Darnell, J. E., Jr., The processing of RNA, *Sci. Am.* **249**(4): 90–100 (1983).

Deutscher, M. P., The metabolic role of RNases, *Trends Biochem. Sci.* **13,** 137–138 (1988).

Gegenheimer, P. and Apiron, D., Processing of prokaryotic ribonucleic acid, *Microbiol. Rev.* **45,** 502–541 (1981).

Guthrie, C. and Patterson, B., Splicosomal snRNAs, *Annu. Rev. Genet.* **22,** 387–419 (1988).

Humphrey, T. and Proudfoot, N. J., A beginning to the biochemistry of polyadenylation, *Trends Genet.* **4,** 243–245 (1988).

Littauer, U. Z. and Soreq, H., The regulatory function of poly(A) and adjacent 3′ sequences in translated RNA, *Prog. Nucleic Acid Res. Biol.* **27,** 53–83 (1982).

Maniatis, T. and Reed, R., The role of small ribonucleoprotein particles in pre-mRNA splicing, *Nature* **325,** 673–678 (1987).

Mattaj, I. W., snRNAs: from gene architecture to RNA processing, *Trends Biochem. Sci.* **9,** 435–437 (1984).

Mowry, K. L. and Steitz, J. A., snRNP mediators of 3′ end processing: functional fossils?, *Trends Biochem. Sci.* **13,** 447–451 (1988).

Ogden, R. C., Knapp, G., Peebles, C. L., Johnson, J., and Abelson, J., The mechanism of tRNA splicing, *Trends Biochem. Sci.* **6,** 154–158 (1981).

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A., Splicing of messenger RNA precursors, *Annu. Rev. Biochem.* **55,** 1119–1150 (1986).

Proudfoot, N. J., How RNA polymerase II terminates transcription in higher eukaryotes, *Trends Biochem. Sci.* **14,** 105–110 (1989).

Sharp, P. A., Splicing of messenger RNA precursors, *Science* **235,** 766–771 (1987).

Söll, D., Abelson, J. N., and Schimmel, P. R. (Eds.), *Transfer RNA: Biological Aspects,* Cold Spring Harbor Laboratory (1980). [Contains several articles on the processing of tRNA and rRNA.]

Steitz, J. A., "Snurps", *Sci. Am.* **258**(6): 56–63 (1988).

Problems

1. Indicate the phenotypes of the following *E. coli lac* partial diploids in terms of inducibility and active enzymes synthesized.
   (a) $I^-P^+O^+Z^+Y^-/I^+P^-O^+Z^+Y^+$
   (b) $I^-P^+O^c\,Z^+Y^-/I^+P^+O^+Z^-Y^+$
   (c) $I^-P^+O^c\,Z^+Y^+/I^-P^+O^+Z^+Y^+$
   (d) $I^+P^-O^c\,Z^+Y^+/I^-P^+O^c\,Z^-Y^-$

2. **Superrepressed** mutants, I^S, *encode lac* repressors that bind operator but do not respond to the presence of inducer. Indicate the phenotypes of the following genotypes in terms of inducibility and enzyme production.
   (a) $I^SO^+Z^+$ (b) $I^SO^cZ^+$ (c) $I^+O^+Z^+/I^SO^+Z^+$

*3. Why do *lacZ*$^-$ *E. coli* fail to show galactoside permease activity after the addition of lactose in the absence of glucose? Why do *lac Y*$^-$ mutants lack β-galactosidase activity under the same conditions?

4. What is the experimental advantage of using IPTG instead of 1,6-allolactose as an inducer of the *lac* operon?

5. Indicate the Pribnow box, −35 region and initiating nucleotide on the antisense strand of the *E. coli* $tRNA^{Tyr}$ promoter shown below.

```
5' CAACGTAACACTTTACAGCGGCGCGTCATTTGATATGATGCGCCCCGCTTCCCGATA 3'
3' GTTGCATTGTGAAATGTCGCCGCGCAGTAAACTATACTACGCGGGGCGAAGGGCTAT 5'
```

*6. Why are *E. coli* that are diploid for rifamycin resistance and rifamycin sensitivity (rif^R/rif^S) sensitive to rifamycin?

7. What is the probability that the 4026-nucleotide DNA sequence coding for the β subunit of *E. coli* RNA polymerase will be transcribed with the correct base sequence. Perform the calculations for the probabilities of 0.0001, 0.001, and 0.01 that each base is incorrectly transcribed.

8. What is the probability that the symmetry of the *lac* operator is merely accidental?

9. Why does the inhibition of DNA gyrase in *E. coli* inhibit the expression of catabolite sensitive operons?

10. Describe the transcription of the *trp* operon in the absence of active ribosomes and tryptophan.

11. Why can't eukaryotic transcription be regulated by attenuation?

12. Charles Yanofsky and his associates have synthesized a 15-nucleotide RNA that is complementary to segment 1 of *trpL* mRNa (but only partially complementary to segment 3). What is its effect on the *in vitro* transcription of *trp* operon? What is its effect if the *trpL* gene contains a mutation in segment 2 that destablizes the 2·3 stem and loop?

13. Why are *relA*$^-$ mutants defective in the *in vivo* transcription of the *his* and *trp* operons?

14. Why aren't primary rRNA transcripts observed in wild-type *E. coli?*

Chapter 30

TRANSLATION

In this chapter we consider **translation,** the mRNA-directed biosynthesis of polypeptides. Although peptide bond formation is a relatively simple reaction, the complexity of the translational process, which involves the coordinated participation of over 100 macromolecules, is mandated by the need to link 20 different amino acid residues accurately in the order specified by a particular mRNA.

We begin by considering the **genetic code,** the correspondence between nucleic acid sequences and polypeptide sequences. Next, we examine the structures and properties of **tRNAs,** the amino acid-bearing entities that mediate the translation process. Following this, we take up what is known about **ribosomes,** the complex molecular machines that catalyze peptide bond formation between the mRNA-specified amino acids. Peptide bond formation, however, does not necessarily yield a

functional protein; many polypeptides must first be post-translationally modified as we discuss in the subsequent section. We then study how cells degrade proteins, a process that must balance protein synthesis, and finally, consider the nonribosomal synthesis of certain small and unusual polypeptides.

1. THE GENETIC CODE

How does DNA encode genetic information? According to the one gene–one polypeptide hypothesis, the genetic message dictates the amino acid sequences of proteins. Since the base sequence of DNA is the only variable element in this otherwise monotonously repeating polymer, the amino acid sequence of a protein must somehow be specified by the base sequence of the corresponding segment of DNA.

A DNA base sequence might specify an amino acid sequence in many conceivable ways. With only 4 bases to code for 20 amino acids, a group of several bases, termed a **codon,** is necessary to specify a single amino acid. A triplet code, that is, one with 3 bases per codon, is minimally required since there are $4^3 = 64$ different triplets of bases whereas there can be only $4^2 = 16$ different doublets, which is insufficient to specify all the amino acids. In a triplet code, as many as 44 codons might not code for amino acids. On the other hand, many amino acids could be specified by more than one codon. Such a code, in a term borrowed from mathematics, is said to be **degenerate.**

Another mystery was, how does the polypeptide synthesizing apparatus group DNA's continuous sequence of bases into codons. For example, the code might be overlapping; that is, in the sequence

ABCDEFGHIJ · · ·

ABC might code for one amino acid, BCD for a second, CDE for a third, *etc.* Alternatively, the code might be nonoverlapping so that ABC specifies one amino acid, DEF a second, HIJ a third, *etc.* The code might also contain internal "punctuation" such as in the nonoverlapping triplet code

ABC,DEF,GHI, · · ·

in which the commas represent particular bases or base sequences. A related question is how does the genetic code specify the beginning and the end of a polypeptide chain.

The genetic code is, in fact, a nonoverlapping, comma-free, degenerate, triplet code. How this was determined and how the genetic code dictionary was elucidated is the subject of this section.

A. Chemical Mutagenesis

The triplet character of the genetic code, as we shall see below, was established through the use of **chemical mutagens,** substances that induce mutations. We therefore precede our study of the genetic code with a discussion of these substances. There are two major classes of mutations:

1. **Point mutations,** in which one base pair replaces another. These are subclassified as:
   (a) **Transitions,** in which one purine (or pyrimidine) is replaced by another.
   (b) **Transversions,** in which a purine is replaced by a pyrimidine or *vice versa.*
2. **Insertion/deletion mutations,** in which one or more nucleotide pairs are inserted in or deleted from DNA.

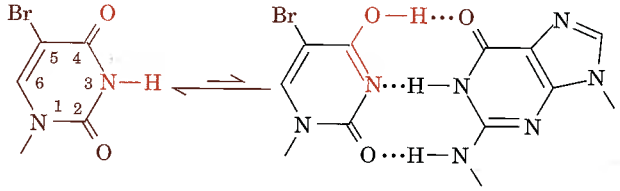
A mutation in any of these three categories may be reversed by a subsequent mutation of the same but not another category.

Point Mutations Are Generated by Altered Bases

Point mutations can result from the treatment of an organism with base analogs or substances that chemically alter bases. For example, the base analog **5-bromouracil (5BU)** sterically resembles thymine (5-methyluracil) but, through the influence of its electronegative Br atom, frequently assumes a tautomeric form that base pairs with guanine instead of adenine (Fig. 30-1). Consequently, when 5BU is incorporated into DNA in place of thymine, as it usually is, it occasionally induces an A·T → G·C transition in subsequent rounds of DNA replication. Occasionally, 5BU is also incorporated into DNA in place of cytosine, which instead generates a G·C → A·T transition.

The adenine analog **2-aminopurine (2AP),** normally base pairs with thymine (Fig. 30-2*a*) but occasionally forms an undistorted but singly hydrogen bonded base pair with cytosine (Fig. 30-2*b*). Thus 2AP also generates A·T → G·C and G·C → A·T transitions.

In aqueous solutions, **nitrous acid** (HNO_2) oxidatively deaminates aromatic primary amines so that it converts cytosine to uracil (Fig. 30-3*a*) and adenine to the guanine-like **hypoxanthine** (which forms two of guanine's three hydrogen bonds with cytosine; Fig.



5-Bromouracil (5BU) (keto tautomer) **5BU (enol tautomer)** **Guanine**

Figure 30-1
The keto form of 5-bromouracil (*left*) is its most common tautomer. However, it frequently assumes the enol form (*right*), which base pairs with guanine.

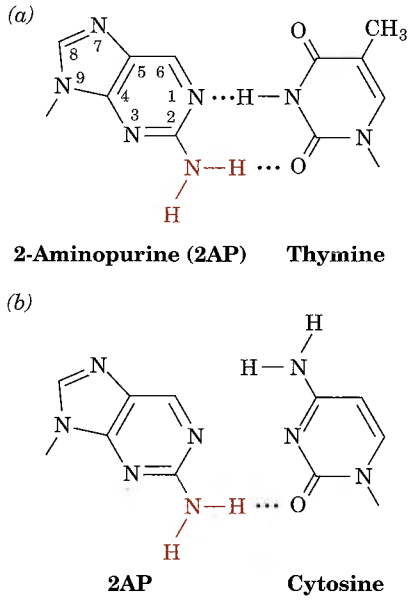


Figure 30-2
The adenine analog 2-aminopurine normally base pairs with (*a*) thymine but occasionally also does so with (*b*) cytosine.

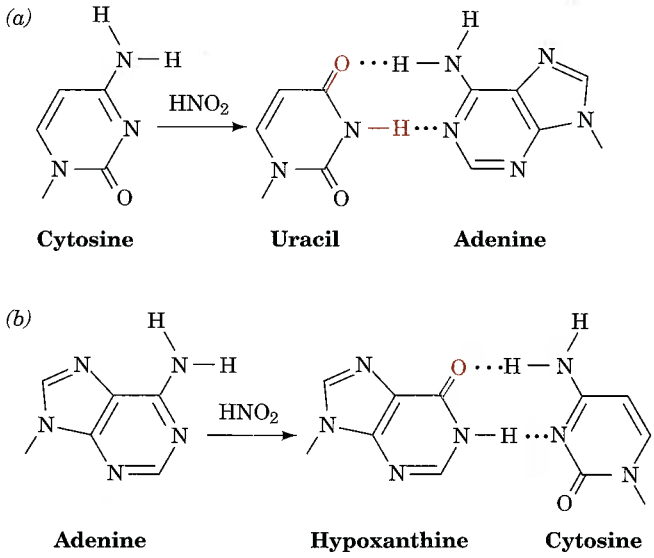
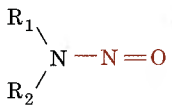


Figure 30-3
Reaction with nitrous acid converts (*a*) cytosine to uracil which base pairs with adenine; and (*b*) adenine to hypoxanthine, a guanine derivative (it lacks guanine's 2-amino group) which base pairs with cytosine.

30-3*b*). Hence, treatment of DNA with nitrous acid, or compounds such as **nitrosamines**

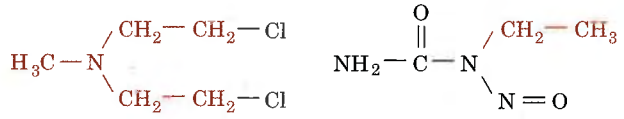


Nitrosoamines

that react to form nitrous acid, results in both A·T → G·C and G·C → A·T transitions.

> **Nitrite,** the conjugate base of nitrous acid, has long been used as a preservative of prepared meats such as frankfurters. However, the observation that many mutagens are also carcinogens (Section 31-5E) suggests that the consumption of nitrite-containing meat is harmful to humans. Proponents of nitrite preservation nevertheless argue that to stop it would result in far more fatalities. This is because lack of such treatment would greatly increase the incidence of **botulism,** an often fatal form of food poisoning caused by the ingestion of protein neurotoxins secreted by the anaerobic bacterium *Clostridium botulinum* (Section 34-4C).

Hydroxylamine (NH_2OH) also induces G·C → A·T transitions by specifically reacting with cytosine to convert it to a compound that base pairs with adenine (Fig. 30-4). The use of alkylating agents such as dimethyl sulfate, **nitrogen mustard,** and **ethylnitrosourea**



Nitrogen mustard Ethylnitrosourea

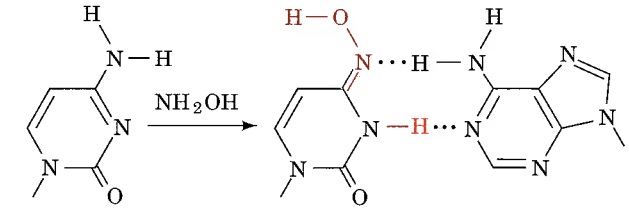


Figure 30-4
Reaction with hydroxylamine converts cytosine to a derivative which base pairs with adenine.

often generates transversions. The alkylation of the N(7) position of a purine nucleotide causes its subsequent depurination in a reaction similar to that diagrammed in Fig. 28-52*a*. The resulting gap in the sequence is filled in by an error-prone enzymatic repair system (Section 31-5B). Transversions arise when the missing purine is replaced by a pyrimidine. The enzymatic repair of DNA that has been damaged by UV radiation may also generate transversions.

Insertion/Deletion Mutations Are Generated by Intercalating Agents

Insertion/deletion mutations may arise from the treatment of DNA with intercalating agents such as acridine orange or proflavin (Section 28-4C). The distance between two consecutive base pairs is doubled by the intercalation of such a molecule between them. The replication of such a distorted DNA occasionally results in the insertion or deletion of one or more nucleotides in the

newly synthesized polynucleotide. (Insertions and deletions of large DNA segments generally arise from aberrant cross-over events; Section 33-2C.)

B. Codons Are Triplets

In 1961, Francis Crick and Sydney Brenner, through genetic invesitgations into the previously unknown character of proflavin-induced mutations, determined the triplet character of the genetic code. In bacteriophage T4, a particular proflavin-induced mutation, designated *FC*0, maps in the *rIIB* cistron (Section 27-1E). The growth of this mutant phage on a permissive host (*E. coli* B) resulted in the occasional spontaneous appearance of phenotypically wild-type phages as was demonstrated by their ability to grow on a restrictive host [*E. coli* K12(λ); recall that *rIIB* mutants form characteristically large plaques on *E. coli* B but cannot lyse *E. coli* K12(λ)]. Yet, these doubly mutated phages are not genotypically wild-type; the simultaneous infection of a permissive host by one of them and true wild-type phage yielded recombinant progeny that have either the *FC*0 mutation or a new mutation designated *FC*1. Thus the phenotypically wild-type phage is a double mutant that actually contains both *FC*0 and *FC*1. *These two genes are therefore* **suppressors** *of one another; that is, they cancel each other's mutant properties.* Furthermore, since they map together in the *rIIB* cistron, they are mutual **intragenic suppressors** (suppressors in the same gene).

The treatment of *FC*1 in a manner identical to that described for *FC*0 provided similar results: the appearance of a new mutant, *FC*2, that is an intragenic suppressor of *FC*1. By proceeding in this iterative manner, Crick and Brenner collected a series of different *rIIB* mutants, *FC*3, *FC*4, *FC*5, *etc.*, in which each mutant *FC*(n) is an intragenic suppressor of its predecessor, *FC*($n-1$). Recombination studies showed, moreover, that odd numbered mutations are intragenic suppressors of even numbered mutations but neither pairs of different odd numbered mutations nor pairs of different even numbered mutations suppress each other. However, recombinants containing three odd-numbered mutations or three even-numbered mutations all are phenotypically wild-type.

Crick and Brenner accounted for these observations by the following set of assumptions:

1. The proflavin-induced mutation *FC*0 is either an insertion or a deletion of one nucleotide pair from the *rIIB* cistron. If it is a deletion then *FC*1 is an insertion, *FC*2 is a deletion, *etc.*, and vice versa.
2. *The code is read in a sequential manner starting from a fixed point in the gene.* The insertion or deletion of a nucleotide shifts the **frame** (grouping) in which succeeding nucleotides are read as codons (insertions or deletions of nucleotides are therefore also known as **frameshift mutations**). Thus the code has no internal punctuation that indicates the reading frame; that is, *the code is comma-free.*
3. *The code is a triplet code.*
4. All or nearly all of the 64 triplet codons code for an amino acid; that is, *the code is degenerate.*

These principles are illustrated by the following analogy. Consider a sentence (gene) in which the words (codons) each consist of 3 letters (bases).

THE BIG RED FOX ATE THE EGG

(Here the spaces separating the words have no physical significance; they are only present to indicate the reading frame.) The deletion of the 4th letter, which shifts the reading frame, changes the sentence to

THE IGR EDF OXA TET HEE GG

so that all words past the point of deletion are unintelligible (specify the wrong amino acids). An insertion of any letter, however, say an X in the 9th position,

THE IGR EDX FOX ATE THE EGG

restores the original reading frame. Consequently, only the words between the two changes (mutations) are altered. As in this example, such a sentence might still be intelligible (the gene could still specify a functional protein), particularly if the changes are close together. Two deletions or two insertions, no matter how close together, would not suppress each other but just shift the reading frame. However, three insertions, say X, Y, and Z in the 5th, 8th, and 12th positions, respectively, would change the sentence to

THE BXI GYR EDZ FOX ATE THE EGG

which, after the third insertion, restores the original reading frame. The same would be true of three deletions. As before, if all three changes were close together, the sentence might still retain its meaning.

Crick and Brenner did not unambiguously demonstrate that the genetic code is a triplet code because they had no proof that their insertions and deletions involved only single nucleotides. Strictly speaking, they showed that a codon consists of $3r$ nucleotides where r is the number of nucleotides in an insertion or deletion. Although it was generally assumed at the time that $r = 1$, proof of this assertion had to await the elucidation of the genetic code (Section 30-1D).

C. Genes Are Colinear with Their Specified Polypeptides

In the early 1960s, Charles Yanofsky demonstrated the colinearity of genes and polypeptides. He did so by isolating a number of mutants of the 268-residue α

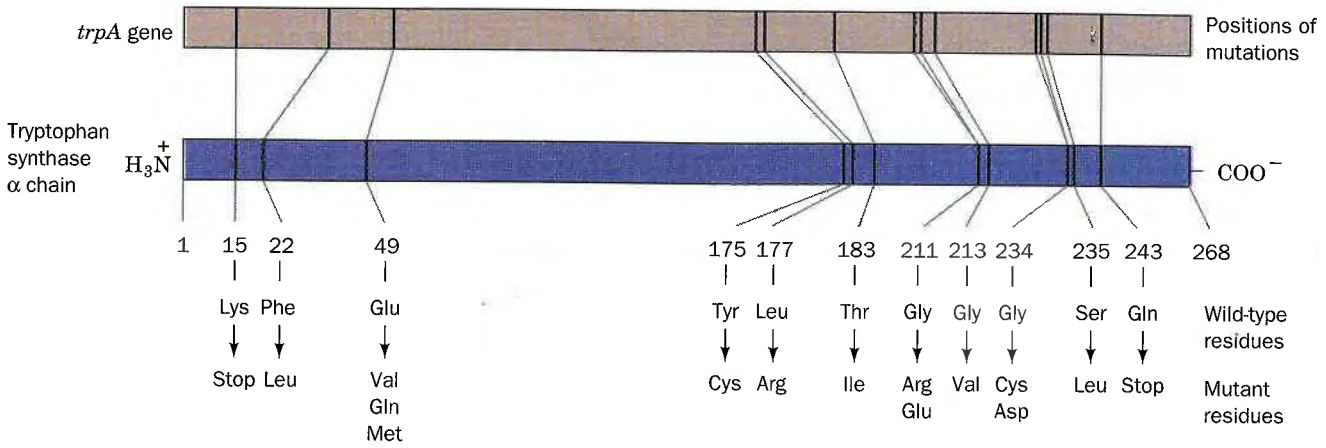


Figure 30-5
The colinearity of the *E. coli trpA* gene with the polypeptide it specifies, the tryptophan synthase α chain. The gene mutation positions, as determined by transductional mapping, have the same order as the corresponding amino acid changes in the polypeptide as determined by fingerprinting.

chain of *E. coli* **tryptophan synthase** (specified by the *trpA* gene; Section 29-3E). The genetic map of these mutants was elucidated by transductional mapping (Section 27-1E) and the amino acid changes to which they give rise were established by fingerprinting (Section 6-1J). The order of the mutants in the gene is the same as the order of the corresponding amino acid changes in the protein (Fig. 30-5). *The E. coli trpA gene is therefore colinear with the polypeptide it specifies.*

D. Deciphering the Genetic Code

In order to understand how the genetic code dictionary was elucidated we must first preview how proteins are synthesized. The mRNAs cannot directly recognize amino acids. Rather, *they specifically bind molecules of tRNA that each carry a corresponding amino acid (Fig. 30-6).* Each tRNA contains a trinucleotide sequence, its **anticodon,** which is complementary to an mRNA codon specifying the tRNA's amino acid. An amino acid is covalently linked to its corresponding tRNA through the action of a specific enzyme that recognizes both of these molecules (a process called "charging" the tRNA). During translation, the mRNA passes through the ribosome such that each codon, in turn, binds its corresponding charged tRNA (Fig. 30-7). As this occurs, the ribosome transfers the tRNA's appended amino acid to the end of the growing polypeptide chain.

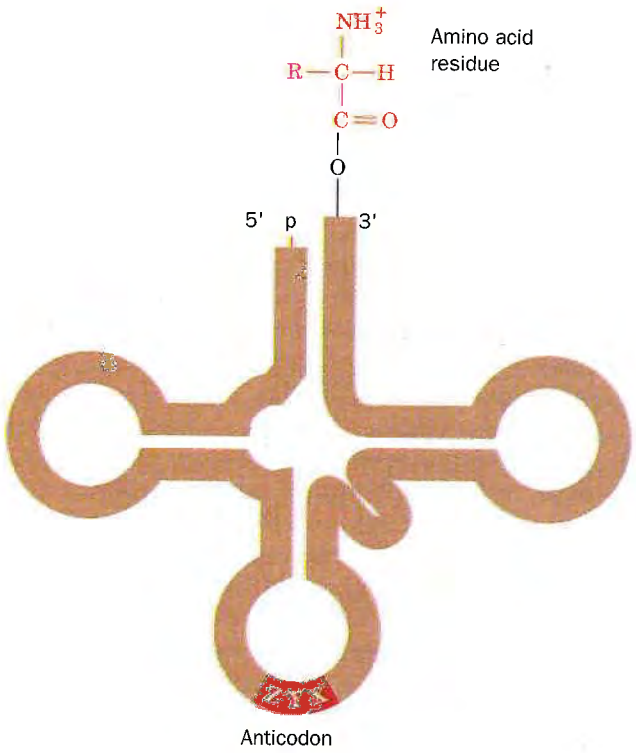


Figure 30-6
Transfer RNA in its "cloverleaf" form showing its covalently linked amino acid residue (*top*) and its anticodon (*bottom;* a trinucleotide segment that base pairs with the complementary mRNA codon during translation).

UUU Specifies Phe

The genetic code could, in principle, be determined by simply comparing the base sequence of an mRNA with the amino acid sequence of the polypeptide it specifies. In the 1960s, however, techniques for isolating and sequencing mRNAs had not yet been developed. The elucidation of the genetic code dictionary therefore proved to be a difficult task.

The major breakthrough in deciphering the genetic code came in 1961 when Marshall Nirenberg and Heinrich Matthaei established that UUU is the codon specifying Phe. They did so by demonstrating that the addition of poly(U) to a cell-free protein synthesizing system stimulates only the synthesis of poly(Phe). The

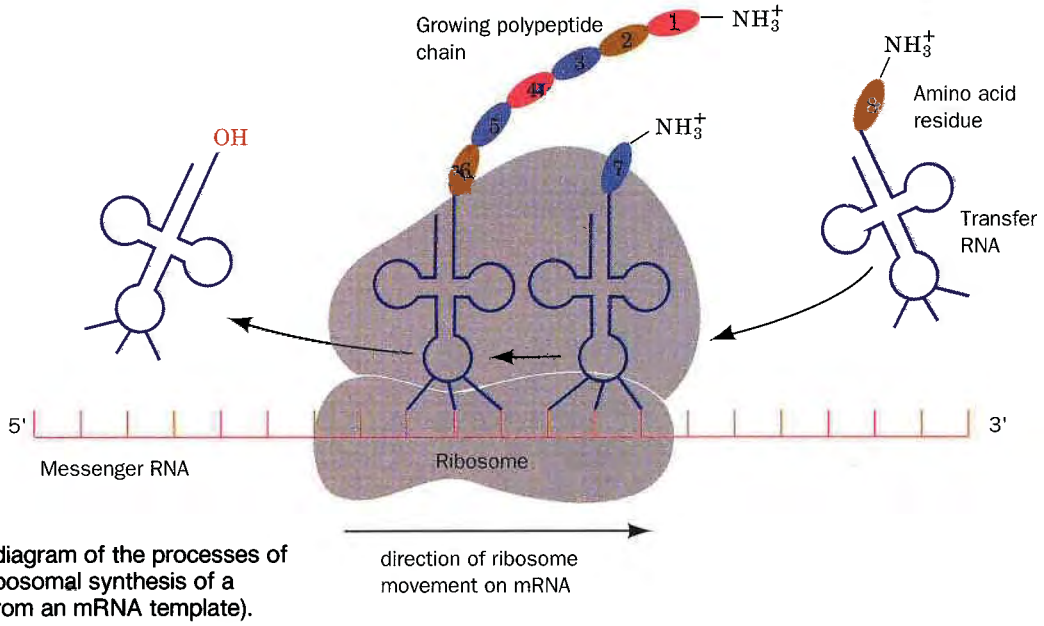


Figure 30-7
A schematic diagram of the processes of translation (ribosomal synthesis of a polypeptide from an mRNA template).

cell-free protein synthesizing system was prepared by gently breaking open *E. coli* cells by grinding them with powdered alumina and centrifuging the resulting cell sap to remove the cell walls and membranes. This extract contained DNA, mRNA, ribosomes, enzymes, and other cell constituents necessary for protein synthesis. When fortified with ATP, GTP, and amino acids, the system synthesized small amounts of proteins. This was demonstrated by the incubation of the system with ^{14}C-labeled amino acids followed by the precipitation of its proteins by the addition of trichloroacetic acid. The precipitate proved to be radioactive.

A cell-free protein synthesizing system, of course, produces proteins specified by the cell's DNA. Upon addition of DNase, however, protein synthesis stops within a few minutes because the system can no longer synthesize mRNA while that originally present is rapidly degraded. Nirenberg found that crude mRNA-containing fractions from other organisms were highly active in stimulating protein synthesis in a DNase-treated protein synthesizing system. This system, is likewise responsive to synthetic mRNAs.

The synthetic mRNAs that Nirenberg used in subsequent experiments were synthesized by the ***Azotobacter vinelandii*** enzyme **polynucleotide phosphorylase.** This enzyme, which was discovered by Severo Ochoa and Marianne Grunberg-Manago, links together nucleotides in the reaction

$$(RNA)_n + NDP \rightleftharpoons (RNA)_{n+1} + P_i$$

where NDP represents a ribonucleoside diphosphate. In contrast to RNA polymerase, however, polynucleotide phosphorylase does not utilize a template. Rather, it randomly links together the available NDPs so that the base composition of the product RNA reflects that of the reactant NDP mixture.

Nirenberg and Matthaei demonstrated that poly(U) stimulates the synthesis of poly(Phe) by incubating poly(U) and a mixture of 1 radioactive and 19 unlabeled amino acids in a DNase treated protein synthesizing system. Significant radioactivity appeared in the protein precipitate only when phenylalanine was labeled. *UUU must therefore be the codon specifying Phe.* In similar experiments using poly(A) and poly(C), it was found that poly(Lys) and poly(Pro), respectively, were synthesized. Thus AAA *specifies* Lys *and* CCC *specifies* Pro. [Poly(G) cannot function as a synthetic mRNA because, even under denaturing conditions, it aggregates to form what is thought to be a four-stranded helix. A mRNA must be single stranded to direct its transcription; Section 30-2D.]

Nirenberg and Ochoa independently employed ribonucleotide copolymers to further elucidate the genetic code. For example, in a poly(UG) composed of 76% U and 24% G, the probability of a given triplet being UUU is $0.76 \times 0.76 \times 0.76 = 0.44$. Likewise, the probability of a triplet consisting of 2U's and 1G; that is, UUG, UGU, or GUU is $0.76 \times 0.76 \times 0.24 = 0.14$. The use of the poly(UG) as a mRNA therefore indicated the base compositions, but not the sequences of the codons specifying several amino acids (Table 30-1). Through the use of copolymers containing 2, 3, and 4 bases, the base compositions of codons specifying each of the 20 amino acids were inferred. Moreover, *these experiments demonstrated that the genetic code is degenerate since, for example, poly(UA), poly(UC), and poly(UG) all direct the incorporation of Leu into a polypeptide.*

The Genetic Code Was Elucidated through Triplet Binding Assays and the Use of Polyribonucleotides with Known Sequences

In the absence of GTP, which is necessary for protein synthesis, trinucleotides but not dinucleotides are almost as effective as mRNAs in promoting the ribosomal

Table 30-1

Amino Acid Incorporation Stimulated by a Random Copolymer of U and G in Mole Ratio 0.76:0.24

Codon	Probability of Occurrence	Relative Incidence[a]	Amino Acid	Relative Amount of Amino Acid Incorporated
UUU	0.44	100	Phe	100
UUG	0.14	32	Leu	36
UGU	0.14	32	Cys	35
GUU	0.14	32	Val	37
UGG	0.04	9	Trp	14
GUG	0.04	9		
GGU	0.04	9	Gly	12
GGG	0.01	2		

[a] Relative incidence is defined here as 100 × probability of occurrence/0.44.

Source: Matthaei, J. H., Jones, O. W., Martin, R. G., and Nirenberg, M., *Proc. Natl. Acad. Sci.* **48**, 666 (1962).

binding of specific tRNAs. This phenomenon, which Nirenberg and Philip Leder discovered in 1964, permitted the various codons to be identified by a simple binding assay. Ribosomes, together with their bound tRNAs, are retained by a nitrocellulose filter but free tRNA is not. The bound tRNA was identified by using charged tRNA mixtures in which only one of the pendent amino acid residues was radioactively labeled. For instance, it was found, as expected, that UUU stimulates the ribosomal binding of only Phe tRNA. Likewise, UUG, UGU, and GUU stimulate the binding of Leu, Cys, and Val tRNAs, respectively. Hence UUG, UGU, and GUU must be codons that specify Leu, Cys, and Val, respectively. In this way, the amino acids specified by some 50 codons were identified. For the remaining codons, the binding assay was either negative (no tRNA bound) or ambiguous.

The genetic code dictionary was completed and previous results confirmed through H. Gobind Khorana's synthesis of polyribonucleotides with specified repeating sequences. In a cell-free protein synthesizing system, UCUCUCUC · · ·, for example, is read

UCU CUC UCU CUC UCU C···

so that it specifies a polypeptide chain of two alternating amino acid residues. In fact, it was observed that this mRNA stimulated the production of

Ser — Leu — Ser — Leu — Ser — Leu — ···

which indicates that either UCU or CUC specifies Ser and the other specifies Leu. This information, together with the tRNA-binding data, permitted the conclusion that UCU codes for Ser and CUC codes for Leu. These data also proved that codons consist of an odd number

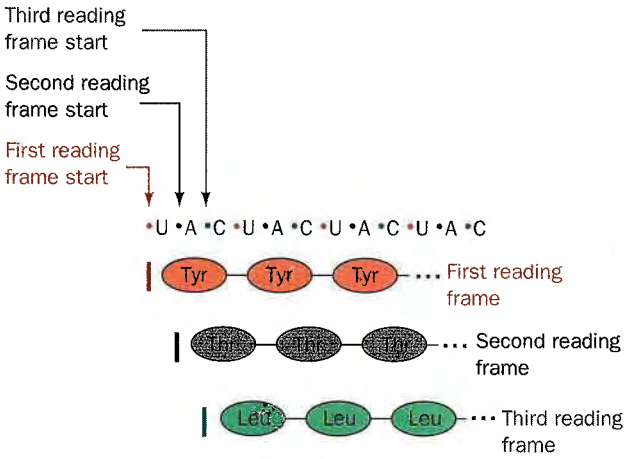


Figure 30-8
An mRNA might be read in any of three reading frames, each of which yields a different polypeptide.

of nucleotides thereby relieving any residual suspicions that codons consist of six rather than three nucleotides.

Alternating sequences of three nucleotides, such as poly(UAC), specify three different homopolypeptides because ribosomes may initiate polypeptide synthesis on these synthetic mRNAs in any of the three possible reading frames (Fig. 30-8). Analyses of the polypeptides specified by various alternating sequences of two and three nucleotides confirmed the identity of many codons and filled out missing portions of the genetic code.

mRNAs Are Read in the 5′ → 3′ Direction

The use of repeating tetranucleotides indicated the reading direction of the code and identified the chain termination codons. Poly(UAUC) specifies, as expected, a polypeptide with a tetrapeptide repeat:

5′ UAU CUA UCU AUC UAU CUA ··· 3′

Tyr — Leu — Ser — Ile — Tyr — Leu — ···

The amino acid sequence of this polypeptide indicates that the mRNA's 5′ end corresponds to the polypeptide's N-terminus; that is, *the mRNA is read in the 5′ → 3′ direction.*

UAG, UAA, and UGA Are Stop Codons

In contrast to the above results, poly(AUAG) yields only dipeptides and tripeptides. This is because *UAG is a signal to the ribosome to terminate protein synthesis:*

AUA GAU AGA UAG AUA GAU···

Ile — Asp — Arg Stop Ile — Asp — ···

Likewise, poly(GUAA) yields dipeptides and tripeptides because UAA is also a chain termination signal:

GUA AGU AAG UAA GUA AGU···

Val — Ser — Lys Stop Val — Ser — ···

UGA is a third stop signal. These stop codons, whose existence was first inferred from genetic experiments, are known, somewhat inappropriately, as **nonsense codons** because they are the only codons that do not specify amino acids. UAG, UAA, and UGA are often referred to as ***amber***, ***ochre***, and ***opal*** codons. [They were so named as the result of a laboratory joke: The German word for amber is Bernstein, the name of an individual who helped discover *amber* mutations (mutations that change some other codon to UAG); *ochre* and *opal* are puns on *amber*.]

