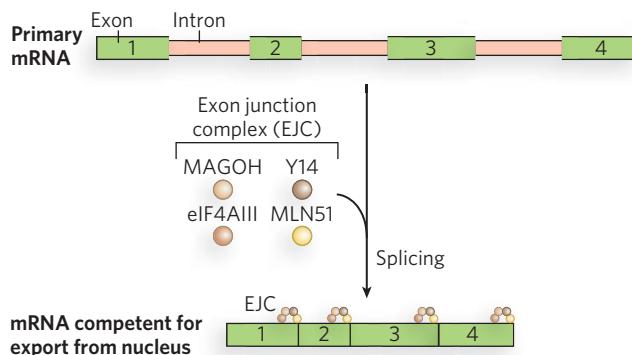


lacking the intron. Furthermore, the spliced mRNA was found to assemble with a different set of proteins than the mRNA that never contained the intron. This led to the conclusion that splicing generates a specific mRNA-protein complex that targets the mRNA for nuclear export, explaining the broader observation that an intron is required for the efficient expression of many eukaryotic protein-coding genes. In this way, only those mRNAs that have the correct end structures and spliced-exon sequence are used for protein synthesis.

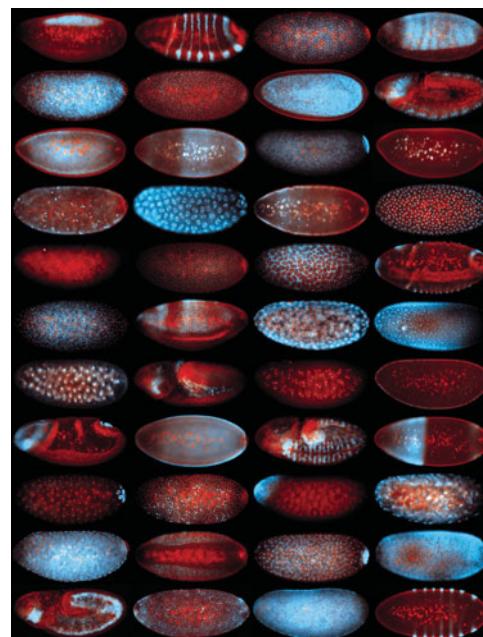
Mature mRNAs produced by splicing end up in different intracellular locations and are differently translated and degraded than are otherwise identical mRNAs produced from non-intron-containing genes. The explanation is that splicing influences the set of proteins that associate with the mRNA in the nucleus to form an mRNP (mRNA ribonucleoprotein particle). These proteins in turn ensure that the mRNA interacts with exportins for shipment out of the nucleus. Chemical cross-linking experiments with cell extracts translating mRNAs with or without introns showed that several proteins bind to exon-exon junctions only as a consequence of splicing. Spliceosomes deposit a complex of proteins called the **exon junction complex (EJC)** on mRNAs at a position 20 to 24 nucleotides upstream of exon-exon junctions (Figure 16-23). At its core, EJC contains four proteins: eIF4AIII, MAGOH, Y14, and MLN51. The bound complexes accompany spliced mRNA into the cytoplasm, where they are removed during the first (“pioneer”) round of translation.



**FIGURE 16-23** The exon junction complex. The EJC is a complex of four proteins responsible for mRNA quality control. An EJC is deposited on the mRNA after splicing, just upstream of each exon-exon junction. The complexes accompany the mature mRNA out of the nucleus and into the cytoplasm.

## Some mRNAs Are Localized to Specific Regions of the Cytoplasm

In specialized cells, including oocytes (egg cells) and neurons, mRNAs are localized to particular sites prior to translation. The mechanism of such mRNA localization is best characterized in the fruit fly *Drosophila melanogaster*, in which maternal mRNAs are trafficked to various parts of the egg to help establish polarity during the early stages of embryo development (Figure 16-24). Shortly after the egg begins to mature, but before fertilization, mRNAs encoding proteins called Oskar and Bicoid bind to proteins that can move along microtubules, filamentous protein polymers that contribute to cell shape and structure. The *oskar* and *bicoid* mRNAs are shuttled to the parts of the egg where their protein products are required to form structures in the developing embryo. Analogous mechanisms of mRNA localization are thought to occur in neurons, in which mRNAs must be moved to parts of the cell very far from the nucleus for the localized protein synthesis required for proper neural function.



**FIGURE 16-24** Transport of mRNA in the *Drosophila* egg.

The building blocks of anterior-posterior axis patterning in *Drosophila* are laid out during egg formation (oogenesis), well before the egg is fertilized and deposited. The developing egg (oocyte) is polarized by differentially localized mRNA molecules. In each photograph, a different mRNA is labeled with a blue fluorescent marker. Nuclear DNA is labeled in red. Each mRNA contributes to pattern formation in the developing embryo. [Source: E. Lecuyer et al., *Cell* 131 : 174-187, 2007. Courtesy of Eric Lecuyer.]

An obvious advantage of regulating gene expression by mRNA localization is that it allows protein production to be spatially restricted within the cytoplasm. In this way, production can be turned on (and off) as required, without waiting for transcription, mRNA export, translation, and subsequent targeting of the protein to the site where it is needed. In addition, localized mRNAs can be translated multiple times to generate many copies of a protein at the required site. Local translation can also protect the rest of the cell from proteins that would be toxic or interfere with functions in other cellular compartments.

### Cellular mRNAs Are Degraded at Different Rates

As for all molecules in the cell, the amount of an mRNA transcript is determined by its relative rates of synthesis and degradation, or decay. Cellular mRNAs are degraded as part of their normal life cycle. **RNA degradation**, catalyzed by ribonucleases, is the complete hydrolysis of RNA molecules into their component nucleotides. When mRNA synthesis and decay rates are balanced, the concentration of the mRNA remains in a steady state. A change in either rate will lead to a net accumulation or depletion of the mRNA, affecting the rate of protein synthesis. Degradative pathways ensure that mRNAs do not build up in the cell and direct the synthesis of unnecessary proteins.

In eukaryotic cells, degradation rates can vary greatly for mRNAs produced from different genes, from a half-life of minutes or even seconds for a gene product that is needed only briefly, to many cell generations for a gene product in constant demand. Bacterial cells grow much faster than eukaryotic cells and must rapidly adapt to changing environmental and metabolic conditions. Their mRNAs are stable for only a few minutes. Degradation rates of an RNA are affected by its primary and secondary structure. For instance, a hairpin structure in bacterial and eukaryotic mRNAs can confer stability. In eukaryotes, sequences rich in A and U residues, known as AU-rich elements (AREs), occur in the 3' untranslated regions (3'UTRs) of some mRNAs. These elements recruit factors such as nucleases or RNA-binding proteins that can enhance or reduce, respectively, the degradation rate of the mRNA.

For mRNAs, degradation in *E. coli* begins with one or a few cuts by an endoribonuclease, followed by 3'→5' degradation by an exoribonuclease. In eukaryotes, the poly(A) tail is shortened, then the 5' cap is removed, before the mRNA can be degraded by ribonucleases. Eukaryotes have a complex of up to 10 conserved 3'→5'



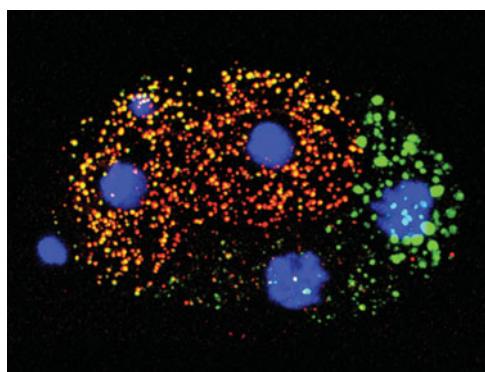
**FIGURE 16-25** The human exosome. The exosome has a ringlike structure. The complex encircles the mRNA and slides along in a 3'→5' direction, degrading the RNA as it moves. [Source: (a) PDB ID 2NN6.]

exoribonucleases, called the **exosome**, which, besides degrading mRNAs, is involved in the processing of the 3' end of rRNAs and tRNAs (Figure 16-25). The exosome is the major path of mRNA degradation in higher eukaryotes, but lower eukaryotes use primarily 5'→3' exonucleases.

Decay rates and half-lives are determined by proteins that enhance or inhibit exosome binding through interactions with the 3'UTR of the mRNA. In addition, two mRNA surveillance pathways are in place to respond to mRNAs that contain a premature stop codon or lack a stop codon. **Nonsense-mediated decay (NMD)** is triggered by exon junction complexes, which, as we've seen, are deposited during pre-mRNA splicing in the nucleus. Normally, EJCs are removed by the ribosome during the first round of translation, but if an EJC is located downstream of a stop codon, its presence indicates that splicing occurred at this position and that there should be additional coding sequence after the stop codon. The EJC thus signals that this is an mRNA with a premature stop codon, or nonsense codon, and triggers mRNA degradation. In contrast, the **non-stop decay** pathway targets mRNAs lacking a stop codon. Ribosomes traversing these mRNAs are released from the 3' end of the message, and the mRNA is shunted to the exosome for degradation.

### Processing Bodies Are the Sites of mRNA Storage and Degradation in Eukaryotic Cells

In eukaryotes, mRNAs that are not engaged in translation are sequestered in localized areas of the cytoplasm called **processing bodies (P bodies)**, which can be observed by light microscopy. P bodies contain proteins that catalyze removal of the mRNA 5' cap and thus are thought to be sites of mRNA degradation. They also appear to be sites where mRNAs are temporarily stored when not being translated, and may therefore play an active role in regulating which proteins are made in response to the cell's needs (Figure 16-26).



**FIGURE 16-26 Processing (P) bodies.** P bodies are sites of mRNA degradation in somatic cells, but in this *C. elegans* egg cell, the P bodies (green) protect mRNAs for later use in embryonic development. [Source: Scott L. Noble et al., *J. Cell Biol.* 182:559–572, 2008. © Rockefeller University Press.]

Recent experimental evidence supports an emerging model of cytoplasmic mRNA function in which translation and degradation rates are influenced by the relative concentrations of mRNA in polyribosomes (groups of ribosomes translating an mRNA; see Chapter 18) and in P bodies. In some cases, mRNA-specific binding factors suppress translation and promote degradation by recruiting P-body proteins to individual mRNAs.

- Bacterial mRNA degradation begins with one or a few cuts by an endoribonuclease, followed by 3'→5' degradation by an exoribonuclease.
- In eukaryotes, a complex of up to 10 conserved 3'→5' exoribonucleases—the exosome—helps degrade mRNAs and processes the 3' end of rRNAs and tRNAs.
- Processing bodies are cytoplasmic locations of mRNA storage and degradation.

## 16.5 Processing of Non-Protein-Coding RNAs

We have seen how the cell produces mature mRNA transcripts through a series of processing mechanisms. Each aspect of mRNA processing, transport, and decay is carefully regulated to control how, when, and where the mRNA is made or translated. RNA processing is not unique to mRNAs. All functional RNAs in cells are produced as precursor transcripts that must be cleaved to form the mature, functional RNA. This processing could be required because RNA polymerases don't always have specific termination sites, so transcripts from the same gene can have heterogeneous ends. Processing also provides convenient opportunities for the cell to regulate RNA levels. We discuss here the processing of three different kinds of functional RNAs central to gene expression: tRNAs, rRNAs, and microRNAs.

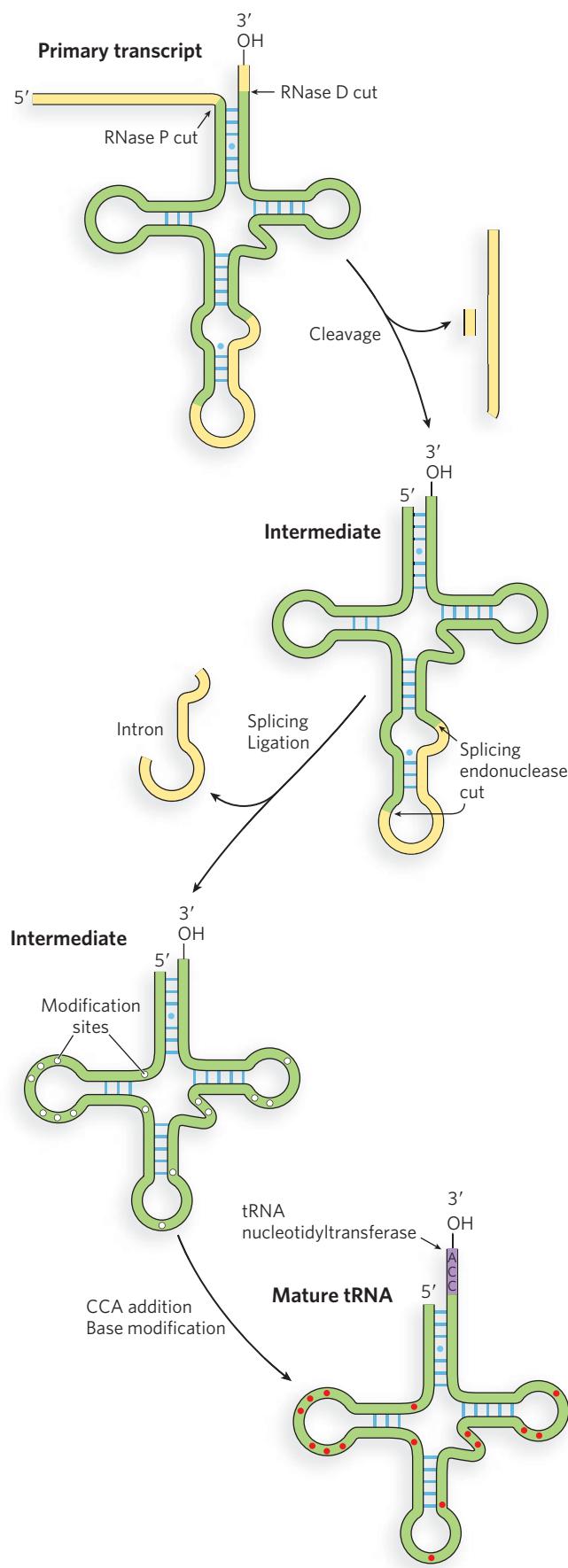
### SECTION 16.4 SUMMARY

- Non-protein-coding RNAs are exported from the nucleus by exportin proteins in a Ran-dependent pathway that requires GTP.
- After splicing, eukaryotic mRNAs are recognized by their processing modifications and then exported to the cytoplasm through a nuclear pore.
- Spliced mRNA exits the nucleus in a Ran-independent pathway involving export factors and other proteins in a large complex called TREX, which coordinates the machineries responsible for transcription, splicing, and export of mRNA.
- Spliceosomes deposit a complex of proteins—the exon junction complex (EJC)—on spliced mRNAs upstream from exon-exon junctions, providing a physical tag indicating that splicing has occurred. The EJC promotes nuclear export and mRNA stability.
- Cellular mRNAs are degraded at different rates, depending on their interactions with ribonucleases. Nonsense-mediated decay and non-stop decay are two pathways that guard against the translation of defective mRNAs.

### Maturation of tRNAs Involves Site-Specific Cleavage and Chemical Modification

Transfer RNAs, the molecules that carry amino acids to the ribosome during protein synthesis (see Chapter 18), are coordinately expressed in response to metabolic needs. In some cases, multiple tRNAs are synthesized as a single primary transcript and are separated by enzymatic cleavage. Even tRNAs synthesized alone are derived from longer RNA precursors by the enzymatic removal of nucleotides from the 5' and 3' ends (Figure 16-27). The endonuclease RNase P, a ribonucleoprotein found in all organisms, removes nucleotides from the 5' end of tRNAs. The RNA component of the enzyme is essential for activity. Indeed, bacterial RNase P can carry out its processing function with precision even in the absence of the protein component. The 3' end of tRNAs is processed by one or more nucleases, including the exonuclease RNase D.

In eukaryotes, a few tRNA transcripts have introns that must be excised. These introns are spliced by an ATP-dependent mechanism distinct from that of the spliceosome: a splicing endonuclease that recognizes



and cleaves the phosphodiester bonds at the intron splice sites (see Figure 16-27). RNA ligase joins the two exons to complete the reaction.

Transfer RNA precursors often undergo further posttranscriptional processing. The terminal CCA-3' to which an amino acid is attached is absent from some bacterial and all eukaryotic tRNA precursors, and is added after the initial transcript is made. This addition is carried out by tRNA nucleotidyltransferase (see Figure 16-27), an unusual enzyme that binds the three ribonucleoside triphosphate precursors in separate active sites and catalyzes the formation of the phosphodiester bonds to produce the CCA-3' sequence. The creation of this defined sequence of nucleotides is therefore not dependent on a DNA or RNA template—the template consists of the three binding sites of the nucleotidyltransferase.

The final type of tRNA processing is the modification of some bases by methylation, deamination, or reduction. In some cases, a base is removed from a specific site in the tRNA sequence and replaced by a different, noncanonical base (Figure 16-28). For example, in some tRNAs, uracil is removed at a particular U residue and reattached to the ribose through C-5 to create pseudouridine ( $\psi$ ). Modified bases often occur at characteristic positions in all tRNAs, suggesting their importance for tRNA structural stability or recognition by other enzymes (see How We Know).

### Maturation of rRNA Involves Site-Specific Cleavage and Chemical Modification

Ribosomal RNAs of bacterial and eukaryotic cells are produced from longer precursors called **preribosomal RNAs (pre-rRNAs)**. The pre-rRNA transcripts, produced by RNA polymerase I (Pol I) in eukaryotes, are coordinately synthesized so that they are present in similar amounts for ribosome assembly. In bacteria, the three rRNAs needed to form a functional ribosome—16S, 23S, and 5S—arise from a single 30S RNA precursor of about 6,500 nucleotides. RNA at both ends of the 30S precursor and segments between the rRNAs are removed during processing (Figure 16-29).

**FIGURE 16-27** The processing of tRNA. Transfer RNAs are trimmed at the 5' and 3' ends by RNase P and RNase D, respectively. Some eukaryotic tRNAs have introns that are excised during processing. The CCA trinucleotide is added to the 3' end, and some bases are modified to provide stability and enhance the functioning of the mature tRNA.

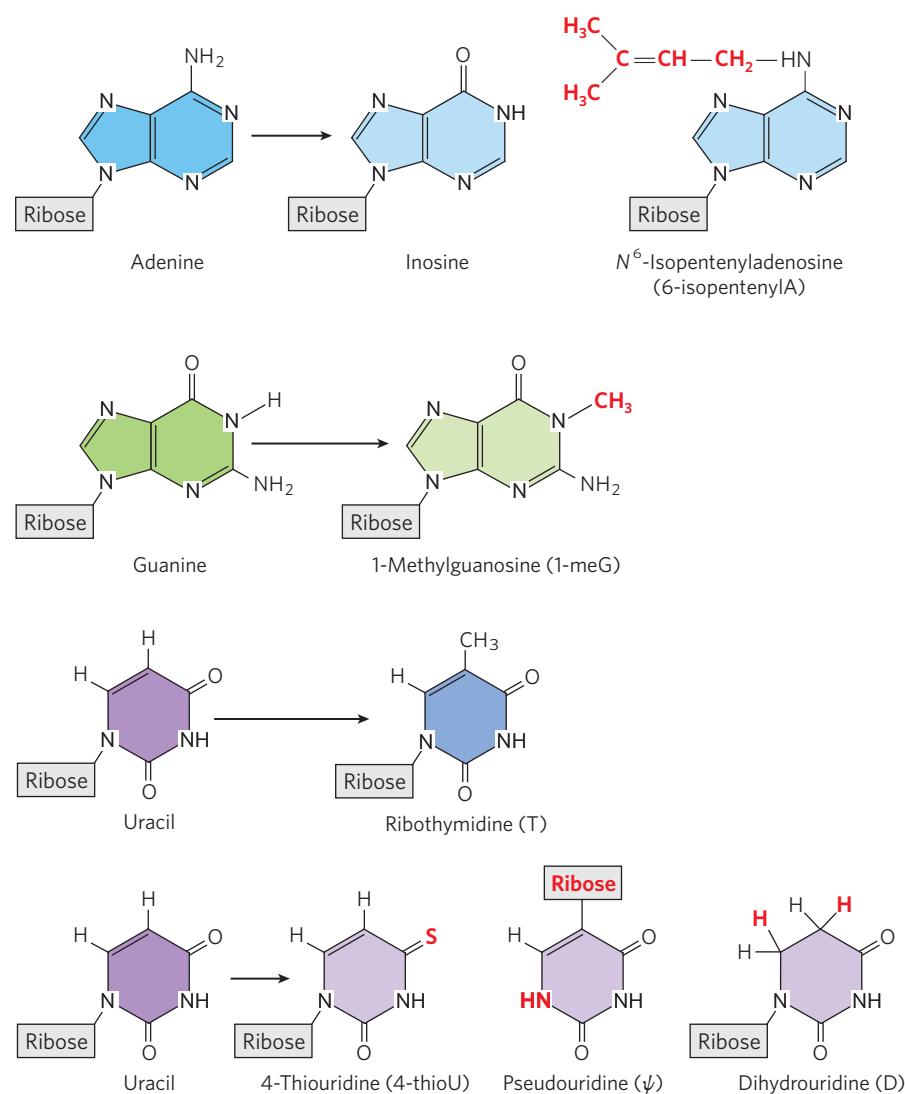
### KEY CONVENTION

For historical reasons stemming from their initial discovery as large, polymeric components of ribosomes, rRNAs are named according to their sedimentation behavior during ultracentrifugation. Thus, 16S refers to 16 svedberg units (S), named for the Swedish physicist Theodor Svedberg (1884–1971) who invented ultracentrifugation. A svedberg describes the sedimentation properties of particles during centrifugation and is defined as exactly  $10^{-13}$  seconds.

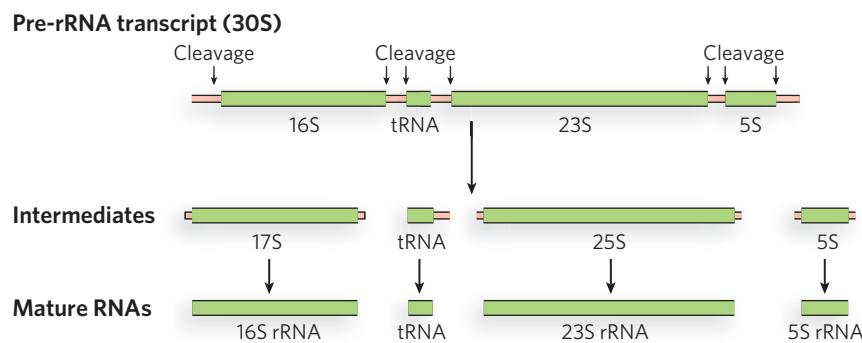
The *E. coli* genome encodes seven pre-rRNA molecules, each with essentially identical rRNA-coding

regions but differing in the segments between these regions. One or two tRNA genes are found between the 16S and 23S rRNA genes, with different tRNAs arising from different pre-rRNA transcripts. Coding sequences for tRNAs are also found on the 3' side of the 5S rRNA in some precursor transcripts.

In eukaryotes, a 45S pre-rRNA transcript is processed in the nucleolus to form the 18S, 28S, and 5.8S rRNAs characteristic of eukaryotic ribosomes. In an interesting quirk of evolution, the 5S rRNA of most eukaryotes is made as a completely separate transcript by a different polymerase (Pol III, rather than Pol I). The enzymes responsible for rRNA processing localize to the nucleolus and are thought to begin cleaving the



**FIGURE 16-28** Modified tRNA bases produced in posttranscriptional reactions. Specific nucleotide residues in tRNAs, and in other RNAs including rRNAs, are chemically modified by enzymes that recognize particular structures and/or sequences.



**FIGURE 16-29** Processing of pre-rRNA transcripts in bacteria. The 3OS pre-rRNA contains the 16S, 23S, and 5S rRNA sequences, as well as a tRNA sequence. During processing, these segments are separated by nuclease activities.

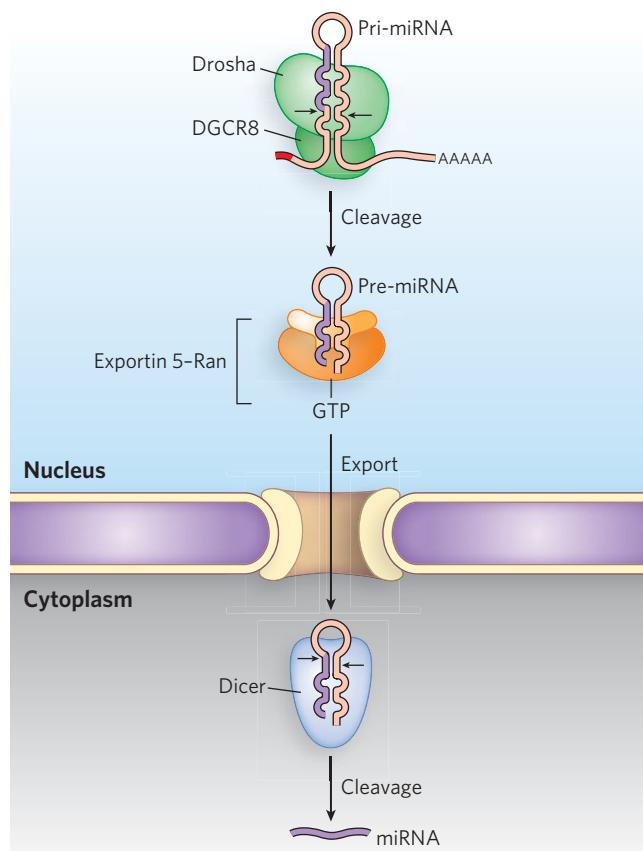
pre-rRNA transcript during transcription. Processing occurs in an ordered set of reactions guided in part by RNAs in **small nucleolar ribonucleoproteins (snoRNPs)**, which base-pair with the pre-rRNA transcript at specific cleavage sites. The binding of snoRNPs also specifies sites where methyl groups are added to 2'-hydroxyl groups in rRNAs. Careful experiments comparing the function of rRNAs with and without particular sites methylated have failed to reveal a dramatic difference in behavior. However, bacteria or yeast cells with all of the naturally occurring sites in rRNA methylated outcompete cells lacking some of these sites.

### Small Regulatory RNAs Are Derived from Larger Precursor Transcripts

Most eukaryotic cells use RNAs called **microRNAs (miRNAs)** (21 to 23 nucleotides) to regulate the expression of many genes. In addition, many eukaryotes also use short interfering RNAs (siRNAs) to prevent protein synthesis from specific mRNAs. These mechanisms of regulation by miRNAs and siRNAs, known as RNA interference, are discussed in detail in Chapter 22.

The miRNAs are encoded by genes that are transcribed but not translated into protein, analogous to tRNA or rRNA. Primary miRNA transcripts (pri-miRNAs) are typically synthesized by Pol II and then capped and polyadenylated. These transcripts have the ability to fold into extended hairpinlike structures with extended single-stranded RNA on the 5' and 3' sides of the hairpin. Such structures are recognized by an RNA-binding protein, which in humans is encoded by *DGCR8* (the *DiGeorge syndrome critical region gene 8*). This binding protein recruits nuclear pri-miRNAs to the enzyme **Drosha** to form a **microprocessor complex** (Figure 16-30). Drosha is an endonuclease that cleaves the pri-miRNA hairpin duplex to produce precursor

miRNAs (pre-miRNAs), ~70 nucleotides in length and containing characteristic dinucleotide 3' overhangs. Pre-miRNAs then assemble with RNA export proteins,



**FIGURE 16-30** Processing of small regulatory RNAs. The small miRNAs and siRNAs are formed by the processing of much larger precursor transcripts by specific endonucleases. [Source: Adapted from V. N. Kim, J. Han, and M. C. Siomi, *Nat. Rev. Mol. Cell Biol.* 10:126–139, 2009, Figs. 2–4.]

including exportin-5, for transport out of the nucleus. Most pre-miRNAs are not perfectly base-paired hairpins, but instead contain one or more unpaired bases, or non-Watson-Crick base pairs, in the stem of the hairpin. This may be because perfectly double-stranded RNAs would activate the cell's antiviral machinery through a pathway known as the interferon response.

Once in the cytoplasm, pre-miRNAs are further cleaved by the double-stranded RNA endonuclease called **Dicer** to release the mature miRNA. Dicer recognizes the dinucleotide overhangs of pre-miRNAs, which help position RNA substrates such that Dicer's cleavage products have a characteristic length of 21 to 23 bp. Dicer can also cleave long double-stranded RNAs generated within the cell or introduced by viral infection or transfection, to produce siRNAs. Dicer assists in loading miRNAs and siRNAs into complexes called RNA-induced silencing complexes (RISCs). Mature miRNAs and siRNAs can base-pair with complementary sequences in the 3'UTRs of specific mRNAs to induce degradation and/or translational silencing of those mRNAs.

### SECTION 16.5 SUMMARY

- Transfer RNAs are produced from precursor transcripts that are enzymatically cleaved at the 5' and 3' ends. In some cases, internal sequences are excised and the flanking sequences ligated to produce the mature tRNA.
- Transfer RNAs contain unusual bases or nucleotide modifications that are formed posttranscriptionally.
- Ribosomal RNAs are produced from much longer precursor transcripts by site-specific cleavage, and particular nucleotides are chemically modified.
- Regulatory RNAs, such as miRNAs, are also processed from larger transcripts.

## 16.6 RNA Catalysis and the RNA World Hypothesis

The study of posttranscriptional processing of RNA molecules led to one of the most exciting discoveries in modern molecular biology: the existence of **ribozymes**, enzymes consisting of RNA. RNA is the only known molecule that is capable of both encoding and actively influencing the expression of genetic information. For this reason, many scientists have proposed that RNA could have formed the basis for the early life forms that

evolved into modern organisms. We discuss here the properties of RNA that support its multiple functions, and we explore the implications of the RNA world hypothesis.

### Ribozyme Diversity Correlates with Function

Self-splicing introns, bacterial RNase P, and several additional classes of naturally occurring ribozymes have been discovered. Experimentally, it has also proved possible to select catalytically active RNA molecules from pools of randomized sequences that are prepared in the laboratory. Ribozymes selected in this way can have a variety of enzymatic functions, demonstrating the inherent catalytic capabilities of RNA. The activities of many ribozymes, natural and selected in the laboratory, consist of breaking or joining phosphodiester bonds in RNA substrates. In some cases, such as self-splicing introns, the breaking and joining are coupled, whereas other ribozymes, such as RNase P, catalyze bond cleavage only.

Because many ribozymes act on an RNA substrate, often a part of the ribozyme itself, base-pairing interactions are critical for binding and positioning the substrate for reaction. Crystallographic structures of ribozymes and their components, determined over the past two decades, show that base pairing is also central to the architecture of ribozyme active sites. The first crystal structure of a large folded RNA, the P4-P6 domain of the *Tetrahymena* group I self-splicing intron (see Figure 16-14), revealed how base-paired RNA helices can pack together through non-Watson-Crick interactions and specifically positioned Mg<sup>2+</sup> ions.

The structural patterns observed in the P4-P6 RNA, including the extensive use of unpaired A residues in the stabilizing, noncovalent contacts essential to the three-dimensional structure, have been observed repeatedly in RNAs ranging from small, self-cleaving ribozymes to ribosomes. Thus, RNA molecules use weak interactions, including the hydrophobic interactions and hydrogen bonding inherent in base pairing and base stacking, along with site-specific metal ion coordination, to form a wide variety of structures with ligand-binding sites and catalytic centers. DNA molecules are inherently less able to form such stable three-dimensional structures, due to subtle but crucial differences in the geometry of the phosphodiester backbone and the lack of a 2' hydroxyl in DNA nucleotides. RNA is therefore uniquely positioned in biology not only to encode genetic information but also to modify it, and possibly even to replicate it.

## HIGHLIGHT 16-3 **EVOLUTION**

### A Viral Ribozyme Derived from the Human Genome?

Ribozymes are thought to have been important in early evolution, because they provide, within the same molecule, the potential to both store and replicate genetic information. However, the age and origin of the known naturally occurring ribozymes are not easy to determine.

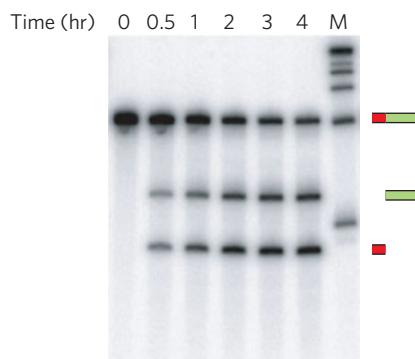
Recent experiments by Jack Szostak and colleagues at Harvard Medical School showed that at least in one case, ribozyme evolution occurred relatively recently. The research group was looking for self-cleaving ribozymes in the human genome. Total human cellular RNA was isolated and tested for the ability to generate shorter fragments in a  $Mg^{2+}$ -dependent reaction. Cleaved fragments were selectively enriched and retested, and after many such



**Jack Szostak** [Source: Courtesy of Jussi Puikkonen.]

cycles, the researchers isolated RNAs that could catalyze autocleavage at specific sites. In an experiment with one such RNA, the starting RNA fragment was found to cleave itself over the course of several hours, resulting in shorter fragments that could be separated by denaturing polyacrylamide gel electrophoresis (**Figure 1**).

Remarkably, the ribozyme discovered in this experiment is structurally and biochemically related to ribozymes first discovered in the human hepatitis delta virus (HDV). Furthermore, sequence comparisons showed that the ribozyme occurs only in mammals, implying that it may have evolved as recently as 200 million years ago. HDV itself may have arisen sometime later, forming from fragments of human RNA.



**FIGURE 1** Gel revealing a viral-like ribozyme in human genomic RNA. Samples of the reaction mixture containing the transcript with the ribozyme sequence and flanking sequences were removed for analysis at different time intervals. Over time, the transcript was processed into shorter fragments corresponding to ribozyme-catalyzed strand cleavage at the junction between the 5' flanking sequence and the boundary of the ribozyme sequence. [Source: K. Salehi-Ashtiani et al., *Science* 313:1788–1792, 2006, doi: 10.1126/science.1129308.]

The known repertoire of ribozymes continues to expand. Some virusoids, small RNAs associated with plant RNA viruses, include a structure that promotes a self-cleavage reaction. (Highlight 16-3 describes the self-cleaving ribozyme of a human virus.) The hammerhead ribozyme is in this class, catalyzing the hydrolysis of an internal phosphodiester bond important for producing unit-length virusoid RNAs. In the eukaryotic spliceosome, the splicing reaction requires a catalytic center formed at least in part by the U2, U5, and U6 snRNAs. And perhaps most importantly, an RNA component of ribosomes catalyzes the synthesis of proteins (see Chapter 18).

Ribozymes vary greatly in size. Ribosomal RNA is thousands of nucleotides long, the *Tetrahymena* self-

splicing group I intron contains ~400 nucleotides, and the smallest active hammerhead ribozyme of virusoids consists of two RNA strands with only 41 nucleotides in all. Experiments have shown that in each case, a ribozyme can be inactivated by heating above its melting temperature or by the addition of denaturing chemicals or complementary oligonucleotides, which disrupt normal base-pairing patterns. Furthermore, ribozymes can be inactivated if essential nucleotides are changed, forming the basis for many experiments demonstrating the importance of particular nucleotides or base-pairing interactions in ribozyme function (see How We Know).

## Could RNA Have Formed the Basis for Early Life on Earth?

The existence of so many kinds of ribozymes has fueled debate about why these catalysts occur in nature, and what they might indicate about the origin of enzymes. Without catalysts, life would not be possible, and an understanding of how enzymes evolved is central to our understanding of life's origins.

The RNA world hypothesis proposes that organisms comprised entirely or mostly of RNA evolved on the early Earth. The base-pairing properties of RNA could have been used to store information, as occurs today in many viruses, and the capacity of RNA to form a variety of structures lent itself to the emergence of ribozymes and perhaps other functional molecules, such as those forming membrane-spanning pores. Using *in vitro* selection methods, Gerald Joyce and his colleagues found that even RNAs containing just two kinds of nucleotides that can base-pair with each other can become catalytically active. This observation indicates that simple nucleic acids that might have been around before life began could have given rise to activities necessary for template-dependent replication.

Eventually, DNA supplanted RNA as a data storage molecule, because of its greater chemical stability. Proteins, with broader catalytic capabilities stemming from chemically diverse amino acids, became the specialized catalytic molecules. The few remaining ribozymes are remnants of the RNA world and provide clues to its former existence.

Several aspects of RNA chemistry and behavior support the RNA world hypothesis. These include the findings that RNA can function as both a genome and an enzyme, that RNA catalyzes peptide bond formation on ribosomes and thus is responsible for protein synthesis, and that components of RNA form spontaneously in “prebiotic soup” experiments designed to replicate conditions on the early Earth. Furthermore, the continuing discovery of RNA molecules that function in fundamental aspects of gene expression and regulation underscores the pervasive and presumably ancient roles of RNA in virtually all aspects of life. More difficult to explain is how and why the specific sugars found in RNA and DNA were selected under prebiotic conditions, and how nucleotides could have been assembled and polymerized without the assistance of enzymes. Researchers continue to investigate these issues, using RNA and related polymers. At present, the RNA world hypothesis is considered to be the most likely explanation for the emergence and evolution of modern organisms.

## SECTION 16.6 SUMMARY

- Ribozymes are important for catalyzing several RNA processing reactions, including self-cleavage of viral RNA replication intermediates and precursor tRNA processing.
- The RNA world hypothesis, based on the special properties of RNA that enable it to form stable functional structures and encode genetic information, postulates that RNA-based life predicated modern DNA-based organisms.

## Unanswered Questions

The study of RNA processing reactions has been a long-standing and active area of research, yet much remains to be deciphered.

- 1. Why do introns exist?** We don't yet know why there are introns and whether introns are ancient or more recent acquisitions in genes. Some introns have been found to encode regulatory RNA molecules that function in the processing of rRNA and in the control of gene expression levels (see Chapter 22). Whether these regulatory RNAs are a cause or a result of the presence of introns is not known. Although the origin of introns may remain uncertain, further insights about their roles in the continuing evolution of genomes will be exciting and may shed light on diseases that result from inaccurate intron removal and processing.
- 2. How does alternative splicing work?** Experimental methods, including microarray technology and genome-wide transcript sequencing, have revealed an abundance of alternative splicing in mammalian cells. However, the mechanics of such molecular gymnastics have yet to be determined. Future research will focus on how splicing is regulated, the frequency with which genes are alternatively spliced, and the roles of splicing regulation in disease.
- 3. How do ribozymes contribute to modern biology?** We don't have clear information on the origin and maintenance of ribozymes in particular biological niches. Some researchers have proposed that the spliceosome is a ribozyme, but this remains unproven. Current evidence suggests that the catalytic center of the spliceosome may, in fact, include both RNA and protein components. If true, such close association of RNA and protein would provide the first example of a true ribonucleoprotein enzyme linking RNA-based and protein-based catalysts.

# How We Know

## Studying Autoimmunity Led to the Discovery of snRNPs

**Lerner, M.R., and J.A. Steitz. 1979.** Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc. Natl. Acad. Sci. USA* 76:5495–5499.

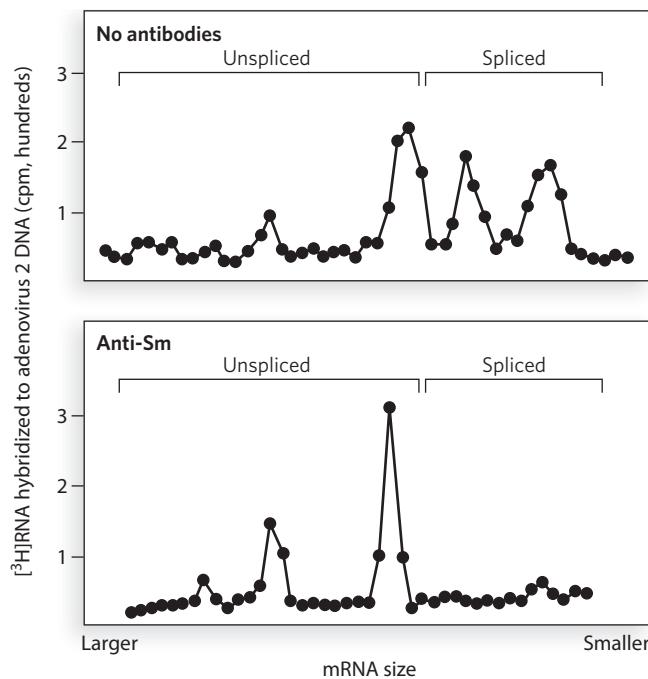


**Joan Steitz** [Source: Courtesy of Joan Steitz.]

autoimmune disease systemic lupus erythematosus, which causes, among other symptoms, a red facial rash, fatigue, and arthritis. Working with partially purified antibodies isolated from the blood of lupus patients, Steitz discovered that the antigens—the binding targets—of these antibodies are normal cellular particles containing a single small nuclear RNA complexed with proteins: snRNPs. Specifically, the autoantibodies associated with lupus recognize a set of snRNP proteins that were eventually named Sm, for the last name of the woman (Smith) whose serum samples were tested.

The snRNPs isolated by precipitation with anti-Sm antibodies were found to bind preferentially to RNAs containing intron-exon junctions, suggesting the involvement of these snRNPs in pre-mRNA splicing. In this experiment, nuclei isolated from cultured human cells were treated with radiolabeled UTP so that any newly synthesized RNA would incorporate the radiolabel. The nuclei were then incubated with anti-Sm antibodies under conditions in which the antibodies could enter the nucleus. After incubation for an hour, RNA was purified from the nuclei and fractionated by size, using agarose gel electrophoresis. The radiolabeled pre-mRNAs and mature RNAs (i.e., unspliced and spliced, respectively) were identified with specific DNA probes, and the relative amount of each was quantified based on the amount of radioactivity in each band in the gel. Radioactivity was plotted as a function of fraction number, which correlates with mRNA size (**Figure 1**). In the absence of added anti-Sm antibodies, both unspliced and spliced mRNAs were detected (top graph). In contrast, the presence of anti-Sm antibodies inhibited the production of spliced mRNAs, such that only a peak corresponding to unspliced pre-mRNAs was detected (bottom graph).

These data led to further studies that elucidated the specific roles of snRNPs in recognizing and catalyzing the removal of introns in the pre-mRNAs of all eukaryotic cells. This latter discovery was recognized by the Nobel Prize in Medicine, awarded to Phillip Sharp and Richard Roberts in 1993. We still don't know how Sm proteins can induce an autoimmune response, given that, presumably, they are sequestered within snRNPs in the nucleus.



**FIGURE 1** Results of labeling experiments showing that processing of pre-mRNA produced from adenovirus DNA is blocked by anti-Sm antibodies. In the absence of antibodies (upper panel), multiple fractions corresponding to spliced mRNAs are detected; in the presence of antibodies (lower panel), only larger, unspliced pre-mRNAs are detected. [Source: Adapted from V. W. Yang et al., *Proc. Natl. Acad. Sci. USA* 78:1371–1375, 1981.]

## RNA Molecules Are Fine-Tuned for Stability or Function

**Cerutti, P., J.W. Holt, and N. Miller. 1968.** Detection and determination of 5,6-dihydrouridine and 4-thiouridine in transfer ribonucleic acid from different sources. *J. Mol. Biol.* 34:505–518.

**Hughes, D.G., and B.E. Maden. 1978.** The pseudouridine contents of the ribosomal ribonucleic acids of three vertebrate species: Numerical correspondence between pseudouridine residues and 2'-O-methyl groups is not always conserved. *Biochem J.* 171:781–786.

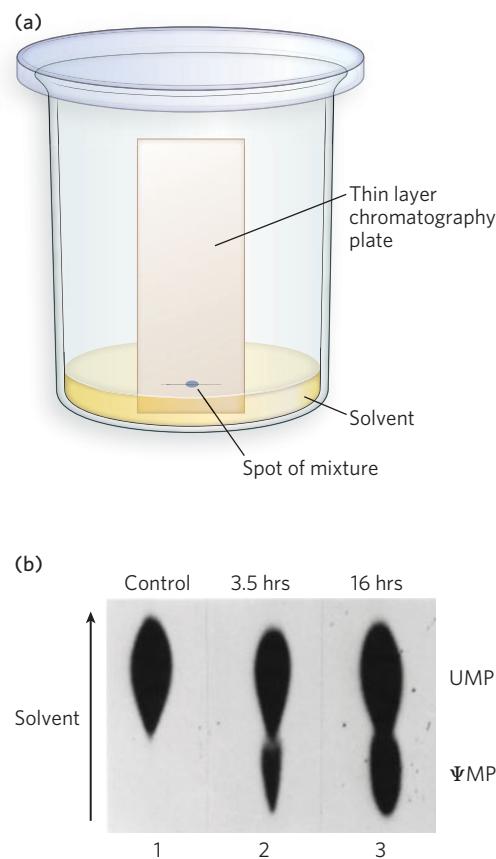
**Kuchino, Y., and E. Borek. 1978.** Tumour-specific phenylalanine tRNA contains two supernumerary methylated bases. *Nature* 271:126–129.

Unusual and chemically modified nucleotides in tRNAs, rRNAs, and a few other kinds of RNA molecules were discovered when these RNAs were purified from cells in sufficient quantity for classical analysis by thin layer chromatography—a standard technique for biochemists. In this method, RNA is hydrolyzed to its single-nucleotide components by ribonucleases, and the digestion products are applied to a glass or plastic plate coated with a thin layer of a chemical that absorbs liquid. One edge of the plate is placed in a solvent that is slowly absorbed upward through the surface layer (Figure 2). As the solvent moves through the sample and up the plate, different nucleotides in the sample move at different rates based on their solubility and their affinity for the surface material. The technique allows clean separation of A, G, C, and U nucleotides. Any chemically modified or noncanonical nucleotides that appear as extra “spots” on the plate can be scraped off for analysis by methods such as mass spectrometry.

In this way, investigators initially discovered that tRNAs contain a few nucleotides other than the standard four. How are such unusual bases synthesized, and why are they maintained? So far, research shows that all cells have sophisticated molecular machinery to produce unusual bases in tRNA and certain other RNAs. Some enzymes recognize a particular type of tRNA, excise the base from a specific nucleotide position, and replace it with another base; other enzymes chemically modify an existing base.

Modified bases seem to contribute in subtle ways to the thermodynamic stability of three-dimensionally structured RNAs such as tRNA and rRNA. For example, there are viable bacterial strains that differ only in their ability to modify rRNA at a specific site, due to the presence or absence of a particular rRNA-modifying enzyme. However, in a culture medium inoculated with equal amounts of the two strains, the strain containing the modifying enzyme eventually takes over the culture.

This result implies that the modification of rRNA contributes to the efficient function of ribosomes, despite requiring extra cellular energy input.



**FIGURE 2** Thin layer chromatography to detect chemically modified nucleotides. (a) RNase-digested tRNA is spotted onto a silica-coated glass plate, and solvent and nucleotides migrate up the plate by capillary action. (b) Modified and unmodified nucleotides—in this case, pseudouridine monophosphate ( $\Psi$ MP) and UMP, as control—migrate at different rates, based on their different solubilities in the solvent and their different affinities for the silica gel. [Source: (b) Y.-T. Yu, M.-D. Shu, and J. A. Steitz, *EMBO J.* 17:5783–5795, 1998, doi:10.1093/emboj/17.19.5783.]

## Ribozyme Form Explains Function

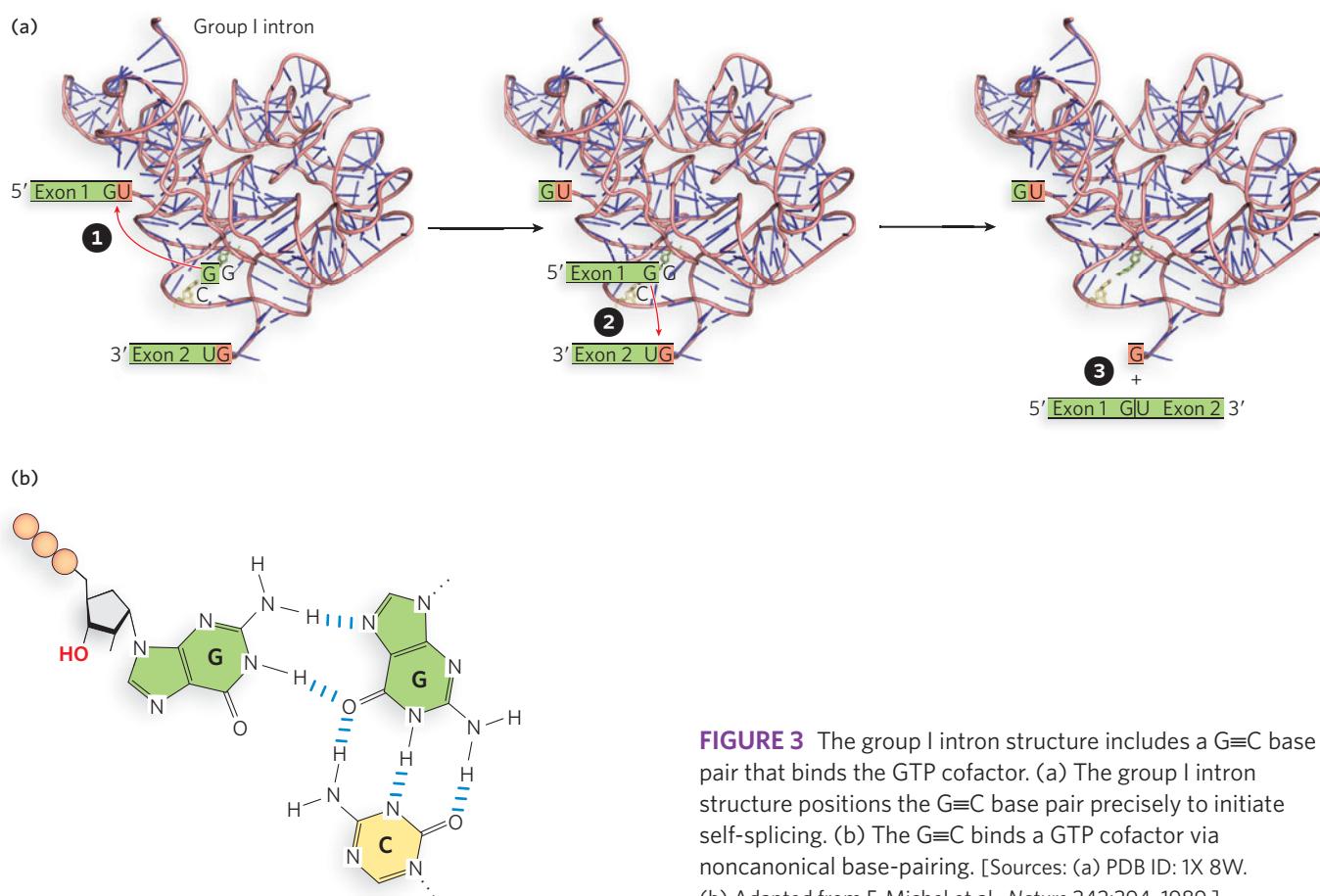
**Michel, F., M. Hanna, R. Green, D.P. Bartel, and J.W. Szostak. 1989.** The guanosine binding site of the *Tetrahymena* ribozyme. *Nature* 342:391–395.

**Stahley, M.R., and S.A. Strobel. 2006.** RNA splicing: Group I intron crystal structures reveal the basis of splice site selection and metal ion catalysis. *Curr. Opin. Struct. Biol.* 16:319–326.

The discovery of ribozymes coincided with an important technological advance for molecular biologists: the ability to transcribe RNA molecules of any sequence *in vitro*, and thereby test the function of RNAs in the complete absence of proteins. Furthermore, they could probe the molecular structure of ribozymes with chemicals that react with RNA nucleotides only when they are not involved in base-pairing interactions or packed against other nucleotides in the folded structure of the RNA.

An early observation was that mutations in the RNA sequence that disrupted parts of the three-dimensional structure also perturbed catalytic activity. For example, researchers noticed that a specific base pair in the

*Tetrahymena* group I self-splicing intron was present in the same place in the secondary structure of all related group I introns. Changing this base pair to any other base combination disrupted the self-splicing reaction, because the intron was no longer capable of binding efficiently to GTP, a cofactor in the splicing reaction (Figure 3). In this way, investigators discovered that the *Tetrahymena* group I intron (and later, other ribozymes) has a defined three-dimensional shape that is essential to catalytic activity. Using *in vitro*-transcribed and purified RNA, it was later possible to crystallize ribozymes and their component domains, revealing how these RNAs form active sites to enhance chemical reaction rates.



**FIGURE 3** The group I intron structure includes a G≡C base pair that binds the GTP cofactor. (a) The group I intron structure positions the G≡C base pair precisely to initiate self-splicing. (b) The G≡C binds a GTP cofactor via noncanonical base-pairing. [Sources: (a) PDB ID: 1X 8W. (b) Adapted from F. Michel et al., *Nature* 342:394, 1989.]

## Key Terms

primary transcript, p. 548	small nuclear ribonucleoprotein (snRNP), p. 557	RNA degradation, p. 571
intron, p. 549	small nuclear RNA (snRNA), p. 557	exosome, p. 571
exon, p. 549	group I intron, p. 561	processing body (P body), p. 571
5' cap, p. 549	group II intron, p. 561	preribosomal RNA (pre-rRNA), p. 573
3' poly(A) tail, p. 551	<i>trans</i> -splicing, p. 563	small nucleolar ribonucleoprotein (snoRNP), p. 575
poly(A) addition site, p. 551	RNA editing, p. 566	microRNA (miRNA), p. 575
RNA splicing, p. 554	editosome, p. 566	Drosha, p. 575
proteome, p. 556	adenosine deaminase acting on RNA (ADAR), p. 566	microprocessor complex, p. 575
alternative splicing, p. 556	exon junction complex (EJC), p. 570	Dicer, p. 576
poly(A) site choice, p. 556		ribozyme, p. 576
branch point, p. 557		

## Problems

- What would be the likely cellular effects of a large deletion in the gene encoding the polymerase responsible for adding 3' poly(A) tails to eukaryotic mRNAs?
- What is the minimum number of transesterification reactions required to splice an intron from a precursor transcript?
- Compare and contrast splicing mechanisms used by spliceosomes, group I introns, and group II introns, with respect to the nucleophiles, proteins, or nucleic acids involved and how the specificity of splice sites is achieved.
- Self-splicing introns do not require an energy source, such as ATP or GTP, to catalyze splicing. How does self-splicing proceed with a reasonable yield of products?
- Given what you have learned about the catalytic mechanism of group I self-splicing introns, propose an experiment to identify new group I introns in the total RNA isolated from an individual organism. Include a unique feature of the group I splicing reaction mechanism in your answer.
- Strains of bacteria lacking the enzymes required for rRNA modification have been engineered in the laboratory and seem to grow normally, despite the absence of modified rRNA. What would happen to these bacterial strains if they were forced to compete with wild-type bacteria? Explain.
- Is it correct to call an RNA molecule that catalyzes a reaction on itself an enzyme? Explain your answer.
- What accounts for the directionality of mRNA transport out of the nucleus?
- What would happen to the lifetime of a human mRNA if a nonhydrolyzable phosphodiester analog were introduced near its 5' end?
- What do the editing reactions catalyzed by APOBEC and ADAR enzymes have in common?
- All living systems share key properties (see Chapter 1). Why did the discovery of RNA catalysis trigger the development of an RNA world hypothesis as a stage in the evolution of life?
- Two different sequences in a primary mRNA transcript are critical to its cleavage in preparation for the addition of a poly(A) tail to the 3' end. What are those sequences, and which of them, if either, is (are) retained in the mature and modified mRNA?
- Following the first step in the splicing of a group II intron, an internal A residue in the intron is linked to other nucleotides by three phosphodiester bonds. What groups on the adenosine are involved in the phosphodiester bonds, and to what is each bond linked?
- If the tRNA nucleotidyltransferase or the enzyme that converts some U residues to pseudouridine in tRNA were inactivated in a eukaryotic cell, which inactivation would most likely be lethal?

## Data Analysis Problem

**Cech, T.R., A.J. Zaug, and P.J. Grabowski. 1981.** In vitro splicing of the ribosomal-RNA precursor of *Tetrahymena*: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 27:487–496.

**Kruger, K., P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling, and T.R. Cech. 1982.** Self-splicing RNA: Auto-excision and auto-cyclization of the ribosomal-RNA intervening sequence of *Tetrahymena*. *Cell* 31:147–157.

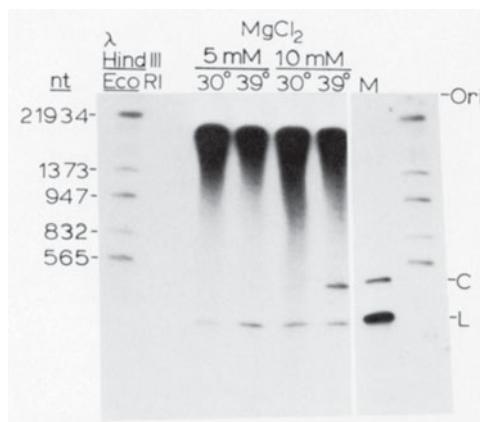
**15.** The discovery of RNA catalysis is sometimes cited as a classic case of serendipity, but it required a thoroughly prepared mind. In the late 1970s and early 1980s, RNA splicing was still a new concept. Thomas Cech and his colleagues set

out to investigate the splicing of an intron in rRNA of the protozoan *Tetrahymena thermophila*. They selected this RNA because rRNAs are much more abundant than most mRNAs, this rRNA has an intron, and *Tetrahymena* is a single-celled eukaryote that can be grown in large quantities. To carry out their early studies, Cech and colleagues devised a method to produce unspliced precursor rRNAs. As described in their 1981 paper, they isolated nuclei from *Tetrahymena* cells, lysed them, and added buffer, labeled rNTPs, and  $\alpha$ -amanitin, then incubated the reaction mix under conditions in which the cellular RNA polymerases would synthesize RNA. They extracted the

labeled RNA by a method that used phenol and chloroform treatments that normally remove most protein.

**(a)** Why did the researchers add  $\alpha$ -amanitin to the transcription reaction?

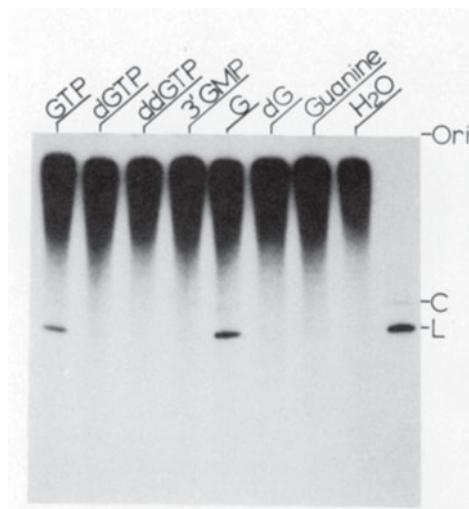
When they incubated the extracted RNA under the right conditions, they saw the production of excised intron RNA. Their result is shown in **Figure 1**. The bands in the first lane are size markers. The large bands at the top are the precursor rRNAs. The L and C labels indicate linear and circular forms of the excised intron RNA. Some different conditions of temperature and  $Mg^{2+}$  ion concentration were tried.



**FIGURE 1**

**(b)** The result demonstrated that the splicing reaction was working, but it did not demonstrate RNA catalysis. Why?

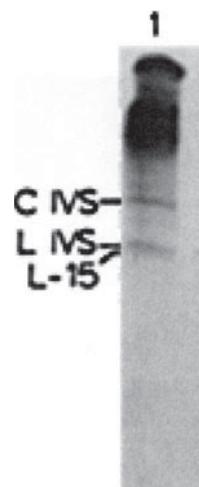
The excision reaction depended entirely on the addition of a guanine nucleotide. The researchers explored this requirement further, and some of their results are shown in **Figure 2**. The rRNA intron excision reaction is shown in the presence of a variety of potential cofactors. The abbreviations for the nucleosides and nucleotides are the standard ones described in Chapter 3. G and dG are the guanosine and nucleosides, with dG the 2'-deoxy form.



**FIGURE 2**

**(c)** From these results, what can you conclude about the nature of the required cofactor?

Intrigued that the splicing reaction did not seem to require the addition of cell extract (proteins), the researchers set out to determine whether proteins were required for the reaction. As described in their 1982 paper, they cloned a segment of the gene for the 26S rRNA, including the intron, in a bacterial plasmid, and expressed this gene segment in vitro using bacterial RNA polymerase. After deproteinizing the RNA product, they incubated the RNA with the buffer components and guanine nucleotide cofactor, as described in the earlier paper. Their results are shown in **Figure 3**. The C and L IVS are the circular and linear forms of the excised intron. The L-15 band is a cleaved form of the L IVS produced in a postsplicing cleavage reaction catalyzed by the IVS itself and not relevant to this problem.



**FIGURE 3**

**(d)** Does this experiment demonstrate self-splicing (RNA catalysis) of the intron? If so, why didn't the experiment described in the first paper do so?

## Additional Reading

### General and Historical

- Berget, S.M., C. Moore, and P.A. Sharp.** 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* 74:3171-3175.
- Chow, L.T., R.E. Glinas, T.R. Broker, and R.J. Roberts.** 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12:1-8.
- Mount, S.M., I. Pettersson, M. Hinterberger, A. Karmas, and J.A. Steitz.** 1983. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* 33:509-518.
- Padgett, R.A., S.M. Mount, J.A. Steitz, and P.A. Sharp.** 1983. Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. *Cell* 35:101-107.
- Steitz, J.A.** 1988. "Snurps." *Sci. Am.* 258(6):56-60, 63. A useful review.

### Pre-mRNA Splicing and RNA Editing

- Collins, C.A., and C. Guthrie.** 2000. The question remains: Is the spliceosome a ribozyme? *Nat. Struct. Biol.* 7:850-854. This short review provides insight into the catalytic mechanism of the spliceosome and its possible evolutionary origins.
- Le Hir, H., A. Nott, and M.J. Moore.** 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* 28:215-220. A summary of the ways in which splicing affects the expression of proteins in eukaryotic cells.
- Nishikura, K.** 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* 79:321-349
- Stuart, K.D., A. Schnaufer, N.L. Ernst, and A.K. Pani-grahi.** 2005. Complex management: RNA editing in trypanosomes. *Trends Biochem. Sci.* 30:97-105. A short and insightful review of RNA editing mechanisms.

### RNA Transport and Degradation

- Belasco, J.G.** 2010. All things must pass: Contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nat. Rev. Mol. Cell Biol.* 11:467-478. A discussion of the mechanistic parallels between the cellular factors and molecular events that govern mRNA degradation in eukaryotes and bacteria.

**Eulalio, A., I. Behm-Ansmant, and E. Izaurralde.** 2007. P bodies: At the crossroads of post-transcriptional pathways. *Nat. Rev. Mol. Cell Biol.* 8:9-22.

**Kindler, S., H. Wang, D. Richter, and H. Tiedge.** 2005. RNA transport and local control of translation. *Annu. Rev. Cell Dev. Biol.* 21:223-245.

**Parker, R., and H. Song.** 2004. The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct. Mol. Biol.* 11:121-127.

### Processing of Non-Protein-Coding RNAs

**Doudna, J.A., and T.R. Cech.** 2002. The chemical repertoire of natural ribozymes. *Nature* 418:222-228.

**Gingeras, T.R.** 2009. Implications of chimaeric non-collinear transcripts. *Nature* 461:206-211.

**Khalil, A.M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B.E. Bernstein, A. van Oudenaarden, et al.** 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* 106:11,667-11,672.

**Wilson, D.S., and J.W. Szostak.** 1999. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* 68:611-647.

### RNA Catalysis and the RNA World Hypothesis

**Hagiwara, Y., M.J. Field, O. Nureki, and M. Tateno.** 2010. Editing mechanism of aminoacyl-tRNA synthetases operates by a hybrid ribozyme/protein catalyst. *J. Am. Chem. Soc.* 132:2751-2758.

**Marvin, M.C., and D.R. Engelke.** 2009. Broadening the mission of an RNA enzyme. *J. Cell. Biochem.* 108: 1244-1251.

**Rios, A.C., and Y. Tor.** 2009. Model systems: How chemical biologists study RNA. *Curr. Opin. Chem. Biol.* 13:660-668.

# The Genetic Code



**Steve Benner** [Source: Courtesy of Steve Benner.]

Death Valley to collect rocks. While there, I began musing about the borate-containing rock samples I was finding, and thinking about the long-known observation that borate can bind to organic molecules that contain 1,2-dihydroxyl groups—exactly the kind of chemical structure present in ribose.

When I returned to the lab, it only took about a day and a half to show experimentally that *ribose could be made stably at high pH in the presence of borate*. Because borate is abundant in nature, it seems likely that it stabilized the prebiotic production of ribose, providing a simple and logical explanation for the presence of ribose on the early Earth. It was satisfying to make this discovery, but also humbling to realize that borate-carbohydrate interactions have been known since the 1950s. So the answer to the ribose stability problem has been staring us in the face all along!

—Steve Benner, on discovering that borate minerals stabilize ribose

## Moment of Discovery

The origin of life has long been an interest of mine, particularly the evolution of nucleic acids and the reason that ribose was selected as the sugar used in RNA. In the 1950s, Stanley Miller and others showed that ribose can be produced abiotically (without enzymes), but Miller and others had noted that ribose is not very stable. This is because ribose and other five-carbon sugars are made under alkaline conditions from simple organic precursors, formaldehyde and glycolaldehyde; a high pH encourages reasonable reaction rates, but the ribose product tends to break down quickly into a brown tar.

Although we weren't actually studying this particular problem, a comment from a colleague about "solving the ribose problem" coincided with a trip I took to

**17.1 Deciphering the Genetic Code: tRNA as Adaptor 586**

**17.2 The Rules of the Code 593**

**17.3 Cracking the Code 596**

**17.4 Exceptions Proving the Rules 601**

The discoveries that DNA is composed of complementary strands and that it holds the instructions for all the proteins in an organism were huge advances in our understanding of the flow of biological information. Proteins and nucleic acids are very different types of chemicals, however, and after the structure of DNA was solved, how the sequence in a chain of nucleotides determines the sequence of amino acids in a protein was not immediately apparent. The next 10 years brought several discoveries that revealed the fascinating processes by which DNA is decoded to produce proteins.

The linear nucleotide sequence of mRNA is translated into protein by tRNA molecules that carry amino acids and contain nucleotide sequences (called anticodons) that pair with complementary sequences (codons) in the mRNA. Different amino acid-carrying tRNAs are lined up according to the mRNA sequence, and the amino acids are stitched together by the ribosome, resulting in a polypeptide with a linear order of amino acids that corresponds to the linear order of codon sequences in the mRNA. The discovery of the translation process and the **genetic code**, the matching of each codon to the amino acid it specifies, is a fascinating story and a landmark in modern science. The genetic code is universal—that is, it is nearly the same in all cells—and thus provides very strong evidence for a common ancestor for all cell types in existence today.

Given that amino acids and nucleotide bases have no obvious chemical relationship, there is no obvious reason to think that a given amino acid should be matched to a particular nucleotide sequence. Yet all organisms—bacteria, yeast, amphibians, plants, archaea, and humans—use the same genetic code. Presumably, once the code had evolved, it resisted change. The universality of the genetic code provides amazingly strong, *molecular* evidence for evolution, much more compelling than arguments based on body shapes and the fossil record.

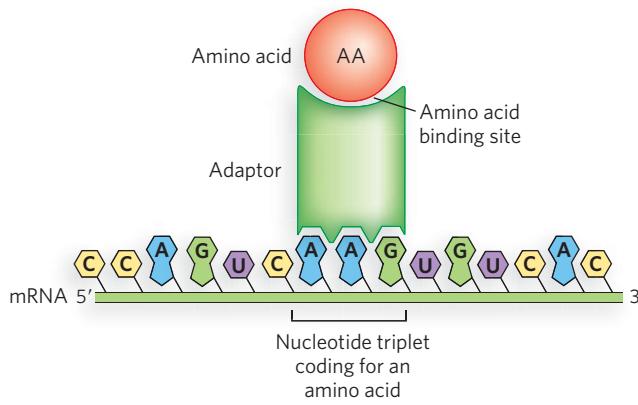
This chapter presents an overview of the genetic code and how it works. We first look at how the tRNA molecule functions in decoding, and how it is exquisitely designed to take advantage of the “degeneracy” of codons, enabling one tRNA to decipher more than one codon. We also examine how the genetic code can resist the harmful effects of single-nucleotide mutations. These special features indicate that the genetic code is not simply an accident of evolution, but has been fine-tuned by natural selection. The last universal common ancestor (LUCA) must have existed for sufficient time to hone the code prior to divergence of the different domains of life as we know them today. Finally, we look at exceptions to the genetic code—variations that only reinforce the idea that all life forms evolved from LUCA and its genetic code. The way in which the genetic code

came into being during evolution is still a perplexing problem. We examine this issue, too, even though there are no clear answers.

## 17.1 Deciphering the Genetic Code: tRNA as Adaptor

DNA and RNA each consist of only four different nucleotides, whereas proteins can have up to 20 different amino acids. For only four nucleotides to specify the 20 common amino acids, multiple nucleotides must be combined to make up a code. Combinations of two nucleotides yield only 16 ( $4^2$ ) different dinucleotide code words, insufficient to encode 20 amino acids. Combinations of three nucleotides yield 64 ( $4^3$ ) code words, more than enough to specify 20 amino acids. Hence, the RNA “code word,” or **codon**, was hypothesized to be a combination of three nucleotides, or possibly more. Insightful experiments, described in this chapter, demonstrated that the code is indeed triplet.

To explain how an RNA sequence codes for a sequence of amino acids, Francis Crick, in 1955, hypothesized the existence of an “adaptor” molecule. He proposed that adaptors can recognize specific codons in the mRNA, and that each adaptor carries a specific amino acid (Figure 17-1). Adaptors line up on the mRNA, thus aligning the sequence of amino acids. Not long after Crick’s adaptor hypothesis, Paul Zamecnik and Mahlon Hoagland discovered a small RNA to which amino acids become covalently attached in an ATP-dependent reaction (see How We Know). These RNA–amino acid hybrids could



**FIGURE 17-1 Crick's adaptor hypothesis.** Adaptor molecules recognize codons in mRNA and carry specific amino acids. Thus, they line up amino acids in an order that depends on the sequence of codons in the mRNA. Today we know that the adaptor is a tRNA molecule. The amino acid is covalently bound at the 3' end of the tRNA molecule, and a specific nucleotide triplet (anticodon) elsewhere in the tRNA interacts with a triplet codon in mRNA through hydrogen bonding of complementary bases.

presumably base-pair with mRNA, because they contain nucleotides and thus fit the description of the adaptor molecule needed to translate the information in mRNA sequence into a polypeptide sequence. This small RNA, later called **transfer RNA (tRNA)**, is aminoacylated at the 3' terminus in an ATP-dependent reaction. A tRNA with its attached amino acid is called an **aminoacyl-tRNA**, and the tRNA is said to be *charged* with that amino acid. The amino acid specificity of tRNAs is provided not by the anticodon in their nucleotide sequence but by the enzymes that attach amino acids to particular tRNAs, enzymes known as **aminoacyl-tRNA synthetases**. Thus, the association between the amino acid and the anticodon is not chemically determined, but rather it evolved.