AUG and GUG Are Chain Initiation Codons

The codons AUG, and less frequently GUG, form part of the chain initiation sequence (Section 30-3C). However, they also specify the amino acid residues Met and Val, respectively, at internal positions of polypeptide chains. (Nirenberg and Matthaei's discovery that UUU specifies Phe was only possible because ribosomes indiscriminately initiate polypeptide synthesis on a mRNA when the Mg^{2+} concentration is unphysiologically high as it was, serendipitously, in their experiments.)

E. The Nature of the Code

The genetic code dictionary, as elucidated by the above methods, is presented in Table 30-2. Examination of this table indicates that the genetic code has several remarkable features:

1. *The code is highly degenerate.* Three amino acids, Arg, Leu, and Ser are each specified by six codons, and most of the rest are specified by either four, three, or two codons. Only Met and Trp are represented by a single codon. Codons that specify the same amino acid are termed **synonyms.**

2. *The arrangement of the code table is nonrandom.* Most synonyms occupy the same box in Table 30-2; that is, they differ only in their third nucleotide. The only exceptions are Arg, Leu, and Ser which, having six codons each, must occupy more than one box. XYU and XYC always specify the same amino acid; XYA and XYG do so in all but two cases. Moreover, changes in the first codon position tend to specify similar (if not the same) amino acids, whereas codons with second position pyrimidines encode mostly hydrophobic amino acids, and those with second position purines encode mostly polar amino acids. Apparently *the code evolved so as to minimize the deleterious effects of mutations.*

Many of the mutations causing amino acid substitutions in a protein can be rationalized, according to the genetic code, as a single point mutation. For instance, all but one of the amino acid substitutions indicated in Fig. 30-5 for the α chain of tryptophan synthase result from single base changes. *As a consequence of the genetic code's degeneracy, however, many point mutations at a third codon position are phenotypically silent; that is, the mutated codon specifies the same amino acid as the wild-type.* Degeneracy may account for as much as 33% of the 25 to 75% range in the G + C content among the DNAs of different organisms (Section 28-1). The frequent occurrence of Arg, Ala, Gly, and Pro also tends to give a high G + C content, whereas Asn, Ile, Lys, Met, Phe, and Tyr contribute to a low G + C content.

Table 30-2
The "Standard" Genetic Code

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

[a] AUG forms part of the initiation signal as well as coding for internal Met residues.

Some Phage DNA Segments Contain Overlapping Genes in Different Reading Frames

Since any nucleotide sequence may have three reading frames, it is possible, at least in principle, for a polynucleotide to encode two or even three different polypeptides. This idea was never seriously entertained, however, because it seemed that the constraints on even two overlapping genes in different reading frames would be too great for them to evolve so that both could specify sensible proteins. It therefore came as a great surprise, in 1976, when Frederick Sanger reported that the DNA of bacteriophage ϕX174 contains two genes that are completely contained within larger genes of different reading frames (Fig. 30-9). Moreover, the end of the overlapping D and E genes contains the control

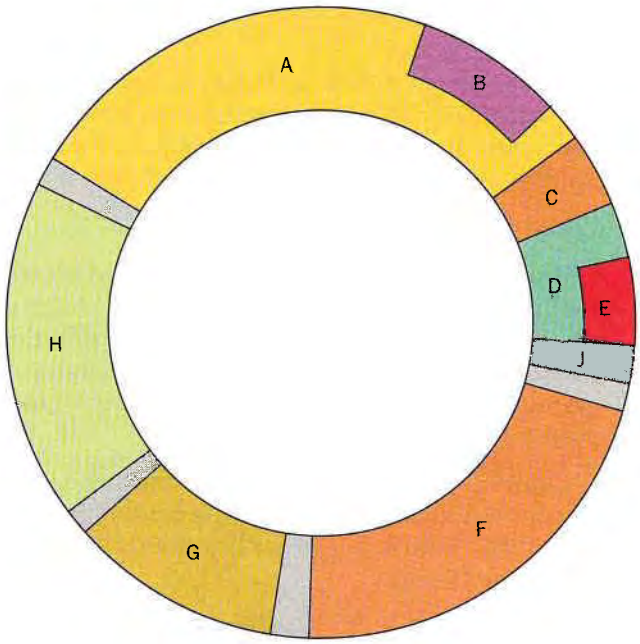


Figure 30-9
The genetic map of bacteriophage ϕX174 as determined by DNA sequence analysis. Genes are labeled A, B, C, *etc.* Note that gene B is wholly contained within gene A and gene E is wholly contained within gene D. These pairs of genes are read in different reading frames and therefore specify unrelated proteins. The unlabeled regions correspond to untranslated control sequences.

sequence for the ribosomal initiation of the J gene so that this short DNA segment performs triple duty. Bacteria also exhibit such coding economy; the ribosomal initiation sequence of one gene in a polycistronic mRNA often overlaps the end of the preceding gene. Nevertheless, completely overlapping genes have only been found in small single-stranded DNA phages, which presumably must make maximal use of the little DNA that they can pack inside their capsids.

The "Standard" Genetic Code Is Widespread but Not Universal

For many years it was thought that the "standard" genetic code (that given in Table 30-2) is universal. This assumption was, in part, based on the observations that one kind of organism (e.g., *E. coli*), can accurately translate the genes from quite different organisms, (e.g., humans). This phenomenon is, in fact, the basis of genetic engineering. Once the "standard" genetic code had been established, presumably during the time of prebiotic evolution (Section 1-4B), any mutation that would alter the way the code is translated would result in numerous, mostly deleterious, protein sequence changes. Undoubtedly there is strong selection against such mutations. DNA sequencing studies in 1981 nevertheless revealed that *the genetic codes of certain mitochondria (mitochondria contain their own genes and protein synthesizing systems that produce 10 to 20 mitochondrial proteins) are variants of the "standard" genetic code (Table 30-3).* For example, in mammalian mitochondria, AUA, as well as the standard AUG, is a Met/initiation codon, UGA specifies Trp rather than "Stop," and AGA and AGG are "Stop" rather than Arg. Note that all mitochondrial genetic codes except those of plants simplify the "standard" code by increasing its degeneracy. For example, in the mammalian mitochondrial code, each amino acid is specified by at least two codons that differ only in their third nucleotide. Apparently the constraints preventing alterations of the genetic code are eased by the small sizes of mitochondrial genomes. More recent studies, however, have revealed that in ciliated protozoa, the codons UAA and UAG specify Gln rather than "Stop." Perhaps UAA and UAG were sufficiently rare codons in a primordial ciliate (which molecular phylogenetic studies indicate branched off very early in eukaryotic evolution) to permit the code change without unacceptable deleterious effects. At any rate, *the "standard" genetic code, although very widely utilized, is not universal.*

Table 30-3
Mitochondrial Deviations from the "Standard" Genetic Code

Mitochondrion	UGA	AUA	CUN[a]	AG$^{A}_{G}$	CGG
Mammalian	Trp	Met[b]		Stop	
Baker's yeast	Trp	Met[b]	Thr		?
Neurospora crassa	Trp				?
Drosophila	Trp	Met[b]		Ser[c]	
Protozoan	Trp				
Plant					Trp
"Standard" code	Stop	Ile	Leu	Arg	Arg

[a] N represents any of the four nucleotides.
[b] Also acts as part of an initiation signal.
[c] AGA only; no AGG codons occur in *Drosophila* mitochondrial DNA.

Source: Breitenberger, C. A. and RajBhandary, U. L., *Trends Biochem. Sci.* **10,** 481 (1985).

2. TRANSFER RNA

The establishment of the genetic function of DNA led to the realization that cells somehow "translate" the language of base sequences into the language of polypeptides. Yet, nucleic acids do not specifically bind amino acids. In 1955, Francis Crick, in what became known as the **adaptor hypothesis,** hypothesized that translation occurs through the mediation of "adaptor" molecules. Each adaptor was postulated to carry a specific enzymatically appended amino acid and to recog-

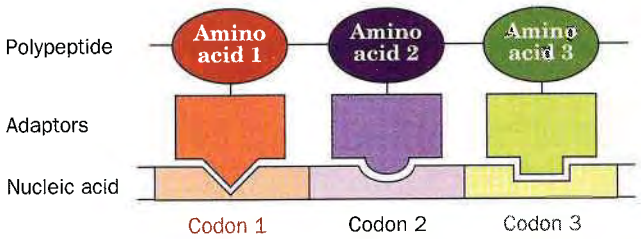


Figure 30-10
The adaptor hypothesis postulates that the genetic code is read by molecules that recognize a particular codon and carry the corresponding amino acid.

nize the corresponding codon (Fig. 30-10). Crick suggested that these adaptors contain RNA because codon recognition could then occur by complementary base pairing. At about this time, Paul Zamecnik and Mahlon Hoagland discovered that in the course of protein synthesis, ^{14}C-labeled amino acids became transiently bound to a low molecular mass fraction of RNA. Further investigations indicated that these RNAs, which at first were called "soluble RNA" or "sRNA" but are now known as transfer RNA (tRNA), are, in fact, Crick's putative adaptor molecules.

A. PRIMARY AND SECONDARY STRUCTURES

In 1965, after a seven year effort, Robert Holley reported the first known base sequence of a biologically significant nucleic acid, that of yeast alanine tRNA ($tRNA^{Ala}$; Fig. 30-11). To do so Holley had to overcome several major obstacles:

1. All organisms contain many species of tRNAs (at least one for each of the 20 amino acids) which, because of their nearly identical properties (see below), are not easily separated. Preparative techniques had to be developed to provide the gram or so of pure yeast $tRNA^{Ala}$ Holley required for its sequence determination.
2. Holley had to invent the methods that were initially used to sequence RNA (Section 28-6).
3. Ten of the 76 bases of yeast $tRNA^{Ala}$ are modified (see below). Their structural formulas had to be elucidated although they were never available in more than milligram quantities.

Since 1965, the techniques for tRNA purification and sequencing have vastly improved. A tRNA may now be sequenced in a few days time with only ~1 μg of material. Presently, the base sequences of ~300 tRNAs from a great variety of organisms are known (many from their corresponding DNA sequences). They vary in length from 60 to 95 nucleotides (18–28 kD) although most have ~76 nucleotides.

Almost all known tRNAs, as Holley first recognized, may be schematically arranged in the so-called cloverleaf secondary structure (Fig. 30-12). Starting from the 5′ end, they have the following common features:

1. A 5′ terminal phosphate group.
2. A 7 bp stem that includes the 5′-terminal nucleotide and which may contain non-Watson–Crick base pairs such as G·U. This assembly is known as the **acceptor** or **amino acid stem** because the amino acid residue carried by the tRNA is appended to its 3′-terminal OH group (Section 30-2C).
3. A 3 or 4 bp stem ending in a loop that frequently contains the modified base **dihydrouridine (D;** see below). This stem and loop are therefore collectively termed the **D arm.**
4. A 5 bp stem ending in a loop that contains the **anticodon,** the triplet of bases that is complementary to the codon specifying the tRNA. These features are known as the **anticodon arm.**

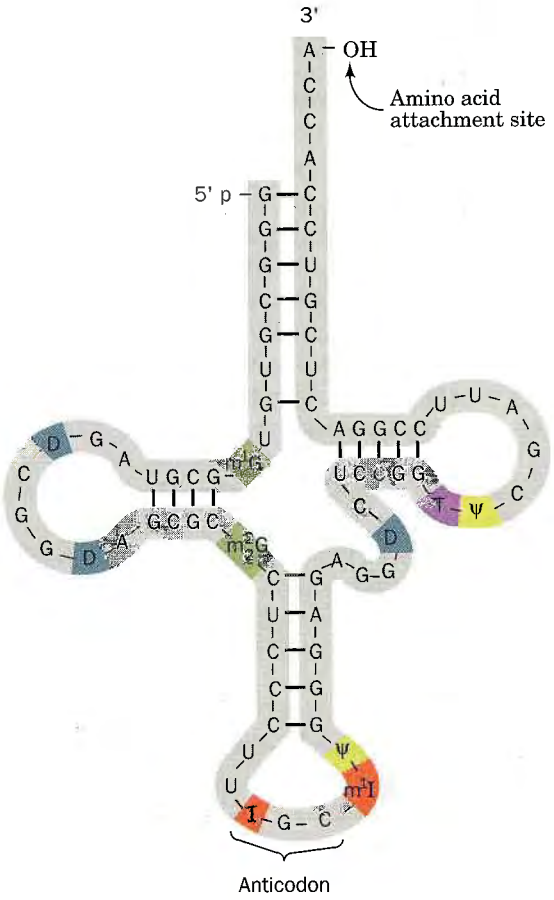


Figure 30-11
The base sequence of yeast $tRNA^{Ala}$ drawn in the cloverleaf form. The symbols for the modified nucleosides (*color*), are explained in Fig. 30-13.

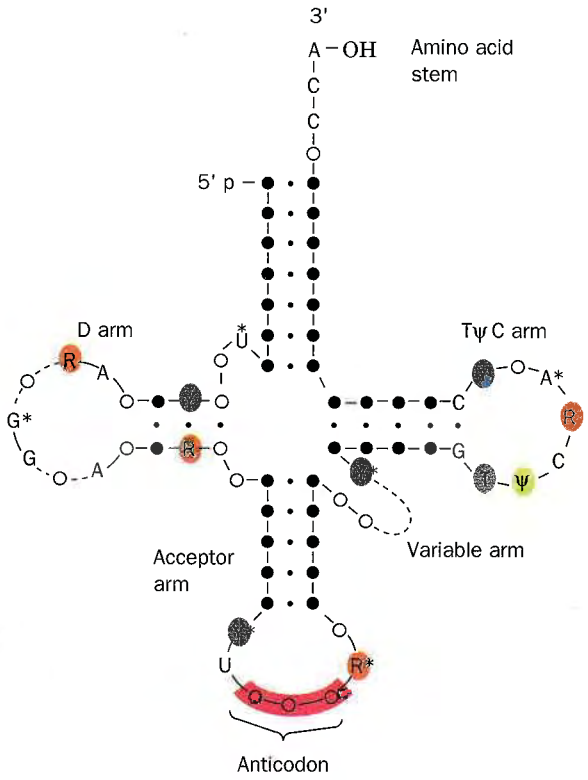


Figure 30-12
The cloverleaf secondary structure of tRNA. Filled circles connected by dots represent Watson–Crick base pairs and open circles in the double helical regions indicate bases involved in non-Watson–Crick base pairing. Invariant positions are indicated: R and Y represent invariant purines and pyrimidines, respectively, ψ signifies pseudouracil. The starred nucleosides are often modified. The dashed regions in the D and variable arms contain different numbers of nucleotides in the various tRNAs.

5. A 5 bp stem ending in a loop that usually contains the sequence TψC (where ψ is the symbol for **pseudouridine;** see below). This assembly is called the **TψC** or **T arm.**

6. All tRNAs terminate in the sequence CCA with a free 3′-OH group. The —CCA may be genetically specified or enzymatically appended to immature tRNA (Section 29-4C).

7. There are 13 invariant positions (always have the same base) and 8 **semiinvariant** positions (only a purine or only a pyrimidine) that occur mostly in the loop regions. These regions also contain **correlated invariants;** that is, pairs of nonstem nucleotides that are base paired in all tRNAs. The purine on the 3′ side of the anticodon is invariably modified. The structural significance of these features is examined in Section 30-2B.

The site of greatest variability among the known tRNAs occurs in the so-called **variable arm.** It has from 3 to 21 nucleotides and may have a stem consisting of up to 7 bp. The D loop also varies in length from 5 to 7 nucleotides.

tRNAs Have Numerous Modified Bases

One of the most striking characteristics of tRNAs is their large proportion, up to 20%, of post-translationally modified or hypermodified bases. A few of the >50 such bases, together with their standard abbreviations, are indicated in Fig. 30-13. Hypermodified nucleosides, such as i^6A, are usually adjacent to the anticodon's 3′ nucleotide when it is A or U. Their low polarities probably strengthen the otherwise relatively weak pairing associations of these bases with the codon thereby increasing translational fidelity. Conversely, certain methylations block base pairing and hence prevent inappropriate structures from forming. Yet, neither of these modifications are essential for maintaining a tRNA's structural integrity (see below), for its proper binding to the ribosome, nor, with one known exception (Section 30-2C), for binding the enzyme that attaches the correct amino acid. The functions of most modified bases therefore remain unknown although mutant bacteria unable to form certain modified bases compete poorly against the corresponding normal bacteria.

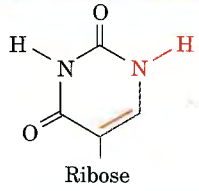
B. Tertiary Structure

The earliest physicochemical investigations of tRNA indicated that it has a well-defined conformation. Yet, despite numerous hydrodynamic, spectroscopic, and chemical cross-linking studies, its three-dimensional structure remained an enigma until 1974. In that year, the 2.5-Å resolution X-ray crystal structure of yeast $tRNA^{Phe}$ was separately elucidated by Alexander Rich in collaboration with Sung Hou Kim and, in a different crystal form, by Aaron Klug. *The molecule assumes an L-shaped conformation in which one leg of the L is formed by the acceptor and T stems folded into a continuous A-RNA-like double helix (Section 28-2B) and the other leg is similarly composed of the D and anticodon stems* (Fig. 30-14). Each leg of the L is ~60 Å long and the anticodon and amino acid acceptor sites are at opposite ends of the molecule, some 76-Å apart. The narrow 20 to 25-Å width of native tRNA is essential to its biological function: During protein synthesis, two RNA molecules must simultaneously bind in close proximity at adjacent codons on mRNA (Section 30-3D).

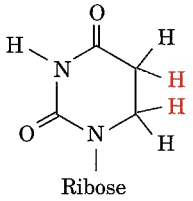
tRNA's Complex Tertiary Structure Is Maintained by Hydrogen Bonding and Stacking Interactions

The structural complexity of yeast $tRNA^{Phe}$ is reminiscent of that of a protein. Although only 42 of its 76 bases occur in double helical stems, *71 of them participate in stacking associations* (*Fig. 30-15*). The structure also contains 9 base pairing interactions that cross-link its tertiary structure (Figs. 30-14*a* and 30-15). Remarkably, all but one of these tertiary interactions, which appear to be

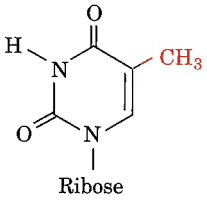
Uracil derivatives



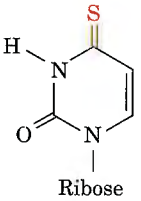
Pseudouridine (ψ)



Dihydrouridine (D)

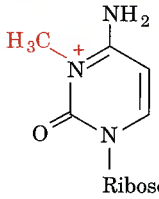


Ribothymidine (T)

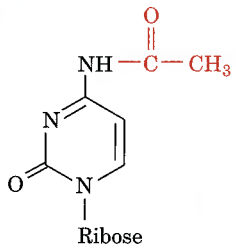


4-Thiouridine (s^4U)

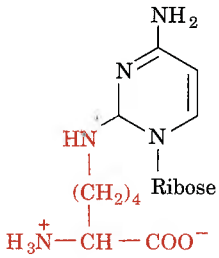
Cytosine derivatives



3-Methylcytidine (m^3C)

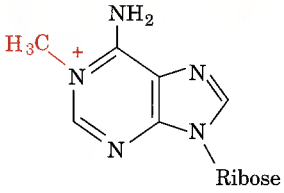


***N*⁴-Acetylcytidine (ac^4C)**

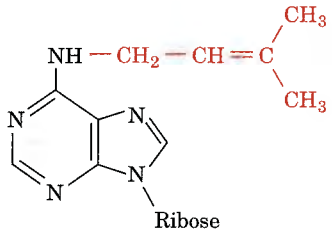


Lysidine (L)

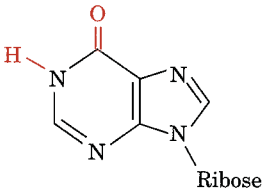
Adenine derivatives



1-Methyladenosine (m^1A)

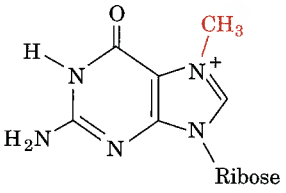


***N*⁶-Isopentenyladenosine (i^6A)**

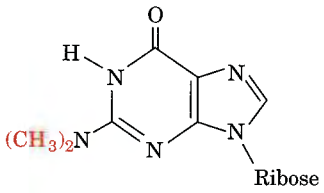


Inosine (I)

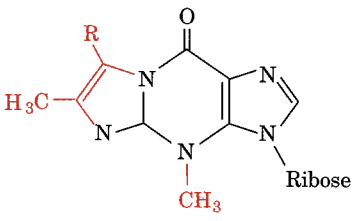
Guanine derivatives



***N*⁷-Methylguanosine (m^7G)**



***N*²,*N*²-Dimethylguanosine (m_2^2G)**



R = H **Wyosine (Wo)**

R =  **Y**

Figure 30-13
A selection of the modified nucleosides that occur in tRNAs together with their standard abbreviations. Note that although inosine chemically resembles guanosine, it is biochemically derived from adenosine. Nucleosides may also be methylated at their ribose 2′ positions to form residues symbolized, for instance, by Cm, Gm, and Um.

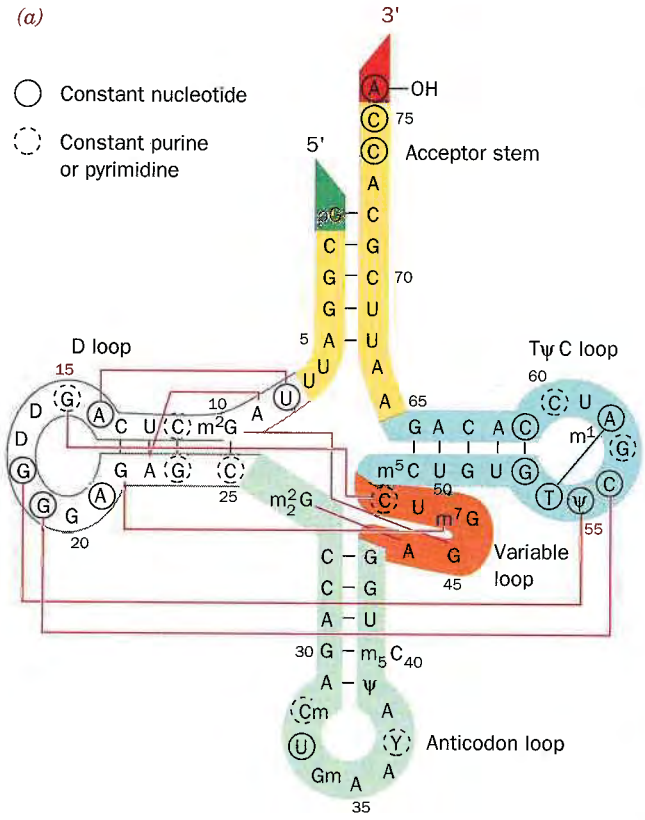


Figure 30-14
The structure of yeast tRNA^{Phe}. (*a*) The base sequence drawn in cloverleaf form. Base pairing interactions are represented by thin red lines connecting the participating bases. Bases that are conserved or semiconserved in all tRNAs are circled by solid and dashed lines, respectively. The 5′ terminus is colored bright green, the acceptor stem is yellow, the D arm is white, the anticodon arm is light green, the variable arm is orange, the TψC arm is light blue, and the 3′ terminus is red. (*b*) The X-ray structure drawn to show how its base paired stems are arranged form the L-shaped molecule. The sugar–phosphate backbone is represented by a ribbon with the same color scheme as that in Part a. [Courtesy of Michael Carson, University of Alabama at Birmingham.]

the mainstays of the molecular structure, are non-Watson–Crick associations. Moreover, most of the bases involved in these interactions are either invariant or semiinvariant, which strongly suggests that all tRNAs have similar conformations (see below). The structure is also stabilized by several unusual hydrogen bonds between bases and either phosphate groups or the 2′-OH groups of ribose residues.

The compact structure of yeast tRNA^{Phe} results from its large number of intramolecular associations, which renders most of its bases inaccessible to solvent. The most notable exceptions to this are the anticodon bases and those of the amino acid-bearing —CCA terminus. No doubt both of these groupings must be accessible in order to carry out their biological functions.

The observation that the molecular structures of yeast tRNA^{Phe} in two different crystal forms are essentially identical lends much credence to the supposition that its crystal structure closely resembles its solution structure. Transfer RNAs other than yeast tRNA^{Phe} have, unfortunately, been notoriously difficult to crystallize. The crystal structures of only three other native species of tRNA, all at resolutions of 3.0 Å or greater, have thus far been reported. The molecular structures of these tRNAs closely resemble that of yeast tRNA^{Phe}. The major structural differences among them result from an apparent flexibility in the anticodon loop and the —CCA terminus as well as from a hingelike mobility between the two legs of the L that gives, for instance, yeast tRNA^{Asp} a boomeranglike shape. Such observations are in accord with the expectation that all tRNAs fit into the same ribosomal cavities.

C. Aminoacyl-tRNA Synthetases

Accurate translation requires two equally important recognition steps: (1) the choice of the correct amino acid for covalent attachment to a tRNA; and (2) the selection of the amino acid-charged tRNA specified by mRNA. The first of these steps, which is catalyzed by amino acid-specific enzymes known as **aminoacyl-tRNA synthetases,** appends an amino acid to the 3′-terminal ribose residue of its cognate tRNA to form an **aminoacyl-tRNA** (Fig.

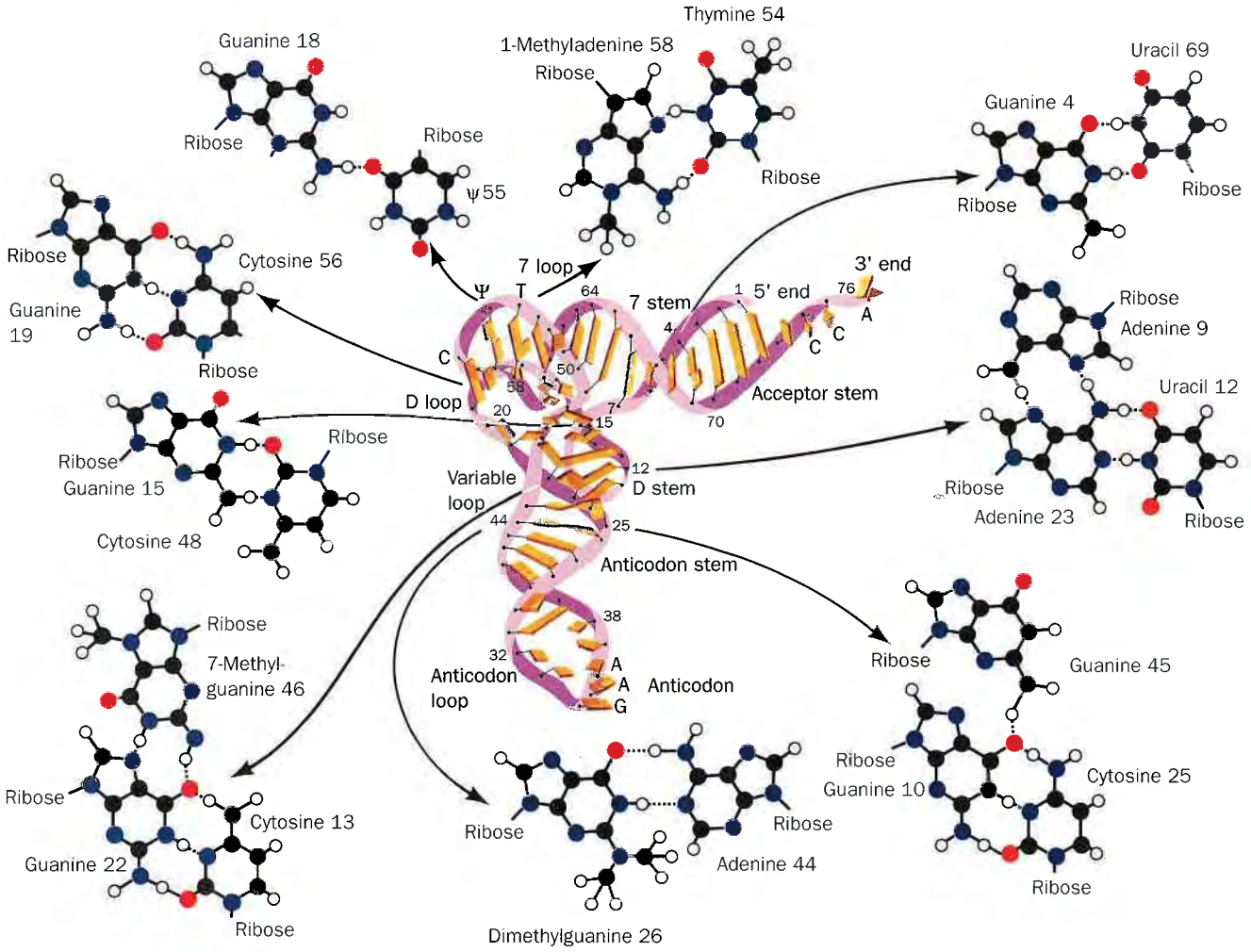
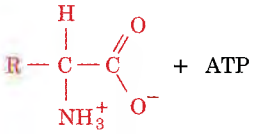


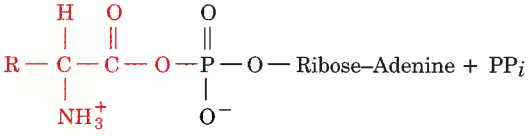
Figure 30-15
The nine tertiary base pairing interactions in yeast $tRNA^{Phe}$. Note that all but one involve non-Watson–Crick pairs and that they are all located near the corner of the L. [After Kim, S. H., *in* Schimmel, P. R., Söll, D., and Abelson, J. N. (Eds.), *Transfer RNA: Structure, Properties and Recognition, p. 87*, Cold Spring Harbor Laboratory (1979).]

30-16). This otherwise unfavorable process is driven by the hydrolysis of ATP in two sequential reactions that are catalyzed by a single enzyme.

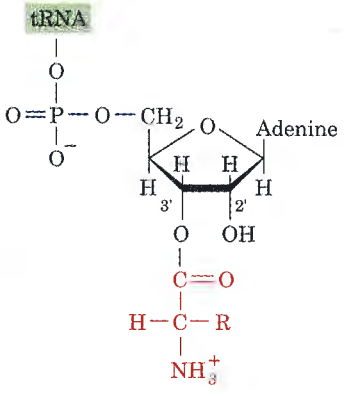
1. The amino acid is first "activated" by reaction with ATP to form an **aminoacyl-adenylate:**



Amino acid



Aminoacyl-adenylate (Aminoacyl-AMP)



Aminoacyl-tRNA

Figure 30-16
In aminoacyl-tRNAs, the amino acid residue is esterified to the tRNA's 3′-terminal nucleoside at either its 3′-OH group, as shown here, or its 2′-OH group.

which, with most aminoacyl-tRNA synthetases in the absence of tRNA, may be isolated although it normally remains tightly bound to the enzyme.

2. This mixed anhydride then reacts with tRNA to form the aminoacyl-tRNA:

$$\text{Aminoacyl-AMP} + \text{tRNA} \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP}$$

Some aminoacyl-tRNA synthetases exclusively append an amino acid to the terminal 2′-OH group of their cognate tRNAs, others do so at the 3′-OH group, and yet others do so at either such position. This selectivity or its absence was established with the use of chemically modified tRNAs that lack either the 2′- or 3′-OH group of their 3′-terminal ribose residue. The use of these derivatives was necessary because, in solution, the aminoacyl group rapidly equilibrates between the 2′ and 3′ positions.

The overall aminoacylation reaction is

$$\text{Amino acid} + \text{tRNA} + \text{ATP} \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP} + PP_i$$

These reaction steps are readily reversible because the free energies of hydrolysis of the bonds formed in both the aminoacyl-adenylate and the aminoacyl-tRNA are comparable to that of ATP hydrolysis. The overall reaction is driven to completion by the inorganic pyrophosphatase-catalyzed hydrolysis of the PP_i generated in the first reaction step. Amino acid activation therefore chemically resembles fatty acid activation (Section 23-2A); the major difference between these two processes, which were both elucidated by Paul Berg, is that tRNA is the acyl acceptor in amino acid activation, whereas CoA performs this function in fatty acid activation.

Different Aminoacyl-tRNA Synthetases Are No More Than Distantly Related

Cells must have at least one aminoacyl-tRNA synthetase for each of the 20 amino acids. The similarity of the reactions catalyzed by these enzymes and the structural resemblance of all tRNAs suggests that all aminoacyl-tRNA synthetases evolved from a common ancestor and should therefore be structurally related. This is not the case, however. In fact, *the aminoacyl-tRNA synthetases form a diverse group of enzymes.* The over 100 such enzymes that have been characterized each have one of four different types of subunit structures, α, α_2, α_4, and $\alpha_2\beta_2$, with subunit sizes ranging from 334 to >1000 residues. Moreover, although synthetases specific for a given amino acid exhibit considerable sequence homology from organism to organism, there is little sequence similarity among synthetases specific for different amino acids. Quite possibly, aminoacyl-tRNA synthetases arose very early in evolution, before the development of the modern protein synthesis apparatus other than tRNAs.

Tyrosyl-tRNA Synthetase Operates via Transition State Binding

The X-ray structure of tyrosyl-tRNA synthetase from *Bacillus stearothermophilus,* determined by David Blow, is illustrated in Fig. 30-17. The 419-residue subunit of this α_2 dimer contains a region of β sheet reminiscent of the dinucleotide-binding fold (Section 7-3B), which forms the tyrosyl adenylate-binding site. This region is remarkably similar in structure to the ATP-binding region of *E. coli* methionyl-tRNA synthetase, the only other aminoacyl-tRNA synthetase of known structure. The C-terminal 99 residues of tyrosyl-tRNA synthetase, as well as three other short segments of its polypeptide chain, are not visible in the crystal structure and are therefore presumed to be conformationally disordered. Each of these segments has several Lys and Arg residues that are implicated in the binding of the polyanionic tRNA molecule. Indeed, the N-terminal 320 residues alone, as generated via protein engineering, catalyzes tyrosine adenylate formation with unchanged k_{cat} and K_M, but neither aminoacylates nor binds $tRNA^{Tyr}$. Most nucleic acid-binding proteins of known structure, as we shall see, have conformationally mobile regions that interact with their corresponding nucleic acid. It has therefore been suggested that these disordered regions function to bind their nuclei acid through flexible interactions.