### KEY CONVENTION

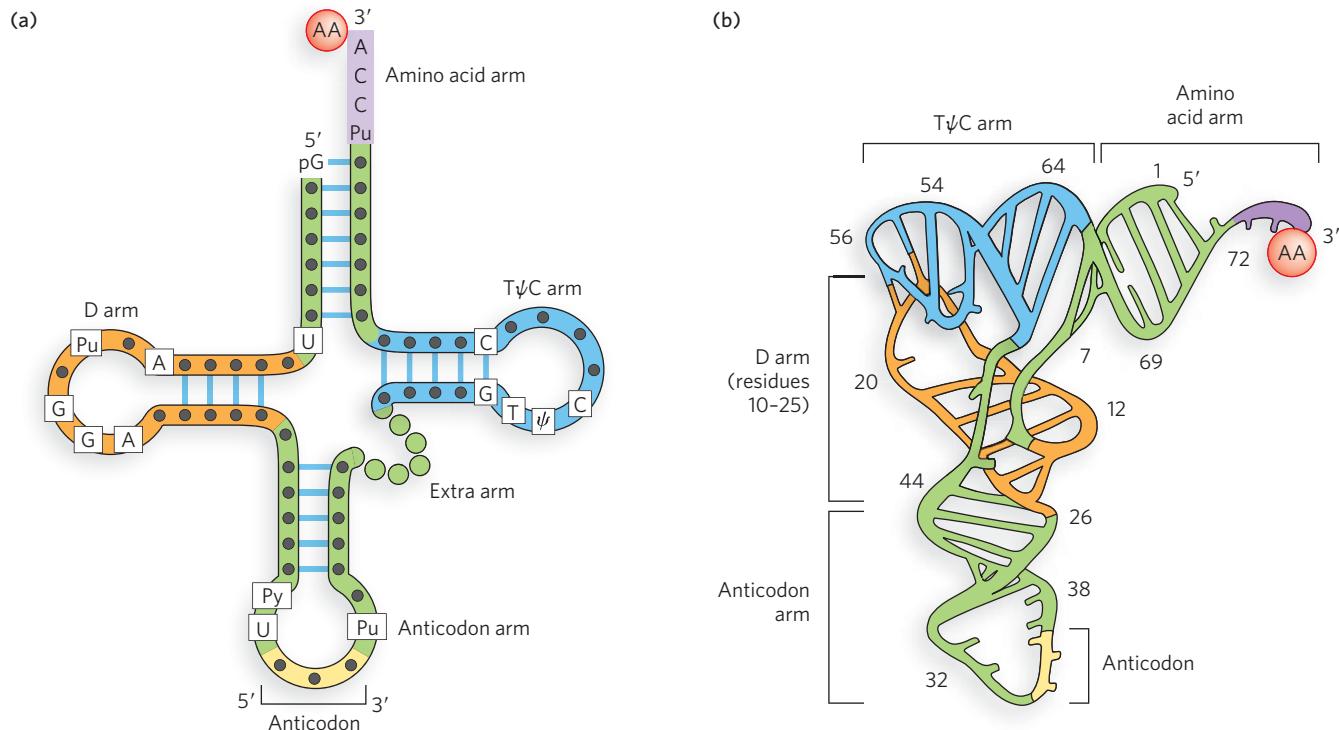
In denoting tRNAs, the specificity is indicated by a superscript, and the aminoacylated-tRNA by a hyphenated name. For example, tRNA<sup>Leu</sup> indicates an uncharged tRNA that is specific for leucine, and leucyl-tRNA<sup>Leu</sup>, or Leu-tRNA<sup>Leu</sup>, indicates a leucine-specific tRNA that is charged with leucine.

### All tRNAs Have a Similar Structure

The structure of the tRNA molecule reveals how it is capable of functioning as an adaptor. We briefly discuss tRNA structure here, and then explore this topic in greater detail in Chapter 18. Transfer RNAs are relatively small, single-stranded RNA molecules.

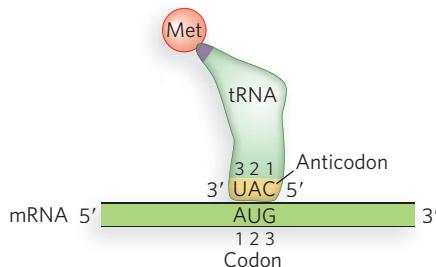
The tRNAs in bacteria and in the cytoplasm of eukaryotic cells are 73 to 93 nucleotide residues long. Mitochondria and chloroplasts contain distinctive, somewhat smaller tRNAs. All tRNAs form intramolecular base pairs and fold up into a precise three-dimensional structure. They contain the trinucleotide sequence CCA at the 3' terminus. The 3'-terminal A residue is the nucleotide to which the amino acid becomes attached.

When drawn in two dimensions, the hydrogen-bonding pattern of all tRNAs forms a cloverleaf structure with four arms; the longer tRNAs have a short fifth arm, or extra arm (Figure 17-2a). In three dimensions, a tRNA folds up further into the form of a twisted L (Figure 17-2b). Two arms of the tRNA are critical for adaptor function, and they are located at the two ends of the L shape. The 3' terminus of the amino acid arm carries a specific amino acid. At the



**FIGURE 17-2** The structure of tRNA. (a) The cloverleaf form. Large dots on the backbone represent the nucleotide residues; blue lines are base pairs. Three nucleotides constitute the anticodon (at the bottom), and amino acids are attached to the 3'-terminal amino acid arm. Unusual modified nucleotides are present in the D arm and T $\psi$ C arm.

Py (pyrimidine) can be either U or C. Pu (purine) can be either A or G. The D arm often contains a dihydrouracil residue (not shown). (b) The three-dimensional structure folds into a twisted L shape. The positions of the anticodon, the 3'-terminal amino acid arm, and the D and T $\psi$ C arms are shown.



**FIGURE 17-3** The pairing relationship of the codon and anticodon. The Met-tRNA<sup>Met</sup> (methionyl-tRNA<sup>Met</sup>) is shown. The nucleotide positions of the codon and anticodon are numbered 1, 2, and 3, in the 5'→3' direction (thus, anticodon nucleotide 3 pairs with codon nucleotide 1).

opposite end is the anticodon arm, so named because it contains the **anticodon**, the three-nucleotide sequence that base-pairs with the complementary codon in mRNA. The other major arms are the D arm, which often contains the unusual nucleotide dihydrouridine (D), and the TψC arm, containing ribothymidine (T) and pseudouridine (ψ), which has an unusual carbon–carbon bond between the base and ribose. The base pairing between the anticodon in the tRNA and the codon in mRNA is antiparallel. For example, the codon for methionine is 5'-AUG, which base-pairs with the tRNA<sup>Met</sup> anticodon 5'-CAU (i.e., 3'-UAC) (Figure 17-3).

#### KEY CONVENTION

The nucleotide positions of the codon (mRNA) and anticodon (tRNA) are numbered 1, 2, and 3, in the 5'→3' direction. Due to the antiparallel base pairing between the anticodon and the codon, the numbering of nucleotides in the anticodon is the reverse of that in the codon. Thus, anticodon nucleotide 3 pairs with codon nucleotide 1.

As shown in Figure 17-2, the anticodon is quite a distance from the 3' terminus of the amino acid arm (where the amino acid is attached), and thus the anticodon cannot directly specify the correct amino acid. Indeed, the ribosome will link any two amino acids lined up correctly on the mRNA, regardless of whether the tRNA is charged with a correct or an incorrect amino acid. It is the function of the aminoacyl-tRNA synthetases to place the correct amino acid onto the tRNA. Therefore, the specificity of the genetic code lies in the accuracy of protein-based aminoacylation of the tRNAs. Most cells contain 20 aminoacyl-tRNA synthetases, one for each amino acid. Because there are more codons than amino acids, some amino acids are specified by

more than one tRNA, yet the same aminoacyl-tRNA synthetase recognizes all tRNAs that specify a given amino acid. The ribosome binds the mRNA and charged tRNAs, bringing the components into proximity for linking together the amino acids attached to adjacent aminoacyl-tRNAs as they align on the mRNA. The entire process of decoding the linear sequence of mRNA into the sequence of a protein is known as **translation** and requires more than 100 different types of protein and RNA molecules (see Chapter 18).

#### The Genetic Code Is Degenerate

As we've seen, there are 64 unique ways to combine four different nucleotides in a triplet codon sequence, yet there are only 20 amino acids. Therefore, either some codons are not found in mRNA sequences, or, as we now know, multiple codons encode the same amino acid. A **degenerate code** is one in which several code words have the same meaning. We refer to the genetic code as degenerate because a single amino acid can be encoded by more than one codon. As we'll see later, the degeneracy of the genetic code is advantageous because it provides the DNA with the ability to absorb single-base mutations with minimal consequences for the protein sequences it encodes.

All 64 codons of the genetic code are used in some fashion: 61 for coding amino acids and 3 for specifying the termination of translation (Figure 17-4). Three

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

**FIGURE 17-4** The genetic code. The codon sequences are written in the 5'→3' direction. The first nucleotide of each codon is shown on the left side of the grid, the second nucleotide at the top, and the third on the right. AUG (shaded green) also serves as the start codon; UAA, UAG, and UGA (red) are stop (or nonsense) codons.

**Table 17-1** The Degeneracy of the Genetic Code

Amino Acid	Number of Codons	Amino Acid	Number of Codons
Arg	6	Asp	2
Leu	6	Cys	2
Ser	6	Gln	2
Ala	4	Glu	2
Gly	4	His	2
Pro	4	Lys	2
Thr	4	Phe	2
Val	4	Tyr	2
Ile	3	Met	1
Asn	2	Trp	1

amino acids, arginine, leucine, and serine, are each specified by six different codons. Five amino acids have four codons, isoleucine has three, and nine amino acids have two codons. Only two amino acids, tryptophan and methionine, are specified by a single codon (Table 17-1).

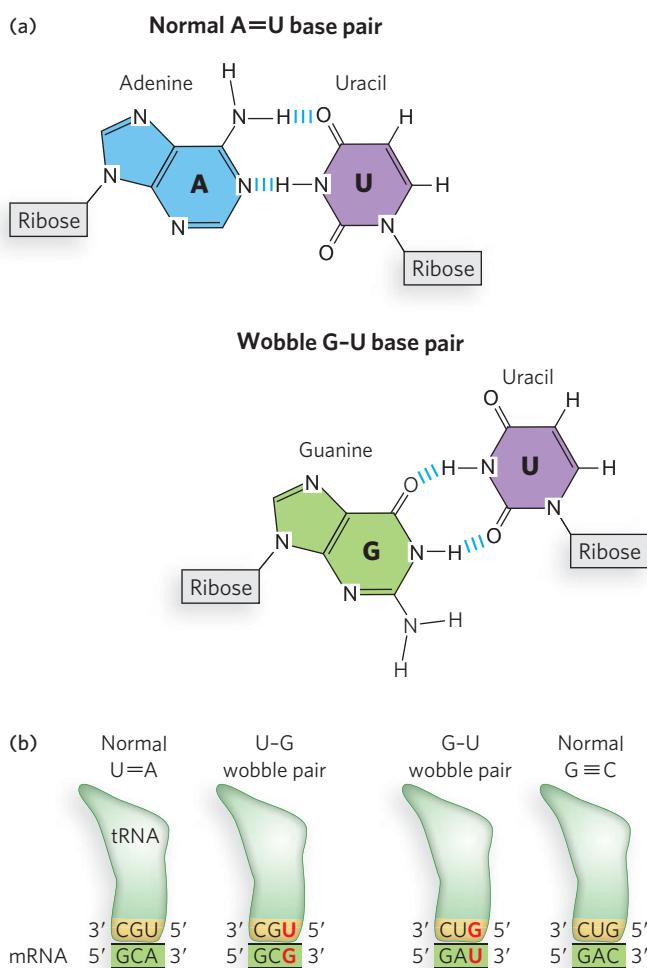
When several different codons specify one amino acid, the first two nucleotides of each codon are the primary determinants of specificity, and the difference between the codons usually lies at the third position. For example, alanine is specified by the triplets GCU, GCC, GCA, and GCG. When four codons specify the same amino acid, they are referred to as a **codon family**. Within a codon family, the first two nucleotides are the same; the nucleotide at the third position does not matter, and base pairing of the first two nucleotides carries the information needed to specify the amino acid. Many amino acids are specified by two codons in which the third nucleotide is either a purine in both or a pyrimidine in both.

### Wobble Enables One tRNA to Recognize Two or More Codons

If all three nucleotides in an mRNA codon were needed to form Watson-Crick base pairs with their counterparts in the tRNA anticodon, 61 different tRNAs would be required in every cell. In fact, only 32 tRNAs are required to recognize all the amino acid codons, because some tRNAs recognize more than one codon. However, some cells contain considerably more than 32 different tRNAs.

As we've seen, when several codons specify the same amino acid, usually the third nucleotide is the only difference. In some cases the cell uses different tRNAs for the different codons that encode the same amino acid, and in

these cases a single aminoacyl-tRNA transferase recognizes the various tRNAs and charges them all with the same amino acid. Many tRNAs can recognize more than one codon, and these tRNAs often contain either a U or a G as the 5' nucleotide of the anticodon (i.e., in position 1, which pairs with the third nucleotide of the codon), because these nucleotides can form noncanonical base pairs: U can pair with either A or G, and G can pair with either C or U (Figure 17-5). We don't see these noncanonical base pairs in DNA because they do not fit within the tight geometric constraints of the DNA duplex, but they are accommodated in the more flexible base pairing that occurs between tRNA and mRNA. The bases that participate in



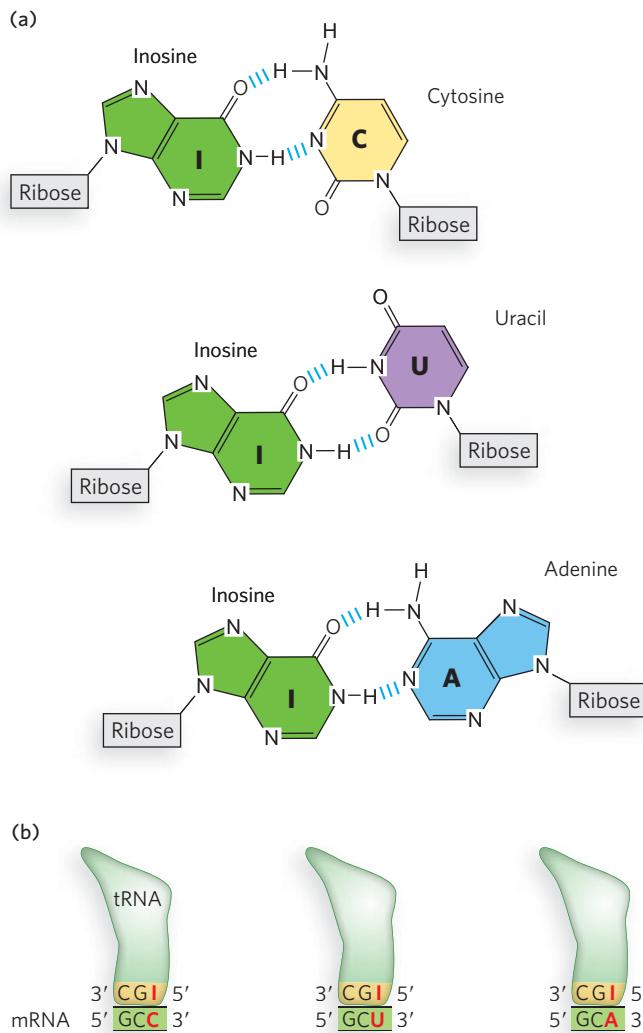
**FIGURE 17-5** Wobble base pairing. (a) Wobble allows one tRNA to recognize two different codons. U normally pairs with A (top) but can form two hydrogen bonds with G to make a weak G-U wobble base pair, which occurs in the third position of the codon (bottom). (b) A tRNA pairs with two different codons through wobble pairing (shown in red) at the third nucleotide of the codon. The G of the G-U pair can be in either the anticodon (left) or the codon (right).

noncanonical base pairs are called **wobble bases**. The wobble bases allow a single tRNA anticodon to bind to more than one mRNA codon. The 5' nucleotide in the anticodon is in the **wobble position**. It is important to note, however, that the structure of tRNA can make the anticodon completely specific for perfect base pairing with one codon. We see this for tryptophan and methionine, each of which has only one codon. Thus, the necessary “flexibility” needed for wobble in tRNA involves more than the anticodon sequence—it also involves the larger structure of the tRNA. The way in which tRNA structure confers wobble on the anticodon is not completely understood.

The anticodon in some tRNAs includes inosine (designated I; this nucleotide residue contains the base hypoxanthine), which can form hydrogen bonds with any of three different nucleotides: U, C, or A (Figure 17-6). These pairings are much weaker than the hydrogen bonds of Watson-Crick base pairs. When Robert Holley sequenced the yeast tRNA<sup>Ala</sup> in 1965, he found inosine at the first position of the anticodon. This explains why the anticodon of yeast tRNA<sup>Ala</sup>, 5'-IGC, can function with three different codons: 5'-GCA, 5'-GCU, and 5'-GCC. Inosine, like the other modified nucleotides in tRNA, is formed posttranscriptionally—adenosine is deaminated by the enzyme adenosine deaminase to produce a keto moiety in place of the amino group.

The process by which some tRNAs can recognize more than one codon was formalized by Crick, who proposed a set of four relationships known as the **wobble hypothesis**:

1. The first two bases of an mRNA codon always form Watson-Crick base pairs with the corresponding bases of the tRNA anticodon, and confer most of the coding specificity.
2. The first base of the anticodon (reading in the 5'→3' direction) pairs with the third base of the codon and determines the number of codons recognized by the tRNA. When the first nucleotide of the anticodon is C or A, base pairing is specific, and only one codon is recognized by that tRNA. When the first nucleotide is U or G, base pairing is less specific, and two different codons may be read by the same tRNA. When the first nucleotide of an anticodon is I, three different codons can be recognized—the maximum number for any tRNA.
3. When an amino acid is specified by several different codons, codons that differ in either of the first two bases require different tRNAs.
4. A minimum of 32 tRNAs are required to translate all 61 codons (31 to encode the amino acids and 1 for initiation).



**FIGURE 17-6** Inosine as a wobble nucleotide. (a) Inosine (I) can form two hydrogen bonds with either C, U, or A. (b) A tRNA containing I in the first position of the anticodon can recognize three different codons, according to the wobble rules. Wobble pairings are shown in red.

## Translation Is Started and Stopped by Specific Codons

As we'll see in Section 17.2 and in further detail in Chapter 18, the codons in an mRNA molecule are read by the ribosome in the 5'→3' direction, without gaps. Because each codon has three nucleotides, an mRNA sequence has the potential to encode three different polypeptide sequences, depending on exactly where translation begins—that is, depending on which register of triplets the translation apparatus acts upon (Figure 17-7). Each register of triplets in mRNA is called a **reading frame**. The amino acid sequence of the protein encoded by the mRNA depends on which reading frame is used.

AUG GUG CGU AGG GUC GAU UGG CGC AGA AAG UUA GUU AGA GAG UAC
Met Val Arg Arg Val Asp Trp Arg Arg Lys Leu Val Arg Glu Tyr
A UGG UGC GUA GGG UCG AUU GGC GCA GAA AGU UAG UUA GAG AGU AC
Trp Cys Val Gly Ser Ile Gly Ala Glu Ser Stop Leu Glu Ser
AU GGU GCG UAG GGU CGA UUG GCG CAG AAA GUU AGU UAG AGA GUA C
Gly Ala Stop Gly Arg Leu Ala Gln Lys Val Ser Stop Arg Val

**FIGURE 17-7** Three possible reading frames. Shown here is a single RNA sequence translated in all three of its reading frames. The protein product is below each sequence.

Specific sequences in mRNA signal the start of translation and thus define the reading frame. Translation almost always starts at an AUG codon, which specifies the amino acid methionine; this codon is referred to as the **initiation codon** or **start codon**. Occasionally, the codon GUG (usually encoding valine) or UUG (usually encoding leucine) is used as an initiation codon. The mRNA can also have internal AUG (or GUG and UUG) codons, yet translation does not begin at these internal positions. In bacteria, there is a specific sequence in the mRNA next to the initiating AUG (or GUG) that binds the ribosome and directs it to start translation. In eukaryotes, the ribosome is directed to the 5' terminus of the mRNA, after which it slides down the mRNA and begins translation at the first AUG codon it encounters.

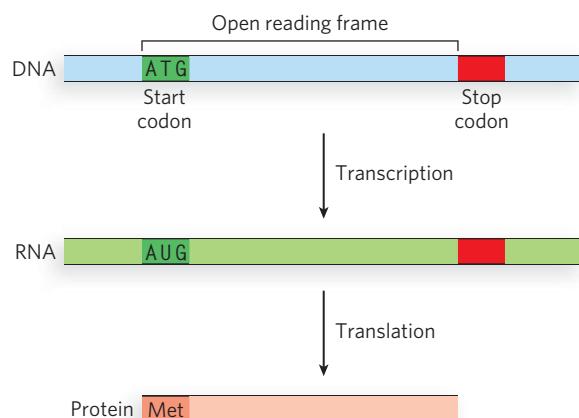
The three codons (UAA, UAG, and UGA) that signal the end of translation and do not specify any amino acid are **termination codons** or **stop codons** (also sometimes called nonsense codons). Termination codons signal the ribosome to dissociate from the newly synthesized polypeptide chain. When the ribosome encounters a stop codon, a release factor associates with the ribosome and terminates protein synthesis. Release factors, even though they recognize specific codons, are actually proteins. In a fascinating display of molecular mimicry, the three-dimensional structure of release factor proteins is very similar to the structure of tRNA.

With 3 of the 64 codons acting as terminators, a random mRNA sequence should contain 1 stop codon about every 20 codons. A long sequence of nucleotide triplets with no stop codons is unlikely to occur by chance, and it generally encodes a protein. Such a sequence is known as an **open reading frame**, or **ORF** (Figure 17-8). For example, the average length of a gene in *E. coli* is 1,000 nucleotides, or about 333 codons clear of termination codons.

### The Genetic Code Resists Single-Base Substitution Mutations

The degeneracy of the genetic code enables it to absorb many types of point mutations without serious consequence. (See Chapter 12 for a fuller discussion of types of mutations.) A single-base substitution that leads to the replacement of one amino acid with another is a **missense mutation**. However, because the genetic code is degenerate, many single-base substitutions are **silent mutations** that do not result in an amino acid replacement. For example, a nucleotide change in the third position of a codon results in an amino acid change only about 25% of the time.

The ability of the code to withstand mutation is even more apparent when we consider that the most



**FIGURE 17-8** Start and stop signals in the open reading frame of a gene. The reading frame of a gene that encodes a protein begins at an ATG start codon in the coding strand of the DNA (AUG in the mRNA) and ends at the first stop codon in the same reading frame as the start codon.

frequent mutation is a **transition mutation**, in which a purine is replaced by another purine (A=T replaced by G≡C, or G≡C by A=T). All three positions of the codon confer some type of protection from deleterious transition mutations. Transition mutations in the third position rarely cause a change at all, due to the wobble rules. Even the functioning of UAA and UAG stop codons is protected from damage by a transition mutation in the third position.

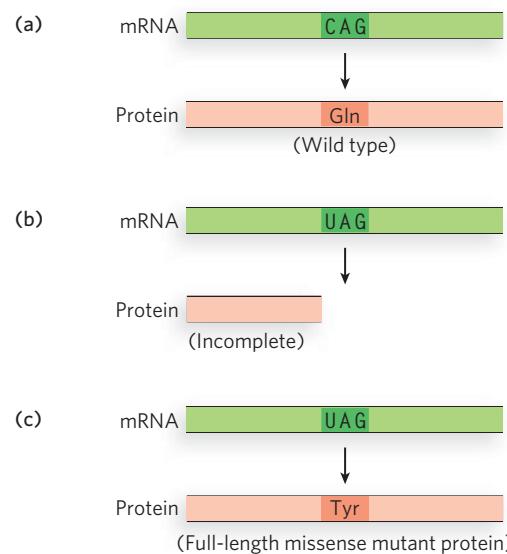
A transition mutation in the first position of most codons does result in an amino acid change, but the change is usually to an amino acid that is chemically similar to the original amino acid. This is especially evident for hydrophobic amino acids, as shown in the left-most column of Figure 17-4. These codons contain U in the second position, and replacement of the first nucleotide results in a codon that specifies another hydrophobic residue. For example, a codon change of GUU to AUU results in an exchange of Ile for Val. Had the GUU codon been altered to CUU, the protein would contain Leu instead of Val. These amino acids have similar chemical properties and thus are much more likely to conserve the protein function than if a hydrophobic residue were replaced by a polar residue. The second position of a codon generally determines whether it encodes a polar (if nucleotide 2 is a purine) or hydrophobic (if a pyrimidine) amino acid. Therefore, transition mutations in the second position also tend to conserve the chemical nature of the protein product.

Errors produced during translation occur most frequently in the codon's first and third nucleotide positions. The redundancy in coding due to wobble in the third position removes most errors. Eight amino acids are specified by codons that contain any of the four nucleotides in the third position. This, coupled with the fact that any purine-pyrimidine mispairing in the wobble position results in the same amino acid in all but three cases, greatly reduces the effect of reading errors at the ribosome. Just as transition mutations generally lead to a conservative change, misreading of purine-pyrimidine codon-anticodon base pairs results in conservative changes.

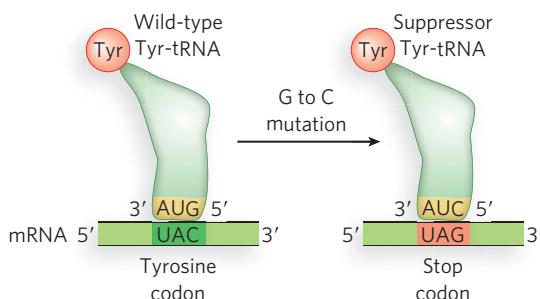
Computational studies that examine the ability of theoretical random genetic codes to withstand the effects of mutation show that most codes arrived at randomly would be much less resistant to mutation than is the code actually used by cells. In fact, the probability of arriving by chance at a code that is as resistant to mutation as the genetic code of living organisms is about one in a million. These considerations suggest that the code was extensively honed by natural selection before the divergence of other life forms from LUCA, the ancestral cell.

## Some Mutations Are Suppressed by Special tRNAs

Far more deleterious than missense mutations are codon changes that result in a termination codon. These **nonsense mutations** abort protein synthesis, resulting in an incomplete protein, which is rarely functional. The gene can be restored to function by a second mutation that converts the nonsense codon to a missense codon or by a mutation in a tRNA that suppresses termination at the nonsense codon by inserting an amino acid at that position (Figure 17-9). Mutant tRNAs that function at a stop codon to allow translation to continue are called **suppressor tRNAs**. For example, a change in the anticodon of tRNA<sup>Tyr</sup> from 5'-GUA to 5'-CUA results in an altered tRNA<sup>Tyr</sup> that inserts tyrosine at a 5'-UAG termination codon (Figure 17-10). Depending on the suppressor tRNA, other amino acids could be inserted at a 5'-UAG termination codon. In theory, any tRNA with an anticodon that is one base pair different from a stop codon could become a suppressor tRNA if a single point mutation occurred in the right place in the anticodon. In fact, suppressor mutations are rare *in vivo*, but this phenomenon has been harnessed as a tool in the molecular biology laboratory.



**FIGURE 17-9 Suppression of a nonsense mutation.** (a) The wild-type mRNA encodes a full-length protein, with CAG encoding glutamine. (b) A nonsense mutation at an internal CAG codon changes it to a UAG termination codon, resulting in an incomplete protein. (c) A tRNA<sup>Tyr</sup> suppressor has a mutant anticodon that pairs with the UAG nonsense (stop) codon, resulting in a full-length protein with a Tyr residue in place of the Gln residue of the wild-type protein.



**FIGURE 17-10** The structure of a suppressor tRNA. The tRNA<sup>Tyr</sup> with anticodon 5'-GUA recognizes the UAC codon. The suppressor tRNA<sup>Tyr</sup> contains a mutation in the anticodon, altering it to 5'-CUA, which base-pairs with the UAG nonsense (stop) codon and inserts a Tyr residue in the protein.

Although suppressor tRNAs usually carry a single-nucleotide change in the anticodon, some mutations in suppressor tRNAs lie outside the anticodon. For example, the suppressor of UGA nonsense codons usually involves a tRNA<sup>Trp</sup> that recognizes UGG. The mutation that provides this ability to recognize UGA (and insert tryptophan at this position) can be in the anticodon, but it can also be due to a change of G to A at nucleotide position 24 in the D arm of tRNA<sup>Trp</sup>. This change is presumed to lead to an altered conformation that can now recognize both the normal UGG codon and the UGA stop codon. There are other instances in which codon recognition by a tRNA is altered by mutations outside the anticodon, and these are probably mediated by effects on the larger tRNA structure, although more research is needed to clarify the mechanism.

Suppression must not be too efficient, otherwise normal termination codons would also be suppressed, leading to abnormally long protein products—an outcome that would be lethal to the cell. Suppression is limited in several ways. Many genes are terminated by more than a single stop codon. But more importantly, there are multiple copies of each tRNA gene, even in cells that are not diploid. Some duplicate tRNA polynucleotide chains are weakly expressed and thus constitute only a small fraction of the tRNA pool for a particular amino acid. Suppressor mutants are typically found in one of these minor tRNA genes, leaving the major tRNA gene to function normally.

An example of suppression in *E. coli* is tRNA<sup>Tyr</sup> with the anticodon 5'-GUA. *E. coli* contains three identical tRNA<sup>Tyr</sup> genes, but one is much more highly transcribed than the others. The tRNA<sup>Tyr</sup> suppressor mutation, which changes the anticodon to 5'-CUA and thus recognizes the 5'-UAG stop codon, is found in one of the minor, less-transcribed tRNA<sup>Tyr</sup> genes. Therefore,

the insertion of tyrosine at UAG stop codons is inefficient, but sufficient full-length protein is produced from a gene with a nonsense mutation to let the cell survive. Furthermore, UAG is used only rarely as a stop codon in *E. coli*. This allows suppression to be reasonably efficient, up to 50%, at UAG stop codons. In comparison, suppression at the more frequently used UAA and UGA stop codons must be kept below 5% to ensure cell viability. There are also examples of suppressor tRNAs for missense mutations, and suppressor tRNAs for frameshift mutations, which place the ribosome in an incorrect reading frame by insertion or deletion of a nucleotide.

## SECTION 17.1 SUMMARY

- Transfer RNAs are small RNA molecules that can covalently attach at the 3' end to an amino acid. The triplet anticodon in tRNA pairs with a triplet codon in mRNA, and this pairing mediates translation of the nucleotide sequence in mRNA into the amino acid sequence of a protein.
- The genetic code is said to be degenerate, because most amino acids are specified by two or more codons. One tRNA often reads two codon sequences, due to noncanonical or wobble base pairing at the third nucleotide position of a codon. When the anticodon contains inosine (I), a modified nucleotide residue, the tRNA recognizes three different codons, ending in A, C, or U.
- An AUG codon, specifying methionine, typically initiates protein synthesis. The three termination codons do not specify an amino acid, but instead instruct the ribosome to stop translation.
- Due to codon assignments and the degeneracy of the genetic code, single-base substitution mutations generally result in codons that specify the same, or similar, amino acids; but nonsense mutations result in a stop codon that can lead to inactive protein. Mutant tRNAs that carry a single-nucleotide change in the anticodon can suppress nonsense mutations by inserting an amino acid in the polypeptide at the mutant termination codon.

## 17.2 The Rules of the Code

The genetic code words—the codons—must follow specific rules for protein synthesis. Even after the discovery of tRNAs, several experiments were required to determine the rules of the genetic code. These classic studies addressed whether codons are read sequentially,

or are overlapping, or have gaps (punctuation) between them, and they looked for confirmation of a triplet codon. Investigators also asked in what direction protein synthesis occurs. As we'll see, these "rules of the code" were addressed by elegant experiments performed even before the code words themselves were understood.

### The Genetic Code Is Nonoverlapping

The triplet codons in mRNA could overlap with one another or could be nonoverlapping. In a nonoverlapping code, each codon would be read as an independent unit; a single-nucleotide substitution in the mRNA would change only one codon, and the mutant protein would have only one amino acid change (Figure 17-11a). In a triplet code with maximal overlap, each codon would share two nucleotides with two other codons; a single-nucleotide change in the mRNA would alter three codons, and the resulting protein would contain three consecutive amino acid changes (Figure 17-11b).

A code with overlapping codons was ruled out experimentally by studies of mutant proteins. Most

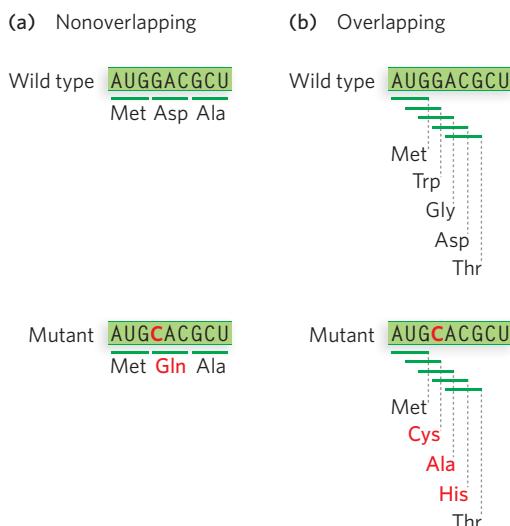
mutations result from single-base substitutions, and therefore in a nonoverlapping code, one amino acid would be changed, whereas in a maximally overlapping code, three consecutive amino acids would be changed. Independent studies of mutants of three different proteins—hemoglobin, tobacco mosaic virus protein, and tryptophan synthetase—demonstrated that the code must be nonoverlapping. In a combined total of nearly 100 different mutants of these proteins, almost all of the mutants had only one amino acid change. Thus, the genetic code is nonoverlapping, and any exceptions in the findings for mutant proteins are probably the result of double mutations.

### There Are No Gaps in the Genetic Code

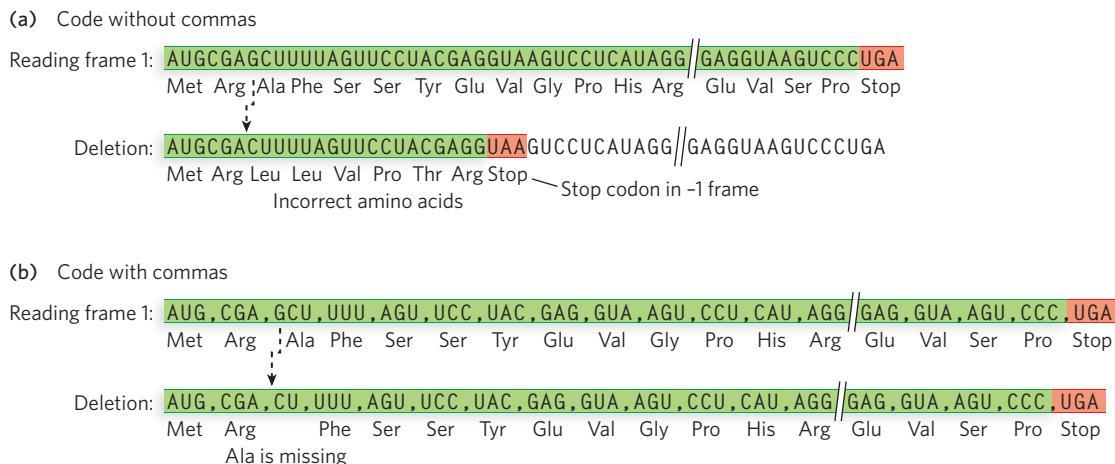
The codons in mRNA could be arranged one after the other, with no separation, or they could be separated by one or more nucleotides acting as punctuation, like a comma. Commas between codons would prevent the deleterious effects of frameshift mutations that result from deletions or insertions of nucleotides. If there were no commas to set codons apart, a frameshift mutation would throw off the entire reading frame (Figure 17-12a). If codons were separated by a non-coding nucleotide, a frameshift mutation would result in only one amino acid change, because the next comma after the altered codon would signal the ribosome to get back on track in the correct reading frame (Figure 17-12b).

Francis Crick and Sydney Brenner performed a series of ingenious experiments in the early 1960s to determine whether the genetic code contains punctuation between codons. They studied the *B* gene of the T4 bacteriophage, which encodes a protein needed for the phage to grow on two different strains of *E. coli*. Severe mutations in the *B* gene restrict phage growth to one *E. coli* strain, but minor alterations produced near one end of the *B* gene are tolerated and preserve the dual-host-range phenotype. Crick and Brenner introduced mutations into the *B* gene chemically, using acridines as mutagens. These planar molecules intercalate into DNA and usually produce mutations through the insertion or deletion of a single base pair. The acridines were used at a concentration low enough to cause an average of one mutation per phage. The researchers tested the effects of these mutations by assaying for plaques (i.e., phage growth) on the two different host strains of *E. coli*.

Most mutations completely inactivated the *B* gene, even when the mutations mapped to the region of the gene that can tolerate minor alterations. These results suggested that not one, but many amino acids were



**FIGURE 17-11 Mutation effects on nonoverlapping and overlapping codes.** (a) In a nonoverlapping code, codons in the mRNA do not share nucleotides, so a single-nucleotide mutation alters only one codon, and the resulting protein has a single amino acid change. (b) In an overlapping code, some nucleotides are shared by several codons. In a triplet code with maximum overlap, a nucleotide can be shared by three codons, so a single-nucleotide mutation results in three codon changes and thus three amino acid alterations in the protein. The genetic code of all living systems is now known to be nonoverlapping.



**FIGURE 17-12** The effect of deletion mutations in codes without and with commas. (a) In a code without commas, a single-nucleotide deletion throws off the reading frame by one nucleotide. All amino acids added after this point are different from those in the wild-type protein. (b) In a code

with commas separating the codons, a deletion should have minimal impact on the protein product because the next comma encountered after the deleted nucleotide would reset the ribosome to the correct reading frame. The genetic code of all living systems is now known to be without commas.

changed by the acridine-induced mutations: the mutations seemed to throw off the entire reading frame, indicating that the code has no commas to bring the ribosome back on track.

To study these mutations further, Crick and Brenner recombined the *B* genes from two mutant T4 phages by mixing them in a host *E. coli* culture, creating phages with a *B* gene containing two mutations instead of one. In some cases, the double-mutated gene restored the dual host range of the wild-type phage, but other pairs of mutations did not restore the dual-range phenotype. Thus, the single-mutant phages used for the recombination could be sorted into two groups, referred to as a plus or a minus group. Double mutants made by crossing two single mutants of like sign could not form a wild-type double mutant, but crossing two single mutants of opposite sign formed double mutants with the dual-host-range phenotype.

Crick and Brenner interpreted their findings as follows. The gain or loss of one base pair results in a shift in the wild-type reading frame (Figure 17-13a), thus changing all the amino acids after the point of mutation (Figure 17-13b, c). When mutants of opposite sign are combined (an insertion and a deletion), the first mutation encountered during translation causes a frameshift, but the second mutation of opposite sign restores the original reading frame (Figure 17-13d). Therefore, the only amino acid alterations that occur are located between the two mutations. Mutants of like sign do not complement one another, because they do

not reestablish the correct reading frame. Overall, these results implied that codons are read in a reading frame without commas. There are no gaps between words in the genetic code.

### The Genetic Code Is Read in Triplets

The first indication that the code is read in groups of three nucleotides came from an extension of the acridine-induced frameshift mutant studies in T4 phage. This time, Crick and Leslie Barnett combined three different mutants. They predicted that if the codons were read in sets of three nucleotides, the crossing of three *B*-gene mutants of T4, all with the same sign, would reestablish the reading frame and produce an active *B*-gene product (Figure 17-13e). Indeed, crossing three *B*-gene mutants of like sign (by mixing in an *E. coli* host culture) restored wild-type, or near wild-type, activity. The results suggested that the original reading frame was restored by either the deletion or the insertion of three nucleotides.



Leslie Barnett, 1920–2002 (left); Francis Crick, 1916–2004 (center); Sydney Brenner (right) [Source: Courtesy of MRC Laboratory of Molecular Biology.]

This was the first evidence that codons consist of three nucleotides. However, this interpretation is based on the assumption that each of the mutations in the three *B*-gene mutants was indeed

(a) Wild type	
DNA:	ATGCTCCGATAATCGTATGGCAGGAG
Protein:	Met Leu Pro Ile Phe Val Ser Asp Glu
(b) Insertion	
	ATG <b>G</b> CTCCGATAATCGTATGGCAGGAG Met Ala Pro Asp Asn Arg Met Ala Gly
(c) Deletion	
	(T) ATGCTCCGATAATCGTAGGCAGGAG Met Leu Pro Ile Phe Val Gly Arg
(d) Insertion plus deletion	
	(T) ATG <b>G</b> CTCCGATAATCGTAGGCAGGAG Met Ala Pro Asp Asn Arg Arg Asp Glu
(e) Triple insertion	
	ATG <b>GGG</b> CTCCGATAATCGTATGGCAGGAG Met Gly Leu Pro Ile Phe Val Ser Asp Glu

**FIGURE 17-13** The effects on a reading frame of combining insertion and deletion mutations. Insertion or deletion of a single nucleotide throws the ribosome into the wrong reading frame and produces a mutant protein (amino acids shown in red). (a) The wild-type protein sequence. (b) The effect of an insertion mutation. (c) The effect of a deletion mutation. (d) Combining an insertion and a deletion affects some amino acids but eventually restores the correct sequence. (e) Combining three consecutive insertion mutations (or three deletions) leaves the remaining triplets intact—evidence that a codon has three, rather than four or five, nucleotides.

a single-nucleotide insertion (or deletion). Without direct sequence information, it remained possible that more than one base pair was added (or removed) in one or more of the mutants. Of course, we now know that codons really do consist of three nucleotides apiece, and that the genetic code is read in triplets.

### Protein Synthesis Is Linear

There are many possible ways in which a chain of amino acids could be assembled on an mRNA transcript. For example, chain growth could initiate at one end, either the N-terminus or the C-terminus, or it could start in the middle and grow outward in both directions. In fact, the chain could even be synthesized in random segments that were then stitched together to form the final product. In 1961, Howard Dintzis performed elegant studies on the synthesis of hemoglobin, using extracts from rabbit reticulocytes (immature red blood cells), and demonstrated that

protein synthesis proceeds linearly, from the N- to the C-terminus (see How We Know).

The triplet sequences of codons in mRNA could be read in either direction. Given that the direction of protein synthesis proceeds from the N-terminus to the C-terminus, determining the direction in which codons are translated becomes a relatively simple matter. Chemical methods of synthesizing RNAs of defined sequence had already been developed to crack the genetic code. For example, a synthetic hexanucleotide RNA should direct the synthesis of two amino acids. Thus, an RNA of sequence 5'-AAAUUU would encode Lys-Phe if the codons were read in the 5'→3' direction, or Phe-Lys if read in the 3'→5' direction (recall that peptides are always written in the N- to C-terminal direction). Experiments demonstrated that codons are read in the 5'→3' direction during translation.

### SECTION 17.2 SUMMARY

- Single-nucleotide changes result in changes in a single amino acid residue in the protein product, demonstrating a nonoverlapping genetic code. Each codon is an independent unit, coding for a single amino acid.
- The genetic code has no commas. Single-nucleotide insertion or deletion mutations result in a complete loss of activity in the mutant gene; these mutations throw off the reading frame from the point of mutation onward, because the code lacks any signal (comma) to reset the reading frame. Double mutations with an insertion and a deletion restore the reading frame.
- Codons are composed of triplet sequences. Triple-insertion mutations and triple-deletion mutations result in active protein.
- Protein synthesis is linear; it proceeds from the N-terminus to the C-terminus, and mRNA is read in the 5'→3' direction.

### 17.3 Cracking the Code

Cracking the genetic code was one of the most significant scientific milestones of the last half of the twentieth century. Today, it would be a simple matter of comparing the sequences of nucleic acids and their corresponding proteins. But in the early 1960s, nucleic acid sequencing had not yet been invented, and although protein sequencing methods were firmly established, the process was laborious. Given the state of sequencing methodology, it is surprising that new and ingenious ways of studying and deciphering the code in its entirety were developed at all.

We briefly review the major experimental strategies that enabled investigators to crack the code.

### Random Synthetic RNA Polymers Direct Protein Synthesis in Cell Extracts

The preparation of cell extracts that could translate mRNA into protein was essential to the method of cracking the code. First, Marshall Nirenberg and Heinrich Matthaei made a simple but remarkable discovery that set into motion the cracking of the genetic code: they could use the enzyme polynucleotide phosphorylase to synthesize RNA templates that would code for protein polymers in cell extracts of *E. coli*. Then, the endogenous mRNA in the cell extracts was removed, to allow the translation machinery to work on the synthetic RNA templates. To do this, the extracts were pre-incubated so that the endogenous ribonuclease (RNase) activity would destroy all existing mRNA, and deoxyribonuclease (DNase) was added to the extracts to prevent further mRNA production. With these treatments, protein synthesis in the cell extracts (or “translation extracts”) was dependent on the addition of the exogenous synthetic RNA.

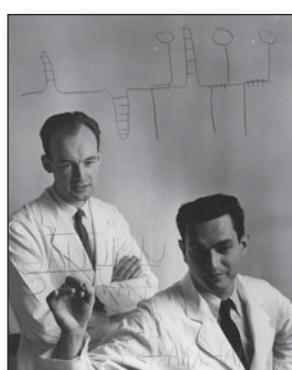
Polynucleotide phosphorylase does not require a template; it uses ribonucleoside diphosphates (NDPs) to make random polymers of RNA:



The intracellular role of polynucleotide phosphorylase is to catalyze the reverse reaction to degrade RNA, using inorganic phosphate to yield NDPs. In vitro, the enzyme can be made to synthesize RNA by the addition of excess NDPs, which at sufficient concentration are polymerized. However, because polynucleotide phosphorylase is not template-directed, the sequence of the RNA polymer is random.

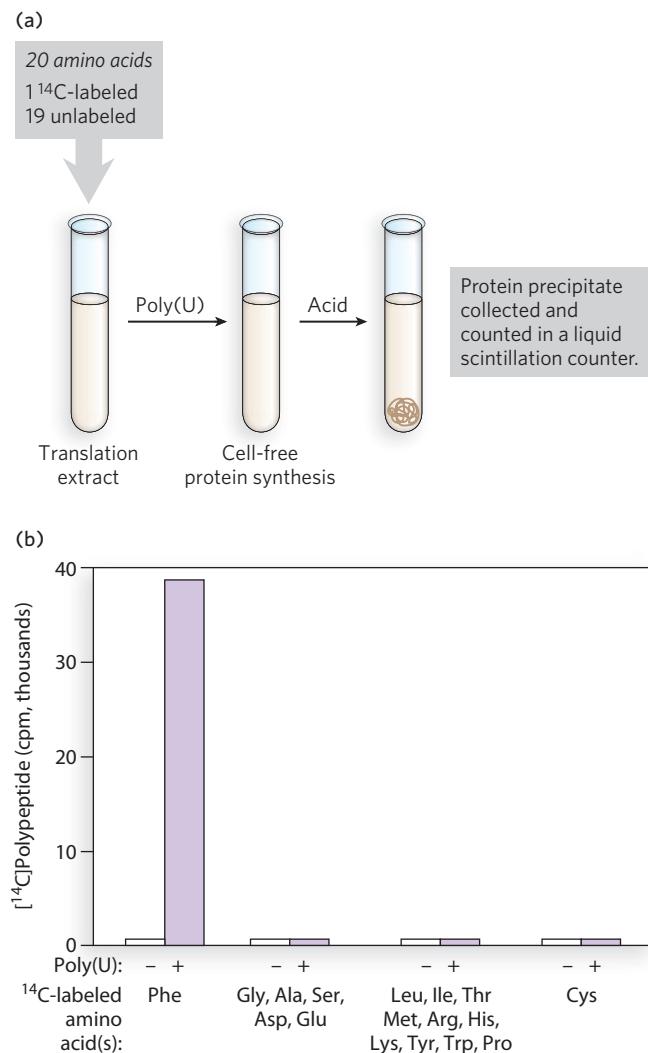
Using UDP as the only ribonucleoside diphosphate substrate, polynucleotide phosphorylase catalyzes the synthesis of poly(U) RNA. To determine what type of protein synthesis poly(U) directs, Nirenberg and Matthaei added poly(U) to 20 reaction mixtures with the *E. coli* cell

extracts and all 20 amino acids, differing only in which amino acid was radioac-



**Heinrich Matthaei (left); Marshall Nirenberg, 1927–2010 (right)** [Source: Courtesy of Marshall W. Nirenberg/The National Library of Medicine.]

tively labeled (Figure 17-14). At the end of the incubation, the reaction mixtures were treated with acid, which precipitates protein polymers but leaves free amino acids in solution. Precipitates were collected and counted in a liquid scintillation counter, which measures radioactivity in counts per minute (cpm). The results showed that poly(U) directed the synthesis of polyphenylalanine, poly(Phe), and therefore the codon for phenylalanine must be UUU. The other homopolymers were synthesized by similar methods. Poly(A) specified the synthesis of poly(Lys), identifying AAA as a codon for lysine.



**FIGURE 17-14** Poly(U)-directed synthesis of poly(Phe).

(a) Twenty different  $^{14}\text{C}$ -labeled amino acids were added, individually or in mixtures, with the rest of the amino acids in unlabeled form, to cell extracts that could synthesize protein in the presence of an added poly(U) template. (b) Only the extracts containing  $[^{14}\text{C}]$ phenylalanine produced a significant amount of radioactive polypeptide in the presence of poly(U). [Source: Adapted from M. Nirenberg and P. Leder, *Science* 145:1399, 1964.]

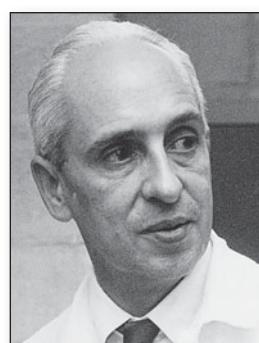
Poly(C) resulted in the production of poly(Pro), and thus CCC is a codon for proline. Unfortunately, poly(G) forms intramolecular hydrogen-bonded structures that prevent its use in these reactions.

The use of polynucleotide phosphorylase was extended to copolymers, RNAs containing more than one type of nucleotide. As an example of this type of analysis, Table 17-2 shows the results obtained by Severo Ochoa's group, using ADP and CDP in a 5:1 molar ratio. Polynucleotide phosphorylase incorporates NDPs into RNA randomly, so these conditions produced RNA with five times more A than C. All possible codons containing A and C were generated in the random RNA copolymer, in the following distribution (setting AAA at 100): AAA (100), A<sub>2</sub>C (60), AC<sub>2</sub> (12), and CCC (0.8). (Note the use of A<sub>2</sub>C and AC<sub>2</sub> to indicate that the *order* of nucleotides is unknown.) Use of this heterogeneous RNA copolymer in translation extracts directed the synthesis of polypeptides containing asparagine, glutamine, histidine, lysine, proline, and threonine. From the relative amounts of these amino acids in the precipitated protein product, the investigators could deduce the respective codon compositions specifying each amino acid. No amino acid was observed at the low frequency expected of a CCC codon (0.8) relative to an AAA codon (100). But proline was observed to be incorporated at a relative frequency of 4.7, which is close to the expected frequency for two codons, CCC and AC<sub>2</sub>.

The researchers could infer from this result that proline is specified by two codons having the compositions AC<sub>2</sub> and CCC, reflecting the degenerate nature of the genetic code. The result was supported by experiments using poly(C), which identified CCC as a codon specifying proline. The other codon, AC<sub>2</sub>, could have

the sequence ACC, CCA, or CAC. Because polynucleotide phosphorylase polymerizes NDPs randomly, one can assign only codon compositions from these results, not codon sequences (except, of course, for codons with only one type of nucleotide).

Numerous experiments were performed using random RNA copolymers, and in this way the nucleotide composition of about 40 codons could be assigned to particular amino acids. However, to identify the nucleotide *sequence* of codons, RNA molecules of defined sequence were required.



**Severo Ochoa, 1905–1993** [Source: AP/Wide World Photos.]

## RNA Polymers of Defined Sequence Complete the Code

Chemical synthesis of short nucleic acids of specific sequence was needed to define the complete genetic code. In 1964, Nirenberg and his coworkers at the National Institutes of Health discovered a novel method to identify an amino acid associated with a short synthetic codon. They found that during protein synthesis in *E. coli* (using translation extracts and <sup>14</sup>C-labeled amino acids), the nascent protein (the protein being synthesized) stayed attached to ribosomes bound to the mRNA and could be separated from unbound amino acids (Figure 17-15a). However, the [<sup>14</sup>C]aminoacyl-tRNA-ribosome-RNA codon complex could not be precipitated by acid treatment, because the complex is dissociated by acid. The researchers

**Table 17-2 Incorporation of Amino Acids into Polypeptides in Response to Random Polymers of RNA**

Amino Acid	Observed Frequency of Incorporation (Lys = 100)	Tentative Assignment for Nucleotide Composition of Corresponding Codon*	Expected Frequency of Incorporation Based on Assignment (Lys = 100)
Asparagine	24	A <sub>2</sub> C	20
Glutamine	24	A <sub>2</sub> C	20
Histidine	6	AC <sub>2</sub>	4
Lysine	100	AAA	100
Proline	4.7	AC <sub>2</sub> , CCC	4.8
Threonine	26	A <sub>2</sub> C, AC <sub>2</sub>	24

\*These designations contain no information on nucleotide sequence (except, of course, AAA and CCC).

developed a filtration method that took advantage of the fact that a nitrocellulose filter binds protein but not RNA (including tRNA, charged or uncharged). More importantly, they found that even RNA polymers as short as a trinucleotide could be used in these experiments. Maxine Singer, also working at the NIH, had developed the capability of synthesizing RNAs of

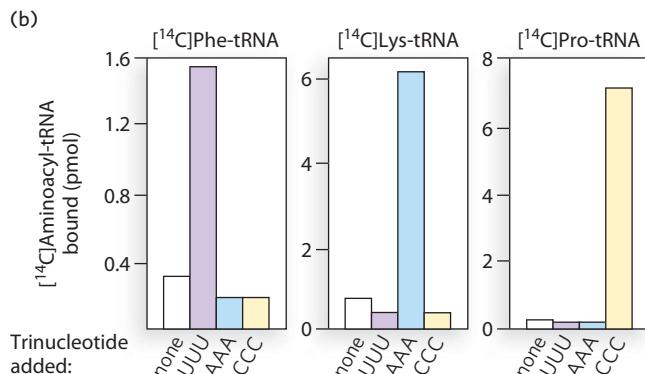
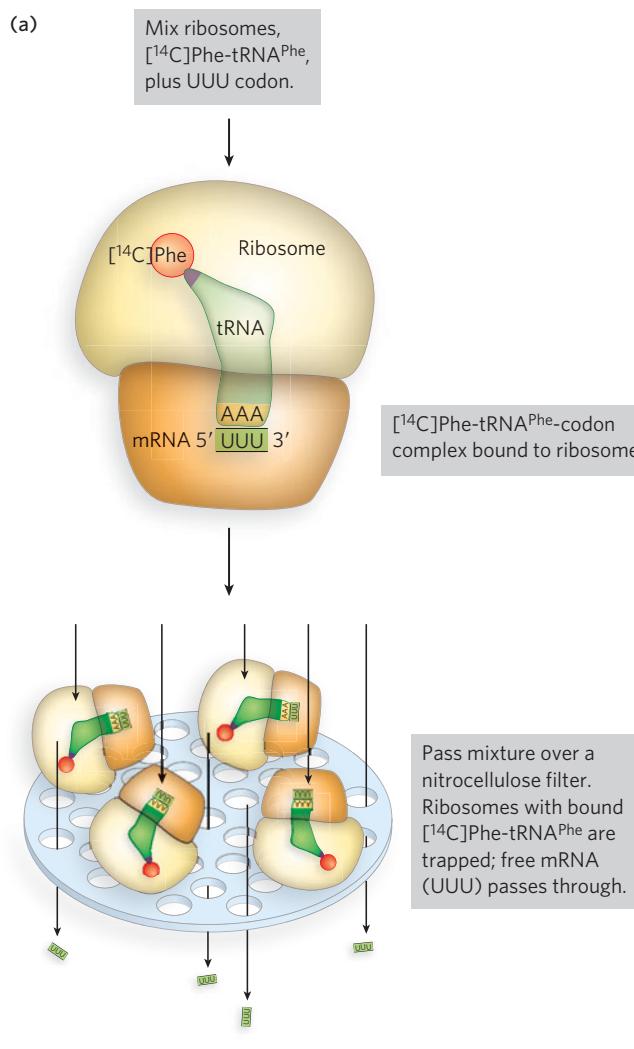
defined sequence. This provided an advance that allowed assignment of several additional codons for amino acids and confirmed assignments made by the precipitation method.



**Maxine Singer** [Source: NIH Photo.]

Results from Nirenberg's initial study, using three different [<sup>14</sup>C]aminoacyl-tRNAs and the synthetic trinucleotides UUU, CCC, or AAA, are shown in Figure 17-15b. Similar studies of all possible triplet RNA sequences identified approximately 50 codon assignments for specific amino acids. However, not all triplet RNA sequences induced tight binding of their corresponding [<sup>14</sup>C]aminoacyl-tRNA to the ribosome, and another technique was required to complete the cracking of the genetic code.

Another breakthrough came from H. Gobind Khorana, who developed chemical methods to synthesize short polyribonucleotides with defined sequences of two, three, or four nucleotide repeats. These short RNAs were then amplified by polymerases to produce long polymers of defined repeating sequence. Results obtained with these long RNA polymers, combined with the earlier results from trinucleotide-induced ribosome binding and the use of random RNA polymers, identified the amino acids specified by all of the codons. For example, the alternating copolymer (AC)<sub>n</sub> contains the alternating codons ACA and CAC. The polypeptide



**H. Gobind Khorana**  
[Source: Courtesy of Archives, University of Wisconsin-Madison.]

**FIGURE 17-15** Use of trinucleotide sequences as mRNA.

(a) Twenty [<sup>14</sup>C]aminoacyl-tRNAs were formed in separate reactions, using the 20 different <sup>14</sup>C-labeled amino acids, and each was added to a preparation of ribosomes, along with a three-nucleotide RNA of defined sequence (a codon). Individual reaction mixtures were then passed through a nitrocellulose filter that binds protein, trapping the [<sup>14</sup>C]aminoacyl-tRNA bound to the triplet RNA codon on the ribosome. The reaction shown here used [<sup>14</sup>C]Phe-tRNA<sup>Phe</sup> and the UUU codon. (b) Results obtained with several codons and [<sup>14</sup>C]aminoacyl-tRNAs. [Source: Adapted from M. Nirenberg and P. Leder, *Science* 145:1399–1407, 1964.]

synthesized on this RNA copolymer contains equal amounts of threonine and histidine. Parallel studies using random RNA copolymers identified the nucleotide composition of a histidine codon as AC<sub>2</sub>, so CAC must code for histidine and ACA for threonine. Examples of results obtained with di-, tri-, and tetranucleotide sequences are shown in Table 17-3.

Repeating trinucleotide sequences yield three types of homopolymeric peptides. For example, for a repeating sequence of UUCs, the mRNA triplets can be read as repeating UUC, or UCU, or CUU codons. Poly(UUC) produced poly(Phe), poly(Lys), and poly(Leu). The ribosome-binding assay showed that UCU encodes serine and CUU encodes leucine. Therefore, the UUC codon can be assigned to phenylalanine. Poly(UAG) yielded only two homopolymers, poly(Ser) and poly(Val), because the UAG reading frame is a string of nonsense

codons that yields no product. The repeating tetranucleotide poly(UAUC) produced a repeating tetrapeptide of Tyr-Leu-Ser-Ile. Knowing the codon assignments for the four amino acids in the repeating tetranucleotide also confirmed that mRNA is read in the 5'→3' direction. If the RNA had been read in the 3'→5' direction, the repeating tetrapeptide would have been Ile-Ser-Leu-Tyr.

Consolidation of the results from many different experiments permitted the assignment of 61 of the 64 possible codons. The other three were identified as termination codons, in part because of the results obtained with synthetic RNAs in which the codons disrupted the amino acid coding patterns. Meanings for all the triplet codons were firmly established by 1966, and they have been verified numerous times, in many different ways. In retrospect, the experiments of Nirenberg and Khorana that identified codons in translation extracts

**Table 17-3 | Polypeptides Produced by Synthetic RNAs of Defined, Repeating Sequence**

Repeating Dinucleotide	Codons	Copolymer
(AC) <sub>n</sub>	ACA-CAC-ACA-CAC-ACA	(Thr-His) <sub>n</sub>
(UC) <sub>n</sub>	UCU-CUC-UCU-CUC-UCU	(Ser-Leu) <sub>n</sub>
Repeating Trinucleotide	Codons	Homopolymer
(UUC) <sub>n</sub>	UUC-UUC-UUC-UUC-UUC UCU-UCU-UCU-UCU-UCU CUU-CUU-CUU-CUU-CUU	Poly(Phe) Poly(Ser) Poly(Leu)
(AAG) <sub>n</sub>	AAG-AAG-AAG-AAG-AAG AGA-AGA-AGA-AGA-AGA GAA-GAA-GAA-GAA-GAA	Poly(Lys) Poly(Arg) Poly(Glu)
(UAG) <sub>n</sub>	UAG-UAG-UAG-UAG-UAG AGU-AGU-AGU-AGU-AGU GUA-GUA-GUA-GUA-GUA	None Poly(Ser) Poly(Val)
Repeating Tetranucleotide	Codons	Repeating copolymer
(UAUC) <sub>n</sub>	UAU-CUA-UCU-AUC AUC-UAU-CUA-UCU UCU-AUC-UAU-CUA CUA-UCU-AUC-UCU	(Tyr-Leu-Ser-Ile) <sub>n</sub> (Tyr-Leu-Ser-Ile) <sub>n</sub> (Tyr-Leu-Ser-Ile) <sub>n</sub> (Tyr-Leu-Ser-Ile) <sub>n</sub>
(GUAA) <sub>n</sub>	GUU-AGU-AAG-UAA-GUA UAA-GUA-AGU-AAG-UAA AAG-UAA-GUA-AGU-AAG AGU-AAG-UAA-GUA-AGU	Val-Ser-Lys-(stop) Val-Ser-Lys-(stop) Val-Ser-Lys-(stop) Val-Ser-Lys-(stop)

should not have worked in the absence of initiation codons. Serendipitously, with the high magnesium concentration used in their in vitro experiments, the normal initiation requirement for protein synthesis was relaxed.

A striking and unexpected finding was the universality of the genetic code in all life forms. Similar experiments using other types of cells gave the same results! As there is no chemical relationship between an amino acid and its codon(s), we might have expected the genetic code to differ among organisms that do not share a common evolutionary lineage. The universality of the code reveals that all organisms are related; they must have evolved from a common ancestor in which the code was already fully developed.

### The Genetic Code Is Validated in Living Cells

One might question the validity of a genetic code determined entirely by in vitro experiments. We now know the sequences of entire genomes of organisms ranging from bacteria to humans, and this enormous body of information confirms that the genetic code determined in the translation extract experiments is indeed interpreted in the same way in living organisms. Even in the 1960s, however, there was evidence that the genetic code was the same in cells as in cell extracts.

By the mid-1960s, more than 100 mutant proteins resulting from single-nucleotide substitutions had been studied, and in all cases the incorrect amino acid in the mutant, relative to the wild type, could be accounted for by a change in a single nucleotide, based on the genetic code. The disease sickle-cell anemia was known to be caused by a single amino acid change, substituting glutamate for valine in human hemoglobin (see Highlight 2-1). Only one nucleotide change is needed to alter the AGU codon specifying glutamate to the UGU codon for valine: these codons specify the same amino acids in extracts of *E. coli* as they do in blood cells. An astute demonstration that the genetic code in the cell is the same as the code determined in cell extracts is described in How We Know.

### SECTION 17.3 SUMMARY

- Researchers cracked the genetic code in experiments that used radioactively labeled amino acids and cell extracts that translate synthetic RNA templates.
- The compositions of many codons specifying amino acids were assigned on the basis of experiments using random RNA polymers synthesized by polynucleotide phosphorylase.

- Two techniques that used RNAs of defined sequence completed the codon table. An assay that induced the binding of [<sup>14</sup>C]aminoacyl-tRNAs and their cognate synthetic trinucleotide RNAs to ribosomes allowed the identification of most of the codons. With the availability of long, synthetic RNA polymers of defined repeating sequence, researchers used the in vitro protein synthesis assay to assign the remaining codons.
- Studies of amino acid replacements in mutant proteins confirmed that the genetic code in living cells is the same as that determined in cell extracts.

## 17.4 Exceptions Proving the Rules

Initial studies of the genetic code suggested that it was universal and without variation. And, as we've seen, this universality implies that life evolved from a common ancestor. Presumably, once the code had developed in LUCA it became locked in place because descendants could not tolerate changes in the genetic code. For example, a change in a codon that specifies lysine to one that specifies leucine would result in a Leu residue replacing every Lys residue in every protein in the cell. Clearly, such a global change would be fatal to the cell, so it is unlikely that such codon changes could occur, even over a long span of evolutionary time. How the translation machinery evolved in LUCA remains one of the greatest questions in evolution, and with no "missing link" cells, we may never know the answer. Even though we don't know how the process of translation evolved, it is interesting to contemplate the evolutionary hurdles that must have been overcome to arrive at this process.

We now know that there *are* some exceptions to the rules of the genetic code. It is not entirely universal after all. Does this mean that the ancestral cell did not perfect the code before modern cells diverged from it? The evidence suggests not. Most of the exceptions support the evolution of a common code in LUCA from which a few changes, in some circumstances, evolved.

### Evolution of the Translation Machinery Is a Mystery

With the discovery of ribozymes, the hypothesis of the RNA world became a very plausible model for the beginning of life. In the RNA world, RNA catalyzes essentially all the chemical reactions needed for life, and there are many examples of catalytic RNAs in cells today—including the ribosome—all of which are

possible vestiges of the RNA world. RNA can also catalyze its own replication. But how do we get from an RNA world to LUCA—a cell with a membrane, DNA for storage, mRNA, ribosomes, tRNA, and protein-based catalysis? Possible ways in which a cell membrane developed are discussed in Chapter 1. For the nucleic acids, we know that RNA is less stable than DNA. Despite Steve Benner's finding that borate can stabilize RNA and was probably plentiful in the prebiotic soup (see Moment of Discovery), it is not hard to imagine the development of DNA as a more stable information storage molecule. But how can we explain the evolution of translation and the genetic code for protein synthesis?

Any hypotheses about how translation evolved must account for each part of the translation machinery. At least 20 tRNAs, 20 aminoacyl-tRNA synthetases, a coding RNA, and the entire ribosome machinery, with its numerous proteins and RNA components, are required to translate the genetic code. These components could not have evolved all at once, so a reasonable hypothesis must either reduce the complexity of the translation process or break it down into individual steps. The hypothesis must also solve the chicken-and-egg problem of how a method of protein synthesis could evolve when the protein components (aminoacyl-tRNA synthetases and ribosomal proteins) could not be synthesized in an RNA world. Moreover, the process was probably not accurate in the beginning; accuracy most likely required evolutionary honing. If accuracy were not required initially, catalytic function is essentially ruled out as the initial role of early proteins. But there must have been an initial benefit to the cell on which natural selection could act. Plausible hypotheses must consider the forces of natural selection that would foster the evolution of proteins before their catalytic role was realized.

Finally, how did the genetic code evolve? Was it a random act of evolution, or did the amino acids somehow participate in generation of the code as we know it today? Did all 20 tRNAs appear individually, or did one appear through random mutation and then diversify into the rest? Did the first code use triplet sequences, or was it simpler, using dinucleotide sequences to code for fewer amino acids? How was a reading frame established? And how did the point of termination develop?

These are some of the challenges to hypothesizing how the translation process evolved. One hypothesis is presented in Highlight 17-1. All cells have the complete translation machinery; there are no cells with “missing links” to tell us the story more directly. However, certain scientific approaches can help

illuminate how the process evolved. For example, there are exceptions to the genetic code, and we can examine them for insights about its evolution. We also know that RNAs can perform many enzymatic reactions, supporting the RNA world hypothesis. Ongoing research is defining the minimal genetic and protein requirements for a living cell, which will identify the essential genetic and protein requirements for life—components that are likely to have been present in LUCA.

## Mitochondrial tRNAs Deviate from the Universal Genetic Code

Phylogeny tells us that the exceptions to the genetic code are derived from a single, universal code, because the exceptions are rare and occur in different branches of the tree of life (see Chapter 8). But in what situations could changes to the genetic code be viable? Even one change would have a global impact on cellular function, because every protein in the cell is made according to the same set of coding instructions. In other words, a single change in a codon–amino acid relationship would cause changes in all proteins encoded by a gene containing that codon.

With this in mind, we can make a few hypotheses about the types of deviations from the genetic code that might be plausible. For example, we can propose that the most easily altered codons are termination codons, which are not located in the middle of genes. If a stop codon were recruited to code for an amino acid by altering the anticodon of a tRNA, that codon could be placed in internal positions of certain genes as the organism evolved (another stop codon could be used for termination). In this way, a particular amino acid would be inserted in the middle of the gene where the stop codon is placed. We can also hypothesize that evolution of this type of exception to the code would have a higher probability in an organism of low genetic complexity (i.e., only a few genes) than in an organism of high complexity, because fewer proteins would be affected. These two hypotheses are largely confirmed.

Mitochondria are a prime example of how the genetic code can be altered. Mitochondria are thought to be the descendants of early bacterial cells that were engulfed by eukaryotic cells and proved beneficial for their unique capacity to perform aerobic metabolism, thereby conferring this advantageous capability on early anaerobic eukaryotes. Over time, this symbiotic relationship relieved the mitochondrial genome of most of its genes, transferring them to the nucleus of the host cell. But mitochondria retained a small

genome of their own, with a limited set of genes. Researchers noticed the first deviations from the genetic code when sequencing mitochondrial DNA (mtDNA). A fascinating aspect of the mitochondrial genome is that it encodes a unique set of tRNAs, just for use in decoding the mtDNA. This feature of mitochondria permits changes in their tRNAs without interfering with the information flow of the cellular genome. As predicted for alterations evolving from the standard code, the most common codon changes in mitochondria involve stop codons.

Genetic code changes in mitochondria are essentially the result of an exquisitely streamlined flow of genetic information. Vertebrate mtDNAs encode 13 proteins, 2 rRNAs, and 22 tRNAs. Instead of the minimum of 32 tRNAs needed for the standard, cellular code, the 22 mitochondrial tRNAs can decipher all possible codons by slight alterations in the rules of the code. For example, only one tRNA is used for each of four codon families. In each case, a single tRNA recognizes the four different codons, each with the same first two nucleotides. Each of these mitochondrial tRNAs has a U in the first (wobble) position of the anticodon (that base-pairs with position 3 of the codon). Using normal base-pairing rules, two tRNAs are needed to decode a codon family, one with a U in the wobble position (pairs with G or A) and one with a G (pairs with U or C). Therefore, the U in these mitochondrial tRNAs is not used to distinguish codons, and base pairing to the first two nucleotides of the codon specifies which amino acid is incorporated.

Other mitochondrial tRNAs function with codons that contain either A or G in the third position, or either U or C, so virtually all the tRNAs recognize two or four codons.

If all mitochondrial tRNAs recognize more than one codon, yet another deviation from the rules of the standard genetic code can be inferred. Normally, tryptophan and methionine are specified by one codon each. In mitochondria, the tRNA specifying tryptophan recognizes the UGG codon, but it also recognizes UGA, which is a termination codon in the standard code. The AUG codon for methionine is used in mitochondria to initiate translation, but the standard isoleucine codon, AUA, specifies methionine at internal positions. In the mitochondria of mammals, codons AGA and AGG, which usually specify arginine, are termination codons. These same mitochondrial codons in the fruit fly specify serine. The known coding variations in mitochondria are summarized in Table 17-4.

The low complexity of the mitochondrial genome has allowed continued evolution, which has resulted in streamlining of the genetic code. However, there are also a few examples in which the code has been altered in free-living cells. The only bacterial variant is the use of the UGA stop codon to encode tryptophan in *Mycoplasma capricolum*. Among eukaryotes, a few species of ciliated protists use the codons UAA and UAG (as generally used only by mitochondria) to specify glutamine. The most perplexing change in the genetic code is found in the yeast *Candida albicans*

**Table 17-4 Known Variant Codon Assignments in Mitochondria**

Organisms	UGA (Stop)	AUA (Ile)	AGA, AGG (Arg)	CUN (Leu)	CGG (Arg)
Animals					
Vertebrates	Trp	Met	(Stop)	+	+
<i>Drosophila</i>	Trp	Met	Ser	+	+
Yeasts					
<i>Saccharomyces cerevisiae</i>	Trp	Met	+	Thr	+
<i>Schizosaccharomyces pombe</i>	Trp	+	+	+	+
Filamentous fungi	Trp	+	+	+	+
Trypanosomes	Trp	+	+	+	+
Higher plants	+	+	+	+	Trp

Note: Only deviations from the standard code assignment are shown; + indicates no deviation. The standard, non-mitochondrial assignments for the codons are shown in parentheses. N = any nucleotide.

## HIGHLIGHT 17-1 EVOLUTION

### The Translation Machinery

How the ribosome, tRNAs, and the genetic code evolved is one of the most perplexing and fascinating areas of evolutionary history. In the RNA world, nucleic acids not only stored genetic information but also performed all the catalytic reactions necessary for life. In modern cells, most catalysts are proteins, which are more efficient catalysts than RNAs. The leap from nucleic acid to protein requires very complicated machinery, and it can't have happened all at once. It presumably evolved in steps. Furthermore, the ancestral cell, LUCA, could not predict that proteins would be superior catalysts to RNAs, so we can presume that the first proteins were made to serve some other purpose. What were the evolutionary steps leading to the translation apparatus, and how did natural selection produce them? As an intellectual exercise, let's consider just one of several possible explanations.

First of all, why would a cell want a protein when it is already using RNA to catalyze cellular metabolism? Harry Noller suggests that the first proteins evolved to help RNA ribozymes fold properly. For example, the high negative charge of the nucleic acid backbone hinders the close approach of nucleic acid helices, but the charge can be mitigated by a protein's basic amino acid side chains. Indeed, most ribozymes in modern cells (including ribosomes) contain a protein component, and many ribosomal proteins are at junctions of RNA helices and may stabilize their proximity. Therein lies a possible selection pressure for an RNA-based cell to develop a way to make proteins: proteins can serve a structural role that helps RNA form a more catalytically competent ribozyme.

What about the building blocks of proteins? Several amino acids were most likely present in the

primordial soup, but how were they linked together in a way that is coded by an RNA without aminoacyl-tRNA synthetases (proteins)? Perhaps the anticodon of tRNA was directly involved in specifying the amino acid it carried. This is easy to envision, given what we know about cellular riboswitches. Riboswitches are small sections of RNA, usually 70 to 170 nucleotides embedded in a larger RNA molecule, that fold into complex structures and bind specific cellular metabolites (usually small molecules, including free amino acids) with high affinity (see Chapter 20). Thus, ancestral tRNAs could have selectively bound particular amino acids. However, modern tRNAs are L-shaped, with an anticodon far removed from the 3' terminus that carries the amino acid. This distance precludes the anticodon from participating in amino acid selection. But perhaps early tRNAs simply folded differently, with the 3' amino acid arm near the anticodon, enabling it to select the correct amino acid and charge itself, thus circumventing the need for aminoacyl-tRNA synthetases (Figure 1).

Now we have the rudiments of a simpler RNA world-based process of translation. The self-charged tRNAs align along mRNA through base pairing, and peptide bonds form—probably spontaneously, given the inherent reactivity of charged tRNAs. Over time, natural selection would direct the evolution of a surface (i.e., the ribosome) on which tRNAs could be more efficiently aligned on the mRNA, and this surface eventually would act as a catalyst for the peptidyl transfer reaction.

The early translation process may have operated at very low fidelity, yet it served the purpose of making structural peptides that enhanced RNA catalytic function. For example, early tRNAs may not have

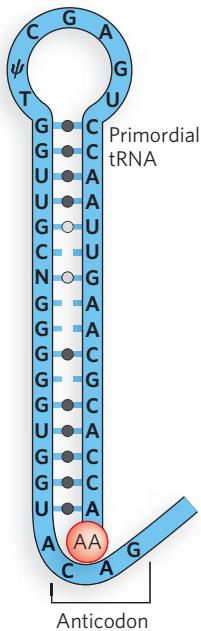
and several related *Candida* species, in which the CUG codon that usually encodes leucine specifies serine instead.

### Initiation and Termination Rules Have Exceptions

Changes in the code need not be absolute; a codon might not always encode the same amino acid. In most organisms we find some examples of amino acids being inserted at positions that are not specified in the standard

code. Two examples are the occasional use of GUG (valine) or UUG (leucine) as an initiation codon. This occurs only for those genes in which the GUG or UUG codon is properly located in the mRNA (see Chapter 18).

Another example is the insertion of selenocysteine (Sec)—sometimes referred to as the twenty-first amino acid, as it is uniquely coded for in all domains of life. When present, selenocysteine is usually found in proteins involved in oxidation-reduction reactions, such as formate dehydrogenase in bacteria and glutathione peroxidase in mammals. These enzymes require the



**FIGURE 1** In this hypothetical primordial tRNA, the anticodon participates directly in amino acid recognition, thus circumventing the need for aminoacyl-tRNA synthetases. [Source: Adapted from J. J. Hopfield, *Proc. Natl. Acad. Sci. USA* 75:4334–4338, 1978.]

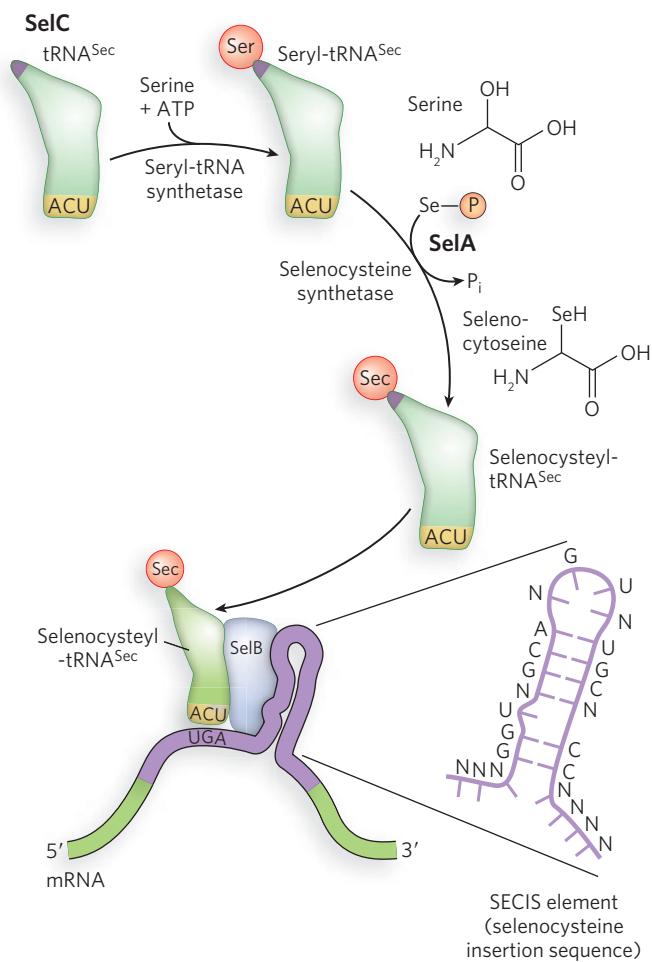
been entirely selective for one amino acid, but may have selected for certain amino acid properties such as charge, polarity, or hydrophobicity. This level of accuracy might be acceptable if the functions of early proteins were purely structural. But even for a structural role, the pressures of natural selection would eventually result in high fidelity in the translation process. Although cells did not “know” that amino acid side chains offered much greater chemical

potential than nucleic acids for both structure and catalysis, they had developed a mechanism to code for protein. And selection could do the rest: the evolution of the protein world would proceed according to the rules of natural selection. Proteins that are better catalysts than their RNA counterparts would eventually take over much of the job of catalytic RNAs, and aminoacyl-tRNA synthetases would evolve.

The evolution of the genetic code is another area of lively debate. Proposals range from a code that simply arose at random and became frozen in time, to one in which the amino acids themselves participated directly in code development. For the latter proposal, consider the hypothetical tRNA structure in Figure 1. The anticodon would participate in selecting the correct amino acid, possibly through favorable contacts between the amino acid and a particular triplet nucleotide sequence, depending on the chemical nature of the amino acid. Indeed, experimental support for this hypothesis exists. The preferential affinity of amino acids for certain nucleic acid sequences has been studied, and it is reported that some amino acids exhibit preferential binding to the anticodons of the tRNAs that encode them. It's also possible that not all 20 amino acids were present early on, and that only a two-nucleotide code was needed for 16 ( $4^2$ ) or fewer amino acids. Some modern amino acids may not have been present in the primordial soup, but instead could have arisen as side products of cellular metabolism. These “cell-invented” amino acids could have been incorporated later, expanding the code to its current triplet form. Regardless of the route taken by natural selection, the genetic code has obviously been honed over the millennia to be robust and resistant to mutations.

element selenium for their activity, generally in the form of a Sec residue in the active site. Although modified amino acids are usually produced by posttranslational reactions, selenocysteine is formed at the level of an aminoacylated tRNA incorporated during translation in response to a UGA (stop) codon (Figure 17-16). A special type of serine-binding tRNA, tRNA<sup>Sec</sup>, present at lower levels than other tRNA<sup>Ser</sup> species, recognizes UGA and no other codons. The tRNA<sup>Sec</sup> is charged with serine, and the serine is enzymatically converted to selenocysteine before it is used at the ribosome. The

charged tRNA does not recognize just any UGA codon; a contextual signal in the mRNA, an element referred to as SECIS (selenocysteine insertion sequence), ensures that this tRNA recognizes only the few UGA codons within certain genes that specify selenocysteine. This process has been extensively studied in *E. coli*. Of the gene products involved, SelC is a tRNA (tRNA<sup>Sec</sup>) that accepts serine initially. SelA is an enzyme that uses selenophosphate to convert the serine to selenocysteine, forming Sec-tRNA<sup>Sec</sup>. SelB is an elongation factor needed by the ribosome to utilize the Sec-tRNA<sup>Sec</sup> and



**FIGURE 17-16** The incorporation of selenocysteine during translation. A serine-charged tRNA (SelC) with the anticodon 5'-UCA recognizes a UGA stop codon. The Ser-tRNA<sup>Sec</sup> is enzymatically converted to selenocysteyl-tRNA<sup>Sec</sup> (Sec-tRNA<sup>Sec</sup>) by SelA, which uses selenophosphate and releases P<sub>i</sub> on attachment of the selenium to the amino acid. SelB is an elongation factor that recognizes a specific hairpin in the mRNA and also binds Ser-tRNA<sup>Sec</sup>, facilitating incorporation of selenocysteine into the protein.

incorporate selenocysteine into the protein. For SelB to function, it must also bind a SECIS element in the mRNA, which has a secondary structure with a bulge that fits into SelB for insertion of a Sec residue. The SECIS element is adjacent to the UGA codon in *E. coli*, but it is often located in the 3'UTRs (3' untranslated regions) of eukaryotes and archaea. The SelB elongation factor is a homolog of elongation factor Tu.

The evolution of genetic code changes in small genomes such as those of mitochondria, the use of a few alternative initiation codons, and the use of a termination codon to incorporate selenocysteine—all these are relatively easy to understand as minor adjustments of a

universal code. But an unusual alteration in the genetic code occurs in many fungal species of the genus *Candida*, as originally discovered for *Candida albicans*. This fungus is an organism of high genomic complexity, yet its genetic code has a dramatic variation: the CUG codon that normally encodes leucine encodes serine instead. The selection pressure for this change is completely unknown. Furthermore, serine and leucine are quite different chemically. Yet, even this change can be understood based on the properties of a universal code.

When several codons encode the same amino acid and require multiple tRNAs, not all of the codons are used with equal frequency. In a phenomenon called **codon bias**, some codons for a particular amino acid are used more frequently (sometimes much more frequently) than others. The tRNAs for the frequently used codons are often present at much higher concentrations than the tRNAs for the rarely used codons. For example, there are six codons for leucine (see Figure 17-4). In bacteria, CUG is used often to encode Leu. However, in fungi closely related to *Candida*, CUG is used only rarely as a Leu codon and is often entirely absent in highly expressed proteins. A change in the coding sense of CUG would thus have a much smaller effect on fungal cell metabolism than might be expected if all Leu codons were used equally.

The coding change for the Leu codons may have occurred by a gradual loss of CUG codons in genes and of the tRNA that recognizes CUG as a Leu codon, followed by a capture event—a mutation in the anticodon of a tRNA<sup>Ser</sup> that allowed it to recognize CUG. Alternatively, there may have been an intermediate stage in which CUG was recognized as both a Leu and Ser codon, perhaps with contextual signals in the mRNAs that helped one tRNA or another recognize specific CUG codons (much like the signals used to insert selenocysteine at a particular stop codon). Phylogenetic analysis (see Chapter 8) indicates that the reassignment of CUG as a Ser codon occurred in *Candida* ancestors about 150 to 170 million years ago.

## SECTION 17.4 SUMMARY

- The genetic code, tRNAs, and translation must have evolved piecemeal, without protein components, and this evolution must have had a selective advantage, even at its earliest stages. Several hypotheses exist, but evidence to support any of them is limited.
- The genetic code is largely universal. Most exceptions occur in mitochondrial DNA, a small genome genetically isolated from the nucleus and relatively free to undergo evolutionary code

changes; many of these changes involve altered stop codons, yielding a streamlined genetic code that requires only 22 tRNAs.

- The few examples of genetic code alterations outside mtDNA usually involve the conversion of termination codons, in keeping with a common ancestry for the genetic code from which all variants are derived.

### Unanswered Questions

The genetic code has been deciphered and is of paramount importance to virtually every investigation in molecular biology. However, important fundamental questions remain. The nature of wobble pairing and the influence of tRNA structure on codon-anticodon pairing most likely will be explained through structural and mutational studies. The evolution of the genetic code remains a mystery, but creative experiments and investigations into exceptions to the code will no doubt provide further insight.

**1. How do nucleotides outside the anticodon influence the structure of tRNA for wobble pairing?** The use of noncanonical base pairs explains how wobble pairing can happen. But nucleotide substitutions outside the anticodon also affect wobble pairing in some way, perhaps through conformational changes when tRNA binds the ribosome.

**2. Why do tRNAs have so many modified bases?**

The proportion of modified bases in tRNAs can approach 20%, and many genes are devoted to synthesizing these modified bases. Yet we still know very little about the functions of these many modifications.

**3. How did the translation machinery evolve?** It is nothing short of mind-boggling to imagine how the translation machinery evolved in the first place. One challenge is the huge number of factors required for translation. The whole process could not have evolved all at once. What were the individual steps, and what forces of natural selection were at work?

# How We Know

## Transfer RNA Connects mRNA and Protein

Hoagland, M.B., M.L. Stephenson, J.F. Scott, L.I. Hecht, and P.C. Zamecnik. 1958. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* 231:241-257.

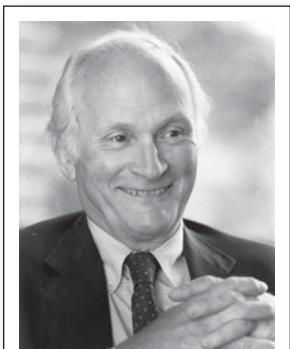


**Paul Zamecnik, 1912-2009**

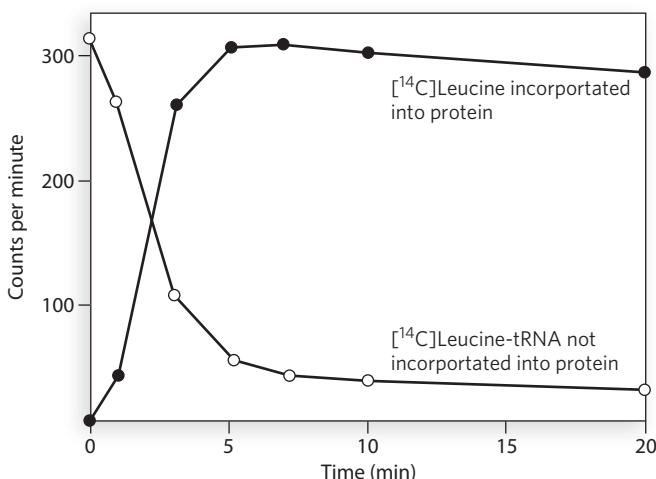
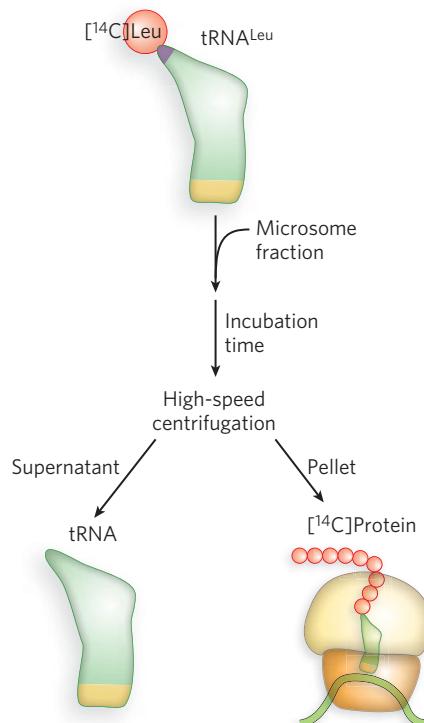
[Source: News Office,  
Massachusetts General  
Hospital.]

Once Francis Crick had hypothesized the existence of an adaptor molecule to bridge the information gap between mRNA and protein, the race was on to find it. It may seem obvious now, but before the discovery of aminoacyl-tRNAs, the nature of the adaptor was mysterious—but Crick did suggest that the adaptor might be, in part, a nucleic acid that could recognize individual codons in mRNA.

The race to find the adaptor was won by Mahlon Hoagland and Paul Zamecnik. They prepared a cell extract that contained the soluble tRNA and the enzymes needed to charge tRNA with amino acids. When [<sup>14</sup>C] leucine was added to the extract, the radiolabeled amino acid became attached to its tRNA. To show that this really was the adaptor molecule, Hoagland and Zamecnik demonstrated that the [<sup>14</sup>C]leucyl-tRNA could incorporate the [<sup>14</sup>C]leucine into a polypeptide chain. They prepared microsomes, a cell fraction containing mostly ribosomes collected as a pellet after high-speed centrifugation. On incubation of the [<sup>14</sup>C]leucyl-tRNA with mRNA and microsomes, [<sup>14</sup>C]leucine was transferred from the tRNA to protein—as was evident from the association of radiolabeled amino acid with ribosomes in the pellet after centrifugation of the reaction mixture (shown in the graph in Figure 1, solid circles), not with the tRNA in the supernatant (open circles). This experiment showing that amino acids are transferred from tRNAs to polypeptides was the first step in deciphering the genetic code.



**Mahlon Hoagland,  
1921-2009** [Source: Chris  
Christo/Worcester Telegram &  
Gazette/The Boston Globe.]



**FIGURE 1** An outline of Hoagland and Zamecnik's experimental method (top) and an example of their results (bottom). During protein synthesis, the radiolabeled amino acid, [<sup>14</sup>C]leucine, is transferred from the tRNA to the ribosome. [Source: M. B. Hoagland et al., *J. Biol. Chem.* 231:241-257, 1958.]

## Proteins Are Synthesized from the N-Terminus to the C-Terminus

Dintzis, H.M. 1961. Assembly of the peptide chains of hemoglobin. *Proc. Natl. Acad. Sci. USA* 47:247-261.

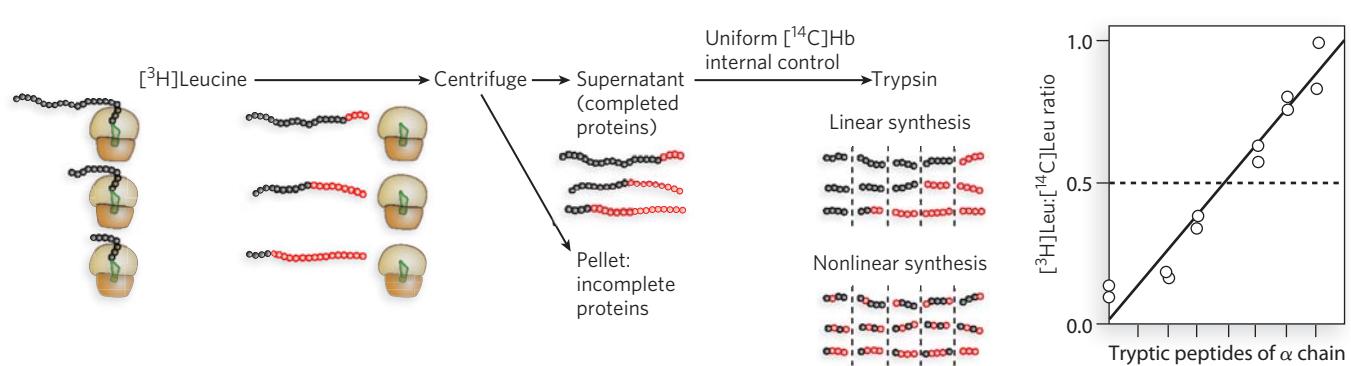


After the adaptor tRNA was discovered, the basic flow of information among the three biopolymers, DNA→RNA→protein, was understood. However, the order in which amino acids were connected to form a protein polymer was still unknown. To solve this, Howard Dintzis designed an ingenious experiment. His idea was to feed [<sup>3</sup>H]leucine to living cells and follow protein synthesis over time. Would incorporation of [<sup>3</sup>H]leucine start at one end of a newly synthesized protein, or would the labeled amino acid appear uniformly throughout the length of the new protein at all time points? This might sound like an easy experiment, but there were big technical hurdles. First, Dintzis had to follow only one protein, yet a cell makes many proteins all at once. Second, protein sequencing had not yet been invented, but Dintzis had to find some way to follow the sequence in which leucine was incorporated in the protein chain.

The beauty in Dintzis's experimental design lies in how he circumvented these difficulties. He knew that immature red blood cells (reticulocytes) turned off the synthesis of most proteins except hemoglobin, which is composed of two chains, α and β. Furthermore, he could use trypsin (a protease) to digest the α chain and could separate the fragments by paper electrophoresis, although he couldn't determine the ordering of the fragments along the protein. So Dintzis added [<sup>3</sup>H]leucine to reticulocytes, lysed cells at various times, separated the α and β chains, and analyzed the α chain by tryptic digestion and electrophoresis. Full-length proteins

would have [<sup>3</sup>H]leucine incorporated only in the part of the protein molecule synthesized last. To ensure that he obtained only full-length α chains, Dintzis removed incomplete chains, which remained bound to ribosomes, by centrifugation. As a control, he preincubated the cells with [<sup>14</sup>C]leucine so that he could compare new synthesis (<sup>3</sup>H]leucine) with overall synthesis (<sup>14</sup>C]leucine), correcting for the different leucine content in each peptide.

If protein chains are not made in a defined order, all peptides should have a similar ratio of <sup>3</sup>H to <sup>14</sup>C (Figure 2). But if proteins are made in a linear order, peptides will vary in their <sup>3</sup>H:<sup>14</sup>C ratio, and the part of the protein made last should contain a higher <sup>3</sup>H:<sup>14</sup>C ratio than the parts made earlier. The results were unambiguous and striking! The peptides differed greatly in <sup>3</sup>H:<sup>14</sup>C ratio, ruling out a random order of synthesis. Further, the tryptic peptides could be ordered to form a gradient of radioactivity. Hence, it was apparent that hemoglobin is synthesized from one end to the other. To determine the direction of synthesis, Dintzis digested the α chain with carboxypeptidase, which specifically removes amino acids from the C-terminus. Only one tryptic peptide was affected by carboxypeptidase treatment, and it was the peptide with the highest <sup>3</sup>H:<sup>14</sup>C ratio. This identified the C-terminus as the part of the protein that is synthesized last. Overall, the results showed that proteins are synthesized from the N-terminus to the C-terminus.



**FIGURE 2** Fragments at the end of the protein have a higher <sup>3</sup>H:<sup>14</sup>C ratio, demonstrating that protein synthesis is linear (open circles). If protein synthesis were random, the <sup>3</sup>H:<sup>14</sup>C

ratio would be constant over the length of the protein (dashed line). [Source: (a) H. M. Dintzis, *Proc. Natl. Acad. Sci. USA* 47:247-261, 1961. By permission of Howard Dintzis.]

## The Genetic Code In Vivo Matches the Genetic Code In Vitro

Terzaghi, E., Y. Okada, G. Streisinger, J. Emrich, M. Inouye, and A. Tsugita. 1966. Change of a sequence of amino acids in phage T4 lysozyme by acridine-induced mutations. *Proc. Natl. Acad. Sci. USA* 56:500-507.



Akira Tsugita, 1928-2007

[Source: Courtesy of Kazuyuki Nakamura.]

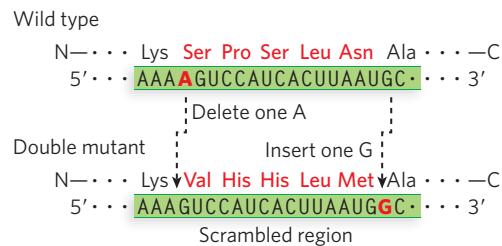
The experiments that defined the genetic code were brilliant, and this was Nobel Prize-winning work. But the investigations that cracked the code, as beautiful as they were in experimental design, were performed outside the context of a living cell, using cell extracts and synthetic mRNA. As a result, many scientists remained skeptical about the relevance of the newly discovered code to the *in vivo* situation. Akira Tsugita's group designed a powerful experiment to address exactly this issue. They studied acridine-induced mutations in the lysozyme gene of phage T4, which inactivate the gene, presumably by inducing the deletion or insertion of a base and thereby throwing off the reading frame. A double mutation in the lysozyme gene creates a pseudo-wild-type T4, consistent with the hypothesis by Crick and his colleagues that an insertion combined with a deletion results in a restored reading frame. According to this hypothesis, the area between the two mutations encodes an amino acid sequence different from that of the wild-type protein. As long as the insertion and deletion are not too far apart in the linear sequence of the gene, the double-mutant protein may retain a significant level of activity.

The investigators used a protease to digest lysozyme, subjected the fragments to electrophoresis, and studied the resulting peptide maps. Comparing maps of the double-mutant and wild-type lysozyme, they identified one peptide with changed electrophoretic mobility. Sequencing of the peptide revealed a five amino acid sequence unique to the mutant:

Wild type: Lys-Ser-Pro-Ser-Leu-Asn-Ala

Mutant: Lys-Val-His-His-Leu-Met-Ala

This result supported the triplet reading frame with no internal punctuation marks, as described by Crick. It could also be used to solve whether the genetic code in living cells is the same as that determined in extracts, by comparing the wild-type and mutant nucleotide sequences. If the code is indeed the same, the codon table established by *in vitro* experiments should produce a nucleotide sequence for the wild type that (1) encodes the wild-type sequence of amino acids, and (2) encodes the double-mutant lysozyme sequence, given either one nucleotide insertion followed by one deletion, or the reverse. Indeed, a solution can be found within these constraints, providing strong evidence at the time of this study that the genetic code derived from *in vitro* studies is the genetic code used in living cells (Figure 3). The result also supported the conclusion that mRNA is read in the 5'→3' direction, as the solution works only if codons are read in this direction. Because of technical limitations, the solution was not validated by sequencing until a decade later.



**FIGURE 3** For the known difference between the wild-type and double-mutant phage T4 lysozyme, the genetic code predicts deletion of an A residue followed by insertion of a G residue in the mutant protein.

## Key Terms

genetic code, p. 586  
 codon, p. 586  
 transfer RNA (tRNA), p. 587  
 aminoacyl-tRNA, p. 587  
 aminoacyl-tRNA synthetase, p. 587  
 anticodon, p. 588  
 translation, p. 588  
 degenerate code, p. 588

codon family, p. 589  
 wobble base, p. 590  
 wobble position, p. 590  
 wobble hypothesis, p. 590  
 reading frame, p. 590  
 initiation codon, p. 591  
 start codon, p. 591  
 termination codon, p. 591

stop codon, p. 591  
 open reading frame (ORF), p. 591  
 missense mutation, p. 591  
 silent mutation, p. 591  
 transition mutation, p. 592  
 nonsense mutation, p. 592  
 suppressor tRNA, p. 592  
 codon bias, p. 606

## Problems

- 1.** The following RNA polymer is added to an *E. coli* extract, where it can be translated in all three possible reading frames. Which amino acids can be polymerized into polypeptides in this system?

5'-AUUAUUAUUAUUAUUAUUAUUAUUAUAU-3'

- 2.** Given a polynucleotide that encodes polymethionine, what other polypeptides will also be produced?  
**3.** Translate the following mRNA into protein, starting from the first initiation codon:

5'-CCGAUGCCAUGGCAGCUCGGUGUUAC  
 AAGGUUGCAUCAGUACCAGUUUGAAUCC-3'

- 4.** From the sequence of a protein, we can gain some information about the gene sequence that encodes it. However, because of the degeneracy of the genetic code, there are many possible nucleotide sequences that could encode a given protein sequence. The usefulness of genomic databases in searching for the genes for proteins of known sequence is made clear by considering the following. How many possible RNA molecules can encode the peptide Met-Asn-Trp-Tyr? How many if a Leu residue is added to the end of the peptide?

- 5.** Shown below is the 5' end of an mRNA molecule. What are the first three (N-terminal) amino acids of its protein product?

5'-AUGUGUUGAUGUAUCAGACCUGUC - - -

- 6.** Translate the following mRNA, starting at the first 5' nucleotide, assuming that translation occurs in an *E. coli* cell. If all tRNAs make maximum use of wobble rules but don't contain inosine, how many distinct tRNAs are required to translate this RNA?

5'-AUGGGUCGUGAGUCAUCGUUAAUUGUAGCU  
 GGAGGGGAGGAAUGA-3'

- 7.** How does the answer to Problem 6 change if the RNA is translated in yeast mitochondria?

- 8.** For the following RNA sequence, which positions can tolerate a mutation without resulting in a change in amino acid sequence? What changes are tolerated at each position?

5'-AUGAUAUUGCUAUCUUGGACU-3'

- 9.** What polypeptide sequence will be made from the following RNA sequence?

5'-AUGCCUCGUCAGGUGUAAAGUCAGGCUUGA-3'

What tRNA<sup>Tyr</sup> suppressor mutation will provide read-through of the first stop codon, and what will the resulting peptide sequence be?

- 10.** What are the sequences of the polypeptides produced from these repeating nucleotide sequences: (a) poly(AG); (b) poly(UG); (c) poly(CAA); (d) poly(AAG); (e) poly(UUAC)?

- 11.** A researcher uses polynucleotide phosphorylase to create random RNA polymers, using a UDP:CDP ratio of 5:1. Codons should be generated in the following proportions, assuming random incorporation of the NDPs by polynucleotide phosphorylase: UUU (83.3), U<sub>2</sub>C (16.7), UC<sub>2</sub> (3.3), and CCC (0.7). The following amino acids are incorporated into protein, in the proportions shown in parentheses: leucine (22.2), phenylalanine (100), proline (5.1), and serine (23.6). What are the probable codon assignments of these four amino acids? Keep in mind that poly(U) codes for poly(Phe), and poly(C) codes for poly(Pro).

- 12.** Polyglycine is translated from the repeating sequence 5'-(GGU-GGC-GGA)<sub>n</sub>-3'. If only one tRNA is needed to make polyglycine, what can you say about the tRNA anticodon?

- 13.** A gene with a frameshift mutation caused by the insertion of one nucleotide produces inactive protein. A second frameshift, caused by the deletion of one nucleotide

at some position downstream of the original mutation, reactivates the gene. The final protein product contains four amino acid residues that differ from the wild-type protein. The two mutations occur in the following sequence:

5'...CATCATCATCATCATCATCATCATCAT...

What is the maximum number of nucleotides between the two point mutations? What is the minimum number?

- 14.** Given the following mRNA sequence, which reading frame is most likely to encode part of a protein?

5'-ACGUCGAGUAGCAGUAUCGAUUGAGC  
UCUUAGAUAGAUCGC

- 15.** Given the wobble rules, at least 31 tRNAs are necessary to decipher the genetic code. Only 6 tRNAs are needed to insert the four amino acids Phe, Leu, Ile, and Met. Using

the table below, hypothesize the anticodon sequences of the 6 tRNAs. Multiple answers are possible.

Amino Acid	Codon
Phe	UUU
	UUC
Leu	UUA
	UUG
	CUU
	CUC
	CUA
	CUG
Ile	AUU
	AUC
	AUA
Met	AUG

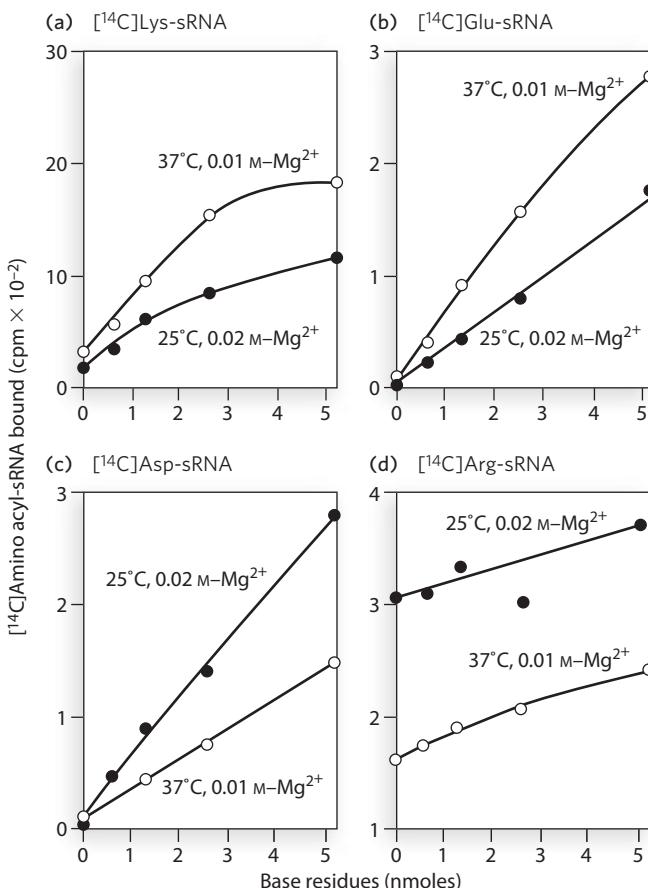
## Data Analysis Problem

**Nishimura, S., D.S. Jones, E. Ohtsuka, H. Hayatsu, T.M. Jacob, and H.G. Khorana. 1965.** Studies on polynucleotides XLVII: The *in vitro* synthesis of homopeptides as directed by a ribopolynucleotide containing a repeating trinucleotide sequence—new codon sequences for lysine, glutamic acid and arginine. *J. Mol. Biol.* 13:283–301.

- 16.** Once researchers had developed a few key strategies, the genetic code was solved within just a few years in the mid-1960s. One chapter of that story is described by Nishimura and coauthors. The work is elegant, while also demonstrating that results obtained in real-world experiments are not always as unambiguous as they may seem when presented in textbooks.

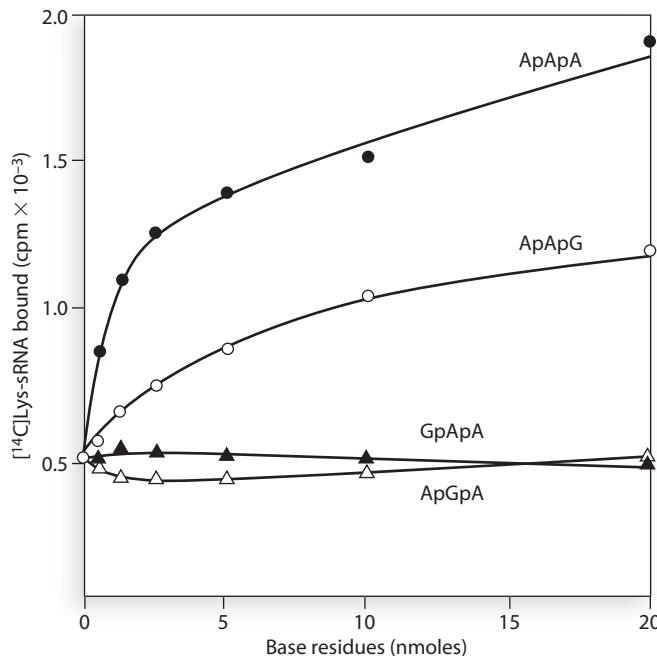
Using methods developed in the Khorana laboratory, Nishimura and colleagues examined the polypeptides generated by an oligonucleotide consisting of AAG repeats. In the first experiment (Figure 1), they examined the binding of radioactively labeled aminoacyl-tRNAs to the ribosome in response to the oligonucleotide. All of the tRNAs were tried, and only the four shown in Figure 1 gave a positive result (i.e., a labeled aminoacyl-tRNA-ribosome complex). Note that in 1965, tRNA was often referred to as sRNA (“soluble” RNA).

- (a) Given our present understanding of the genetic code, what is the maximum number of different labeled aminoacyl-tRNAs that could be bound to the ribosome in response to this oligonucleotide?  
 (b) Given the tenuous understanding of the code in 1965, can you suggest an explanation for the positive results obtained with the four different tRNAs in this experiment?



**FIGURE 1**

One of the possible codons present in the  $(AAG)_n$  oligonucleotide was assigned to lysine, based on a series of experiments including the one shown in **Figure 2**. The researchers used the method advanced by Nirenberg to determine which trinucleotide would stimulate binding of [ $^{14}\text{C}$ ]Lys-tRNA<sup>Lys</sup> to a ribosome.



**FIGURE 2**

- (c) Which codon did the researchers assign to lysine?
- (d) The researchers also tested the trinucleotide AAA for its response to [ $^{14}\text{C}$ ]Lys-tRNA<sup>Lys</sup> (see Figure 2). Suggest why they would include AAA in their experiment, even though AAA was not represented in the  $(AAG)_n$  repeating oligonucleotide.

## Additional Reading

### General

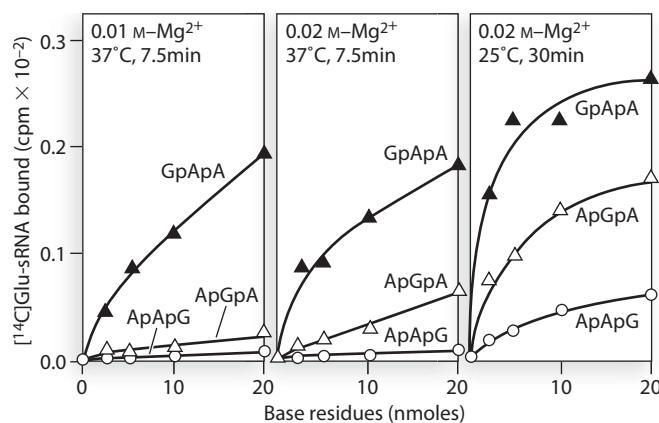
#### Cold Spring Harbor Symposia on Quantitative Biology.

**1966.** *The Genetic Code*. Vol. 31. Cold Spring Harbor, NY: Cold Spring Harbor Laboratories. The symposium was held when the genetic code was being completed; the introduction by Francis Crick gives a fascinating portrait of where things stood and where they were going.

**Crick, F.H.C. 1988.** *What Mad Pursuit: A Personal View of Scientific Discovery*. New York: Basic Books, New York. A wonderful account of Crick's personal odyssey in science.

**Vogel, G. 1998.** Tracking the history of the genetic code. *Science* 281:329–331.

Next, the researchers turned their attention to the tRNAs for glutamine. With [ $^{14}\text{C}$ ]Glu-tRNA<sup>Glu</sup>, they obtained the results shown in **Figure 3**.



**FIGURE 3**

- (e) From these results, which codon did the authors assign to Glu?
- (f) Suggest why all three codons elicited a positive response in the experiment shown in the rightmost panel of Figure 3.
- In their tests, the researchers then examined the production of homopolymers of amino acids in response to  $(AAG)_n$ . They found poly(Glu), poly(Arg), and poly(Lys), but no poly(Asp).
- (g) Based on all of the results shown here, what are the most likely codon assignments for the codons present in  $(AAG)_n$ ?
- (h) Go back to question (b) and suggest why Asp-tRNA<sup>Asp</sup> bound to ribosomes in response to this oligonucleotide in the first experiment (see Figure 1).

### Deciphering the Genetic Code: tRNA as Adaptor

**Crick, F.H.C. 1970.** Central dogma of molecular biology. *Nature* 227:561–563. The classic paper in which Crick proposes the central dogma of information flow in biology.

**Hoagland, M.B., M.L. Stephenson, J.F. Scott, L.I. Hecht, and P.C. Zamecnik. 1958.** A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* 231:241–257. The report that identified aminoacylated tRNA as Crick's adaptor molecule.

**Holley, R.W., J.H. Apgar, G.A. Everett, J.T. Madison, M. Marquisse, S.H. Merrill, J.R. Penswick, and A. Zamir. 1965.** Structure of a ribonucleic acid. *Science*

147:1462–1465. This documents the first structure of a tRNA, yeast tRNA<sup>Ala</sup>; the putative anticodon contains the unusual nucleotide inosine.

### The Rules of the Code

**Brenner, S., A.O. Stretton, and S. Kaplan, S. 1965.** Genetic code: The “nonsense” triplets for chain termination and their suppression. *Nature* 206:994–998.

**Dintzis, H.M. 1961.** Assembly of the peptide chains of hemoglobin. *Proc. Natl. Acad. Sci. USA* 47:247–261. Ingenious experiments determined the direction in which proteins are synthesized.

**Yanofsky, C., B.C. Carlton, J.R. Guest, D.R. Helinski, and U. Henning. 1964.** On the colinearity of gene structure and protein structure. *Proc. Natl. Acad. Sci. USA* 51:266–272. Different mutations in an enzyme are used to deduce that the linear order of mutations in a gene correspond to the linear order in the protein.

### Cracking the Code

**Khorana, H.G., H. Buchi, H. Ghosh, N. Gupta, T.M. Jacob, H. Kossel, R. Morgan, S.A. Narang, E. Ohtsuka,**

**and R.D. Wells. 1966.** Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 31:39–49.

**Nirenberg, M.W., and J.H. Matthaei. 1961.** The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* 47:1588–1602. The first study that outlined an experimental approach to crack the genetic code.

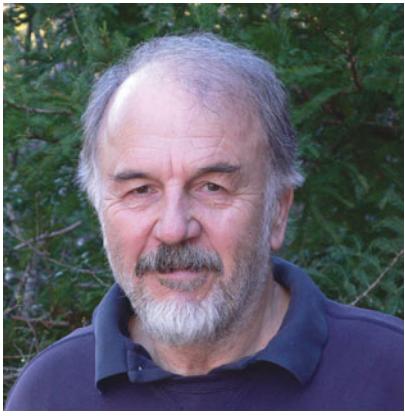
**Speyer, J.F., P. Lengyel, C. Basilio, A.J. Wahba, R.S. Gardner, and S. Ochoa. 1963.** Synthetic polynucleotides and the amino acid code. *Cold Spring Harb. Symp. Quant. Biol.* 28:559–567.

### Exceptions Proving the Rules

**Fox, T.D. 1987.** Natural variation in the genetic code. *Annu. Rev. Genet.* 21:67–91.

**Knight, R.D., S.J. Freeland, and L.F. Landweber. 2001.** Rewiring the keyboard: Evolvability of the genetic code. *Nat. Rev. Genet.* 2:49–58.

# Protein Synthesis



**Harry Noller** [Source: Courtesy of Harry Noller.]

inactivate protein enzymes, hoping to recover nonfunctional ribosomes that could be analyzed to determine which proteins had been affected and thereby discover those responsible for activity. The problem was, the ribosome withstood nearly all the standard chemical treatments we tried!

An unexpected result from one experiment led me to suggest to an undergraduate researcher, Brad Chaires, to try the RNA-specific reagent kethoxal, which reacts with guanine bases to produce adducts that disrupt base pairing. To our great surprise, ribosomes were inactivated by modification of just six G residues out of the more than 4,000 nucleotides in the ribosomal RNA.

When Brad graduated, I followed up with reconstitution experiments showing that it was, in fact, the ribosomal RNA that had been functionally inactivated, and that tRNA could protect ribosomes from kethoxal inactivation. Protection resulted from the tRNA binding to the ribosome and blocking kethoxal's access to parts of the rRNA, so they couldn't be chemically modified. These results led us to propose that ribosomal RNA, rather than ribosomal protein, was responsible for the functional activity of the ribosome. Many colleagues considered this "a crackpot idea," which I found frustrating but, in the antiestablishment spirit of the 1970s, also motivating.

Our findings led us to sequence the ribosomal RNAs (which Carl Woese referred to as "Sacred Scrolls"), and we were excited to find that the sites of chemical inactivation corresponded to the most evolutionarily conserved parts of rRNA. We then embarked on a fruitful collaboration with Woese to determine the secondary structures of the ribosomal RNAs, which ultimately led us in the direction of working out the three-dimensional structure of the ribosome. After 35 years, we can at last see that ribosomal RNA is indeed the functional core of the ribosome.

—**Harry Noller**, on discovering the functional importance of ribosomal RNA

- 18.1 The Ribosome 616**
- 18.2 Activation of Amino Acids for Protein Synthesis 624**
- 18.3 Initiation of Protein Synthesis 629**
- 18.4 Elongation of the Polypeptide Chain 638**
- 18.5 Termination of Protein Synthesis and Recycling of the Synthesis Machinery 642**
- 18.6 Translation-Coupled Removal of Defective mRNA 647**
- 18.7 Protein Folding, Covalent Modification, and Targeting 654**

The synthesis of proteins is the final step in the flow of genetic information, beginning with DNA replication and continuing with transcription into mRNA. Because of the abundance of proteins—making up roughly 44% of the dry weight of a human body, for example—and their central catalytic, transport, structural, and regulatory roles in all organisms, substantial cellular resources are devoted to protein synthesis. Like DNA and RNA synthesis, the process of protein synthesis can be considered in terms of initiation, elongation, and termination stages. Furthermore, as in nucleic acid synthesis, substrates must be activated before polymerization, and the completed product, in this case the polypeptide chain, must be chemically modified, folded, and targeted to the correct intracellular or extracellular location to become a functional protein.

The transfer of information from the 4-letter nucleotide sequence of an mRNA into the 20-letter amino acid sequence of a protein, however, is a fundamentally more complex task than nucleotide synthesis. In contrast to the transcription of DNA into RNA, in which a direct correspondence forms by hydrogen bonding between the sequence of the template strand and that of the synthesized RNA strand, there is no obvious chemical correspondence between the three-nucleotide mRNA codons and the amino acids they represent. In the 1950s, Francis Crick suggested that amino acids were attached to “adaptor” RNA molecules that could provide direct base-pairing complementarity with each codon in an mRNA sequence. And indeed, Paul Zamecnik and colleagues later showed that amino acids are covalently attached to RNA molecules, subsequently identified as tRNAs (see Chapter 17). The aminoacyl-tRNA molecules associate with the ribosomes, which were shown to be composed of both RNA and protein. Together with mRNAs, tRNAs, and aminoacyl-tRNA synthetases, ribosomes carry out the coupled tasks of recognizing each three-nucleotide codon of the genetic code and incorporating the specified amino acid into a growing polypeptide chain. Experiments in many laboratories showed that the molecular machinery of translation is essentially the same in all cells, although, as we’ll see, some of the mechanistic details of protein synthesis differ in bacteria and eukaryotes.

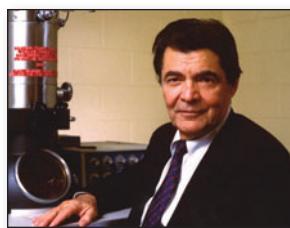
In Chapter 17 we introduced the genetic code and the mechanism of decoding by tRNAs. In this chapter, we discuss the structures and activities of ribosomes and aminoacyl-tRNA synthetases. We then explore the ways in which these remarkable molecules work together with mRNA and tRNA to carry out fast, accurate protein synthesis.

## 18.1 The Ribosome

In 1974, George Palade received a Nobel Prize for his discovery, in the 1950s, of **ribosomes**, the macromolecular machines responsible for protein synthesis (Figure 18.1). Although historically considered organelles, ribosomes are not encapsulated in a lipid membrane and thus are more accurately described as very large (macro) molecules. Ribosomes are present in the cytosol of all cells, as well as in the matrix of mitochondria and the stroma of chloroplasts. Because protein synthesis is common to virtually all life forms, ribosomes are evolutionarily well-conserved. In bacteria, they can translate mRNA as it is being transcribed from DNA, but in eukaryotes, the nuclear envelope and mRNA processing steps separate RNA synthesis from translation (see Chapter 16). Ribosomes are abundant; each *E. coli* cell contains approximately 15,000 ribosomes, making up almost 25% of the dry weight of the cell. The large number of these macromolecules is essential to the cell’s ability to produce proteins when needed. The structure of the ribosome facilitates protein synthesis by bringing together the mRNA codon and the corresponding charged tRNA adaptor, and then catalyzing peptide bond formation. Although the general function of ribosomes has been well-established for some time, recent advances in understanding the structure and assembly of the ribosome have elucidated more of the details of protein synthesis. Here we provide an overview of ribosomal structure and function that will serve as a framework for the details in the rest of the chapter.

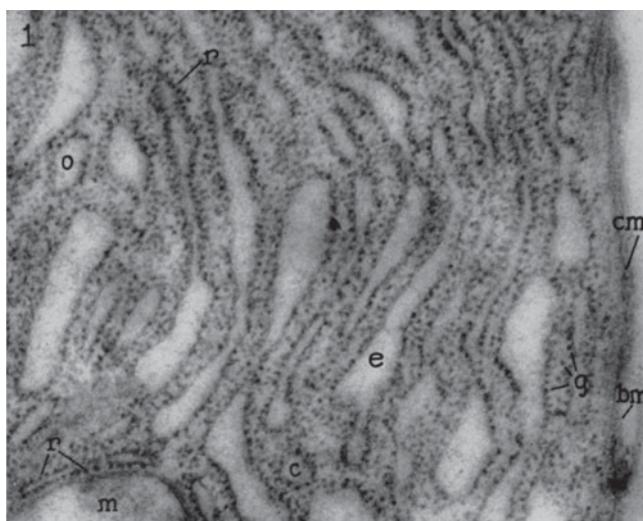
### The Ribosome Is an RNA-Protein Complex Composed of Two Subunits

Bacterial ribosomes contain about 60% ribosomal RNA and 40% protein, organized into two unequal subunits that are named according to their sedimentation coefficients (in svedberg units, S; see Chapter 16). The 50S subunit, the larger of the two, contains the peptidyl transferase center, which catalyzes peptide bond formation between adjacent amino acids in a growing polypeptide chain. The 30S subunit contains the decoding center where aminoacylated tRNAs “read” the genetic code by base pairing with each triplet codon in the mRNA. The assembled ribosome, with a



**George Palade, 1912-2008**

[Source: Courtesy of University of California, San Diego.]



**FIGURE 18-1 Ribosomes.** In this electron micrograph from 1955, most of the ribosomes are attached to a membrane, the endoplasmic reticulum. [Source: © The Rockefeller University Press. *The Journal of Cell Biology* 1:59–68, 1955.]

combined sedimentation coefficient of 70S, smoothly integrates the functions of each subunit to ensure rapid, accurate protein synthesis.

Each ribosomal subunit contains dozens of **ribosomal proteins (r-proteins)** and one or more large **ribosomal RNAs (rRNAs)** (Table 18-1). Like the subunits themselves, rRNAs are named in svedberg units. The 50S subunit contains a 5S and a 23S rRNA, and the 30S subunit contains a single 16S rRNA, together comprising more than 4,500 nucleotides. The 50S subunit in *E. coli* also contains more than 30 different r-proteins, named L1 through L36, and the 30S subunit contains a single copy of each of 21 different r-proteins, named S1 through S21 (L denotes large subunit, S denotes small). Although there are many more individual r-proteins than RNAs in the ribosome, the relatively small size of most r-proteins (~15 kDa, on average, in bacteria) means that the proteins contribute only about one-third of the overall mass of the ribosome.

Decades of research on the structure and function of ribosomal proteins and RNAs have shifted the focus from the proteins to the rRNA. Despite the complexity of

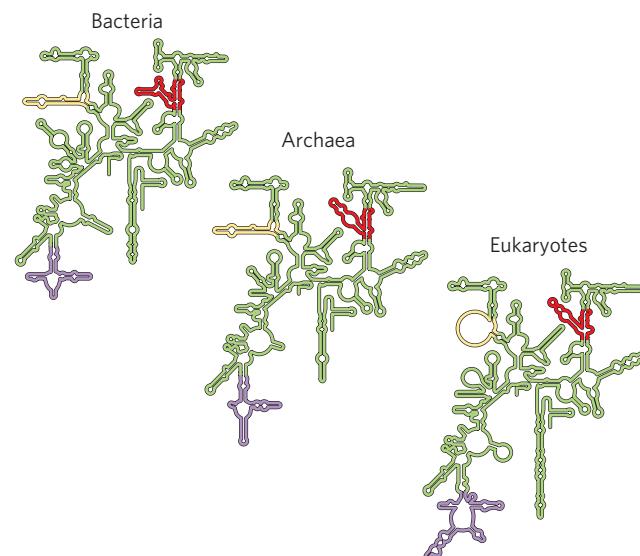
ribosome structure, Masayasu Nomura and colleagues demonstrated in the late 1960s that both bacterial ribosomal subunits can be broken down into their RNA and protein components, then reconstituted in vitro. Under appropriate experimental conditions, the RNAs and proteins spontaneously reassemble to form 50S or 30S subunits nearly identical in sedimentation behavior and activity to native subunits. Se-

quencing of rRNAs in the 1970s, and the secondary structure proposed for 16S rRNA by Harry Noller and Carl Woese in 1981, showed that rRNAs have been highly conserved by evolution, particularly in molecular regions implicated in the critical functions of the ribosome (Figure 18-2). Noller's biochemical analysis of functional



**Masayasu Nomura**

[Source: Courtesy of Masayasu Nomura.]



**FIGURE 18-2 Conservation in the secondary structures of small-subunit rRNAs from the three domains of life.** The red, yellow, and purple indicate areas where the structures of the rRNAs from bacteria, archaea, and eukaryotes have diverged; conserved regions are shown in green.

**Table 18-1 Protein and RNA Components of the *E. coli* Ribosome**

Ribosomal Subunit	Number of Different Proteins	Total Number of Protein Subunits	Protein Designations	rRNAs
50S	33	36	L1-L36	5S rRNA and 23S rRNA
30S	21	21	S1-S21	16S rRNA



**Tom Steitz (left), Ada Yonath (center), and Venki Ramakrishnan (right).** [Source: AP Photo/Scanpix Sweden/Bertil Ericson.]

data also provided an important tool for analyzing evolutionary relationships among species.

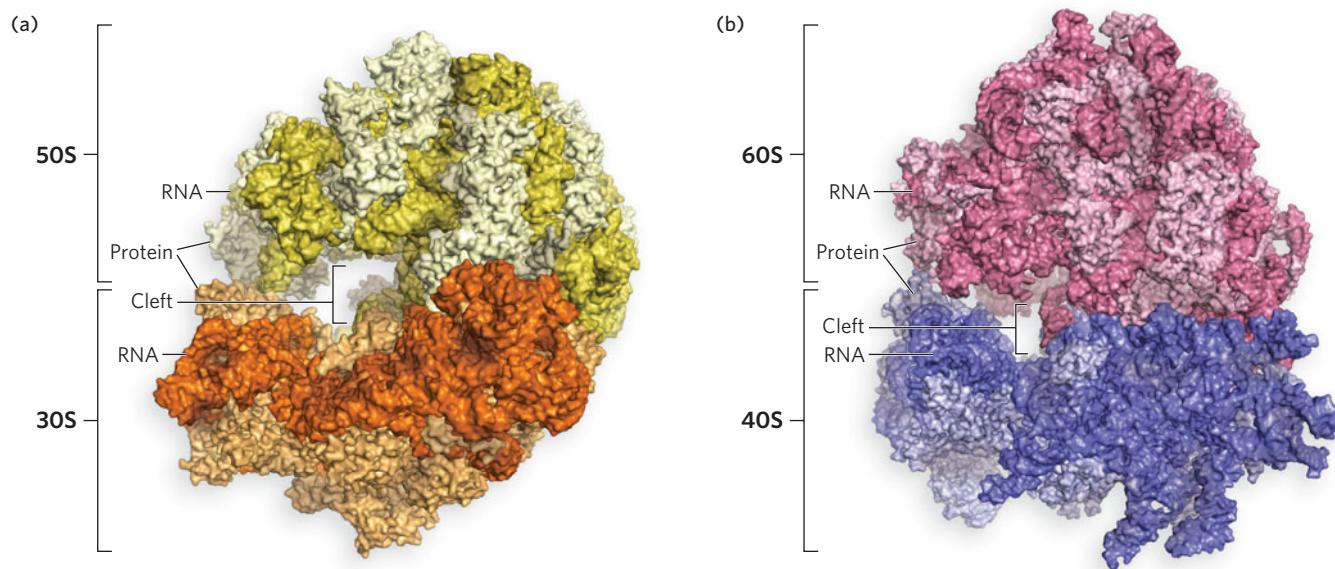
During the past decade—in the culmination of many years of work in multiple laboratories—cryo-electron microscopy (using frozen samples) and x-ray crystallography have revealed the atomic-resolution structure of the bacterial ribosome and its subunits in exquisite detail (**Figure 18-3a**). Venki Ramakrishnan, Tom Steitz, and Ada Yonath received a Nobel Prize in 2009 for their crystallographic work. More recently, the crystal structure of a eukaryotic ribosome was solved, revealing a similar overall structure but with more complexity that reflects additional levels of regulation (**Figure 18-3b**).

With a combined molecular weight of  $\sim 2.5 \times 10^6$  Da, the bacterial ribosome is far more complex than the DNA or RNA polymerases, and knowledge of its structure has provided a wealth of insights about its function

and evolutionary origins. The two irregularly shaped ribosomal subunits fit together to form a cleft through which the mRNA passes during translation. Although the proteins in bacterial ribosomes vary in size and structure, most have globular domains arranged on the ribosomal surface. Some proteins also have snakelike extensions that protrude into the rRNA core of the ribosome, stabilizing its structure. The functions of some of these proteins have not yet been determined, but a structural role seems likely for many of them.

Each of the three single-stranded rRNAs of *E. coli* has a specific three-dimensional conformation with extensive intrachain base pairing. This secondary structure was predicted based on the available sequences of rRNAs from many organisms (see Figure 18-2.) These structures have largely been confirmed in high-resolution three-dimensional models, yet they fail to convey the extensive network of tertiary interactions evident in the complete structure.

Extensive structural, biochemical, and genetic data support the conclusion that rRNA is responsible for all ribosomal functions, including tRNA binding and peptide bond formation. The predominant location of r-proteins on the outer surface of the ribosome, away from its RNA-rich functional interface, underscores this idea (see Figure 18-3). In addition to dominating the functional centers within each ribosomal subunit, rRNA provides most of the contacts between the two subunits in the intact ribosome. The intersubunit



**FIGURE 18-3 Crystal structure of the bacterial ribosome.** (a) The 50S (large) and 30S (small) subunits together form the 70S ribosome. The interface between the two subunits forms a cleft where the peptidyl transferase reaction (as described later in the text) occurs. (b) The yeast ribosome has a similar structure with increased complexity. [Sources: (a) PDB ID 1SVA and 2OW8; (b) PDB ID 3O58 and 3O2Z.]

bridges don't just hold the two subunits together; they foster relative movement between the subunits that is integral to the process of polypeptide elongation (see Section 18.4).

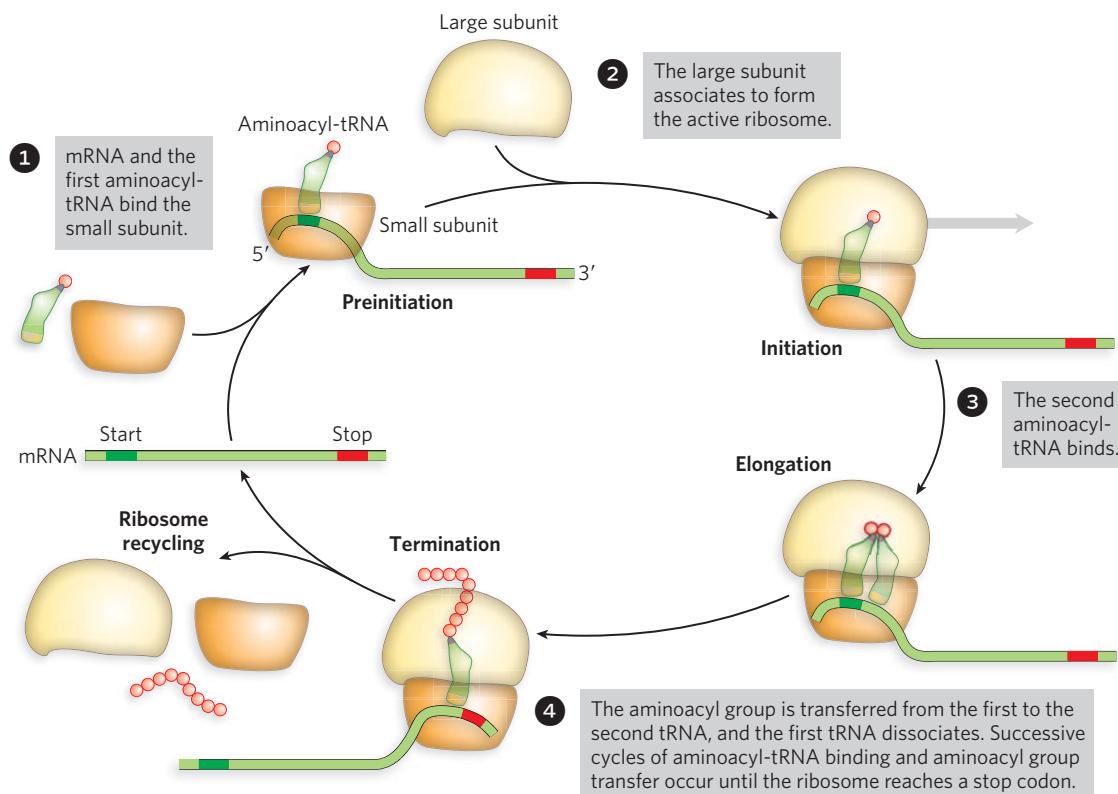
The ribosomes of eukaryotic cells (other than the mitochondrial and chloroplast ribosomes) are larger and more complex than bacterial ribosomes. Electron microscopic and centrifugation studies show that these ribosomes have a diameter of about 23 nm and a sedimentation coefficient of about 80S. Like their bacterial counterparts, eukaryotic ribosomes have two subunits. The subunits vary in size among species but, on average, are 60S and 40S. Altogether, the cytosolic ribosomes contain more than 80 different proteins and four types of rRNA; unlike bacterial ribosomes, they cannot be spontaneously reconstituted *in vitro*, suggesting a more complex assembly process. In contrast, the ribosomes of mitochondria and chloroplasts are somewhat smaller and simpler than bacterial ribosomes. Nevertheless, the rRNAs of all cell types are conserved (and as we saw in Chapter 17, all tRNAs have similar sizes and shapes), indicating that ribosomal

structure and function are strikingly similar in all organisms and organelles.

## Ribosomal Subunits Associate and Dissociate in Each Cycle of Translation

The association of ribosomal subunits at the start of protein synthesis and their dissociation on release of the completed polypeptide are fundamental to the process of translation in all cells. Because the subunits are initially separated, the initiation of protein synthesis is inherently regulated by the assembly of active ribosomes on an mRNA, together with tRNAs. The recruitment of the small ribosomal subunit to an mRNA is a first essential step, and cells and viruses have a variety of mechanisms for controlling how and when this happens.

An overview of translation is shown in **Figure 18-4**. In general, translation begins with mRNA and an initiator tRNA binding to the small ribosomal subunit. Once the small subunit is positioned at the beginning of the coding sequence of the mRNA, the large



**FIGURE 18-4** An overview of the main events in translation. Translation is initiated by the pairing of an mRNA and tRNA on the ribosome. In elongation, the ribosome moves along the mRNA, matching tRNAs to each

codon and catalyzing peptide bond formation. Translation terminates at a stop codon, and the ribosomal subunits are released for another round of synthesis.

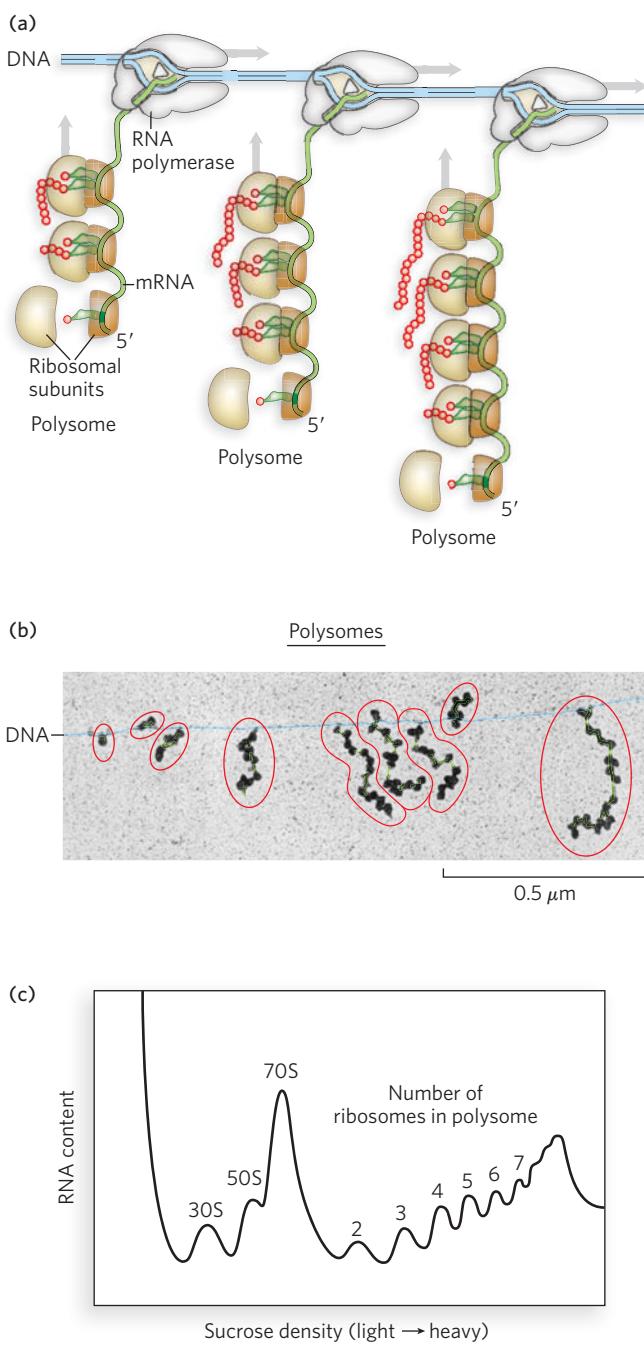
ribosomal subunit joins noncovalently to form an active ribosome that begins reading each mRNA codon in sequence in the 5'→3' direction. As each codon is encountered, a matching tRNA bearing its covalently attached amino acid enters the decoding and peptidyl transferase centers of the ribosome. After peptide bond formation between the C-terminal end of the growing polypeptide chain and the amino acid of the incoming aminoacyl-tRNA, the ribosome translocates to the next codon, and the cycle repeats. On encountering a stop codon, the ribosome is released and its subunits dissociate, ready to begin translation on another mRNA.