Although the way in which tyrosyl-tRNA synthetase interacts with $tRNA^{Tyr}$ remains obscure, model building coupled with protein engineering studies have revealed how this enzyme adenylates tyrosine. Chemical stud-

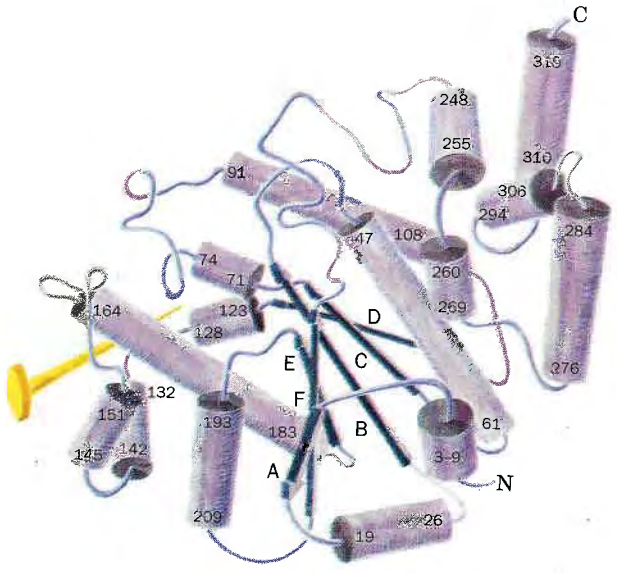


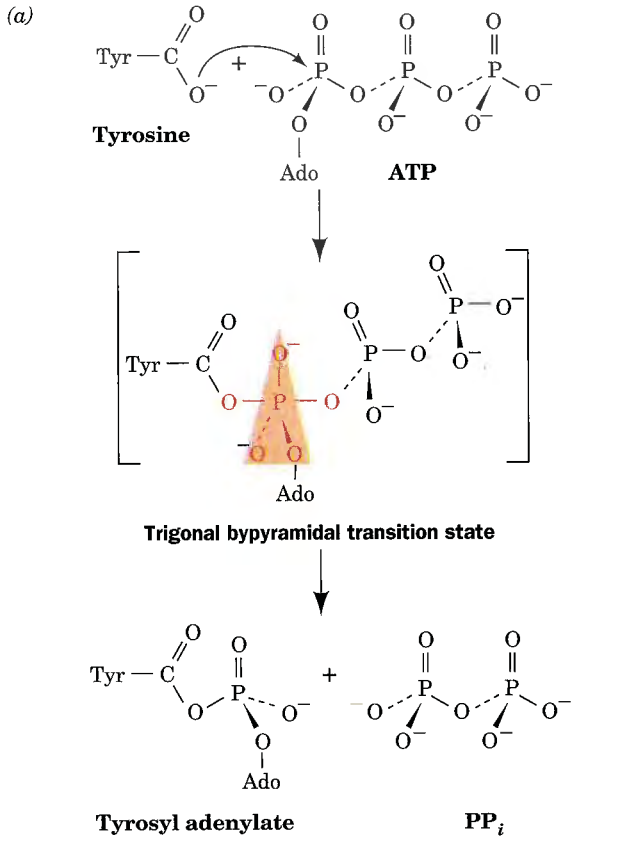
Figure 30-17
The X-ray structure of residues 1 to 320 of tyrosyl-tRNA synthetase. The position of the molecular twofold axis of this dimeric protein is indicated on the lower left. [After Blow, D. M. and Brick, P., *in* Jurnak, F. A. and McPherson, A., *Biological Macromolecules and Assembly,* Vol 2: *Nucleic Acids and Interactive Proteins, p.* 448, Wiley (1985).]

ies have demonstrated that this reaction proceeds via inversion of configuration at ATP's α phosphorus. This observation implies that the reaction involves a single displacement in which the tyrosyl carboxylate group is the nucleophile and PP_i is the leaving group (Fig. 30-18*a*). Model building studies by Alan Fersht and Greg Winter based on this premise, together with the X-ray structure of tyrosyl-tRNA synthetase's tyrosyl adenylate complex, indicate that the enzyme operates by preferentially binding the transition state (Section 14-1F): The γ phosphate in the reaction's pentacoordinate transition state, but not its reactants or products, hydrogen bonds to the enzyme's Thr 40 and His 45 side chains (Fig. 30-18*b*). Fersht and Winter confirmed this conclusion through protein engineering studies in which they replaced Thr 40 with Ala and/or His 45 (which is evolutionarily conserved in aminoacyl-tRNA synthetases) with Gly. All of these mutant enzymes have greatly reduced catalytic activities (a 3×10^5-fold reduction in k_{cat} in the double mutant) even though they all bind both tyrosine and ATP with nearly undiminished affinities. Note that the interactions stabilizing the transition state occur at some distance from the α phosphorous reaction site. Moreover, the enzyme has no catalytically active functional groups, such as general acids or bases, in the vicinity of the reaction site. Evidently, the inherent reactivities of the nucleophilic tyrosyl carboxyl group and ATP's activated PP_i leaving group are sufficient to drive the reaction at a satisfactorily high rate (an $\sim 10^9$-fold increase over the uncatalyzed reaction) with only transition state binding combined with reactant proximity and orientation effects (Section 14-1E).

The Structural Features Recognized by Aminoacyl-tRNA Synthetases May Be Quite Simple

Considerable effort has been expended in elucidating the manner in which aminoacyl-tRNA synthetases recognize their corresponding tRNAs. The methods used to do so include the use of specific tRNA fragments, mutationally altered tRNAs, and chemical cross-linking agents. The most common synthetase contact sites on tRNA occur on the inner (concave) face of the L. Other than that, there appears to be little regularity in how the various tRNAs recognize their cognate synthetases. Indeed, the smaller synthetases appear to recognize only the acceptor region of their tRNAs, whereas the larger enzymes contact much of their tRNA's inner surface. Thus, *the anticodon does not necessarily participate in this recognition process.*

The foregoing suggests that the features of a tRNA recognized by its cognate aminoacyl-tRNA synthetase are idiosyncratic. Genetic manipulations by Paul Schimmel revealed that these features, for at least one type of tRNA, are surprisingly simple. Numerous sequence alterations of *E. coli* $tRNA^{Ala}$ do not appreciably affect its capacity to be aminoacylated with alanine. Yet,



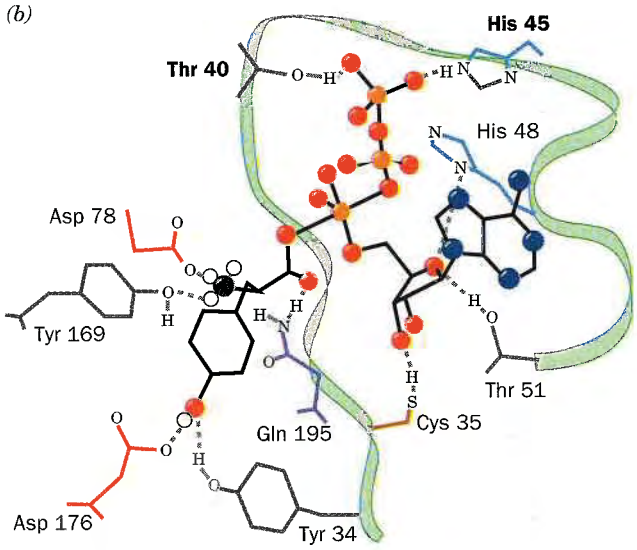


Figure 30-18
The mechanism of tyrosyl adenylate formation as catalyzed by tyrosyl-tRNA synthetase. (*a*) In the reaction's chemical mechanism, the tyrosyl carboxylate group nucleophilically attacks ATP's α phosphorus (*top*) in a single displacement reaction that proceeds via a trigonal bipyramidal transition state (*middle*) to yield tyrosyl adenylate and PP_i (*bottom*). (*b*) Model building studies based on the X-ray structure of tyrosyl-tRNA synthetase in complex with tyrosyl adenylate (a stable complex in the absence of $tRNA^{Tyr}$) indicate that the γ phosphate of the reacting ATP hydrogen bonds to Thr 40 and His 45 (*top*) only in the reaction's transition state. Tyrosyl adenylate also makes 12 hydrogen bonds with the enzyme (several of which are indicated here by dashed lines) that apparently are not significantly disturbed in the transition state. [After Leatherbarrow, R. J., Fersht, A. R., and Winter, G., *Proc. Natl. Acad. Sci.* **82,** 7841 (1985).]

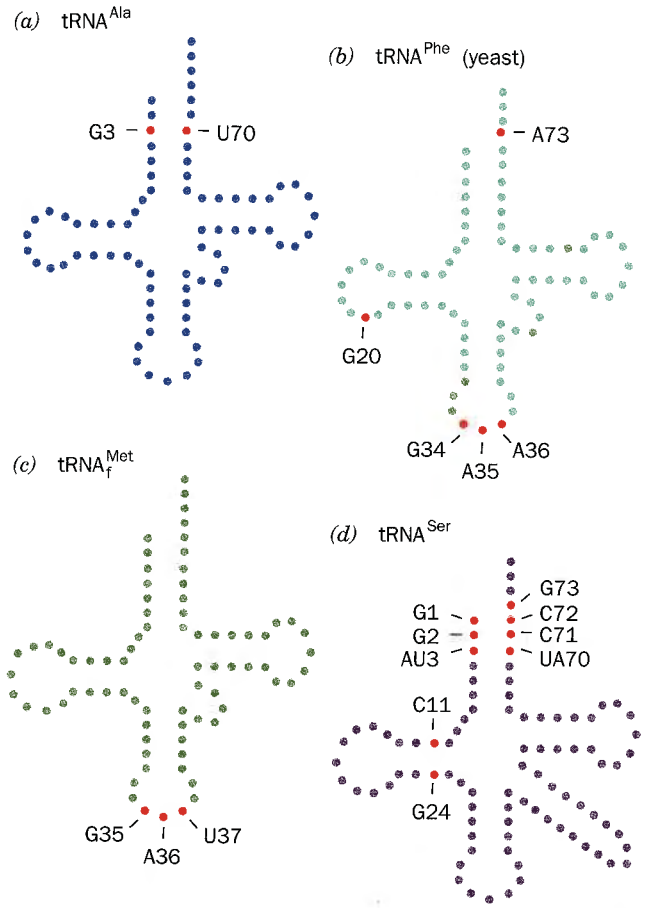


Figure 30-19
Major identity elements in four tRNAs. Each base in the tRNA is represented by a circle. Red circles indicate positions that have been shown to be identity elements for the recognition of the tRNA by its cognate aminoacyl-tRNA synthetase. In each case, other identity elements may yet be discovered. [After Schulman, L. H. and Abelson, J., *Science* **240,** 1592 (1988).]

most base substitutions in the G3 · U70 base pair located in the tRNA's acceptor stem (Fig. 30-19*a*) greatly diminish this reaction. Moreover, the introduction of a G · U base pair into the analogous position of $tRNA^{Cys}$ and $tRNA^{Phe}$ causes them to be aminoacylated with alanine even though there are few other sequence similarities between these mutant tRNAs and $tRNA^{Ala}$ (e.g., Fig. 30-20). In fact, *E. coli* alanyl-tRNA synthetase even efficiently aminoacylates a 24-nucleotide "microhelix" derived from only the G3 · U70-containing acceptor stem of *E. coli* $tRNA^{Ala}$. Since the only known *E. coli* tRNAs that normally have a G3 · U70 base pair are the $tRNA^{Ala}$, and this base pair is also present in the $tRNA^{Ala}$ from many organisms including yeast (Fig. 30-11), the foregoing observations strongly suggest that *the G3 · U70 base pair is a major feature recognized by alanyl-tRNA synthetases.* These enzymes presumably recognize the distorted shape of the G · U base pair (Fig. 30-15), an idea corroborated by the observation that base changes at G3 · U70, which least affect the acceptor identity of $tRNA^{Ala}$ yield base pairs that structurally resemble G · U.

The elements of three other tRNAs, which are recognized by their cognate tRNA synthetases, are indicated in Fig. 30-19. As with $tRNA^{Ala}$, these identifiers appear to comprise only a few bases. Note that the anticodon is an identifier in two of these tRNAs. In another example of an anticodon identifier, the *E. coli* $tRNA^{Ile}$ specific for the codon AUA has the anticodon LAU where L is **lysidine,** a modified cytosine whose 2-keto group is replaced by the amino acid lysine (Fig. 30-13). The L in this context pairs with A rather than G, a unique case of base modification altering base pairing specificity. The replacement of this L with unmodified C, as expected, yields a tRNA which recognizes the Met codon AUG (codons bind anticodons in an antiparallel fashion). Surprisingly, however, this altered $tRNA^{Ile}$ is also a much better substrate for methionyl-tRNA synthetase than it is for isoleucyl-tRNA synthetase. Thus, both the codon and the amino acid specificity of this tRNA are changed by a single post-transcriptional modification.

Proofreading Enhances the Fidelity of Amino Acid Attachment to tRNA

The charging of a tRNA with its cognate amino acid is a remarkably accurate process. Experimental measure-

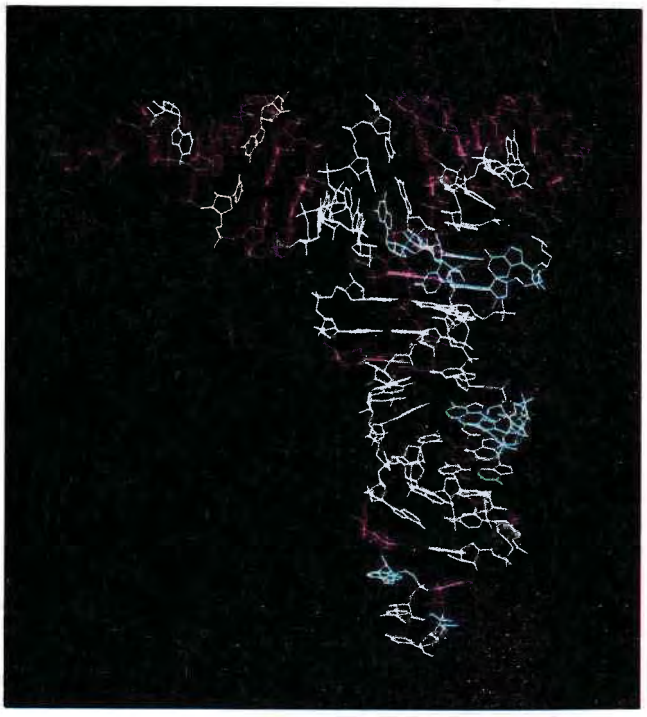


Figure 30-20
A three-dimensional model of *E. coli* $tRNA^{Ala}$ based on the X-ray structure of yeast $tRNA^{Phe}$ (Fig. 30-14) in which the nucleotides that are different in *E. coli* $tRNA^{Cys}$ are highlighted in blue-white and the G3 · U70 base pair is highlighted in ivory. [Courtesy of Ya-Ming Hou, MIT.]

ments indicate, for example, that, at equal concentrations of isoleucine and valine, isoleucyl-tRNA synthetase transfers ~50,000 isoleucines to $tRNA^{Ile}$ for every valine it so-transfers. Yet, *there are insufficient structural differences between Val and Ile to permit such a high degree of accuracy in the direct generation of aminoacyl-tRNAs.* It seems likely that isoleucyl-tRNA synthetase has a binding site of sufficient size to admit isoleucine but which would exclude larger amino acids. On the other hand, valine, which differs from isoleucine by only the lack of a single methylene group, fits into the isoleucine-binding site. The binding free energy of a methylene group is estimated to be ~12 kJ·mol^{-1}. Equation [3.16] indicates that the ratio f of the equilibrium constants, K_1 and K_2, with which two substances bind to a given binding site is given by

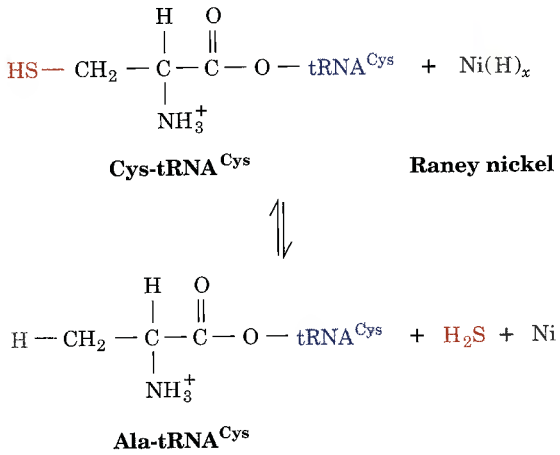
$$f = \frac{K_1}{K_2} = \frac{e^{-\Delta G_1/RT}}{e^{-\Delta G_2/RT}} = e^{-\Delta\Delta G/RT} \qquad [30.1]$$

where $\Delta\Delta G = \Delta G_1 - \Delta G_2$ is the difference between the free energies of binding of the two substances. It is therefore estimated that isoleucyl-tRNA synthetase could discriminate between isoleucine and valine by no more than a factor of ~100.

Berg resolved this apparent paradox by demonstrating that, in the presence of $tRNA^{Ile}$, isoleucyl-tRNA synthetase catalyzes the quantitative hydrolysis of valine-adenylate to valine + AMP rather than forming Val-$tRNA^{Ile}$. Thus, *isoleucyl-tRNA synthetase subjects aminoacyl-adenylates to a* ***proofreading*** *or* ***editing*** *step that has been shown to occur at a separate catalytic site.* This site presumably binds Val residues but excludes the larger Ile residues. *The enzymes's overall selectivity is therefore the product of the selectivities of its adenylation and proofreading steps, thereby accounting for the high fidelity of translation.* Many other synthetases discriminate against noncognate amino acids in a similar fashion. However, synthetases that have adequate selectivity for their corresponding amino acid (e.g., tyrosyl-tRNA synthetase discriminates between tyrosine and phenylalanine through hydrogen bonding with the tyrosine-OH group), lack editing functions. Note that *editing occurs at the expense of ATP hydrolysis, the thermodynamic price of high fidelity (increased order).*

D. Codon–Anticodon Interactions

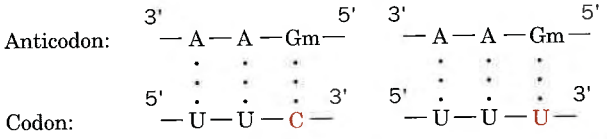
In protein synthesis, the proper tRNA is selected only through codon–anticodon interactions; the aminoacyl group does not participate in this process. This phenomenon was demonstrated as follows. Cys-$tRNA^{Cys}$, in which the Cys residue was ^{14}C labeled, was reductively desulfurized with Raney nickel so as to convert the Cys residue to Ala:



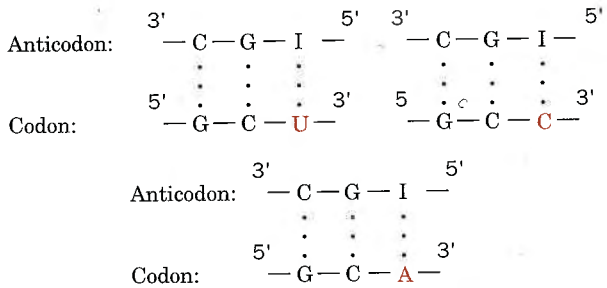
The resulting ^{14}C-labeled hybrid, Ala-$tRNA^{Cys}$, was added to a cell-free protein synthesizing system extracted from rabbit reticulocytes. The product hemoglobin α chain's only radioactive tryptic peptide was the one that normally contains the subunit's only Cys. No radioactivity was found in the peptides that normally contain Ala but no Cys. Evidently, only the anticodons of aminoacyl-tRNAs are involved in codon recognition.

Genetic Code Degeneracy Is Largely Mediated by Variable Third Position Codon–Anticodon Interactions

One might naively guess that each of the 61 codons specifying an amino acid would be read by a different tRNA. Yet, even though most cells contain several groups of **isoaccepting tRNAs** (different tRNAs that are specific for the same amino acid), *many tRNAs bind to two or three of the codons specifying their cognate amino acids.* For example, yeast $tRNA^{Phe}$, which has the anticodon GmAA, recognizes the codons UUC and UUU (remember that the anticodon pairs with the codon in an antiparallel fashion),



and yeast $tRNA^{Ala}$, which has the anticodon IGC, recognizes the codons GCU, GCC, and GCA.



It therefore seems that non-Watson–Crick base pairing can occur at the third codon–anticodon position (the anticodon's first position is defined as its 3′ nucleotide), the site of most codon degeneracy (Table 30-2). Note also that the third (5′) anticodon position commonly contains a modified base such as Gm or I.

The Wobble Hypothesis Structurally Accounts for Codon Degeneracy

By combining structural insight with logical deduction, Crick proposed, in what he named the **wobble hypothesis,** how a tRNA can recognize several degenerate codons. He assumed that the first two codon–anticodon pairings have normal Watson–Crick geometry. The structural constraints that this places on the third codon–anticodon pairing ensure that its conformation does not drastically differ from that of a Watson–Crick pair. Crick then proposed that there could be a small amount of play or "wobble" in the third codon position which allows limited conformational adjustments in its pairing geometry. This permits the formation of several non-Watson–Crick pairs such as U·G and I·A (Fig. 30-21*a*). The allowed "wobble" pairings are indicated in Fig. 30-21*b*. Then, by analyzing the known pattern of codon–anticodon pairing, Crick deduced the most plausible sets of pairing combinations in the third codon–anticodon position (Table 30-4). Thus, an anticodon with C or A in its third position can only pair with its Watson–Crick complementary codon. If U, G, or I occupies the third anticodon position, two, two, or three codons are recognized, respectively.

No prokaryotic or eukaryotic cytoplasmic tRNA is known to participate in a nonwobble pairing combination. There is, however, no known instance of such a tRNA with an A in its third anticodon position which suggests that the consequent U·A pair is not permitted. The structural basis of wobble pairing is poorly understood although it is clear that it is influenced by base modifications.

A consideration of the various wobble pairings indicates that at least 31 tRNAs are required to translate all 61 coding triplets of the genetic code (there are 32 tRNAs in the minimal set because translational initiation requires a separate tRNA; Section 30-3C). Most cells have >32 tRNAs, some of which have identical anticodons. Nevertheless, *all isoaccepting tRNAs in a cell are recognized by a single aminoacyl-tRNA synthetase.*

Table 30-4
Allowed Wobble Pairing Combinations in the Third Codon–Anticodon Position

5′-Anticodon Base	3′-Codon Base
C	G
A	U
U	A or G
G	U or C
I	U, C, or A

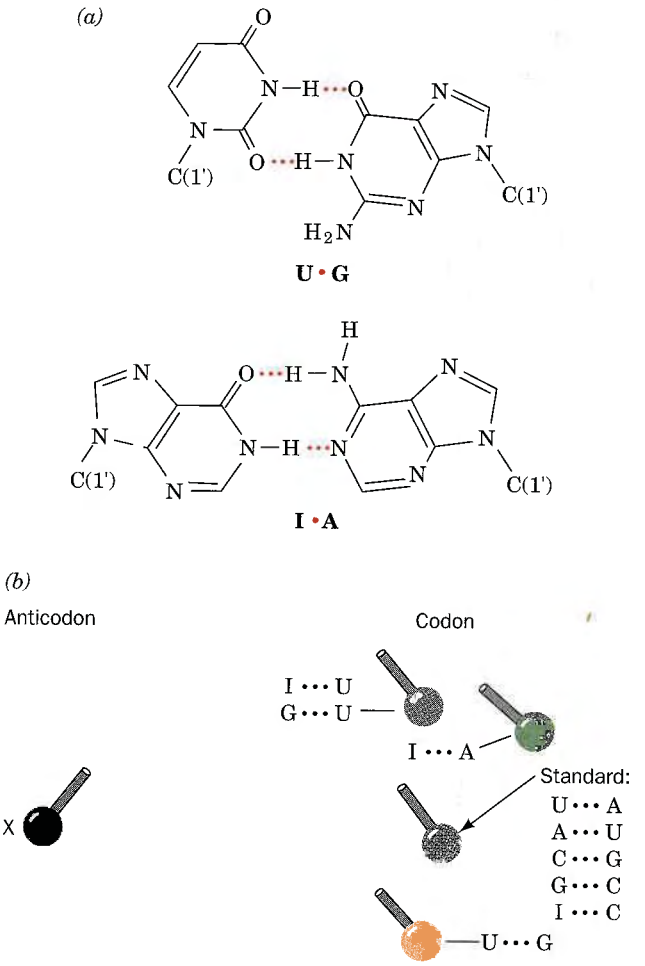


Figure 30-21
Wobble pairing. (*a*) U·G and I·A wobble pairs. (*b*) The geometry of wobble pairing. The spheres and their attached bonds represent the positions of ribose C(1′) atoms with their accompanying glycosidic bonds. X (*left*) designates the nucleoside at the 5′ end of the anticodon (tRNA). The positions on the right are those of the 3′ nucleoside of the codon (mRNA) in the indicated wobble pairings. [After Crick, F. H. C., *J Mol. Biol.* **19,** 55 (1966).]

Some Mitochondrial tRNAs Have More Permissive Wobble Pairings Than Other tRNAs

The codon recognition properties of mitochondrial tRNAs must reflect the fact that mitochondrial genetic codes are variants of the "standard" genetic code (Table 30-3). For instance, the human mitochondrial genome, which consists of only 16,569 bp, encodes 22 tRNAs (together with 2 ribosomal RNAs and 13 proteins). Fourteen of these tRNAs each read one of the synonymous pairs of codons indicated in Tables 30-2 and 30-3 (MNX, where X is either C or U or else A or G) according to normal G·U wobble rules: The tRNAs have either a G or a modified U in their third anticodon position that,

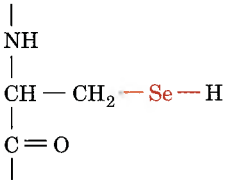
respectively, permits them to pair with codons having X = C or U or else X = A or G. The remaining 8 tRNAs, which, contrary to wobble rules, each recognize 1 of the groups of 4 synonymous codons (MNY, where Y = A, C, G, or U), all have anticodons with a U in their third position. Either this U can somehow pair with any of the 4 bases or these tRNAs read only the first two codon positions and ignore the third. Thus, not surprisingly, many mitochondrial tRNAs have unusual structures in which, for example, the GTψCRA sequence (Fig. 30-12) is missing, or, in the most bizarre case, a $tRNA^{Ser}$ lacks the entire D arm.

Frequently Used Codons Are Complementary to the Most Abundant tRNA Species

The analysis of the base sequences of several highly expressed structural genes of baker's yeast, *Saccharomyces cerevisiae,* has revealed a remarkable bias in their codon usage. Only 25 of the 61 coding triplets are commonly used. *The preferred codons are those that are most nearly complementary, in the Watson–Crick sense, to the anticodons in the most abundant species in each set of isoaccepting tRNAs.* Furthermore, codons that bind anticodons with two consecutive G·C pairs or three A·U pairs are avoided so that the preferred codon–anticodon complexes all have approximately the same binding free energies. A similar phenomenon occurs in *E. coli* although several of its 22 preferred codons differ from those in yeast. The degree with which the preferred codons occur in a given gene is strongly correlated, in both organisms, with the gene's level of expression. This, it has been proposed, permits the mRNAs of proteins that are required in high abundance to be rapidly and smoothly translated.

Selenocysteine Is Specified by a tRNA

Although it is widely stated, even in this text, that proteins are synthesized from the 20 "standard" amino acids, that is, those specified by the "standard" genetic code, some organisms, in fact, use a 21st amino acid, **selenocysteine,** in synthesizing a few of their proteins.



The selenocysteine residue

Selenium, a biologically essential trace element, is a component of several enzymes in both prokaryotes and eukaryotes. *E. coli* contains two selenoproteins, both **formate dehydrogenases,** which each contain a selenocysteine residue. The selenocysteine residues are ribosomally incorporated into these proteins by a unique tRNA bearing a UCA anticodon that is specified by a particular (in the mRNA) UGA codon (normally the *opal* stop codon). How the ribosomal system differentiates this UGA from normal *opal* stop codons is unknown although mRNA context effects and the physiological state of the cell are probably involved. The selenocysteinyl-tRNA is synthesized by the aminoacylation of its tRNA with L-serine by the same aminoacyl-tRNA synthetase that charges $tRNA^{Ser}$, followed by the enzymatic selenylation of the resulting Ser residue.

E. Nonsense Suppression

Nonsense mutations are usually lethal when they prematurely terminate the synthesis of an essential protein. An organism with such a mutation may nevertheless be "rescued" by a second mutation on another part of the genome. For many years after their discovery, the existence of such **intergenic suppressors** was quite puzzling. It is now known, however, that they usually arise from mutations in a tRNA gene that cause the tRNA to recognize a nonsense codon. Such a **nonsense suppressor** tRNA appends its amino acid (which is the same as that carried by the corresponding wild-type tRNA) to a growing polypeptide in response to the recognized stop codon thereby preventing chain termination. For example, the *E. coli amber* suppressor known as *su*3 is a $tRNA^{Tyr}$ whose anticodon has mutated from the wild-type GUA (which reads the Tyr codons UAU and UAC) to CUA (which recognizes the *amber* stop codon UAG). An $su3^+$ *E. coli* with an otherwise lethal *amber* mutation in a gene coding for an essential protein would be viable if the replacement of the wild-type amino acid residue by Tyr does not inactivate the protein.

There are several well-characterized examples of *amber* (UAG), *ochre* (UAA), and *opal* (UGA) suppressors in *E. coli* (Table 30-5). Most of them, as expected, have mutated anticodons. UGA-1 tRNA, however, differs from the wild-type only by a G → A mutation in its D stem, which changes a G·U pair to a stronger A·U pair.

Table 30-5
Some *E. coli* Nonsense Suppressors

Name	Codon Suppressed	Amino Acid Inserted
*su*1	UAG	Ser
*su*2	UAG	Gln
*su*3	UAG	Tyr
*su*4	UAA, UAG	Tyr
*su*5	UAA, UAG	Lys
*su*6	UAA	Leu
*su*7	UAA	Gln
UGA-1	UGA	Trp
UGA-2	UGA	Trp

Source: Körner, A. M., Feinstein, S. I., and Altman, S., *in* Altman, S. (Ed.), *Transfer RNA, p.* 109, MIT Press (1978).

This mutation apparently alters the conformation of the tRNA's CCA anticodon so that it can form an unusual wobble pairing with UGA as well as with its normal codon, UGG. Nonsense suppressors also occur in yeast.

Suppressor tRNAs Are Mutants of Minor tRNAs

How do cells tolerate a mutation that both eliminates a normal tRNA and prevents the termination of polypeptide synthesis? They survive because the mutated tRNA is usually a minor member of a set of isoaccepting tRNAs and because nonsense suppressor tRNAs must compete for stop codons with the protein factors that mediate the termination of polypeptide synthesis (Section 30-E3). Consequently, the rate of suppressor-mediated synthesis of active proteins with either UAG or UGA nonsense mutations rarely exceeds 50% of the wild-type rate whereas mutants with UAA, the most common termination codon, have suppression efficiencies of <5%. Many mRNAs, moreover, have two tandem stop codons so that even if their first stop codon was suppressed, termination could occur at the second. Nevertheless, many suppressor-rescued mutants grow relatively slowly because they cannot make an otherwise prematurely terminated protein as efficiently as do wild-type cells.

Other types of suppressor tRNAs are also known. **Missense suppressors** act similarly to nonsense suppressors but substitute one amino acid in place of another. **Frameshift suppressors** have eight nucleotides in their anticodon loops rather than the normal seven. They read a four base codon beyond a base insertion thereby restoring the wild-type reading frame.

3. RIBOSOMES

Ribosomes were first seen in cellular homogenates by dark field microscopy in the late 1930s by Albert Claude who referred to them as "microsomes". It was not until the mid-1950s, however, that George Palade observed them in cells by electron microscopy thereby disposing of the contention that they are merely artifacts of cell disruption. The name ribosome derives from the fact that these particles in *E. coli* consist of ~⅔ RNA and ⅓ protein. (**Microsomes** are now defined as the artifactual vesicles formed by the endoplasmic reticulum upon cell disruption. They are easily isolated by differential centrifugation and are rich in ribosomes.) The correlation between the amount of RNA in a cell and the rate at which it synthesizes protein led to the suspicion that ribosomes are the site of protein synthesis. This hypothesis was confirmed in 1955 by Paul Zamecnik who demonstrated that ^{14}C-labeled amino acids are transiently associated with ribosomes before they appear in free proteins. Further research showed that ribosomal polypeptide synthesis has three distinct phases: (1) chain initiation, (2) chain elongation, and (3) chain termination.

In this section we examine the structure of the ribosome, insofar as it is known, and then outline the ribosomal mechanism of polypeptide synthesis. In doing so we shall compare the properties of ribosomes from prokaryotes (mostly *E. coli*) with those of eukaryotes (mostly rat liver cytoplasm).

A. Ribosome Structure

The *E. coli* ribosome, which has a particle mass of ~2.5 × 10^6 D and a sedimentation coefficient of 70S, is a spheroidal particle that is ~250 Å across in its largest dimension. It may be dissociated, as James Watson discovered, into two unequal subunits (Table 30-6). The small (30S) subunit consists of a 16S rRNA molecule and 21 different polypeptides, whereas the large (50S) subunit contains a 5S and a 23S rRNA together with 32

Table 30-6
Components of *E. coli* Ribosomes

	Ribosome	Small Subunit	Large Subunit
Sedimentation coefficient	70S	30S	50S
Mass (kD)	2520	930	1590
RNA			
Major		16S, 1542 nucleotides	23S, 2904 nucleotides
Minor			5S, 120 nucleotides
RNA mass (kD)	1664	560	1104
Proportion of mass	66%	60%	70%
Proteins		21 polypeptides	31 polypeptides
Protein mass (kD)	857	370	487
Proportion of mass	34%	40%	30%

Source: Lewin, B., *Genes* (3rd ed.), *p.* 145, Wiley (1987).

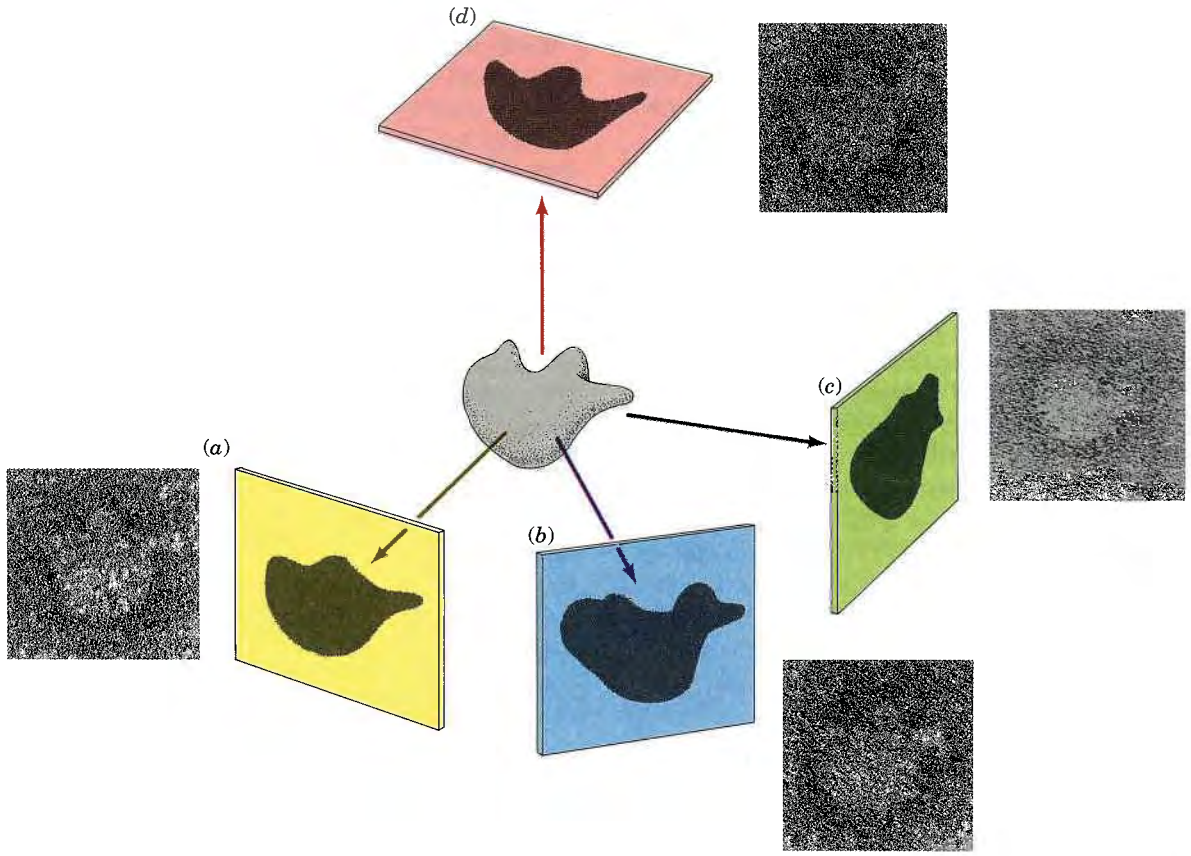


Figure 30-22
The three-dimensional model of the large ribosomal subunit was deduced by mathematically combining its two-dimensional electron microscope images as viewed from different directions. The model of the small subunit was similarly determined. [Courtesy of James Lake, UCLA.]

different polypeptides. The up to 20,000 ribosomes in an *E. coli* cell account for ~80% of its RNA content and 10% of its protein.