Because mRNAs are usually at least 300 nucleotides long and extend well beyond the ~200 Å girth of the ribosome, multiple ribosomes can occupy each mRNA during translation, forming a **polysome**, or **polyribosome** (Figure 18-5a). Polysomes can be directly visualized by electron microscopy (Figure 18-5b). Because they form very high molecular weight particles, they can also be readily detected in cell extracts by analyzing their sedimentation in sucrose or glycerol density gradients (Figure 18-5c). Polysome formation lets each mRNA molecule provide the template for multiple copies of a protein molecule at once, thus allowing the efficient use of each mRNA.

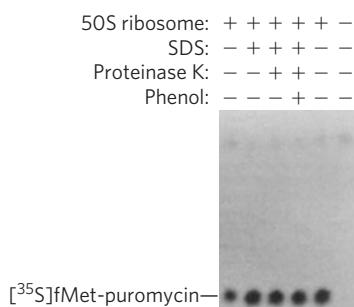
### The Ribosome Is a Ribozyme

Noller concluded that ribosomal RNA has a fundamental importance in catalyzing protein synthesis on the basis of the results of two experiments. First, he inactivated the ribosome by making base changes in rRNA; second, he removed proteins from the ribosome and found that the deproteinized ribosome retained peptidyl transferase activity (see How We Know).

The object of the second experiment was to remove proteins without disrupting rRNA structure, which precluded the use of extraction reagents such as acid. Instead, detergents, a protease, and phenol were used, all of which are effective deproteinizing treatments that do not degrade or denature RNA. To monitor the activity of ribosomes after protein removal, Noller and colleagues used a simplified peptidyl transferase reaction, referred to as the fragment reaction, which requires only the 50S subunit and does not need initiation or elongation factor proteins (these factors are discussed later in the chapter); the assay monitors the addition of [<sup>35</sup>S]methionine to an amino acid mimic, puromycin. Using the fragment reaction, the researchers found that the 50S subunit of *Thermus aquaticus*, a bacterium that grows at elevated temperatures (a



**FIGURE 18-5 Polysomes.** (a) Polysomes consist of multiple ribosomes associated with a single mRNA. (b) The *E. coli* polysomes (circled in red) visible in this electron micrograph are translating mRNA (green) as it is being transcribed from DNA (blue line). The promoter is at the left. (c) Polysomes can be separated on a sucrose density gradient. The polysome profile indicates total RNA content at increasing sucrose density. The results show that 30S and 50S ribosomal subunits as well as 70S ribosomes migrate near the top of the gradient, whereas polysomes comprising two or more ribosomes associated with mRNA transcripts migrate in the heavier (more dense) fractions of the gradient. [Source: (b) L. O. Miller et al., *Science* 169:392–395, 1970.]

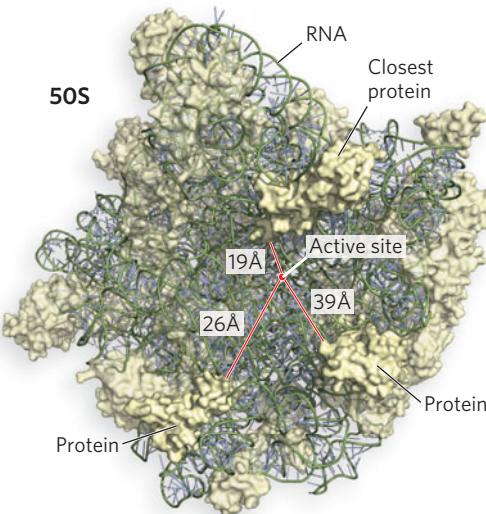


**FIGURE 18-6** The effects of method of protein extraction on ribosomal function. Ribosomes from the bacterium *Thermus aquaticus* were treated with the detergent sodium dodecyl sulfate (SDS), the enzyme proteinase K, or phenol, then tested for peptidyl transferase activity using the fragment reaction. The product of the reaction, [<sup>35</sup>S]fMet-puromycin, was assayed by paper electrophoresis and autoradiography. (fMet is *N*-formylmethionine, the initiating amino acid in bacterial protein synthesis, as described later in the chapter.) [Source: H. F. Noller, V. Hoffarth, and L. Zimniak, *Science* 256:1416–1419, 1992, Fig. 2b.]

thermophile), retained peptidyl transferase activity even after proteins were extracted by treatments with detergent, proteinase K (a nonspecific protease that degrades virtually all proteins), and phenol (Figure 18-6). The thermophilic ribosomes were selected because they were expected to be inherently more stable than ribosomes from organisms such as *E. coli* that grow at moderate temperatures. However, even *E. coli* 50S subunits remained active after treatment with detergent or proteinase K, suggesting that the rRNA structure was largely intact.

These important findings foreshadowed the discovery in the crystal structure of the 50S ribosomal subunit that there is no protein within 18 Å of the peptidyl transferase active site (Figure 18-7). The high-resolution structure thus confirmed what had been suspected for more than a decade: the ribosome is a ribozyme. The ribosomal RNA, not protein, is responsible for catalysis of peptide bond formation.

In addition to confirming the central role of rRNA in peptide bond formation, crystal structures of the ribosome and its subunits show that most of the contacts between tRNA and the ribosome involve contacts with 16S or 23S rRNA, not with protein. Thus, with these new data, the traditional focus on the protein components of ribosomes shifted. In addition to providing the structural core, rRNAs form the functional core of the ribosome, implying that r-proteins are secondary elements in the complex, playing a stabilizing or



**FIGURE 18-7** Ribosomal RNA in the ribosomal active site.

The active site—where the peptidyl transferase forms peptide bonds—is 18 Å away from the closest r-protein, evidence that the ribosome is a ribozyme. [Source: PDB ID 1Q7Y.]

regulatory role rather than a catalytic role in the business of translation.

What do these findings suggest about the origin of this most fundamental cellular process? Francis Crick wondered in the 1960s about the possible existence of an all-RNA ribosome at some early point in evolution. The importance of rRNA in modern ribosomes supports this idea and is consistent with the addition of proteins to a preexisting rRNA-containing ribosome over the course of evolution. Recent studies of mitochondrial ribosomes suggest that r-proteins may restrict the variety of proteins that a ribosome can synthesize (Highlight 18-1).

### The Ribosome Structure Facilitates Peptide Bond Formation

The ribosome must bind simultaneously to at least two tRNAs during each cycle of amino acid addition to the C-terminus of a growing polypeptide chain. In fact, equilibrium binding studies and ribonuclease protection experiments showed that ribosomes contain binding sites for three tRNAs (Figure 18-8). The **A site** is the location of aminoacyl-tRNA binding, the **P site** is the location of peptidyl-tRNA binding, and the **E site** is the exit site, occupied by the tRNA molecule released after the growing polypeptide chain is transferred to the aminoacyl-tRNA. Each tRNA starts in the A site, moves to the P site after peptide bond formation, then exits through the E site.

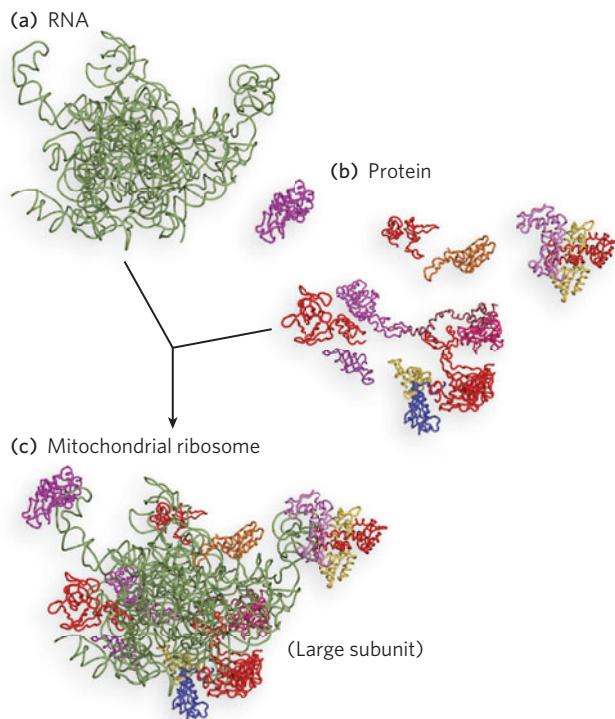
## HIGHLIGHT 18-1 EVOLUTION

### Mitochondrial Ribosomes: A Window into Ribosome Evolution?

Mitochondria encode their own ribosomes for synthesizing some of the mitochondrial proteins that carry out oxidative phosphorylation to produce the cell's ATP. Although mitochondria are thought to have descended from symbiotic bacteria, their ribosomes differ substantially from those of modern bacteria. In particular, the ratio of protein to rRNA in animal mitochondrial ribosomes is 2 : 1 by mass, rather than the 1 : 2 ratio observed for bacterial and archaeal ribosomes. Because the ribosomes have roughly the same mass, this means that mitochondrial ribosomes are two-thirds protein and one-third rRNA, whereas bacterial ribosomes are one-third protein and two-thirds rRNA by mass. What does this imply about the origins of the ribosome and the importance of rRNA in its function?

To begin to answer these questions, Rajendra Agrawal and Stephen Harvey proposed a structural model of the mitochondrial large ribosomal subunit, using a combination of cryo-electron microscopy and molecular modeling. Although the resolution of the electron microscopy-derived electron density, 12.1 Å, was lower than that of x-ray crystallography, it was possible to model higher-resolution bits of structure into the electron density map by using various molecular landmarks. The resulting model predicts the arrangement of individual protein and rRNA components in the large subunit of the mitochondrial ribosome (Figure 1). Although there is much less rRNA than r-protein, the rRNA forms the same overall architecture and occupies the same positions known to be essential for protein synthesis in bacterial and archaeal ribosomes. However, the r-proteins encroach on the all-RNA core of the mitochondrial ribosome, where they substitute for segments of rRNA that are missing relative to bacterial ribosomes.

Because mtDNA has a higher mutation rate and therefore evolves faster than the cellular genome, the mitochondrial ribosome presumably is now farther from its all-RNA origins than is the bacterial ribosome. This may reflect the greater structural and functional diversity of proteins relative to RNA. Far in the future, mitochondrial and, eventually, all ribosomes might consist primarily of protein, as their rRNA components are replaced over time. Mammalian mitochondria have evolved to synthesize just 13 proteins, so their ribosomes may be subject to different selective



**FIGURE 1** (a) The RNA portions and (b) the protein portions of the large ribosomal subunit of mammalian mitochondria, determined by cryo-electron microscopy and modeled using available crystal structures of the bacterial 50S ribosomal subunit. (c) The proteins and RNAs making up the complete large subunit. Notice how protein segments penetrate the interior of the subunit. [Source: PDB ID 2FTC.]

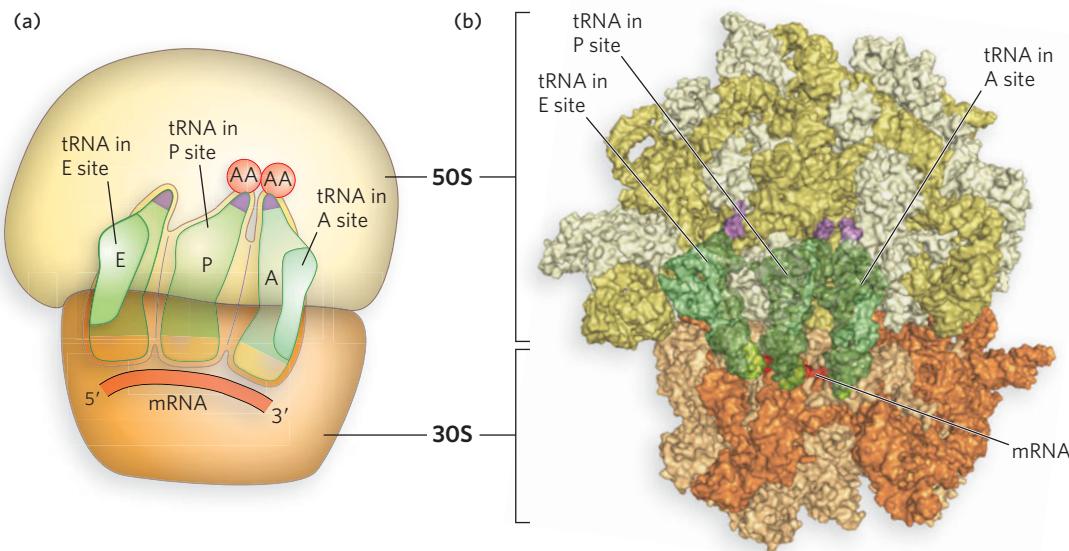
pressures from those at work on cytosolic ribosomes. Comparative studies of mitochondrial and cytosolic ribosome structure and activity may provide unexpected insights into the changing roles of RNA and protein in ribosome form and function.



**Rajendra Agrawal** [Source: Courtesy of Rajendra K. Agrawal.]



**Stephen Harvey** [Source: Courtesy of Georgia Research Alliance.]

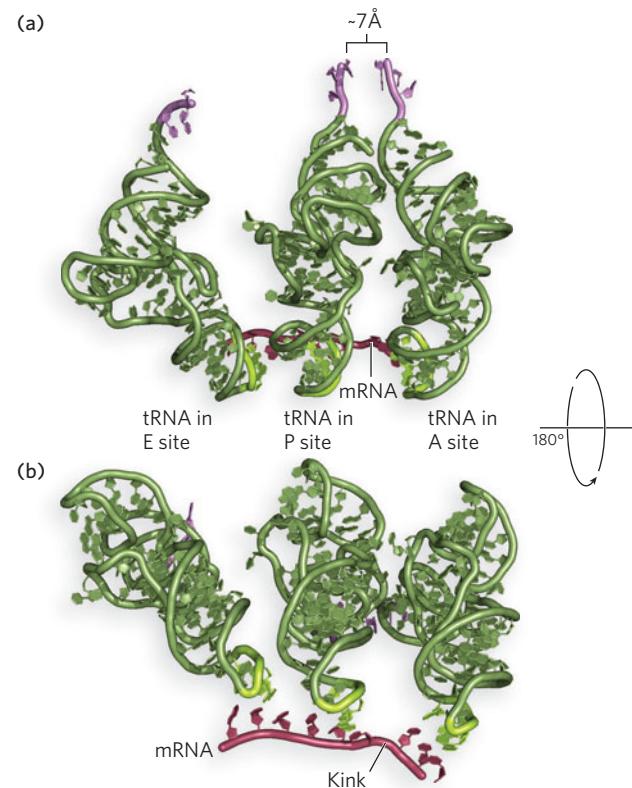


**FIGURE 18-8 The A, P, and E sites of the ribosome.** (a) The A, P, and E sites in relation to bound mRNA. Both ribosomal subunits are shown. (b) A crystal structure of the *E. coli* 50S subunit with bound tRNAs (representing aminoacyl-tRNA, peptidyl-tRNA, and free tRNA, respectively) in the A, P, and E sites, viewed from the 30S interface. [Sources: (b) PDB 1SVA and PDB 2OW8.]

Each of these sites spans the two ribosomal subunits and thereby functionally links the decoding center of the small subunit with the peptidyl transferase center of the large subunit. The anticodon loop at one end of each L-shaped tRNA molecule makes contact with the mRNA positioned in the small-subunit decoding site, and the aminoacylated 3' end of the tRNA occupies the peptidyl transferase center in the large subunit, 70 Å away.

During elongation, the ribosome houses a growing polypeptide chain, which is covalently linked to the tRNA in the P site, as the peptidyl-tRNA. Each time the ribosome shifts from one mRNA codon to the next, it makes room for a new aminoacyl-tRNA in the A site. The peptidyl transferase center of the ribosome catalyzes nucleophilic attack by the amino group of an incoming aminoacyl-tRNA on the terminal carbonyl group of the growing polypeptide. After the peptide bond forms, the polypeptide chain transfers from one tRNA to the other, with the aminoacyl-tRNA becoming the peptidyl-tRNA as the growing polypeptide chain is added to it.

The rate of peptide bond formation is enhanced on the ribosome largely because the 3' ends of the two reacting tRNAs are positioned optimally for the chemical reaction to occur (Figure 18-9a). Note that there is no accompanying hydrolysis of a high-energy bond, such as that of a nucleoside triphosphate, at this stage of protein synthesis. This is because each amino acid has



**FIGURE 18-9 Alignment of tRNAs on mRNA and the peptidyl transferase site.** (a) The peptidyl-tRNA and the aminoacyl-tRNA are positioned so that the amino acids are at the optimal distance for peptide bond formation. (b) The anticodons kink the mRNA. [Sources: PDB ID 1GIX and PDB ID 2OW8.]

already been activated by its attachment to tRNA in an aminoacylation reaction. As we'll see, the aminoacylation step requires ATP hydrolysis, so the energy cost has already been paid (see Section 18.2). Thus, each acyl group provides the high-energy bond that is hydrolyzed to drive the formation of a new peptide bond during the peptidyl transferase reaction.

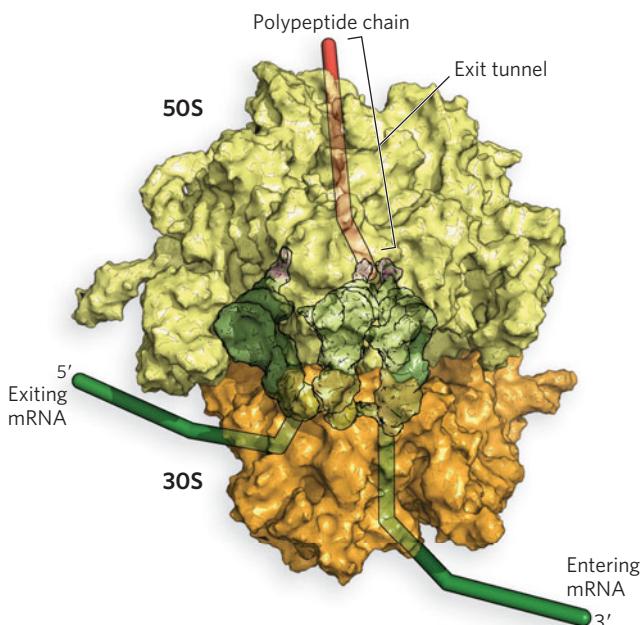
Crystal structures of the ribosome revealed the presence of channels for the movement of mRNAs and polypeptides during protein synthesis. In the small subunit, the mRNA entry and exit channels are narrow clefts with space to accommodate only a single strand of RNA. Thus, an mRNA must be unfolded and available for base pairing to tRNAs when it enters the decoding center. The A and P sites in the small subunit provide room for tRNAs to base-pair with two adjacent codons in the mRNA sequence. Here, the ribosome induces a kink in the mRNA between the two codons to help ensure accuracy of the reading frame (see Figure 18.9b). This kink may also contribute to correct positioning of the aminoacylated ends of the tRNA in the peptidyl transferase center of the large subunit.

In the large subunit, a tunnel (50 Å long) allows the exit of polypeptides during translation (Figure 18.10). The diameter of the exit tunnel seems to preclude formation of structures wider than an  $\alpha$  helix as the

polypeptide is being synthesized. Thus, nascent proteins lack tertiary structure and must assemble into their correct three-dimensional structure after exiting the ribosome.

## SECTION 18.1 SUMMARY

- Protein synthesis occurs on ribosomes, large complexes of protein and rRNA. Bacteria have 70S ribosomes, with a large (50S) and a small (30S) subunit. Eukaryotic ribosomes are significantly larger, 80S (with 60S and 40S subunits), and contain more proteins.
- Protein synthesis is regulated by the ability of the small ribosomal subunit to associate independently with the mRNA before translation begins. Once the ribosome is fully assembled, it moves along the mRNA, matching a tRNA to each codon and catalyzing peptide bond formation. Multiple ribosomes can occupy a single mRNA, forming a polysome.
- In all organisms, the rRNA of the large ribosomal subunit catalyzes peptide bond formation, and the small ribosomal subunit reads the genetic code and ensures that the correct amino acid is added to the growing polypeptide chain.
- Aminoacyl-tRNAs first bind to the A site of the ribosome. As the peptide bond is formed between the amino acid and the growing peptide chain, the newly formed peptidyl-tRNA moves to the P site and the free tRNA exits through the E site.
- The mRNA and growing polypeptide pass through separate channels in the ribosome that require them to be unfolded. The ribosome induces a kink in the mRNA between the A and P sites to allow base-pairing of the tRNAs. Polypeptides form their functional three-dimensional structure after emerging from the ribosomal exit tunnel.



**FIGURE 18.10** The exit tunnel for protein in the 50S subunit.

**subunit.** The protein exit tunnel is adjacent to the P site and only wide enough to allow an unfolded polypeptide to pass through. [Source: (b) PDB ID 1VSA, 2OW8, 1GIX.]

## 18.2 Activation of Amino Acids for Protein Synthesis

For the synthesis of a polypeptide with a sequence defined by an mRNA, two chemical requirements must be met: (1) the carboxyl group of each amino acid must be activated to facilitate the formation of a peptide bond, and (2) a link must be established between each new amino acid and the information in the mRNA that encodes it. Both of these are satisfied by covalent attachment of the amino acid to a tRNA prior to protein

synthesis. Attaching each amino acid to its corresponding tRNA is critical. This reaction takes place in the cytosol, not on the ribosome. Each of the 20 amino acids is covalently linked to a specific tRNA at the expense of ATP hydrolysis, catalyzed by  $Mg^{2+}$ -dependent activating enzymes known as aminoacyl-tRNA synthetases. When attached to an amino acid, a tRNA is said to be *charged*.

### Amino Acids Are Activated and Linked to Specific tRNAs

The charging of a tRNA requires the formation of an acyl linkage between the carboxyl group of an amino acid and the free 2'- or 3'-hydroxyl end of the tRNA. All tRNAs share a similar structure, including three or four arms and a 3'-terminal CCA sequence (see Figures 6-22 and 17-2). The acylation reaction results in attachment of the amino acid to the 3'-terminal adenosine.

Aminoacyl-tRNA synthetases must activate the amino acid before it is attached to the tRNA. This reaction occurs in two steps in the enzyme's active site. In the **adenylation step** (Figure 18-11), an enzyme-bound intermediate, 5'-aminoacyl adenylate (5'-aminoacyl-AMP), forms when the carboxyl group of the amino acid reacts with the  $\alpha$ -phosphoryl group of ATP to form a phosphoanhydride linkage, with displacement of pyrophosphate. In the subsequent **tRNA-charging step**, the aminoacyl group is transferred from enzyme-bound aminoacyl-AMP to its specific tRNA. The amino acid can be transferred to either the 2'-OH or the 3'-OH (left and right paths, respectively, in Figure 18-11) of the 3'-terminal adenosine of the tRNA, depending on the type of aminoacyl-tRNA synthetase. Class I synthetases attach the amino acid to the 2'-OH, and class II synthetases attach the amino acid to the 3'-OH. In the class I pathway, the aminoacyl ester linkage migrates to the 3'-OH position spontaneously by a transesterification reaction.

The resulting ester linkage between the amino acid and the tRNA has a highly negative standard free energy of hydrolysis ( $\Delta G^{\circ} = -29 \text{ kJ/mol}$ ). Because its hydrolysis is energetically favorable, this bond between the amino acid and the tRNA provides an energetic driving force for translation. The first part of the activation reaction separates two phosphates from ATP instead of just one (as in many other ATP-driven reactions). Ultimately, the energy stored in the bond between the two phosphates is released on hydrolysis by the enzyme inorganic pyrophosphatase. Thus, two high-energy phosphate bonds are ultimately expended for each amino acid molecule activated, rendering the overall reaction for amino acid activation essentially irreversible.

### Aminoacyl-tRNA Synthetases Attach the Correct Amino Acids to Their tRNAs

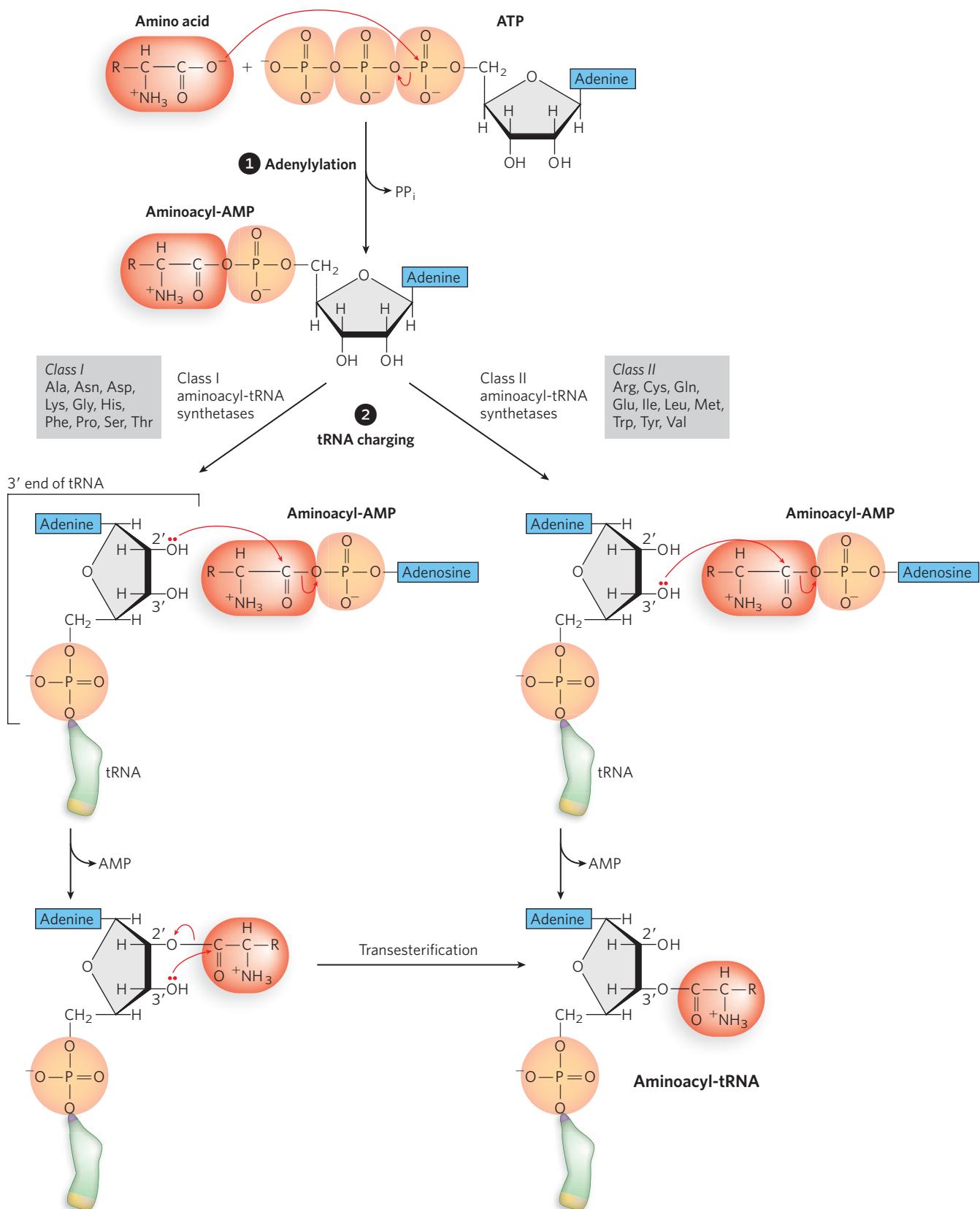
A distinct aminoacyl-tRNA synthetase is responsible for attaching each of the 20 amino acids to its corresponding tRNA. Each enzyme is specific for one amino acid but can recognize more than one tRNA, because most amino acids are specified by more than one codon (see Chapter 17). The structures of all the aminoacyl-tRNA synthetases of *E. coli* have been determined, and the structures reflect their mode of tRNA recognition and catalytic activity (Figure 18-12). Class I enzymes are typically monomeric and attach the amino acid to the 2'-OH of the 3'-terminal adenine of tRNA. As noted above, the aminoacyl group spontaneously migrates to the 3'-OH position. Class II aminoacyl-tRNA synthetases are sometimes multimeric, and they approach their tRNA substrate from a different side than the class I synthetases, typically attaching the amino acid to the 3'-OH. These two classes of aminoacyl-tRNA synthetases are the same in all organisms. There is no evidence for a common ancestor from which the two classes diverged, and the biological, chemical, or evolutionary reasons for two enzyme classes for essentially identical processes remain obscure.

There are exceptions to the rule of one aminoacyl-tRNA synthetase for one amino acid. Because the synthetase sequences are evolutionarily related, they are usually readily identified in genomic sequence data, and the absence of a gene encoding the aminoacyl-tRNA synthetase specific for glutamine in some bacteria was puzzling. Biochemical and genetic experiments revealed that in these bacteria, a single enzyme charges both tRNA<sup>Gln</sup> and tRNA<sup>Glu</sup> with glutamate. A second enzyme then catalyzes an amination reaction that converts the glutamate of Glu-tRNA<sup>Gln</sup> to glutamine.

### The Structure of tRNA Allows Accurate Recognition by tRNA Synthetases

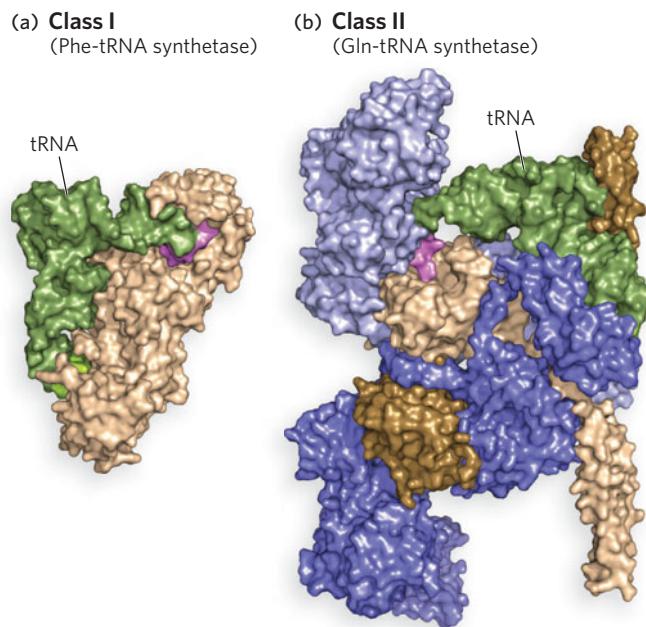
The overall fidelity of protein synthesis requires that each individual aminoacyl-tRNA synthetase must be specific for a single amino acid and for certain tRNAs. The interaction of aminoacyl-tRNA synthetases and tRNAs has been referred to as the “second genetic code,” reflecting its critical role in maintaining the accuracy of protein synthesis. The “coding” rules seem to be more complex than those of the “first” code.

By observing changes in nucleotides that alter substrate specificity, researchers have identified nucleotide positions that are involved in substrate discrimination by the aminoacyl-tRNA synthetases (Figure 18-13). Some nucleotides are conserved in all tRNAs and therefore cannot be used for discrimination. Although in some cases the nucleotides of the anticodon itself are

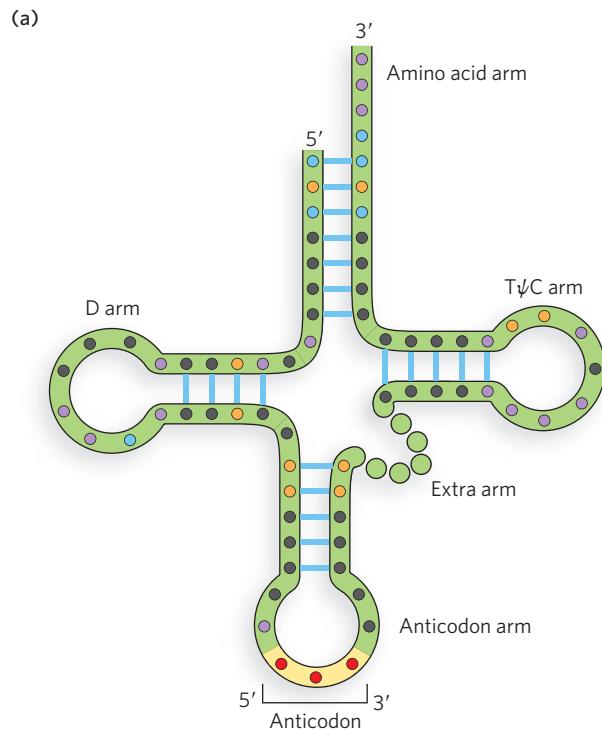


**FIGURE 18-11 Charging of tRNAs by aminoacyl-tRNA synthetases.** In step 1, the adenylylation step, the amino acid is linked to adenylate, forming aminoacyl-AMP. In step 2, the amino acid is transferred from the AMP to the tRNA in one of two pathways, catalyzed by a class I or class II aminoacyl-tRNA synthetase, as described in the text. The synthetases in each class (denoted by their amino acid) are listed.

of two pathways, catalyzed by a class I or class II aminoacyl-tRNA synthetase, as described in the text. The synthetases in each class (denoted by their amino acid) are listed.



**FIGURE 18-12** Crystal structures of aminoacyl-tRNA synthetases with bound tRNA. (a) Phe-tRNA synthetase is a class I and (b) Gln-tRNA synthetase is a class II aminoacyl-tRNA synthetase. For each structure, the bound tRNA is in green, the 3' end of the tRNA to which the amino acid attaches is in purple. [Sources: (a) PDB ID 1EUQ. (b) PDB ID 1EIY.]

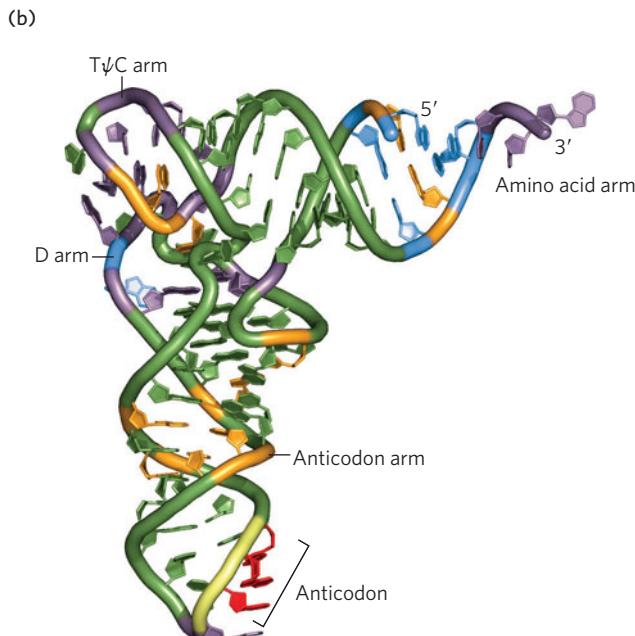


recognized, nucleotide positions conferring synthetase specificity tend to be concentrated in the amino acid arm and elsewhere in the anticodon arm, as well as other parts of the tRNA molecule. Determination of the crystal structures of aminoacyl-tRNA synthetases complexed with their cognate tRNAs and ATP has added a great deal to our understanding of these interactions.

Ten or more specific nucleotides may be involved in the recognition of a tRNA by its specific aminoacyl-tRNA synthetase. But in a few cases, the recognition mechanism is quite simple. For example, for alanyl-tRNA synthetase (or Ala-tRNA synthetase—a shorthand commonly used for these enzymes), across a range of organisms from bacteria to humans, the primary determinant of recognition of tRNA<sup>Ala</sup> is a single G-U base pair in its amino acid arm. A short RNA with as few as 7 bp arranged in a simple hairpin mini-helix is efficiently aminoacylated by the Ala-tRNA synthetase, as long as the RNA contains the critical G-U.

### Proofreading Ensures the Fidelity of Aminoacyl-tRNA Synthetases

The aminoacylation of tRNA both activates an amino acid for peptide bond formation and appends the amino acid to an adaptor tRNA that ensures appropriate



**FIGURE 18-13** Sequence features of a tRNA that are recognized by aminoacyl-tRNA synthetases. (a) The two-dimensional and (b) three-dimensional structures of tRNA, showing the nucleotides recognized by only one (orange

residues) or by two or more (blue residues) aminoacyl-tRNA synthetases. Nucleotides shown in purple are common to all tRNAs and therefore cannot be used to distinguish among them. [Source: (b) PDB ID 1EHZ.]

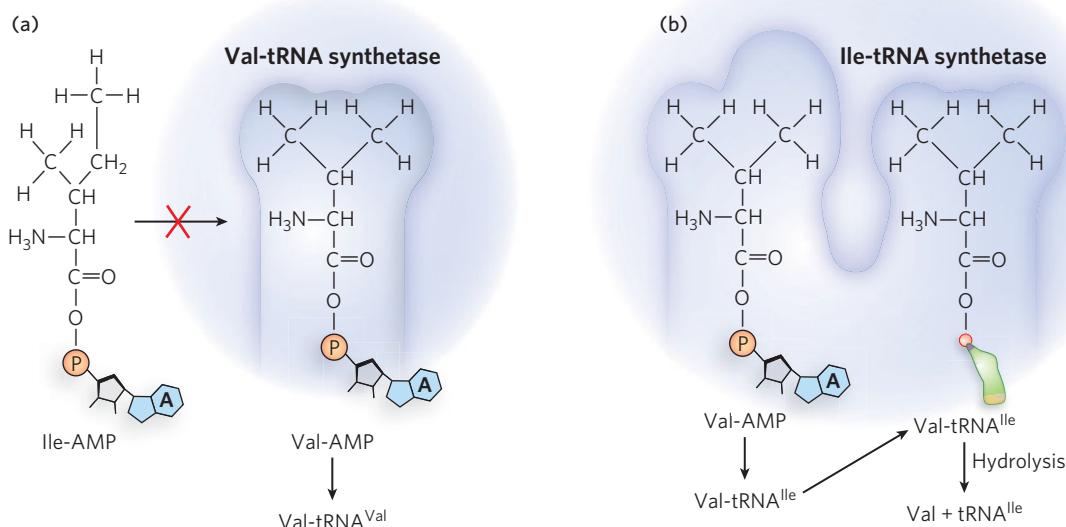
placement of the amino acid in a growing polypeptide. The identity of the amino acid attached to a tRNA is not checked on the ribosome, however, so attachment of the correct amino acid to the tRNA is essential to the fidelity of protein synthesis.

Discrimination between two similar amino acid substrates has been studied in detail in the case of Ile-tRNA synthetase, which must distinguish between valine and isoleucine, amino acids that differ by just a single methylene group ( $-\text{CH}_2-$ ). Because valine is smaller, Val-tRNA synthetase can discriminate between valine and isoleucine by having a binding pocket too small for isoleucine to bind (Figure 18-14a). However, the correlating strategy cannot work for Ile-tRNA synthetase, because the small valine molecule can fit in the big isoleucine pocket. Instead, Ile-tRNA synthetase must rely on the energetics of substrate binding and proofreading. Ile-tRNA synthetase favors activation of isoleucine (to form Ile-AMP) over valine by a factor of 200—as we might expect, given the amount by which a methylene group (in Ile) could enhance substrate binding. Yet a Val residue is erroneously incorporated into proteins in positions normally occupied by an Ile residue at a frequency of only about 1 in 3,000. How is this more than tenfold increase in accuracy brought about? Ile-tRNA synthetase, like some other aminoacyl-tRNA synthetases, has a proofreading function.

A general principle of proofreading by enzymes is that if available binding interactions do not provide sufficient discrimination between two substrates, the neces-

sary specificity can be achieved by substrate-specific binding in *two successive steps*. The effect of forcing the system through two consecutive filters is multiplicative. In the case of Ile-tRNA synthetase, the first filter is the initial binding and activation of the amino acid to form the aminoacyl-AMP, followed by transfer of the amino acid to the tRNA. The second is the binding of any *incorrect* aminoacyl-tRNA products to a separate active site on the enzyme; a substrate that binds in this second active site is hydrolyzed (Figure 18-14b). Because the R group of valine is slightly smaller than that of isoleucine, Val-tRNA fits the hydrolytic (proofreading) site of the Ile-tRNA synthetase but Ile-tRNA does not. Thus, only Val-tRNA is hydrolyzed in the proofreading active site.

The greatly accelerated rate of hydrolysis of incorrectly charged tRNAs provides an important mechanism of enhancing the fidelity of the overall process. This is an example of **kinetic proofreading**, in which a complex process occurs in multiple steps, the rates of which are tuned to maximize the speed of correct reactions while stalling and reversing incorrect reactions. Note that this proofreading mechanism requires energy, because it requires a second round of aminoacylation (with the correct amino acid). The few aminoacyl-tRNA synthetases that activate amino acids with no close structural relatives (e.g., Cys-tRNA synthetase) demonstrate little or no proofreading activity; in these cases, the active site for aminoacylation can sufficiently discriminate between the proper substrate and any other, incorrect amino acid.



**FIGURE 18-14** Proofreading by an aminoacyl-tRNA synthetase. (a) tRNA<sup>Val</sup> is charged in the acylation site of the Val-tRNA synthetase. Isoleucine is larger than valine and does not fit in the acylation site. (b) tRNA<sup>Ile</sup> is charged in the acylation site of the Ile-tRNA synthetase. Because valine is

smaller than isoleucine, however, tRNA<sup>Ile</sup> is sometimes charged with valine. This incorrectly charged Val-tRNA<sup>Ile</sup> fits into the synthetase's proofreading site, where it is hydrolyzed to release valine from the tRNA.

In summary, translation relies on aminoacyl-tRNA synthetases to ensure the correct charging of tRNAs, because the ribosome does not distinguish between correctly and incorrectly charged tRNAs during protein synthesis. The decoding center of the ribosome is designed to detect and favor proper codon-anticodon base pairing, but does not link this information to the identity of the amino acid in the peptidyl transferase center at the other end of the tRNA (see How We Know). As a result, the overall error rate of protein synthesis (~1 mistake per  $10^4$  amino acids incorporated) is significantly higher than that of DNA replication. (Modern protein technology has been designed to exploit this lack of proofreading so that ribosomes can incorporate synthetic amino acids into proteins, as described in Highlight 18-2.) Flaws in a protein are eliminated when the protein is degraded and are not passed on to future generations, so they have less biological significance than errors in DNA. The degree of fidelity in protein synthesis is sufficient to ensure that most proteins contain no mistakes and that the large amount of energy required to synthesize a protein is rarely wasted. One defective protein molecule is usually unimportant when the cell contains many correct copies of the same protein.

## SECTION 18.2 SUMMARY

- Aminoacyl-tRNA synthetases covalently link amino acids to tRNAs to create the substrates used by the ribosome during protein synthesis. The acylation reaction results in a high-energy bond that supplies the energetic driving force for translation.
- A different aminoacyl-tRNA synthetase exists for each amino acid. Because multiple codons can specify a single amino acid, most aminoacyl-tRNA synthetases can recognize multiple tRNAs bearing anticodons complementary to the codons for a particular amino acid.
- The anticodon is responsible for the specificity of interaction between the aminoacyl-tRNA and the complementary mRNA codon, but it rarely provides the primary site for synthetase recognition.
- Two successive binding steps allow kinetic proofreading that increases the fidelity of tRNA aminoacylation. This is essential because ribosomes do not discriminate between correctly and incorrectly charged tRNAs during protein synthesis.

## 18.3 Initiation of Protein Synthesis

Having described the components of the translation machinery, we now turn to a detailed discussion of the initiation stage of protein synthesis—the most highly

regulated step of translation. Translation **initiation** includes recruitment of the small ribosomal subunit to the mRNA; identification of the **initiation codon**, or **start codon**; association of the charged initiator tRNA with the mRNA; and recruitment of the large ribosomal subunit to form an active ribosome (Figure 18-15). Each step of protein synthesis, in both *E. coli* and eukaryotes, requires several protein factors to facilitate the reaction (Table 18-2). **Initiation factors** (denoted IF in bacteria and eIF in eukaryotes) are critical to enhancing the rate and fidelity of all steps in the process, fine-tuning the underlying rRNA-based activities and the interactions among mRNA, tRNA, and rRNA. Other steps in protein synthesis are generally the same in bacteria and eukaryotes, but some aspects of initiation differ across these two groups.

### Base Pairing Recruits the Small Ribosomal Subunit to Bacterial mRNAs

Translation in all organisms begins with binding of the ribosome's small subunit to an mRNA. In bacteria, the initiating 5'-AUG is guided to its correct position on the ribosome by the **Shine-Dalgarno sequence** (named for Australian researchers John Shine and Lynn Dalgarno, who identified it), also called the **ribosome-binding site** (RBS). This consensus sequence is an initiation signal of 4 to 9 purine residues, situated 8 to 13 nucleotides on the 5' side of the start codon (Figure 18-16a). The sequence base-pairs with a complementary pyrimidine-rich sequence near the 3' end of the 16S rRNA of the 30S ribosomal subunit. This mRNA-rRNA interaction positions the initiating 5'-AUG sequence of the mRNA in the precise location on the 30S subunit where it is required for translation initiation. Higher degrees of base-pairing complementarity and optimal spacing between the Shine-Dalgarno sequence and the initiation codon increase the efficiency of translation of a particular mRNA. The specific 5'-AUG where the initiator aminoacyl-tRNA—a special type of methionyl-tRNA—is to be bound is distinguished from other methionine AUG codons by its proximity to the Shine-Dalgarno sequence.

The Shine-Dalgarno sequence can be used to initiate synthesis of more than one protein, if they are encoded in a single transcript called a polycistronic mRNA. In some bacterial polycistronic mRNAs, the open reading frames overlap (Figure 18-16b, bottom). Despite lacking a Shine-Dalgarno sequence for each internal start site, the internal open reading frames can be translated efficiently because of overlapping start and stop codons, typically 5'-AUGA. Ribosomes terminating translation of the upstream message can initiate the downstream message simply by shifting their reading frame.

## HIGHLIGHT 18-2 TECHNOLOGY

### Genetic Incorporation of Unnatural Amino Acids into Proteins

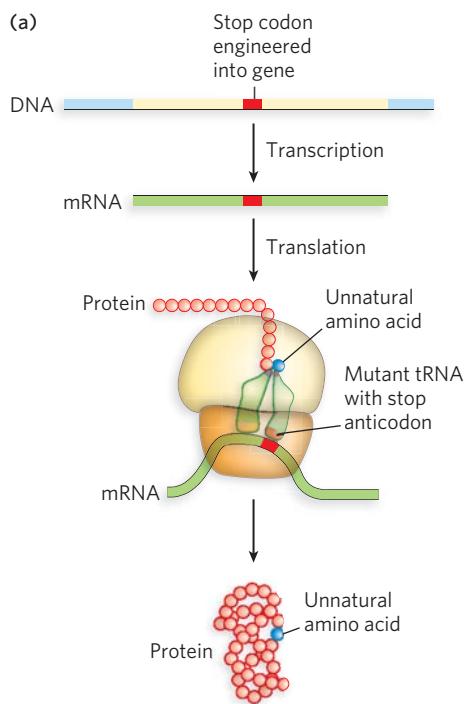
Peter Schultz and his colleagues at the Scripps Research Institute wondered whether the specificity of aminoacyl-tRNA synthetases together with the lack of discrimination of mischarged tRNAs by ribosomes could be exploited to incorporate unnatural amino acids into cellular proteins. First, they introduced stop codons at sites in an mRNA where they hoped to introduce an unnatural amino acid in the corre-

sponding polypeptide. Next, they engineered suppressor tRNAs (see Chapter 17) with an anticodon sequence complementary to the stop codon. Taking advantage of the known structural determinants of the synthetases' recognition of tRNA, the researchers designed suppressor tRNAs to be charged by aminoacyl-tRNA synthetases with mutations that caused them to use unnatural amino acid substrates (Figure 1). Schultz's research group showed that bacterial, yeast, and mammalian cells containing the engineered mRNAs, tRNAs, and aminoacyl-tRNA synthetases, and unnatural amino acids, would produce proteins with the unnatural amino acid residues at the planned positions.

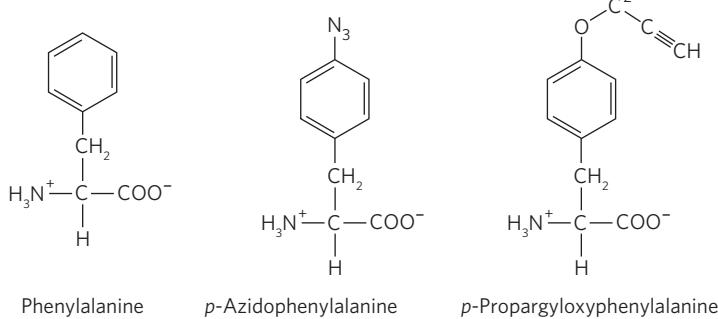


**Peter Schultz** [Source: Courtesy of Peter Schultz.]

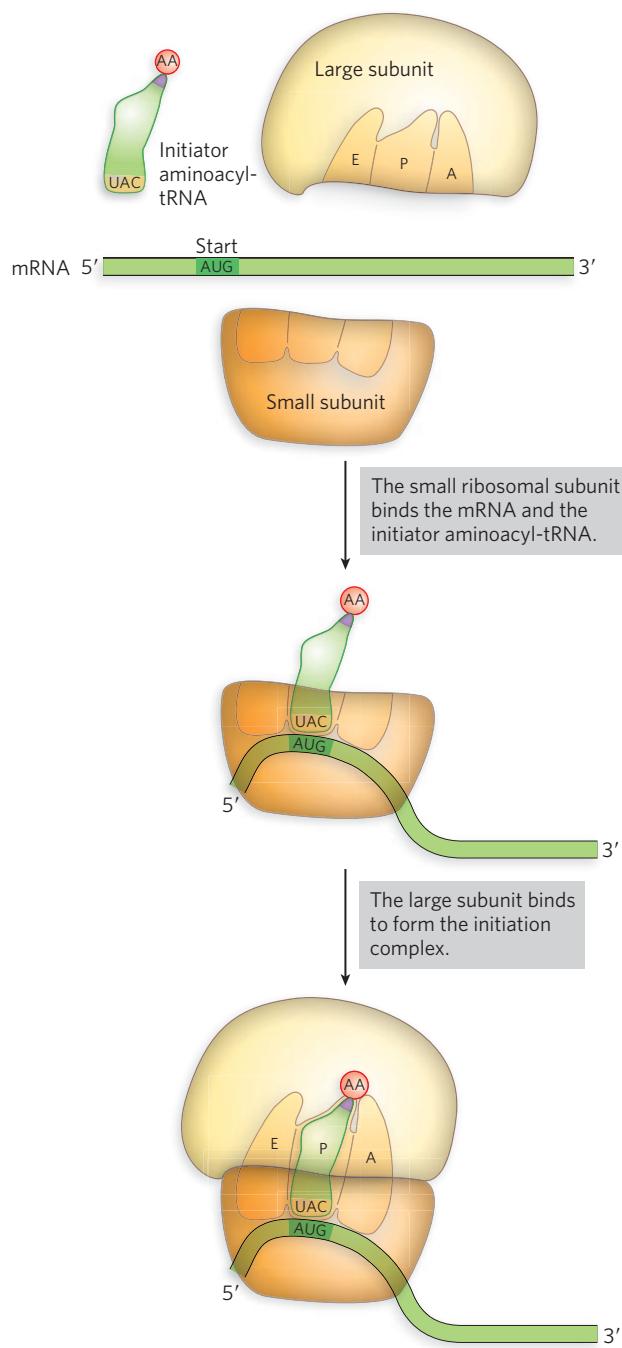
This method enables the incorporation into proteins of fluorescent, glycosylated, sulfated, metal ion-binding, and redox-active (i.e., electron-transferring) amino acids, as well as amino acids with new chemical and photochemical reactivity. Why might such technology be useful? Schultz hopes to use the approach to explore protein structure and function, both *in vitro* and *in vivo*, by incorporating chemical tags or probes that can report on their local molecular/structural environment or provide molecular "beacons" in cells. Furthermore, with the ability to generate proteins with new or enhanced properties by making use of amino acids not found in nature, it may eventually be possible to enable cells to synthesize therapeutic proteins.



#### (b) Some Unnatural Amino Acids



**FIGURE 1** (a) An mRNA stop codon and a mutated tRNA (with a stop anticodon) are used to introduce unnatural amino acids into proteins. (b) Two unnatural amino acids, derivatives of phenylalanine (which is also shown for comparison), that can be incorporated by this method.



**FIGURE 18-15** An overview of the events in translation initiation.

### Eukaryotic mRNAs Recruit the Small Ribosomal Subunit Indirectly

The 5' cap and poly(A) tail modifications on eukaryotic mRNAs serve three purposes: they protect the ends from degradation, facilitate nuclear export of the mRNA, and promote translation by binding initiation factors that form a link between the mRNA and the ribosome. The 5' terminus of the mRNA contains a

modified G residue, the 5' cap, that can be bound by the cap-binding protein, eIF4E, which in turn binds other proteins that recruit the small ribosomal subunit to the mRNA. Once associated with the 5' end of the mRNA, the small subunit locates the 5'-AUG start codon by sampling the RNA in the 5'-3' direction.

In addition to the 5' cap, the presence of a purine nucleotide three residues before the start codon and a G residue immediately following the start codon is thought to enhance translation through contact with the initiator tRNA. This **Kozak sequence** was discovered by Marilyn Kozak during analysis of the sequence features of eukaryotic mRNAs that increased translation efficiency (Table 18-3). At the 3' terminus of eukaryotic mRNAs, the poly(A) tail stimulates translation efficiency by fostering reinitiation after completion of a polypeptide chain.

### A Specific Amino Acid Initiates Protein Synthesis

Protein synthesis begins at the N-terminal end and proceeds by the stepwise addition of amino acids to the C-terminal end of the growing polypeptide, as found by Howard Dintzis in 1961 (see Chapter 17, How We Know). The 5'-AUG initiation codon specifies an N-terminal Met residue. Although methionine has only one codon, AUG, all organisms have two tRNAs for methionine. One is used exclusively when 5'-AUG is the initiation codon for protein synthesis; the other is used to code for a Met residue in an internal position in a polypeptide. An initiation factor specifically binds to the initiator tRNA and delivers it to the ribosome, thereby enabling cells to separate translation initiation from elongation.

The distinction between an initiating 5'-AUG and an internal one is straightforward. In bacteria, the two types of tRNA specific for methionine are designated tRNA<sup>Met</sup> and tRNA<sup>fMet</sup>. The amino acid incorporated in response to the 5'-AUG initiation codon is *N*-formylmethionine (fMet). It arrives at the ribosome as **N-formylmethionyl-tRNA<sup>fMet</sup>** (fMet-tRNA<sup>fMet</sup>), which is formed in two successive reactions. First, methionine is attached to tRNA<sup>fMet</sup> by the Met-tRNA synthetase (which in *E. coli* aminoacylates both tRNA<sup>fMet</sup> and tRNA<sup>Met</sup>), in a reaction such as that shown in Figure 18-11. Second, a transformylase enzyme transfers a formyl group from *N*<sup>10</sup>-formyltetrahydrofolate to the amino group of the methionyl moiety (Figure 18-17). The transformylase is more selective than the synthetase: it is specific for methionyl moieties attached to tRNA<sup>fMet</sup>, presumably recognizing some unique structural feature of that tRNA. Addition of the *N*-formyl group to methionine prevents fMet from

**Table 18-2 Essential Components of the Main Stages of Protein Synthesis in *E. coli* and Eukaryotes**

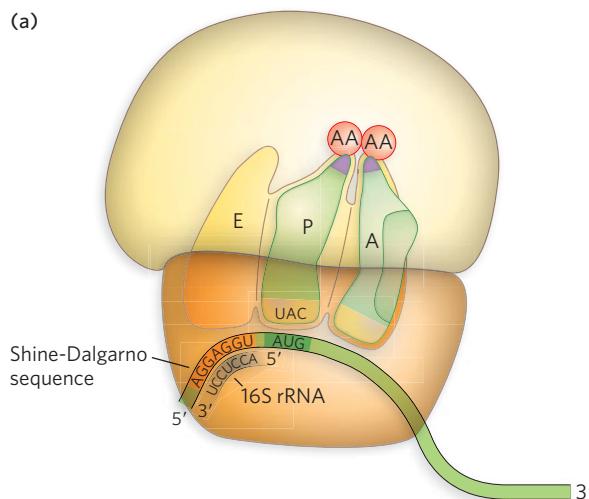
Stage	<i>E. coli</i>	Eukaryotes
1. Activation of amino acids	20 amino acids 20 aminoacyl-tRNA synthetases 32 or more tRNAs ATP $Mg^{2+}$	20 amino acids 20 aminoacyl-tRNA synthetases 32 or more tRNAs ATP $Mg^{2+}$
2. Initiation	mRNA <i>N</i> -Formylmethionyl-tRNA <sup>fmet</sup> Start codon in mRNA (AUG) 30S ribosomal subunit 50S ribosomal subunit Initiation factors (IF-1, IF-2, IF-3) GTP $Mg^{2+}$	Methionyl-tRNA <sup>met</sup> Start codon in mRNA (AUG) 40S ribosomal subunit 60S ribosomal subunit Initiation factors (eIF1, eIF1A, eIF2, eIF3, eIF4B, eIF4F (complex of eIF4E, eIF4A, eIF4G), eIF5, eIF5B) GTP $Mg^{2+}$
3. Elongation	Functional 70S ribosome (initiation complex) Aminoacyl-tRNAs specified by codons Elongation factors (EF-Tu, EF-Ts, EF-G) GTP $Mg^{2+}$	Functional 80S ribosome (initiation complex) Aminoacyl-tRNAs specified by codons Elongation factors (eEF1 $\alpha$ , eEF1 $\beta\gamma$ , eEF2) GTP $Mg^{2+}$
4. Termination and release	Stop codon in mRNA Release factors (RF-1, RF-2, RF-3) EF-G IF-3	Stop codon in mRNA Release factors (eRF1, eRF3)
5. Folding and posttranslational processing	Specific enzymes, cofactors, and other components for removal of initiating residues and signal sequences, additional proteolytic processing, modification of terminal residues, and attachment of phosphate, acetyl, methyl, carboxyl, carbohydrate, or prosthetic groups	Specific enzymes, cofactors, and other components for removal of initiating residues and signal sequences, additional proteolytic processing, modification of terminal residues, and attachment of phosphate, acetyl, methyl, carboxyl, carbohydrate, or prosthetic groups

entering interior positions in a polypeptide (only Met-tRNA<sup>Met</sup> inserts methionine in interior positions) while allowing fMet-tRNA<sup>fMet</sup> to bind a specific ribosomal initiation site that accepts neither Met-tRNA<sup>Met</sup> nor any other aminoacyl-tRNA.

In eukaryotic cells, all polypeptides synthesized by cytosolic ribosomes begin with a Met residue (rather than fMet); however, as in bacteria, the cell uses a special initiator tRNA, tRNA<sub>i</sub><sup>Met</sup>, distinct from the tRNA<sup>Met</sup> used at AUG codons in interior positions in the mRNA. The Met-tRNA<sub>i</sub><sup>Met</sup> is distinct because it has a specific sequence in the anticodon arm that is recognized by an

initiation factor protein. Polypeptides synthesized by mitochondrial and chloroplast ribosomes, however, begin with *N*-formylmethionine. This strongly supports the view that mitochondria and chloroplasts originated from bacterial ancestors symbiotically incorporated into precursor eukaryotic cells at an early stage of evolution.

In bacteria, the enzyme deformylase typically removes the formyl group from the N-terminal Met residue during or shortly after production of the polypeptide. In both bacteria and eukaryotes, enzymes called aminopeptidases frequently remove the entire



(b) Non-overlapping genes

Shine-Dalgarno sequence	Protein-coding region 1	Protein-coding region 2
UUUGAGGAGGUACGUACUAC	AUGGCUGA AUCGUUAACGGGAGGAGGU	UGGGAA AUGAAGCC AGCAAUAGCUGACGUACA

Overlapping genes

Shine-Dalgarno sequence	Protein-coding region 1	Protein-coding region 2
UUUGAGGAGGUACGUACUAC	AUGGCUGA GGGAA AUGAAGCC	UUUGGUACGUACUAC AGGAGGUACGUACUAC UGGGAA AUGAAGCC CUGGUAGCUGACGUACA

**TABLE 18-3 The Kozak Sequence**

Organism(s)	Consensus Sequence*
Vertebrates	GCCRCC <b>ATGG</b>
Terrestrial plants	AACA <b>ATGGC</b>
<i>Drosophila melanogaster</i> (fruit fly)	CAAA <b>ATG</b>
<i>Saccharomyces cerevisiae</i> (baker's yeast)	AAAAAA <b>ATG</b> TCT
<i>Dictyostelium discoideum</i> (slime mold)	AAAAAA <b>ATG</b> RNA
<i>Plasmodium</i> spp. (malarial protozoa)	TAAAAAA <b>ATGAAN</b>

\*R = purine; N = any base.

N-terminal methionine, and sometimes one or two additional amino acids, from newly synthesized polypeptides. Thus, many mature proteins do not have a Met residue at their N-terminal end.

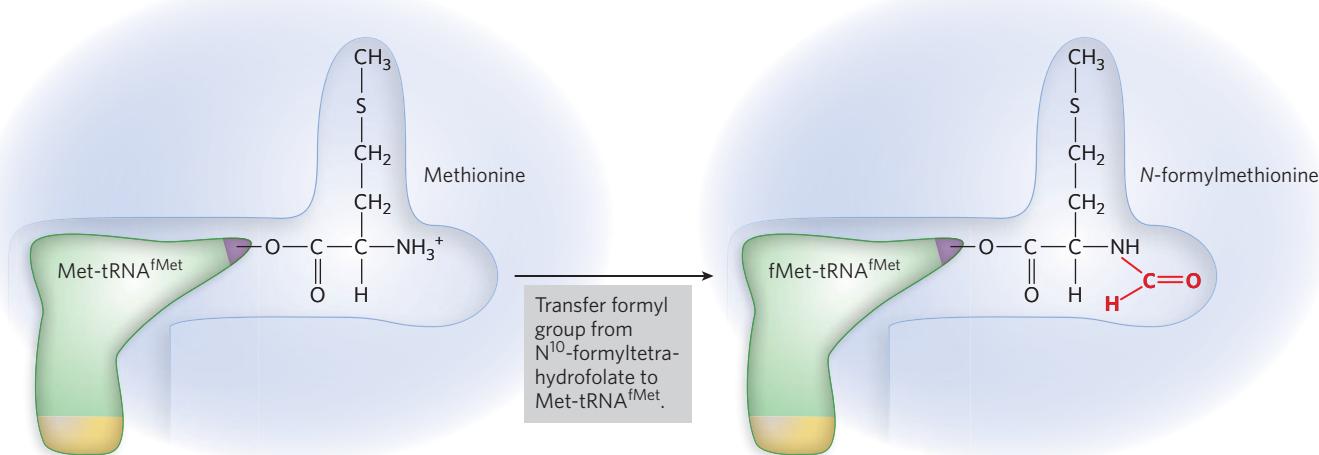
### Initiation in Bacterial Cells Requires Three Initiation Factors

As discussed earlier, ribosomes have three sites that bind aminoacyl-tRNAs: the aminoacyl (A) site, the peptidyl (P) site, and the exit (E) site. In addition to the

**FIGURE 18-16 The Shine-Dalgarno sequence.** (a) The Shine-Dalgarno sequence positions the mRNA on the bacterial ribosome by binding to the 16S rRNA. (b) In bacteria, some polycistronic genes have overlapping start and stop codons (bottom), allowing the ribosome to initiate synthesis in the absence of a Shine-Dalgarno sequence.

30S subunit, an mRNA, and an initiating fMet-tRNA<sup>fMet</sup>, the initiation of polypeptide synthesis in bacteria requires a set of three initiation factor proteins known as **IF-1**, **IF-2**, and **IF-3**. Each plays a specific role in assembling the small ribosomal subunit, with the mRNA and initiator tRNA in place, and the large ribosomal subunit in a process controlled by GTP hydrolysis.

Formation of the initiation complex, consisting of an active 70S ribosome, takes place in three steps, as shown in **Figure 18-18**. In step 1, the 30S subunit binds two initiation factors, IF-1 and IF-3. IF-3 prevents premature combination of the 30S and 50S subunits. IF-1 binds at the A site and blocks tRNA binding during initiation. The mRNA then binds to the 30S subunit through base pairing of the Shine-Dalgarno sequence with 16S rRNA. This short mRNA-rRNA helix is bound in a cleft in the 30S subunit, which precisely positions the mRNA adjacent to the P site, thus accounting for the accuracy of start codon selection. The initiating 5'-AUG is now positioned at the P site. This is the only site to which fMet-tRNA<sup>fMet</sup> can bind, and fMet-tRNA<sup>fMet</sup> is the only aminoacyl-tRNA that can bind first to the P site. During the subsequent elongation stage, all other incoming aminoacyl-tRNAs (including the so-called elongator Met-tRNA<sup>Met</sup> that binds interior AUG codons) bind first to the A site and only subsequently transfer to the P and E sites.



**FIGURE 18-17 Formation of fMet-tRNA<sup>fMet</sup>.** The tRNA<sup>fMet</sup> is first charged with methionine (not shown), then the Met is converted to fMet by methionyl-tRNA formyltransferase (transformylase).

In step 2, the complex consisting of the 30S ribosomal subunit, IF-1, IF-3, and mRNA is joined by both GTP-bound IF-2 and the initiating fMet-tRNA<sup>fMet</sup>. The anticodon of this tRNA now pairs with the mRNA's initiation codon in the P site. X-ray crystallography has revealed contacts between rRNA bases and three G≡C base pairs in the anticodon arm of initiator, but not elongator, Met-tRNAs. This observation suggests a mechanism by which initiator tRNA is favored in the P site during the initiation process. In step 3, a conformational change in the 30S subunit triggers the release of IF-3, enabling association with the 50S subunit; simultaneously, the GTP bound to IF-2 is hydrolyzed to GDP and P<sub>i</sub>, which are released from the complex. All three initiation factors depart from the ribosome at this point.

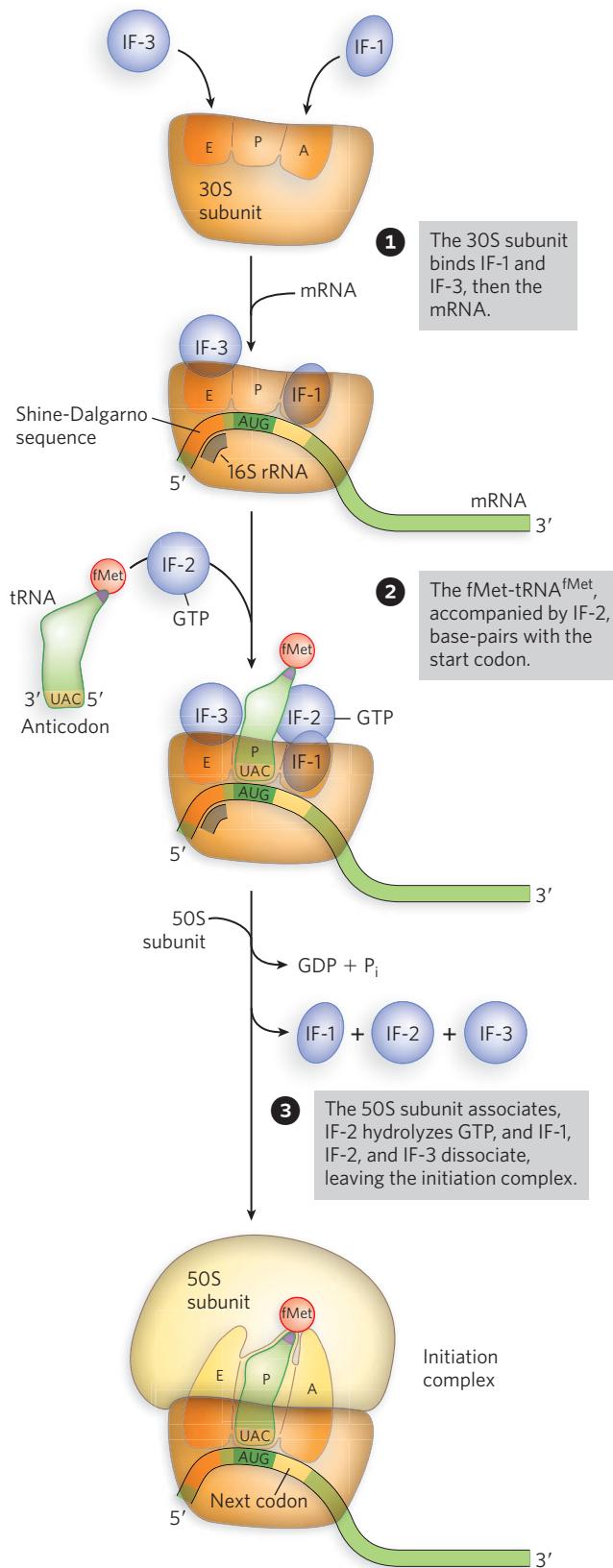
Completion of the steps in Figure 18-18 produces a functional 70S ribosome called the **initiation complex**, containing the mRNA and the initiating fMet-tRNA<sup>fMet</sup>. The correct location of fMet-tRNA<sup>fMet</sup> in the P site in the initiation complex is assured by at least three points of recognition and attachment: the codon-anticodon interaction involving the initiation 5'-AUG fixed in the small subunit portion of the P site, interaction between the Shine-Dalgarno sequence in the mRNA and the 16S rRNA of the small subunit, and binding interactions between the large-subunit portion of the P site and the fMet-tRNA<sup>fMet</sup>. The initiation complex is now ready for elongation.

### Initiation in Eukaryotic Cells Requires Additional Initiation Factors

Translation is generally similar in eukaryotic and bacterial cells; most of the significant differences are in the mechanism of initiation. Eukaryotic initiation requires, besides separate small and large ribosomal subunits, at

least 12 initiation factors and the binding and hydrolysis of ATP and GTP, as illustrated in Figure 18-19. Before translation begins, in step 1, the ribosomal subunits are separated by initiation factors eIF3 and eIF1A, the functional homologs of IF-3 and IF-1 in bacteria, preventing premature subunit joining and blocking initiator tRNA binding to the ribosomal A site, respectively. A third initiation factor, eIF1, binds to the E site. Meanwhile (step 2), the GTP-binding factor eIF2—containing three subunits, eIF2 $\alpha$ , eIF2 $\beta$ , and eIF2 $\gamma$ —associates with GTP and a charged initiator tRNA (Met-tRNA<sup>iMet</sup>) to form a ternary complex. An A=U base pair near the amino acid-binding site of the amino acid arm of Met-tRNA<sup>iMet</sup>, but not present in elongator Met-tRNA<sup>Met</sup>, is critical for binding to eIF2-GTP. Two other proteins, eIF5 (not shown) and eIF5B, which are involved in later steps of ribosomal assembly, also associate at this point. Three factors, eIF1, eIF1A and eIF3, mediate interaction between the ternary complex and the 40S subunit to form a 43S preinitiation complex.