Although the ribosome has recently been crystallized by Ada Yonath, it is such a complex entity that it will be many years before its structure is known in molecular detail. However, the low resolution structures of the ribosome and its subunits have been determined through image reconstruction techniques pioneered by Aaron Klug in which electron micrographs of a single particle or ordered sheets of particles taken from several directions are combined to yield its three-dimensional image (Fig. 30-22). The small subunit is a roughly mitten-shaped particle, whereas the large subunit is spheroidal with three protuberances on one side (Fig. 30-23). The large subunit also contains a tunnel, up to 25 Å in diameter and 100 to 120 Å long, that extends from a cleft between the subunit's three protuberances and is postulated to provide the nascent polypeptide's exit path (Fig. 30-24).

Ribosomal RNAs Have Evolutionarily Conserved Secondary Structures

The *E. coli* 16S rRNA, which was sequenced by Harry Noller, consists of 1542 nucleotides. A computerized search of this sequence for stable double helical segments yielded many plausible but often mutually exclusive secondary structures. However, the comparison of the sequences of 16S rRNAs from several prokaryotes, under the assumption that their structures have been evolutionarily conserved, led to the flowerlike secondary structure for 16S rRNA proposed in Fig. 30-25. This four-domain structure, which is 46% base paired, is reasonably consistent with the results of nuclease digestion and chemical modification studies. Its double helical stems tend to be short (<8 bp) and many of them are imperfect. Intriguingly, electron micographs of the 16S rRNA resemble those of the complete 30S subunit, thereby suggesting that the 30S subunit's overall shape is largely determined by the 16S rRNA.

The large ribosomal subunit's 5S and 23S rRNAs,

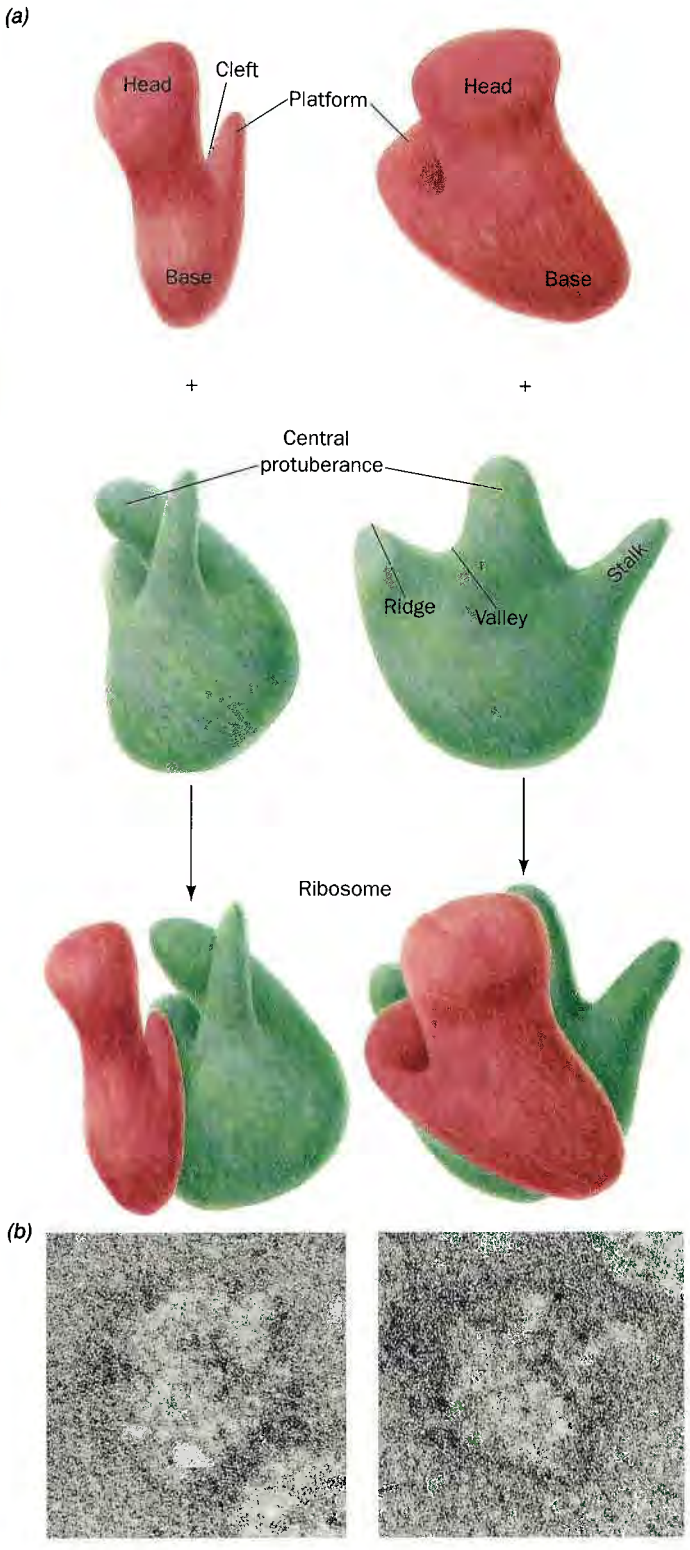


Figure 30-23
(*a*) A three-dimensional model of the *E. coli* ribosome deduced as indicated in Fig. 30-22. The small subunit (*top*) combines with the large subunit (*middle*) to form the complete ribosome (*bottom*). The two views of the ribosome match those seen in (*b*) the electron micrographs. [Courtesy of James Lake, UCLA.]

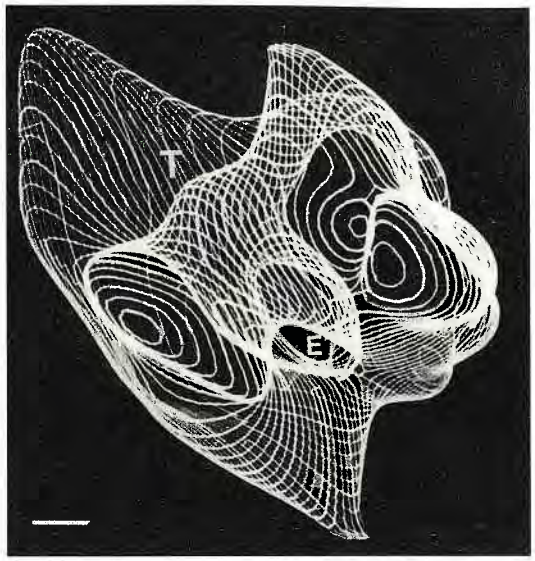


Figure 30-24
A computer-generated image of the large ribosomal subunit from *Bacillus stearothermophilus* as determined by electron micrographic image reconstruction of oriented two-dimensional arrays of particles. An ~25 Å in diameter tunnel extends ~100 Å from the cleft between the subunit's three protrusions (T) to the nascent polypeptide's probable exit site (E). The bar is 20 Å long. [Courtesy of Ada Yonath, Weizmann Institute of Science.]

which consist of 120 and 2904 nucleotides, respectively, have also been sequenced. As with the 16S rRNA, they appear to have extensive secondary structures. That proposed for 5S rRNA is shown in Fig. 30-26. Of course, as we have seen for tRNA, the secondary structure of an RNA provides little indication of its three-dimensional structure (but see below).

Ribosomal Proteins Have Been Partially Characterized

Ribosomal proteins are difficult to separate because most of them are insoluble in ordinary buffers. By convention, ribosomal proteins from the small and large subunits are designated with the prefixes S and L, respectively, followed by a number indicating their position, from upper left to lower right, on a two-dimensional gel electrophoretogram (roughly in order of decreasing molecular mass; Fig. 30-27). Only protein S20/L26 is common to both subunits. It is apparently located at the interface between the two subunits. One of the large subunit proteins is partially acetylated at its N-terminus so that it gives rise to two electrophoretic spots (L7/L12). Four copies of this protein are present in the large subunit. Moreover, these four copies of L7/L12 aggregate with L10 to form a stable complex that was initially thought to be a unique protein, "L8."

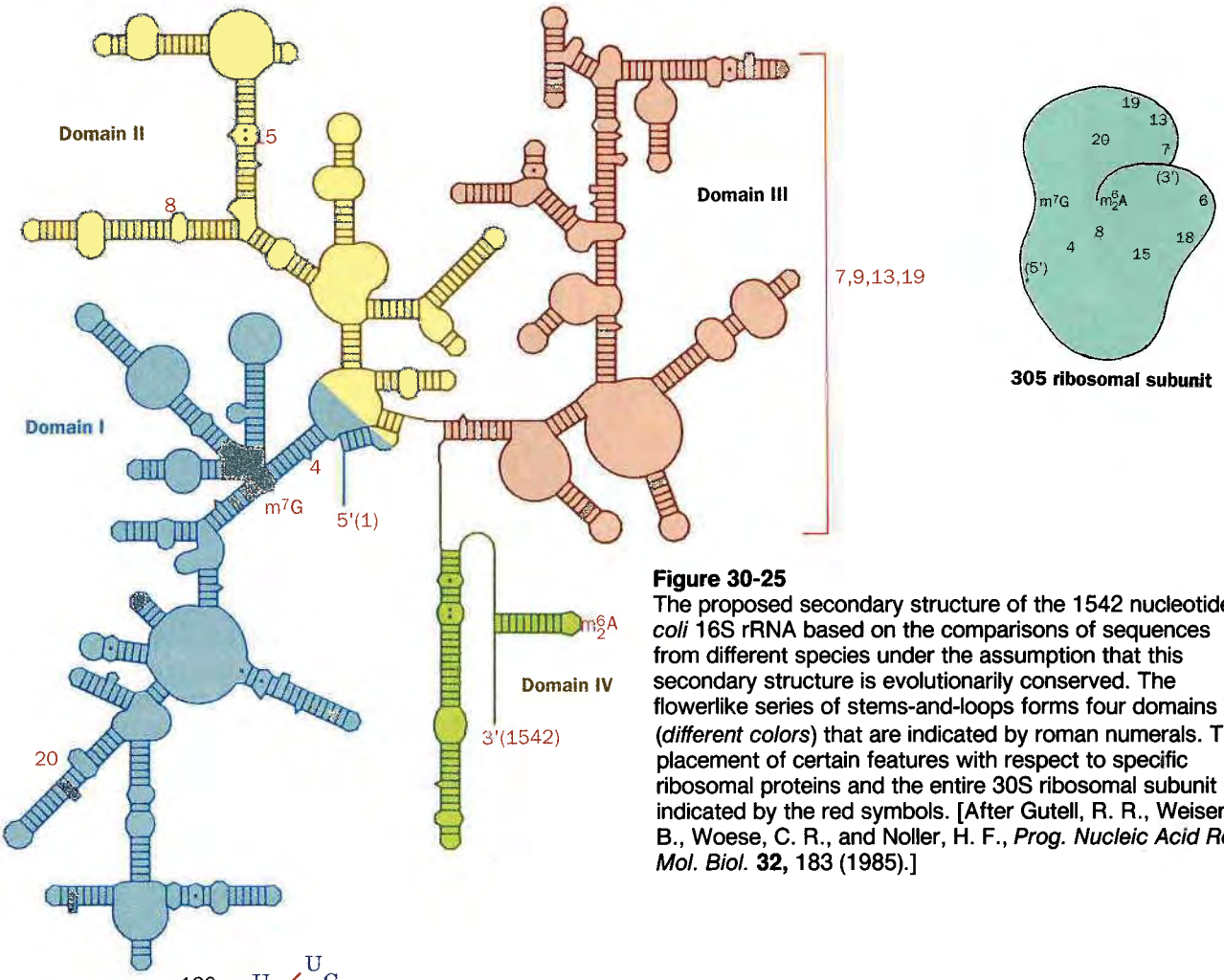


Figure 30-25
The proposed secondary structure of the 1542 nucleotide *E. coli* 16S rRNA based on the comparisons of sequences from different species under the assumption that this secondary structure is evolutionarily conserved. The flowerlike series of stems-and-loops forms four domains (*different colors*) that are indicated by roman numerals. The placement of certain features with respect to specific ribosomal proteins and the entire 30S ribosomal subunit are indicated by the red symbols. [After Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F., *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 183 (1985).]

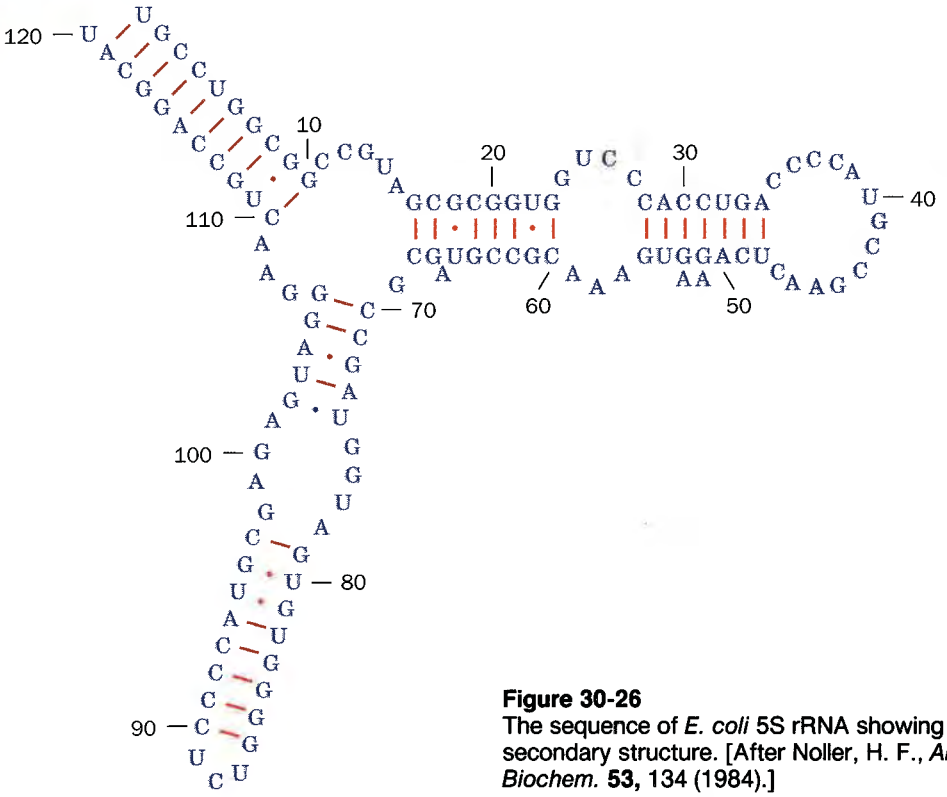


Figure 30-26
The sequence of *E. coli* 5S rRNA showing its proposed secondary structure. [After Noller, H. F., *Annu. Rev. Biochem.* **53,** 134 (1984).]

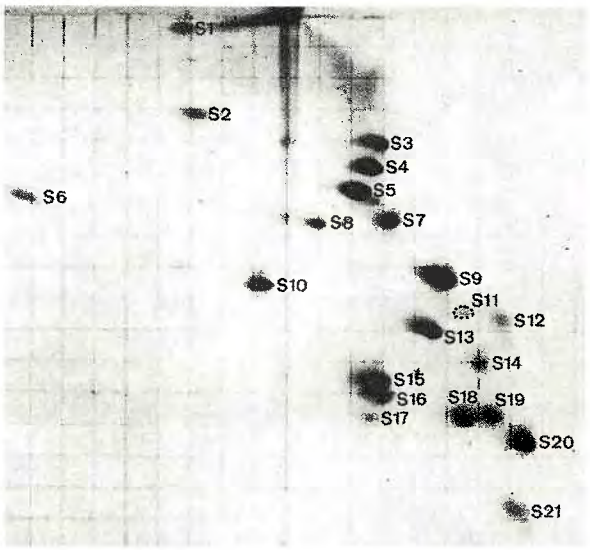


Figure 30-27
A two-dimensional gel electrophoretogram of *E. coli* small subunit proteins. First dimension (*vertical*): 8% acrylamide, pH 8.6; second dimension (*horizontal*): 18% acrylamide; pH 4.6. [From Kaltschmidt, E. and Wittmann, H. G., *Proc. Natl. Acad. Sci.* **67,** 1277 (1970).]

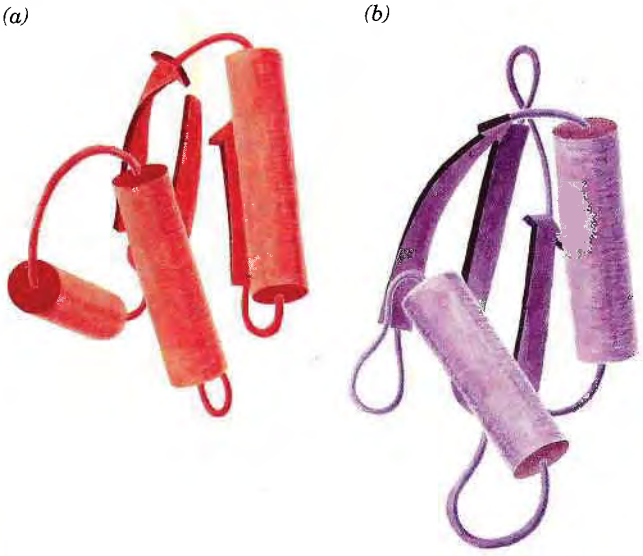


Figure 30-28
The X-ray structures of two ribosomal proteins: (*a*) The 74-residue C-terminal fragment of *E. coli* L7/L12. (*b*) *Bacillus stearothermophilus* L30 (61 residues). The two protein molecules are oriented so as to show their closely similar backbone comformations. [After Leijonmarck, M., Appelt, K., Badger, J., Liljas, A., Wilson, K. S., and White, S. W., *Proteins* **3,** 244 (1988).]

All the other ribosomal proteins occur in only one copy per subunit.

The amino acid sequences of all 52 *E. coli* ribosomal proteins have been elucidated, mainly by Heinz-Günter Wittmann and Brigitte Wittmann-Liebold. They range in size from 46 residues for L34 to 557 residues for S1. Most of these proteins, which exhibit little sequence similarity with one another, are rich in the basic amino acids Lys and Arg and contain few aromatic residues as expected for proteins that are closely associated with polyanionic RNA molecules. The X-ray structure of only two ribosomal proteins have so far been reported: those of L30 and a C-terminal segment of L7/L12 (Fig. 30-28). These proteins have remarkably similar structures despite their only 14% amino acid sequence identity.

Ribosomal Subunits Are Self-Assembling

Ribosomal subunits form, under proper conditions, from mixtures of their numerous macromolecular components. *Ribosomal subunits are therefore self-assembling entities.* Masayasu Nomura determined how this occurs through partial reconstitution experiments. If one macromolecular component is left out of an otherwise self-assembling mixture of proteins and RNA, the other components that fail to bind to the resulting partially assembled subunit must somehow interact with the omitted component. Through the analysis of a series of such partial reconstitution experiments, Nomura constructed an assembly map of the small subunit (Fig. 30-29). This map indicates that the first steps in small subunit assembly are the independent binding of cer-

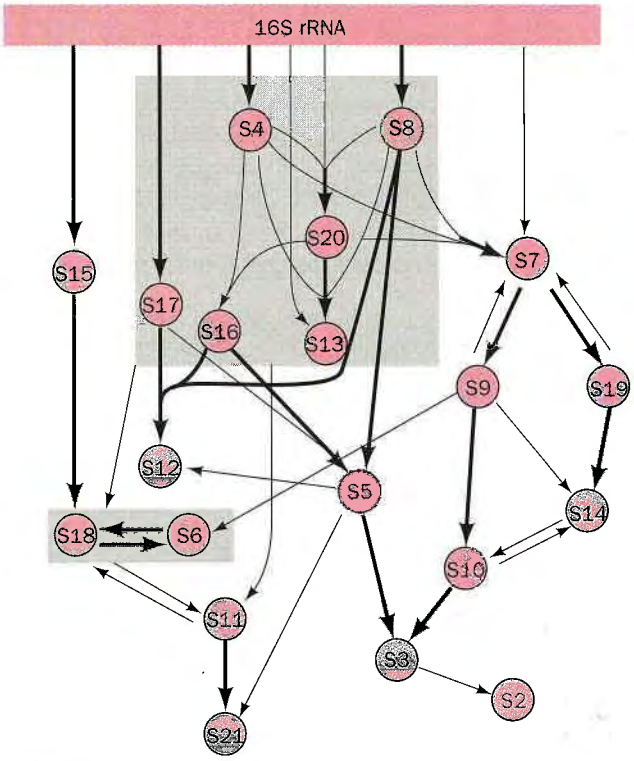


Figure 30-29
The assembly map of the *E. coli* small subunit. Thick and thin arrows between components indicate strong and weak facilitation of binding, respectively. For example, the thick arrow from 16S rRNA to S4 indicates that S4 binds directly to 16S rRNA in the absence of other proteins, whereas the thin arrows from 16S rRNA, S4, S8, S9, S19, and S20 to S7 indicate that the former components all participate in binding S7. [After Held, W. A., Ballou, B., Mizushima, S., and Nomura, M., *J. Biol. Chem.* **249,** 3109 (1974).]

tain proteins to 16S rRNA. The resulting assembly intermediates provide the molecular scaffolding for binding other proteins. At one stage of the assembly process, an intermediate particle must undergo a marked conformational change before assembly can continue. The large subunit self-assembles in a similar manner. The observation that similar assembly intermediates occur *in vivo* and *in vitro* suggests that *in vivo* and *in vitro* assembly processes are much alike.

Ribosomal Architecture Has Been Deduced through Immune Electron Microscopy and Neutron Diffraction Studies

The positions of most ribosomal components have been determined through a variety of physical and chemical techniques. Many proteins have been located by James Lake and by Georg Stöffler through **immune electron microscopy.** Rabbit antibodies [**immunoglobulin G (IgG);** Section 34-2A] raised against a specific ribosomal protein bind to this protein where it is exposed on the surface of its subunit. Electron microscopy of the ribosomal subunit·IgG complex indicates the point of attachment of the IgG and hence the site of the ribosomal protein to which it binds (Figs. 30-30 and 30-31). These results have been confirmed and extended through neutron diffraction measurements of 30S subunits conducted by Donald Engleman and Peter Moore (Fig. 30-32*a*) and by similar studies on 50S subunits by Knud Nierhaus. The protein positions indicated in Figs. 30-31 and 30-32*a* are consistent with the subunit assembly map shown in Fig. 30-29 in that pairs of proteins that must interact for proper subunit assembly (although not necessarily by direct contact) are in close proximity.

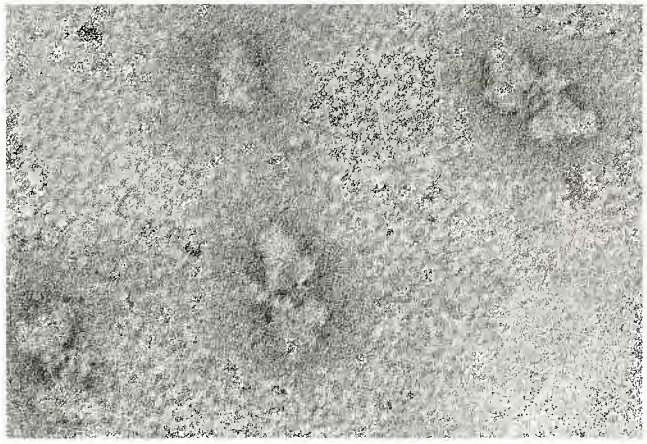


Figure 30-30
Immune electron microscopy reveals the positions of ribosomal proteins. Immunoglobin G (IgG) raised against a particular ribosomal protein, here S6, is mixed with ribosomes. The IgG, which is a Y-shaped protein (Section 34-2B), binds to its corresponding antigen at the ends of the two short prongs of the Y, thereby binding together two ribosomes. The position of the protein on the surface of the ribosome is indicated by the point of attachment of the IgG. [Courtesy of James Lake, UCLA.]

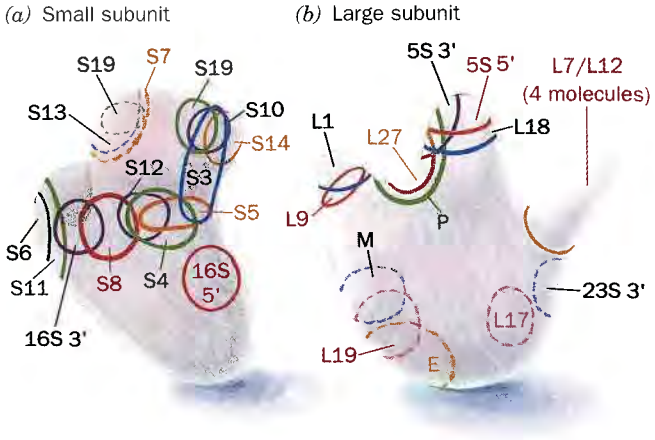


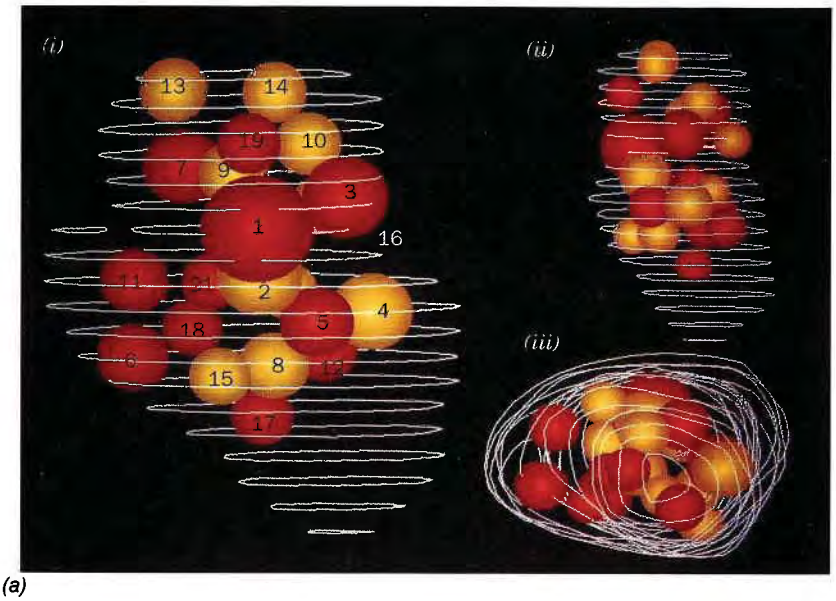
Figure 30-31
Maps of the *E. coli* ribosomal (*a*) small and (*b*) large subunits indicating the locations of some of their component proteins as determined by immune electron microscopy. Sites that are dashed are located on the back side of the subunit. On the small subunit, the symbols 16S 3′ and 16S 5′ mark the ends of the 16S RNA. On the large subunit, P indicates the peptidyl transferase site, E marks the site where the nascent polypeptide emerges from the ribosome (the end of the tunnel in Fig. 30-24), M specifies the ribosome's membrane anchor site, and 5S 3′ and 5S 5′ mark the ends of the 5S rRNA. [After Lake, J. A., *Annu. Rev. Biochem.* **54,** 512 (1985).]

Many of the secondary structural elements of the 16S rRNA have been located on the small subunit (Figs. 30-25 and 30-32*b*). Their positions were indirectly established from the known positions of proteins that nuclease protection and RNA–protein cross-linking experiments indicate bind to these elements. Thus, we now have a complete, albeit crude, model of the *E. coli* 30S ribosomal subunit.

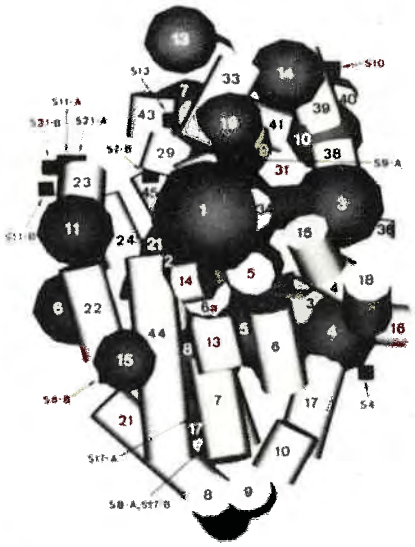
Affinity Labeling Has Helped Identify the Ribosome's Functional Components

Considerable effort has gone into identifying the ribosomes's functional components such as the peptidyl transferase center that catalyzes peptide bond formation (Section 30-3D). Many of these investigations have involved **affinity labeling,** a technique in which a reactive group is attached to a natural ligand of the system of interest such as an antibiotic that binds to the ribosome (Section 30-3G). The reactive group, which may be spontaneously reactive or photolabile so that it only reacts upon UV illumination **(photoaffinity labeling),** is carried to the ligand-binding site where it reacts to cross-link the ligand to the surrounding groups. Dissociation of the resulting particle permits the identification of the components with which the usually radioactive affinity label has reacted.

The results of affinity labeling the ribosome have often been difficult to interpret because its various functions each appear to involve several ribosomal components. For example, mRNA binding apparently involves



(a)



(b)

Figure 30-32
The structure of the 30S ribosomal subunit. (*a*) The relative positions of all 21 proteins of the 30S ribosomal subunit superimposed on its surface outline. Calling Part (*i*) the front view, then Parts (*ii*) and (*iii*) are the left side and bottom views, respectively. The proteins are assigned their standard numbers in Part (*i*) in which S20 is directly behind S3 (the different colors of spheres are only a viewing aid). The distances between pairs of these proteins were determined from the neutron scattering of concentrated solutions of reconstituted 30S subunits in which the two proteins of interest were heavily deuterated while all other subunit components were normally protonated (deuterons scatter neutrons quite differently from protons). Such measurements on many different pairs of proteins permitted the construction of this map in which the volume of each sphere is proportional to the corresponding protein's mass and its position marks the protein's center of mass. Compare this map with Fig. 30-31*a*. [Courtesy of Peter Moore, Yale University and Malcolm Capel, Brookhaven National Laboratory.] (*b*) The locations of the double helical elements of the 16S RNA (*cylinders*) relative to the 30S subunit proteins (*spheres*) as deduced from protein-RNA cross-linking studies. The view is the same as in Figs. 30-31*a* and 30-32(*a*,*i*). [From Schüler, D. and Brinacombe, R., *EMBO J.* **7**, 1512 (1988).]

proteins S1, S3, S4, S5, S9, S12, and S18 as well as the 16S rRNA, whereas proteins L2, L11, L15, L16, L18, L23, and L27, and the 23S RNA are implicated in the peptidyl transferase function. To further confuse matters, studies with mutants deficient in various ribosomal proteins have revealed that the absence of any of at least 15 of the 52 *E. coli* ribosomal proteins does not greatly affect the ribosome's translational ability. Nevertheless, the following functionalities have been located (Figs. 30-23 and 30-31):

1. The 3′ end of the 16S rRNA, which is known to participate in mRNA binding (Section 30-3C), is located on the small subunit's "platform." The locations of the proteins implicated in ribosomal mRNA binding, together with the observation that the ribosome protects an ~40 nucleotide mRNA segment from RNase digestion, indicates that mRNA binds to the small subunit across the region connecting its "head" to its "base."
2. The anticodon-binding sites occur in the small subunit's "cleft" region.
3. The four L7/L12 subunits forming the large subunit's "stalk" participate in the ribosome's various GTPase reactions.
4. The peptidyl transferase function (P) occupies the "valley" between the large subunit's other two protuberances.
5. The site that binds ribosomes to membranes (E; Section 11-3F), occurs on the large subunit adjacent to the polypeptide exit tunnel.

Thus, *the large subunit appears to be mainly involved in mediating biochemical tasks such as catalyzing the reactions of polypeptide elongation, whereas the small subunit is the major actor in ribosomal recognition processes such as mRNA and tRNA binding (although the large subunit is also implicated in tRNA binding).* Note that rRNA probably has a major functional role in many, if not all, ribosomal processes (recall that RNA has demonstrated catalytic properties; Sections 29-4A and C).

Eukaryotic Ribosomes Are Larger and More Complex Than Prokaryotic Ribosomes

Although eukaryotic and prokaryotic ribosomes resemble each other in both structure and function, they

Table 30-7
Components of Rat Liver Cytoplasmic Ribosomes

	Ribosome	Small Subunit	Large Subunit
Sedimentation coefficient	80S	40S	60S
Mass (kD)	4220	1400	2820
RNA			
Major		18S, 1874 nucleotides	28S, 4718 nucleotides
Minor			5.8S, 160 nucleotides
			5S, 120 nucleotides
RNA mass (kD)	2520	700	1820
Proportion of mass	60%	50%	65%
Proteins		33 polypeptides	49 polypeptides
Protein mass (kD)	1700	700	1000
Proportion of mass	40%	50%	35%

Source: Lewin, B., *Genes* (3rd ed.), *p.* 146, Wiley (1987).

differ in nearly all details. Eukaryotic ribosomes have particle masses in the range 3.9 to 4.5 $\times$ 10^6 D and have a nominal sedimentation coefficient of 80S. They dissociate into two unequal subunits that have compositions that are distinctly different from those of prokaryotes (Table 30-7; compare with Table 30-6). The small (40S) subunit of the rat liver cytoplasmic ribosome, the most well-characterized eukaryotic ribosome, consists of 33 unique polypeptides and an 18S rRNA. Its large (60S) subunit contains 49 different polypeptides and three rRNAs of 28S, 5.8S, and 5S. Electron microscopy indicates that these subunits, as well as the intact ribosome, have shapes that are similar to those of their prokaryotic counterparts.