Binding of the 43S preinitiation complex to an mRNA (step 3) is mediated by a complex called eIF4F. It contains eIF4E, a factor that binds the 5' cap; eIF4A, an ATPase and RNA helicase; and eIF4G, which binds both eIF4E and eIF3 to provide a link between the 43S complex and the mRNA. The eIF4G also binds to poly(A) binding protein (PABP), which is associated with the 3' poly(A) tail of the mRNA, and this eIF4G-PABP association brings the 5' and 3' ends of the mRNA together (Figure 18-20). Circularization of the eukaryotic mRNA by the eIF4G-PABP interaction facilitates the translational regulation of gene expression, as well (see Chapter 21). Following association of the eIF4F complex, another initiation factor, eIF4B, binds (not shown in Figure 18-19); its function is less clear.



**FIGURE 18-18 Translation initiation in bacteria.** Initiation occurs in three steps, as described in the text. Initiation factors IF-1 and IF-3 prevent premature binding of the elongation aminoacyl-tRNAs (i.e., those used at the elongation stage) and of the 50S subunit, respectively. IF-2 accompanies fMet-tRNA<sup>fMet</sup> to the initiation site.

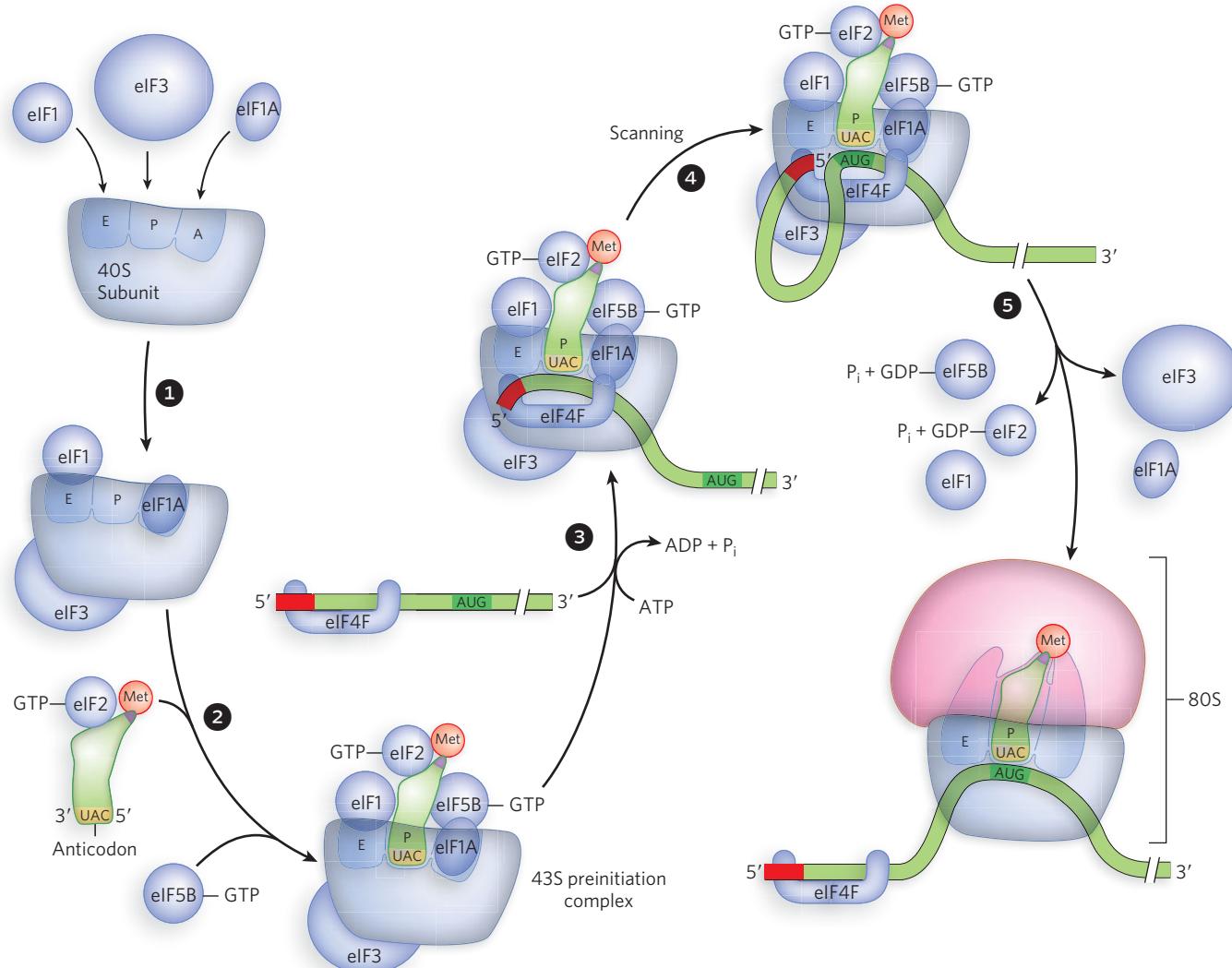
This larger complex, containing the 43S components, eIF4F, and mRNA, is a stable 48S particle. The 48S particle is capable of **scanning**, sliding along the mRNA in search of a start codon. Once assembled on the 5' end of the mRNA (step 4 in Figure 18-19), the 43S complex scans along the mRNA in the 5'→3' direction to the first AUG, which is recognized as the start site for translation initiation. The eIF4F complex is probably involved in this scanning process, perhaps using the RNA helicase activity of eIF4A to eliminate secondary structure in the 5' untranslated portion of the mRNA. Scanning is also facilitated by eIF4B, but the details of this process are unknown. Base pairing between the start codon and the anticodon of the initiator tRNA is primarily responsible for identification of the initiation site. Factor eIF1 is also critical to proper AUG recognition by preventing stable ribosomal association with non-AUG codons.

When the 43S complex has located the AUG start codon in a bound mRNA, the complex joins the 60S subunit to form an active 80S ribosome (step 5 in Figure 18-19). Because eIF1, eIF1A, eIF2, and eIF3 occlude parts of the 40S subunit surface that must interface with the 60S subunit, they must be displaced before subunit joining. This process requires eIF5 and eIF5B. The GTPase-activating eIF5 is specific for eIF2, stimulating hydrolysis of the eIF2-bound GTP and thus reducing eIF2's affinity for the initiator tRNA. Finally, eIF5B, a ribosome-dependent GTPase homologous to bacterial IF-2, hydrolyzes GTP and triggers dissociation of eIF2-GDP and other initiation factors from the 40S subunit, with concomitant association of the 60S subunit.

The initiation complex is now complete. The efficiency of translation is affected by many properties of the mRNA and proteins in this complex, including the length of the 3' poly(A) tract (in most cases, longer is better).

### Some mRNAs Use 5' End-Independent Mechanisms of Initiation

Some viral and eukaryotic mRNAs lack a 5' cap, but still rely on the eukaryotic translation machinery to produce their proteins. They accomplish this with an RNA segment called an **internal ribosome entry site**



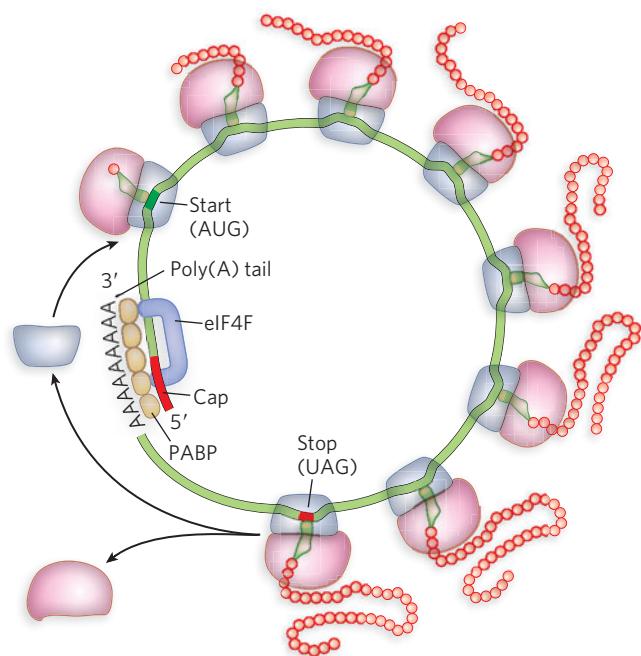
**FIGURE 18-19** Translation initiation in eukaryotes. Initiation is more complex in eukaryotes; the five steps are described in the text. Eukaryotic initiation factors (eIFs) promote the assembly of the 43S preinitiation complex with an mRNA and subsequent scanning of the mRNA to identify the start codon.

**(IRES)**, located on the 5' side of the start codon; it recruits the 40S subunit through direct interaction with the subunit or with eIF4F (Figure 18-21a). Cap-independent binding can also involve other proteins, such as La protein, which binds a pyrimidine-rich segment in the 5' region of the mRNA.

The first IRES was discovered in poliovirus, when researchers noticed that the viral mRNA is efficiently translated despite lacking a 5' cap. On infection, poliovirus produces a protease that cleaves the host cell's eIF4G into two fragments. This renders eIF4G useless for the host cell's protein synthesis, but does not compromise viral protein synthesis because the poliovirus IRES requires just one fragment of eIF4G to initiate

translation. In this way, the virus can simultaneously down-regulate host protein expression and maintain its own protein synthesis in infected cells. Different viral IRES subtypes seem to form distinct three-dimensional structures that enable direct interactions with the host translation machinery. For example, hepatitis C virus (HCV) mRNA has a pseudoknot adjacent to two stem-loop structures (Figure 18-21b). The resulting three-dimensional structure binds to the host cell's 80S ribosome such that the start codon is positioned in the P site (Figure 18-21c).

Some cellular mRNAs, although they contain a 5' cap, also contain IRES segments that enable translation under conditions that normally block initiation. Cellular

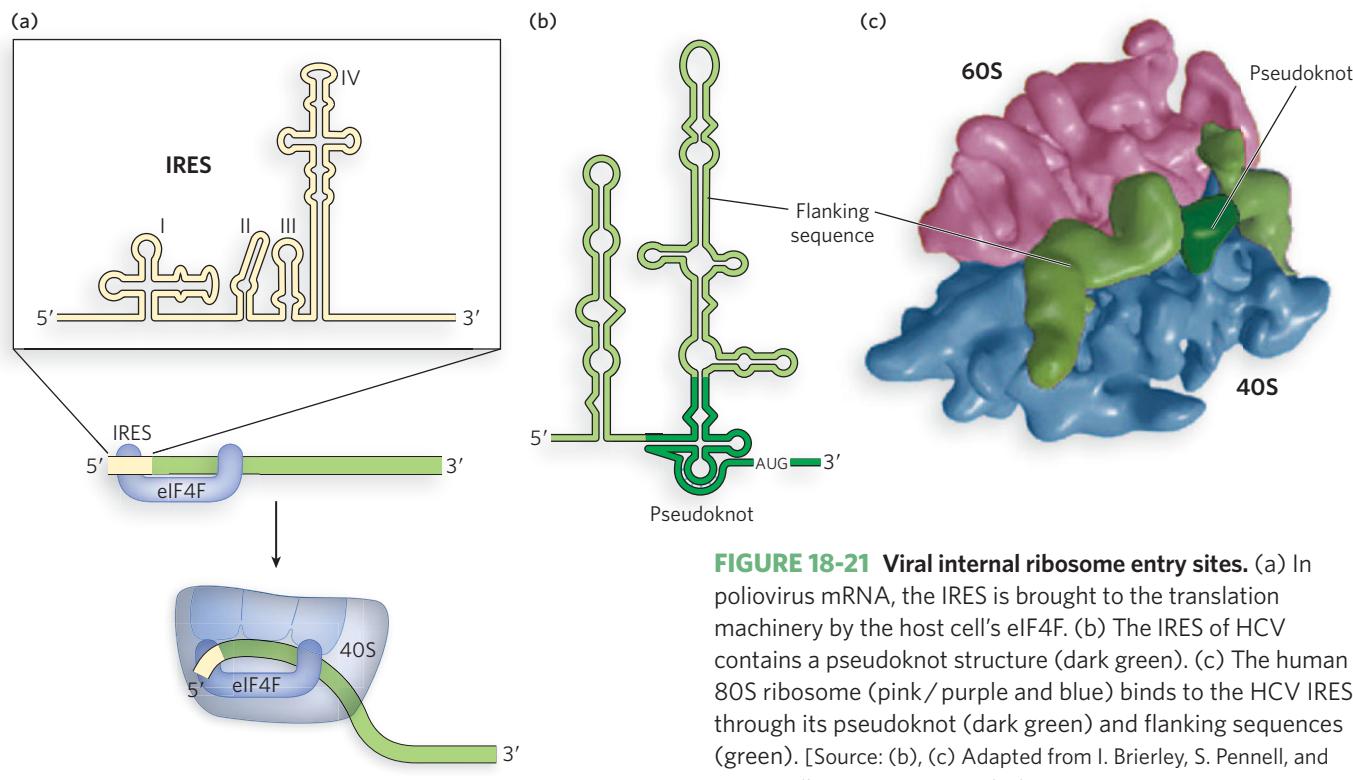


**FIGURE 18-20 Circularization of eukaryotic mRNA.** eIF4F (purple) consists of elF4A, elF4E, and elF4G. The elF4G binds to the 5' cap and to the poly(A) binding protein (PABP), effectively circularizing the mRNA.

mRNAs bearing an IRES can be translated even during viral infection and in other situations in which most protein synthesis is shut off, such as during starvation and programmed cell death. This is an important advantage of the IRES-mediated initiation mechanism. Internal initiation avoids the use of the 5' cap and the initiation factors needed to recognize it. The IRES positions the mRNA start codon correctly on the 40S subunit to ensure correct initiation during each round of translation.

### SECTION 18.3 SUMMARY

- In bacteria, ribosomes are recruited to the mRNA by the Shine-Dalgarno sequence, 4 to 9 purine residues located 8 to 13 nucleotides on the 5' side of the start codon.
- In eukaryotes, the mRNA 5' cap is bound by initiation factor eIF4E, which then binds other proteins that recruit the ribosomal subunits. The ribosome identifies the initiation codon by scanning the mRNA in a 5'→3' direction.
- In bacteria, the initiating aminoacyl-tRNA in all proteins is *N*-formylmethionyl-tRNA<sup>fMet</sup>. In eukaryotic cells, it is a special form of methionyl-tRNA, Met-tRNA<sub>i</sub><sup>Met</sup>.



**FIGURE 18-21 Viral internal ribosome entry sites.** (a) In poliovirus mRNA, the IRES is brought to the translation machinery by the host cell's eIF4F. (b) The IRES of HCV contains a pseudoknot structure (dark green). (c) The human 80S ribosome (pink/purple and blue) binds to the HCV IRES through its pseudoknot (dark green) and flanking sequences (green). [Source: (b), (c) Adapted from I. Brierley, S. Pennell, and R. J. C. Gilbert, *Nat. Rev. Microbiol.* 5:598–610, 2007.]

- Bacterial initiation factors IF-1, IF-2, and IF-3 promote assembly of the mRNA, fMet-tRNA<sup>fMet</sup>, and both ribosomal subunits to form the initiation complex. IF-1 and IF-3 prevent premature binding of the large subunit and elongator tRNAs, respectively. IF-2, a GTPase, recruits fMet-tRNA<sup>fMet</sup> to the P site. GTP hydrolysis allows the release of all three initiation factors and promotes association of the large subunit.
- Initiation in eukaryotes involves a host of initiation factors: eIF1, eIF1A, and eIF3 promote association of the small ribosomal subunit with eIF2-bound Met-tRNA<sub>i</sub><sup>Met</sup>, eIF5, and eIF5B, forming the 43S preinitiation complex. The mRNA 5' cap is bound by the eIF4F complex, which also binds to the small ribosomal subunit. Once this 48S particle is formed, the small subunit scans the mRNA for the start codon, after which the large subunit associates to form an active 80S ribosome.
- Some viral and eukaryotic mRNAs do not depend on the 5' cap for translation initiation, but instead bind eukaryotic initiation factors or the 40S ribosomal subunit at an internal ribosome entry site (IRES) downstream from the 5' end.

## 18.4 Elongation of the Polypeptide Chain

In the second stage of protein synthesis, **elongation**, the nascent polypeptide is lengthened by the covalent attachment of successive amino acid units, each carried to the ribosome and correctly positioned by its tRNA, which base-pairs to its corresponding codon in the mRNA. Elongation requires cytosolic proteins known as **elongation factors**. The binding of each incoming aminoacyl-tRNA and the movement of the ribosome along the mRNA are facilitated by the hydrolysis of GTP as each residue is added to the growing polypeptide. These steps are generally the same in bacteria and eukaryotes, just the names of some elongation factors differ. Because bacterial elongation is better defined, we describe that process here; we'll mainly focus on formation of the first peptide bond: conversion of the fMet-tRNA to a dipeptidyl-tRNA.

### Peptide Bonds Are Formed in the Translation Elongation Stage

In bacteria, elongation requires the initiation complex described in Section 18.3, aminoacyl-tRNAs, elongation factors **EF-Tu**, **EF-Ts**, and **EF-G**, and GTP.

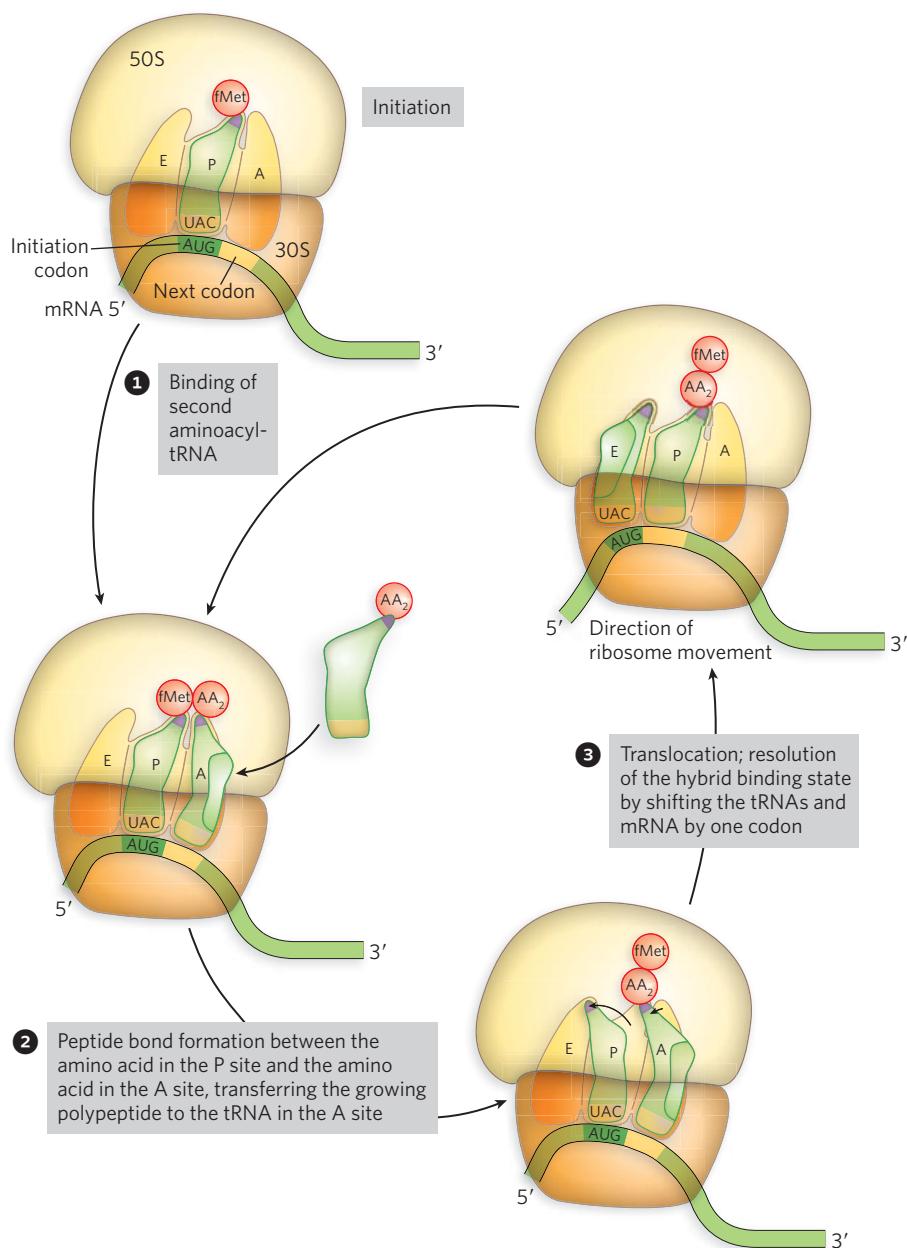
Cells use three steps to add each amino acid residue, and the steps are repeated as many times as there are residues to be added, as summarized in [Figure 18-22](#). In step 1, the appropriate incoming aminoacyl-tRNA binds to the ribosome with the help of EF-Tu ([Figure 18-23](#)). The aminoacyl-tRNA binds first to a complex of GTP-bound EF-Tu, and then to the A site of the 70S initiation complex. Next, the GTP is hydrolyzed and an EF-Tu-GDP complex is released from the 70S ribosome. The EF-Tu-GTP complex is regenerated in a process involving EF-Ts and GTP.

The GTPase activity of EF-Tu during the first step of elongation in bacterial cells makes an important contribution to the rate and fidelity of protein synthesis. Both the EF-Tu-GTP and the EF-Tu-GDP exist for a few milliseconds before they dissociate. These two intervals provide opportunities for proofreading of the codon-anticodon interactions. Once EF-Tu-GTP is hydrolyzed to EF-Tu-GDP, it loses binding affinity for the aminoacyl-tRNA. When EF-Tu-GDP releases the aminoacyl end of the tRNA within the ribosome, that end is still far from the active site of peptide bond formation. Correct codon-anticodon interactions in the ribosome rotate the tRNA into position for reaction with the growing polypeptide chain (or with the fMet group, for formation of the first peptide bond), in a process called **accommodation**. Incorrect aminoacyl-tRNAs normally dissociate from the A site at this time. If, in an *in vitro* experiment, the GTP analog guanosine 5'-O-(3-thiophosphosphate) (GTP $\gamma$ S) is used in place of GTP, hydrolysis is slowed, improving the fidelity (by increasing the proofreading interval) but reducing the rate of protein synthesis.

The process of protein synthesis (including the characteristics of codon-anticodon pairing) has clearly been optimized through evolution to balance the requirements for both speed and fidelity. Improved fidelity might diminish speed, whereas increased speed would probably compromise fidelity. Notice that the proofreading mechanism on the ribosome establishes only that the proper codon-anticodon pairing has taken place. As we saw in Section 18.2, the identity of the amino acid attached to a tRNA is not checked on the ribosome.

### Substrate Positioning and the Incoming tRNA Contribute to Peptide Bond Formation

In the second step of elongation, a peptide bond is formed between the two amino acids bound by their tRNAs to the A and P sites on the ribosome ([Figure 18-24](#)). Formation of the first peptide bond occurs by the transfer of the initiating N-formylmethionyl group from its tRNA in the P site to the amino group of

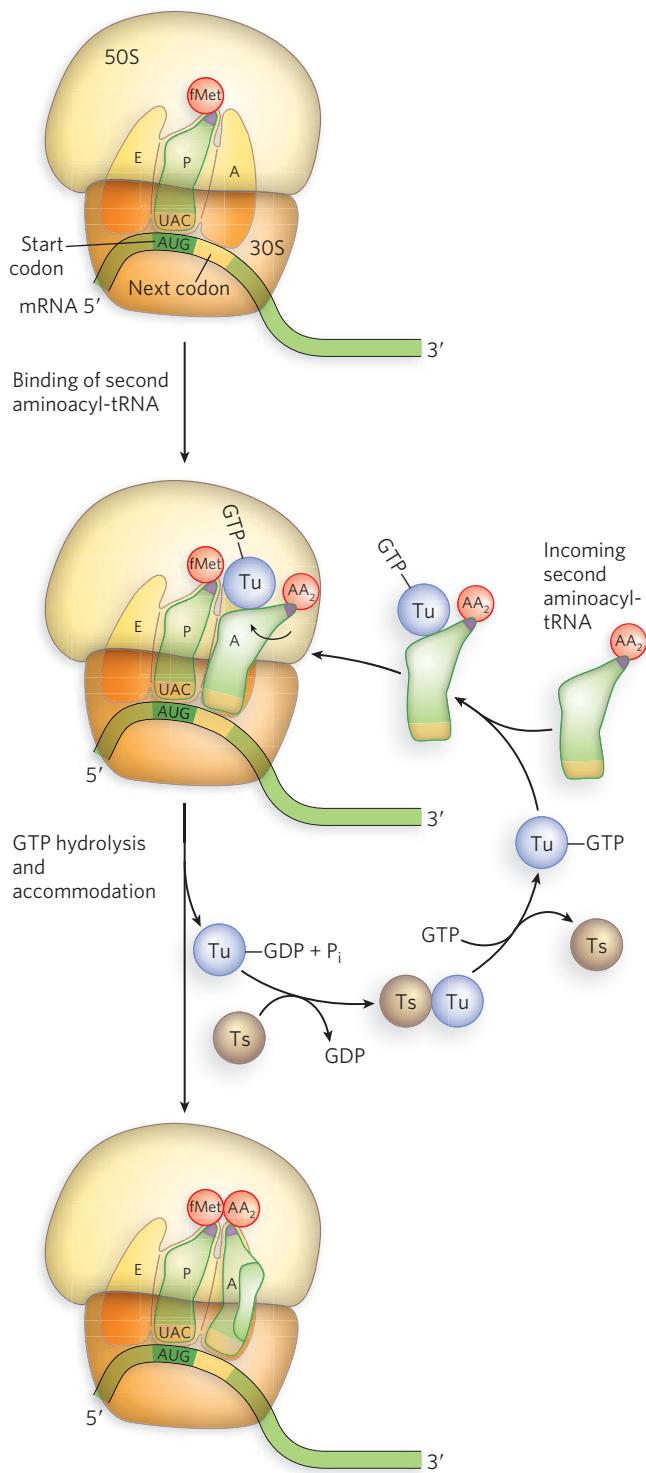


**FIGURE 18-22** An overview of the translation elongation cycle. The details of these three steps are shown in Figures 18-23, 18-24, and 18-26.

the second amino acid on its tRNA in the A site. The  $\alpha$ -amino group of the amino acid in the A site acts as a nucleophile, displacing the tRNA in the P site to form the peptide bond. This reaction produces a dipeptidyl-tRNA in the A site, leaving an uncharged (deacylated) tRNA<sup>fMet</sup> bound to the P site. The tRNAs then shift to a hybrid binding state, with elements of each spanning two different sites on the ribosome. To resolve this hybrid binding state, in the third, or translocation, step of elongation (described in detail below), the ribosome moves down the mRNA one codon, so the deacylated tRNA<sup>fMet</sup> is repositioned to the E site, and the

dipeptidyl-tRNA is repositioned to the P site, making the A site available for another aminoacyl-tRNA.

The enzyme that catalyzes peptide bond formation, peptidyl transferase, is intrinsic to the 23S rRNA of the large ribosomal subunit. The **peptidyl transferase reaction** entails nucleophilic attack of the  $\alpha$ -amino group of the A-site aminoacyl-tRNA on the carbonyl carbon of the ester bond linking the fMet (or the growing peptide chain) to the P-site tRNA. The ribosome increases the rate of peptide bond formation 10<sup>6</sup>- to 10<sup>7</sup>-fold above the intrinsic (uncatalyzed) rate of  $\sim 10^{-4} \text{ M}^{-1} \text{s}^{-1}$ . Combined evidence from x-ray crystallographic, genetic,



**FIGURE 18-23** EF-Tu-mediated binding of aminoacyl-tRNA to the ribosome. In the first step of elongation, the incoming aminoacyl-tRNA is bound by EF-Tu-GTP and inserted into the A site (EF-Tu is shown simply as Tu). GTP hydrolysis releases EF-Tu-GDP, leaving the tRNA in place. Through accommodation, the tRNA base-pairs with the mRNA codon and shifts by 70 Å into the correct position for the peptidyl transferase reaction. EF-Ts (shown as Ts) recycles EF-Tu by regenerating the EF-Tu-GTP complex.

and biochemical studies supports the idea that the A- and P-site substrates are precisely aligned in the active site by interactions of the 3'-terminal CCA sequences of the tRNAs and the nucleophilic  $\alpha$ -amino group with nucleotides in the 23S rRNA.

The proposed catalytic pathway of the peptidyl transferase reaction involves a six-membered transition state in which proton shuttling occurs by way of the 2'-OH of the terminal A residue (A76) of the P-site tRNA. In addition to the close positioning and orientation of the reactive groups relative to each other, the ribosome provides an electrostatic environment that reduces the energetic cost of forming the highly polar transition state by shielding the reaction from bulk water. In this organized environment, the ribosome prevents the extensive solvent rearrangement and consequent decrease in entropy that would occur if the same reaction took place in solution. In this way, the reaction on the ribosome is driven by a favorable entropy change.

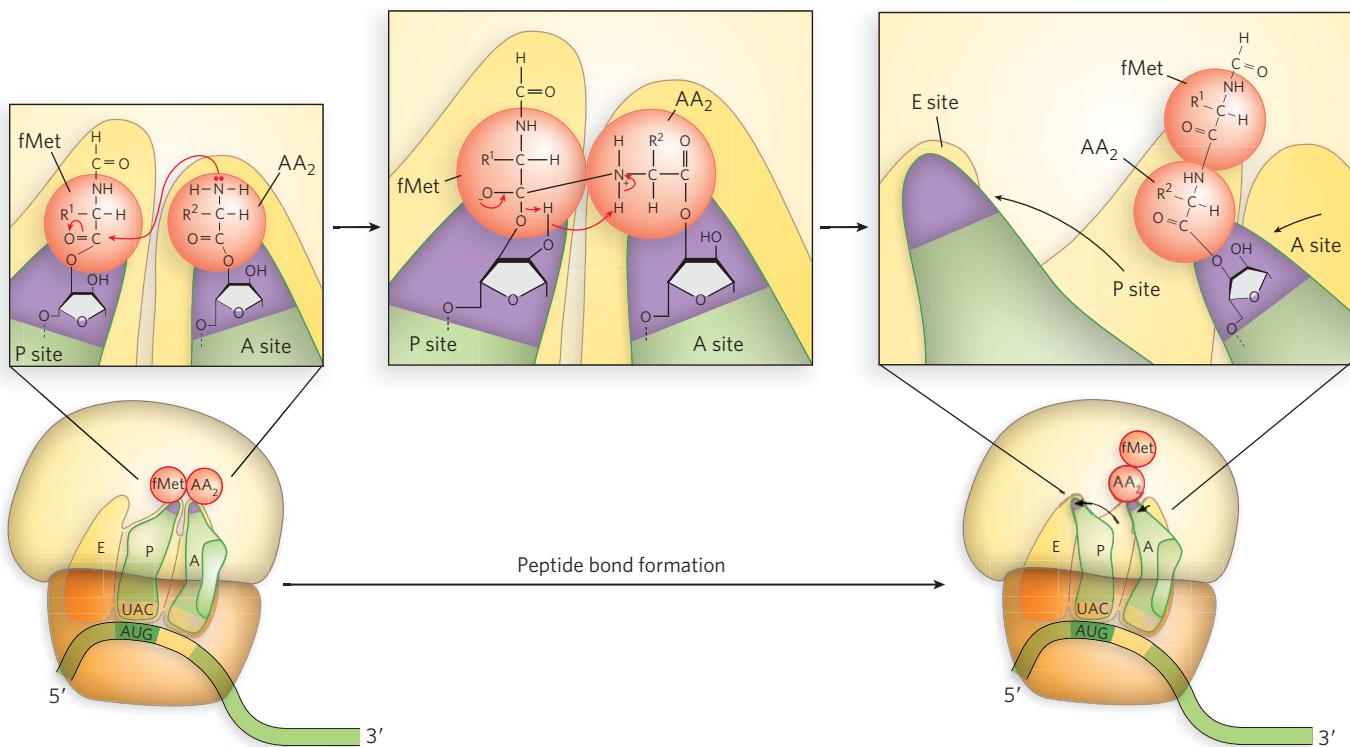
### The GTPase EF-G Drives Translocation by Displacing the A-Site tRNA

Immediately following formation of the first peptide bond, the ribosome moves one codon toward the 3' end of the mRNA in a process called **translocation**. This movement shifts the anticodon of the dipeptidyl-tRNA, which is still attached to the second codon of the mRNA, from the A site to the P site, and shifts the deacylated tRNA from the P site to the E site, from which the tRNA is released into the cytosol. The third codon of the mRNA now lies in the A site and the second codon in the P site.

Movement of the ribosome along the mRNA requires the energy provided by hydrolysis of another molecule of GTP by the GTPase EF-G. EF-G is similar in structure to the EF-Tu–aminoacyl-tRNA complex (Figure 18-25) and can bind to the large subunit side of the A site, displacing the peptidyl-tRNA. On binding of EF-G-GTP to the A site, interactions with a region of the large subunit trigger GTP hydrolysis (Figure 18-26). When GTP is hydrolyzed, the EF-G structure changes so that it can make contact with the small subunit and bring about translocation of the A-site tRNA. It is thought that the change in EF-G structure as GTP is hydrolyzed leads to a change in the three-dimensional conformation of the entire ribosome, resulting in its movement along the mRNA.

### GTP Binding and Hydrolysis Regulate Successive Elongation Cycles

The ribosome, with its attached dipeptidyl-tRNA and mRNA, is now ready for addition of a third amino acid residue and the next elongation cycle. This process



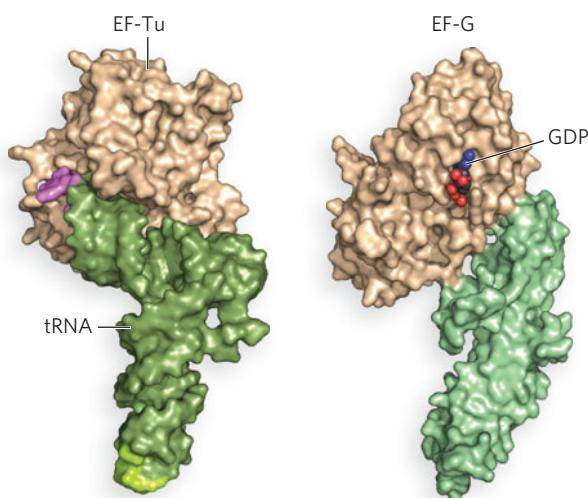
**FIGURE 18-24** The peptidyl transferase reaction. In the second step of elongation, the α-amino group of the A-site aminoacyl-tRNA attacks the carbonyl carbon of the P-site fMet-tRNA, shifting the fMet to the A-site tRNA to form a dipeptidyl-tRNA. This reaction is catalyzed by the 23S rRNA.

occurs in the same way as addition of the second residue: EF-Tu-GTP delivers a charged tRNA to the A site, the rRNA catalyzes peptide bond formation, then EF-G-GTP displaces the peptidyl-tRNA into the P site. For each amino acid residue correctly added to the growing polypeptide, two GTPs are hydrolyzed to GDP

and P<sub>i</sub> as the ribosome moves from codon to codon along the mRNA toward the 3' end.

The polypeptide remains attached to the tRNA of the most recent amino acid to be inserted. This association maintains the functional connection between the information in the mRNA and its polypeptide product. At the same time, the ester linkage between this tRNA and the C-terminus of the growing polypeptide activates the terminal carboxyl group for nucleophilic attack by the incoming amino acid to form a new peptide bond. As the existing ester linkage between the polypeptide and tRNA is broken during peptide bond formation, the linkage between the polypeptide and the information in the mRNA persists, because each newly added amino acid is still attached to its tRNA.

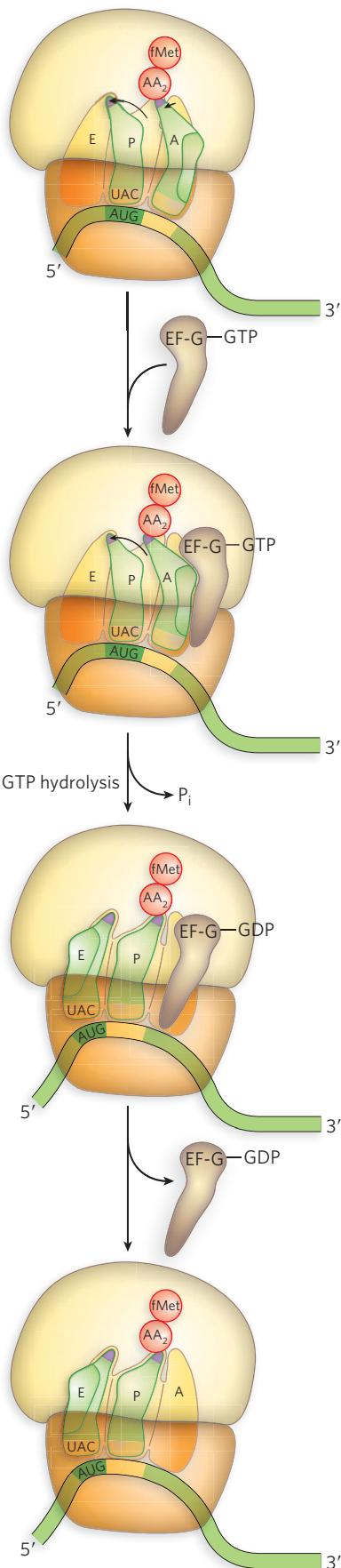
The elongation cycle in eukaryotes is quite similar to that in bacteria. Three eukaryotic elongation factors—**eEF1α**, **eEF1βγ**, and **eEF2**—have functions analogous to those of the bacterial elongation factors EF-Tu, EF-Ts, and EF-G, respectively.



**FIGURE 18-25** EF-Tu-aminoacyl-tRNA and EF-G. The structure of EF-G mimics that of the EF-Tu-aminoacyl-tRNA complex, allowing EF-G to fit in the A site. [Source: PDB ID 1623 and 1DAR.]

## SECTION 18.4 SUMMARY

- In the first step of elongation in bacteria, the incoming aminoacyl-tRNA binds to the ribosome with the help of EF-Tu. This step requires GTP



**FIGURE 18-26 Translocation promoted by EF-G binding and GTP hydrolysis.** In the third step of elongation, the hybrid binding state formed in the peptidyl transferase reaction (see Figure 18-24) is resolved by the binding of EF-G-GTP at the A site. GTP hydrolysis causes a conformational change in EF-G that displaces the A-site peptidyl-tRNA into the P site, causing the ribosome to move forward by one codon on the mRNA. Once EF-G-GDP is released, the A site is available for another aminoacyl-tRNA.

hydrolysis and allows time for proofreading of the codon-anticodon interaction. In the second step, a peptide bond is formed between the two amino acids bound by their tRNAs to the ribosomal A and P sites.

- Peptide bond formation involves nucleophilic attack of the  $\alpha$ -amino group of the A-site aminoacyl-tRNA on the carbonyl carbon of the ester bond linking the growing peptide chain to the P-site tRNA. This reaction is driven by a favorable change in entropy.
- The third step of elongation is translocation of the peptidyl-tRNA from the A site into the P site. This is accomplished with the help of EF-G, a GTPase that is a structural analog of the EF-Tu–aminoacyl-tRNA complex.
- The three steps of elongation are repeated for each codon in the mRNA; each cycle consumes two GTP molecules.

## 18.5 Termination of Protein Synthesis and Recycling of the Synthesis Machinery

The completion of a polypeptide is signaled by one of three mRNA codons (UAA, UAG, and UGA) that act as **termination codons**, or **stop codons**. As the translation product is released, the ribosomes and associated factors become available for another round of translation. After discussing these events of termination and recycling, we also consider the overall energy requirements of protein synthesis and some effects of antibiotics on the protein synthesis machinery.

### Completion of a Polypeptide Chain Is Signaled by an mRNA Stop Codon

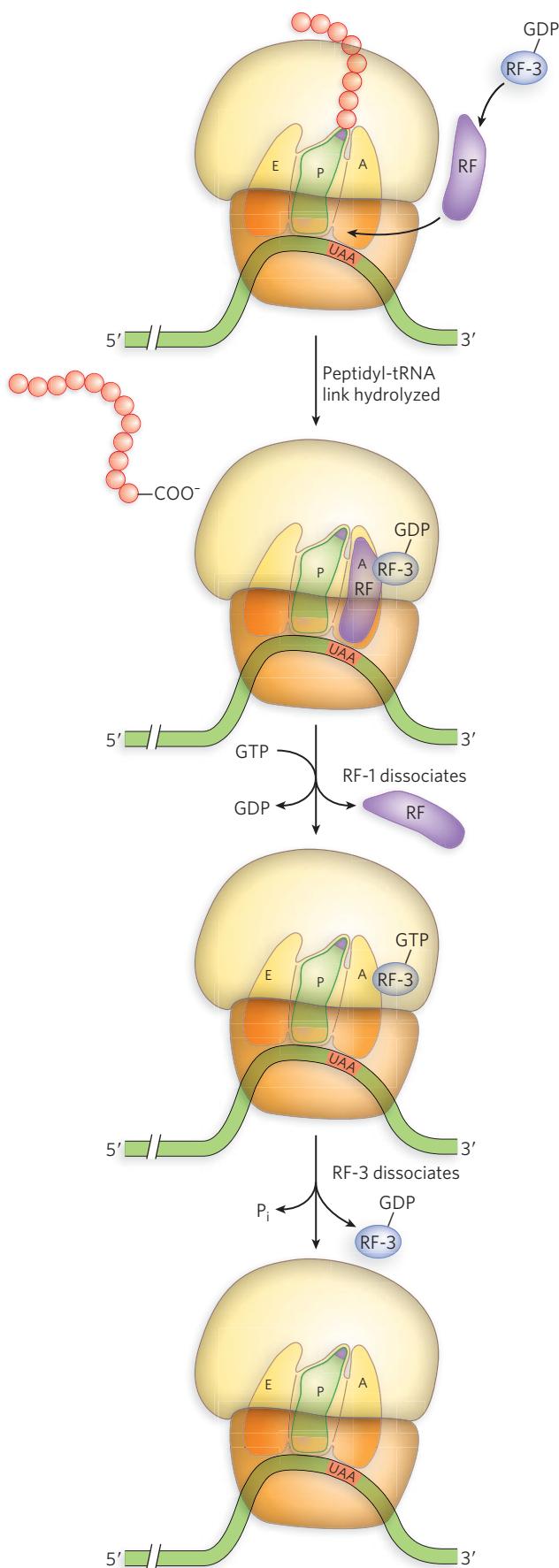
Elongation continues until the ribosome adds the last amino acid coded by the mRNA. **Termination** is signaled by the presence of a stop codon in the mRNA. Mutations in a tRNA anticodon that allow an amino

acid to be inserted at a termination codon are generally deleterious to the cell.

In bacteria, once a stop codon occupies the ribosomal A site, three **termination factors**, or **release factors**—the proteins RF-1, RF-2, and RF-3—contribute to (1) hydrolysis of the terminal peptidyl-tRNA bond; (2) release of the free polypeptide and the last tRNA, now uncharged, from the P site; and (3) dissociation of the 70S ribosome into its 30S and 50S subunits, ready to start a new cycle (Figure 18-27). RF-1 and RF-2 are related factors and are referred to as class I release factors; RF-3 acts in a different way and is referred to as a class II release factor. RF-1 and RF-2 recognize termination codons and bind the ribosome in much the same way as tRNAs. RF-1 recognizes stop codons UAG and UAA, and RF-2 recognizes UGA and UAA. Either RF-1 or RF-2 (depending on which codon is present) binds at the stop codon and induces peptidyl transferase to transfer the growing polypeptide to a water molecule rather than to another amino acid.

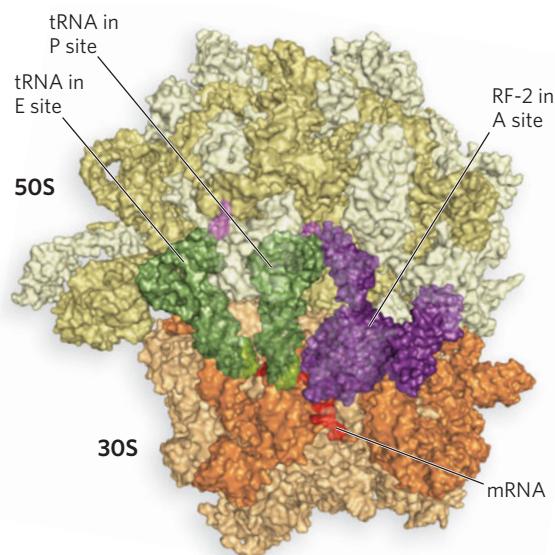
RF-1 and RF-2 have domains that mimic the structure of a tRNA (Figure 18-28). RF-3, a GTPase, catalyzes the dissociation of RF-1 and RF-2 from the ribosome following release of the polypeptide chain. Unlike the other GTP-binding factors that regulate translation, RF-3 binds with higher affinity to GDP than to GTP. For this reason, RF-3-GDP is the predominant form of the factor that binds initially to the ribosome. However, the association between RF-3-GDP and the ribosome is weak unless one of the other release factors, RF-1 or RF-2, is present. On polypeptide release, stimulated by RF-1 or RF-2, an associated ribosomal conformational change triggers exchange of the RF-3-bound GDP for GTP. The RF-3-GTP has a much higher affinity for the ribosome, leading to displacement of RF-1 or RF-2 and creating contact between RF-3 and the factor-binding center of the large ribosomal subunit. As observed for other GTPases that bind the ribosome, this interaction on the large subunit stimulates GTP hydrolysis, allowing RF-3-GDP to fall off.

In eukaryotes, a single release factor, eRF1, recognizes all three termination codons. A second release factor, eRF3, catalyzes GTP-dependent release of eRF1 from the ribosome.



**FIGURE 18-27** Termination of translation in bacteria.

When the ribosome comes to a stop codon, RF-1 or RF-2 binds to the A site and induces polypeptide chain release. RF-3-GDP then binds to the ribosome and exchanges GTP for GDP, displacing the RF-1 or RF-2. The RF-3-GTP is tightly bound to the ribosome, but GTP hydrolysis weakens its affinity and RF-3 is released.



**FIGURE 18-28** The crystal structure of RF-2 bound in the ribosomal A site. RF-1 (not shown) and RF-2 are proteins that mimic tRNAs, binding to the A site and mRNA, displacing the last tRNA from the A site into the P site, and releasing the polypeptide from the ribosome. [Source: PDB ID 3FIE and 3FIF.]

### Ribosome Recycling Factor Prepares Ribosomes for New Rounds of Translation

After release of the polypeptide and the termination factors, the ribosome remains bound to the mRNA and contains deacylated tRNA in the P and E sites. **Ribosome recycling** removes the mRNA and tRNAs and separates the ribosome into its subunits in preparation for new rounds of translation. In bacteria, **ribosome recycling factor** (RRF) binds to the empty ribosomal A site and recruits EF-G to stimulate release of the uncharged tRNAs in the P and E sites, in a process mimicking EF-G-stimulated polypeptide elongation (Figure 18-29).

In a continuing story of molecular mimicry, RRF, like EF-G and EF-Tu-tRNA, fits into the ribosome's tRNA-binding sites. GTP hydrolysis by EF-G is thought to result in translocation and displacement of the tRNAs, as occurs during translation. However, when RRF occupies the A site and translocates to the P site, the ribosome releases EF-G and RRF. IF-3 then binds the small ribosomal subunit and disaggregates the ribosome, thereby triggering release of the mRNA and preparing the small subunit for a new round of translation. The apparent absence of a ribosome recycling factor in eukaryotes suggests that termination of translation in higher organisms may be somewhat different.

### Fast and Accurate Protein Synthesis Requires Energy

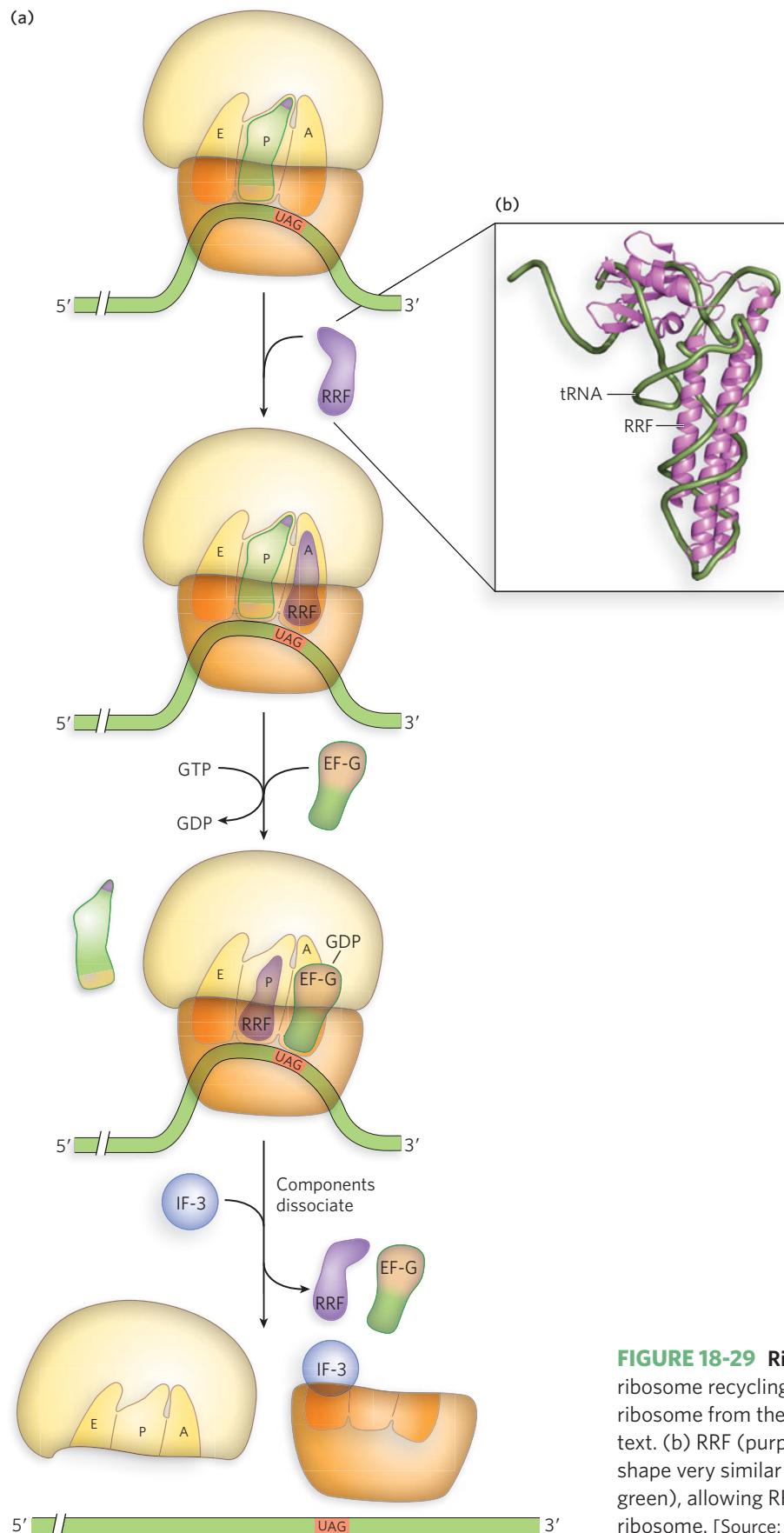
The synthesis of a protein true to the information specified in the cell's mRNA requires energy. Forming each aminoacyl-tRNA requires two high-energy phosphate groups. An additional ATP is consumed each time an incorrectly activated amino acid is hydrolyzed by the deacylation activity of an aminoacyl-tRNA synthetase, as part of its proofreading activity. One GTP is cleaved to GDP and  $P_i$  during the first elongation step, and another during the translocation step. Thus, on average, the energy derived from the hydrolysis of more than four NTPs to NDPs is required for the formation of each peptide bond.

This represents a large thermodynamic drive in the direction of synthesis: at least  $4 \times 30.5 \text{ kJ/mol} = 122 \text{ kJ/mol}$  of phosphodiester bond energy to generate a peptide bond, which has a standard free energy of formation of about 21 kJ/mol. The net free-energy change during peptide bond synthesis is thus  $-101 \text{ kJ/mol}$ . Why would the cell need to expend so much energy on protein synthesis?

Proteins are information-containing polymers. The biochemical goal in the peptidyl transferase reaction is not simply the formation of a peptide bond but the formation of a peptide bond between two *specified* amino acids. Each of the high-energy phosphate compounds expended in this process plays a critical role in maintaining proper alignment between each new codon in the mRNA and its associated amino acid at the growing end of the polypeptide. This energy input permits very high fidelity in the biological translation of the genetic message of mRNA into the amino acid sequence of proteins.

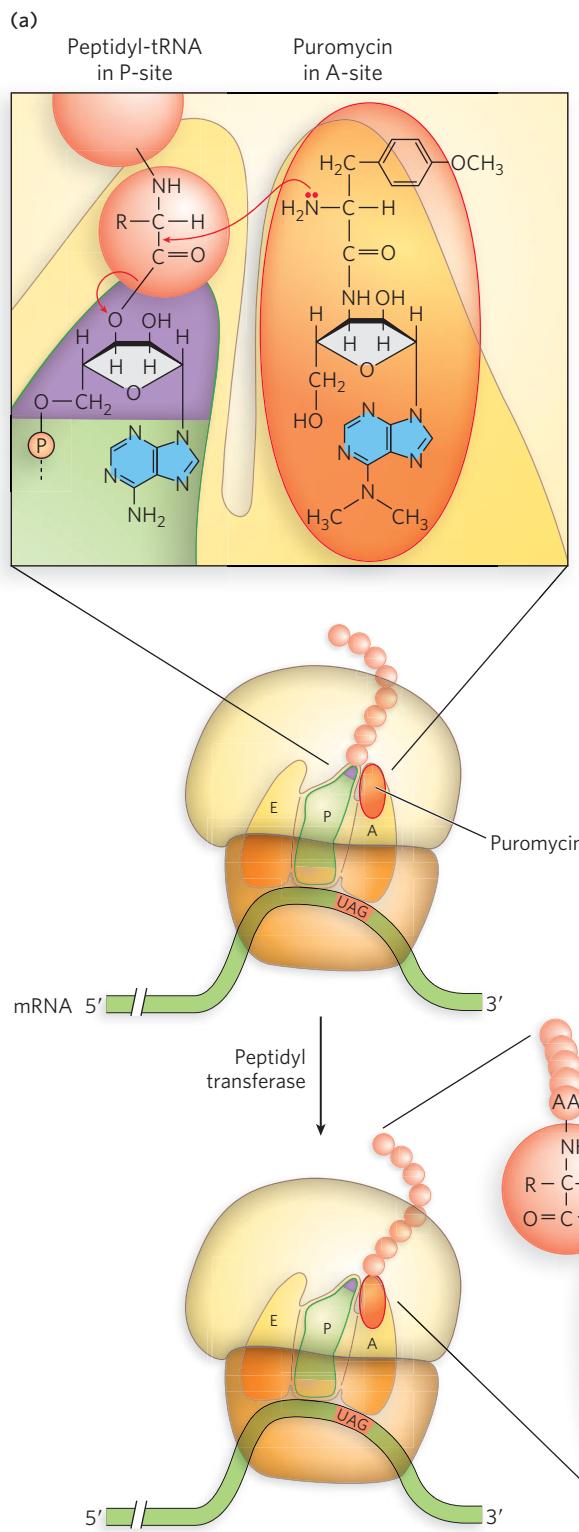
### Antibiotics and Toxins Frequently Target the Protein Synthesis Cycle

Protein synthesis is the primary target of many naturally occurring antibiotics and toxins. Antibiotics are produced by bacteria or other microorganisms to inhibit protein synthesis in other bacteria—that is, these biochemical weapons are synthesized by some microorganisms and are extremely toxic to others. The differences between bacterial and eukaryotic protein synthesis, though often subtle, are sufficient that most (though by no means all) of the antibiotics discussed here are relatively harmless to eukaryotic cells. Because nearly every step in protein synthesis can be specifically inhibited by one antibiotic/toxin or another, antibiotics have become valuable tools in the study of protein synthesis.



**FIGURE 18-29** Ribosome recycling. (a) In bacteria, ribosome recycling factor (RRF) and EF-G separate the ribosome from the mRNA and tRNAs, as described in the text. (b) RRF (purple) is a protein with a three-dimensional shape very similar to that of a tRNA (superimposed in green), allowing RRF to bind to the tRNA sites on the ribosome. [Source: (b) Adapted from PDB ID 1DD5.]

**Puromycin**, made by the mold *Streptomyces alboniger*, is one of the best-understood inhibitory antibiotics. Its structure is very similar to the 3' end of an aminoacyl-tRNA, so puromycin can bind to the ribosomal A site and

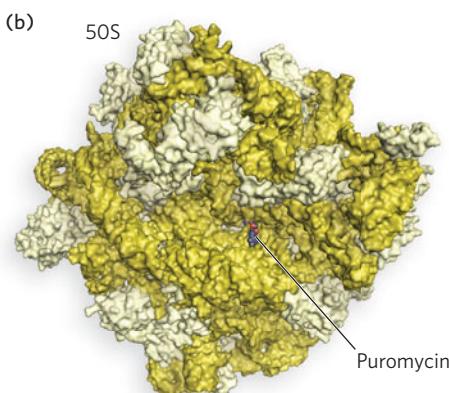


participate in peptide bond formation, producing peptidyl-puromycin (Figure 18-30). However, because puromycin resembles only the 3' end of the tRNA, it does not engage in translocation and dissociates from the ribosome shortly after it is linked to the C-terminus of the peptide. This prematurely stops protein synthesis.

**Tetracyclines** inhibit bacterial protein synthesis by blocking the ribosomal A site, preventing the binding of aminoacyl-tRNAs. **Chloramphenicol** inhibits protein synthesis by bacterial (and mitochondrial and chloroplast) ribosomes by blocking peptidyl transfer; it does not affect cytosolic protein synthesis in eukaryotes. Conversely, **cycloheximide** blocks the peptidyl transferase of 80S eukaryotic ribosomes but not that of 70S bacterial (or mitochondrial and chloroplast) ribosomes. **Streptomycin**, a trisaccharide, causes misreading of the genetic code (in bacteria) at relatively low concentrations and inhibits initiation at higher concentrations.

Several other inhibitors of protein synthesis are notable because of their toxicity to humans and other mammals. **Diphtheria toxin** is an enzyme that catalyzes the ADP-ribosylation of a diphthamide (a modified histidine) residue of the eukaryotic elongation factor eEF2, thereby inactivating it. The toxin is secreted by the bacterium *Corynebacterium diphtheriae*. Infected individuals experience fever, chills, neck swelling, and a fast heart rate. Although common historically, diphtheria has largely been eradicated in developed countries through

**FIGURE 18-30** Puromycin as inhibitor of translation.  
 (a) Puromycin inhibits translation by binding to the A site and mimicking the 3' end of an aminoacyl-tRNA. The puromycin participates in the peptidyl transfer reaction, but because it is not bound to a tRNA, it is not anchored to the ribosome. (b) Crystal structure showing puromycin bound to the peptidyl transferase center of the bacterial 50S ribosomal subunit. [Source: PDB ID 1Q7Y.]



## HIGHLIGHT 18-3 MEDICINE

### Toxins That Target the Ribosome

The toxic effects of two small proteins called ricin and abrin, derived from castor beans and jequirity peas, respectively, have been known and studied since ancient times. After a period of extensive study of these toxins at the end of the nineteenth century, the scientific community largely lost interest—until experiments published in the late 1960s and early 1970s showed that ricin and abrin inhibited protein synthesis in rats and cultured cells. The toxins did not interfere with the structure of polysomes, so researchers concluded that they acted on a component required for elongation of the polypeptide.

Studies showed that ricin and abrin required some time to exert their effects, suggesting that they might have enzymatic activity. Both toxins consist of two distinct, disulfide-linked polypeptide chains. The finding that reducing agents enhanced the toxins' ability to inhibit protein synthesis *in vitro* suggested that the activity lies in one of the individual chains; in both toxins, this was found to be the smaller chain, termed the A chain. After cell fractionation to isolate ribosomes and smaller protein factors, experiments showed that the A chains act on

the ribosomes. Further studies revealed that the target is the 60S ribosomal subunit.

In elegant studies in the 1980s, Ira Wool and Y. Endo found that the toxin A chain is a glycosidase (an enzyme that cleaves sugars at the glycosidic linkage) that removes the adenine from an A residue in an exposed loop of 28S rRNA (residue A4,324 in rat 28S RNA). Because this loop is involved in the binding of an elongation factor, the modified ribosomes are unable to support protein synthesis.

Ricin is legendary for its use as a murder weapon. In perhaps the most famous case, known as the Umbrella Murder, journalist Georgi Markov was killed on a London street, reportedly by the Bulgarian secret police. Markov felt a pain like a bee sting on the back of his thigh. Turning around, he saw a man pick up an umbrella from the ground and then quickly step into a taxi, which drove away. When Markov arrived at his office at the BBC World Service, he noticed a red pimple forming at the site of the sting, and the pain was increasing. Later that day he developed a fever and was admitted to a hospital, where he died a few days later. The cause of death was poisoning from a ricin-filled pellet, presumably fired from an umbrella tip.

widespread vaccination. Among the plant-produced toxins, **ricin**, an extremely toxic protein of the castor bean, inactivates the 60S subunit of eukaryotic ribosomes by depurinating a specific adenosine in 23S rRNA (Highlight 18-3). Table 18-4 lists some antibiotics and toxins that target translation, their functional consequences, and some clinical and other uses.

### SECTION 18.5 SUMMARY

- After many elongation cycles, release factors recognize the stop codon and terminate polypeptide synthesis by releasing the polypeptide, displacing the tRNAs, and separating the ribosomal subunits. In bacteria, ribosome recycling factor stimulates release of the tRNAs in the P and E sites, leading to release of the mRNA.
- At least four high-energy phosphate bonds (from ATP and GTP) must be broken to generate each peptide bond, an energy investment necessary for guaranteeing the fidelity of translation.
- Many well-studied and clinically important antibiotics and toxins inhibit some aspect of protein synthesis.

## 18.6 Translation-Coupled Removal of Defective mRNA

Occasionally, mRNAs with premature stop codons (nonsense mutations; see Chapter 17) or with no stop codon at all (so-called non-stop mRNAs) arise from errors in DNA replication or transcription, or from mRNA degradation. Such defective mRNAs have the potential to produce nonfunctional or even toxic proteins, and they can also prevent efficient termination of translation and recycling of the protein synthesis machinery. For these reasons, elegant mechanisms have evolved to deal with such aberrant mRNAs during translation.

### Ribosomes Stalled on Truncated mRNAs Are Rescued by tmRNA

Truncated, or non-stop, mRNAs occur when DNA transcription ends prematurely or when, due to a mutation, an mRNA lacks a stop codon. Truncated proteins produced by incomplete mRNAs, if inactive, could spell disaster for the cell, because they might contain the binding determinants that allow them to take the place

**Table 18-4 Some Antibiotics and Toxins That Inhibit Translation: Sources, Actions, and Uses**

Class	Examples	Source(s)	Targets	Mode of Action	Uses
<i>Inhibitors of initiation</i>					
Edeine	Edeine	<i>Bacillus brevis</i>	Bacteria and eukaryotic cells	Binds small-subunit P and E sites, destabilizing interaction between mRNA and initiator aminoacyl-tRNA (prevents initiator aminoacyl-tRNA binding)	Formerly an agricultural pesticide; now banned in most countries
Kasugamycin	Kasugamycin	<i>Streptomyces kasugiensis</i>	Bacteria and fungi; low toxicity to higher eukaryotes	Binds small-subunit P and E sites, destabilizing interaction between mRNA and initiator aminoacyl-tRNA (prevents initiator aminoacyl-tRNA binding)	Agricultural fungicide to prevent rice blast infection
Orthosomycins	Avilamycin, evernimicin,	<i>Streptomyces viridochromogenes</i> , <i>Micromonospora carbonaceae</i> , respectively	Bacteria	Bind 50S subunit, preventing association with preinitiation complex	Avilamycin used in animal feeds to promote growth; evernimicin developed for treatment of bacterial infections (e.g., MRSA), but caused reproductive defects in rats
Pactamycin	Pactamycin	<i>Streptomyces pactum</i>	Bacteria and eukaryotic cells	Inhibits initiation and elongation; mechanism unknown	Potential antitumor agent
<i>Inhibitors of elongation</i>					
Aminoglycosides	Gentamicin, hygromycin B, kanamycin, neomycin, paromomycin	<i>Streptomyces</i> spp.	Bacteria and eukaryotic cells	Promote errors in decoding by stabilizing incorrect codon-anticodon pairings	Prevent various bacterial infections; used topically (e.g., Neosporin), orally, by injection, or during surgery
Aminoglycosides	Streptomycin	<i>Streptomyces griseus</i>	Bacteria	Promotes errors in decoding by altering rate of GTPase activation in EF-Tu	Broad-spectrum antibiotic to treat tuberculosis and plague; also used as pesticide in agriculture

(continued)

**Table 18-4 Some Antibiotics and Toxins That Inhibit Translation: Sources, Actions, and Uses (continued)**

Class	Examples	Source(s)	Targets	Mode of Action	Uses
Amphenicols	Azidamfenicol, chloramphenicol, florfenicol, thiamphenicol	<i>Streptomyces venezuelae</i> and synthetic	Bacteria	Bind A site, inhibiting peptidyl transferase reaction	Broad-spectrum antibiotics to treat serious bacterial infections
Enacyloxins and kirromycins	Aurodox, azdimycin, delvomycin, efrotomycin, heneicomycin, kirromycin, mocimycin,	<i>Streptomyces</i> spp.	Bacteria	Stall ternary complex on ribosome by preventing conformational change in EF-Tu	Veterinary medicines and food additives to promote animal growth
Lincosamides	Clindamycin, lincomycin	Synthetic and actinomycetes, respectively	Bacteria	Bind A site	Treatment of bacterial infections, protozoal diseases, malaria, toxic shock syndrome
Macrolides	Azithromycin, carbomycin, clarithromycin, erythromycin, roxithromycin, spiramycin, telithromycin, tylosin	<i>Streptomyces</i> spp.	Bacteria	Bind exit tunnel, preventing exit of polypeptide from ribosome	Treatment of bacterial infections
Oxazolidinones	Cycloserine, linezolid, ranbezolid	Synthetic	Bacteria	Bind A site; exact mechanism unknown	Treatment of bacterial infections resistant to other antibiotics
Pleuromutilin	Retapamulin, tiamulin, valnemulin	<i>Clitopilus scypoides</i> and synthetic	Bacteria	Inhibit initiation and peptidyl transferase reaction by binding to A and P sites at the same time	Treatment of skin infections; veterinary medicine
Sparsomycin	Sparsomycin	<i>Streptomyces sparsogenes</i>	Bacteria and eukaryotic cells	Inhibits tRNA binding to A site and enhances tRNA binding to P site	Antitumor drug
Streptogramins	Dalfopristin, pristinamycin, quinupristin, streptogramin A, streptogramin B	<i>Streptomyces</i> spp.	Bacteria	Bind distinct, adjacent locations on peptidyl transferase center and act synergistically to block both A and P sites in initiation and elongation	Treatment of skin and other infections

(continued)

**Table 18-4 Some Antibiotics and Toxins That Inhibit Translation: Sources, Actions, and Uses (continued)**

Class	Examples	Source(s)	Targets	Mode of Action	Uses
Tetracyclines	Chlortetracycline, demeclocycline, doxycycline, lymecycline, mecloxycline, methacycline, minocycline, oxytetracycline, rolitetracycline, tetracycline	<i>Streptomyces</i> spp. and synthetic	Bacteria	Bind 30S subunit, preventing ternary complex binding to ribosome	Broad-spectrum antibiotics to treat pneumonia, acne, skin infections, genital and urinary tract infections, ulcers, Lyme disease, anthrax, and other disorders
<i>Inhibitors of translocation</i>					
Aminoglycosides	Spectinomycin	<i>Streptomyces spectabilis</i>	Bacteria	Inhibits translocation by stabilizing an intermediate	Treatment of gonorrhea; no longer available in United States
Fusidic acid (steroid)	Fusidic acid	<i>Fusidium coccineum</i>	Bacteria	Prevents EF-G dissociation by binding to EF-G-GTP in complex with ribosome	Treatment of bacterial infections
Ricin (protein)	Ricin	<i>Ricinus communis</i>	Bacteria and eukaryotic cells	Depurinates 23S rRNA, probably disrupting GTPase-stimulating activity	Developed, but never used, by United States for biological/chemical warfare
Thiopeptides	Micrococcin, nosiheptide, thiostrepton	<i>Streptomyces azureus</i>	Bacteria	Bind A site, inhibiting tRNA-IF-Tu binding and EF-G binding (inhibiting initiation and translocation)	Veterinary medicine
Tuberactinomycins	Capreomycin, eniromycin, spectinomycin, viomycin	<i>Streptomyces</i> spp.	Bacteria	Inhibit translocation by stabilizing an intermediate	Treatment of tuberculosis

Note: "Class" includes some individual antibiotics, categorized separately because of their particular mode of action. "Examples" includes just a sample of antibiotics in that class. "Uses" lists treatments for humans, unless noted as veterinary or agricultural. Although not noted, many of these antibiotics/toxins are also used in the laboratory as selective markers and/or to study protein synthesis.

of the active protein in a cellular activity. The ribosome takes care of this problem in a process that recognizes and removes the defective mRNA, as well as the defective protein.

When a ribosome reaches the 3' end of a truncated mRNA, it stalls, unable to recruit either an appropriate tRNA or the proper release factors. In

bacteria and some eukaryotic organelles, a fascinating quality-control pathway solves this problem with a 457-nucleotide RNA called **tmRNA**, also known as SsrA RNA or 10Sa RNA. A versatile, evolutionarily conserved bacterial molecule, tmRNA has the combined structural and functional properties of both a tRNA and an mRNA.