Sequence comparisons of the corresponding rRNAs from various species indicates that evolution has conserved their secondary structures rather than their base sequences (Figs. 30-25 and 30-33). For example, a G·C in a base paired stem of *E. coli* 16S rRNA has been replaced by an A·U in the analogous stem of yeast 18S

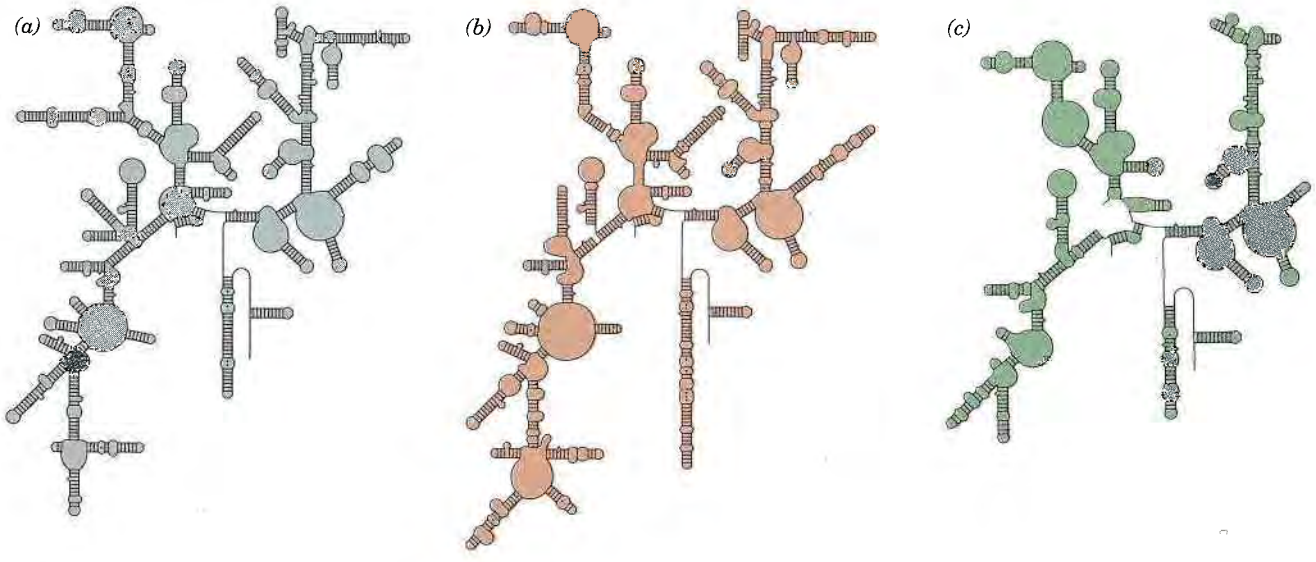


Figure 30-33
The predicted secondary structures of evolutionarily distant 16S-like rRNAs from (*a*) archaebacteria (*Halobacterium volcanii*), (*b*) eukaryotes (baker's yeast), and (*c*) mammalian mitochondria (bovine). Compare them with Fig. 30-25, the predicted secondary structure of 16S RNA from eubacteria (*E. coli*). Note the close similarities of these assemblies; they differ mostly by insertions and deletions of stem-and-loop structures. The 23S-like rRNAs from a variety of species likewise have similar secondary structures. [After Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F., *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 183 (1985).]

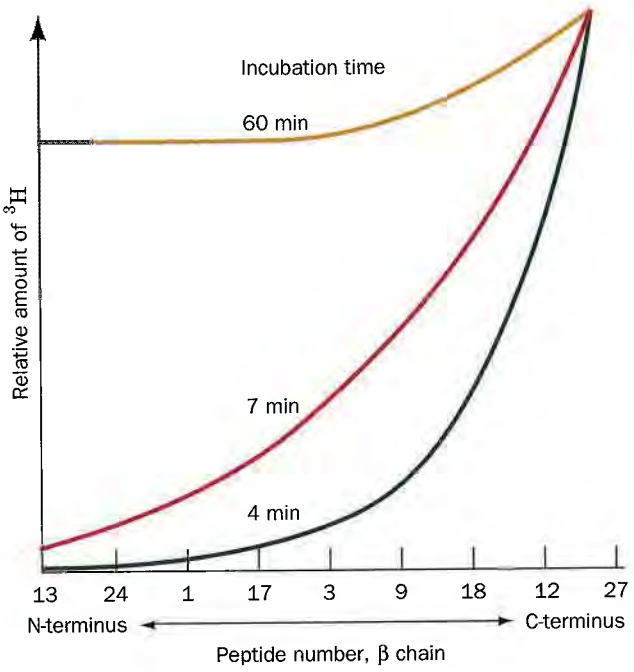


Figure 30-34
Distribution of [^{3}H]Leu among the tryptic peptides from the β subunit of soluble rabbit hemoglobin after the incubation of rabbit reticulocytes with [^{3}H]leucine for the indicated times. [After Dintzis, H. M., *Proc. Natl. Acad. Sci.* **47**, 255 (1961).]

rRNA. The 5.8S rRNA, which occurs in the large eukaryotic subunit in base paired complex with 28S rRNA, is homologous in sequence to the 5′ end of prokaryotic 23S rRNA. Apparently 5.8S RNA arose through mutations that altered rRNA's post-transcriptional processing producing a fourth rRNA.

B. Polypeptide Synthesis: An Overview

Before we commence our detailed discussion of polypeptide synthesis, it will be helpful to outline some of its major features.

Polypeptide Synthesis Proceeds from N-Terminus to C-Terminus

The direction of ribosomal polypeptide synthesis was established, in 1961, by Howard Dintzis through radioactive labeling experiments. He exposed reticulocytes that were actively synthesizing hemoglobin to ^{3}H-labeled leucine for times less than that required to make an entire polypeptide. The extent that the tryptic peptides from the soluble (completed) hemoglobin molecules were labeled increased with their proximity to the C-terminus (Fig. 30-34). Incoming amino acids must therefore be appended to a growing polypeptide's C-terminus; that is, *polypeptide synthesis proceeds from N-terminus to C-terminus.*

Ribosomes Read mRNA in the 5′ → 3′ Direction

The direction that the ribosome reads mRNAs was determined through the use of a cell-free protein synthesizing system in which the mRNA was poly(A) with a 3′-terminal C.

$$5'\ \text{A—A—A—}\cdots\text{—A—A—A—C}\ 3'$$

Such a system synthesizes a poly(Lys) that has a C-terminal Asn.

$$\overset{+}{\text{H}_3\text{N}}\text{—Lys—Lys—Lys—}\cdots\text{—Lys—Lys—Asn—COO}^-$$

This, together with the knowledge that AAA and AAC code for Lys and Asn and the polarity of polypeptide synthesis, indicates that *the ribosome reads mRNA in the 5′ → 3′ direction.* Since mRNA is synthesized in the 5′ → 3′ direction, this accounts for the observation that, in prokaryotes, ribosomes initiate translation on nascent mRNAs (Section 29-3).

Active Translation Occurs on Polyribosomes

Electron micrographs reveal that ribosomes engaged in protein synthesis are tandemly arranged on mRNAs like beads on a string (Figs. 30-35 and 29-16). The indi-

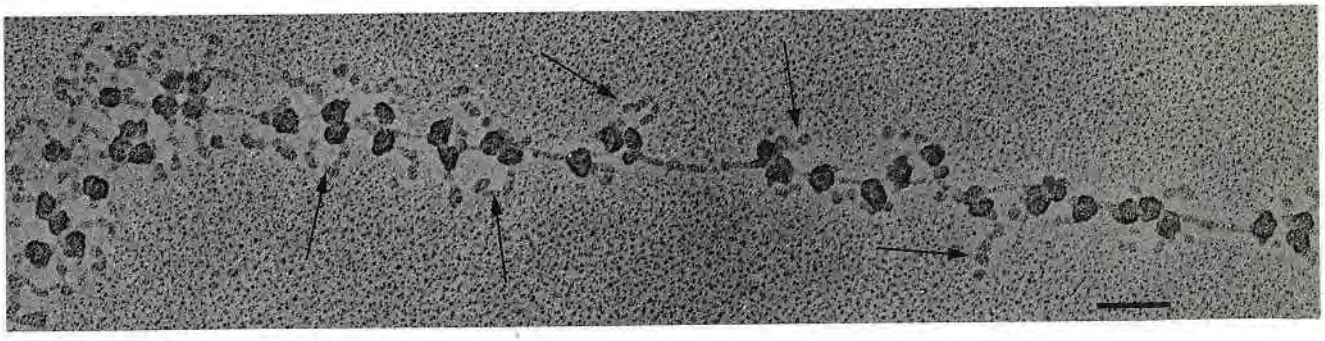
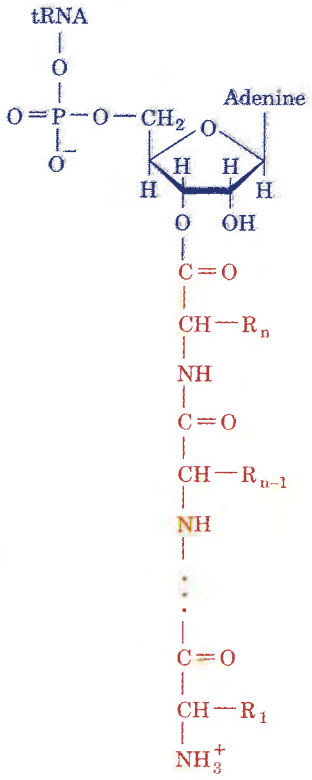


Figure 30-35
Electron micrographs of polysomes from silk gland cells of the silkworm *Bombyx mori.* The 3′ end of the mRNA is on the right. Arrows point to the silk fibroin polypeptides. The bar represents 0.1 μm. [Courtesy of Oscar L. Miller, Jr., University of Virginia.]

vidual ribosomes in these **polyribosomes (polysomes)** are separated by gaps of 50 to 150 Å so that they have a maximum density on mRNA of ~1 ribosome per 80 nucleotides. Polysomes arise because once an active ribosome has cleared its initiation site, a second ribosome can initiate at that site.

Chain Elongation Occurs by the Linkage of the Growing Polypeptide to the Incoming tRNA's Amino Acid Residue

During polypeptide synthesis, amino acid residues are sequentially added to the C-terminus of the nascent, ribosomally bound polypeptide chain. If the growing polypeptide is released from the ribosome by treatment with high salt concentrations, its C-terminal residue is invariably esterified to a tRNA molecule as a **peptidyl-tRNA.**



Peptidyl-tRNA

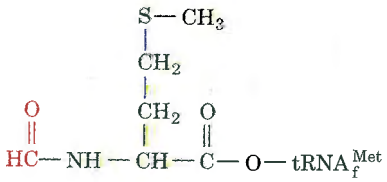
The nascent polypeptide must therefore grow by being transferred from the peptidyl-tRNA to the incoming aminoacyl-tRNA to form a peptidyl-tRNA with one more residue (*Fig. 30-36*). Apparently, the ribosome has at least two tRNA-binding sites: the so-called **P site,** which binds the peptidyl-tRNA, and the **A site,** which binds the incoming aminoacyl-tRNA (Fig. 30-36). Consequently, after the formation of a peptide bond, the newly deacylated P-site tRNA must be released and replaced by the newly formed peptidyl-tRNA from the A site thereby permitting a new round of peptide bond formation. The recent finding that each ribosome can bind up to three deacylated tRNAs but only two aminoacyl-tRNAs indicates, however, that the ribosome has a third tRNA-binding site: the **exit** or **E site,** which transiently binds the outgoing tRNA.

The details of the chain elongation process are discussed in Section 30-3D. Chain initiation and chain termination, which are special processes, are examined in Sections 30-3C and 30-3E, respectively. In all of these sections we shall first consider the process of interest in *E. coli* and then compare it with the analogous eukaryotic activity.

C. Chain Initiation

fMet Is the N-Terminal Residue of Prokaryotic Polypeptides

The first indication that the initiation of translation requires a special codon, since identified as AUG (and, in prokaryotes, occasionally GUG), was the observation that almost one half of the *E. coli* proteins begin with the otherwise uncommon amino acid Met. This was followed by the discovery of a peculiar form of Met-tRNA^{Met} in which the Met residue is *N*-formylated.



Formylmethionine-tRNA$_f^{Met}$ (fMet-tRNA$_f^{Met}$)

The ***N*-formylmethionine** residue **(fMet),** which already has an amide bond, can therefore only be the N-terminal residue of a polypeptide. In fact, polypeptides synthesized in an *E. coli* derived cell-free protein synthesizing system always have a leading fMet residue. *fMet must therefore be E. coli's initiating residue.*

The tRNA that recognizes the initiation codon, **tRNA$_f^{Met}$** (Fig. 30-37), differs from the tRNA that carries internal Met residues, **tRNA$_m^{Met}$,** although they both recognize the same codon. In *E. coli,* uncharged (deacylated) tRNA$_f^{Met}$ is first aminoacylated with Met by the same aminoacyl-tRNA synthetase that charges tRNA$_m^{Met}$. The resulting Met-tRNA$_f^{Met}$ is specifically *N*-formylated to yield fMet-tRNA$_f^{Met}$ in an enzymatic reaction which employs N^{10}-formyltetrahydrofolate (Section 24-4D) as its formyl donor. The formylation enzyme does not recognize Met-tRNA$_m^{Met}$. The X-ray structures of *E. coli* tRNA$_f^{Met}$ and yeast tRNA^{Phe} (Fig. 30-14) are largely similar but differ conformationally in

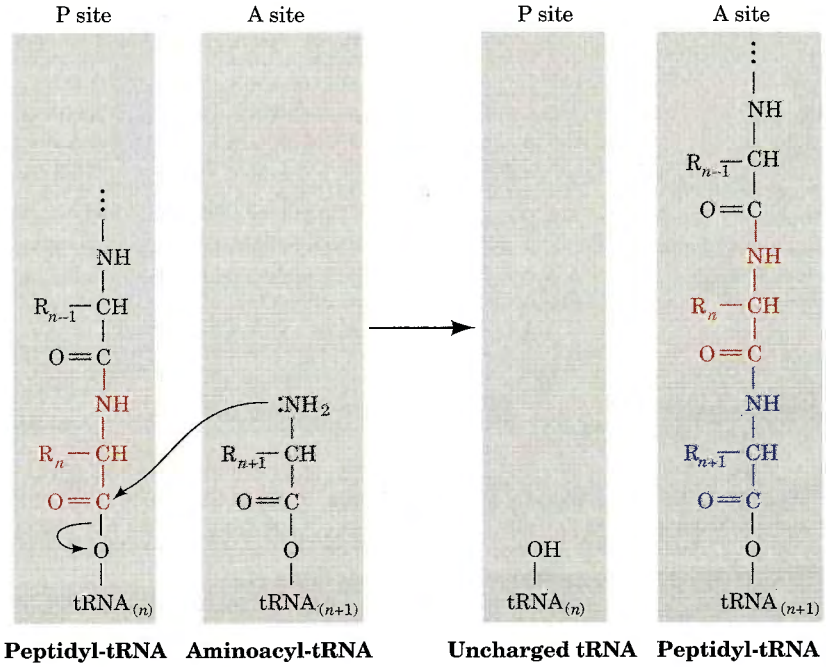


Figure 30-36
The ribosomal peptidyl transferase reaction forming a peptide bond. The amino group of the aminoacyl-tRNA in the A site nucleophilically displaces the tRNA of the peptidyl-tRNA in the P site thereby transferring the nascent polypeptide to the A site tRNA.

their acceptor stems and anticodon loops. Perhaps these structural differences permit $tRNA_f^{Met}$ to be distinguished from $tRNA_m^{Met}$ in the reactions of chain initiation and elongation (see Section 30-3D).

E. coli proteins are post-translationally modified by deformylation of their fMet residue and, in many proteins, by the subsequent removal of the resulting N-terminal Met. This processing usually occurs on the nascent polypeptide, which accounts for the observation that *E. coli* proteins all lack fMet.

Base Pairing between mRNA and the 16S rRNA Helps Select the Translational Initiation Site

AUG codes for internal Met residues as well as the initiating Met residue of a polypeptide. Moreover, mRNAs usually contain many AUGs (and GUGs) in different reading frames. Clearly, *a translational initiation site must be specified by more than just an initiation codon.*

In *E. coli,* the 16S rRNA contains a pyrimidine-rich sequence at its 3′ end. This sequence, as John Shine and Lynn Dalgarno pointed out in 1974, is partially complementary to a purine-rich tract of 3 to 10 nucleotides, the **Shine-Dalgarno sequence,** that is centered ~10 nucleotides upstream from the start codon of nearly all

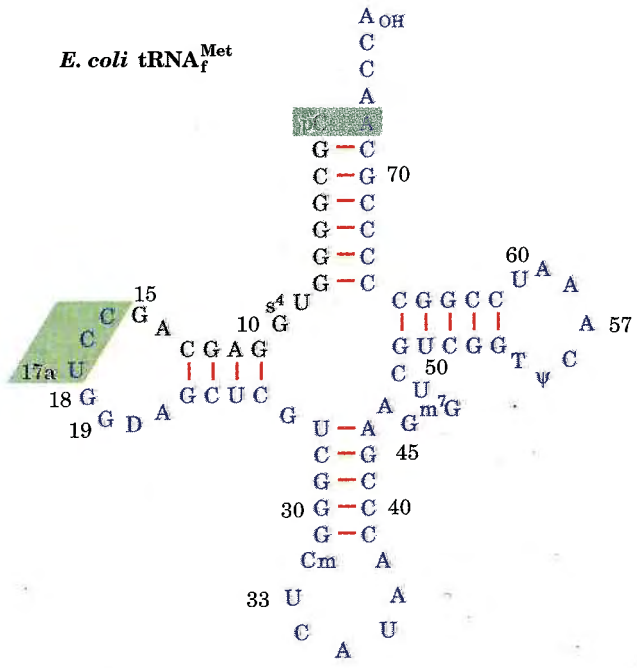


Figure 30-37
The nucleotide sequence of *E. coli* $tRNA_f^{Met}$ shown in cloverleaf form. The shaded boxes indicate the significant differences between this initiator tRNA and noninitiator tRNAs such as yeast $tRNA^{Ala}$ (Fig. 30-11). [After Woo, N. M., Roe, B. A., and Rich, A., *Nature* **286,** 346 (1980).]

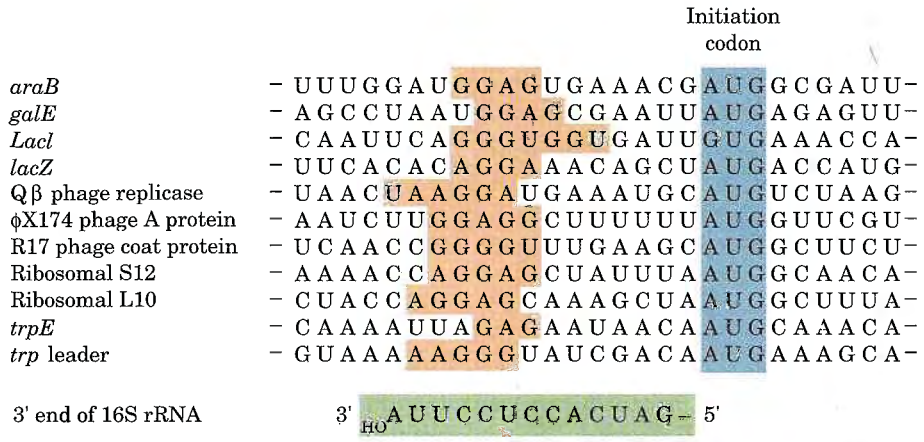


Figure 30-38
Some translational initiation sequences recognized by *E. coli* ribosomes. The mRNAs are aligned according to their initiation codons (*blue shading*). Their Shine–Dalgarno sequences (*red shading*) are complementary, counting G·U pairs, to a portion of the 16S rRNA's 3′ end (*below*). [After Steitz, J. A., *in* Chambliss, G., Craven, G. R., Davies, J., Davis, K., Kahan, L., and Nomura, M. (Eds.), *Ribosomes. Structure, Function and Genetics, pp.* 481–482, University Park Press (1979).]

known prokaryotic mRNAs (Fig. 30-38). *Base pairing interactions between a mRNA's Shine–Dalgarno sequence and the 16S rRNA apparently permit the ribosome to select the proper initiation codon.* Thus ribosomes with mutationally altered anti-Shine–Dalgarno sequences often have greatly reduced ability to recognize natural mRNAs, although they efficiently translate mRNAs whose Shine–Dalgarno sequences have been made complementary to the altered anti-Shine–Dalgarno sequences. Moreover, treatment of ribosomes with the bacteriocidal protein **colicin E3** (produced by *E. coli* strains carrying the E3 plasmid), which specifically cleaves a 49-nucleotide fragment from the 3′ terminus of 16S rRNA, yields ribosomes that cannot initiate new polypeptide synthesis but can complete the synthesis of a previously initiated chain. In fact, when ribosomes that have bound a fragment of **R17 phage** mRNA containing the initiation sequence for its so-called A protein are treated with colicin E3 and then dissociated in 1% SDS, the mRNA fragment is released in complex with the 49-nucleotide rRNA fragment (Fig. 30-39).

Initiation Is a Three-Stage Process that Requires the Participation of Soluble Protein Initiation Factors

Intact ribosomes do not directly bind mRNA so as to initiate polypeptide synthesis. Rather, *initiation is a complex process in which the two ribosomal subunits and fMet–tRNA$_f^{Met}$ assemble on a properly aligned mRNA to form a complex that is competent to commence chain elongation. This assembly process also requires the participation of protein* ***initiation factors*** *that are not permanently associated with the ribosome.* Initiation in *E. coli* involves three initiation factors designated **IF-1, IF-2,** and **IF-3** (Table 30-8). Their existence was discovered when it was found that washing small ribosomal subunits with 1*M* ammonium chloride solution, which removes the initiation factors but not the "permanent" ribosomal proteins, prevents initiation.

The initiation sequence in *E. coli* ribosomes has three stages (Fig. 30-40):

1. Upon completing a cycle of polypeptide synthesis, the 30S and 50S subunits remain associated as inactive 70S ribosomes. IF-3 binds to the 30S subunit so as to promote the dissociation of this complex. IF-1 increases this dissociation rate, perhaps by assisting the binding of IF-3.

2. GTP, mRNA, and a complex of IF-2 with fMet–tRNA$_f^{Met}$ subsequently bind to the 30S subunit in unknown order. Hence, fMet–tRNA$_f^{Met}$ recognition must not be mediated by a codon–anticodon interaction; it is the only tRNA–ribosome association not to require one. This interaction, nevertheless, helps

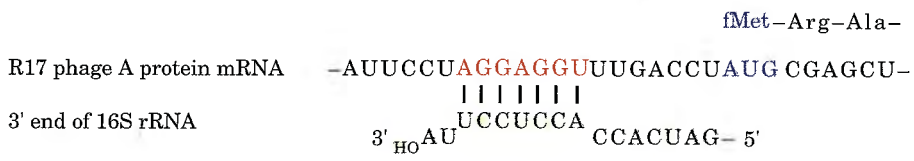


Figure 30-39
Base pairing interactions between the colicin E3 fragment of *E. coli* 16S rRNA and the R17 phage A protein initiator region. [After Steitz, J. A. and Jakes, K., *Proc. Natl. Acad. Sci.* **72,** 4735 (1975).]

Table 30-8
The Soluble Protein Factors of *E. coli* Protein Synthesis

Factor	Mass (kD)	Function
Initiation factors		
IF-1	9	Assists IF-3 binding
IF-2	97	Binds initiator tRNA and GTP
IF-3	22	Releases 30S subunit from inactive ribosome and aids mRNA binding
Elongation factors		
EF-Tu	43	Binds aminoacyl-tRNA and GTP
EF-Ts	74	Displaces GDP from EF-Tu
EF-G	77	Promotes translocation by binding GTP to the ribosome
Release factors		
RF-1	36	Recognizes UAA and UAG Stop codons
RF-2	38	Recognizes UAA and UGA Stop codons
RF-3	46	Binds GTP and stmulates RF-1 and RF-2 binding

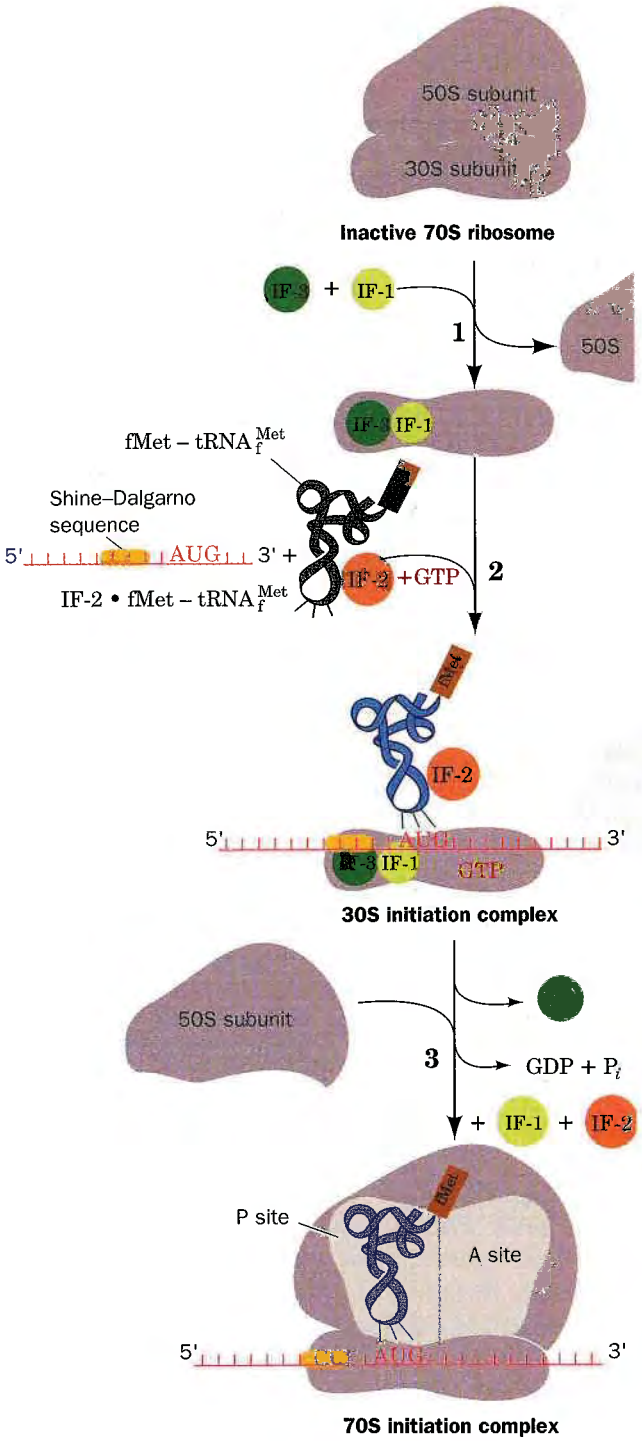


Figure 30-40
The initiation pathway in *E. coli* ribosomes.

bind fMet-tRNA$_f^{Met}$ to the ribosome. IF-3 also functions in this stage of the initiation process: It assists the 30S subunit in binding the mRNA's Shine–Dalgarno sequence.

3. Lastly, in a process that is preceded by IF-3 release, the 50S subunit joins the 30S initiation complex in a manner that hydrolyzes its bound GTP to GDP + P_i. This irreversible reaction conformationally rearranges the 30S subunit and releases IF-1 and IF-2 for participation in further initiation reactions.

Initiation results in the formation of an fMet–tRNA$_f^{Met}$·mRNA·ribosome complex in which the fMet–tRNA$_f^{Met}$ occupies the ribosome's P site while its A site is poised to accept an incoming aminoacyl-tRNA (an arrangement analogous to that at the conclusion of a round of elongation: Section 30-3D). This arrangement was established through the use of the antibiotic **puromycin** as is discussed in Section 30-3D. Note that tRNA$_f^{Met}$ is the only tRNA that directly enters the P site. All other tRNAs must do so via the A site during chain elongation (Section 30-3D).

Eukaryotic Initiation Resembles that of Prokaryotes

Eukaryotic initiation resembles the overall prokaryotic process but differs from it in detail. Ribosomes have a far more extensive "zoo" of initiation factors (designated eIF-*n*; "e" for eukaryotic) than do prokaryotes. Over 10, many with multiple subunits, occur in some eukaryotic systems although they are more difficult to distinguish from ribosomal proteins than are prokaryotic initiation factors.

The most striking difference between eukaryotic and prokaryotic ribosomal initiation occurs in the second

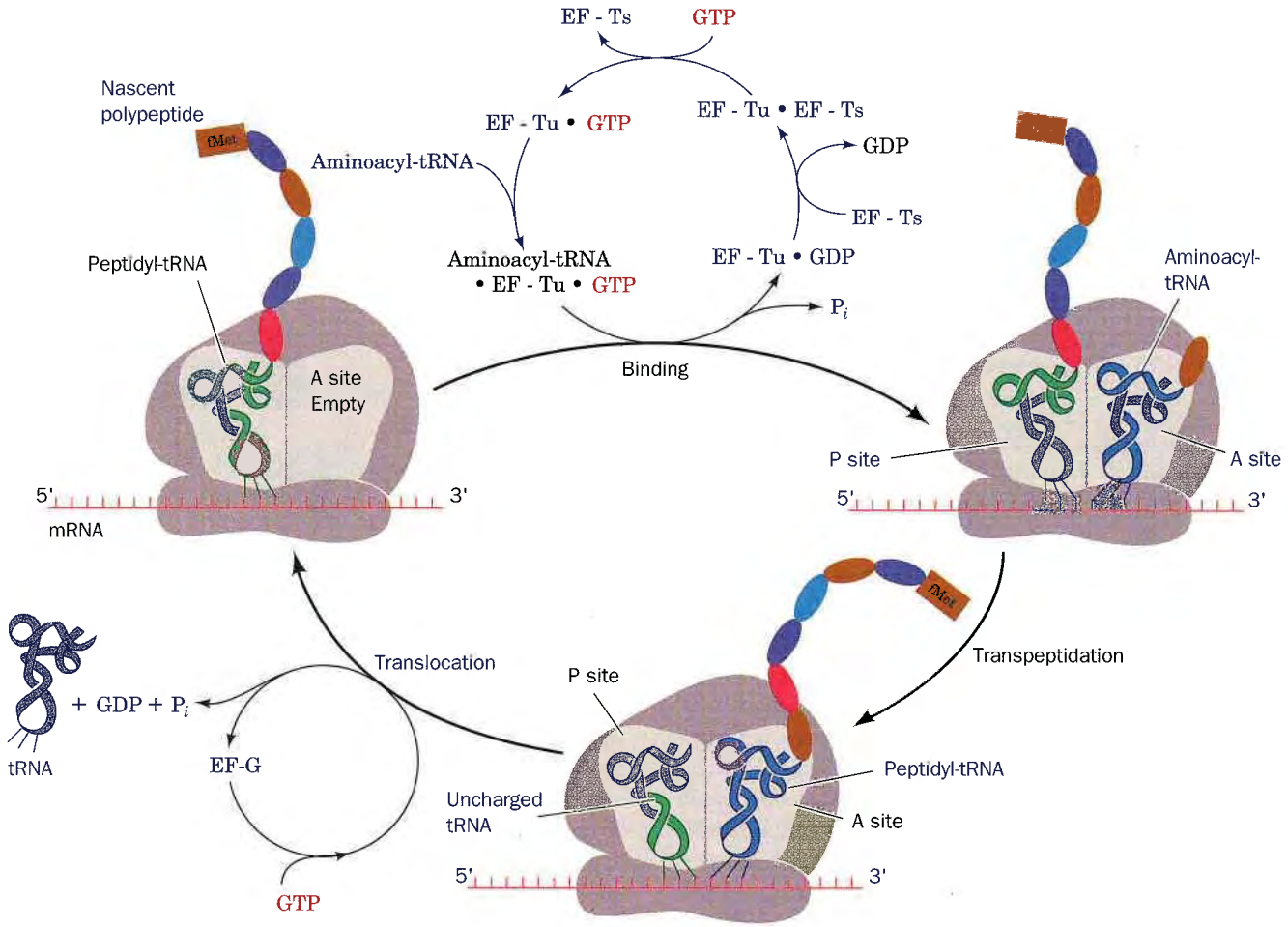


Figure 30-41
The elongation cycle in *E. coli* ribosomes. Eukaryotic elongation follows a similar cycle but EF-Tu and EF-Ts are replaced by a single multisubunit protein, eEF-1, and EF-G is replaced by eEF-2.

stage of the process, the binding of mRNA and a complex of **eIF-2,** GTP, and **Met-tRNA$_i^{Met}$** to the 40S ribosomal subunit (here the subscript "i" distinguishes eukaryotic initiator tRNA, whose appended Met residue is never *N*-formylated, from that of prokaryotes; both species are, nevertheless, readily interchangeable *in vitro*). Eukaryotic mRNAs lack the complementary sequences to bind 18S rRNA in the Shine–Dalgarno manner. Rather, *translation of eukaryotic mRNAs, which are invariably monocistronic, almost always starts at their first AUG.* This, together with the observations that (1) prokaryotic but not eukaryotic ribosomes can inititiate on circular RNAs, and (2) a subunit of **eIF-4F** is a **cap binding protein,** suggests that the 40S subunit binds at or near eukaryotic mRNA's 5′ cap (Section 29-4A) and migrates downstream until it encounters the first AUG. This hypothesis explains the greatly reduced initiation rates of improperly capped mRNAs.

D. Chain Elongation

Ribosomes elongate polypeptide chains in a three-stage reaction cycle that adds amino acid residues to a growing polypeptide's C-terminus (*Fig. 30-41*). This process, which occurs at a rate of up to 40 residues/s, involves the participation of several nonribosomal proteins known as **elongation factors** (Table 30-8).

Aminoacyl-tRNA Binding

In the "binding" stage of the *E. coli* elongation cycle, a binary complex of GTP with the elongation factor **EF-Tu** combines with an aminoacyl-tRNA. The resulting ternary complex binds to the ribosome and, in a reaction that hydrolyzes the GTP to GDP + P_i, the aminoacyl-tRNA is bound in a codon–anticodon complex to the ribosomal A site and EF-Tu·GDP + P_i is released. In the remainder of this stage, which serves to regenerate the EF-Tu·GTP complex, GDP is displaced from EF-Tu·GDP by the elongation factor **EF-Ts** which, in turn, is displaced by GTP.

Aminoacyl-tRNAs can bind to the ribosomal A site without the mediation of EF-Tu but at a rate too slow to support cell growth. The importance of EF-Tu is indicated by the fact that it is the most abundant *E. coli* protein; it is present in ~100,000 copies per cell (>5%

of the cell's protein), which is approximately the number of tRNA molecules in the cell. Consequently, *the cell's entire complement of aminoacyl-tRNAs is essentially sequestered by EF-Tu.*

EF-Tu binds neither formylated nor unformylated Met-tRNA$_f^{Met}$, which is why the initiator tRNA never reads internal AUG or GUG codons. What is the structural basis of this discrimination? *E. coli* tRNA$_f^{Met}$ differs from other *E. coli* tRNAs by the absence of a base pair at the end of its amino acid stem (Fig. 30-37). The conversion of its 5′-terminal C residue to U by bisulfite treatment, which reestablishes the missing base pair as U · A, allows EF-Tu binding. Evidently, EF-Tu recognizes the amino acid stem of noninitiator tRNAs. However, the initiator tRNAs from several other sources do have fully base paired amino acid stems.

Transpeptidation

The peptide bond is formed in the second stage of the elongation cycle through the nucleophilic displacement of the P-site tRNA by the amino group of a 3′-linked aminoacyl-tRNA in the A site (Fig. 30-36). The nascent polypeptide chain is thereby lengthened at its C-terminus by one residue and transferred to the A-site tRNA, a process called **transpeptidation.** The reaction occurs without the need of activating cofactors such as ATP because the ester linkage between the nascent polypeptide and the P-site tRNA is a "high-energy" bond. The peptidyl transferase center that catalyzes peptide bond formation is located entirely on the large subunit as is demonstrated by the observation that in high concentrations of organic solvents such as ethanol, the large subunit alone catalyzes peptide bond formation. The organic solvent apparently distorts the large subunit in a way that mimics the effect of small subunit binding. Peptidyl transferase activity seems to arise from the juxtaposition of several polypeptide chains in the large subunit together with the 23S RNA (Section 30-3A).