The 5' terminus of tmRNA mimics the structure of tRNA<sup>Ala</sup> and is charged with alanine by the Ala-tRNA synthetase. The alanine-charged tmRNA binds like a tRNA to the A site of the stalled ribosome, together with EF-Tu-GTP (Figure 18-31). The Ala-tmRNA donates its alanine to the nascent polypeptide chain in the P site, in a standard peptidyl transferase reaction. The tmRNA then takes the place of mRNA in the vacated mRNA channel of the ribosome stalled at the P site. The placement of another RNA molecule instead of an mRNA in this ribosomal channel could not normally occur, because an intact mRNA would occupy the channel. But a prematurely terminated mRNA presents a vacant mRNA site for a tmRNA to occupy. Thus, tmRNA can act as a surrogate mRNA, replacing the truncated mRNA, with its self-encoded peptide reading frame directing synthesis of a 9 amino acid sequence before encountering a stop codon. Thus, counting the alanine, the tmRNA incorporates 10 amino acids at the C-terminus of the truncated protein.

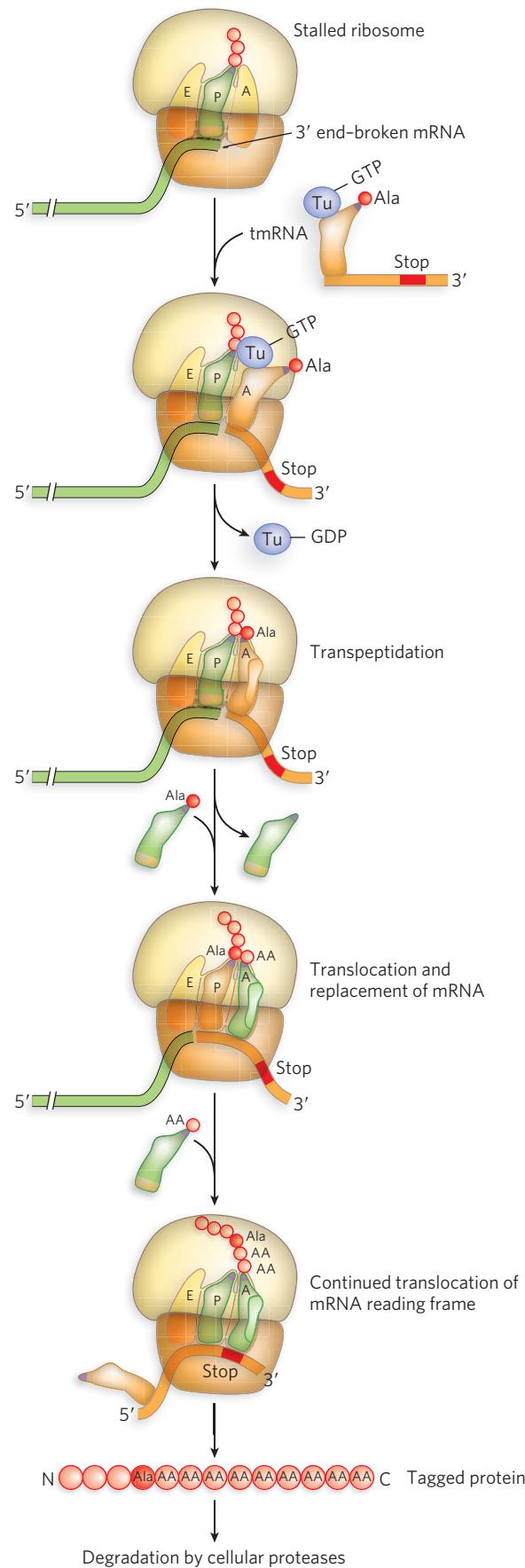
Translation terminates at the tmRNA stop codon, permitting disassembly and recycling of the ribosomal subunits. And in a remarkably strategic cellular maneuver, the 10 amino acid sequence encoded by tmRNA is a cellular signal that tags the protein for degradation—that is, the 10-residue degradation tag at the C-terminus is recognized by C-terminal-specific cellular proteases. The proteases denature and degrade tagged proteins in both the cytoplasm and the periplasmic space (between the inner plasma membrane and outer membrane of some bacteria). Furthermore, in addition to ribosome rescue and protein tagging, the tmRNA salvage system facilitates degradation of the defective mRNA by the enzyme RNase R.

Although not essential in *E. coli*, tmRNA activity is required for bacterial survival under adverse conditions and for virulence in some (perhaps all) pathogenic bacteria. Recent evidence suggests that in addition to its quality-control function, the tmRNA system might play a key role in regulating proteins for which cellular concentrations are particularly sensitive to the balance between proteolysis and the cell's translational efficiency.

## Eukaryotes Have Other Mechanisms to Detect Defective mRNAs

Eukaryotes respond to non-stop mRNAs in other ways, not with the bacterial tmRNA mechanism described

**FIGURE 18-31** The rescue of stalled bacterial ribosomes by tmRNA. In bacteria, tmRNA rescues stalled ribosomes by mimicking both tRNA and mRNA, allowing release of the faulty mRNA while at the same time marking the truncated polypeptide for degradation.

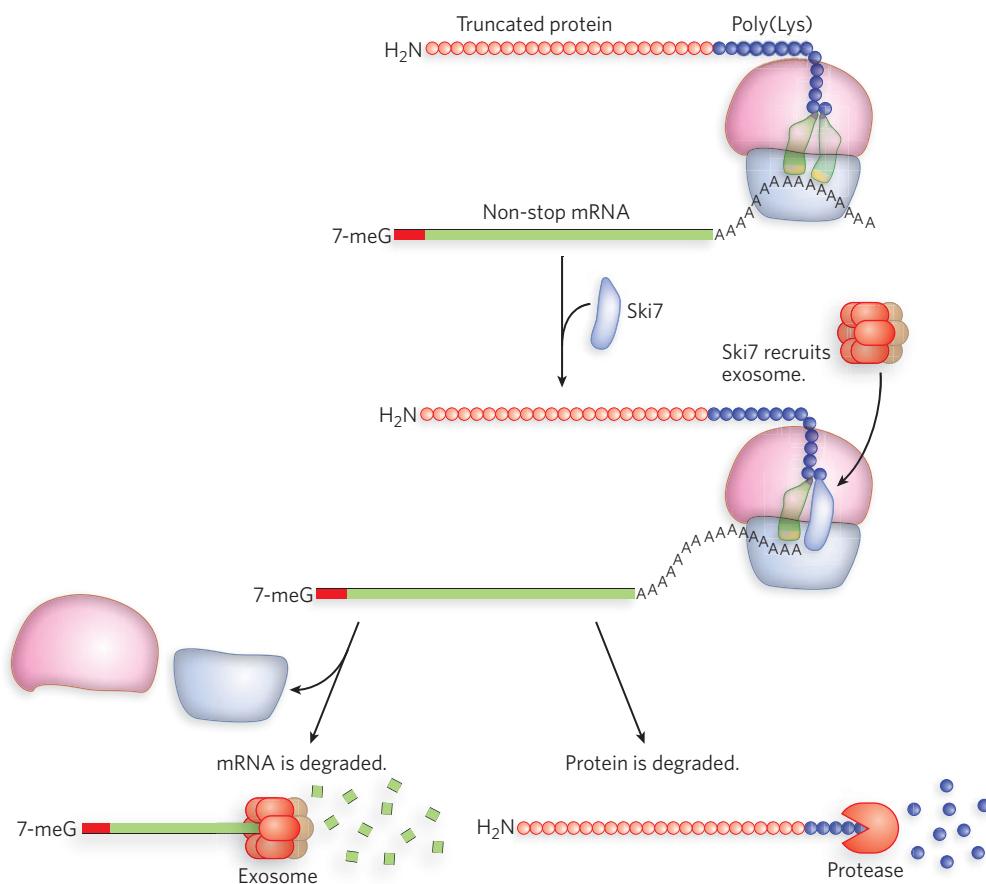


above. Because eukaryotic mRNAs contain a 3' poly(A) tail, an mRNA without a stop codon is translated through the tail to produce a string of Lys residues at the C-terminus of the polypeptide (AAA encodes lysine). The ribosome stalled at the end of the non-stop mRNA binds to the protein Ski7 (related to eRF3), which initiates **non-stop mRNA decay** (Figure 18-32). Ski7 triggers dissociation of the ribosome and degradation of the non-stop mRNA by recruiting the exosome ribonuclease that cleaves in the 3'→5' direction (see Figure 16-25). The defective polypeptide is rapidly degraded by a protease that recognizes the C-terminal poly(Lys) tag.

A process known as **nonsense-mediated mRNA decay** has evolved in eukaryotes to detect and destroy mRNAs that contain a premature stop, or nonsense, codon (Figure 18-33). This process works through the splicing machinery in the nucleus, before the mRNA is

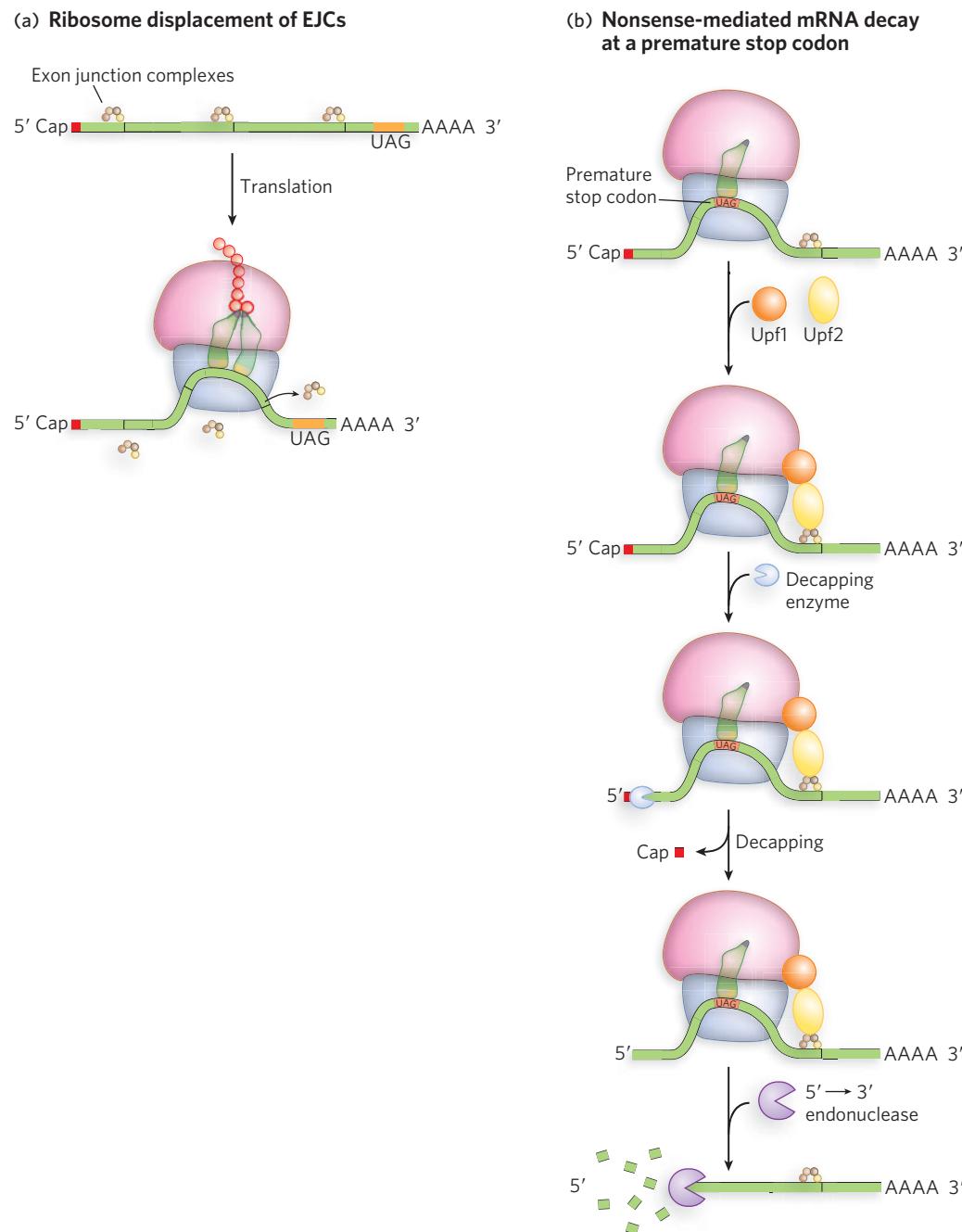
exported to the cytoplasm for translation. During pre-mRNA splicing, the site of each excised intron is marked by the presence of an exon junction complex (EJC), positioned on the 5' side of the exon-exon boundary (see Chapter 16). During the first round of translation, the ribosome removes the EJCs as it moves along the mRNA from the start to the end of the coding region. If a premature stop codon is encountered, however, the ribosome is released early, before all the EJCs have been removed. In this event, Upf1 and Upf2 proteins are recruited by the leftover EJC(s), which in turn recruit an enzyme that cleaves the 5' cap from the mRNA. Without the cap to protect it, the end of the RNA is highly susceptible to degradation by an exoribonuclease that hydrolyzes RNA from the free 5' end.

Note that both of the above processes for removing damaged or truncated mRNA are performed by the ribosome, and therefore the ribosome is an effective



**FIGURE 18-32** The rescue of stalled eukaryotic ribosomes by non-stop mRNA decay.

Ribosomes continue to translate through the poly(A) tail of a non-stop mRNA, resulting in a poly(Lys) protein tail. Ski7 binds to the ribosome and recruits an exosome to degrade the non-stop mRNA and a protease to degrade the protein.



**FIGURE 18-33** Eukaryotic nonsense-mediated mRNA decay. (a) The spliceosome deposits exon junction complexes (EJCs) at the splice sites of an mRNA (see Chapter 16), and the EJCs are normally displaced by the ribosome as it travels along the mRNA. (b) If the ribosome encounters a premature stop codon upstream of an EJC, then Upf1 and Upf2 recruit a decapping enzyme and a 5'→3' exonuclease to promote degradation of the mRNA.

proofreader. The recognition of defective mRNAs thus requires active translation. Recent evidence suggests that such mechanisms play significant roles in the regulation and evolution of protein function, as we'll discuss in detail in Chapter 22.

### SECTION 18.6 SUMMARY

- In bacteria, tmRNA rescues ribosomes stalled on non-stop mRNAs (i.e., lacking a stop codon), playing the role of both tRNA and mRNA in the A site. Translation continues, using the tmRNA

as template and ending at the tmRNA stop codon.

- Eukaryotic non-stop mRNAs are translated through the poly(A) tail, encoding a string of Lys residues that tag the protein for degradation.
- In eukaryotes, pre-mRNA splicing leaves an exon junction complex on spliced mRNAs. If a ribosome encounters a stop codon before it has removed all the EJCs, the stop codon is identified as premature, and the mRNA is tagged for degradation.

## 18.7 Protein Folding, Covalent Modification, and Targeting

To achieve their biologically active forms, new polypeptides must fold into the correct three-dimensional conformation. Before or after folding, the new polypeptide might undergo enzymatic processing, including the removal of one or more amino acids (usually from the N-terminus); the addition of acetyl, phosphoryl, methyl, carboxyl, or other groups to certain amino acid residues; proteolytic cleavage; and/or the attachment of oligosaccharides or prosthetic groups. In this way, the linear, or one-dimensional, genetic message in the mRNA is converted into the three-dimensional structure of the protein.

### Some Proteins Fold Spontaneously, and Others Need Help from Molecular Chaperones

Most proteins fold during translation (after emerging from the ribosome) or immediately after translation, typically beginning with the formation of local secondary structures, including  $\alpha$  helices and  $\beta$  sheets. In a cooperative process, these secondary structural elements then interact, often through hydrophobic interactions, to produce the stable three-dimensional structure of the active protein. In many cases, proteins called **chaperones** catalyze local unfolding and refolding of polypeptide chains to enhance the rate and accuracy of overall folding. By mechanisms sometimes coupled to ATP hydrolysis, chaperones bind transiently to hydrophobic protein segments during folding to ensure that interactions form in the proper order (see Figures 4-23 and 4-24).

After folding into their native conformations, some proteins form intrachain or interchain disulfide bonds, or bridges, between Cys residues. In eukaryotes, disulfide bonds are common in proteins to be exported from cells. The cross-links formed in this way help protect

the molecule's native conformation from denaturation in the extracellular environment, which can differ greatly from the intracellular environment and is generally oxidizing (see Chapter 4, How We Know).

### Covalent Modifications Are Common in Newly Synthesized Proteins

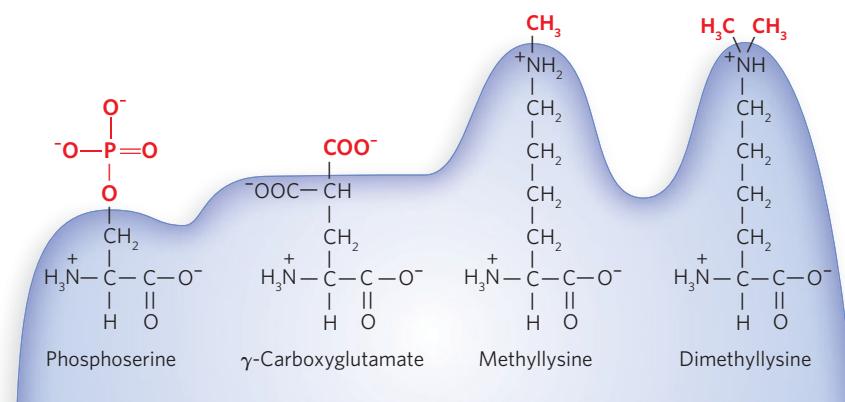
Some newly made proteins, both bacterial and eukaryotic, do not attain their final biologically active conformation until they have been altered by one or more processing reactions known as **posttranslational modifications**. As we've seen, the first residue inserted in all polypeptides is *N*-formylmethionine (in bacteria) or methionine (in eukaryotes). The formyl group, the N-terminal Met residue, and often additional N-terminal (and, in some cases, C-terminal) residues may be removed enzymatically in forming the final functional protein. In as many as 50% of eukaryotic proteins, the amino group of the N-terminal residue is *N*-acetylated after translation. Carboxyl-terminal residues are also sometimes modified. Scientists are still learning the rules to determine which proteins will have which amino acids removed or modified.

The 15 to 30 residues at the N-terminal end of some proteins play a role in directing the protein to its ultimate destination in the cell. After this protein trafficking, these residues are often removed by specific peptidases.

Individual amino acid residues can be modified, either permanently or transiently, with significant effects on functionality, increasing or decreasing the protein's ability to bind other molecules (Figure 18-34). The hydroxyl groups of certain Ser, Thr, and Tyr residues of some proteins are enzymatically phosphorylated by ATP; extra carboxyl groups may be added to Glu residues of some proteins, and Lys, Arg, or Glu residues can be methylated.

The carbohydrate side chains of glycoproteins are attached covalently during or after synthesis of the polypeptide. In some glycoproteins, the carbohydrate side chain is attached enzymatically to Asn residues (*N*-linked oligosaccharides), in others to Ser or Thr residues (*O*-linked oligosaccharides). Many proteins that function extracellularly, as well as the proteoglycans that coat and lubricate mucous membranes, contain oligosaccharide side chains.

Other covalent modifications to proteins include the addition of isoprenyl groups or prosthetic groups, and cleavage by proteases. Many bacterial and eukaryotic proteins require covalently bound prosthetic groups for their activity. Finally, many proteins are initially synthesized as large, inactive precursor polypeptides that are proteolytically trimmed to their smaller, active forms.



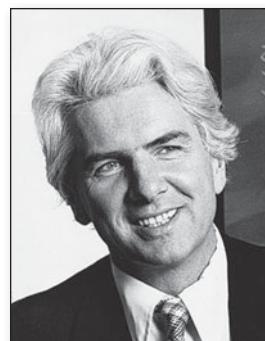
**FIGURE 18-34 Modified amino acid residues.** Amino acid residues can be phosphorylated, carboxylated, or methylated to alter protein function.

### Proteins Are Targeted to Correct Locations during or after Synthesis

Cells are made up of many structures and compartments, and, in the case of eukaryotic cells, they contain organelles, each with specific functions that require distinct sets of proteins and enzymes. These proteins (with the exception of those produced in mitochondria and chloroplasts) are synthesized on ribosomes in the cytosol or on the endoplasmic reticulum (ER). How are proteins directed to their final cellular destinations? Proteins destined for secretion, integration into the plasma membrane, or inclusion in lysosomes generally share the first few steps of a pathway that begins in the ER. Proteins destined for mitochondria, chloroplasts, or the nucleus use three separate mechanisms. And proteins destined for the cytosol simply remain where they are synthesized.

The most important element in many of these targeting pathways is a short sequence of amino acids called a **signal sequence**, whose function was first postulated

by Günter Blobel and colleagues in 1970. The signal sequence directs a protein to its appropriate location in the cell and, for many proteins, is removed during transport or after arrival at its final destination. In proteins directed to mitochondria, chloroplasts, or the ER, the signal sequence is at the N-terminus of the newly synthesized polypeptide. Some signal sequences promote transport to the ER or mitochondria, others to the nucleus.



**Günter Blobel** [Source: Courtesy of Günter Blobel, Rockefeller University.]

Like a promoter sequence, the strength of a signal sequence depends on how similar it is to an unknown, hypothetical ideal sequence. In many cases, the targeting capacity of a particular signal sequence has been confirmed by fusing the sequence from one protein to a second protein and showing that the signal directs the second protein to the location where the first protein is normally found.

### Posttranslational Modification of Many Eukaryotic Proteins Begins in the Endoplasmic Reticulum

The best-characterized targeting system begins in the ER. Most lysosomal, membrane, or secreted proteins are synthesized on ribosomes attached to the ER. These proteins have an N-terminal signal sequence that marks them for translocation into the ER or its lumen; hundreds of such signal sequences have been determined (Figure 18-35). Signal sequences vary in length from 13 to 36 amino acid residues, but all contain 10 to 15 hydrophobic residues, one or more positively charged residues near the N-terminus, and a short polar sequence near the cleavage site (for eventual removal of the signal) at the C-terminus.

The signal sequence itself helps direct the ribosome to the ER, as shown in Figure 18-36. The targeting pathway begins with initiation of protein synthesis on cytosolic ribosomes (step 1). The signal sequence forms early in the synthesis process because it is at the N-terminus, which is synthesized first. As the signal sequence emerges from the ribosome (step 2), the signal and the ribosome itself are bound by the large **signal recognition particle (SRP)**. Then SRP binds GTP (step 3), and halts elongation of the polypeptide when it is about 70 amino acids long and the signal

**Transmembrane proteins**

Human influenza virus A hemagglutinin

Met Lys Ala Lys Leu Leu Val Leu Leu Tyr Ala Phe Val Ala Gly Asp Glu --

Lipoprotein

Met Lys Ala Thr Lys Leu Val Leu Gly Ala Val Ile Leu Gly Ser Thr Leu Leu Ala Gly Cys Ser --

**Secreted proteins**

Bovine growth hormone

Met Met Ala Ala Gly Pro Arg Thr Ser Leu Leu Leu Ala Phe Ala Leu Leu Cys Leu Pro Trp Thr Gln Val Val Gly Ala Phe --

Bee promelittin

Met Lys Phe Leu Val Asn Val Ala Leu Val Phe Met Val Val Tyr Ile Ser Tyr Ile Tyr Ala Ala Phe --

Drosophila glue protein

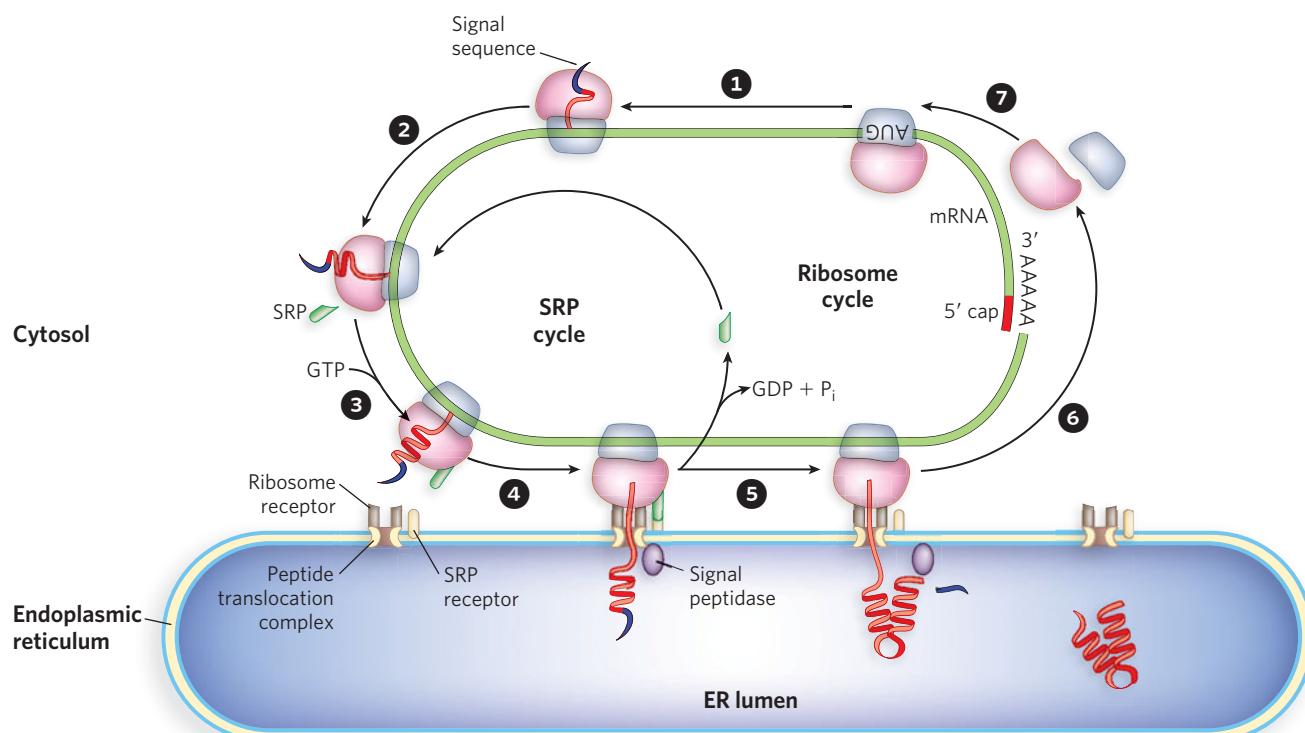
Met Lys Leu Val Val Ala Val Ile Ala Cys Met Leu Ile Gly Phe Ala Asp Pro Ala Ser Gly Cys Lys --

Human preproinsulin

Met Ala Leu Trp Met Arg Leu Leu Pro Leu Leu Ala Leu Leu Ala Leu Trp Gly Pro Asp Pro Ala Ala Phe Val --

**FIGURE 18-35 Signal sequences.** Just a few of the hundreds of known signal sequences are shown here. Hemagglutinin is a transmembrane protein on the surface of the influenza virus. Lipoproteins are components of plasma membranes and organelle membranes, and transport fats in the bloodstream. Proteins destined for secretion include

insulin (as its precursor, preproinsulin), growth hormone, the bee venom toxin melittin (as its precursor promelittin), and a *Drosophila* glue protein used in forming the pupa. Hydrophobic residues are shown in yellow; charged residues in blue.



**FIGURE 18-36 Trafficking of proteins from the cytosol into the ER.** The signal sequence of the nascent polypeptide is bound by SRP, which targets the elongating protein to the ER

lumen. After the polypeptide has been synthesized, the ribosomal subunits dissociate and are recycled. The steps are described in the text.

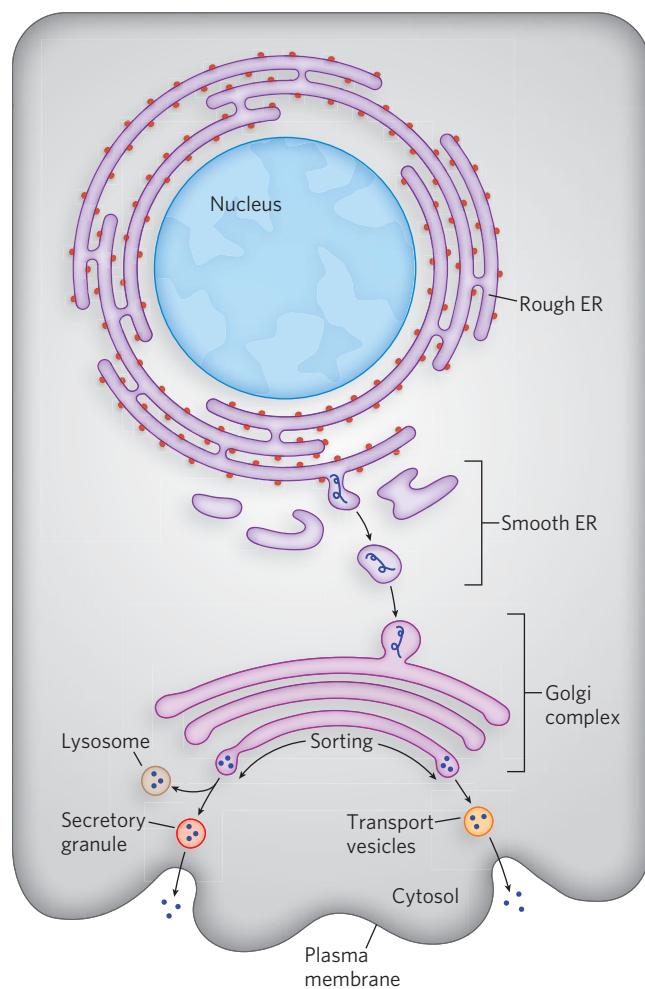
sequence has completely emerged from the ribosome. The GTP-bound SRP directs the ribosome (still bound to the mRNA) and the incomplete polypeptide to GTP-bound SRP receptors in the cytosolic face of the ER (step 4). The nascent polypeptide is delivered to a **peptide translocation complex** in the ER, which may interact directly with the ribosome. SRP dissociates from the ribosome, accompanied by hydrolysis of the GTP bound to the SRP and to the SRP receptor. Elongation of the polypeptide resumes (step 5), with the ATP-driven translocation complex feeding the growing polypeptide into the ER lumen until the complete protein has been synthesized. The signal sequence is removed by a signal peptidase in the ER lumen. The ribosome dissociates (step 6) and is recycled (step 7).

### Glycosylation Plays a Key Role in Eukaryotic Protein Targeting

In the ER lumen, newly synthesized proteins are further modified in several ways. Following removal of signal sequences, polypeptides are folded, disulfide bonds are formed, and many proteins are glycosylated. In many glycoproteins, Asn residues are *N*-linked to a wide variety of oligosaccharides, but the pathways by which they form share common steps. Several antibiotics, such as **tunicamycin**, act by interfering with this process and have aided in elucidating the steps of protein glycosylation. A few proteins are *O*-glycosylated in the ER, but most *O*-glycosylation occurs in the Golgi complex (Golgi apparatus) or in the cytosol (for proteins that do not enter the ER).

Once a protein is suitably modified, it can move to its final intracellular destination. Proteins travel from the ER to the Golgi complex in transport vesicles (Figure 18-37). In the Golgi complex, some proteins, as we've noted, are *O*-glycosylated, and some *N*-linked oligosaccharides are further modified. By mechanisms not yet fully understood, the Golgi complex also sorts proteins and sends them to their final destinations. The processes that segregate proteins targeted for secretion from those targeted for the plasma membrane or lysosomes must distinguish among these proteins on the basis of structural features other than signal sequences, which were removed in the ER lumen.

The pathways that target proteins to mitochondria and chloroplasts also rely on N-terminal signal sequences. Although mitochondria and chloroplasts contain DNA, most of their proteins are encoded by nuclear DNA and must be targeted to the appropriate organelle. Unlike other targeting pathways, however, the mitochondrial and chloroplast pathways begin only



**FIGURE 18-37** Movement of proteins destined for membranes or secretion. After synthesis on ribosomes of the rough ER and targeting to the ER lumen, proteins travel in transport vesicles through the Golgi complex. Within the Golgi complex, the proteins may be further modified before they are sorted and shipped to their final destinations in secretory granules or transport vesicles.

after protein synthesis is complete. Cytosolic chaperone proteins bind to precursor proteins and deliver them to receptors on the exterior surface of the target organelle. Specialized translocation mechanisms then transport each protein to its final destination in the organelle, after which the signal sequence is removed.

### Signal Sequences for Nuclear Transport Are Not Removed

Many proteins and nucleic acids move into and out of the nucleus through nuclear pores. RNA molecules synthesized in the nucleus are exported to the cytosol

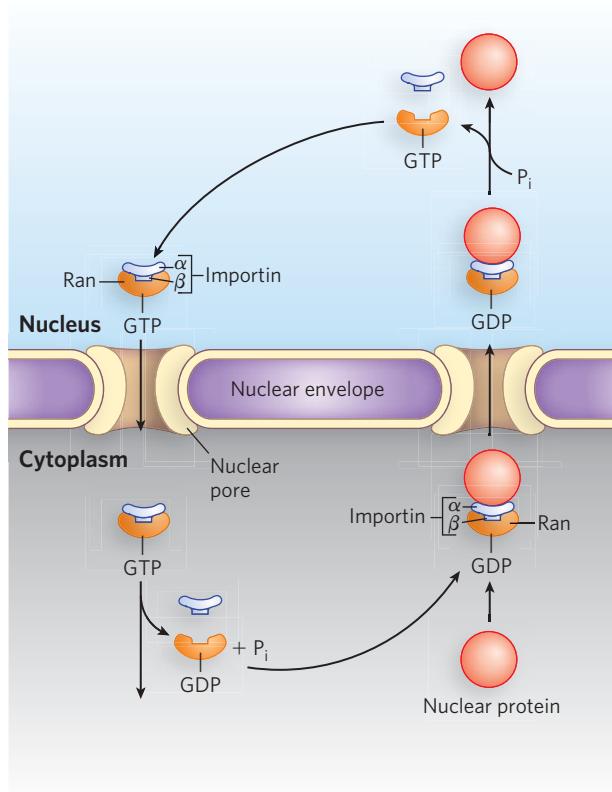
for translation (see Chapter 16). Ribosomal proteins synthesized on cytosolic ribosomes are imported into the nucleus and assembled into 60S and 40S ribosomal subunits in the nucleolus, where the rRNAs are produced. Completed subunits are then exported back to the cytosol. A variety of nuclear proteins are synthesized in the cytosol and imported into the nucleus (e.g., RNA and DNA polymerases, histones, topoisomerases, and proteins that regulate gene expression). All of this traffic is modulated by a complex system of molecular signals and transport proteins.

In multicellular eukaryotes, cell division poses a problem for nuclear proteins. At each cell division, the nuclear envelope breaks down, and after division is completed and the nuclear envelope re-forms, the dispersed nuclear proteins must be re-imported. To allow this repeated nuclear importation, the signal sequence that targets a protein to the nucleus—the **nuclear localization sequence (NLS)**—must remain on the protein after it arrives at its destination. An NLS, unlike other signal sequences, may be located almost anywhere along the primary sequence of the protein. The amino acid sequences of NLSs can vary considerably, but many consist of 4 to 8 residues and include several consecutive basic residues (Arg or Lys).

Nuclear import is mediated by proteins that cycle between the cytosol and the nucleus, including importin  $\alpha$  and  $\beta$  and the Ran GTPase (Figure 18-38), in a mechanism like that discussed for nuclear RNA export and import in Chapter 16. A heterodimer of importin  $\alpha$  and  $\beta$  functions as a carrier for cargo proteins targeted to the nucleus, with the  $\alpha$  subunit binding cargo proteins in the cytosol. The importin-cargo complex docks at a nuclear pore and is translocated through the pore by an energy-dependent mechanism that requires the Ran GTPase. Once inside the nucleus, interaction with Ran-GTP triggers a change in the conformation of importin that leads to release of the cargo protein. The importin-Ran-GTP complex then passes through the nuclear pore back into the cytosol. Here, the Ran-binding protein binds to Ran and releases importin, and a GTPase-activating protein stimulates conversion of Ran-GTP to Ran-GDP.

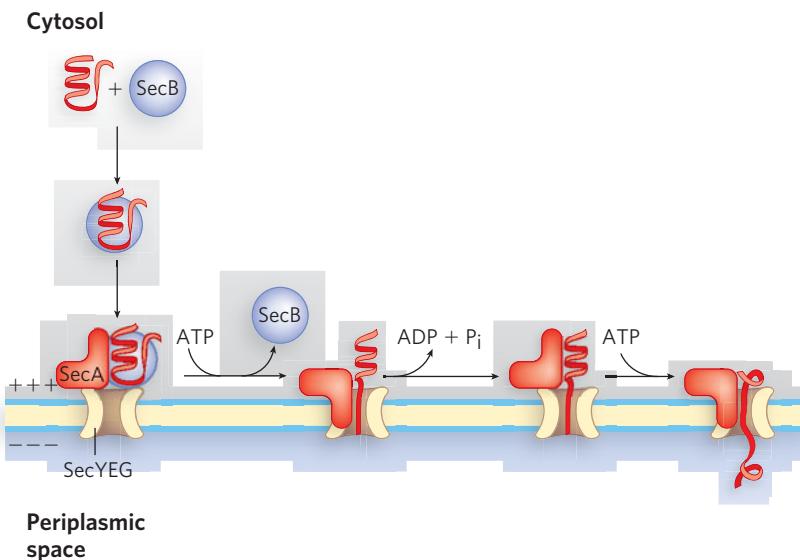
## Bacteria Also Use Signal Sequences for Protein Targeting

Bacteria can target proteins to their inner or outer membranes, to the periplasmic space, or to the extracellular medium. They use signal sequences at the N-terminus of the proteins, much like those on eukaryotic proteins targeted to the ER, mitochondria, and chloroplasts.



**FIGURE 18-38** Targeting of nuclear proteins. Importins bind a nuclear protein through its NLS and transport the protein through a nuclear pore, with the help of Ran-GDP. In the nucleus, Ran exchanges its GDP for GTP, facilitating release of the nuclear protein. The importins and Ran-GTP then shuttle back to the cytoplasm through a nuclear pore.

Most proteins exported from *E. coli* make use of the pathway shown in Figure 18-39. Following translation, the N-terminal signal sequence may impede folding of a protein to be exported. A soluble chaperone protein, SecB, binds to the signal sequence or to other features of the protein's incompletely folded structure. The bound protein is then delivered to SecA, a protein associated with the inner surface of the plasma membrane. SecA acts as both a receptor and a translocating ATPase. Released from SecB and bound to SecA, the protein is delivered to a translocation complex in the membrane, made up of SecY, SecE, and SecG, and is translocated stepwise through the membrane at the SecYEG complex, about 20 amino acid residues at a time. Each step is facilitated by the hydrolysis of ATP, catalyzed by SecA. Although most exported bacterial proteins use this pathway, some follow an alternative route that uses signal recognition and receptor proteins homologous to components of the eukaryotic SRP and SRP receptor.



**FIGURE 18-39** A model for protein export in bacteria. A partially folded protein with a signal sequence is bound by SecB, then transferred to SecA and SecYEG, the latter a component of the bacterial plasma membrane. SecYEG pushes the protein through the membrane stepwise, and the protein folds on the other side of the membrane, in the periplasmic space.

### SECTION 18.7 SUMMARY

- Polypeptides fold into their active, three-dimensional forms during or immediately after synthesis, often with the help of ATP-dependent chaperone proteins. Many proteins are further processed by posttranslational modification reactions that add functional groups, such as phosphates or sugars.
- During or immediately following synthesis, many proteins are directed to specified cellular locations. One targeting mechanism involves a peptide signal sequence, generally at the N-terminus of a newly synthesized protein.
- In eukaryotic cells, one class of signal sequences is recognized by the signal recognition particle, which binds the signal sequence as soon as it appears on the ribosome and transfers the entire ribosome and incomplete polypeptide to the ER. The peptides are moved into the ER lumen, where they may be modified and moved to the Golgi complex, then sorted and sent to lysosomes, the plasma membrane, or transport vesicles.
- Proteins targeted to mitochondria and chloroplasts, and those destined for export in bacterial cells, also make use of an N-terminal signal sequence. Specific enzymes remove the signal once the protein reaches its destination.

- Nuclear localization signals are not removed, because nuclear proteins must be relocalized to the nucleus each time the cell divides. Protein import requires importins, the Ran GTPase, and GTP.
- Bacterial proteins may be targeted to the plasma membrane by a signal sequence.

### Unanswered Questions

Although many details of bacterial protein synthesis are known in exquisite detail, how translation assists in protein folding and how ribosomes work together within polysomes are still important mechanisms to solve. In addition, researchers have yet to determine the mechanisms by which eukaryotic translation is initiated and regulated.

#### 1. How is translation rate coupled to protein folding?

The physics and kinetics of translation clearly affect how proteins fold. For example, many mRNAs include rare codons for which there are few available matching tRNAs, in order to stall ribosomes and hence provide time for proteins to fold. Understanding how this works is important for determining how proteins attain their correct structure in cells. Forces produced by ribosomes as they traverse an mRNA may also be important for melting RNA structures that could otherwise impede translation.

**2. How do the crowded conditions inside cells influence translation rates and accuracy?** Most of the experiments that have probed translation mechanisms have been performed using purified ribosomes under relatively dilute conditions ( $\sim 1 \mu\text{M}$ ). In rapidly growing cells, however, ribosomes may be present at up to 100 times this concentration. Computer simulations that model the process of translation in the presence of various cellular factors may help determine the impact of

molecular crowding on the rate of protein synthesis.

**3. How is eukaryotic translation initiation regulated?**

Translation initiation is much more complex in eukaryotic cells than in bacterial cells. It is critical to understand how the eukaryotic system works, because much protein synthesis regulation occurs at the level of initiation. Molecular structures of the eukaryotic ribosome, coupled with more detailed biochemical studies, will help reveal the details of this process.

# How We Know

## The Ribosome Is a Ribozyme

**Noller, H.F., and J.B. Chaires.** 1972. Functional modification of 16S ribosomal RNA by kethoxal. *Proc. Natl. Acad. Sci. USA* 69:3115–3118.

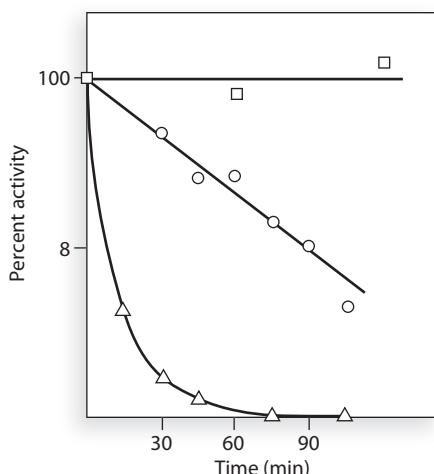
**Noller H.F., V. Hoffarth, and L. Zimniak.** 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* 256:1416–1419.

A series of experiments conducted in the early 1970s by Harry Noller provided the first evidence that ribosomal RNA, rather than ribosomal protein, is responsible for catalyzing peptide bond formation during protein synthesis. Using chemicals that react with side chains in proteins or nucleotides, Noller discovered that the 30S ribosomal subunit of bacteria could be inactivated by a reagent called kethoxal, which primarily attacks guanosine nucleotides in RNA.

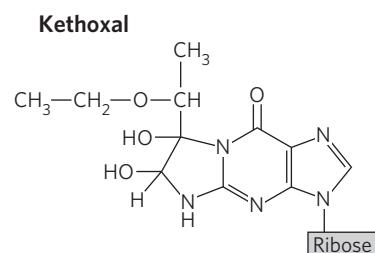
Bacterial ribosomes were purified and separated into their two subunits. After addition of kethoxal to the 30S subunits, these were mixed with unmodified 50S subunits and the ribosomes tested for the ability to stimulate *in vitro* protein synthesis, using tRNA and a poly(U) mRNA template. In contrast to samples with untreated 30S subunits, the kethoxal-treated sample rapidly lost activity (Figure 1). Analysis of the inactivated

ribosomes showed that modification of just six G nucleotides in the rRNA was sufficient to inhibit the peptidyl transferase reaction. Inactivation was slower in the presence of bound tRNA, however, leading to the conclusion that kethoxal interferes with protein synthesis by modifying the binding site for tRNA on the 30S subunit.

In 1992, Noller and his colleagues published a study showing that virtually all the r-proteins could be removed from the ribosome with only small effects on protein-synthesizing activity, whereas damage to the rRNA destroyed the ribosome's catalytic properties. In 2000, the high-resolution crystal structure of the 50S ribosomal subunit revealed that the active site responsible for peptide bond formation is composed entirely of rRNA. Thus, the structural data confirmed what had long been suspected based on biochemical evidence: the ribosome is a ribozyme.



**FIGURE 1** Chemical modification of rRNA inactivates the 30S ribosomal subunit. The graph shows the percentage of normal activity (synthesis of polypeptide) over time. In the absence of kethoxal (squares), the 30S subunit retained activity; in the presence of kethoxal (circles), activity was



impaired over time. A control reaction lacking tRNA and mRNA had virtually no activity (triangles). The chemical structure of a kethoxal-modified G nucleotide is also shown. [Source: Adapted from H. F. Noller and J. B. Chaires, *Proc. Natl. Acad. Sci. USA* 69:3115–3118, 1972.]

## Ribosomes Check the Accuracy of Codon-Anticodon Pairing, but Not the Identity of the Amino Acid

Chapeville, F., F. Lipmann, G. Von Ehrenstein, B. Weisblum, W.J. Ray Jr., and S. Benzer.

1962. On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. USA* 48:1086–1092.

Zaher, H.S., and R. Green. 2009. Quality control by the ribosome following peptide bond formation. *Nature* 457:161–166.



Rachel Green

[Source:

© Paul Fetter 2007.]

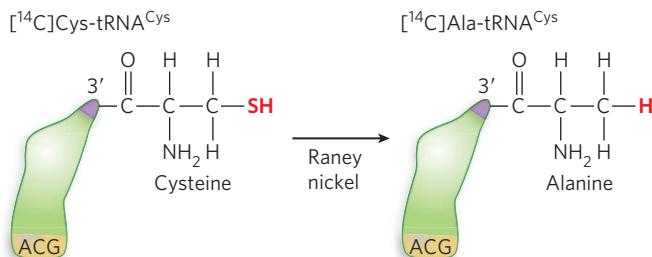
Translation relies on aminoacyl-tRNA synthetases to ensure the correct charging of tRNAs, because the ribosome does not discriminate between correctly and incorrectly charged tRNAs during protein synthesis. A classic experiment by Seymour Benzer and his colleagues demonstrated that if a tRNA is aminoacylated with the wrong amino acid, this incorrect amino acid is efficiently incorporated into a protein in response to the codon normally recognized by the tRNA.

This experiment, performed in 1962, used  $^{14}\text{C}$  labeling to track how amino acids attached to particular tRNAs were incorporated into polypeptides. For example, correctly charged  $[^{14}\text{C}]\text{Cys-tRNA}^{\text{Cys}}$  was chemically treated with Raney nickel (a nickel-aluminum alloy catalyst) to form a mischarged  $[^{14}\text{C}]\text{Ala-tRNA}^{\text{Cys}}$  (Figure 2a). Poly(UG) mRNA has UGU codons that, with wobble base pairing in the third position, match the ACG anticodon for  $\text{tRNA}^{\text{Cys}}$ , but has no codons for alanine. Using this mRNA, the researchers found that the ribosome efficiently incorporated  $[^{14}\text{C}]$ alanine into acid-insoluble polypeptide (Figure 2b). This result was an elegant and straightforward demonstration that the ribosome does not proofread tRNAs for correct aminoacylation.

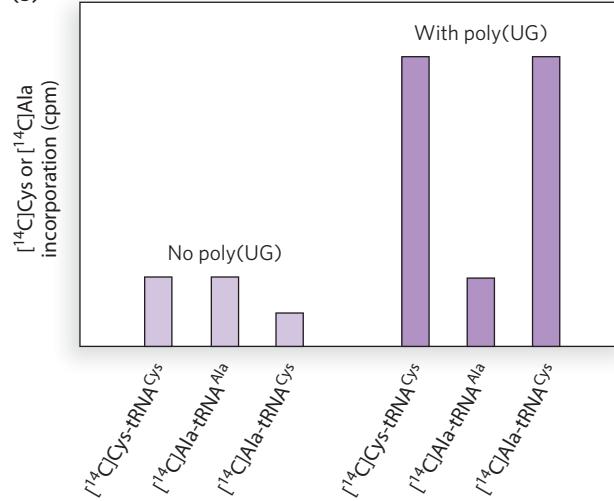
For many years, protein synthesis was thought to rely on the combined accuracy of tRNA aminoacylation and aminoacyl-tRNA selection by the ribosome in cooperation with the GTPase elongation factor EF-Tu (see Section 18.4). These two processes operate before peptide bond formation to ensure that only correctly charged and correctly matched tRNAs enter the ribosomal A site.

More recently, an additional mechanism occurring after peptidyl transfer was found to contribute to the accuracy of protein synthesis. Using a well-defined in vitro bacterial translation system, Rachel Green and Hani Zaher showed that incorporation of an amino acid from a mismatched aminoacyl-tRNA into the elongating polypeptide leads to a general loss of specificity in the ribosomal A site. The resulting propagation of errors leads to early termination of protein synthesis, avoiding the production of a complete protein containing incorrect amino acids.

### (a) Conversion of Cys to Ala by Raney nickel



### (b)



**FIGURE 2** The ribosome does not proofread aminoacylated tRNAs. (a) The  $^{14}\text{C}$ -labeled Cys residue of  $\text{Cys-tRNA}^{\text{Cys}}$  is converted to Ala by Raney nickel, which removes the sulfur from cysteine. (b) Polypeptide synthesis with and without poly(UG) mRNA, using correctly charged  $[^{14}\text{C}]\text{Cys-tRNA}^{\text{Cys}}$ , or  $[^{14}\text{C}]\text{Ala-tRNA}^{\text{Ala}}$  (as control), or mischarged  $[^{14}\text{C}]\text{Ala-tRNA}^{\text{Cys}}$ . The results in the presence of UGU codons (right three columns) show incorporation of labeled Cys with Cys-tRNA<sup>Cys</sup> (left); negligible incorporation of labeled Ala with Ala-tRNA<sup>Ala</sup> (middle), and incorporation of labeled Ala with Ala-tRNA<sup>Cys</sup> (right). [Source: Adapted from F. Chapeville et al., *Proc. Natl. Acad. Sci. USA* 48:1086–1092, 1962.]

## Key Terms

ribosome, p. 616	Shine-Dalgarno sequence, p. 629	termination factor, p. 643
ribosomal protein (r-protein), p. 617	Kozak sequence, p. 631	release factor, p. 643
ribosomal RNA (rRNA), p. 617	<i>N</i> -formylmethionyl-tRNA <sup>fMet</sup> , p. 631	tmRNA, p. 650
polysome, p. 620	initiation complex, p. 634	nonsense-mediated mRNA decay, p. 652
polyribosome, p. 620	scanning, p. 635	chaperone, p. 654
A site, p. 621	elongation, p. 638	posttranslational modification, p. 654
P site, p. 621	elongation factor, p. 638	signal sequence, p. 655
E site, p. 621	peptidyl transferase reaction, p. 639	signal recognition particle (SRP), p. 655
initiation, p. 629	translocation, p. 640	nuclear localization sequence (NLS), p. 658
initiation codon, p. 629	termination codon, p. 642	
start codon, p. 629	stop codon, p. 642	
initiation factor, p. 629	termination, p. 642	

## Problems

- On average, how many phosphoanhydride bonds are hydrolyzed in the course of synthesizing a 400 amino acid protein? Assume that you begin with the mature mRNA, ribosomal subunits, tRNAs, free amino acids, and all necessary factors.
- Name the type of chemical bonds that link (a) adjacent amino acids in a protein; (b) an amino acid to tRNA; (c) adjacent nucleotides in RNA; (d) a codon in mRNA to an anticodon in tRNA; (e) the two subunits of a ribosome.
- Discuss the advantages to the cell of having multiple ribosomes translating a single mRNA molecule.
- The amino acid hydroxyproline, which is critical to the structure of collagen and certain other proteins, has no representative codon in the genetic code. How might it be incorporated into proteins?
- The isoleucyl-tRNA synthetase has a proofreading function that improves the fidelity of the aminoacylation reaction, whereas the histidyl-tRNA synthetase lacks such a proofreading function. Explain why.
- As described in Chapter 11, some DNA polymerases have proofreading activities. After a nucleotide is added to a growing nucleic acid chain, it can be removed (if incorrectly paired with the template) by hydrolysis of the phosphodiester bond that links it to the growing polymer. Ribosomes do not have similar proofreading activities; they cannot remove the last amino acid added to a growing polypeptide, regardless of whether it was correctly added or not. If ribosomes possessed such a proofreading function, would cleavage of the bond linking the last amino acid to the polymer have any effect on the rest of the polypeptide? Why or why not?
- A researcher isolates mutant variants of the bacterial translation factors IF-2, EF-Tu, and EF-G. In each case, the mutation allows proper folding of the protein and binding of GTP, but does not allow GTP hydrolysis. At what stage would translation be blocked by each mutant protein?
- Some aminoacyl-tRNA synthetases do not recognize and bind the anticodon of their cognate tRNAs; they use other structural features of the tRNAs to impart binding specificity. The tRNAs for alanine apparently fall into this category.
  - What features of tRNA<sup>Ala</sup> are recognized by Ala-tRNA synthetase?
  - Describe the consequences of a C-to-G mutation in the third position of the anticodon of tRNA<sup>Ala</sup>.
  - What other kinds of mutations might have similar effects?
  - Mutations of these types are never found in natural populations of organisms. Why? (Hint: Consider what might happen both to individual proteins and to the organism as a whole.)
- The gene for a eukaryotic polypeptide of 300 amino acid residues is altered so that the polypeptide has an N-terminal signal sequence recognized by SRP and an internal nuclear localization signal, beginning at residue 150. Where is the protein likely to be found in the cell?
- Chloramphenicol binds to bacterial ribosomes and is a potent inhibitor of bacterial protein synthesis, but it does not inhibit the cytosolic ribosomes in eukaryotes. However, chloramphenicol is rarely used as a human antibiotic because of its severe toxicity. Suggest a reason for chloramphenicol toxicity in humans.

## Data Analysis Problems

**Chapeville, F., F. Lipmann, G. Von Ehrenstein, B.**

**Weisblum, W.J. Ray Jr., and S. Benzer. 1962.** On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. USA* 48:1086-1092.

**11.** The experiments demonstrating that ribosomes check the anticodon of the tRNA, but not the amino acid attached to it, were carried out well before the entire genetic code was defined. The research team, led by Seymour Benzer, chose tRNA<sup>Cys</sup> for their demonstration, for several reasons—including the ability to reduce Cys to Ala in a simple chemical reaction using Raney nickel. For their experiment showing that Ala could be incorporated into polypeptide in place of Cys if present in the translation reaction mix as Cys-tRNA<sup>Ala</sup>, the researchers first used polynucleotide phosphorylase to synthesize an RNA polymer consisting of only U and G residues. They knew from previous reports that this polymer would yield polypeptides with the amino acid composition they required.

- (a) Which amino acids are incorporated into a polypeptide specified by a random sequence of U and G?
- (b) Why was tRNA<sup>Cys</sup> a good choice for the experiment?

Benzer and colleagues showed that incorporation of radioactively labeled amino acids into polypeptide was linearly related to the amount of labeled aminoacyl-tRNA added to the experiment. They treated one preparation of [<sup>14</sup>C]Cys-tRNA<sup>Cys</sup> with Raney nickel; about 60% of the charged tRNA was reduced to [<sup>14</sup>C]Ala-tRNA<sup>Cys</sup>. They then plotted incorporation of radioactivity, over time, in preparations with [<sup>14</sup>C]Cys-tRNA<sup>Cys</sup> (Figure 1) and preparations with [<sup>14</sup>C]Ala-tRNA<sup>Cys</sup> (Figure 2), measured in

counts per minute (cpm) in the trichloroacetic acid-precipitated (TCA-ppt) preparations.

- (c) Given the amount of labeled aminoacyl-tRNA added (indicated on the vertical axis of the graphs) and the amount incorporated into polypeptide, does this result support the contention that Ala-tRNA<sup>Cys</sup> is incorporated at codons specifying Cys? Why or why not?

The researchers next treated a preparation of [<sup>14</sup>C]Phe-tRNA<sup>Phe</sup> with Raney nickel, then compared the incorporation of [<sup>14</sup>C]Phe into polypeptide in response to a poly(U) RNA before and after Raney nickel treatment. No effect of the Raney nickel treatment was seen.

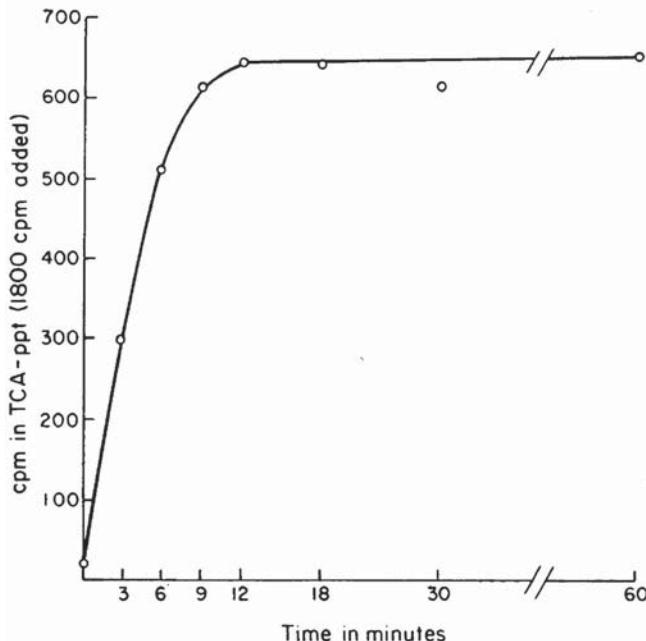
- (d) Why was this experiment carried out?

The polypeptides synthesized using poly(UG) and Ala-tRNA<sup>Cys</sup> were then digested to single amino acids, and the products were analyzed to directly demonstrate that about 60% of the labeled residues were Ala.

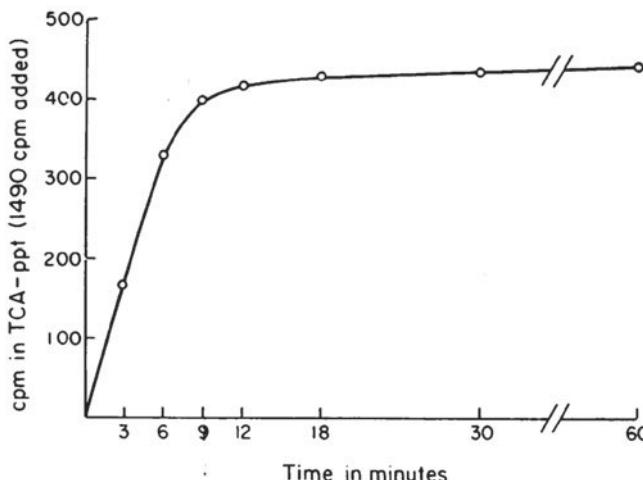
- (e) Why was this experiment needed, given the results shown in Figures 1 and 2?

**Noller, H.F., V. Hoffarth, and L. Zimniak. 1992.** Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* 256:1416-1419.

- 12.** Experiments strongly suggesting that the peptidyl transferase activity of ribosomes is due to rRNA, not r-proteins, were carried out by Harry Noller and colleagues in 1992 (see Moment of Discovery and How We Know). The work required the removal of most or all of the proteins in a ribosome without eliminating peptidyl transferase activity. The basic reaction used is shown in Figure 1. A <sup>35</sup>S-labeled fMet residue (shown here as f-[<sup>35</sup>S]Met) was attached to the 3' end of the hexameric oligonucleotide



**FIGURE 1** [<sup>14</sup>C]Cys-tRNA<sup>Cys</sup> incorporation.



**FIGURE 2** [<sup>14</sup>C]Ala-tRNA<sup>Cys</sup> incorporation.

5'-CAACCA. This labeled fMet was transferred to puromycin in the presence of the 50S bacterial ribosome subunit, when the solvent contained about 33% methanol (with Mg<sup>2+</sup> and K<sup>+</sup> ions).

- (a) What do the substrates required for this in vitro reaction suggest about the binding properties of the 50S ribosomal subunit?
- (b) What is special about the hexanucleotide sequence attached to the labeled fMet residue?
- (c) In which ribosomal site must the fMet-oligonucleotide bind, and in which site must the puromycin bind?
- (d) What advantages does this reaction have for this particular study, relative to protein synthesis with complete ribosomes and mRNA and charged tRNAs?

Using this assay, the researchers demonstrated that the 50S subunit or the intact 70S ribosome from *E. coli* would catalyze the reaction even after extraction with sodium dodecyl sulfate (a detergent that denatures proteins) and extensive treatment with proteinase K (which degrades virtually all proteins). Extraction with phenol (a treatment that separates protein from nucleic acid),

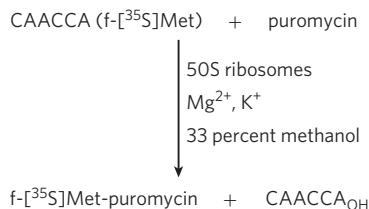


FIGURE 1

however, led to a loss in activity. Subsequently, the investigators found that 50S subunits derived from the thermophilic bacterium *Thermus aquaticus* did *not* lose activity when extracted with phenol.

- (e) Why did the researchers explore the 50S subunits from *T. aquaticus*?

Noller and colleagues went on to test the sensitivity of the reaction to chloramphenicol, carbomycin (another ribosome-binding inhibitor with a mechanism similar to chloramphenicol), and a general ribonuclease called RNase T1, with or without (+ or -) SDS, proteinase K (PK), or phenol treatment. The results are shown in Figure 2, an autoradiograph of the product after high-voltage paper electrophoresis.

- (f) What conclusions can you derive from the results of this experiment?
- (g) The conclusion that ribosomes are ribozymes did not achieve general acceptance until elucidation of the three-dimensional structure of a bacterial 50S ribosomal subunit. Suggest a reason for the delay in general acceptance.

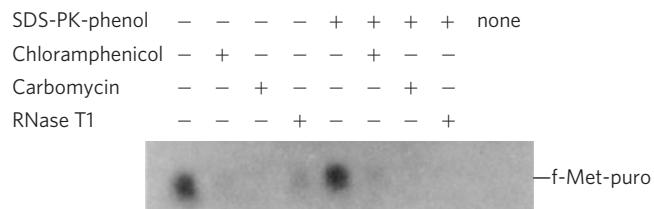


FIGURE 2

## Additional Reading

### The Ribosome

- Ban, N., P. Nissen, J. Hansen, P.B. Moore, and T.A. Steitz. 2000.** The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289:905–920.
- Mathews, M.B., N. Sonenberg, and J.W.B. Hershey. 2006.** *Translational Control in Biology and Medicine*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Schluelenzen, F., A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. 2000.** Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* 102:615–623.

**Schuwirth, B.S., M.A. Borovinskaya, C.W. Hau, W. Zhang, A. Vila-Sanjurjo, J.M. Holton, and J.H. Cate. 2005.** Structures of the bacterial ribosome at 3.5 Å resolution. *Science* 310:827–834.

**Wimberly, B.T., D.E. Brodersen, W.M. Clemons Jr., R.J. Morgan-Warren, A.P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. 2000.** Structure of the 30S ribosomal subunit. *Nature* 407:327–339.

### Initiation of Protein Synthesis

- Fraser, C.S., and J.A. Doudna. 2007.** Structural and mechanistic insights into hepatitis C viral translation initiation. *Nat. Rev. Microbiol.* 5:29–38.

**Jackson, R.J., C.U. Hellen, and T.V. Pestova.** 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* 11:113–127.

#### Elongation of the Polypeptide Chain

**Rodnina, M.V., M. Beringer, and W. Wintermeyer.** 2007. How ribosomes make peptide bonds. *Trends Biochem. Sci.* 32:20–26.

#### Termination of Protein Synthesis and Recycling of Synthesis Machinery

**Youngman, E.M., M.E. McDonald, and R. Green.** 2008. Peptide release on the ribosome: Mechanism and implica-

tions for translational control. *Annu. Rev. Microbiol.* 62:353–373.

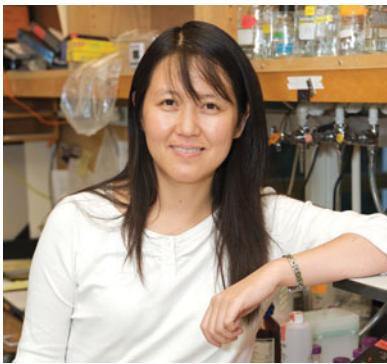
#### Translation-Coupled Removal of Defective mRNA

**Isken, O., and L.E. Maquat.** 2007. Quality control of eukaryotic mRNA: Safeguarding cells from abnormal mRNA function. *Genes Dev.* 21:1833–1856.

#### Protein Folding, Covalent Modification, and Targeting

**Reed, R., and H. Cheng.** 2005. TREX, SR proteins and export of mRNA. *Curr. Opin. Cell Biol.* 17:269–273.

# Regulating the Flow of Information



**Lin He** [Source: Courtesy of Lin He.]

## Moment of Discovery

One of the most exciting moments in my career happened when I was working as a postdoctoral fellow with Greg Hannon at the Cold Spring Harbor Laboratories. We wondered whether microRNAs—short, regulatory RNAs that control the expression of some eukaryotic genes—might also be involved in promoting the development of cancer.

Indeed, one cluster of microRNAs, the miR17-92 polycistron, is located in a region of DNA that is amplified in human B-cell lymphomas. Furthermore, we found that B-cell lymphoma tissue samples and cultured cells contained much higher levels of primary or mature microRNAs derived from the miR17-92 locus compared with those found in normal, noncancerous tissues.

To test whether miR17-92 overexpression could actually accelerate tumor development, I used a mouse model in which hematopoietic stem cells, the precursors to B cells, were infected with a retrovirus encoding the miR17-92 cluster, along with the gene for green fluorescent protein (GFP), which serves as a convenient marker of cells expressing the infected virus because the cells turn green. Our first experiments yielded only three mice that developed tumors. The next step was to determine whether the tumors came specifically from hematopoietic cells that were overexpressing the miR17-92 RNA. By the time I dissected the mouse tumors, made a suspension of the cells, and sent the samples to be analyzed by fluorescence-activated cell sorting (FACS), it was well past midnight—and the FACS sorting couldn't be completed until the next day.

After many anxious hours of waiting, we got the results the next morning: all the tumor cells were green! This was incredibly exciting because it indicated that these tumors came from cells that overexpressed the miR17-92 cluster of microRNAs, one of the first examples where functional RNA genes can promote tumorigenesis. At that moment I knew this would be a very promising direction for future research on cancer development.

—**Lin He**, on discovering that microRNA overexpression accelerates tumor development

**19.1 Regulation of Transcription Initiation 669**

**19.2 The Structural Basis of Transcriptional Regulation 678**

**19.3 Posttranscriptional Regulation of Gene Expression 684**

All cells, whether single-celled bacteria or components of a complex multicellular organism such as a human, contain every gene in that organism's genome. However, cells need the products of only some of these genes, and even those that they do need, they may need only under certain conditions. A bacterium doesn't need the enzymes required to metabolize lactose when its current food source is glucose. Different cell types of multicellular organisms use different subsets of gene products to carry out their various functions—a liver cell does not need the same gene products as a muscle cell.

The relative amounts of each gene product required by a cell can also vary considerably. In an actively growing cell, for example, ribosomes are in high demand and can account for almost half of the cell's dry weight, whereas only a few molecules of certain DNA repair proteins are required to do the necessary repair job. From an energy standpoint, having all gene products present in the highest possible amounts at all times would overburden the cell, considering the resources required to synthesize RNA and protein. Therefore, the expression of genes must be regulated so that their products are present in the right amount and only when they are needed.

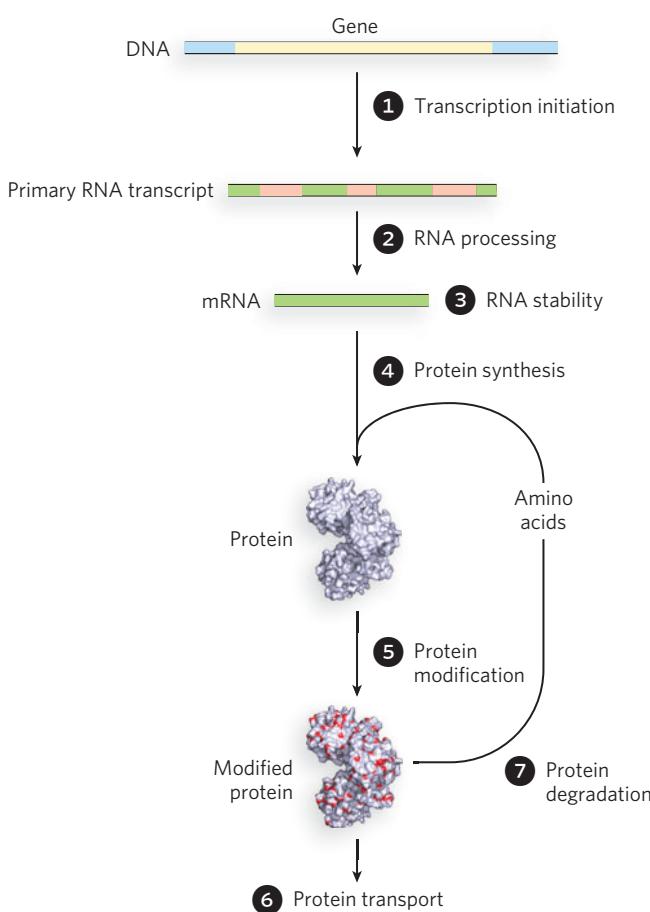
Cells have evolved to respond to environmental changes and to adapt quickly to new growth conditions. This is how organisms can colonize a wide variety of biological habitats. Changing conditions range from the availability and use of different food sources to complex developmental regulatory programs in multicellular organisms, for which some gene products may be needed for a surprisingly brief time and in only a few cells.

Gene regulation is also important to the prevention of diseases, including cancer. In multicellular organisms, selective cell proliferation and destruction are critical for maintaining the proper levels of each cell type. The cell division cycle and programmed cell death pathways are exquisitely controlled by genes that promote or prevent these processes in response to cellular signals. When these regulatory genes are compromised, uncontrolled cell division can lead to the development of tumors. For example, *p53* is a regulatory gene that plays a role in initiating programmed cell death. Loss-of-function mutations in *p53* are found in about 50% of all lung cancers, 70% of colon cancers, and 30% to 50% of breast cancers.