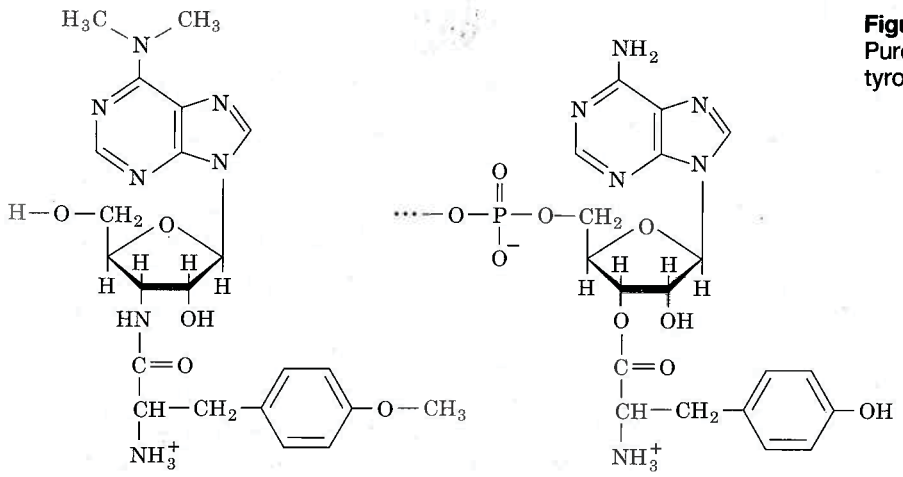
Translocation

In the final stage of the elongation cycle, the now uncharged P-site tRNA (at first tRNA$_f^{Met}$ but subsequently a noninitiator tRNA) is expelled (or, perhaps, transferred to the E site and expelled in the next binding reaction) and, in a process known as ***translocation,*** *the peptidyl-tRNA in the A site, together with its bound mRNA, is moved to the P site.* This prepares the ribosome for the next elongation cycle. The maintenance of the peptidyl-tRNA's codon–anticodon association is no longer necessary for amino acid specification. Rather, it probably acts as a placekeeper that permits the ribosome to precisely step off the three nucleotides along the mRNA required to preserve the reading frame. Indeed, the observation that frameshift suppressor tRNAs induce a four nucleotide translocation (Section 30-2E) indicates that mRNA movement is directly coupled to tRNA movement.

The translocation process requires the participation of an elongation factor, **EF-G,** that binds to the ribosome together with GTP and is only released upon hydrolysis of the GTP to GDP + P_i. EF-G release is prerequiste for beginning the next elongation cycle because the ribosomal binding of EF-G and EF-Tu are mutually exclusive. Translocation is clearly a highly complex mechanical process and, unfortunately, one that is but poorly understood.

Puromycin Is an Aminoacyl-tRNA Analog

The ribosomal elongation cycle was originally characterized through the use of the antibiotic **puromycin** (Fig. 30-42). *This substance, which resembles the 3′ end of Tyr-tRNA, causes the premature termination of polypeptide chain synthesis.* Puromycin, in competition with the specified aminoacyl-tRNA but without the need of elongation factors, binds to the ribosomal A site which, in turn, catalyzes a normal transpeptidation reaction to form peptidyl-puromycin. Yet, the ribosome cannot catalyze the transpeptidation reaction in the next elonga-



Puromycin **Tyrosyl-tRNA**

Figure 30-42
Puromycin (*left*) resembles the 3′ terminus of tyrosyl-tRNA (*right*).

tion cycle because puromycin's "amino acid residue" is linked to its "tRNA" via an amide rather than an ester bond. Polypeptide synthesis is therefore aborted and the peptidyl-puromycin is released.

In the absence of EF-G and GTP, an active ribosome cannot bind puromycin because its A site is already occupied by a peptidyl-tRNA. A newly initiated ribosome, however, violates this rule; it catalyzes fMet-puromycin formation. *These observations demonstrated the functional existence of the ribosomal P and A sites and established that fMet-tRNA$_f^{Met}$ binds directly to the P site, whereas other aminoacyl-tRNAs must first enter the A site.*

The Eukaryotic Elongation Cycle Resembles that of Prokaryotes

The eukaryotic elongation cycle closely resembles that of prokaryotes. In eukaryotes, the functions of EF-Tu and EF-Ts are assumed by two different subunits of the eukaryotic elongation factor **eEF-1.** Likewise **eEF-2** functions in a manner analogous to EF-G. However, the corresponding eukaryotic and prokaryotic elongation factors are not interchangable.

E. Chain Termination

Polypeptide synthesis under the direction of synthetic mRNAs such as poly(U) terminates with a peptidyl-tRNA in association with the ribosome. However, *the translation of natural mRNAs, which contain the termination codons UAA, UGA, or UAG, results in the production of free polypeptides* (*Fig. 30-43*). In *E. coli*, the termination codons, the only codons that normally have no corresponding tRNAs, are recognized by protein **release factors** (Table 30-8): **RF-1** recognizes UAA and UAG, whereas **RF-2** recognizes UAA and UGA. Neither of these release factors can bind to the ribosome simultaneously with EF-G. A third release factor, **RF-3,** which binds GTP, stimulates the ribosomal binding of RF-1 and RF-2. The release factors act at the ribosomal A site as is indicated by the observation that they compete with suppressor tRNAs for termination codons.

The binding of a release factor to the appropriate termination codon induces the ribosomal peptidyl transferase to transfer the peptidyl group to water rather than to an aminoacyl-tRNA (*Fig. 30-44*). The consequent uncharged tRNA subsequently dissociates from the ribosome and the release factors are expelled with the concomitant hydrolysis of GTP to GDP + P_i. The resulting inactive ribosome then releases its bound mRNA preparatory for a new round of polypeptide synthesis.

Termination in eukaryotes resembles that in prokaryotes but requires only a single release factor, **eRF,** that binds to the ribosome together with GTP. This GTP is hydrolyzed to GDP + P_i in a reaction that is thought to trigger eRF's dissociation from the ribosome.

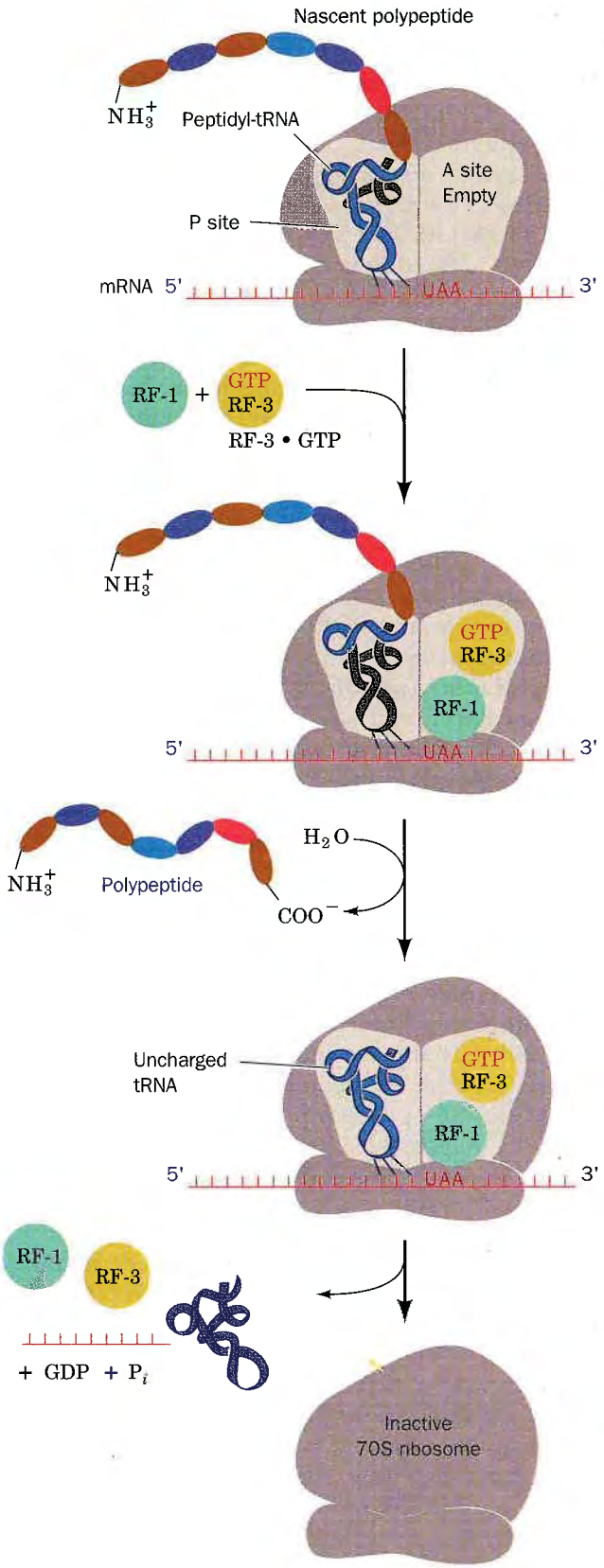


Figure 30-43
The termination sequence in *E. coli* ribosomes. RF-1 recognizes the termination codons UAA and UAG, whereas RF-2 recognizes UAA and UGA. Eukaryotic termination follows an analogous pathway but requires only a single release factor, eRF, that recognizes all three termination codons.

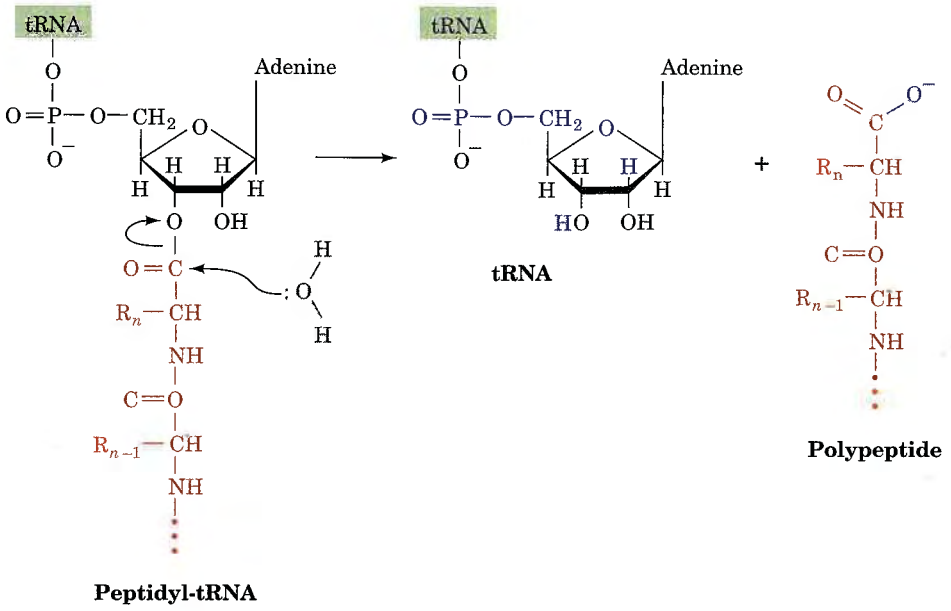


Figure 30-44
The ribosome catalyzed hydrolysis of peptidyl-tRNA to form a polypeptide and free tRNA.

GTP Hydrolysis Speeds Up Ribosomal Processes

What is the role of the various GTP hydrolysis reactions that are essential for normal ribosomal function? Translation occurs in the absence of GTP, albeit extremely slowly, so that the free energy of the transpeptidation reaction is sufficient to drive the entire translational process. Moreover, none of the GTP hydrolysis reactions yield a "high-energy" covalent intermediate as does, say ATP hydrolysis in numerous biosynthetic reactions. It is therefore thought that GTP binding allosterically causes ribosomal components to change their conformations in a way that facilitates a particular process such as translocation. This conformational change also catalyzes GTP hydrolysis which, in turn, permits the ribosome to relax to its initial conformation with the concomitant release of products including GDP + P_i. *The high rate and irreversibility of the GTP hydrolysis reaction therefore ensures that the various complex ribosomal processes to which it is coupled, initiation, elongation, and termination, will themselves be fast and irreversible.* GTP hydrolysis also facilitates translational accuracy (see below).

F. Translational Accuracy

The genetic code is normally translated with remarkable fidelity. We have already seen that transcription and tRNA aminoacylation both proceed with high accuracy (Sections 29-2D and 30-2C). The accuracy of ribosomal mRNA decoding was estimated from the rate of misincorporation of ^{35}S-Cys into highly purified **flagellin,** an *E. coli* protein (Section 34-3G) that normally lacks Cys. These measurements indicated that the mistranslation rate is $\sim 10^{-4}$ errors/per codon. This rate is greatly increased in the presence of **streptomycin,** an antibiotic that increases the rate of ribosomal misreading (Section 30-3G). From the types of reading errors that streptomycin is known to induce, it was concluded that the mistranslation arose almost entirely from the confusion of the Arg codons CGU and CGC for the Cys codons UGU and UGC. The above error rate is therefore largely caused by mistakes in ribosomal decoding.

Aminoacyl-tRNAs are selected by the ribosome only according to their anticodon. Yet, the binding energy loss arising from a single base mismatch in a codon–anticodon interaction is estimated to be ~ 12 kJ·mol^{-1} which, according to Eq. [30.1], cannot account for a ribosomal decoding accuracy of less than $\sim 10^{-2}$ errors per codon. Evidently, the ribosome has some sort of proofreading mechanism that increases its overall decoding accuracy.

How might a ribosome proofread a codon–anticodon interaction? Two types of mechanisms can be envisaged: (1) a selective binding mechanism, such as those of aminoacyl-tRNA synthetases (Section 30-2C); and (2) a kinetic mechanism. The problem with a ribosomal selective binding mechanism is that there is little evidence indicating the existence of a second aminoacyl-tRNA binding site that functions to exclude improper codon–anticodon interactions. Evidence is accumulating, however, that is consistent with a **kinetic proofreading** mechanism.

Kinetic Proofreading Requires a Branched Reaction Path

Kinetic proofreading models of tRNA selection require only one-binding site. John Hopfield theorized that such a process can occur via a branched reaction mechanism of polypeptide chain elongation such as is diagrammed in Fig. 30-45:

1. The initial-binding reaction discriminates, as discussed, between cognate (specified) and noncognate (unspecified) tRNAs according to their codon–anticodon binding energies.
2. Following this, the bound GTP irreversibly hydrolyzes yielding an activated intermediate as is indicated by the star. This complex can then react in one of two ways:
   - **(a)** EF-Tu·GDP can dissociate from the ribosome with rate constant k_3 thereby committing the ribosome to form a peptide bond.
   - **(b)** Alternatively, the aminoacyl-tRNA can dissociate from the ribosome with rate constant k_4 thereby aborting the elongation step. EF-Tu·GDP subsequently dissociates from the ribosome permitting it to reinitiate the elongation step.

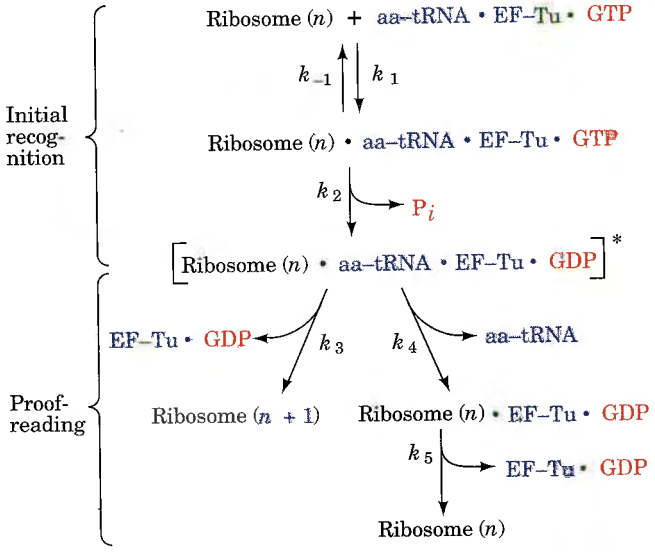


Figure 30-45
A kinetic proofreading mechanism for selecting a correct codon–anticodon interaction. The initial recognition reaction screens the aminoacyl-tRNA (aa-tRNA) for the correct codon–anticodon interaction. The resulting complex converts, in a GTP-driven process, to a "high-energy" intermediate (*) which, in turn, either releases EF-Tu·GDP preparatory to forming a peptide bond or releases aminoacyl-tRNA before EF-Tu·GDP is released. If k_4/k_3 is greater for a codon–anticodon mismatch than it is for a match, then these latter steps constitute a proofreading mechanism for proper tRNA binding.

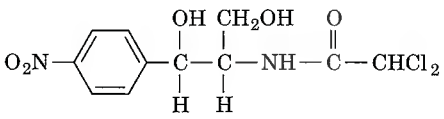
Table 30-9
Some Ribosomal Inhibitors

Inhibitor	Action
Chloramphenicol	Inhibits peptidyl transferase on the prokaryotic large subunit
Cycloheximide	Inhibits peptidyl transferase on the eukaryotic large subunit
Erythromycin	Inhibits translocation by the prokaryotic large subunit
Fusidic acid	Inhibits elongation in prokaryotes by preventing EF-G·GDP dissociation from the large subunit
Puromycin	An aminoacyl-tRNA analog that causes premature chain termination in prokaryotes and eukaryotes
Streptomycin	Causes mRNA misreading and inhibits chain initiation in prokaryotes
Tetracycline	Inhibits the binding of aminoacyl-tRNAs to the prokaryotic small subunit
Diphtheria toxin	Catalytically inactivates eEF-2 by ADP-ribosylation
Ricin/Abrin	Poisonous plant proteins that catalytically inactivate the eukaryotic large subunit

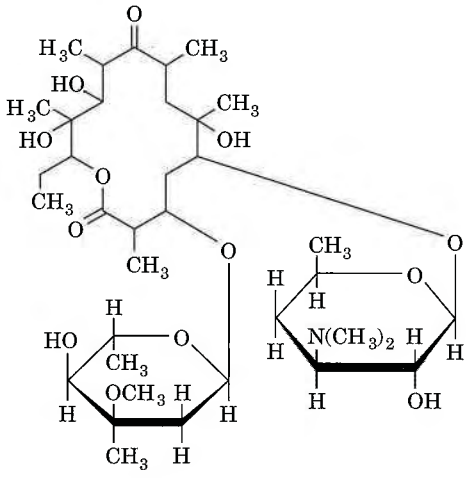
If the ratio k_4/k_3 is greater for noncognate than for cognate tRNAs, then a second screening will have occurred. It is thought that the physical basis of this second screening is that k_3 is independent of the tRNA's identity, whereas k_4 is larger for a relatively weakly bound noncognate tRNA than it is for a cognate tRNA. The rate of EF-Tu·GDP dissociation therefore provides a countdown clock against which the ribosome measures the rate of tRNA dissociation: A noncognate tRNA usually dissociates from the ribosome before EF-Tu·GDP does (an average period of several milliseconds), whereas a cognate tRNA usually remains bound. The activated intermediate is essential for this process because otherwise its tRNA dissociation step (that characterized by k_4) would be identical to that of the initial recognition step (that characterized by k_{-1}). GTP hydrolysis therefore provides the second context necessary for proofreading.

The kinetic proofreading model is supported by the following observations:

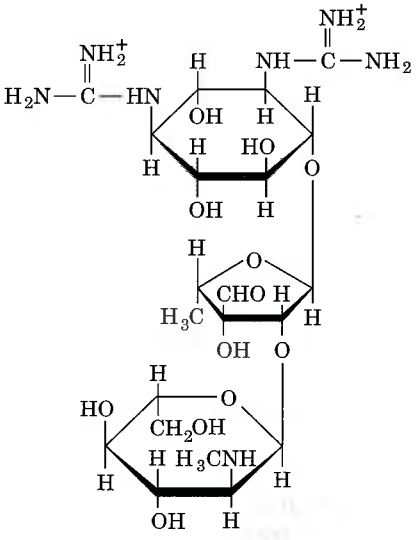
1. More GTP is hydrolyzed per peptide bond formed with noncognate than with cognate tRNAs (although this observation is also consistent with selective binding models).



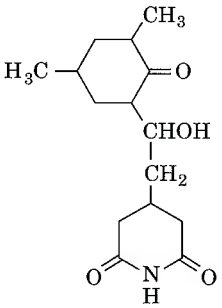
Chloramphenicol



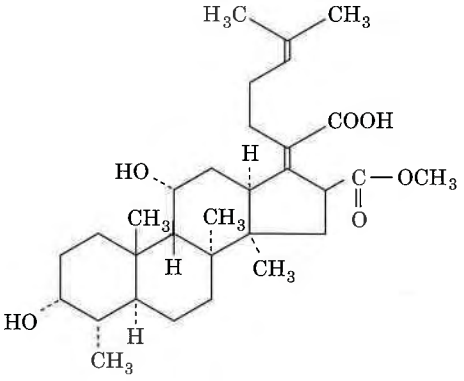
Erythromycin



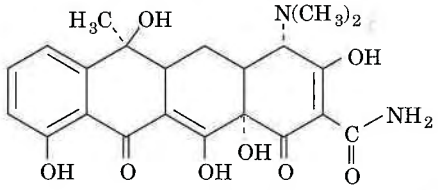
Streptomycin



Cycloheximide



Fusidic acid



Tetracycline

Figure 30-46
A selection of antibiotics that act as translational inhibitors.

2. The rate of EF-Tu · GDP dissociation from a ribosome is, in fact, independent of its bound aminoacyl-tRNA's identity.

G. Protein Synthesis Inhibitors: Antibiotics

Antibiotics are bacterially or fungally produced substances that inhibit the growth of other organisms. Antibiotics are known to inhibit a variety of essential biological processes including DNA replication (e.g., novobiocin, Section 28-5C), transcription (e.g., rifamycin B; Section 29-2C), and bacterial cell wall synthesis (e.g., penicillin; Section 10-3B). However, *the majority of known antibiotics, including a great variety of medically useful substances, block translation.* This situation is presumably a consequence of the translational machinery's enormous complexity, which makes it vulnerable to disruption in many ways. Antibiotics have also been useful in analyzing ribosomal mechanisms because, as we have seen for puromycin (Section 30-3D), the blockade of a specific function often permits its biochemical dissection into its component steps. Table 30-9 and Fig. 30-46 present

several medically significant and/or biochemically useful translational inhibitors. We study the mechanisms of a few of the best characterized of them below.

Streptomycin

Streptomycin, which was discovered in 1944 by Selman Waksman, is a medically important member of a family of antibiotics known as **aminoglycosides** that inhibit prokaryotic ribosomes in a variety of ways. At low concentrations, streptomycin induces the ribosome to characteristically misread mRNA: One pyrimidine may be mistaken for the other in the first and second codon positions and either pyrimidine may be mistaken for adenine in the first position. This inhibits the growth of susceptible cells but does not kill them. At higher concentrations, however, streptomycin prevents proper chain initiation and thereby causes cell death.

Certain streptomycin resistant mutants (str^R) have ribosomes with an altered protein S12 compared with streptomycin sensitive bacteria (str^S). Intriguingly, a change in base C912 of 16S rRNA (which lies in the central loop in Fig. 30-25) also confers streptomycin resistance. (Some mutant bacteria are not only resistant to streptomycin but dependent on it; they require it for growth.) In partial diploid bacteria that are heterozygous for streptomycin resistance (str^R/str^S), streptomycin sensitivity is dominant. This puzzling observation is explained by the finding that, in the presence of streptomycin, str^S ribosomes remain bound to initiation sites thereby excluding str^R ribosomes from these sites. Moreover, the mRNAs in these blocked complexes are degraded after a few minutes, which alows the str^S ribosomes to bind to newly synthesized mRNAs as well.

Chloramphenicol

Chloramphenicol, the first of the "broad-spectrum" antibiotics, inhibits the peptidyl transferase activity on the large subunit of prokaryotic ribosomes. However, its clinical uses are limited to only severe infections because of its toxic side effects which are caused, at least in part, by the chloramphenicol sensitivity of mitochondrial ribosomes. Binding experiments with reconstituted 50S subunits that are missing one or another component suggest that protein L16 is necessary for chloramphenicol binding. This is corroborated by affinity labeling experiments indicating that L16, as well as several other large subunit proteins and the 23S RNA, are in proximity to bound chloramphenicol. The 23S RNA is also implicated in chloramphenicol resistance by the observation that some of its mutants are chloramphenicol resistant. Chloramphenicol's binding-site must lie near the large subunit's A site since chloramphenicol competes for binding with puromycin and the 3′ end of aminoacyl-tRNAs but not with peptidyl-tRNAs. This observation suggests that chloramphenicol inhibits peptidyl transfer by interfering with the interactions of ribosomes with A site-bound aminoacyl-tRNAs.

Tetracycline

Tetracycline and its derivatives are broad-spectrum antibiotics that bind to the small subunit of prokaryotic ribosomes where they inhibit aminoacyl-tRNA binding. Tetracycline also blocks the stringent response (Section 29-3F) by inhibiting ppGpp synthesis. This indicates that deacylated tRNA must bind to the A site in order to activate stringent factor.

Tetracycline-resistant bacterial strains have become quite common thereby precipitating a serious clinical problem. Most often, however, resistance is conferred by a decrease in bacterial cell membrane permeability to tetracycline rather than any alteration of ribosomal components.

Diphtheria Toxin

Diphtheria is a disease that results from bacterial infection by *Corynebacterium diphtheriae* that harbor the bacteriophage **corynephage *β*.** Diphtheria was a leading cause of childhood death until early in this century when immunization became prevalent. Although the bacterial infection is usually confined to the upper respiratory tract, the bacteria secrete a phage-encoded protein, known as **diphtheria toxin,** that is responsible for the disease's lethal effects. *Diphtheria toxin specifically inactivates the eukaryotic elongation factor eEF-2 thereby inhibiting eukaryotic protein synthesis.*

The pathogenic effects of diphtheria are prevented, as was discovered in the 1880s, by immunization with **toxoid,** formaldehyde inactivated toxin. Individuals who have contracted diphtheria are treated with antitoxin from horse serum, which binds to and thereby inactivates diphtheria toxin, as well as with antibiotics to combat the bacterial infection.

Diphtheria toxin acts in a particularly interesting way. It is a monomeric 58-kD protein that is readily cleaved by trypsin and trypsin-like enzymes into two fragments, A and B. The B domain of intact toxin binds to an unknown receptor on the plasma membrane of susceptible cells. The toxin is then proteolytically cleaved whereupon the B fragment facilitates the A fragment's cytosolic uptake via receptor-mediated endocytosis (free fragment A is devoid of toxic activity).

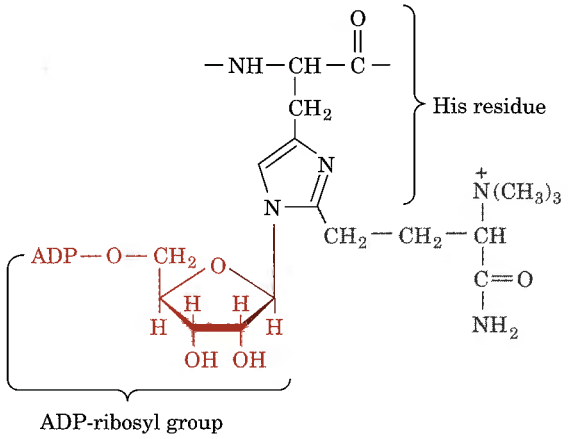
Within the cytosol, the A fragment catalyzes the **ADP-ribosylation** of eEF-2 by NAD^+,

$$\underset{\text{(active)}}{\text{eEF-2}} + \text{NAD}^+ \xrightarrow{\text{diphtheria toxin}} \underset{\text{(inactive)}}{\text{ADP-ribosyl-eEF-2}} + \text{Nicotinamide} + \text{H}^+$$

thereby inactivating this elongation factor. Since the A fragment acts catalytically, *one molecule is sufficient to ADP-ribosylate all of a cell's eEF-2s, which halts protein synthesis and kills the cell.* Only a few micrograms of

diphtheria toxin are therefore sufficient to kill an unimmunized individual.

Diphtheria toxin specifically ADP-ribosylates a modified His residue on eEF-2 known as **diphthamide:**



ADP-Ribosylated diphthamide

Diphthamide occurs only in eEF-2 (not even in its bacterial counterpart, EF-G), which accounts for the specificity of diphtheria toxin in exclusively modifying eEF-2. The observation that diphthamide occurs in all eukaryotic eEF-2's suggests that this residue is essential to eEF-2 activity. Yet, certain mutant cultured animal cells, which have unimpaired capacity to synthesize proteins, lack the enzymes that post-translationally modify His to diphthamide. Diphthamide's normal biological role is therefore a mystery.

4. CONTROL OF EUKARYOTIC TRANSLATION

The rates of ribosomal initiation on prokaryotic mRNAs vary by factors of up to 100. For example, the proteins specified by the *E. coli lac* operon, β-galactosidase, galactose permease, and thiogalactoside transacetylase, are synthesized in molar ratios of 10:5:2. This variation is probably a consequence of their different Shine–Dalgarno sequences. Alternatively, ribosomes may attach to *lac* mRNA only at its β-galactosidase gene and occasionally detach in response to a chain termination signal (thereby accounting for the decreasing translational rates along the operon). At any rate, there is no evidence that prokaryotic translation rates are responsive to environmental changes. *Genetic expression in prokaryotes is therefore almost entirely transcriptionally controlled (Section 29-3).* Of course, since their mRNAs have lifetimes of only a few minutes, it would seem that prokaryotes have little need of translational controls.

Eukaryotic transcriptional control, although far more complex than that in prokaryotes, is largely reserved for regulating cell differentiation (Section 33-3). There are, however, increasing indications that eukaryotic cells respond to their needs, at least in part, through translational control. This is feasible because the lifetimes of eukaryotic mRNAs are generally hours or days. In this section, we examine the two best-characterized eukaryotic translational control mechanisms: (1) the regulation of hemoglobin synthesis by heme, and (2) the effects of virus-induced proteins known as **interferons.** We also consider the phenomenon of mRNA masking.

A. Translational Control by Heme

Reticulocytes synthesize protein, almost exclusively hemoglobin, at an exceedingly high rate and are therefore a favorite subject for the study of eukaryotic translation. Hemoglobin synthesis in fresh reticulocyte lysates proceeds normally for several minutes but then abruptly stops because of the inhibition of translational initiation and the consequent polysome disaggregation. This effect is prevented by the addition of heme [a mitochondrial product (Section 24-4A) that this *in vitro* system cannot synthesize] thereby indicating that *globin synthesis is regulated by heme availability.* The inhibition of globin translational initiation is also reversed by the addition of the eukaryotic initiation factor eIF-2 and by high levels of GTP.

In the absence of heme, reticulocyte lysates accumulate a protein, **heme-controlled inhibitor (HCI),** that phosphorylates a specific Ser residue on the α subunit of eIF-2 (eIF-2 is an αβγ trimer that carries GTP and Met-tRNA$_i^{Met}$ to the 40S ribosomal subunit; Section 30-3C). HCI is generated, in the absence of heme, from a preexisting proinhibitor by a poorly characterized process that probably involves at least one other protein.

Phosphorylated eIF-2 can participate in the ribosomal initiation process in much the same way as unphosphorylated eIF-2. This puzzling situation was clarified by the discovery that GDP does not spontaneously dissociate from eIF-2 at the completion of initiation as it does from IF-2 in the corresponding prokaryotic process (Fig. 30-40). Rather, eIF-2 exchanges its GDP for GTP in a reaction mediated by another initiation factor, **eIF-2B** (Fig. 30-47). It turns out that phosphorylated eIF-2 forms a much tighter complex with eIF-2B than does unphosphorylated eIF-2. This sequesters eIF-2B (Fig.

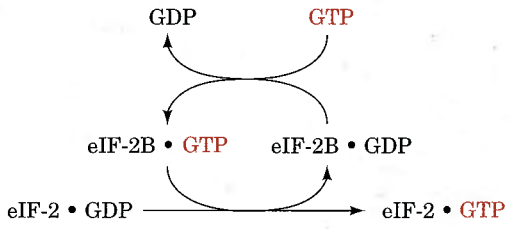


Figure 30-47
The eIF-2·GDP product of eukaryotic ribosomal initiation is regenerated by GDP–GTP interchange with eIF-2B·GTP.

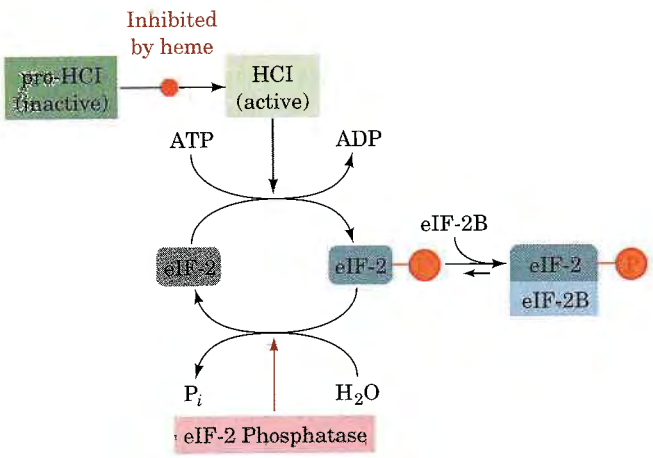


Figure 30-48
A model for heme-controlled protein synthesis in reticulocytes.

30-48), which is present in lesser amounts than is eIF-2, thereby preventing regeneration of the eIF-2·GTP required for translational initiation. The presence of heme reverses translational inhibition by inhibiting HCI. The phosphorylated eIF-2 molecules are reactivated through the action of **eIF-2 phosphatase,** which is unaffected by heme. Reticulocytes, in addition, contain a recently discovered 67-kD protein that protects eIF-2 from HCI-catalyzed phosphorylation.

B. Interferon

***Interferons** are glycoproteins that are secreted by virus infected vertebrate cells. Upon binding to surface receptors of other cells, interferons convert them to an antiviral state, which impairs the replication of a wide variety of RNA and DNA viruses.* Indeed, the discovery of interferon in the 1950s arose from the observation that virus-infected individuals are resistant to infection by a second type of virus.

There are three families of interferons: **type α** or **leucocyte interferon** (leucocytes are white blood cells), the related **type β** or **fibroblast interferon** (fibroblasts are connective tissue cells), and **type γ** or **lymphocyte interferon** (lymphocytes are immune system cells). *Interferon synthesis is induced by double-stranded RNA (dsRNA), which is probably generated during infection by both DNA and RNA viruses, as well as by the synthetic dsRNA poly(I)·poly(C).* Interferons are effective antiviral agents in concentrations as low as $3 \times 10^{-14} M$, which makes them among the most potent biological substances. Moreover, they have far wider specificities than antibodies raised against a particular virus. They have therefore elicited great medical interest, particularly since some cancers are virally induced (Section 33-4C). Indeed, they are in clinical use against certain tumors and viral infections. These treatments are made possible by the production of large quantities of these otherwise quite scarce proteins through molecular cloning techniques (Section 28-8).

Interferons prevent viral proliferation largely by inhibiting protein synthesis in infected cells (lymphocyte interferon also modulates the immune response). They do so in two independent ways (Fig. 30-49):

1. Interferons induce the production of a protein kinase that, in the presence of dsRNA, phosphorylates the eIF-2 α subunit identically to the action of HCI in reticulocytes, thereby inhibiting ribosomal initiation. This observation suggests that eIF-2 phosphorylation may be a general mechanism of eukaryotic translational control.
2. Interferons also induce the synthesis of **(2′,5′)-**

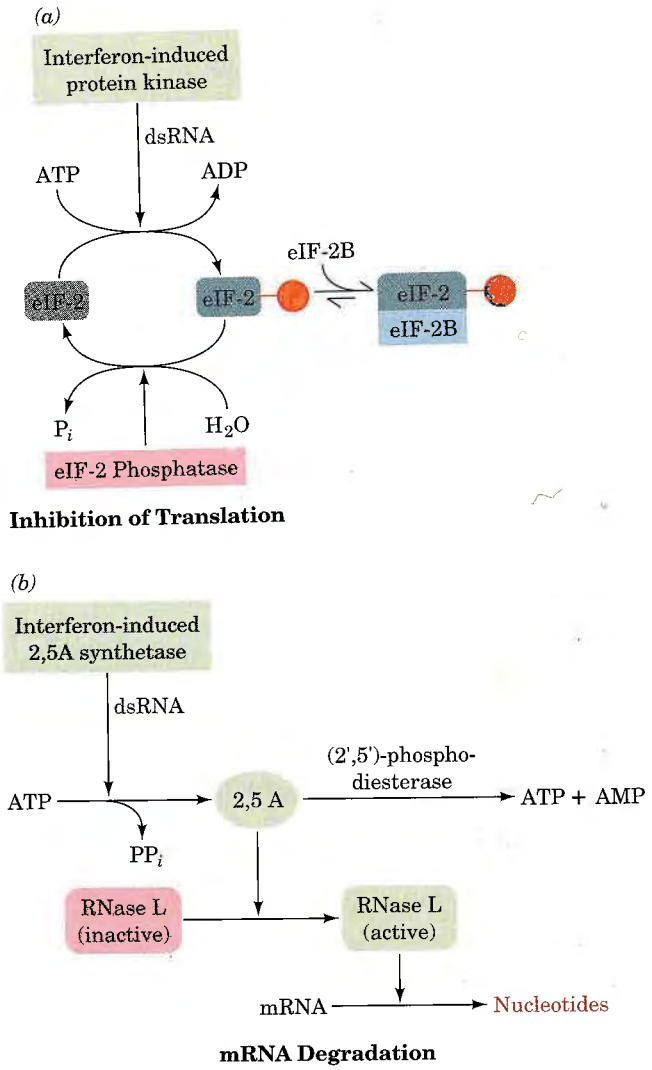


Figure 30-49
In interferon treated cells, the presence of dsRNA, which normally results from a viral infection, causes (*a*) the inhibition of translational initiation, and (*b*) the degradation of mRNA, thereby blocking translation and preventing virus replication.

oligoadenylate synthetase. In the presence of dsRNA, this enzyme catalyzes the synthesis from ATP of the unusual oligonucleotide **pppA(2′p5′A)$_n$** where $n = 1$ to 10. *This compound, 2,5-A, activates a preexisting endonuclease, RNase L, to degrade mRNA thereby inhibiting protein synthesis.* 2,5-A is itself rapidly degraded by an enzyme named **(2′,5′)-phosphodiesterase** so that it must be continually synthesized to maintain its effect.

The independence of the 2,5-A and the interferon-induced protein kinase systems is demonstrated by the observation that the effect of 2,5-A on protein synthesis is reversed by added mRNA but not by eIF-2.

C. mRNA Masking

It has been known since the previous century that early embryonic development in organisms such as sea urchins is governed almost entirely by information present in the egg before fertilization. Indeed, sea urchin embryos exposed to sufficient actinomycin D (Section 29-2D) to inhibit RNA synthesis without blocking DNA synthesis develop normally through their early stages without a change in their protein synthesis program. This is because *an unfertilized egg contains large quantities of mRNA that is "masked" by associated proteins so as to prevent its association with the ribosomes that are also present. Upon fertilization, this mRNA is somehow "unmasked" in a controlled fashion and commences directing protein synthesis.* Development of the embryo can therefore start immediately upon fertilization rather than waiting for the generation of paternally specified mRNAs.

The cytoplasms of many eukaryotic cells contain large amounts of protein-complexed mRNAs that are not associated with ribosomes. It remains to be seen, however, whether mRNA masking is used for translational control in nonembryonic tissues.

5. POST-TRANSLATIONAL MODIFICIATION

To become mature proteins, polypeptides must fold to their native conformations, their disulfide bonds, if any, must form, and, in the case of multisubunit proteins, the subunits must properly combine. Moreover, as we have seen throughout this text, many proteins are modified in enzymatic reactions that proteolytically cleave certain peptide bonds and/or derivatize specific residues. In this section we shall review some of these **post-translational modifications.**

A. Proteolytic Cleavage

Proteolytic cleavage is the most common type of post-translational modification. Probably all mature proteins have been so-modified, if by nothing else than the endoproteolytic removal of their leading Met (or fMet) residue shortly after it emerges from the ribosome. Many proteins, which are involved in a wide variety of biological processes, are synthesized as inactive precursors that are activated under proper conditions by limited proteolysis. Some examples of this phenomenon that we have encountered are the conversion of trypsinogen and chymotrypsinogen to their active forms by tryptic cleavages of specific peptide bonds (Section 14-3E), and the formation of active insulin from the 84-residue proinsulin by excision of an internal 33-residue polypeptide (Section 8-1A). Inactive proteins that are activated by removal of polypeptides are called **proproteins,** whereas the excised polypeptides are termed **propeptides.**

Propeptides Direct Collagen Assembly

Collagen biosynthesis is illustrative of many facets of post-translational modification. Recall that collagen, a major extracellular component of connective tissue, is a fibrous triple helical protein whose polypeptides largely consist of the repeating amino acid sequence (Gly-X-Y)$_n$ where X is often Pro, Y is often 4-hydroxyproline (Hyp), and $n \approx 340$ (Section 7-2C). The polypeptides of **procollagen** (Fig. 30-50) differ from those of the mature protein by the presence of both N- and C-terminal propeptides of ~100 residues whose sequences, for the

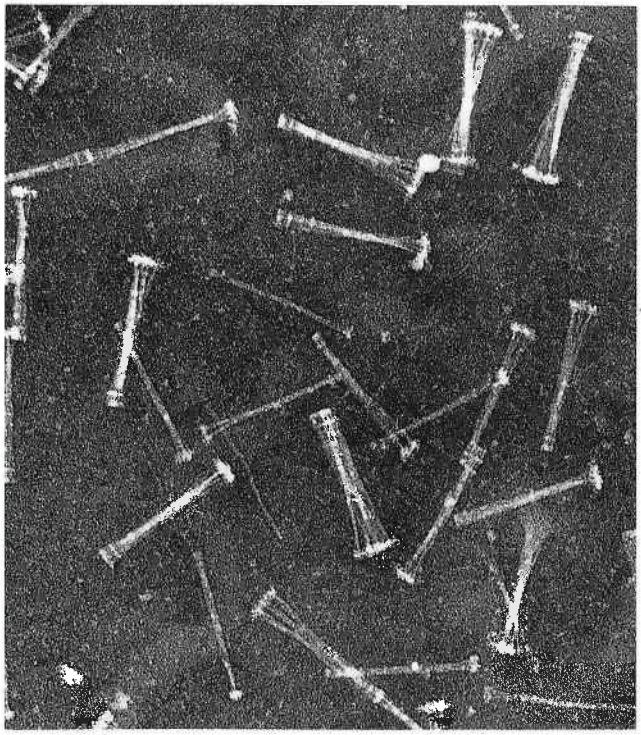


Figure 30-50
An electron micrograph of procollagen aggregates that have been secreted into the extracellular medium. [Courtesy of Jerome Gross, Harvard Medical School.]

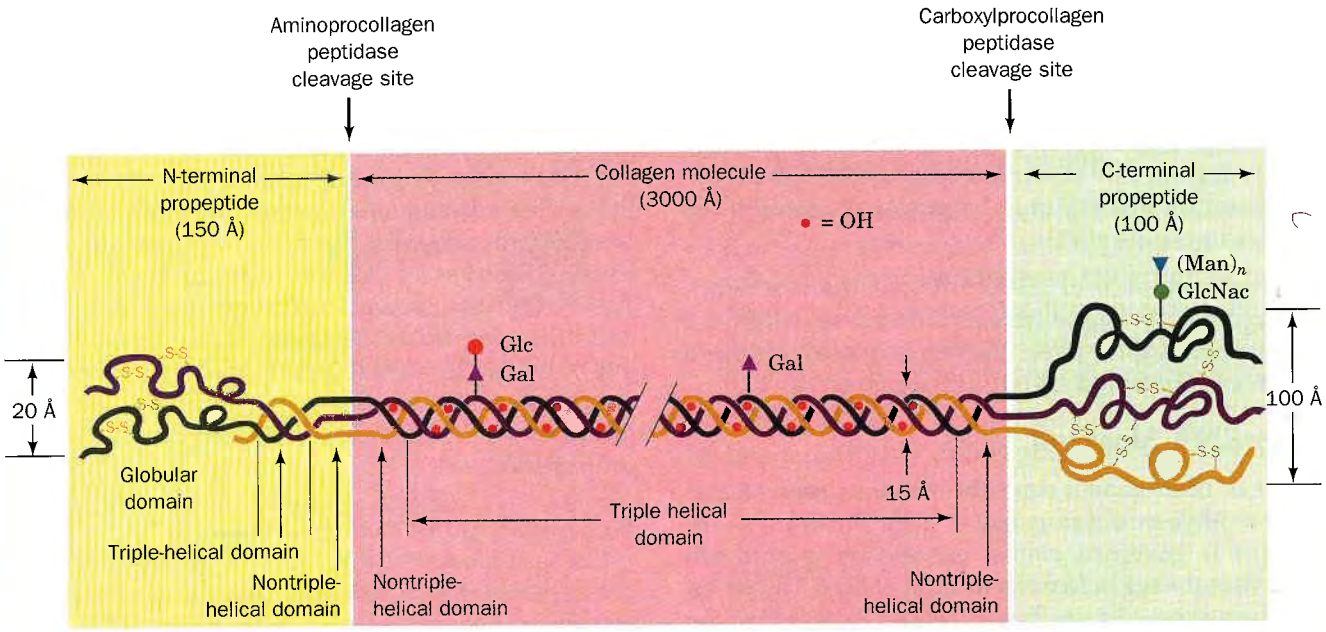


Figure 30-51
A schematic representation of the procollagen molecule. Gal, Glc, GlcNac, and Man, respectively, denote galactose, glucose, *N*-acetylglucosamine, and mannose residues. Note that the aminopropeptide has intrachain disulfide bonds while the carboxylpropeptide has both intrachain and interchain disulfide bonds. [After Prockop, D. J., Kivirikko, K. I., Tuderman, L., and Guzman, N. A., *New Engl. J. Med.* **301,** 16 (1979).]

most part, are unlike those of mature collagen. The procollagen polypeptides rapidly assemble, *in vitro* as well as *in vivo,* to form a collagen triple helix. In contrast, polypeptides extracted from mature collagen will only reassemble over a period of days, if at all. *The collagen propeptides are apparently necessary for proper procollagen folding.*

The N- and C-terminal propeptides of procollagen are removed by **amino-** and **carboxylprocollagen peptidases** (Fig. 30-51), which may also be specific for the different collagen types. An inherited defect of aminoprocollagen peptidase in cattle and sheep results in a bizarre condition, **dermatosparaxis,** that is characterized by extremely fragile skin. An analogous disease in man, **Ehler-Danlos syndrome VII,** is caused by a mutation in one of the procollagen polypeptides that inhibits the enzymatic removal of its aminopropeptide. Collagen molecules normally spontaneously aggregate to form collagen fibrils (Figs. 7-33 and 7-34). However, electron micrographs of dermatosparaxic skin show sparse and disorganized collagen fibrils. *The retention of collagen's aminopropeptides apparently interferes with proper fibril formation.* (The dermatosparaxis gene was bred into some cattle herds because heterozygotes produce tender meat.)

Signal Peptides Are Removed from Nascent Proteins by a Signal Peptidase

Many trans-membrane proteins or proteins that are destined to be secreted are synthesized with an N-terminal **signal peptide** of 13 to 36 predominantly hydrophobic residues. According to the **signal hypothesis** (Section 11-3F), a signal peptide is recognized by a **signal recognition particle (SRP).** The SRP binds a ribosome synthesizing a signal peptide to a receptor on the membrane [the rough endoplasmic reticulum (RER) in eukaryotes and the plasma membrane in bacteria] and conducts the signal peptide and the following nascent polypeptide through it.

Proteins bearing a signal peptide are known as **preproteins** or, if they also contain propeptides, as **preproproteins.** Once the signal peptide has passed through the membrane, it is specifically cleaved from the nascent polypeptide by a **signal peptidase.** Both insulin and collagen are secreted proteins and are therefore synthesized with leading signal peptides in the form of **preproinsulin** and **preprocollagen.** These and many other proteins are therefore subject to three sets of sequential proteolytic cleavages: (1) the deletion of their initiating Met residue, (2) the removal of their signal peptides, and (3) the excision of their propeptides.

Polyproteins

Some proteins are synthesized as segments of **polyproteins,** polypeptides that contain the sequences of two or more proteins. Examples include most polypeptide hormones (Section 33-3C), the proteins synthesized by many viruses including those causing polio (Section 32-2C) and AIDS, and **ubiquitin,** a highly conserved eukaryotic protein involved in protein degradation

(Section 30-6B). Specific proteases post-translationally cleave polyproteins to their component proteins, presumably through the recognition of the cleavage site sequences. Some of these proteases are conserved over remarkable evolutionary distances. For instance, ubiquitin is synthesized as several tandem repeates (polyubiquitin) that *E. coli* properly cleave even though prokaryotes lack ubiquitin.

B. Covalent Modification

Proteins are subject to specific chemical derivatizations, both at the functional groups of their side chains and at their terminal amino and carboxyl groups. Over 150 different types of side chain modifications, involving all side chains but those of Ala, Gly, Ile, Leu, Met, and Val, are known (Section 4-3A). These include acetylations, glycosylations, hydroxylations, methylations, nucleotidylylations, phosphorylations, and ADP-ribosylations as well as numerous "miscellaneous" modifications.

Some protein modifications, such as the phosphorylation of glycogen phosphorylase (Section 17-1A) and the ADP-ribosylation of eEF-2 (Section 30-3G), modulate protein activity. Several side chain modifications convalently bond cofactors to enzymes, presumably to increase their catalytic efficiency. Examples of linked cofactors that we have encountered are N^{ϵ}-lipoyllysine in dihydrolipoyl transacetylase (Section 19-2A) and 8α-histidylflavin in succinate dehydrogenase (Section 19-3F). The attachment of complex carbohydrates, which occur in almost infinite variety, alter the structural properties of proteins and form recognition markers in various types of targeting and cell–cell interactions (Sections 10-3C, 11-3D, and 21-3B). Modifications that cross-link proteins, such as occur in collagen and elastin (Sections 7-2C and D), stabilize supramolecular aggregates. The functions of most side chain modifications, however, remain enigmatic.

Collagen Assembly Requires Chemical Modification

Collagen biosynthesis (Fig. 30-52) is illustrative of protein maturation through chemical modification. As the nascent procollagen polypeptides pass into the RER of the fibroblasts that synthesized them, the Pro and Lys residues are hydroxylated to Hyp, 3-hydroxy-Pro, and 5-hydroxy-Lys. The enzymes that do so are sequence specific: **prolyl 4-hydroxylase** and **lysyl hydroxylase** act only on the Y residues of the Gly-X-Y sequences, whereas **prolyl 3-hydroxylase** acts on the X residues but only if Y is Hyp. Glycosylation, which also occurs in the RER, subsequently attaches sugar residues to 5-hydroxy-Lys residues (Section 7-2C). The folding of three polypeptides into the collagen triple helix must follow hydroxylation and glycosylation because the hydroxylases and glycosylases do not act on helical substrates. Moreover, the collagen triple helix denatures below physiological temperatures unless stabilized by hydrogen bonding interactions involving Hyp residues (Section 7-2C). Folding is also preceded by the formation of specific interchain disulfide bonds between the carboxylpropeptides. This observation bolsters the previously discussed conclusion that collagen propeptides help select and align the three collagen polypeptides for proper folding.

The procollagen molecules pass into the Golgi apparatus where they are packaged into secretory granules (Sections 11-3F and 21-3B) and secreted into the extracellular spaces of connective tissue. The aminopropeptides are excised just after procollagen leaves the cell and

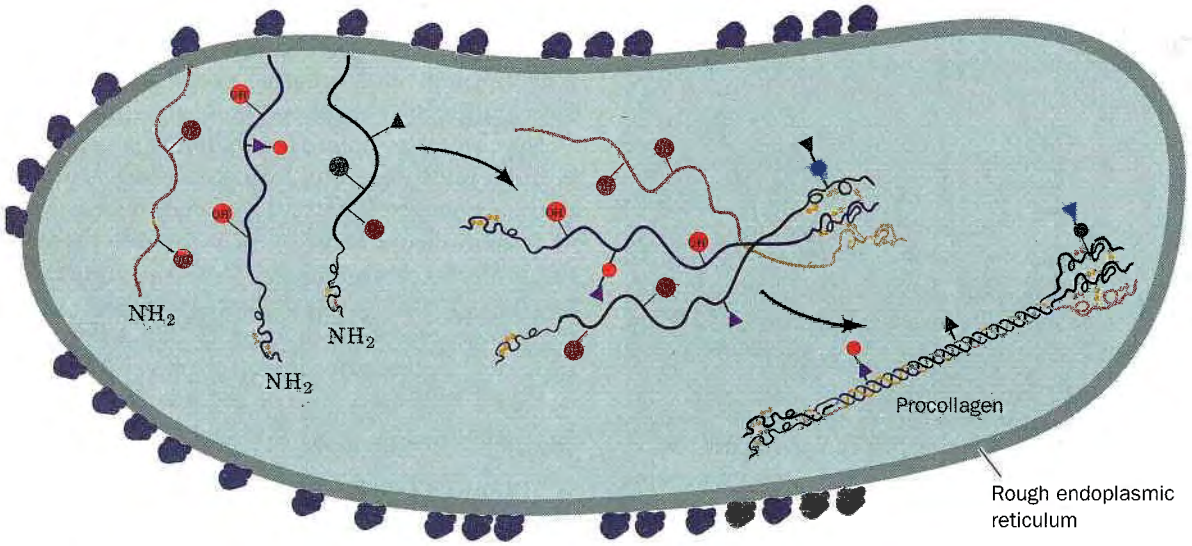


Figure 30-52
A schematic representation of procollagen biosynthesis. The diagram does not indicate the removal of signal peptides.

[After Prockop, D. J., Kivirikko, K. I., Tuderman, L., and Guzman, N. A., *New Engl. J. Med.* **301,** 18 (1979).]

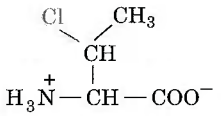
the carboxylpropeptides are removed sometime later. The collagen molecules then spontaneously assemble into fibrils, which suggests that an important propeptide function is to prevent intracellular fibril formation. Finally, after the action of the extracellular enzyme lysyl oxidase, the collagen molecules in the fibrils spontaneously cross-link (Fig. 7-35).

6. PROTEIN DEGRADATION

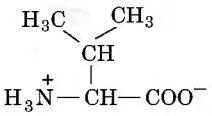
The pioneering work of Henry Borsook and Rudolf Schoenheimer around 1940 demonstrated that the components of living cells are constantly turning over. Proteins have lifetimes that range from as short as a few minutes to weeks or more. In any case, *cells continuously synthesize proteins from and degrade them to their component amino acids.* The function of this seemingly wasteful process is twofold: (1) to eliminate abnormal proteins whose accumulation would be harmful to the cell, and (2) to permit the regulation of cellular metabolism by eliminating superfluous enzymes and regulatory proteins. Indeed, since the level of an enzyme depends on its rate of degradation as well as its rate of synthesis, *controlling a protein's rate of degradation is as important to the cellular economy as is controlling its rate of synthesis.* In this section we consider the processes of intracellular protein degradation and their consequences.

A. Degradation Specificity

Cells selectively degrade abnormal proteins. For example, hemoglobin that has been synthesized with the valine analog α-amino-β-chlorobutyrate



α-Amino-β-Chorobutyrate



Valine

has a half-life in reticulocytes of ~10 min, whereas normal hemoglobin lasts the 120-day lifetime of the red cell (which makes it perhaps the longest lived cytoplasmic protein). Likewise, unstable mutant hemoglobins are degraded soon after their synthesis which, for reasons explained in Section 9-3A, results in the hemolytic anemia characteristic of these molecular disease agents. Bacteria also selectively degrade abnormal proteins. For instance, *amber* and *ochre* mutants of β-galactosidase have half-lives in *E. coli* of only a few minutes, whereas the wild-type enzyme is almost indefinitely stable. Most abnormal proteins, however, probably arise from the chemical modification and/or spontaneous denaturation of these fragile molecules in the cell's reactive environment rather than by mutations or the rare errors in transcription or translation. *The ability to eliminate damaged proteins selectively is therefore an essential recycling mechanism that prevents the buildup of substances that would otherwise interfere with cellular processes.*

Table 30-10
Half-Lives of Some Rat Liver Enzymes

Enzyme	Half-Life (h)
Short-lived enzymes	
Ornithine decarboxylase	0.2
RNA polymerase I	1.3
Tyrosine aminotransferase	2.0
Serine hydratase	4.0
PEP carboxylase	5.0
Long-lived enzymes	
Aldolase	118
GAPDH	130
Cytochrome *b*	130
LDH	130
Cytochrome *c*	150

Source: Dice, J. F. and Goldberg, A. L., *Arch. Biochem. Biophys.* 170, 214 (1975).

Normal intracellular proteins are eliminated at rates that depend on their identities. A given protein is eliminated with first-order kinetics indicating that the molecules being degraded are chosen at random rather than according to their age. The half-lives of different enzymes in a given tissue vary substantially as is indicated for rat liver in Table 30-10. Remarkably, *the most rapidly degraded enzymes all occupy important metabolic control points, whereas the relatively stable enzymes have nearly constant catalytic activities under all physiological conditions. The susceptibilities of enzymes to degradation have evidently evolved together with their catalytic and allosteric properties so that cells can efficiently respond to environmental changes and metabolic requirements.* The criteria through which native proteins are selected for degradation are considered in Section 30-6B.

The rate of protein degradation in a cell also varies with its nutritional and hormonal state. Under conditions of nutritional deprivation, cells increase their rate of protein degradation so as to provide the necessary nutrients for indispensible metabolic processes. The mechanism that increases degradative rates in *E. coli* is the stringent response (Section 29-3F). A similar mechanism may be operative in eukaryotes since, as happens in *E. coli*, increased rates of degradation are prevented by antibiotics that block protein synthesis.

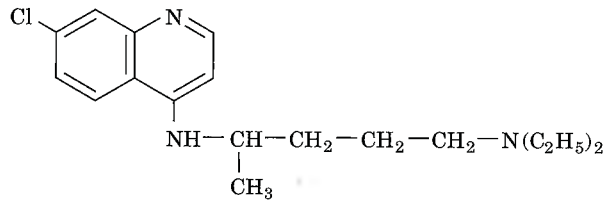
B. Degradation Mechanisms

Eukaryotic cells have dual systems for protein degradation, a lysosomal mechanism and an ATP-dependent cytosolically based mechanism. We consider both mechanisms below.

Lysosomes Degrade Proteins Nonselectively

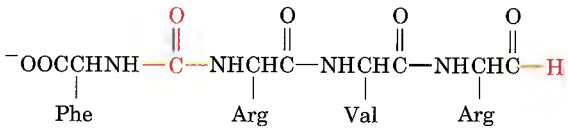
Lysosomes are membrane-encapsulated organelles (Section 1-2A) that contain ~50 hydrolytic enzymes, including a variety of proteases known as **cathepsins.** The lysosome maintains an internal pH of ~5 and its enzymes have acidic pH optima. This situation presumably protects the cell against accidental lysosomal leakage since lysosomal enzymes are largely inactive at cytosolic pH's.

Lysosomes recycle intracellular constituents by fusing with membrane-enclosed bits of cytoplasm known as **autophagic vacuoles** and subsequently breaking down their contents. They similarly degrade substances that the cell takes up via endocytosis (Section 11-4B). The existence of these processes has been demonstrated through the use of lysosomal inhibitors. For example, the antimalarial drug **chloroquine**



Chloroquine

is a weak base that freely penetrates the lysosome in uncharged form where it accumulates in charged form thereby increasing the intralysosomal pH and inhibiting lysosomal function. The treatment of cells with chloroquine reduces their rate of protein degradation. Similar effects arise from treatment of cells with cathepsin inhibitors such as the polypeptide antibiotic **antipain.**



Lysosomal protein degradation appears to be nonselective. Lysosomal inhibitors do not affect the rapid degradation of abnormal proteins or short-lived enzymes. However, they prevent the acceleration of protein breakdown upon starvation.

Many normal and pathological processes are associated with increased lysosomal activity. **Diabetes mellitus** (Section 25-3B) stimulates the lysosomal breakdown of proteins. Similarly, muscle wastage caused by disuse, denervation, or traumatic injury arises from increased lysosomal activity. The regression of the uterus after childbirth, in which this muscular organ reduces its mass from 2 kg to 50 g in 9 days, is a striking example of this process. Many chronic inflammatory diseases, such as **rheumatoid arthritis,** involve the extracellular release of lysosomal enzymes which break down the surrounding tissues.

Ubiquitin Marks Proteins Selected for Degradation

It was initially assumed that protein degradation in eukaryotic cells is primarily a lysosomal process, Yet, reticulocytes, which lack lysosomes, selectively degrade abnormal proteins. The observation that protein breakdown is inhibited under anaerobic conditions led to the discovery of a cytosolically based ATP-dependent proteolytic system that is independent of the lysosomal system. This phenomenon was thermodynamically unexpected since peptide hydrolysis is an exergonic process.

Analysis of a cell-free rabbit reticulocyte system has demonstrated that **ubiquitin,** a protein of previously unknown function (Fig. 30-53), is required for ATP-dependent protein degradation. *This 76-residue monomeric protein, so-named because it is ubiquitous as well as abundant in eukaryotes, is the most highly conserved protein known:* It is identical in such diverse organisms as humans, toad, trout, and *Drosophila,* and differs in only three residues between humans and yeast. Evidently, ubiquitin is all but uniquely suited for some essential cellular process.

Proteins that are selected for degradation are so-marked by covalently linking them to ubiquitin. This process, which is reminiscent of amino acid activation (Section 30-2C), occurs in three steps (Fig. 30-54):

1. In an ATP-requiring reaction, ubiquitin's terminal carboxyl group is conjugated, via a thioester bond, to **ubiquitin-activating enzyme (E1),** a 105-kD dimer of identical subunits.

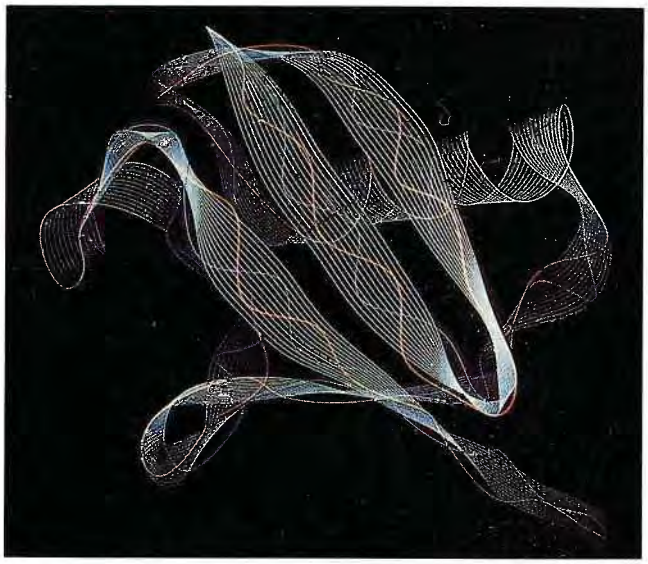


Figure 30-53
The X-ray structure of ubiquitin. The white ribbon represents the polypeptide backbone and the red and blue curves, respectively, indicate the directions of the carbonyl and amide groups. [Courtesy of Michael Carson, University of Alabama at Birmingham. X-ray structure determined by Charles Bugg.]

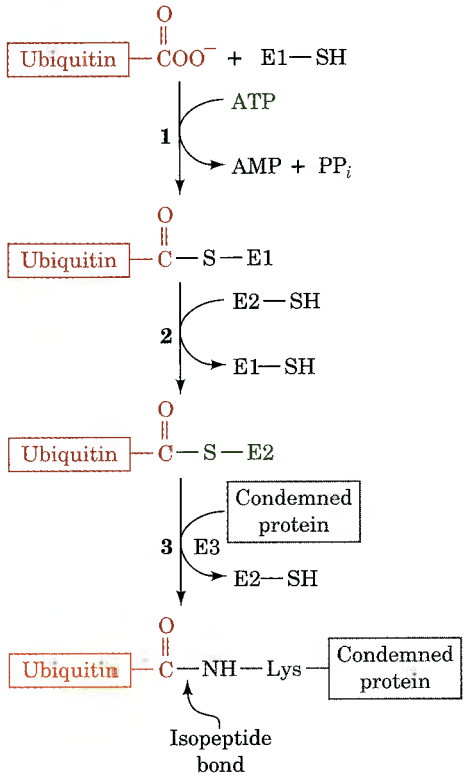


Figure 30-54
The reactions involved in the attachment of ubiquitin to a protein. In the first part of the process, ubiquitin's terminal carboxyl group is joined, via a thioester linkage, to E1 in a reaction driven by ATP hydrolysis. The activated ubiquitin is subsequently transferred to a sulfhydryl group of E2 and then, in a reaction catalyzed by E3, to a Lys ε-amino group on a condemned protein thereby flagging the protein for proteolytic degradation by UCDEN.

Table 30-11
The Half-Lives of Cytoplasmic Enzymes as a Function of Their N-Terminal Residues

N-Terminal Residue	Half-Life
Stabilizing	
Met	>20 h
Ser	
Ala	
Thr	
Val	
Gly	
Destabilizing	
Ile	~30 min
Glu	
Tyr	~10 min
Gln	
Highly destabilizing	
Phe	~3 min
Leu	
Asp	
Lys	
Arg	~2 min

Source: Bachmair, A., Finley, D., and Varshavsky, A., *Science* **234**, 180 (1986).

2. The ubiquitin is then transferred to a sulfhydryl group of one of several small proteins (25–70 kD) named **ubiquitin-carrier proteins (E2's).**

3. Finally, **ubiquitin-protein ligase (E3;** ~180 kD) transfers the activated ubiquitin from E2 to a Lys ε-amino group of a previously bound protein thereby forming an **isopeptide bond.** E3 therefore appears to have a key role in selecting the protein to be degraded. Usually, several ubiquitin molecules are so-linked to this condemned protein. In addition, as many as 20 ubiquitin molecules may be tandemly linked to a target protein to form a multiubiquitin chain in which Lys 48 of each ubiquitin forms an isopeptide bond with the C-terminal carboxyl group of the following ubiquitin.

The ubiquitinated protein is proteolytically degraded in an ATP-dependent process mediated by a large (≥1000 kD) but otherwise poorly characterized multiprotein complex named ***ubiquitin-conjugate degrading enzyme (UCDEN).*** This protease only degrades ubiquitin-linked proteins.

A Protein's Half-Life Is Partially Determined by Its N-Terminal Residue

The structural features that E3 uses to select at least native proteins for destruction may be remarkably simple. *The half-life of a cytoplasmic protein varies with the identity of its N-terminal residue (Table 30-11).* Indeed, in a selection of 208 cytoplasmic proteins known to be long lived, all have a "stabilizing" residue, Met, Ser, Ala, Thr, Val, Gly, or Cys, at their N-termini. This is true for both eukaryotes and prokaryotes, which suggests the system that selects proteins for degradation is conserved in eukaryotes and prokaryotes, even though prokaryotes lack ubiquitin. Nevertheless, there are clear indications that other, more complex signals are also important in the selection of proteins for degradation. For instance, proteins with segments rich in Pro (P), Glu (E), Ser (S), and Thr (T) residues are rapidly degraded although how these so-called **PEST proteins** are recognized is unknown. Likewise, the criteria by which cells select defective proteins for degradation are unknown.

7. NONRIBOSOMAL POLYPEPTIDE SYNTHESIS

Several hundred polypeptide antibiotics, such as actinomycin D (Section 29-2D) and gramicidin A (Section 18-2C), have been characterized. These often cyclic molecules consist of rarely more than 20, often unusual, amino acid residues. *Many polypeptide antibiotics are synthesized by soluble enzymes rather than ribosomally from mRNA templates.* The synthesis of these substances is, consequently, unaffected by ribosomal inhibitors such as chloramphenicol (Section 30-3G) that arrest protein synthesis. In this section, we consider the mechanism of biosynthesis of the channel-forming ionophore **gramicidin S,** which is representative of the synthesis of many other polypeptide antibiotics.

Gramicidin S, a product of *Bacillus brevis,* is a cyclic decapeptide that consists of two identical pentapeptides joined head to tail (Fig. 30-55). Fritz Lipmann demonstrated that this antibiotic is synthesized by two enzymes, E_I (280 kD) and E_{II} (100 kD) that activate the amino acids indicated in Fig. 30-56. Each of the amino acids of gramicidin S is activated by the ATP-driven linkage of the amino acid via a thioester bond to its corresponding enzyme. E_{II} binds only a D-Phe residue, whereas E_I simultaneously binds the other four gramicidin S residues.

The polymerization process begins when E_{II} transfers its D-Phe residue to the E_I-bound Pro residue to form a dipeptide (Fig. 30-57*a*). The growing oligopeptide is then sequentially transferred to the remaining amino acid residues of the pentapeptide (Fig. 30-57*b*). The absence of any amino acid from the *in vitro* reaction mixture results in premature termination of the reaction at

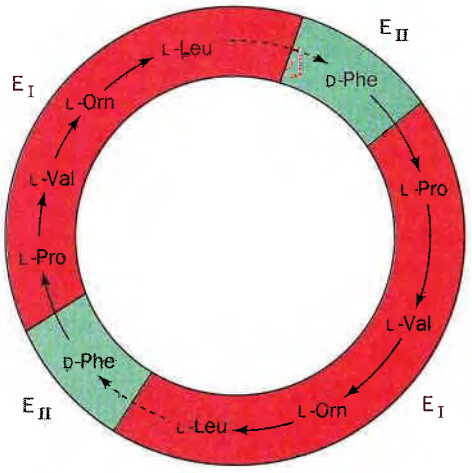


Figure 30-55
The amino acid sequence of gramicidin S. The amino acids activated by E_I and E_{II} are shaded red and green. Dashed arrows indicate the points of cyclization. [After Lipmann, F., *Acc. Chem. Res.* **11,** 363 (1971).]

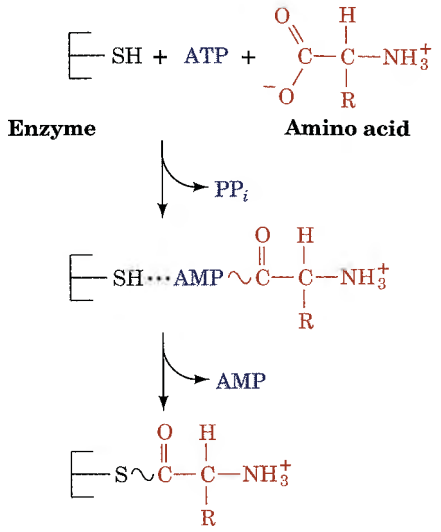


Figure 30-56
The activation of amino acids by the enzymes that synthesize gramicidin S. In the first step of the reaction an enzyme bound aminoacyl-adenylate is formed as it is in the aminoacyl-tRNA synthetase reaction (Section 30-2C). In the second reaction step, however, the amino acid residue is linked to the enzyme via a thioester bond rather than to a tRNA.