Gene expression can be regulated at many different points in the synthesis of a functional RNA or protein. Transcription initiation is the most widely used regulatory point in both bacteria and eukaryotes, as this is the least costly way to control a gene. Initiation of

transcription occurs at the very beginning of the synthetic pathway, before the investment in energy needed to make either RNA or protein. Nevertheless, mechanisms to regulate gene expression are found at virtually every point along the biosynthetic pathway. The points of regulation, shown in Figure 19-1, include (1) transcription initiation, (2) RNA processing, (3) RNA stability, (4) protein synthesis, (5) protein modification, (6) protein transport, and (7) protein degradation.

It is obviously important for cells to use their resources efficiently and not to waste energy synthesizing gene products that they do not need in a particular growth environment. But just as critical as efficiency is adaptability and thus control: cells must be able to respond rapidly to changes in the need for various gene products. In other words, cells are control freaks! One could argue that control is, ultimately, more important than energy efficiency for the cell's adaptation and survival. Although this point is often lost in



**FIGURE 19-1** Seven points at which gene expression can be regulated.

discussions of gene regulation, it is central to biology. As you can see from Figure 19-1, and as we'll learn in this and the following chapters, some regulatory mechanisms are directed at mRNA or even at the protein products of mRNA translation. Why do cells "waste" their efforts in this way? Such pathways provide a means of rapidly altering the levels of active proteins in response to the cell's needs. Over the course of evolution, cells and organisms with such capability have won out over those that may have been more efficient but less able to adapt to changing conditions. Thus, gene regulation involves a fine balance between efficiency and adaptability. New and surprising regulatory mechanisms continue to be discovered, and newly discovered posttranscriptional and translational regulatory processes are proving to be highly important, especially in eukaryotes.

In Chapters 15–18 we learned about the mechanics of transcription and translation, processes critical to the flow of biological information. Now we turn our attention to how these processes are regulated by the cell to conserve resources while effectively responding to changing environmental conditions and achieving optimum cell function. This chapter presents some general principles of gene regulation that are common to the mechanisms used by both bacteria and eukaryotes. We start by examining the protein-DNA interactions that hold the key to transcriptional regulation. We then discuss principles of posttranscriptional regulation, to provide a more complete overview of the rich complexity of regulatory mechanisms. Chapter 20 gives us a more in-depth look into bacterial gene regulation, and Chapter 21 and Chapter 22 address the complex regulation of gene expression in eukaryotes.

## 19.1 Regulation of Transcription Initiation

Regulatory processes operating at the level of transcription initiation are the best documented, and probably the most common. Elaborate mechanisms have evolved to regulate the process of transcription initiation—before large amounts of cellular energy are invested in the production of mRNAs and their protein products. But however diverse, these control mechanisms are really just variations on a common theme and boil down to simple protein-protein and protein-DNA interactions. Indeed, regulation at the step of transcription initiation can be explained simply by changes in how RNA polymerase interacts with the DNA at promoter sequences.

Regulatory proteins that bind DNA can have profound effects on the affinity of RNA polymerase for a promoter. These effects can flow in either direction, either enhancing or preventing RNA polymerase function. Given both the energy expenditure of gene expression and the need for cells to be able to respond quickly to changes in their environment, one can only imagine the enormous evolutionary pressure placed on regulatory mechanisms. We provide here an overview of the transcriptional regulatory mechanisms used by cells. The most detailed information derives from studies in bacterial systems, but eukaryotic mechanisms of gene regulation, although more complex and using somewhat different strategies, can be explained by the same basic principles.

### Activators and Repressors Control RNA Polymerase Function at a Promoter

The most basic mechanism for regulation of transcription initiation is encoded in the DNA sequence of the promoter. RNA polymerase has different intrinsic affinities for promoters of different sequence. In the absence of other controls, these differences in promoter strength correlate with the efficiency with which the genes are transcribed. Genes for products that are required at all times, such as the enzymes of central metabolic pathways, are expressed at a nearly constant level. These genes are often referred to as **housekeeping genes**, and unvarying expression is called **constitutive gene expression**. Although housekeeping genes are expressed constitutively, the expression levels of different housekeeping genes vary widely. For these genes, the RNA polymerase-promoter interaction strongly influences the rate of transcription initiation; with differences in promoter sequences, the cell can synthesize the appropriate level of each housekeeping gene product.

When the level of a gene product rises and falls with a cell's changing needs, this is known as **regulated gene expression**. **Activation** is an increase in expression and **repression** is a decrease in expression of a gene in response to a change in environmental conditions. The mechanisms of gene activation and repression, in both bacteria and eukaryotes, require the assistance of **transcription factors** (also called transcription regulators), proteins that alter the affinity of the RNA polymerase for the promoter. Transcription factors that enhance gene expression are called **activators**, and those that reduce expression are called **repressors**. As we'll see later in the chapter, bacterial and eukaryotic transcription factors have many common structural and functional features. These

regulators act by binding to specific DNA sequences known as **regulatory sites**.

A gene is said to be under **positive regulation** when binding of an activator protein promotes or increases expression of that gene. Conversely, a gene is under **negative regulation** when binding of a repressor protein prevents or decreases expression. Thus, positive and negative regulation refer to the type of regulatory protein involved: the bound protein either facilitates or inhibits transcription.

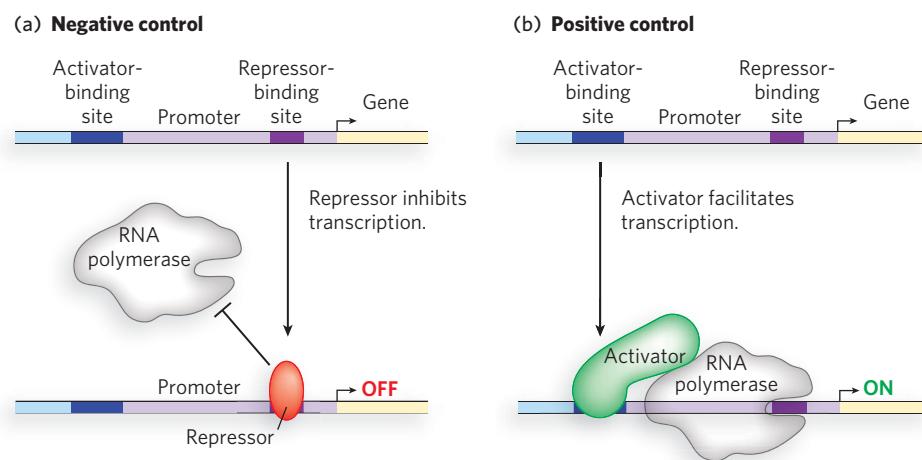
A repressor can lower the rate of gene transcription if its regulatory site overlaps the gene's promoter and repressor binding sterically occludes binding of the RNA polymerase to the promoter (Figure 19-2a). Repressors can act in other ways as well. Some prevent transcription by binding a regulatory site that is near the promoter but does not block RNA polymerase binding. Recall that the RNA polymerase-promoter complex converts from the closed complex to the open complex, in which the DNA strands are melted locally, opening up the duplex for transcription (see Figures 15-13 and 15-22). Repressors can block the closed-to-open transition, thereby preventing transcription. Instead of steric occlusion, the underlying principle in these mechanisms is allostery: a conformational change in the RNA polymerase-DNA complex prevents formation of the open complex. Repressors of this type can act on the RNA polymerase or directly on the DNA to stabilize the closed complex over the open complex. Other types of repressors act by holding the RNA

polymerase to the promoter site, preventing its escape from the promoter.

Activators provide a molecular counterpoint to repressors: they bind to DNA and enhance the activity of RNA polymerase at a promoter. For example, an activator may induce a conformational change in the polymerase that accelerates transition to the open complex. Alternatively, an activator may alter the torsion of DNA, making it more likely to unwind and form the open complex. Probably the most common way that activators function is through cooperativity (Figure 19-2b). In this case, the activator binds both the RNA polymerase and a DNA regulatory site next to the promoter, thereby increasing the affinity of the polymerase for the promoter; this activation process is referred to as recruitment.

### Transcription Factors Can Function by DNA Looping

Binding sites for activators and repressors are often found at or near the promoter, particularly in bacteria. However, regulatory sites can also be found far away from the promoter. In fact, in eukaryotes, regulatory sites are sometimes thousands of base pairs upstream or downstream from a promoter. How do transcription factors exert their effects on RNA polymerase when their binding sites are so far away from the gene's promoter? Experiments directed at understanding this “action at a distance” have



**FIGURE 19-2** Negative and positive transcriptional regulation. (a) In this example of negative regulation, a repressor-binding site overlaps the promoter. When the repressor protein binds, RNA polymerase cannot initiate transcription and no mRNA is produced. (b) In this example of positive regulation, an activator protein binds near the promoter and recruits RNA polymerase to the site. Transcription is initiated and mRNA is synthesized.

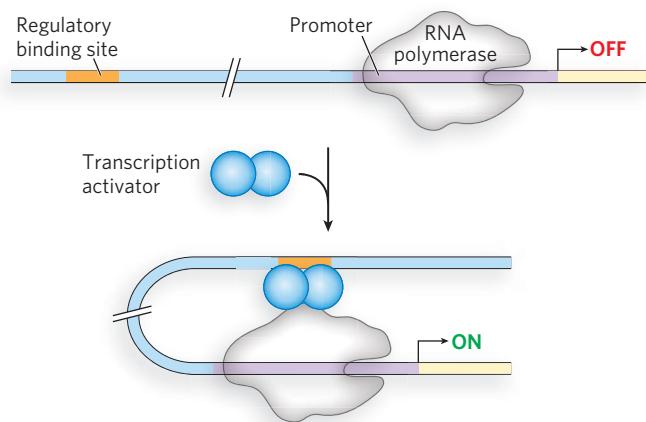
demonstrated that distant regulatory sites can often be placed closer to or farther from the promoter and still retain function. Not only is distance from the promoter of little consequence, but the regulatory sites also retain their function regardless of experimental changes in their sequence orientation relative to the promoter.

When distant regulatory sites were first discovered, scientists imagined that the regulatory proteins might bind to these sites and then slide along the DNA until they reached RNA polymerase at the promoter. However, experiments revealed that this was not the case (see How We Know). Instead, the DNA between the regulatory site and the RNA polymerase-promoter complex loops out to bring the regulatory protein and RNA polymerase together (Figure 19-3). This looping can be observed directly in the electron microscope (Figure 19-4). **DNA looping** is facilitated by proteins called architectural regulators that bind to DNA sequences between the regulatory site and the promoter, bending the DNA (Figure 19-5). The use of DNA looping is common in eukaryotic gene regulation. Some bacterial regulatory proteins also function through DNA looping, facilitated by architectural regulators. In eukaryotes, the distant regulatory sites that bind transcription factors and function over long distances from the promoter are called **enhancers**.

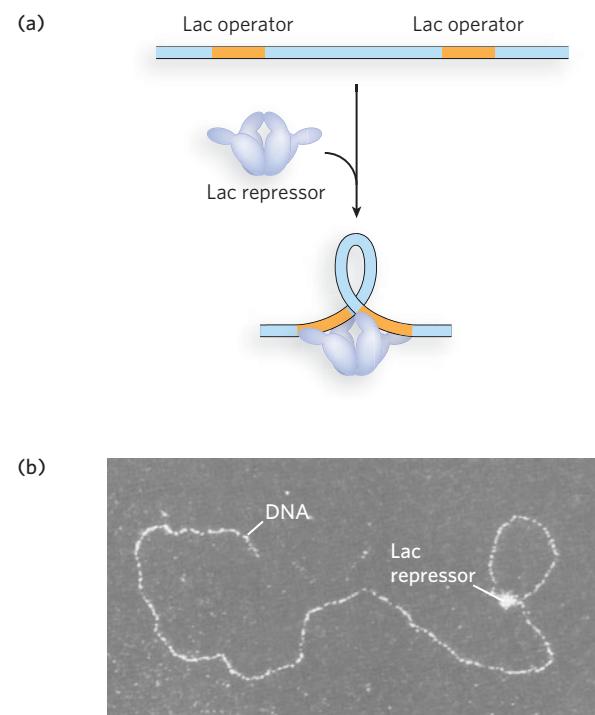
Gene regulation by DNA looping can result in either activation or repression. Activation can recruit RNA polymerase to the promoter through cooperativity, in much the same way as an activator that binds

near the promoter. Recruitment of RNA polymerase through DNA looping can also be mediated by a protein “bridge” between the activator and the polymerase (Figure 19-6a). Proteins that act by bridging activators and RNA polymerase, but do not bind DNA directly, are called **coactivators**. For example, the eukaryotic protein complex Mediator acts as a bridge between RNA polymerase II and regulatory proteins bound to distant sites and is essential for transcription activation (see Chapter 15). Repression can also occur through proteins that do not bind the DNA directly but instead bind activator proteins and prevent the recruitment of RNA polymerase (Figure 19-6b). Repressors that act through protein-protein interaction rather than by binding DNA directly are called **corepressors**.

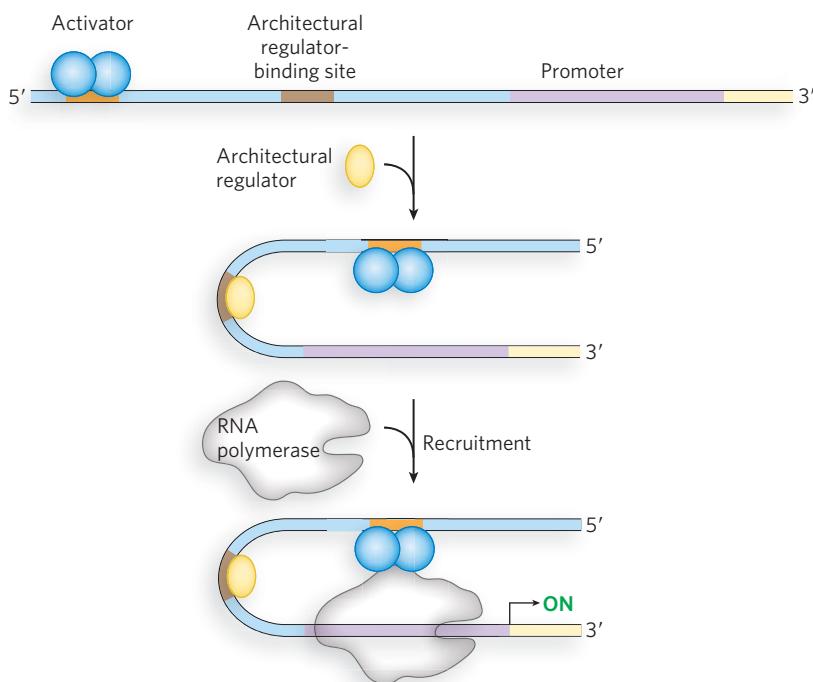
There could be an unintended consequence of gene regulation by forming DNA loops over large distances. A regulator meant to target a distant promoter could act instead on a different promoter located in the opposite direction. In eukaryotes, this problem is solved



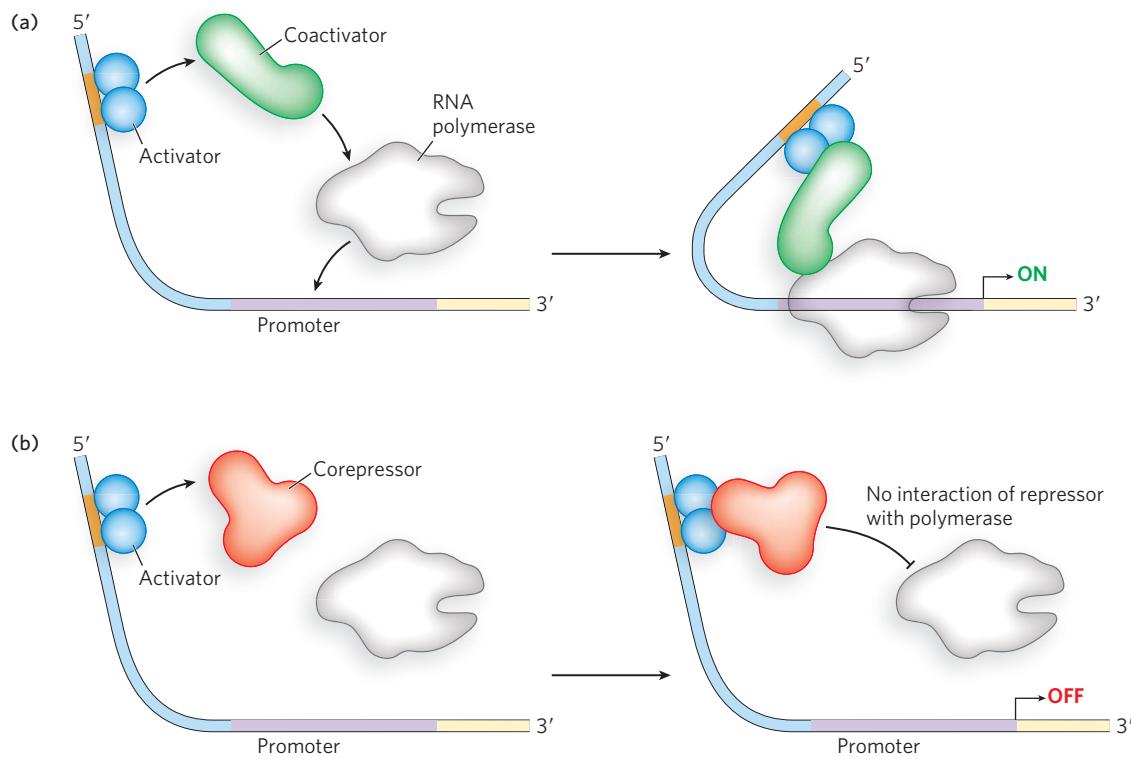
**FIGURE 19-3 Action at a distance: DNA looping.** After the binding of a transcription activator (blue) to a distant regulatory site, the activator also binds the promoter-bound RNA polymerase through protein-protein interactions, forming a DNA loop and activating the polymerase.



**FIGURE 19-4 DNA looping mediated by a single transcription factor.** (a) The bacterial Lac repressor protein, a tetramer of identical subunits, binds two distant sites on a single DNA molecule, forming a DNA loop. (b) The DNA loop is visible in this micrograph, negatively stained with uranyl acetate and imaged by dark-field electron microscopy. [Source: (b) H. Kramer et al., *EMBO J.* 6:1481–1491, 1987.]

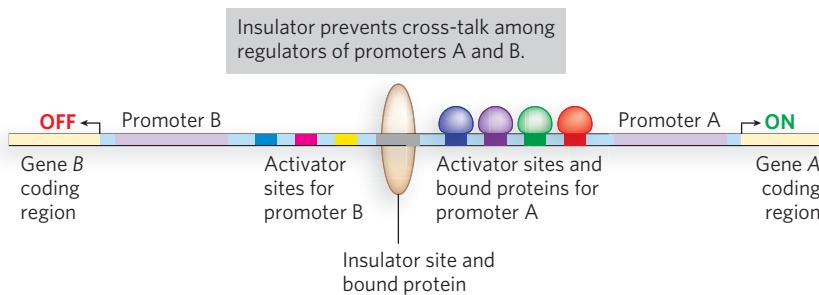


**FIGURE 19-5 Transcription factors playing an architectural role.** Some transcription factors, known as architectural regulators, bend the DNA when they bind their DNA site, thus promoting looping. Here, the regulator facilitates looping for recruitment of RNA polymerase by an upstream activator.



**FIGURE 19-6 Transcription coactivators and corepressors acting as bridges.** Coactivators and corepressors act indirectly, binding regulatory proteins without making direct contact with DNA. (a) Coactivators bind transcription activators and facilitate their function in activating RNA

polymerase. (b) Corepressors bind transcription activators and inactivate their polymerase-activating function. In these examples, the activators are bound upstream from the promoter, but activator sites can also be located downstream.



**FIGURE 19-7 Insulators.** Eukaryotic promoters have many regulatory elements that require DNA looping across long distances. Shown here are two genes, A and B, each with several activator-binding sites. When the regulatory sites for gene A are filled, the activators act on the promoter of gene A, but the insulator sequence blocks their action on the promoter of gene B. Insulators have bound proteins that enable the insulator function (see Chapter 21).

by the presence of **insulators**, short sequences of DNA that prevent inappropriate cross-signaling (Figure 19-7). (Insulators are discussed further in Chapter 21.)

### Regulators Often Work Together for Signal Integration

Activators and repressors often function at the same promoter. The use of multiple transcription factors allows the expression of a gene to be affected by more than one environmental condition. **Signal integration**, occurring in both eukaryotes and bacteria, is the control of a gene by multiple regulators in response to more than one environmental signal. A simple example in bacteria is the regulation of genes that encode products responsible for metabolizing sugar, the main energy source for bacteria (Figure 19-8).

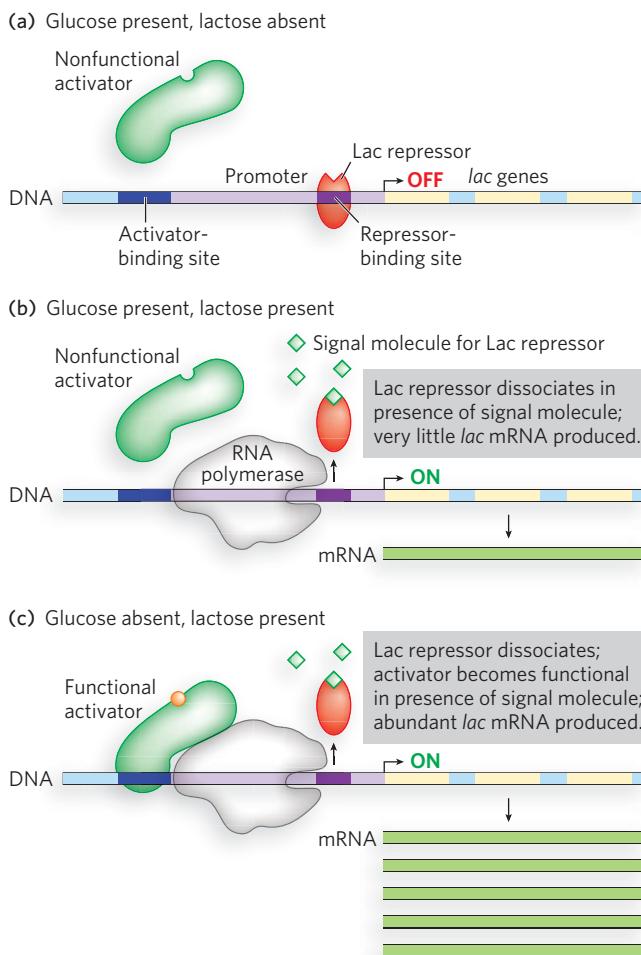
Bacteria are capable of deriving energy from many different sugars, and they have sets of genes for metabolizing each one. But it would be a waste of cellular resources to express all of these genes all the time, and systems of regulation have evolved in which the genes for metabolizing a given sugar are expressed only when that sugar is present in the environment. Take, for example, the *lac* operon, a set of genes for the metabolism of lactose (see Chapter 5 and Chapter 15; operons are more fully defined below). When lactose is not present, the Lac repressor protein is bound to the operon DNA at a sequence called the operator and ensures that the genes for lactose metabolism are not transcribed. When lactose is present, the cell sends a signal for the Lac repressor to dissociate from the operator, allowing transcription of the genes encoding lactose-metabolizing enzymes.

Even though bacteria can metabolize many different sugars, their best energy source is glucose. When both glucose and lactose are present in the environment, the cell preferentially metabolizes the glucose. It would be a waste of energy to continue producing the lactose-metabolizing enzymes, but the presence of lactose causes dissociation of the Lac repressor from the DNA. And yet, under these conditions, the genes encoding lactose-metabolizing enzymes are not highly transcribed. How does the cell do this? This is where signal integration comes into the picture. The lactose-metabolizing genes are also under the control of an activator protein needed for the efficient transcription of the *lac* operon genes, even in the absence of the Lac repressor (see Figure 19-8). When glucose is present, the activator protein is kept in a nonfunctional form. But in the absence of glucose, the activator becomes functional and, provided lactose is present (and thus Lac repressor is not bound to the operator), the genes for lactose metabolism are expressed at a high level.

This exquisite control, achieved by two different transcription factors working together, is an example of signal integration. The cell can adjust its energy resources by taking into account more than one environmental condition (the availability of glucose and of lactose).

### Gene Expression Is Regulated through Feedback Loops

The regulation of gene expression usually operates as a feedback circuit. This is easier to explain in bacteria than in eukaryotes, although similar principles apply in both. Recall that genes for the metabolism of lactose are



**FIGURE 19-8 Signal integration in gene expression.** Two transcription factors, an activator and a repressor, integrate two different environmental signals (the presence of glucose and lactose) for fine control of gene expression in the lactose-metabolic pathway of bacteria. (a) When glucose is present and lactose is absent, the Lac repressor binds the promoter and blocks RNA polymerase; there is no gene expression. (b) In the presence of both glucose and lactose, the repressor binds a small signal molecule, changes shape, and is released from the DNA. The *lac* genes are now transcribed at a low, basal level, because RNA polymerase has a low intrinsic affinity for the promoter sequence. The presence of glucose keeps the activator in a nonfunctional state. (c) In the absence of glucose and presence of lactose, the activator binds a different small signal molecule, changes conformation, binds DNA near the promoter, and recruits RNA polymerase for high levels of gene expression.

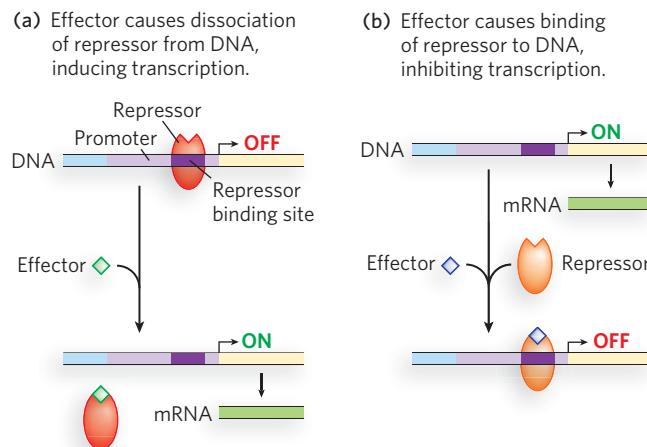
controlled by multiple transcription factors. The repressors and activators either bind DNA or not, depending on signals received from the environment. The binding of a repressor or activator to DNA is often regu-

lated by a molecular signal called an **effector**, usually a small molecule or another protein that binds the activator or repressor and causes a conformational change that results in an increase or decrease in transcription.

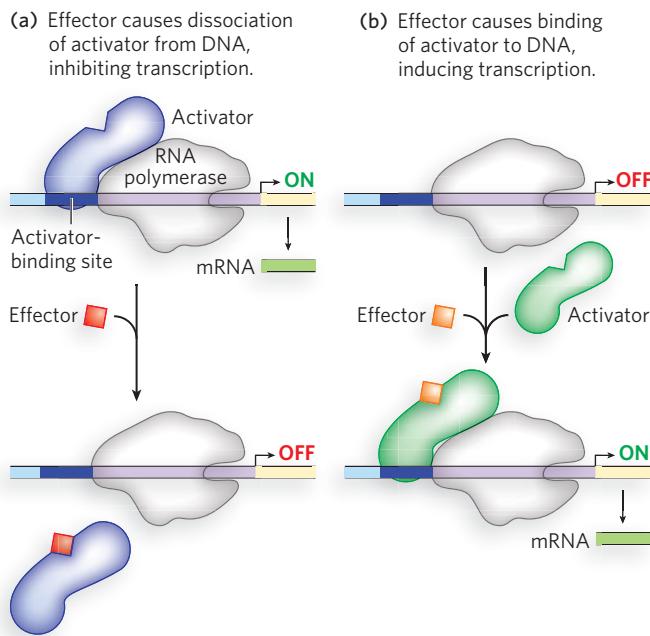
Repressors can be activated or inactivated by effectors. In one scenario, the effector binds to the repressor and induces a conformational change that results in dissociation of the repressor from its binding site on the DNA, allowing transcription to proceed (Figure 19-9a). Alternatively, the interaction of an inactive repressor and a signal molecule could cause the repressor to bind to DNA, shutting down transcription (Figure 19-9b).

The same considerations apply to activators. Some activators bind DNA and enhance transcription until dissociation of the activator is triggered by the binding of a signal molecule (Figure 19-10a). In other cases, the activator binds to DNA only after interaction with a signal molecule (Figure 19-10b). Signal molecules that bind activators can therefore increase or decrease transcription, depending on how they affect the activator.

Given the allosteric control of activators and repressors, we can understand how a regulatory feedback loop functions. In the bacterial *lac* operon, Lac repressor binds DNA in the absence of an effector, preventing the expression of genes required for the metabolism of lactose. The effector for the Lac repressor



**FIGURE 19-9 The role of effectors in negative regulation.** The binding of signal molecules (known as effectors) to repressors can (a) relieve or (b) enhance repression. In (a), the repressor (red) binds DNA in the absence of the effector; the external signal causes dissociation of the repressor to permit transcription. In (b), the repressor (orange) binds DNA in the presence of the signal, shutting down transcription. The repressor dissociates and transcription ensues only when the signal is removed (not shown).



**FIGURE 19-10** The role of effectors in positive regulation. The binding of effectors to activators can (a) inhibit or (b) enhance activation. In (a), the activator (blue) binds in the absence of the effector and transcription proceeds; when the signal is present, the activator dissociates and transcription is inhibited. In (b), the activator (green) binds in the presence of the signal to stimulate transcription. The activator dissociates and transcription ceases only when the signal is removed (not shown).

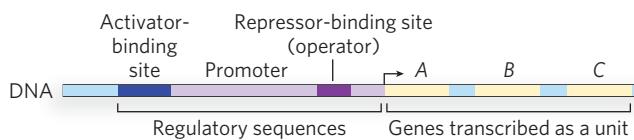
is allolactose, a minor by-product of lactose metabolism. Therefore, when lactose is present in the environment, the signal molecule is formed and binds the Lac repressor, causing it to dissociate from the DNA. This gives RNA polymerase access to the promoter of the *lac* operon for a low, basal level of transcription. Transcription of the operon is greatly enhanced by binding of the activator cAMP receptor protein (CRP). CRP does not bind its regulatory site when glucose is available. In the absence of glucose, however, cells produce cAMP (cyclic AMP), which is an allosteric effector of CRP, producing a conformational change that enables CRP to bind its regulatory site. The bound activator then recruits RNA polymerase and boosts gene expression from the *lac* operon.

When lactose is depleted, allolactose is also depleted, and in the absence of this effector the Lac repressor again binds the operator site, preventing RNA polymerase from transcribing the *lac* operon. Likewise, when glucose becomes available, cAMP levels diminish and CRP no longer binds DNA. Regulatory feedback loops like these are common in all cells.

## Related Sets of Genes Are Often Regulated Together

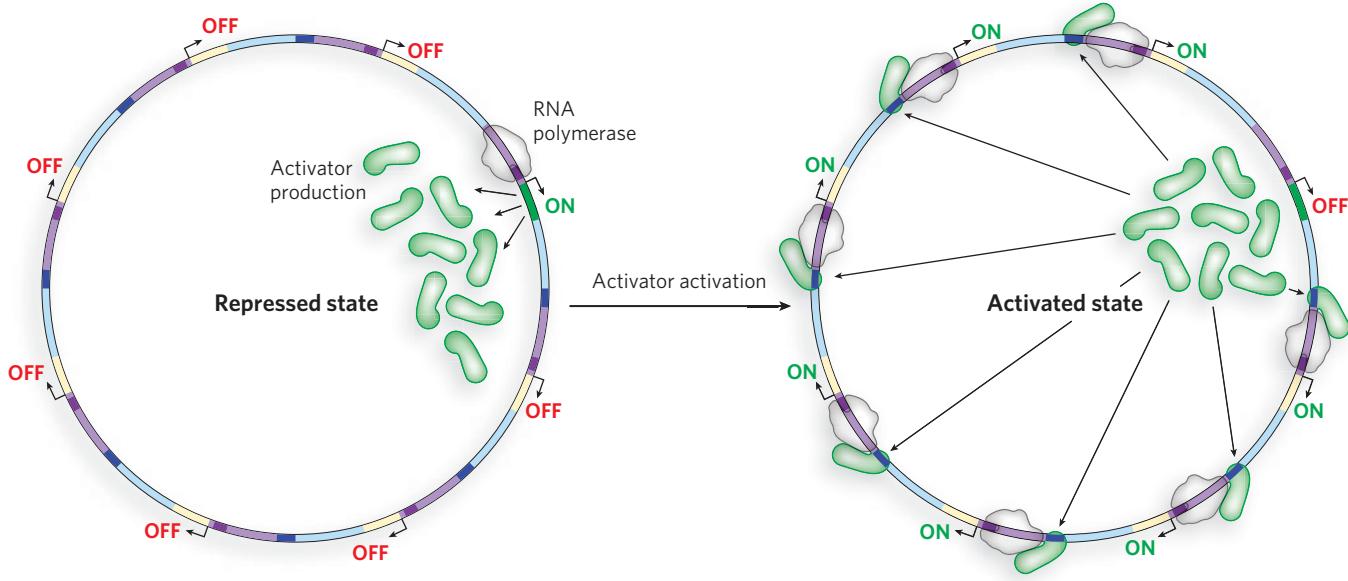
Bacterial promoters are often positioned upstream from several genes that operate in a common metabolic pathway. Transcription produces a long **polycistronic mRNA** that contains multiple genes in one transcript. The single promoter that initiates transcription of the cluster is the site of regulation for all the genes in the polycistronic message. The polycistronic DNA, its promoter, and all the additional sequences that function together in regulating its transcription are called an **operon** (Figure 19-11). Most operons contain 2 to 6 genes, but some have more than 20 genes. Often, the different genes of the polycistronic mRNA are translated separately by different ribosomes that assemble at internal Shine-Dalgarno sequences, just upstream from each gene, that determine individual levels of translation efficiency.

The organization of bacterial genes into operons allows small sets of genes that function together to be regulated together. But there are also instances in which multiple operons are controlled in a coordinated fashion. A group of operons with a common regulator is called a **regulon**. This arrangement allows for shifts in cellular functions that can require the activation of hundreds of genes—a major theme in the regulated expression of dispersed networks of genes in bacteria. Eukaryotes also exhibit global regulation of genes; genes that function together are dispersed over different chromosomes, yet are typically controlled in a coordinated way through common control elements and transcription factors. Figure 19-12 shows a generalized view of global regulation, in which multiple genes may be turned on by the presence of the same activator or by the removal of a common repressor. (Mechanisms of global transcriptional gene regulation in bacteria and eukaryotes are described in detail in Chapter 20 and Chapter 21.)

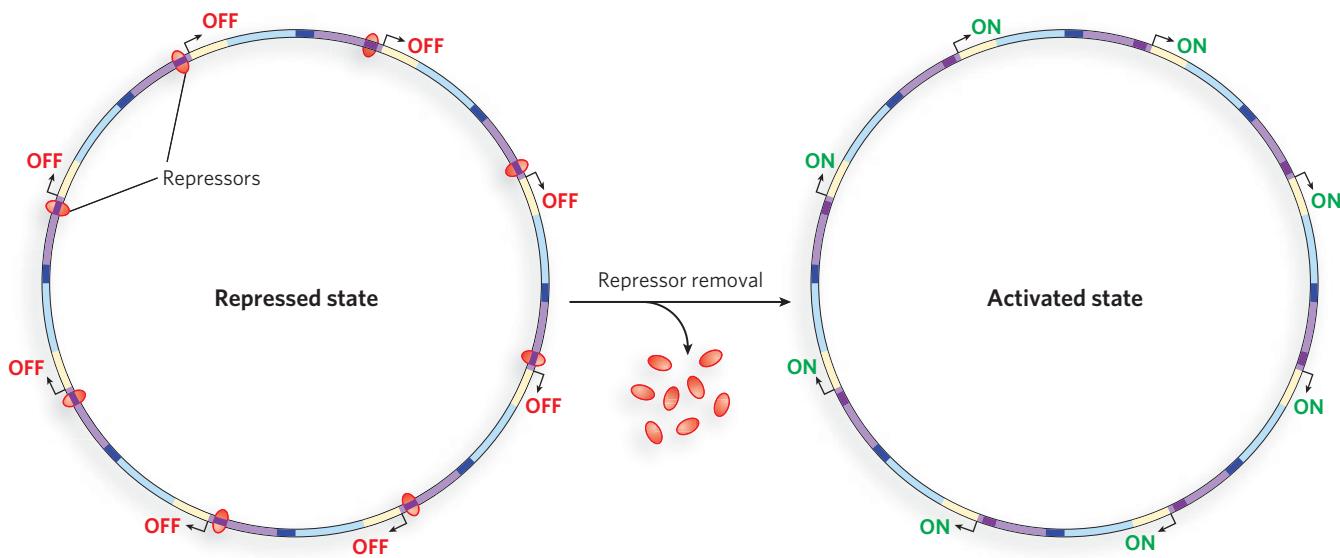


**FIGURE 19-11** A bacterial operon. In this hypothetical operon, genes A, B, and C are transcribed as a single unit: a polycistronic mRNA. Typical regulatory sequences in the operon include binding sites for proteins that either activate or repress transcription from the promoter.

## (a) Positive regulation by activators



## (b) Regulation by destruction of repressors

**FIGURE 19-12** Global regulation of groups of genes.

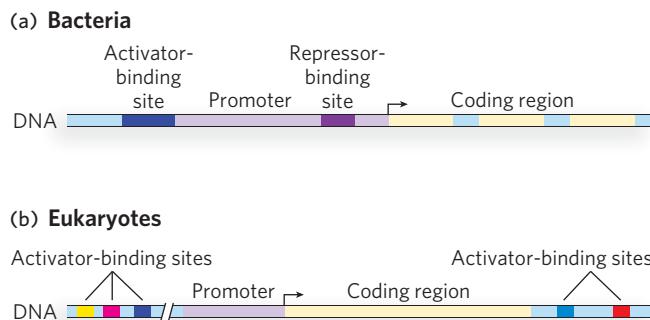
(a) Global regulation can occur through the binding of a common transcription activator (green). When needed, the activator may be produced de novo by expression of its gene (green), as shown, or an existing activator protein may become active for DNA binding through

interaction with another protein or a small effector molecule. (b) Alternatively, global regulation can result from the removal of a common repressor (red) bound to DNA sites, either by an allosteric change induced by binding of a small effector molecule or by proteolytic digestion of the repressor.

## Eukaryotic Promoters Use More Regulators Than Bacterial Promoters

Signal integration is important to gene regulation in both bacteria and eukaryotes. However, eukaryotic promoters for Pol II, the RNA polymerase that transcribes protein-coding genes, typically contain more

regulatory-binding sites than do bacterial promoters (Figure 19-13). The use of more transcription factors in eukaryotic gene control reflects the greater need for gene regulation in a more complex organism with a larger genome. For example, nonspecific DNA binding of regulatory proteins could become a problem in the much larger genomes of higher eukaryotes, because the



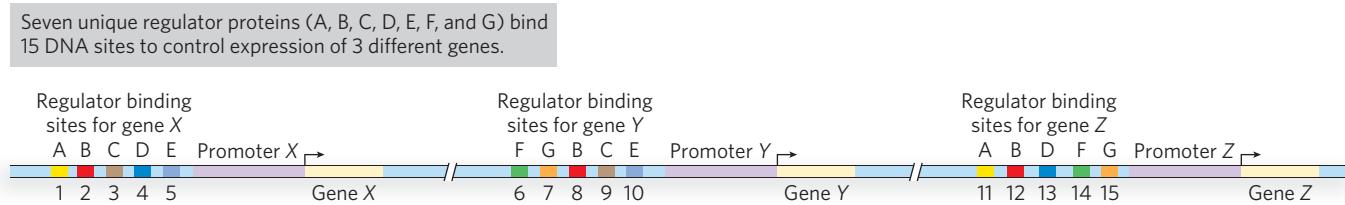
**FIGURE 19-13 Bacterial and eukaryotic regulatory regions compared.** (a) Bacterial promoters are usually regulated by only one or two transcription factors, and their binding sites are typically near, or overlap, the promoter. (b) Eukaryotic genes, especially those of multicellular organisms, usually have numerous regulator-binding sites spanning a large region (sometimes more than 50 kbp) located upstream and/or downstream from the promoter, or even within the coding sequence of the gene itself (not shown).

chance that a specific binding sequence will occur randomly at an inappropriate site increases with genome size. Indeed, the number of transcription factor-binding sites in eukaryotic promoter regions varies with the complexity of the organism. Genes in single-celled yeasts have only a few regulator sites and are not much more complicated than bacterial genes, whereas the promoters in multicellular organisms can have 10 or more regulator-binding sites spaced over long distances, 50 kbp or more away from the transcription start site. Specificity for transcriptional regulation is improved by multiple regulatory proteins that must bind DNA and form a multiprotein complex to become active. This multiprotein requirement vastly reduces the probability of random gene activation or repression.

## Multiple Regulators Provide Combinatorial Control

Using multiple transcription factors for every gene in a genome would be energetically costly if every gene required unique regulators, but using different combinations of a limited set of transcription factors to differentially regulate many genes provides an opportunity for efficiency. This is made possible by **combinatorial control**—the need for specific combinations of factors to unlock each particular gene (Figure 19-14). Consider the hypothetical genes *A*, *B*, and *C*, each of which requires five transcription factors. If each factor were distinct, the cell would require 15 transcription factors to control expression of these three genes. But if genes *A* and *B* used three of the same factors, and a combination of the factors for genes *A* and *B* is used to regulate gene *C*, then differential regulation of these three genes would require only 7 different proteins instead of 15.

Combinatorial control occurs in bacteria as well as in eukaryotes, and we've already seen an example in bacteria in the case of the two regulatory elements of the genes involved in lactose metabolism. Recall that these genes are controlled by a repressor that senses the presence of lactose and by an activator that senses the presence of glucose. The genes encoding proteins for metabolism of other sugars have their own repressors, but use the same activator. For instance, the digestion of galactose requires removal of the galactose repressor from the DNA, and this occurs only when galactose is present. However, as with the lactose genes, high expression of the galactose genes is achieved only when glucose is absent from the environment. The same protein activator used at the *lac* operon also regulates the galactose genes: CRP, which becomes functional by binding its effector molecule cAMP when glucose is not present. The regulation of the genes for different sugar-metabolizing pathways by a common activator is an example of combinatorial control.



**FIGURE 19-14 Combinatorial control in gene regulation.** Each of these three hypothetical eukaryotic promoters requires five different regulatory proteins, to bind a total of 15 regulatory sites. Each color represents a particular transcription factor and its regulatory binding site. Each gene uses different combinations of transcription factors, and some factors are used for more than one gene. In total, there are seven unique regulatory sequences, and thus seven unique transcription factors, controlling expression of all three genes.

## Regulation by Nucleosomes Is Specific to Eukaryotes

In eukaryotes, transcription initiation almost always depends on the action of activator proteins. One important reason for the apparent predominance of positive regulation seems obvious: packaging of DNA into chromatin renders most promoters inaccessible, and thus their associated genes are silent. Chromatin structure affects access to some promoters more than others, but generally, repressors that prevent the access of RNA polymerase to DNA would be redundant. Therefore, eukaryotic genes are constitutively repressed and require activation in order to be transcribed. Recall that transcription is regulated by different types of change in chromatin structure (see Chapter 10). The chromatin state can be either open or closed. Open chromatin is often (but not always) associated with acetylation of nucleosomes, whereas closed chromatin is associated with methylation of nucleosomes. Thus eukaryotic activators and repressors can act through modification of nucleosomes that alter chromatin structure, rather than by recruiting RNA polymerase or preventing polymerase binding to DNA.

Bacterial RNA polymerase generally has access to every promoter, and most bacterial genes are controlled by specific repressors. In eukaryotes, however, general repression by nucleosomes, combined with the use of activators to regulate transcription, is more efficient than the use of specific repressors. If the 20,000 to 25,000 genes in the human genome were negatively regulated, each cell would have to constantly synthesize specific repressors to prevent the transcription of a great many genes. Instead, the nucleosomes that function to condense DNA also repress most genes, and the cell only has to synthesize the activators needed to promote transcription of the subset of genes required at a particular time. These arguments notwithstanding, there are examples of negative regulation in eukaryotes, from yeast to humans.

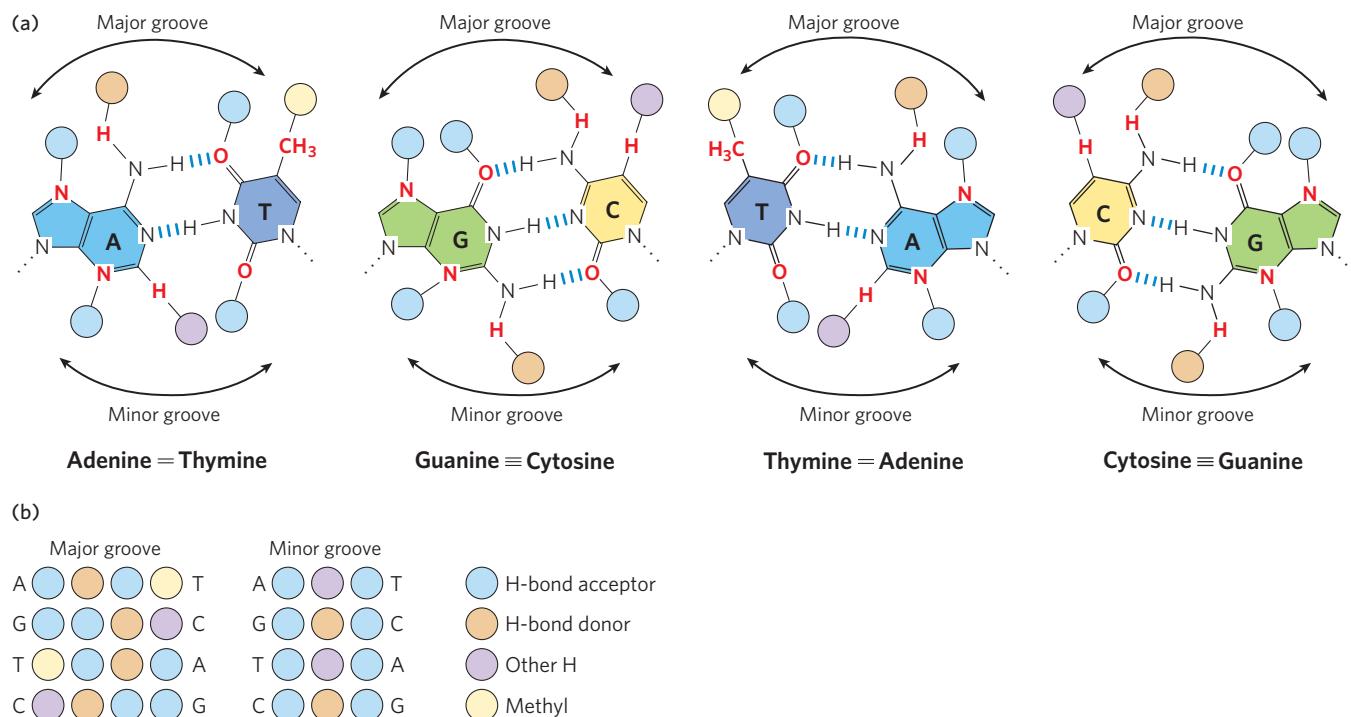
- Repressors can hinder transcription by binding DNA at a site that prevents RNA polymerase binding or by preventing the closed-to-open complex transition of the polymerase-promoter complex (negative regulation).
- Activators promote RNA polymerase binding through cooperativity or promote formation of the open complex by causing a conformational change in the promoter or the polymerase (positive regulation).
- Binding sites for transcription factors need not be close to the transcription start site, and in eukaryotes they are often located thousands of base pairs from the promoter. Regulatory proteins that bind sites distant from the promoter exert their effects through DNA looping.
- Promoters may be controlled by two or more transcription factors, allowing integration of signals from more than one environmental variable.
- Small signal molecules (effectors) allosterically regulate the function of activators and repressors.
- Sets of genes that function in one pathway are often controlled together, at the same time.
- Eukaryotes generally have more transcription factors than bacteria, reflecting the greater need for regulated gene expression in a complex multicellular organism. Specificity of gene expression is enhanced by the use of multiple regulators.
- In combinatorial control, the same regulatory protein is used to control different genes in combination with other regulators, forming a multiprotein regulator that is specific for individual genes.
- Chromatin structure renders most eukaryotic promoters inaccessible to RNA polymerase and plays an important role in gene expression. Gene expression typically requires proteins that modify nucleosomes and open up the chromatin structure to transcriptional activation.

### SECTION 19.1 SUMMARY

- The various mechanisms of transcription initiation are among the most well-documented regulated processes in gene expression. Transcription initiation is the step most often regulated, and regulation at this point is the most energy efficient because it occurs before the investment of energy in mRNA and protein synthesis.
- Transcription initiation is mediated by intrinsic promoter affinity for RNA polymerase or by repressor and activator proteins that modulate promoter affinity for the polymerase.

## 19.2 The Structural Basis of Transcriptional Regulation

As we saw in Chapter 4, protein structures come in an amazingly wide assortment of shapes and sizes, but these structures can often be broken down into discrete functional domains formed by common structural motifs. Indeed, most transcription factors use a surprisingly small subset of structural motifs to interact with regulatory sites on DNA, with RNA polymerase, or with other regulatory proteins. In fact, some of the common motifs in the architecture of transcription factors can



**FIGURE 19-15 Hydrogen-bond donor and acceptor atoms in the major and minor grooves of DNA.** Shown here are the functional groups (in red) of all four base pairs as displayed in the major and minor grooves. Hydrogen-bond donor and acceptor atoms that can be used for base-pair recognition

be determined from analysis of the protein's primary sequence. We explore here the most common transcription factor domain structures that are involved in DNA binding and interaction with other proteins.

### Transcription Factors Interact with DNA and Proteins through Structural Motifs

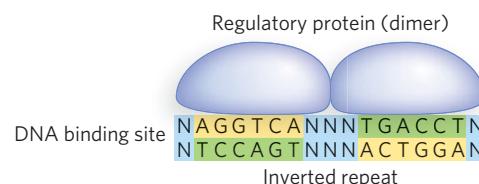
The recognition of DNA by a regulatory protein is almost always through certain amino acid side chains of an  $\alpha$  helix referred to as the **recognition helix**. A limited set of structural motifs function to present the recognition helix to the DNA. Amino acid side chains in this helix usually “read” the DNA sequence along the major groove, because (as you’ll recall from Chapter 5) more hydrogen-bond donor and acceptor atoms of nucleotide bases are found in the major groove than in the minor groove (Figure 19-15).

The DNA binding sites for regulatory proteins are often short inverted nucleotide repeats at which multiple (usually two) subunits of the regulatory protein bind cooperatively (Figure 19-16). Accordingly, many bacterial and eukaryotic activators and repressors are dimers. Crystal structures of activators and repressors bound to DNA show that each monomer of a homodimer

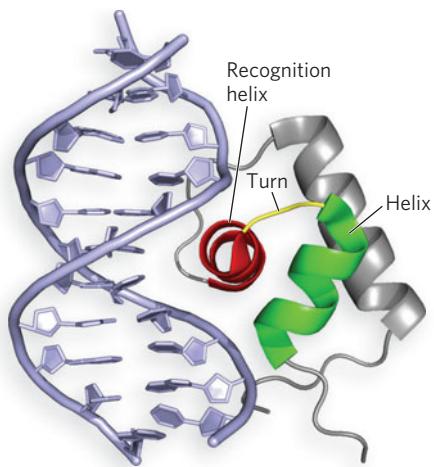
by proteins are marked by red and blue circles, respectively. Other hydrogen atoms are marked with purple circles, and methyl groups are marked with yellow circles. Notice how the four possible base pairs are chemically distinct in the major groove compared with the minor groove.

binds the same nucleotide sequence within the inverted repeat. Regulatory proteins use several structural motifs to promote dimerization.

Examples of DNA-binding and protein-dimerization motifs are described in Chapter 4. Here we focus on those that play prominent roles in the function of regulatory proteins.



**FIGURE 19-16 An inverted repeat at the site of transcription factor binding.** A nucleotide sequence followed by the reverse, complementary sequence is known as an inverted repeat. It can have a variable number of base pairs that are not part of the repeat between the two repeated sequences. A palindrome is an inverted repeat with no base pairs between the two repeat sequences. Proteins that bind to inverted repeats are dimeric, and each subunit binds to one half of the repeat. Illustrated here is a homodimer binding an inverted repeat (N = any nucleotide).



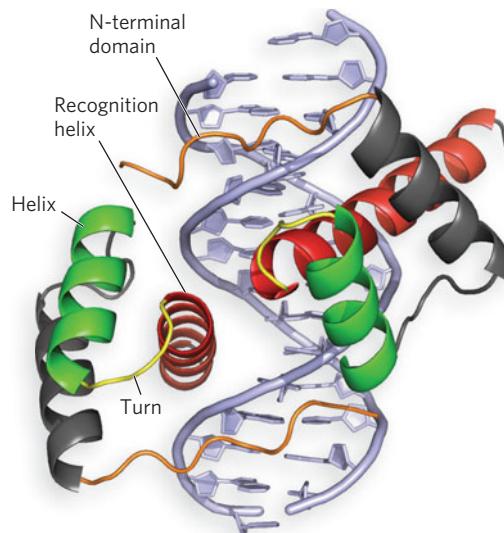
**FIGURE 19-17** The helix-turn-helix DNA-binding motif.

This molecular structure shows the DNA-binding domain of the bacterial Lac repressor (as a ribbon structure) interacting with the major groove of DNA. The helix-turn-helix motif is red and green; the DNA recognition helix is red. [Source: PDB ID 1LCC.]

**Helix-Turn-Helix Motif** Bacterial regulators most commonly use the **helix-turn-helix motif** to present the recognition helix to DNA, and several eukaryotic regulatory proteins also interact with DNA through this motif. The helix-turn-helix motif consists of about 20 amino acid residues that form two short  $\alpha$  helices connected by a  $\beta$  turn (Figure 19-17). This motif lacks intrinsic stability and is generally part of a somewhat larger DNA-binding domain. Only one of the two  $\alpha$ -helical segments serves as the recognition helix; it packs against other regions of the protein and protrudes from the protein surface for insertion into the major groove.

**Homeodomain Motif** Researchers first identified the **homeodomain motif** as a conserved 60 amino acid sequence in transcription activators encoded by genes that regulate body pattern development in fruit flies. We now know that the homeodomain is found in proteins from a wide variety of multicellular organisms, including humans. When the structure of the homeodomain was determined, it was found to contain a helix-turn-helix motif—but with some important differences. First, the homeodomain is composed of three  $\alpha$  helices, only two of which (helices 2 and 3) correspond to the helix-turn-helix motif. Second, the N-terminal residues of the homeodomain reach around the DNA and interact with the minor groove (Figure 19-18).

**Basic Leucine Zipper and Basic Helix-Loop-Helix Motifs** The **leucine zipper motif** is an amphipathic

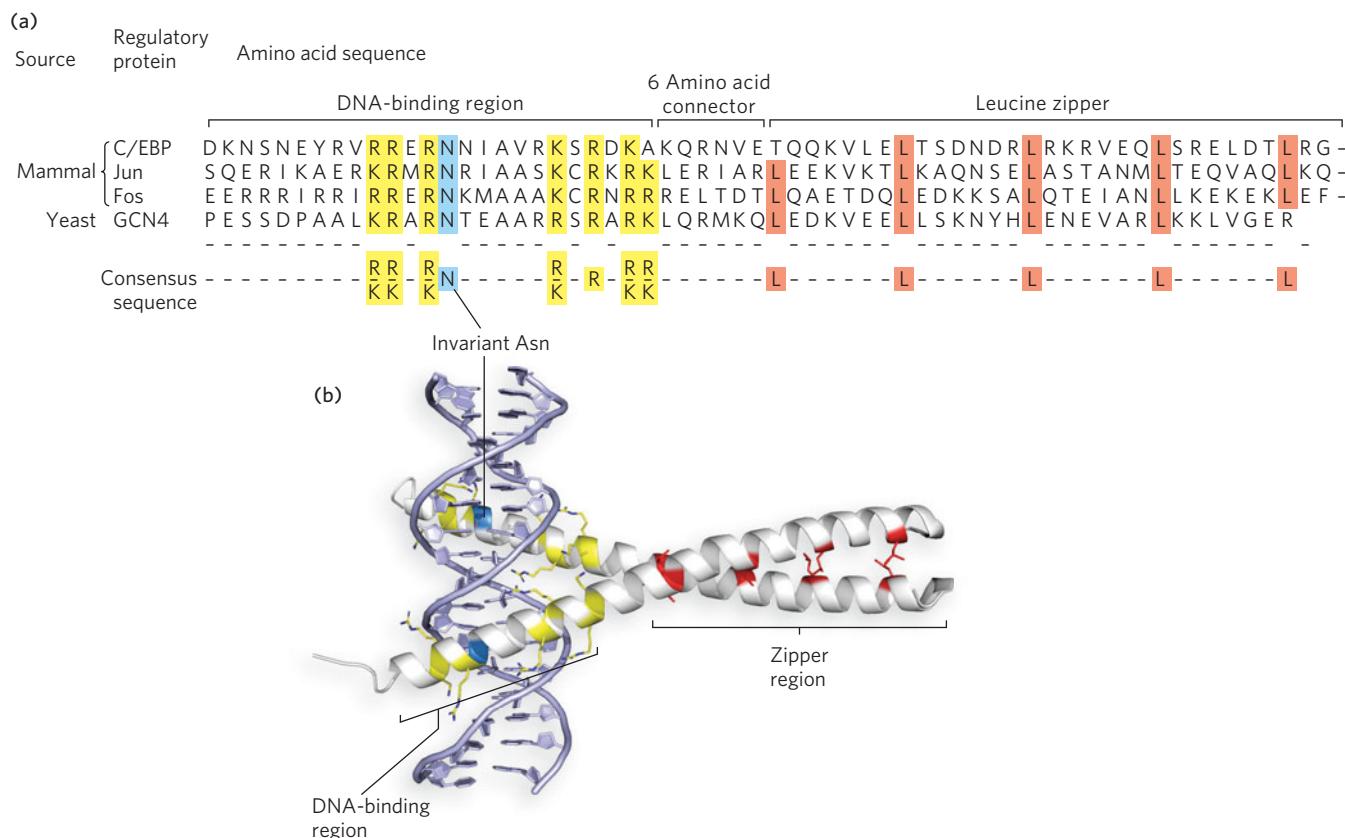


**FIGURE 19-18** The homeodomain DNA-binding motif.

This molecular structure shows the homeodomain motifs of the *Drosophila* transcription factor known as Paired, a dimeric protein (only a small part of the much larger Paired dimer is shown). The recognition helix in each subunit is stacked on two other  $\alpha$  helices and can be seen protruding into the major groove. The N-terminal sequence inserts into the minor groove. [Source: PDB ID 1FJL.]

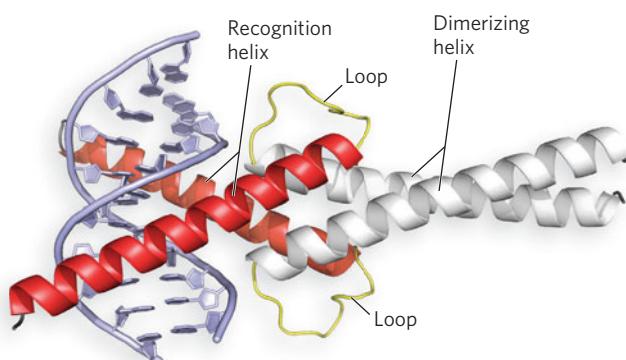
$\alpha$  helix, with a series of hydrophobic amino acid residues concentrated on one side of the helix. A striking feature of this  $\alpha$  helix is the occurrence of Leu residues at every seventh position, forming a hydrophobic surface along one side of the helix, the site where two identical subunits dimerize. Researchers initially thought that dimerization was caused by interdigitating Leu residues (hence the name “zipper”). We now know that the two subunits dimerize through packing of the residues along the inner surface of the interface where they form a coiled-coil structure. Certain transcription factors use leucine zippers in combination with basic residues at one end of the  $\alpha$  helices that make up the recognition helices, forming **basic leucine zipper motifs** (Figure 19-19). These basic leucine zipper regulators are sometimes referred to as bZIP proteins. The crystal structure of bZIPs shows that they grip and bind DNA like a set of tongs (Figure 19-19b). However, there are many regulatory proteins that use leucine zipper helices only for dimerization and contain a separate motif for DNA binding. Leucine zippers are found in many eukaryotic transcription activators and in a few bacterial regulators.

A somewhat similar structural motif in some eukaryotic transcription factors is the **basic helix-loop-helix motif** (Figure 19-20). These proteins share a conserved region of about 50 residues that are



**FIGURE 19-19 The basic leucine zipper motif.** This motif is often used to mediate protein-protein interactions in eukaryotic transcription factors. (a) The amino acid sequences of several basic leucine zipper (bZIP) proteins. Notice the Leu (L) residues at every seventh position in the zipper region and the number of basic residues (Lys (K) and Arg (R)), and one invariant Asn (N) in the DNA-binding region. A consensus sequence is

shown at the bottom. (b) A basic leucine zipper from the yeast activator protein GCN4. Only the two “zippered”  $\alpha$  helices, each from a different subunit of the dimeric protein, are shown. The helices wrap around each other in a coiled-coil. The interacting Leu residues are shown in red. The basic residues in the DNA-binding region are shown in yellow. The invariant Asn residue is shown in blue. [Source: (b) PDB ID 1YSA.]



**FIGURE 19-20 The basic helix-loop-helix motif.** This ribbon model shows the human homodimeric transcription factor Max bound to its DNA target site. The two amphipathic  $\alpha$  helices of each Max subunit are shown in red and white; the loop is yellow in both. The two subunits form a four-helix bundle through association of their dimerizing  $\alpha$  helices (white), while the DNA-binding  $\alpha$  helices (red) extend from the bundle. [Source: PDB ID 1HLO.]

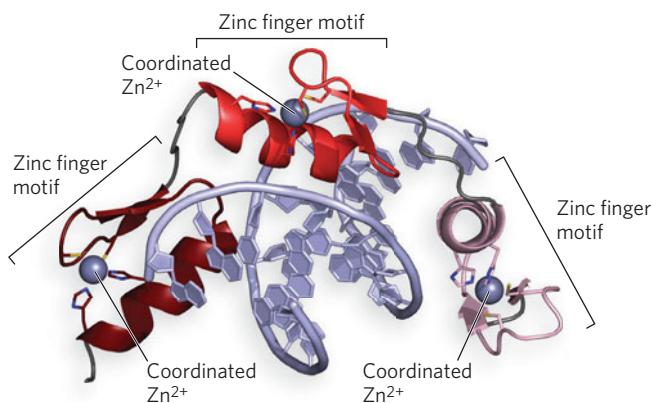
important for both DNA binding and protein dimerization. The basic helix-loop-helix region contains two amphipathic  $\alpha$  helices—one of which contains basic residues—linked by a loop of variable length. Dimer formation is mediated by one set of amphipathic  $\alpha$  helices, and DNA binding is mediated by the amphipathic  $\alpha$  helices that contain basic residues. The recognition helices grip the binding sequence in DNA in much the same way as the basic leucine zipper.

**Zinc Finger Motif** The **zinc finger motif** comes in a few varieties, two of which are discussed here. A zinc finger domain consists of about 30 residues that form an elongated loop held together at the base by a single  $Zn^{2+}$  ion. The  $Zn^{2+}$  ion is coordinated to four amino acid side chains, usually four Cys residues or two Cys and two His residues. The zinc functions to stabilize the motif, which presents a recognition helix to DNA; the

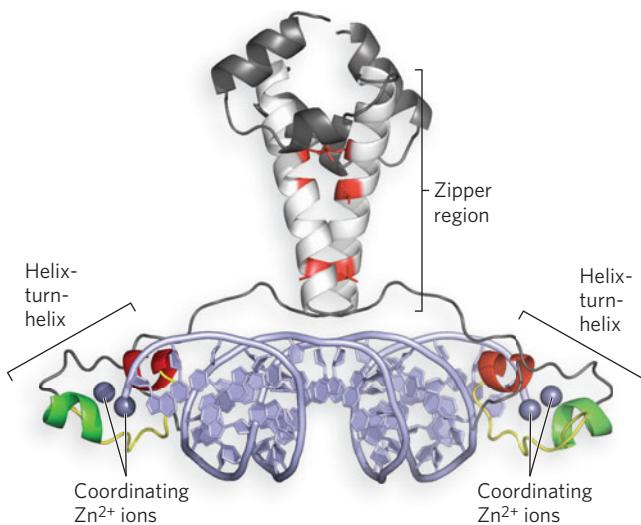
$Zn^{2+}$  does not interact with the DNA directly. The interaction of a single zinc finger with DNA is typically weak, and this particular binding motif has the unique feature that the protein can have multiple copies of the motif that act together as a chain. Multiple zinc fingers are found in many DNA-binding proteins, and they substantially enhance binding affinity by interacting simultaneously with the DNA. In fact, one DNA-binding protein of the frog *Xenopus laevis* has 37 zinc fingers! In the mouse regulatory protein Zif268, the zinc finger domains present the recognition helix so that it winds around the DNA following the major groove (Figure 19-21). Zinc finger proteins in this class are among the few regulatory proteins that function as monomers. They do not require the DNA binding site to be an inverted repeat and can recognize a long sequence of DNA that contains no internal repeat sequences, because each recognition helix is unique.

Another type of regulatory zinc finger protein combines the  $Zn^{2+}$ -binding motif with the helix-turn-helix motif (Figure 19-22). This type of zinc finger protein uses two  $Zn^{2+}$  ions to stabilize the DNA-binding domain, which has a helix-turn-helix motif. These proteins bind to the DNA as dimers; the example shown in Figure 19-22 uses a leucine zipper to mediate the dimer contacts. Zinc finger motifs are common in eukaryotes, and there are a few examples among bacterial regulators.

**Transcription-Activation Motifs** In addition to structural domains devoted to DNA binding and protein dimerization, transcription activators contain regions



**FIGURE 19-21 The zinc finger motif.** This ribbon structure of a fragment from the mouse regulatory protein Zif268 shows three zinc fingers (colored differently) arranged one after the other in the protein. Each  $Zn^{2+}$  ion is shown as a small sphere. The Zif268 recognition helices enter the major groove of the DNA, and the three fingers wind around the DNA helix. [Source: PDB ID 1ZAA.]



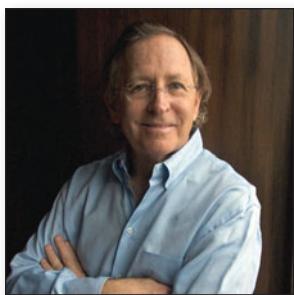
**FIGURE 19-22 Zinc finger, helix-turn-helix, and leucine zipper motifs in the same regulatory protein.** The Gal4 protein (Gal4p), a yeast transcription activator, is a dimer, held together by a leucine zipper. Each of the two DNA-binding domains contains two  $Zn^{2+}$  ions, which help hold the recognition helix (red) of the helix-turn-helix motif in the proper geometry for DNA recognition. [Source: PDB ID 3COQ.]

used for recruiting RNA polymerase or other protein factors. These recruitment regions are thought to be relatively unstructured. The first one was noted in the Gal4 protein (Gal4p), a yeast transcription activator. Researchers observed that several acidic residues were associated with the activating function, and the region could be altered by mutation without much effect on function; the region was referred to as an “acid blob.”

Other activation domains of transcription factors have also been shown to contain unstructured regions characterized (like Gal4p) by acidic residues or by other types of amino acids. For example, certain activation domains are glutamine-rich or proline-rich, and also appear to be unstructured. The current understanding is that these activation regions function as many short sections of amino acid residues that act together like a patch of Velcro: the more there are, the greater their effect. This somewhat unstructured approach to achieving activation might enable transcription factors to extend their range of protein-protein interactions for combinatorial control (see Section 19.1).

## Transcription Activators Have Separate DNA-Binding and Regulatory Domains

Transcription activators typically contain a regulatory domain that is separate from and functionally independent of the DNA-binding domain. An early and now classic experiment demonstrating this

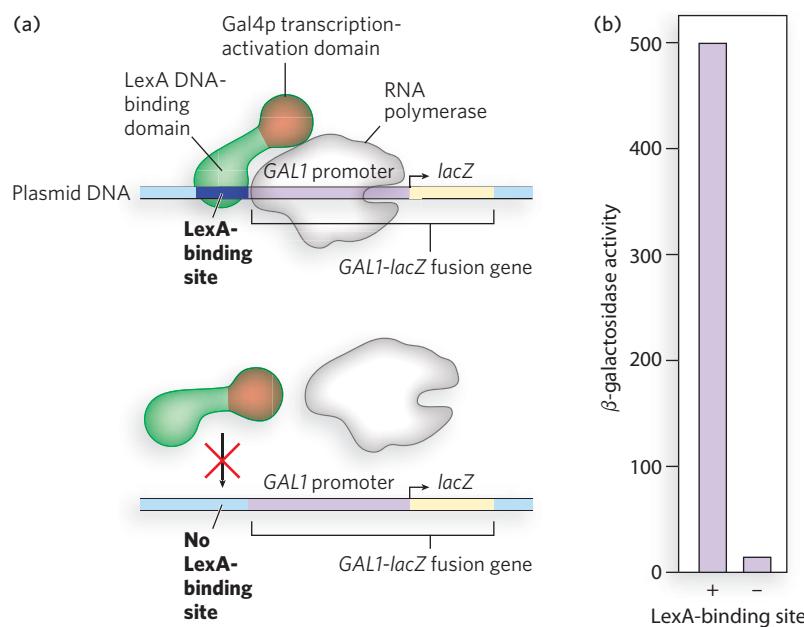


**Mark Ptashne** [Source: Courtesy of Mark Ptashne.]

property of transcription activators was performed by Mark Ptashne and his colleagues (**Figure 19-23**). They proposed that the regulatory and DNA-binding functions of a transcription activator are independent and separable. To test this idea, they spliced together DNA encoding the transcription-activation domain of a eukaryotic activator and DNA encoding the DNA-binding domain of a bacterial repressor. The prediction was that the fusion protein would activate transcription of a gene under the control of the eukaryotic activator, provided that the gene contained an upstream binding sequence recognized by the bacterial repressor. The researchers chose for their study the yeast transcription activator Gal4p, which drives the expression of genes for galactose metabolism, including the gene *GAL1*. The

C-terminal activation sequences of Gal4p were fused to the N-terminal DNA-binding region of an *E. coli* repressor, LexA. The new gene encoding the LexA-Gal4 fusion protein was inserted into a plasmid and transferred to yeast, along with a second plasmid that contained the *GAL1* gene promoter fused to the bacterial  $\beta$ -galactosidase gene, *lacZ*. As predicted, in the transformed yeast, the new LexA-Gal4 fusion protein activated the *GAL1-lacZ* gene construct containing an upstream LexA-binding sequence, but did not activate the *GAL1-lacZ* gene on the plasmid that lacked the LexA-binding site.