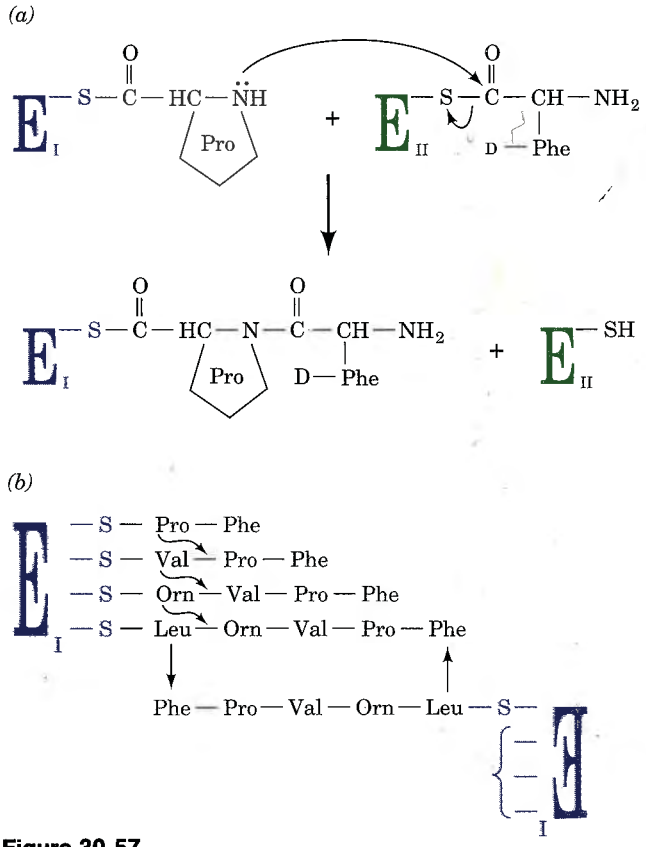


Figure 30-57
The biosynthesis of gramicidin S. (*a*) The initial transfer of Phe to an E_I-linked Pro residue to form a peptide bond. (*b*) The elongation (*top*) and cyclization (*bottom*) reactions on E_I (here the arrows indicate group transfer, not electron transfer). [After Lipmann, F., *Science* **173,** 878 (1971).]

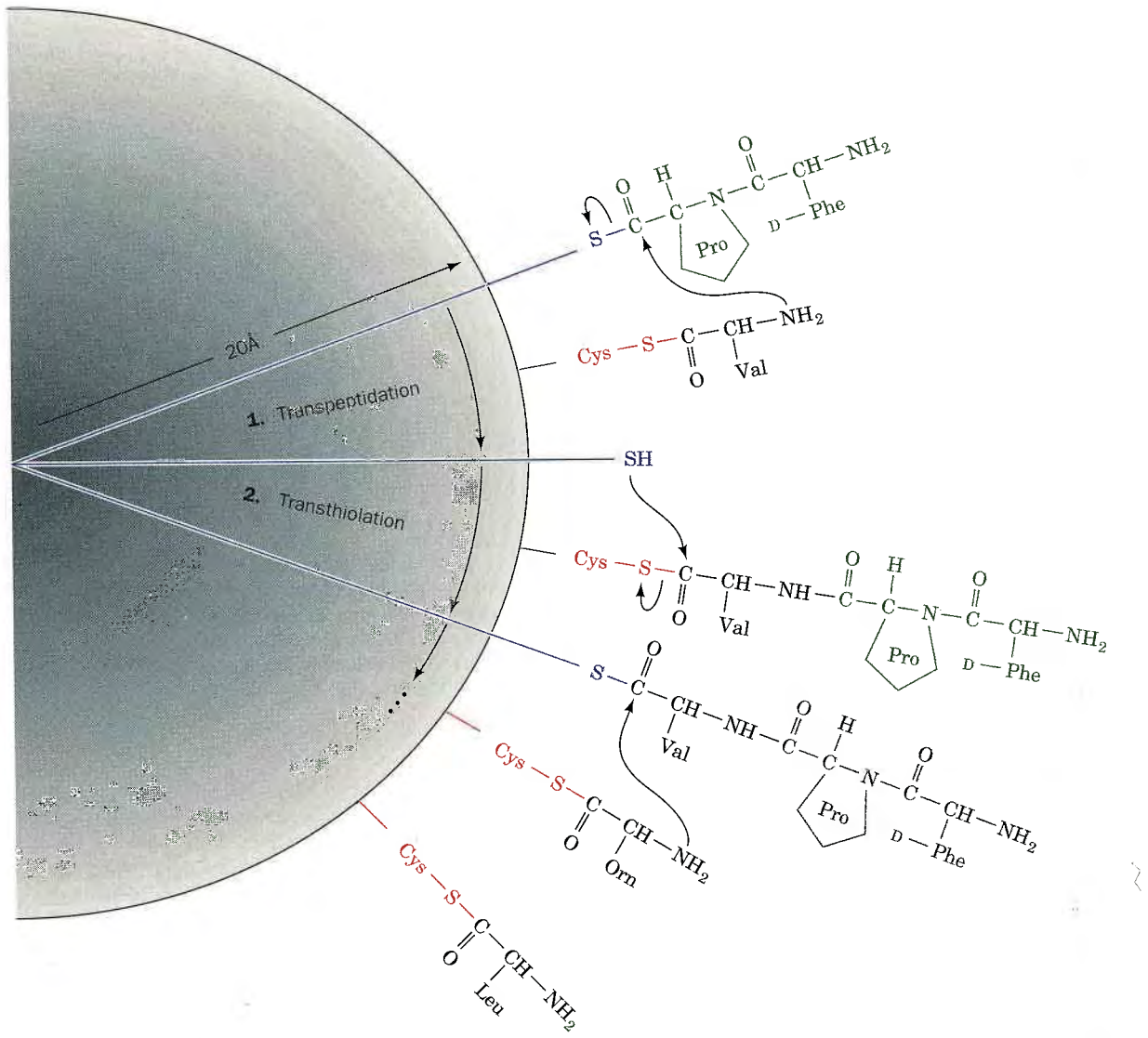
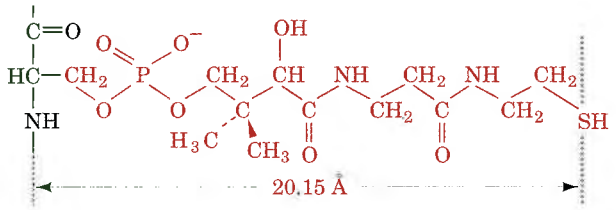


Figure 30-58
The proposed scheme for the participation of the pantetheine residue (*blue*) in the biosynthesis of gramicidin S. The circular arrows indicate the movement of the pantetheine in collecting the amino acid residues that are linked to the enzyme via Cys thioester bonds. The transpeptidation and transthiolation reactions alternate to synthesize the pentapeptide. [After Lipmann, F., *Acc. Chem. Res.* **6**, 366 (1973).]

that point. Note that chain elongation proceeds towards the C-terminus as it does in ribosomal polypeptide synthesis. The final enzyme-linked pentapeptide reacts in a head-to-tail fashion with a second such assembly to form the decapeptide product (Fig. 30-55).

The resemblance of the above reaction sequence to that of fatty acid synthesis led Lipmann to propose that phosphopantetheine is a cofactor in polypeptide synthesis as it is in acyl carrier protein (Section 23-4A). In fact, E_I contains a single Ser-linked phosphopantetheine.



This 20-Å long residue is thought to sequentially collect the enxyme-linked amino acids on the growing oligopeptide through alternating transpeptidation and transthiolation reactions (Fig. 30-58). Transthiolation in polypeptide synthesis is therefore analogous to translocation in ribosomal chain elongation (Section 30-3D).

Chapter Summary

Point mutations are caused by either base analogs that mispair during DNA replication or by substances that react with bases to form products that mispair. Insertion/deletion (frameshift) mutations arise from the association of DNA with intercalating agents that distort the DNA structure. The analysis of a series of frameshift mutations that supressed one another established that the genetic code is an unpunctuated triplet code. Fine structure genetic mapping combined with amino acid sequence analyses demonstrated that genes are colinear with the polypeptides they specify. In a cell-free protein synthesizing system, poly(U) directs the synthesis of poly(Phe) thereby demonstrating that UUU is the codon specifying Phe. The genetic code was established through the use of polynucleotides of known composition but random sequence, the ability of defined triplets to promote the ribosomal binding of tRNAs bearing specific amino acids, and through the use of synthetic mRNAs of known alternating sequences. The latter investigations also demonstrated that the 5′ end of mRNA corresponds to the N-terminus of the polypeptide it specifies and established the sequences of the stop codons. Degenerate codons differ mostly in the identities of their third base. Small single-stranded DNA phages such as ϕX174 contain overlapping genes in different reading frames. The genetic code used by mitochondria differs in several respects from the "standard" genetic code.

Transfer RNAs consist of 60 to 95 nucleotides that can be arranged in the cloverleaf secondary structure. As many as 10% of a tRNA's bases may be modified. Yeast $tRNA^{Phe}$ forms a narrow L-shaped three-dimensional structure that resembles that of other tRNAs. Most of the bases are involved in stacking and base pairing associations including nine tertiary interactions that appear to be essential for maintaining the molecule's native conformation. Amino acids are appended to their cognate tRNAs in a two-stage reaction catalyzed by the corresponding aminoacyl-tRNA synthetase. The great accuracy of tRNA charging arises from the proofreading of the bound amino acid by aminoacyl-tRNA synthetase at the expense of ATP hydrolysis. Ribosomes select tRNAs solely on the basis of their anticodons. Sets of degenerate codons are read by a single tRNA through wobble pairing. Nonsense mutations may be suppressed by tRNAs whose anticodons have mutated to recognize a Stop codon.

The ribosome consists of a small and a large subunit whose complex shapes have been revealed by electron microscopy. The three RNAs and 52 proteins comprising the *E. coli* ribosome self-assemble under proper conditions. The positions of many ribosomal components relative to the subunit surfaces have been largely determined by immune electron microscopy and neutron diffraction measurements. Affinity labeling experiments have identified the ribosomal components in the vicinity of various ribosomal binding sites and catalytic centers. Ribosomal polypeptide synthesis proceeds by the addition of amino acid residues to the C-terminal end of the nascent polypeptide. The mRNAs are read in the 5′ → 3′ direction. mRNAs are usually simultaneously translated by several ribosomes in the form of polysomes. The ribosome has at least three tRNA-binding sites: the P site, which binds the peptidyl-tRNA, the A site, which binds the incoming aminoacyl-tRNA, and the E site, which transiently binds the outgoing tRNA. During polypeptide synthesis, the nascent polypeptide is transferred to the aminoacyl-tRNA thereby lengthening the nascent polypeptide by one residue. The discharged tRNA is then released and the new peptidyl-tRNA, with its associated codon, is translocated to the P site. In prokaryotes, the initiation sites on mRNA are recognized through their Shine–Dalgarno sequences and by their initiating codon. Prokaryotic initiating codons specify fMet–$tRNA_f^{Met}$. Initiation is a complex process involving the participation of three initiation factors that induce the assembly of the ribosomal subunits with mRNA and fMet–$tRNA_f^{Met}$. The eukaryotic initiation site is usually the first AUG downstream from the 5′-terminal cap and this AUG specifies unformylated Met-$tRNA_i^{Met}$. Polypeptides are elongated in a three-part cycle, aminoacyl-tRNA binding, transpeptidation, and translocation, that requires the participation of elongation factors and is vectorially driven by GTP hydrolysis. Termination codons bind release factors that induce the peptidyl transferase to hydrolyze the peptidyl-tRNA bond. The high accuracy of translation indicates that the ribosome proofreads the codon–anticodon interaction, most probably via a kinetic mechanism. Ribosomal inhibitors, many of which are antibiotics, are medically important and biochemically useful in elucidating ribosomal function. Streptomycin causes mRNA misreading and inhibits prokaryotic chain initiation, chloramphenicol inhibits prokaryotic peptidyl transferase, tetracycline inhibits aminoacyl-tRNA binding to the prokaryotic small subunit, and diphtheria toxin ADP-ribosylates eEF-2.

Several mechanisms of translational control have been elucidated in eukaryotes. Hemoglobin synthesis in reticulocytes is inhibited in the absence of heme by heme-controlled inhibitor. This enzyme catalyzes the phosphorylation of eIF-2, which then tightly binds eIF-2B, thereby blocking translational initiation. In the presence of dsRNA, cells treated with interferon are translationally inhibited. This happens through two independent mechanisms: the induction of a protein kinase that phosphorylates eIF-2, and the induction of 2,5-A synthetase whose product 2,5-A, activates an endonuclease that degrades mRNA. Translation is also inhibited by mRNA masking, at least in certain embryos.

Proteins may be post-translationally modified in a variety of ways. Protelytic cleavages, usually by specific peptidases, activate proproteins. The signal peptides of preproteins are removed by signal peptidases. Covalent modifications alter many types of side chains in a variety of ways that modulate the catalytic activities of enzymes, provide recognition markers, and stabilize protein structures.

Proteins in living cells are continually turning over. This controls the level of regulatory enzymes and disposes of abnormal proteins that would otherwise interfere with cellular processes. Proteins are degraded by lysosomes in a nonspecific process that is stimulated during starvation, as well as by various pathological and normal states. A cytosolically based ATP-dependent system degrades normal as well as abnormal proteins in a process that flags these proteins by the covalent attachment of ubiquitin.

The biosynthesis of gramicidin S, which is representative of

the synthesis of many other polypeptide antibiotics, is mediated by soluble enzymes rather than by ribosomes. The proper amino acids are linked as thioesters to E_I and E_{II} in reactions driven by ATP hydrolysis. Chain elongation occurs by the pantetheine-mediated transfer of the growing oligopeptides to the activated amino acid residues. The reaction of two enzyme-linked pentapeptides yields the product cyclic decapeptide.

References

General

Lewin, B., *Genes* (3rd ed.), Chapters 5–7, Wiley (1987).

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M., *Molecular Biology of the Gene* (4th ed.), Chapters 14 and 15, Benjamin/Cummings (1987).

The Genetic Code

Attardi, G., Animal mitochondrial DNA: an extreme example of genetic economy, *Int. Rev. Cytol.* **93,** 93–145 (1985).

Benzer, S., The fine structure of the gene, *Sci. Am.* **206**(1): 70–84 (1962).

The Genetic Code, Cold Spring Harbor Symp. Quant. Biol. **31** (1966). [A collection of papers describing the establishment of the genetic code. See especially the articles by Crick, Nirenberg, and Khorana.]

Crick, F. H. C., Burnett, L., Brenner, S., and Watts-Tobin, R. J., General nature of the genetic code for proteins, *Nature* **192,** 1227–1232 (1961).

Crick, F. H. C., The genetic code, *Sci. Am.* **207**(4): 66–74 (1962). [The structure of the code as determined by phage genetics.]

Crick, F. H. C., The genetic code: III, *Sci. Am.* **215**(4): 55–62 (1966). [A description of the nature of the code after its elucidation was almost complete.]

Fox, T. D., Natural variation in the genetic code, *Annu. Rev. Genet.* **21,** 67–91 (1987).

Garen, A., Sense and nonsense in the genetic code, *Science* **160,** 149–159 (1968). [A discussion of the genetic aspects of nonsense mutations.]

Judson, J. F., *The Eighth Day of Creation,* Part II, Simon & Schuster (1979). [A fascinating historical narrative on the elucidation of the genetic code.]

Khorana, H. G., Nucleic acid synthesis in the study of the genetic code, *Nobel Lectures in Molecular Biology, 1933–1975, pp.* 303–331, Elsevier (1977).

Nirenberg, M. W. and Matthaei, J. H., The dependence of cell-free protein synthesis in *E. Coli* upon naturally occurring or synthetic polyribonucleotides, *Proc. Natl. Acad. Sci.* **47,** 1588–1602 (1961). [The landmark paper reporting the finding that poly(U) stimulates the synthesis of poly(Phe).]

Nirenberg, M. W., The genetic code: II, *Sci. Am.* **208:** 80–94 (1963). [Discusses the use of synthetic mRNAs to analyze the genetic code.]

Nirenberg, M. and Leder, P., RNA code words and protein synthesis, *Science* **145,** 1399–1407 (1964). [The determination of the genetic code by the ribosomal binding of tRNAs using specific trinucleotides.]

Nirenberg, M., The genetic code, *Nobel Lectures in Molecular Biology, 1933–1975, pp.* 335–360, Elsevier (1977).

Singer, B. and Kuśmierek, J. T., Chemical mutagenesis, *Annu. Rev. Biochem.* **51,** 655–693 (1982).

Yanofsky, C., Carlton, B. C., Guest, J. R., Helinski, D. R., and Henning, U., On the colinearity of gene structure and protein structure, *Proc. Natl. Acad. Sci.* **51,** 266–272 (1964).

Yanofsky, C., Gene structure and protein structure, *Sci. Am.* **216**(5): 80–94 (1967).

Transfer RNA

Altman, S. (Ed.), *Transfer RNA,* MIT Press (1978). [A series of general reviews on tRNAs.]

Björk, G. R., Ericson, J. U., Gustafsson, C. E. D., Hagervall, T. G., Jösson, Y. H., and Wikström, P. M., Transfer RNA modification, *Annu. Rev. Biochem.* **56,** 263–287 (1987).

Blow, D. M. and Brick, P., Aminoacyl-tRNA synthetases, *in* Jurnak, F. A. and McPherson, A. (Eds.), *Biological Macromolecules and Assembly,* Vol. 2: *Nucleic Acids and Interactive Proteins, pp.* 441–469, Wiley (1985).

Crick, F. H. C., Codon–anticodon pairing: the wobble hypothesis, *J. Mol. Biol.* **19,** 548–555 (1966).

Bennetzen, J. L. and Hall, B. D., Codon selection in yeast, *J. Biol. Chem.* **257,** 3026–3031 (1982).

Freist, W., Isoleucyl-tRNA synthetase: an enzyme with several catalytic cycles displaying variation in specificity and energy consumption, *Angew. Chem. Int. Ed. Engl.* **27,** 773–788 (1988).

Kersten, H., On the biological significance of modified nucleosides in tRNA, *Prog. Nucleic Acid Res. Mol. Biol.* **31,** 59–114 (1984).

Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. M. J., Seeman, N. C., and Rich, A., Three-dimensional tertiary structure of yeast phenylalanine transfer RNA, *Science* **185,** 435–440 (1974); *and* Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C., and Klug, A., Structure of yeast phenylalanine tRNA at 3 Å resolution, *Nature* **250,** 546–551 (1974). [The landmark papers describing the high resolution structure of a tRNA.]

Kline, L. K. and Söll, D., Nucleotide modifications in RNA, *in* Boyer, P. D. (Ed.), *The Enzymes* (3rd ed.), Vol. 15, *pp.* 567–582, Academic Press (1982).

Leatherbarrow, R. J., Fersht, A. R., and Winter, G., Transition-state stabilization in the mechanisim of tyrosyl-tRNA synthetase revealed by protein engineering, *Proc. Natl. Acad. Sci.* **82,** 7840–7844 (1985).

Leinfelder, W., Zehelein, E., Mandrand-Berthelot, M.-A., and Böck, A., Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine, *Nature* **331,** 723–725 (1988).

Muramatsu, T., Nishakawa, K., Nemoto, F., Kuchino, Y., Nishamura, S., Miyazawa, T., and Yokoyama, S., Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification, *Nature* **336,** 179–181 (1988).

Rich, A. and Kim, S. H., The three-dimensional structure of transfer RNA, *Sci. Am.* **238**(1): 52–62 (1978).

Schimmel, P., Parameters for the molecular recognition of tRNAs, *Biochemistry* **28,** 2747–2759 (1989).

Schimmel, P. R., Aminoacyl tRNA synthetases: general scheme of structure-function relationships in the polypeptides and recognition of transfer RNAs, *Annu. Rev. Biochem.* **56,** 125–158 (1987).

Schimmel, P. R., Söll, D., and Abelson, J. N. (Eds.), *Transfer RNA,* Cold Spring Harbor Laboratory (1979). [A two volume treatise on tRNA. Part 1 deals with structure, properties, and recognition; Part 2 treats biological aspects.]

Schulman, L. H. and Abelson, J., Recent excitement in understanding transfer RNA identity, *Science* **240,** 1591–1592 (1988).

Sprinzl, M., Moll, J., Meissner, F., and Hartmann, T., Compilation of tRNA sequences, *Nucleic Acids Res.* **13,** r1–r49 (1985).

Steege, D. A. and Söll, D. G., Suppression, *in* Goldberger, R. F. (Ed.), *Biological Regulation and Development,* Vol. 1., *pp.* 433–485, Plenum Press (1979).

von Ehrenstein, G., Weisblum, B., and Benzer, S., The function of sRNA as amino acid adapter in the synthesis of hemoglobin, *Proc. Natl. Acad. Sci.* **49,** 669–675 (1963).

Ribosomes

Bielka, H. (Ed.), *The Eukaryotic Ribosome,* Springer–Verlag (1982). [A detailed treatise.]

Brinacombe, R., The emerging three-dimensional structure and function of 16S ribosomal RNA, *Biochemistry* **27,** 4207–4214 (1988).

Capel, M. S., Engelman, D. M., Freeborn, B. R., Kjeldgaard, M., Langer, J. A., Ramakrishnan, V., Schindler, D. G., Schneider, D. K., Schoenborn, B. P., Siller, I.-Y., Yabuki, S., and Moore, P. B., A complete mapping of the proteins in the small ribosomal subunit of *Escherichia coli, Science* **238,** 1403–1406 (1987).

Caskey, C. Th., Peptide chain termination, *Trends Biochem. Sci.* **5,** 234–237 (1980).

Chambliss, G., Craven, G. R., Davies, J., Davis, K., Kahan, L., and Nomura, M. (Eds.), *Ribosomes. Structure, Function and Genetics,* University Park Press (1979). [An informative series of reviews.]

Clark, B., The elongation step of protein biosynthesis, *Trends Biochem. Sci.* **5,** 207–210 (1980).

Dahlberg, A. E., The functional role of ribosomal RNA in protein synthesis, *Cell* **57,** 525–529 (1989).

Dintzis, H. M., Assembly of the peptide chains of hemoglobin, *Proc. Natl. Acad. Sci.* **47,** 247–261 (1961). [The determination of the direction of polypeptide biosynthesis.]

Edelmann, P. and Gallant, J., Mistranslation in E. coli, *Cell* **10,** 131–137 (1977).

Fersht. A., *Enzyme Structure and Mechanism* (2nd ed.), Chapter 13, Freeman (1985). [A discussion of enzymatic specificity and editing mechanisms.]

Gale, E. F., Cundliffe, E., Reynolds, P. E., Richmond, M. H., and Waring, M. J., *The Molecular Basis of Antibiotic Action,* Chapter 6, Wiley (1981).

Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F., Comparitive anatomy of 16-S-like ribosomal RNA, *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 155–216 (1985).

Hardesty, B. and Kramer, G. (Eds.), *Structure, Function, and Genetics of Ribosomes,* Springer–Verlag (1985). [A series of useful reviews on current aspects of ribosomology.]

Held, W. A., Ballou, B., Mizushima, S., and Nomura, M., Assembly mapping of 30S ribosomal proteins from *Escherichia coli, J. Biol. Chem.* **249,** 3103–3111 (1974).

Hunt, T., The initiation of protein synthesis, *Trends Biochem. Sci.* **5,** 178–181 (1980).

Kurland, C. G. and Ehrenberg, M., Optimization of translation, *Prog. Nucleic Acid Res. Mol. Biol.* **31,** 191–219 (1984).

Lake, J. A., Evolving ribosome structure: domains in archaebacteria, eubacteria, eocytes and eukaryotes, *Annu. Rev. Biochem.* **54,** 507–530 (1985).

Lake, J. A., The ribosome, *Sci. Am.* **245**(2): 84–97 (1981).

Liljas, A., Structural studies of ribosomes, *Prog. Biophys. Mol. Biol.* **40,** 161–228 (1982).

Maitra, U., Stringer, E. A., and Chaudhuri, A., Initiation factors in protein biosynthesis, *Annu. Rev. Biochem.* **51,** 869–900 (1982).

Moldave, K., Eukaryotic protein synthesis, *Annu. Rev. Biochem.* **54,** 109–149 (1983).

Moore, P. B. and Capel, M. S., Structure-function correlations in the small ribosomal subunit from *Escherichia coli, Annu. Rev. Biophys. Biophys. Chem.* **17,** 349–367 (1988).

Moore, P. B., The ribosome returns, *Nature* **331,** 223–227 (1988).

Nagano, K. and Harel, M., Approaches to a three-dimensional model of the *E. coli* ribosome, *Prog. Biophys. Mol. Biol.* **48,** 67–101 (1986).

Nierhaus, K. H., Structure, function and assembly of ribosomes, *Curr. Top. Microbiol. Immunol.* **97,** 81–155 (1982). [A thorough and enlightening review.]

Nierhaus, K. H. and Wittman, H. G., Ribosomal function and its inhibition by antibiotics in prokaryotes, *Naturwissenschaften* **67,** 234–250 (1980).

Noller, H. F., Jr., and Moldave, K., (Eds.), Ribosomes, *Methods Enzymol.* **164** (1988). [Contains numerous articles on ribosome methodology.]

Noller, H. F., Structure of ribosomal RNA, *Annu. Rev. Biochem.* **53,** 119–162 (1984).

Pappenheimer, A. M., Jr., Diphtheria toxin, *Annu. Rev. Biochem.* **46,** 69–94 (1977).

Prince, J. B., Gutell, R. R., and Garrett, R. A., A consensus model of the *E. coli* ribosome, *Trends Biochem. Sci.* **8,** 359–363 (1983).

Rané, H. A., Klootwijk, J., and Musters, W., Evolutionary conservation of structure and function of high molecular weight ribosomal RNA, *Prog. Biophys. Mol. Biol.* **51,** 77–129 (1988).

Rhoads, R. E., Cap recognition and the entry of mRNA into the protein initiation cycle, *Trends Biochem. Sci.* **13,** 52–56 (1988).

Schüler, D. and Brinacombe, R., The *Escherichia coli* 30S ribosomal subunit; an optimized three-dimensional fit between the ribosomal proteins and the 16S RNA, *EMBO J.* **7,** 1509–1513 (1988).

Shatkin, A. J., mRNA cap binding proteins: essential factors for initiating translation, *Cell* **40,** 223–224 (1985).

Shine, J. and Dalgarno, L., The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites, *Prog. Natl. Acad. Sci.* **71,** 1342–1346 (1974).

Spirin, A. S., Ribosomal translocation: facts and models, *Prog. Nucleic. Res. Mol. Biol.* **32,** 75–114 (1985).

Steitz, J. A. and Jakes, K., How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S RNA and the mRNA during initiation of protein synthesis in *Escherichia coli, Proc. Natl. Acad. Sci.* **72,** 4734–4738 (1975).

Stern, S., Powers, T., Changchien, L.-I., and Noller, H. F., RNA–protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA, *Science* **244,** 783–790 (1989).

Stöffler, G. and Stöffler-Meilicke, M., Immunoelectron microscopy of ribosomes, *Annu. Rev. Biophys. Bioeng.* **13,** 303–330 (1984).

Thompson, R. C., EFTu provides an internal kinetic standard for translational accuracy, *Trends Biochem. Sci.* **13,** 91–93 (1988).

Control of Translation

Jagus, R., Anderson, W. F., and Safer, B., The regulation of initiation of mammalian protein synthesis, *Prog. Nucleic Acid Res. Mol. Biol.* **25,** 127–183 (1981).

Kaempfer, R., Regulation of eukaryotic translation, *in* Fraenkel-Conrat, H. and Wagner, R. R., (Eds.), *Comprehensive Virology,* Vol. 19, *pp.* 99–175, Plenum Press (1984).

Lengyel, P., Biochemistry of interferons and their actions, *Annu. Rev. Biochem.* **51,** 251–282 (1982).

Pestka, S., Langer, J. A., Zoon, K. C., and Samuel, C. E., Interferons and their actions, *Annu. Rev. Biochem.* **56,** 757–777 (1987).

Proud, C. G., Guanine nucleotides, protein phosphorylation and control of translation, *Trends Biochem. Sci.* **11,** 73–77 (1986).

Raff, R. A., Masked messenger RNA and the regulation of protein synthesis in eggs and embryos, *in* Prescott, D. M., and Goldstein, L. (Eds.), *Cell Biol.* Vol. 4, *pp.* 107–136, Academic Press (1980).

Williams, B. R. G. and Kerr, I. M., The 2–5A($pppA^{2'}p^{5'}A^{2'}p^{5'}A$) system in interferon-treated and control cells, *Trends Biochem. Sci.* **5,** 138–140 (1980).

Post-Translational Modification

Fessler, J. H. and Fessler, L. I., Biosynthesis of procollagen, *Annu. Rev. Biochem.* **47,** 129–162 (1978).

Prockop, D. J., Kivirikko, K. I., Tuderma, L., and Guzman, N. A., The biosynthesis of collagen and its disorders, *New Engl. J. Med.* **301,** 13–23 (1979).

Wold, F., In vivo chemical modification of proteins, *Annu. Rev. Biochem.* **50,** 783–814 (1981).

Wold, F. and Moldave, K. (Eds.), Posttranslational Modifications, Parts A and B, *Methods Enzymol.* **106** and **107** (1984). [Contains extensive descriptions of the amino acid "zoo".]

Protein Degradation

Bachmair, A., Finley, D., and Varshavsky, A., In vivo half-life of a protein is a function of its amino terminal residue, *Science* **234,** 179–186 (1986).

Chau, V., Tobias, J. W., Bachmair, A., Marriott, D., Ecker, D. J., Gonda, D. J., and Varshavsky, A., A multiubiquitin chain is confined to a specific lysine in a targeted short-lived protein. *Science* **243,** 1576–1583 (1989).

Finley, D. and Varshavsky, A., The ubiquitin system: functions and mechanisms, *Trends Biochem. Sci.* **10,** 343–347 (1985).

Goldberg, A. L. and Dice, J. F., *and* Goldberg, A. L. and St. John, A. C., Intracellular protein degradation in mammalian and bacterial cells: parts I *and* II, *Annu. Rev. Biochem.* **42,** 835–869 (1974) *and* **45,** 747–803 (1976).

Hershko, A., Ubiquitin-mediated protein degradation, *J. Biol. Chem.* **262,** 15237–15240 (1988).

Hershko, A. and Ciechanover, A., The ubiquitin pathway for the degradation of intracellular proteins, *Prog. Nucleic Acid Res. Mol. Biol.* **33,** 19–56 (1986).

Rechsteiner, M., *Ubiquitin,* Plenum Press (1988). [A series of authoritative articles on ubiquitin structure and function.]

Rechsteiner, M., Rogers, S., and Rote, K., Protein structure and intracellular stability, *Trends Biochem. Sci.* **12,** 390–394 (1987).

Senahdi, V.-J., Bugg, C. E., Wilkinson, K. D., and Cools, W. J., Three-dimensional structure of ubiquitin at 2.8 Å resolution, *Prog. Natl. Acad. Sci.* **82,** 3582–3585 (1985).

Waxman, L., Fagan, J. M., and Goldberg, A. L., Demonstration of two distinct high molecular weight proteases in rabbit reticulocytes, one of which degrades ubiquitin conjugates, *J. Biol. Chem.* **262,** 2451–2457 (1987).

Nonribosomal Polypeptide Synthesis

Kleinkauf, H. and Döhren, H., Nonribosomal polypeptide formation on multifunctional proteins, *Trends Biochem. Sci.* **8,** 281–283 (1983).

Lipmann, F., Attempts to map a process evolution of peptide biosynthesis, *Science* **173,** 875–884 (1971).

Lipmann, F., Nonribosomal peptide synthesis on polyenzyme templates, *Acc. Chem. Res.* **11,** 361–367 (1973).

Problems

1. What is the product of reacting guanine with nitrous acid? Is the reaction mutagenic? Explain.
2. What is the polypeptide specified by the following DNA antisense strand? Assume translation starts after the first initiation codon.

   5′-TCTGACTATTGAGCTCTCTGGCACATAGCA-3′

*3. The fingerprint of a protein from a phenotypically revertant mutant of bacteriophage T4 indicates the presence of an altered tryptic peptide with respect to the wild-type. The wild-type and mutant peptides have the following sequences:

Wild-type	Cys-Glu-Asp-His-Val-Pro-Gln-Tyr-Arg
Mutant	Cys-Glu-Thr-Met-Ser-His-Ser-Tyr-Arg

   Indicate how the mutant could have arisen and give the base sequences, as far as possible, of the mRNAs specifying the two peptides. Comment on the function of the peptide in the protein.

4. Explain why the various classes of mutations can reverse a mutation of the same class but not a different class.
5. Which amino acids are specified by codons that can be changed to an *amber* codon by a single point mutation?
6. The mRNA specifying the α chain of human hemoglobin contains the base sequence

   $\cdots$UCCAAAUACCGUUAAGCUGGA$\cdots$

   The C-terminal tetrapeptide of the normal α chain, which is specified by part of this sequence, is

   -Ser-Lys-Tyr-Arg

   In hemoglobin Constant Spring, the corresponding region of the α chain has the sequence

   -Ser-Lys-Tyr-Arg-Gln-Ala-Gly-$\cdots$

   Specify the mutation that causes hemoglobin Constant Spring.

7. Explain why a minimum of 32 tRNAs are required to translate the "standard" genetic code?
8. Draw the wobble pairings not in Fig. 30-21*a*.
9. A colleague of yours claims that by exposing *E. coli* to HNO_2 she has mutated a $tRNA^{Gly}$ to an *amber* suppressor. Do you believe this claim? Explain.

*10. Deduce the anticodon sequences of all suppressors listed in Table 30-5 except UGA-1 and indicate the mutations that caused them.

11. How many different types of macromolecules must be minimally contained in a cell-free protein synthesizing system from *E. coli* ? Count each type of ribosomal component as a different macromolecule.
12. Why do oligonucleotides containing Shine–Dalgarno sequences inhibit translation in prokaryotes? Why don't they do so in eukaryotes?
13. Why does m^7GTP inhibit translation in eukaryotes? Why doesn't it do so in prokaryotes?
14. What would be the distribution of radioactivity in the completed hemoglobin chains upon exposing reticulocytes to ^3H-labeled leucine for a short time followed by a chase with unlabeled leucine?
15. Design an mRNA with the necessary prokaryotic control sites that codes for the octapeptide Lys-Pro-Ala-Gly-Thr-Glu-Asn-Ser.
16. Indicate the translational control sites in and the amino acid sequence specified by the following prokaryotic mRNA.

   5′-CUGAUAAGGAUUUAAAUUAUGUGUCAAUCA-CGAAUGCUAAUCGAGGCUCCAUAAUAACACUU-CGAC-3′

17. What is the energetic cost, in ATPs, for the *E. coli* synthesis of a polypeptide chain of 100 residues starting from amino acids and mRNA? Assume that no losses are incurred as a result of proofreading.

*18. It has been suggested that Gly-tRNA synthetase does not require an editing mechanism. Why?

19. An antibiotic named fixmycin, which was isolated from a fungus growing on ripe passion fruit, is effective in curing many types of veneral disease. In characterizing fixmycin's mode of action, you have found that it is a bacterial translational inhibitor that binds exclusively to the large subunit of *E. coli* ribosomes. The initiation of protein synthesis in the presence of fixmycin results in the generation of dipeptides that remain associated with the ribosome. Suggest a mechanism of fixmycin action.
20. Heme inhibits protein degradation in reticulocytes by allosterically regulating ubiquitin-activating enzyme. What physiological function might this serve?
21. Genbux Inc., a genetic engineering firm, has cloned the gene encoding an industrially valuable enzyme into *E. coli* such that the enzyme is produced in large quantities. However, since the firm wishes to produce the enzyme in ton quantities, the expense of isolating it would be greatly reduced if the bacterium could be made to secrete it. As a high priced consultant, what general advice would you offer to solve this problem?