As shown in Figure 19-23, the LexA-Gal4 fusion protein activated the production of  $\beta$ -galactosidase more than 500-fold when the LexA-binding site was present (compared with when absent) in the *GAL1-lacZ* gene construct. Production of  $\beta$ -galactosidase is easily measured by using a synthetic substrate that yields a blue product when cleaved by the enzyme. The LexA-Gal4 fusion protein did not activate transcription from the wild-type *GAL1* promoter, because the protein no longer recognized the Gal4p-binding site upstream from the *GAL1* gene.



**FIGURE 19-23** Experiment demonstrating the separate DNA-binding and regulatory domains of transcription activators. (a) A *GAL1-lacZ* fusion gene is inserted in a plasmid downstream from a LexA-binding site (top), or downstream from a DNA segment lacking the LexA-binding site (bottom). In both cases, cells are transformed with a plasmid encoding a fusion protein containing the bacterial LexA DNA-binding

domain (green) fused to the transcription-activation region of yeast Gal4p (orange). (b) Expression of the *GAL1-lacZ* fusion gene is measured by the blue color released on hydrolysis of the synthetic substrate X-gal by  $\beta$ -galactosidase, the protein product of *lacZ*. Expression of  $\beta$ -galactosidase is induced in cells containing plasmids with the LexA-binding site. [Source: Adapted from R. Brent and M. Ptashne, *Cell* 43:729–736, 1985.]

This experiment not only demonstrated the modular nature of regulatory proteins, but also gave molecular biologists yet another tool for studying the inner workings of the cell. The widely used yeast two-hybrid assay makes use of the fact that the DNA-binding and transcription-activation regions of the Gal4p regulatory protein are stable, separate domains (see Chapter 7).

## SECTION 19.2 SUMMARY

- The DNA-binding domains of transcription factors are usually constructed from a limited set of structural motifs, including the helix-turn-helix, homeodomain, leucine zipper, helix-loop-helix, and zinc finger motifs.
- Many transcription factors form dimers and bind inverted repeat sequences, thereby increasing their affinity for DNA.
- The basic leucine zipper and basic helix-loop-helix motifs facilitate both DNA binding and protein dimerization.
- Many transcription-activation domains are composed of acidic, proline-rich, or glutamine-rich regions.
- Transcription activators typically contain separable and functionally independent DNA-binding and regulatory domains.

## 19.3 Posttranscriptional Regulation of Gene Expression

Thus far we have discussed the regulatory mechanisms involved in initiation of transcription, but as Figure 19-1 shows, regulatory mechanisms operate at many steps following transcription. RNA processing and translation of mRNA into protein are regulated at several points. The regulation of protein synthesis at the initiation stage is quite prevalent because, much like the strategy of regulating transcription at the initiation step, it saves the cell the huge energy investment of synthesizing the protein product. Nevertheless, there are some posttranslational regulatory mechanisms.

The regulation of gene expression *after* production of a functional protein does have several advantages. For example, it takes substantial time to produce mRNA and translate it into protein, so one benefit of regulating a pathway by acting on a fully formed protein is the speed with which changes in the amount or activity of the protein can be implemented. Covalent modification can turn a protein on or off very rapidly. Some types of gene regulation can be inherited through generations

of cell division through imprinting (see Chapter 21) and epigenetics (see Chapter 10).

We explore here some of the main mechanisms of posttranscriptional regulation, through mRNA processing and mRNA stability (RNA interference), translation initiation, covalent modification, cellular localization, and protein degradation.

### Some Regulatory Mechanisms Act on the Nascent RNA Transcript

After transcription initiation, there are several ways in which a gene can be regulated before the mature mRNA transcript is produced. As an overview, we briefly describe three steps at which regulation can occur: transcript elongation, mRNA splicing, and modification of mRNA termini. These, and other examples, are discussed in more detail in Chapter 20 (for bacteria) and Chapter 22 (for eukaryotes).

**Transcript Elongation** One bacterial example of regulation affecting transcript elongation is a process known as attenuation. Attenuation prevents movement of the transcribing RNA polymerase into the first gene of an operon unless the proper conditions have been met. Controls to stop attenuation and thus proceed with transcription involve a delicate balance of metabolites, proteins, and mRNA structure. Attenuation is relatively common for the operons of amino acid biosynthesis and is particularly well documented for the *trp* operon of *E. coli* (see Chapter 20). In eukaryotes, many factors affect transcript elongation, and these elongation factors can be targets of control.

**mRNA Splicing** Many eukaryotic RNA transcripts contain introns that are spliced out in forming the mature mRNA (see Chapter 16). The splicing process is performed by a multiprotein spliceosome in the nucleus. Sometimes an mRNA has alternative splice junctions to choose from, which result in different products. The choice of splice site is regulated by repressors, activators, and enhancers in ways that seem to be mechanistically similar to the regulation of transcription initiation. Alternative splicing choice is thus another point at which gene expression can be regulated.

**Modification of mRNA Termini** Both the 5' and 3' ends of eukaryotic mRNAs are highly modified in multistep reactions (see Chapter 16). The 5' terminus is modified by the addition of nucleotides connected by unusual phosphodiester bonds, referred to as the 5' cap. The 3' terminus is cleaved at a particular site prior to transcription termination, then multiple AMP residues are

added to form a poly(A) tail. Specific proteins recognize and bind to these modifications, which are important in mRNA transport from the nucleus, mRNA stability in the cell, and efficient association of the mRNA with ribosomes and its use in translation. Exciting new discoveries are being made about control mechanisms at the level of these mRNA modification and transport steps.

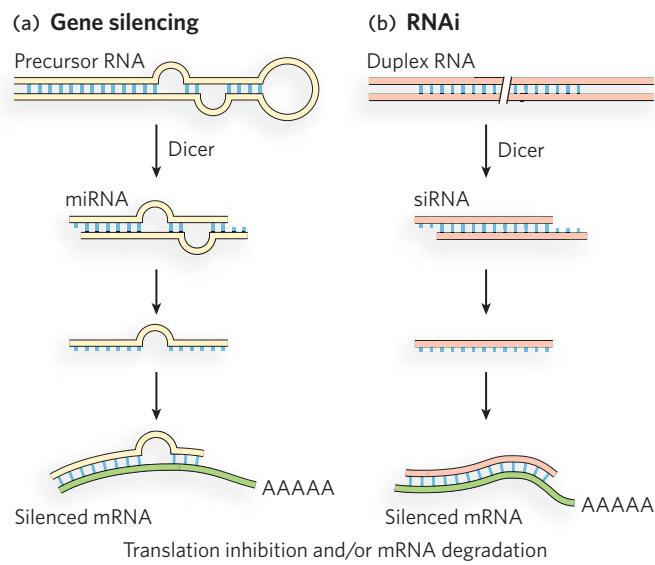
### Small RNAs Sometimes Affect mRNA Stability

The amount of protein generated from a gene is dependent on the stability of the RNA message, and regulatory mechanisms have evolved to control mRNA stability. In higher eukaryotes, certain genes are “silenced” by a class of RNAs that interact with mRNAs, resulting in degradation of the mRNA or inhibition of translation. This form of gene regulation uses small RNAs. A variety of small RNAs can control developmental timing, repress the activity of transposons, or destroy invading RNA viruses, especially in plants, which lack an immune system. Small RNAs may also play a role in heterochromatin formation, which silences all the genes contained in the heterochromatin. The role of these small mRNAs in eukaryotic gene regulation is explored further in Chapter 22. Bacteria also contain a variety of small RNAs that act at several levels to regulate gene expression.

Small RNAs are sometimes called microRNAs (miRNAs). When present only temporarily, such as during development, transient small RNAs are referred to as small temporal RNAs (stRNAs). Hundreds of different miRNAs have been identified in the more complex eukaryotes. They are transcribed as precursor RNAs, about 70 nucleotides long, that form hairpinlike structures (Figure 19-24a). An endonuclease trims precursor RNAs to form short duplexes of 20 to 25 nucleotides, one strand of which anneals to the target mRNA. The best characterized of these endonucleases is Dicer; endonucleases in the Dicer family are widely distributed in eukaryotes.

When the expression of genes producing miRNAs goes awry, tumors can result. As described in Moment of Discovery, overexpression of the miR17-92 cluster of miRNAs results in tumor formation in mice. This result was one of the first to associate functional RNAs with tumorigenesis, implicating them in the development of cancer.

Gene regulation mechanisms involving Dicer, besides their important physiological role, also have a very useful practical application. If an investigator introduces into an organism a duplex RNA corresponding to a target mRNA, Dicer cleaves the duplex into short



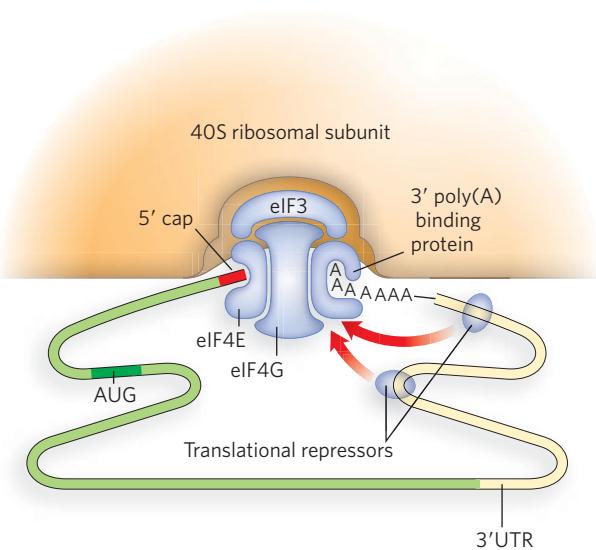
**FIGURE 19-24** Gene silencing and RNA interference.

(a) Dicer cleaves hairpin-shaped precursor RNAs into microRNAs (miRNAs), which bind to and silence mRNA by inhibition of translation. (b) Synthetic double-stranded RNA can also result in RNA interference (RNAi). When the double-stranded RNA is injected into a cell, Dicer cleaves it into small interfering RNAs (siRNAs), which interact with the target mRNA; the mRNA is degraded or its translation is inhibited.

segments called small interfering RNAs (siRNAs), which bind to and silence the mRNA (Figure 19-24b). This laboratory technique is called **RNA interference (RNAi)**. In plants, almost any gene can be shut down in this way. In nematodes, simply feeding the duplex RNA to the worm silences the target gene. The technique is a very important tool in studies of gene function, because any gene can be silenced without constructing a mutant organism. Study of functional RNAs such as miRNAs is an exciting and relatively new area of molecular biology—a field to watch for future medical advances.

### Some Genes Are Regulated at the Level of Translation

Some translational regulation does occur in bacteria, but it is a much more common occurrence in eukaryotes because of the long half-lives of many eukaryotic mRNAs. Most translational regulation occurs at the initiation step, for efficiency and energy conservation (Figure 19-25). For instance, translation of a eukaryotic mRNA requires several different initiation factors that assemble at the 5' region of the mRNA on the 40S ribosomal subunit (see Chapter 18). Many initiation factors are subject to a variety of regulatory mechanisms, which in turn modulate translation at the initiation step.



**FIGURE 19–25 Regulation of translation initiation in eukaryotes.** One of the most important mechanisms for translational regulation in eukaryotes involves the binding of repressors to specific sites in the 3' untranslated region (3'UTR) of the mRNA. The repressors interact with initiation factors eIF3, eIF4E, and eIF4G or with the ribosome to prevent or slow translation. The factor eIF4G mediates an interaction between eIF4E at the 5' cap and poly(A) binding protein at the 3' poly(A) site (see Chapter 18). This interaction is needed for efficient translation, and factors that disrupt it repress translation.

Global translational control also exists in bacteria and eukaryotes. For example, the ribosomal apparatus represents a large energy investment for the cell, and synthesis of the many components of the ribosome is regulated in processes linked to the cellular demand for proteins. Translational control of dozens of genes encoding ribosomal components are regulated by protein binding to the translation start sites in the mRNAs. Furthermore, if rRNAs are not present in sufficient amounts to match ribosomal protein subunits, the excess unassembled ribosomal proteins bind and inhibit translation of their respective mRNAs, forming a feedback translational control circuit.

### Some Covalent Modifications Regulate Protein Function

Protein function can be dramatically altered by many types of covalent modification. Protein modifications include phosphorylation, acetylation, methylation, glycosylation, ubiquitination, and sumoylation (further discussed below). The modifications can have various effects: they may render the protein active or inactive; result in a change in oligomeric state, with

functional consequences; alter the protein's affinity for DNA or for another protein; or affect the protein's stability in the cell.

An example of proteins that are highly regulated by covalent modification is the subunits of nucleosomes. Recall from Chapter 10 that nucleosomal proteins have long N-terminal tails that are often covalently modified by phosphorylation, methylation, and acetylation. These modifications regulate transcription by changing the level of chromatin compaction, thus controlling the access of RNA polymerase and other proteins to the DNA. About 10% of the chromatin in a typical eukaryotic cell is in a more condensed form (heterochromatin) than the rest of the chromatin, and genes in these regions are strongly repressed. Most of the remaining, less-condensed chromatin (euchromatin) is transcriptionally active. Histones found in condensed and less-condensed chromatin differ in their patterns of covalent modification. These modification patterns are probably recognized by enzymes that alter the structure of chromatin (see Chapter 10). The effects of nucleosome modification on chromosome structure, and therefore on gene regulation, have no clear parallel in bacteria because bacterial chromosomes are not packaged in this way.

Modifications associated with the activation of transcription are recognized by enzymes that make the chromatin more accessible to the transcriptional machinery. When transcription of a gene is no longer required, certain modifications are enzymatically removed and others are added, marking the chromatin as transcriptionally inactive. The effect of histone modification on gene expression is discussed further in Chapter 21. Other examples of covalent modifications that direct gene expression are briefly described below and are expanded upon in Chapters 20–22.

### Gene Expression Can Be Regulated by Intracellular Localization

In bacteria, transcription repressors and activators can undergo an allosteric change on binding a small effector molecule (such as allolactose or cAMP) that acts as a signal of environmental conditions, and in this way gene expression is repressed or activated in response to the signal (see Section 19.1). The compartmentation of eukaryotic cells affects the way in which gene expression can respond to environmental signals and provides opportunities for regulation at the level of transfer of proteins between intracellular locations. A prevalent pathway for communication with the extracellular environment in eukaryotes is through cell surface receptors. The receptors bind a signal molecule and relay its

message through the plasma membrane by complex signal transduction pathways, eventually resulting in transcriptional regulation in the nucleus.

A relatively simple example of a signal transduction pathway is the JAK-STAT pathway (Figure 19-26). This consists of a transmembrane receptor, a protein kinase called JAK (Janus kinase), and a transcription factor called STAT (signal transducer and activator of transcription). The transmembrane receptor binds cytokines (e.g., interferon and interleukin), small molecules that signal cells to grow or differentiate. On cytokine binding, the receptor activates JAK, which phosphorylates the receptor. This, in turn, promotes binding and phosphorylation of STAT by the phosphorylated receptor. Once phosphorylated, STAT dimerizes and enters the nucleus, where it binds DNA and activates the expression of genes involved in cell growth and differentiation. There are at least seven different STATs in mammals, each binding a different DNA sequence. The

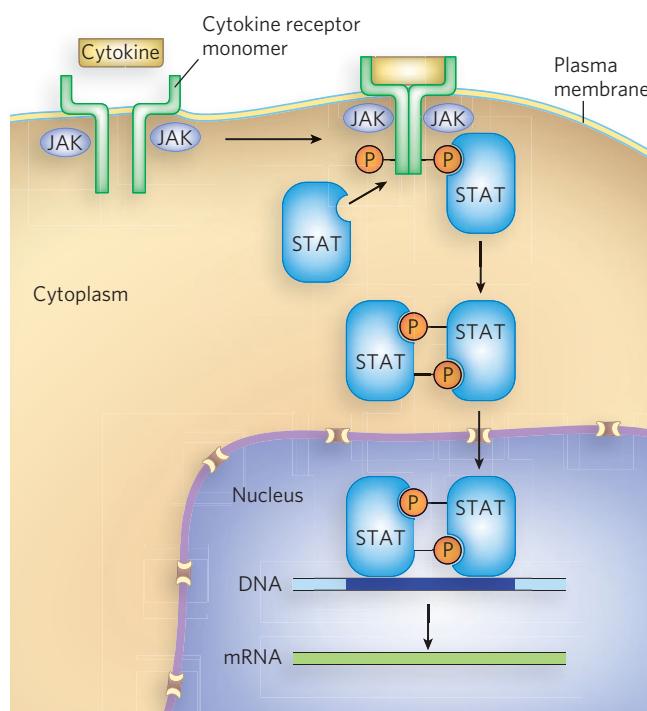
JAK-STAT pathway is conserved in organisms ranging from worms to mammals, indicating its importance to cellular function. Genetic defects in this pathway are associated with immune diseases and cancer.

Dephosphorylation by specific protein phosphatases is also important in the regulation of transcription activator or repressor activity, and this sometimes involves cellular localization as a means of regulating gene expression—that is, access of the transcription factors to the nucleus can be regulated by their state of phosphorylation.

Phosphorylation-dephosphorylation often causes conformational changes that alter the activity of a regulatory protein other than an activator or repressor. For example, the target could be a protein that binds an activator and masks its function. When the masking protein is phosphorylated, it dissociates from the activator and gene expression is thus enhanced. The hormone insulin regulates gene expression in this way, by phosphorylation-dephosphorylation of proteins involved in glucose metabolism. These mechanisms short-circuit the need for changes in mRNA or protein synthesis (Highlight 19-1). Insulin also regulates gene expression through a protein kinase signaling mechanism that ultimately activates the transcription of numerous genes involved in cell metabolism.

Steroid hormone receptors are another example of regulation by intracellular localization. These receptors are transcription activators, held in the cytoplasm by association with a heat shock protein, Hsp70. Steroid hormones are soluble in lipids and can pass through the plasma membrane without a specific transporter. On entry of the steroid hormone into a cell that expresses the particular steroid-binding receptor, and binding to the receptor, Hsp70 dissociates and the receptor-hormone complex dimerizes and enters the nucleus (Figure 19-27).

The cell can also control the intracellular localization of a regulatory protein in the absence of signal transduction. Nuclear proteins, newly synthesized in the cytoplasm, contain a localization sequence that targets them to the nucleus. Cellular localization of a regulatory protein can thus be achieved by masking or unmasking the nuclear localization sequence, controlling access of the protein to the nucleus. Cells also regulate the localization of some proteins through covalent modification by the 101-residue polypeptide known as SUMO (small ubiquitinlike modifier). When the SUMO polypeptide is attached to Lys residues of a protein, the sumoylated protein is transported to a subcompartment of the nucleus, where it is sequestered and unable to function until the SUMO polypeptide is removed.



**FIGURE 19-26 Signal transduction by the JAK-STAT pathway.** Cytokines signal a cell to increase transcription, and they act through a membrane-bound receptor protein. Cytokine binding to the receptor causes two receptor molecules to form a dimer, resulting in a conformational change that enables the JAK kinase to phosphorylate the receptor. This attracts the STAT protein, which in turn becomes phosphorylated and dimerizes, whereupon it enters the nucleus and activates the transcription of specific STAT-regulated genes.

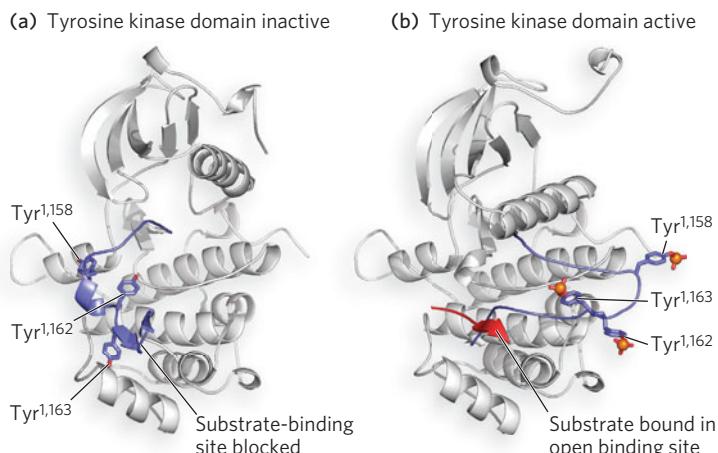
## HIGHLIGHT 19-1 MEDICINE

### Insulin Regulation: Control by Phosphorylation

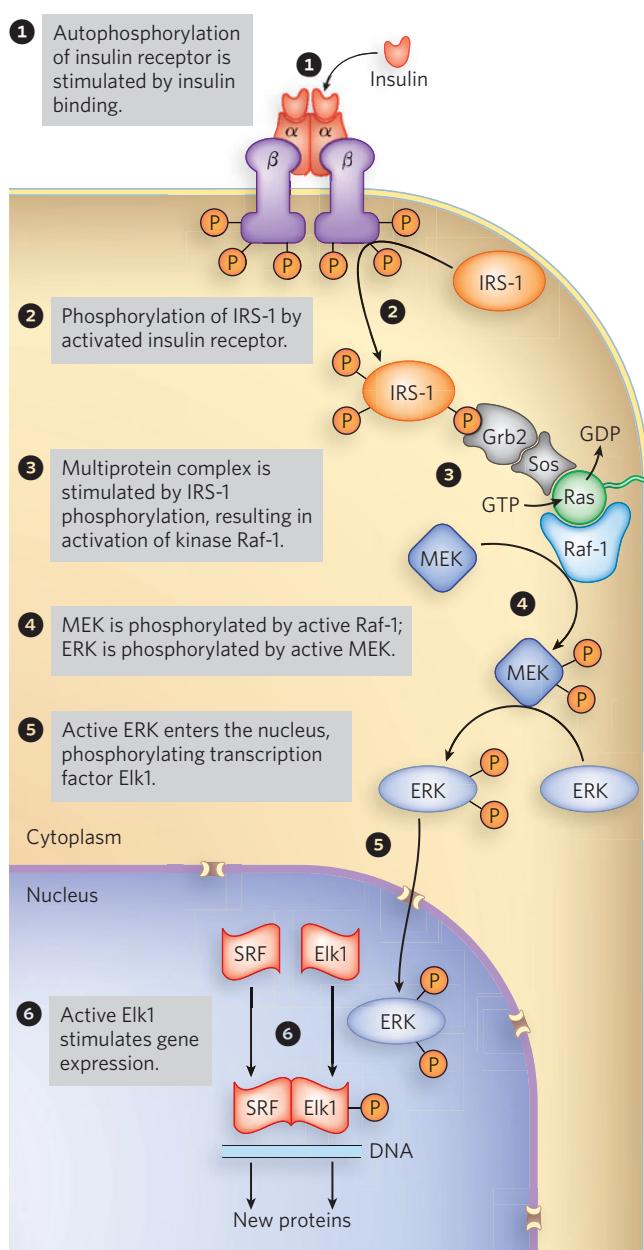
Insulin is a small (51-residue) peptide hormone, produced in the pancreas, that is central to the control of energy and glucose metabolism. For example, insulin stimulates glucose uptake from the blood by muscle, fat, and liver cells, and the use of glucose, in preference to fat, as an energy source. In these duties, insulin acts as a regulator of gene transcription.

The insulin signaling pathway involves an extensive protein kinase cascade, resulting in activation of more than 100 genes. The signaling cascade is initiated when insulin binds to the membrane-bound insulin receptor, which induces the receptor to autophosphorylate specific Tyr residues within a dimer of the receptor. This, in turn, activates the receptor to phosphorylate other proteins. Structural analysis of the tyrosine kinase domain of the insulin receptor reveals the basis for the regulation of activity by autophosphorylation (Figure 1).

One of the target proteins in the insulin protein kinase cascade is insulin receptor substrate-1 (IRS-1).



**FIGURE 1** The tyrosine kinase domain of the insulin receptor is activated through autophosphorylation. (a) When the tyrosine kinase domain is inactive, the activation loop (blue) sits in the active site and none of the critical Tyr residues are phosphorylated. (b) When insulin binds the receptor, the tyrosine kinase activity phosphorylates Tyr<sup>1,158</sup>, Tyr<sup>1,162</sup>, and Tyr<sup>1,163</sup> (phosphate groups are orange). Introducing these three phosphate groups results in a 30 Å movement in the activation loop, shifting it out of the substrate-binding site, which becomes available to phosphorylate target proteins (red). [Sources: (a) PDB ID 1IRK. (b) PDB ID 1IR3.]

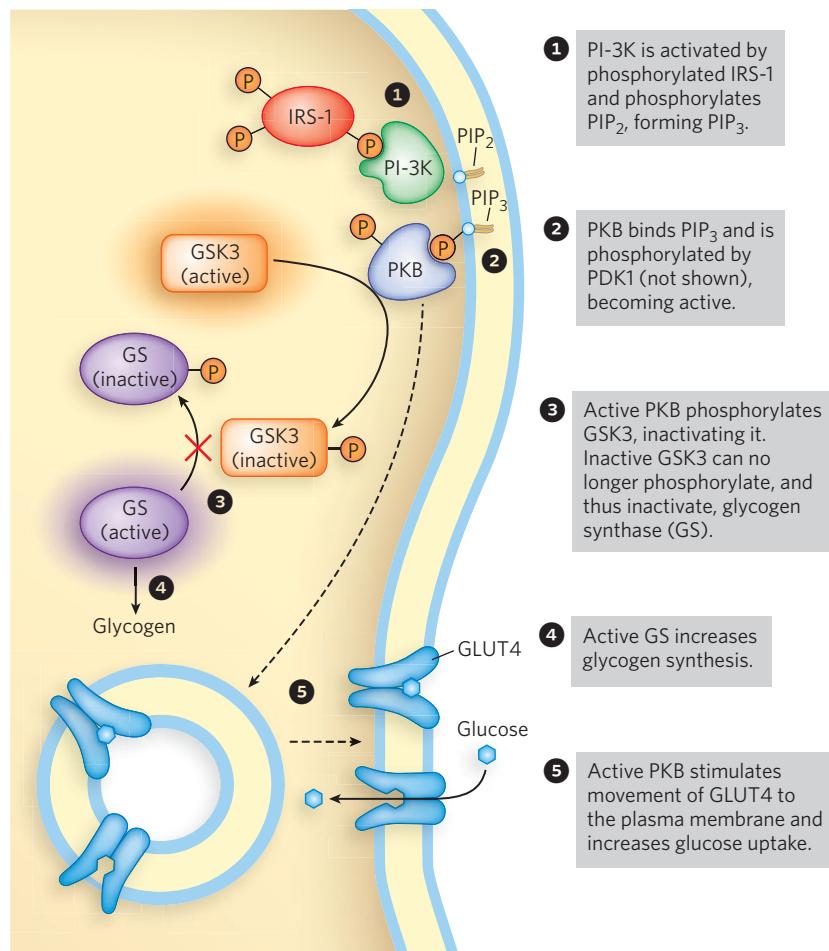


**FIGURE 2** The MAPK cascade initiated by insulin regulates gene expression. Binding of insulin to the insulin receptor triggers autophosphorylation, activating the protein kinase domain in the cytoplasmic region of the receptor. The tyrosine kinase phosphorylates IRS-1, activating it to bind other proteins that then phosphorylate yet other proteins, creating a cascade of protein phosphorylation events that amplifies the original signal. Numerous “middle” factors (e.g., Grb2, Sos, Ras, Raf-1, MEK, ERK, SRF), several of which are kinases, are required to transduce and amplify the original signal. The end result is phosphorylation and activation of Elk1, a transcription factor that stimulates gene expression.

On phosphorylation by the activated insulin receptor, IRS-1 nucleates the formation of a protein complex that results in phosphorylation of a series of protein kinases. Through this protein phosphorylation cascade, the original signal of insulin binding to its membrane receptor is amplified by many orders of magnitude. The protein kinase cascade initiated by insulin is sometimes referred to as a MAPK (mitogen-activated protein kinases) cascade. The cascade ultimately leads to phosphorylation of a transcription activator, Elk1, that initiates gene transcription (Figure 2).

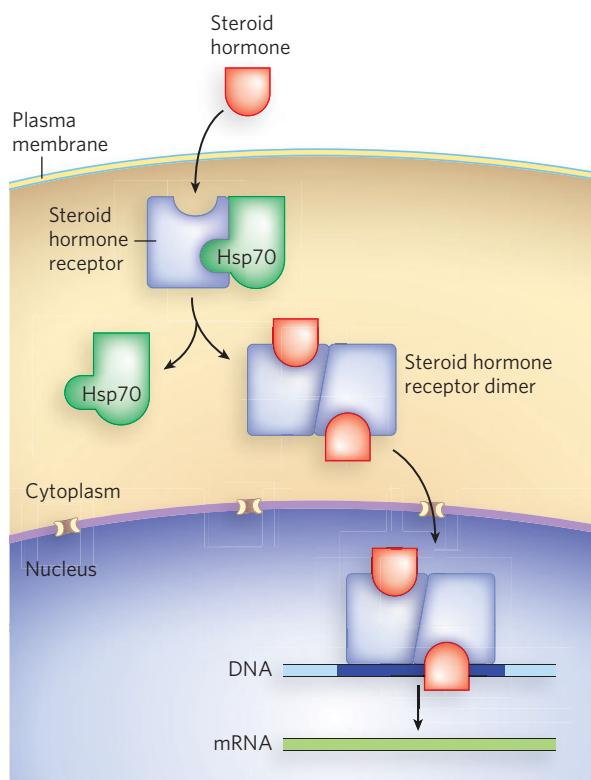
Insulin and the insulin receptor also regulate glycogen (a storage form of glucose) metabolism through another phosphorylation pathway that short-circuits the need for RNA or protein synthesis in the regulation of gene expression (Figure 3). As in the pathway described above, IRS-1 is phosphorylated by the activated insulin receptor. From here, the pathways diverge. IRS-1 binds and activates the enzyme phosphoinositide 3-kinase (PI-3K), which initiates a cascade of phosphorylation events ultimately resulting in phosphorylation of glycogen synthase kinase 3 (GSK3). Active, unphosphorylated GSK3 contributes to the slowing of glycogen synthesis by inactivating the enzyme glycogen synthase. When phosphorylated, GSK3 is inactivated, and glycogen synthase remains active in liver and muscle cells. The phosphorylation cascade initiated by insulin thus results in increased glycogen synthesis. Given that glycogen is the storage form of glucose, the insulin signal effectively removes glucose from the blood by promoting cellular glycogen production.

Glucose import into cells is yet another pathway controlled by insulin in a way that short-circuits RNA and protein synthesis (see Figure 3). Cell surface receptors for glucose uptake are controlled by insulin in response to blood glucose levels. Glucose uptake is mediated by the glucose transporter protein GLUT4, which is mainly stored in intracellular vesicles. Insulin release from the pancreas in response to high blood glucose results in fusion of



**FIGURE 3** Insulin rapidly controls changes in the cell's glycogen metabolism (by increasing glycogen synthase activity) and glucose import (by moving the receptor GLUT4 to the plasma membrane), without the need for new protein synthesis. As with the MAPK cascade, many "middle" factors, as shown here, are involved in this signal transduction pathway.

these cytoplasmic vesicles with the plasma membrane, introducing GLUT4 to the membrane and permitting glucose import. When blood glucose returns to normal, the GLUT4 receptors are returned to intracellular vesicles. In type 1 diabetes, the inability to release insulin (and thus to mobilize glucose transporters) results in low rates of glucose uptake into muscle and adipose tissue. One consequence is a prolonged period of high blood glucose after a carbohydrate-rich meal, which can lead to organ damage (and is also the basis for the glucose tolerance test for diagnosing diabetes).



**FIGURE 19-27** The regulation of a steroid hormone receptor by cellular localization. A steroid hormone enters the cell and binds its receptor, which is held in the cytoplasm by interaction with heat shock protein Hsp70. Hormone binding stimulates dissociation of the Hsp70 and dimerization of the hormone-receptor complex, which migrates into the nucleus and binds its regulatory site, activating gene transcription.

### Protein Degradation by Ubiquitination Modulates Gene Expression

Once a protein has been produced in response to an environmental signal, it is important that the protein can be removed when it is no longer needed. Cells have a regulated mechanism for targeting proteins for removal through a protein degradation pathway. An efficient mechanism for proteolysis is also important for the turnover of misfolded or unfolded proteins, enabling recycling of their amino acids for the synthesis of new proteins. For protein removal, both bacteria and eukaryotes use a large, multisubunit, barrel-shaped, ATP-dependent protease with a central chamber where proteins are degraded. The access of proteins to this protease machine is restricted to those specifically targeted for permanent removal.

Although we don't yet understand all the signals that trigger recognition of a protein for degradation,

**Table 19-1** The Relationship between N-Terminal Amino Acid Residue and Protein Half-Life

N-Terminal Residue	Half-Life
<b>Protein-stabilizing residues</b>	
Ala, Gly, Met, Ser, Thr, Val	>20 h
<b>Protein-destabilizing residues</b>	
Gln, Ile	~30 min
Glu, Tyr	~10 min
Pro	~7 min
Asp, Leu, Lys, Phe	~3 min
Arg	~2 min

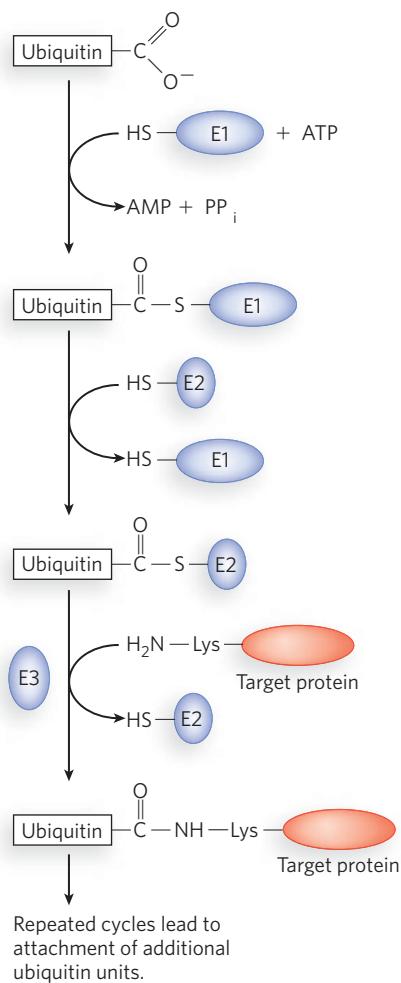
Source: Adapted from A. Bachmair, D. Finley, and A. Varshavsky, *Science* 234:179–186, 1986.

Note: Half-lives were measured in yeast for the  $\beta$ -galactosidase protein modified so that, in each experiment, it had a different N-terminal residue. Half-lives may vary among proteins and among organisms, but this general pattern seems to hold for all species.

one simple signal has been found. For many proteins, the identity of the first amino acid residue—the one that remains after removal of the N-terminal Met residue and any other posttranslational proteolytic processing of the N-terminal end (see Chapter 16)—has a profound influence on half-life (Table 19-1). These N-terminal signals have been conserved over billions of years and are the same in the protein degradation systems of bacteria and eukaryotes.

In eukaryotes, but not bacteria, regulated protein degradation is directed by the attachment of the 76-residue polypeptide ubiquitin, which, as its name suggests, is ubiquitous among eukaryotes. Ubiquitin is highly conserved; it is essentially identical in organisms as different as yeast and humans. Three enzymes are involved in the covalent attachment of ubiquitin to a protein (Figure 19-28). Two belong to large protein families that have different specificities for target proteins and therefore regulate different cellular processes. Once a protein is ubiquitinated, repeated cycles produce a long polyubiquitin chain.

Ubiquitinated proteins are degraded by the **26S proteasome** ( $M_r$ ,  $2.5 \times 10^6$ ), shown in Figure 19-29. The proteasome consists of two copies each of at least 32 different subunits, which assort into two main subcomplexes: a barrel-like core particle and a regulatory particle at each end of the barrel. The 20S core particle consists of four rings; the outer rings are formed from seven  $\alpha$  subunits and the inner rings from seven



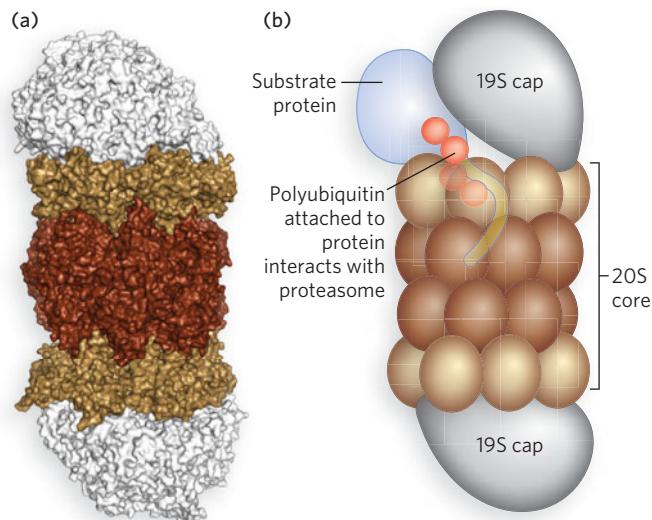
**FIGURE 19-28** The protein ubiquitination pathway. In eukaryotes, three enzymes (E1, E2, and E3) carry out the polyubiquitination of proteins in a process that involves ATP and two enzyme-ubiquitin intermediates. The free carboxyl group of the ubiquitin C-terminal Gly residue is linked through an amide bond to the  $\epsilon$  amino group of a Lys residue of the target protein. Additional cycles produce polyubiquitin, a covalent polymer that targets the protein for destruction.

$\beta$  subunits. Three of the seven subunits in each  $\beta$  ring have protease activity, each with different substrate specificity. The stacked rings of the core particle form the barrel-like structure within which target proteins are degraded. The 19S regulatory particle at each end of the core particle contains 18 subunits, including some that recognize and bind to ubiquitinated proteins. Six of the subunits are ATPases that probably function in unfolding the ubiquitinated proteins and translocating them into the core particle for degradation.

Not surprisingly, defects in the ubiquitination pathway have been implicated in a wide range of disease

states. The inability to degrade certain proteins that activate cell division can lead to tumor formation, and the too rapid degradation of proteins that act as tumor suppressors can have the same effect. The ineffective or overly rapid degradation of cellular proteins also appears to play a role in a range of other conditions: renal diseases, asthma, neurodegenerative disorders (e.g., Alzheimer disease, Parkinson disease), cystic fibrosis (sometimes caused by overly rapid degradation of a chloride ion channel), and Liddle syndrome (in which a sodium channel in the kidney is not degraded, leading to excessive  $\text{Na}^+$  absorption and early-onset hypertension). Drugs designed to inhibit proteasome function are being developed as potential treatments for some of these conditions. In a changing metabolic environment, protein degradation is as important to cell survival as is protein synthesis, and much remains to be learned about these pathways.

Bacteria and eukaryotic organelles that evolved from bacteria also have proteasome-like particles; these include ClpAP, ClpXP, HslUV, Lon, and FtsH proteases.



**FIGURE 19-29** The regulation of proteolysis by the proteasome. (a) The three-dimensional structure of the 26S proteasome is highly conserved in all eukaryotes. The two subassemblies are the 20S core particle (light and dark brown) and the 19S regulatory particle (gray), one at each end of the core. (b) The core particle consists of four rings arranged in a barrel-like structure. Each inner ring has seven different  $\beta$  subunits (dark brown), three of which have protease activity; each outer ring has seven different  $\alpha$  subunits (light brown). At each end of the core, the regulatory particle (gray) forms a cap (composed of base and lid segments). The 19S regulatory particles are thought to unfold ubiquitinated proteins (blue) and translocate them into the core particle for destruction. [Source: PDB ID 3L5Q.]

Most bacteria do not use a protein such as ubiquitin to tag proteins for degradation (although some do use a protein-tagging strategy), but their proteasomal analogs look surprisingly similar to the eukaryotic proteasome.

### SECTION 19.3 SUMMARY

- Gene regulation can occur at various steps after transcription initiation. Points of regulation involving the RNA transcript include transcript elongation, splicing, modification, and stability. The stability of mRNAs can be affected by microRNAs.
- Control of gene expression can occur at the level of translational initiation or elongation. Eukaryotes are particularly adept at regulating the initiation step.
- Gene expression is also controlled at the level of protein products by several types of covalent modification, such as phosphorylation, acetylation, and methylation. Covalent modification carries the advantage of rapidly altering protein activity without waiting for changes in transcription and translation.
- Protein targeting to particular intracellular compartments is another mechanism of gene regulation. Transcription factors can be excluded from the nucleus by phosphorylation or by binding of a regulatory protein that masks a nuclear localization signal. With degradation or modification of the regulatory protein, the transcription factor can enter the nucleus.
- Gene expression can be regulated at the level of protein stability, which typically involves degradation by protease machinery. In eukaryotes, ubiquitination is used to direct proteins to the proteasome complex for degradation.

### Unanswered Questions

The many levels of gene regulation required in cellular function and adaptation to changing conditions are coming into focus for molecular biologists. But the extra levels and structural complexities required for the development of multicellular organisms such as

humans, with 50 trillion cells, still defy the imagination. As sophisticated as our current state of knowledge is, when we look back some years from now, it will probably appear quite primitive.

1. **How extensive are the roles of microRNAs?** New miRNAs are being found frequently. They function in various ways, but the details are still scarce and the diversity of functional mechanisms is only now becoming apparent. Some miRNAs are clearly implicated in cancer, making the understanding of these small regulatory molecules extremely important to human health.
2. **How often is intracellular localization used to regulate protein function?** Regulatory mechanisms at steps other than transcription, such as intracellular localization, are being discovered at a rapid pace, and are proving to be of great importance to cellular function. Modifications that lead to compartmentalization of a protein can be quickly implemented, enabling rapid changes in the cell, and just as rapidly reversed, conserving the protein for repeated use. Because proteins and mRNAs are neither formed nor lost in this form of regulation, it provides a fuel-efficient regulatory mechanism that may have more widespread use than is currently appreciated.
3. **How do regulatory mechanisms function together in the cell or whole organism?** Our understanding of regulatory mechanisms for individual genes, and sometimes for several genes in a specific pathway, is growing. But it seems likely that for a cell to function efficiently in a complex environment, it must be capable of integrating sensory inputs of many sorts. We could hypothesize that different regulatory mechanisms engage in cross-talk, possibly resulting in vast regulatory networks. We currently know little about how different regulatory paths communicate or interconnect in the cell. Further improvements in genomic techniques for systems biology, and increased computational power to categorize and analyze the data, are likely to have a huge impact on our understanding of how whole networks of interrelated proteins are regulated during cellular function and the development of complex organisms.

# How We Know

## Plasmids Have the Answer to Enhancer Action

**Dunaway, M., and P. Dröge. 1989.** Transactivation of the *Xenopus* rRNA gene promoter by its enhancer. *Nature* 341:657–659.



**Marietta Dunaway**

[Source: Courtesy of Marietta Dunaway.]

sequences simultaneously. A simple and clever test to distinguish between these two models was performed by Marietta Dunaway and Peter Dröge in a study involving plasmids in yeast.

They placed an enhancer on one plasmid and a promoter on another plasmid, then topologically linked the plasmids together. They transferred these linked plasmids into *Xenopus* oocytes, along with a control plasmid containing the same promoter but no enhancer. If the enhancer-binding protein functioned through space, topological linkage would result in preferential activation of the promoter on the linked plasmid over that on the unlinked plasmid. But if the protein slid from the enhancer site to reach the promoter, it would not activate the promoter on either plasmid. The two promoter-containing plasmids had identical promoters but different gene sequences, which allowed Dunaway and Dröge to distinguish the level of transcription from each plasmid by a method called quantitative S1 mapping. In this method, cells are lysed and a  $^{32}\text{P}$ -labeled DNA probe is hybridized to the mRNA, then the hybrid is digested with S1 nuclease, which degrades single-stranded DNA and RNA. The DNA-RNA duplex formed by the portion of the mRNA that hybridized with the probe is protected from lysis, and its length can be observed by gel electrophoresis.



**Peter Dröge** [Source: Courtesy of Peter Dröge.]

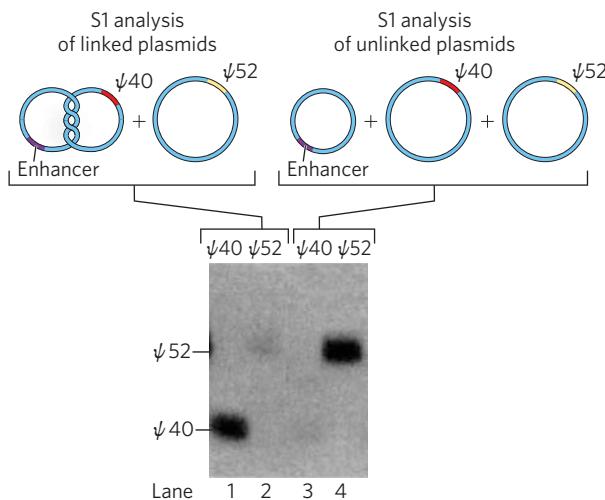
During early studies of enhancer action, two major models were proposed for how enhancer-binding proteins might work to activate a promoter at a distance. The protein could either slide along the DNA to the promoter or it could act “through space,” probably by looping out the intervening DNA so that the protein contacts the promoter and enhancer sequences simultaneously.

A simple and clever test to distinguish between these two models was performed by Marietta Dunaway and Peter Dröge in a study involving plasmids in yeast.

They placed an enhancer on one plasmid and a promoter on another plasmid, then topologically linked the plasmids together. They transferred these linked plasmids into *Xenopus* oocytes, along with a control plasmid containing the same promoter but no enhancer. If the enhancer-binding protein functioned through space, topological linkage would result in preferential activation of the promoter on the linked plasmid over that on the unlinked plasmid. But if the protein slid from the enhancer site to reach the promoter, it would not activate the promoter on either plasmid. The two promoter-containing plasmids had identical promoters but different gene sequences, which allowed Dunaway and Dröge to distinguish the level of transcription from each plasmid by a method called quantitative S1 mapping. In this method, cells are lysed and a  $^{32}\text{P}$ -labeled DNA probe is hybridized to the mRNA, then the hybrid is digested with S1 nuclease, which degrades single-stranded DNA and RNA. The DNA-RNA duplex formed by the portion of the mRNA that hybridized with the probe is protected from lysis, and its length can be observed by gel electrophoresis.

To quantify transcription from the topologically

linked plasmid versus the control plasmid, Dunaway and Dröge divided the lysate and performed S1 analysis using either a 40-nucleotide probe ( $\psi 40$ ) that specifically hybridized to the mRNA transcribed from the linked plasmid, or a 52-nucleotide probe ( $\psi 52$ ) that specifically hybridized to the mRNA transcribed from the control plasmid (Figure 1, top left). The samples were then subjected to agarose gel electrophoresis. The results revealed that when the enhancer-containing and promoter-containing plasmids are intertwined (by topological linkage), the enhancer-binding protein preferentially stimulates transcription of the gene on the linked plasmid over the gene on the control plasmid (compare lanes 1 and 2 in Figure 1). In a control experiment (see Figure 1, top right), in which all plasmids were unlinked, transcription from the control plasmid ( $\psi 52$ ) was detected more than transcription from the other, also unlinked plasmid ( $\psi 40$ ) (compare lanes 3 and 4). In all their experiments, Dunaway and Dröge used a  $\psi 52$  probe that was more radioactive than the  $\psi 40$  probe. Further experiments (not shown) revealed that transcription from both promoters in the original control experiment (lanes 3 and 4 in Figure 1) was actually about equal. Overall, these results demonstrated that the enhancer acts through space and does not need to slide along DNA to activate the promoter.



**FIGURE 1** An enhancer functions through space to activate a promoter, as shown in this experiment using topologically linked plasmids. [Source: Adapted from M. Dunaway and P. Dröge, *Nature* 341:657–659, 1989.]

## Key Terms

housekeeping gene, p. 669	negative regulation, p. 670	regulon, p. 675
constitutive gene expression, p. 669	DNA looping, p. 671	combinatorial control, p. 677
regulated gene expression, p. 669	enhancer, p. 671	recognition helix, p. 679
activation, p. 669	coactivator, p. 671	helix-turn-helix motif, p. 680
repression, p. 669	corepressor, p. 671	homeodomain motif, p. 680
transcription factor, p. 669	insulator, p. 673	basic leucine zipper motif, p. 680
activator, p. 669	signal integration, p. 673	basic helix-loop-helix motif, p. 680
repressor, p. 669	effector, p. 674	zinc finger motif, p. 681
regulatory site, p. 670	polycistronic mRNA, p. 675	RNA interference (RNAi), p. 685
positive regulation, p. 670	operon, p. 675	26S proteasome, p. 690

## Problems

- Suppose you are planning to use the yeast two-hybrid assay to identify proteins that interact with a particular target protein (see Chapter 7). The assay makes use of the ability to separate the DNA-binding domain of a typical eukaryotic activator protein from its activation domain. You genetically fuse the gene encoding the protein you are studying (the “bait”) to the gene encoding the DNA-binding domain of the bacterial protein LexA, so that they are expressed as a single fusion protein. You place the binding site for LexA upstream from *lacZ* (encoding β-galactosidase) as a reporter gene—its expression can be selected for and easily detected. How might you design the rest of this genetic screen to identify the genes encoding proteins that interact with your bait protein?
- Activator proteins A and B are required to express gene X. Analysis of the DNA upstream from the gene X promoter identified an 18 bp sequence with near twofold symmetry that is required for activation. Purification of the gene A and gene B products showed that both proteins form homodimers, but neither the A nor the B homodimer binds the 18 bp site. What are the possible functions of the A and B activators with respect to the 18 bp site? Propose a test of one of your ideas.
- Briefly describe the relationship between chromatin structure and transcription in eukaryotes.
- MicroRNAs known as small temporal RNAs (stRNAs) have been discovered in higher eukaryotes. Describe their characteristics and general function.
- An effector molecule binds to an activator protein, changing the activator’s conformation so that it is no longer active. Transcription of the gene is thus shut down. Is this positive or negative regulation?
- A transcription activator contains the following sequence:  
IARLEEKVCTLKAQNSELASTANMLTEQVAQLKQ  
The sequence includes a motif that may be used by certain transcription factors. What is this motif called? How does it function?
- In one bacterial species, investigators find a regulon that coordinates the expression of 17 genes, and identify a repressor that binds a defined site upstream from all the regulon genes. When the investigators inactivate the repressor protein, transcription of 4 of the genes increases. However, no transcription of the other 13 genes is observed, despite the presence of good promoters for RNA polymerase binding. Suggest a reason for the lack of transcription of these genes.
- A repressor protein effectively blocks transcription from bacterial gene X. A mutant form of the repressor is engineered with an altered DNA-binding site in the helix-turn-helix motif. This mutant repressor does not repress transcription from gene X. When the mutant repressor is expressed at high levels on a plasmid that is introduced into the bacterial cell, transcription of X is increased even though the wild-type repressor (capable of binding its normal DNA binding site and shutting down transcription) is present in the same cell. Explain.
- Zinc finger motifs have been appropriated for use in biotechnology. Several of these motifs can be strung together in an engineered protein, together with a fused nuclease domain, to create what has been dubbed a zinc finger nuclease. Such nucleases can be constructed to recognize and cleave almost any DNA sequence with high specificity. Explain why zinc finger motifs have been adapted for this purpose, rather than helix-turn-helix, helix-loop-helix, or homeodomain motifs.
- Steroid hormone receptors are located in the cytoplasm, where they can interact with incoming hormones. However, steroid hormones act by regulating gene function, and genes are in the nucleus. How is this regulation achieved?
- Expression of the CRP transcription activator in *E. coli* readily leads to transcription of the lactose metabolism genes when lactose is present and glucose is not. If a particular eukaryotic activator is expressed in the appropriate eukaryotic cell, introduced on an engineered virus or plasmid, it often does not trigger transcription of its target gene. Explain.

## Data Analysis Problem

- Brent, R., and M. Ptashne, M. 1985.** A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* 43:729–736.
- 12.** The concept that eukaryotic regulatory proteins have multiple functional domains was arrived at in stages. However, a few experiments stand out, such as the study by Roger Brent and Mark Ptashne published in 1985. When the study began, one known mechanism for activation of transcription by an activator protein was simply direct interaction with RNA polymerase. The investigators also considered an alternative mechanism: that the transcription activator functioned by altering the structure of the DNA to which it bound, facilitating the binding of RNA polymerase.

The study focused on two different regulatory proteins. The first was a well-characterized bacterial repressor called LexA. The LexA repressor controls a regulon in *E. coli*, the SOS response, that is activated when cellular DNA is subjected to extensive damage. The sequence of its binding site on DNA was known, and the protein had been studied by the Ptashne group and others. The second regulatory protein was a eukaryotic gene activator protein from yeast, Gal4p, which activates transcription of the *GAL1* gene when yeast cells are grown on galactose. Ptashne and his coworkers knew that the DNA-binding element of Gal4p was located in the N-terminal 74 amino acids of the protein. They deleted these amino acids and replaced them with the first 87 amino acids of the LexA protein, which they knew contained the DNA-binding elements of that protein. They then expressed this fusion protein, LexA-Gal4, in both yeast and *E. coli*. They separately expressed the native LexA protein by itself. To monitor the effects of the fusion protein in yeast, the researchers needed to construct several variants of a second plasmid containing the  $\beta$ -galactosidase gene (the *lacZ* gene, encoding an enzyme activity that is easy to measure) fused to an unrelated yeast gene, *CYC1*. The different constructs contained a variety of regulatory sequences upstream from the fusion genes.

- (a)** Suggest why the investigators did not simply examine the effects of the LexA-Gal4 fusion protein on the *GAL1* gene already in yeast cells.

The investigators first tested the LexA-Gal4 protein in *E. coli* cells that lacked their own LexA-encoding gene, and showed that cells containing the fusion protein repressed transcription of genes normally repressed by LexA.

- (b)** Why was this control experiment undertaken?

Next, they carried out a series of measurements of  $\beta$ -galactosidase activity with the two-plasmid system in yeast cells, with the results shown in Table 1 (adapted

from their published table). In the table,  $\beta$ -galactosidase activity is given in units of blue color produced in conversion of substrate to product. UAS is upstream activator sequence; UAS<sub>C1</sub> and UAS<sub>C2</sub> are binding sites for activator proteins that function at the *CYC1* gene; and UAS<sub>G</sub> is the normal binding site for Gal4p. UAS<sub>G</sub> consists of four separate binding sites for Gal4p, each 17 bp long. The 17mer is a site with just one of these sequences. The abbreviation “op” means operator, which is the LexA-binding site; -178 and -577 indicate the distance in base pairs between the operator and the transcription start site. In the yeast cells used for the study, the genes encoding the endogenous Gal4p and the *CYC1* gene activators were all present and functional.

**Table 1** The LexA-Gal4 Fusion Protein Activates Transcription of a *CYC1-lacZ* Fusion Gene

Growth Medium	Upstream Element	$\beta$ -Galactosidase Activity of Regulatory Protein	
		LexA	LexA-Gal4
Galactose	No UAS	<1	<1
	<i>lexA</i> op at -178	<1	590
	<i>lexA</i> op at -577	<1	420
	UAS <sub>C1</sub> and UAS <sub>C2</sub>	550	500
	UAS <sub>G</sub>	950	950
	17mer	600	620
Glucose	No UAS	<1	<1
	<i>lexA</i> op at -178	<1	210
	<i>lexA</i> op at -577	<1	140
	UAS <sub>C1</sub> and UAS <sub>C2</sub>	180	160
	UAS <sub>G</sub>	<1	<1
	17mer	<1	<1

- (c)** How effective is the LexA-Gal4 fusion protein (acting at the LexA operator) at activating gene expression, relative to the cellular Gal4p acting at UAS<sub>G</sub>?
- (d)** Is transcription activated by the LexA protein by itself?
- (e)** Does the location of the LexA operator affect the activity of the LexA-Gal4 fusion protein?
- (f)** When UAS<sub>G</sub> or the 17mer is upstream from the *CYC1-lacZ* reporter gene, why is expression seen only when the cells were grown in galactose?
- (g)** What result in Table 1 indicates that the LexA-Gal4 fusion protein is activating transcription by direct interaction with RNA polymerase, not by altering the structure of the DNA to which the polymerase binds?

## Additional Reading

### General

- D'Alessio, J.A., K.J. Wright, and R. Tjian.** 2009. Shifting players and paradigms in cell-specific transcription. *Mol. Cell* 236:924–931.
- Ptashne, M.** 2005. Regulation of transcription: From lambda to eukaryotes. *Trends Biochem. Sci.* 30:275–279.

### Regulation of Transcription Initiation

- Juven-Gershon, T., and J.T. Kadonaga.** 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol.* 339:225–229.
- Pan, Y., C.J. Tsai, B. Ma, and R. Nussinov.** 2010. Mechanisms of transcription factor selectivity. *Trends Genet.* 26:75–83.
- Ross, W., and R.L. Gourse.** 2009. Analysis of RNA polymerase-promoter complex formation. *Methods* 47:13–24.
- Wade, J.T., and K. Struhl.** 2008. The transition from transcriptional initiation to elongation. *Curr. Opin. Genet. Dev.* 18:130–136.

### The Structural Basis of Transcriptional Regulation

- Christensen, K.L., A.N. Patrick, E.L. McCoy and H.L. Ford.** 2008. The six family of homeobox genes in development and cancer. *Adv. Cancer Res.* 101:93–126.

**Elhiti, M., and C. Stasolla.** 2009. Structure and function of homodomain-leucine zipper (HD-Zip) proteins. *Plant Signal Behav.* 4:86–88.

**He, X., L. He, and G.J. Hannon.** 2007. The guardian's little helper: MicroRNAs in the p53 tumor suppressor network. *Cancer Res.* 67:11,099–11,101.

**Huffman, J.L., and R.G. Brennan.** 2002. Prokaryotic transcription regulators: More than just the helix-turn-helix motif. *Curr. Opin. Struct. Biol.* 12:98–106.

### Posttranscriptional Regulation of Gene Expression

- Breitkreutz, D., L. Braiman-Wiksman, N. Daum, M.E. Denning, and T. Tennenbaum.** 2007. Protein kinase C family: On the crossroads of cell signaling in skin and tumor epithelium. *J. Cancer Res. Clin. Oncol.* 133:793–808.
- Deng, S., G.A. Calin, C.M. Croce, G. Coukos, and L. Zhang.** 2008. Mechanisms of microRNA deregulation in human cancer. *Cell Cycle* 7:2643–2646.
- Shi, X.B., C.G. Tepper, and R.W. deVere White.** 2008. Cancerous miRNAs and their regulation. *Cell Cycle* 7:1529–1538.

# The Regulation of Gene Expression in Bacteria



**Bonnie Bassler** [Source: Paul Fetters Photography.]

population-wide changes in behavior; community behavior allows bacteria to carry out tasks that could never be accomplished if a single bacterium acted alone. We suspect that the evolution of cell-cell communication in bacteria is one of the first steps in the development of multicellular organisms.

*Vibrio harveyi* is a bioluminescent gram-negative marine bacterium that regulates light production in response to two distinct chemical “words,” or autoinducers. As a new professor, I wanted to answer a question that had baffled the field for several years: *Why does V. harveyi need two chemical signals for communication, when one should be sufficient?* The identity of one autoinducer, AI-1 (autoinducer-1), had been determined, but the other, AI-2, remained an enigma. Our lab cloned the gene responsible for synthesizing AI-2, and sequenced the gene. At that time there were no extensive databases of bacterial genome sequences available, only partial genomes for 40 or 50 different bacterial species. Nonetheless, we searched the incomplete database for a match to our sequence.

I recall sitting in front of the computer as the name of bacterium after bacterium scrolled up onto the screen. In the end, every single bacterium in the database contained a gene that closely matched our sequence! I realized at that moment that the bacteria were talking across species. The mysterious autoinducer AI-2 was in fact a chemical that enables different species to communicate with each other, a system that would obviously be very useful in natural settings where many different kinds of bacteria live together. This discovery changed the entire course of research in the field, and led me to focus on the mechanisms of interspecies communication over the past decade.

—Bonnie Bassler, on her discovery of *interspecies quorum sensing*

- 20.1 Transcriptional Regulation** 698
- 20.2 Beyond Transcription: Control of Other Steps in the Gene Expression Pathway** 712
- 20.3 Control of Gene Expression in Bacteriophages** 720

As we learned in Chapter 19, cells typically express just a subset of their genes at any given time. Some gene products are synthesized in large amounts, and many others in only small amounts, or not at all, depending on the needs of the cell. For example, in bacteria, proteins directly involved in DNA replication and protein biosynthesis are required continuously during active growth, whereas proteins that mediate DNA repair or the metabolism of rare sugars may typically be present only at low levels. Furthermore, requirements for many gene products change over time. The need for enzymes that participate in various metabolic pathways increases or decreases as nutrient types and levels change. The regulation of gene expression is essential for making optimal use of available energy and, more importantly, for enabling cells to adapt to a wide variety of environmental changes.

Much of what we know about gene regulation comes from studies that focused, at least initially, on bacterial systems. Microbes are masters at regulating gene expression, due to their need to adapt quickly to changing conditions. Thus they have provided investigators with many opportunities to discover fundamental mechanisms that, as it turns out, also characterize the gene regulatory pathways in humans, plants, and other eukaryotes. Recall that there are seven points in the flow of biological information where regulation can take place (see Figure 19-1). Not all of these occur in bacteria, however, due to the absence or rarity of certain of the processes, such as pre-mRNA splicing, in bacterial cells.

In this chapter, we focus on some of the central aspects of bacterial gene regulation by examining specific examples. Although much of the classic research in this field concentrated on the regulation of transcription, more recent investigations have revealed that other stages of gene expression, notably translation, provide bacterial cells with exquisite tools to fine-tune their protein levels. In addition to the roles of regulatory proteins in altering gene expression levels, regulatory RNAs—as researchers are now discovering—are ubiquitous in controlling how, when, and where proteins are made. The multiple levels of gene regulation observed in the bacteriophage  $\lambda$  infection and replication cycle set the stage for explaining the kinds of complex regulatory interactions found in eukaryotes, which we'll turn to in Chapter 21 and Chapter 22.

## 20.1 Transcriptional Regulation

Cells need to maintain control of their growth and must be able to adapt quickly to a changing environment, but they also strive for energy efficiency. For this reason, transcription is a common site of regulation, because

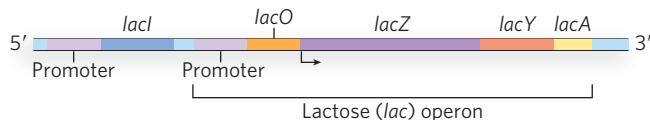
unnecessary downstream steps in gene expression can be avoided, thus conserving energy. Turning transcription off—or down—can alter protein levels without involving the cell's protein biosynthetic machinery at all. The control of transcription also permits the synchronized regulation of multiple genes encoding products with interdependent activities. For example, when their DNA is heavily damaged, bacterial cells require a coordinated increase in the levels of many DNA repair enzymes. Interactions between proteins and DNA are the key to transcriptional regulation, and much is now known about the specificity of transcriptional control.

As we discussed in Chapter 15 and Chapter 19, bacterial genes contain relatively simple promoters with sequences that allow RNA polymerase to bind and initiate transcription at specific sites. Variations in promoter sequences, or in the space between promoters and the gene(s) they control, affect transcription efficiency. In addition, bacteria have mechanisms for regulating groups of genes. For example, functionally related genes frequently cluster together in operons (see Figure 19-11), where they can be controlled by a single promoter. Alternative sigma factors, which bind and regulate RNA polymerase, also contribute to the global control of transcription (see Table 15-2). In some cases, activator and repressor proteins confer further levels of regulation by altering transcription in response to metabolites. The activities of multiple activators and repressors can converge on a single promoter to fine-tune transcription in response to various stimuli.

In this section we discuss the regulation of two bacterial operons for which mechanistic details have been particularly well established. The lactose (*lac*) and tryptophan (*trp*) operons both require multiple regulatory proteins, but the overall mechanisms of regulation are distinct. We then consider the SOS response in *E. coli*, illustrating how genes scattered throughout the genome can be coordinately regulated. Throughout this discussion we describe some of the experimental approaches that have provided insights into these regulatory pathways.

### The *lac* Operon Is Subject to Negative Regulation

Many of the principles of bacterial gene expression were first discovered in studies of sugar metabolism in *Escherichia coli*. This bacterium can use a variety of sugars as an energy source, depending on what is available in its environment. Metabolism of each sugar type requires a unique set of enzymes. The genes encoding this set of enzymes are often grouped together into an operon, which allows the genes to be coordinately



**FIGURE 20-1** The lactose (*lac*) operon of *E. coli*. The three genes of the *lac* operon are transcribed as a single unit from a single promoter. The operator region regulates transcription through interaction with the Lac repressor protein, encoded by *lacI*. The repressor is transcribed separately from the operon (i.e., has a separate promoter) and is constitutively expressed.

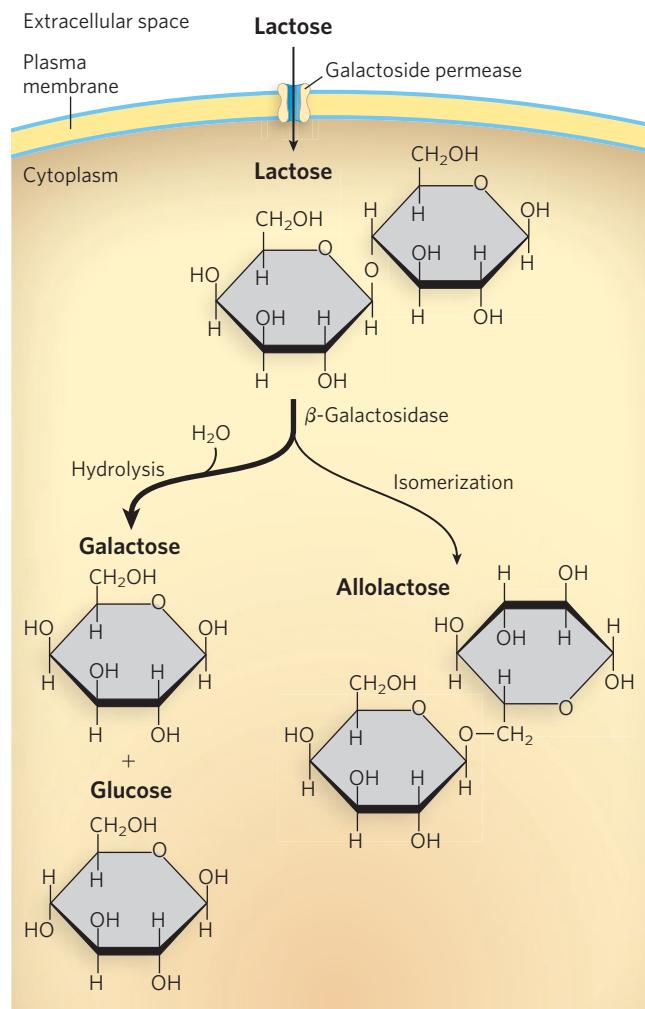
regulated. In the 1960s, two French scientists, François Jacob and Jacques Monod, examined the *E. coli* genes involved in metabolizing the sugar lactose. Through clever genetic experiments, they determined how expression of these genes is coordinately regulated in response to the presence or absence of lactose. This work, for which Jacob and Monod won the 1965 Nobel Prize in Physiology or Medicine, uncovered one of the central themes in molecular biology: some genes encode proteins with the sole function of regulating the expression of other genes.

The *lac* operon (Figure 20-1) includes the genes for β-galactosidase (*lacZ*), galactoside permease (*lacY*), and thiogalactoside transacetylase (*lacA*)—sometimes referred to collectively as the *lac* genes. Although the operon is transcribed as a single unit (i.e., the mRNA is polycistronic), the transcript contains three ribosome-binding sites, one preceding each open reading frame, that allow independent translation of each protein product. Each resulting protein functions in the metabolism of lactose. β-Galactosidase catalyzes cleavage of lactose into its components, glucose and galactose (Figure 20-2), which can then be metabolized further to generate ATP. The galactoside permease protein inserts into the plasma membrane and imports lactose into the cell. Thiogalactoside transacetylase modifies toxic galactosides that are imported along with lactose, facilitating their removal from the cell. When lactose is available, wild-type *E. coli* expresses these three genes. When lactose is unavailable, transcription from the *lac* operon is greatly reduced (Figure 20-3a).

Jacob and Monod isolated mutants of *E. coli* with defective regulation of the *lac* operon. They identified *lacI* and *lacO*, two DNA regions where mutations led to constitutive expression of the operon, whether or not lactose was present (Figure 20-3b). To understand how *lacI* and *lacO* worked, the researchers performed merodiploid analysis, a procedure that essentially makes the

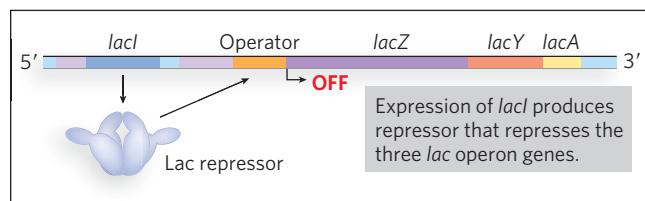
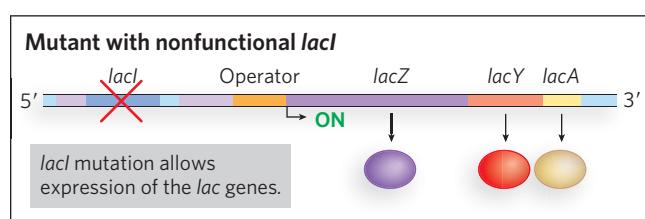
bacterial cell diploid for the *lac* operon locus (see Chapter 5, How We Know).

The first partial diploids they created combined wild-type strains with either the *lacI* or *lacO* mutants (Figure 20-3c, top and middle). The wild-type *lacI* allele was able to make up for (or “rescue”) the defect in the *lacI* mutant; these partial diploids had normal regulation of both sets of lactose metabolism genes. Jacob and Monod hypothesized that the *lacI* locus produced a diffusible product that could act on any DNA molecule, not just the DNA from which it was generated. In this case, the *lacI* gene product from the wild-type strain successfully regulated the *lac* operon DNA of the *lacI*

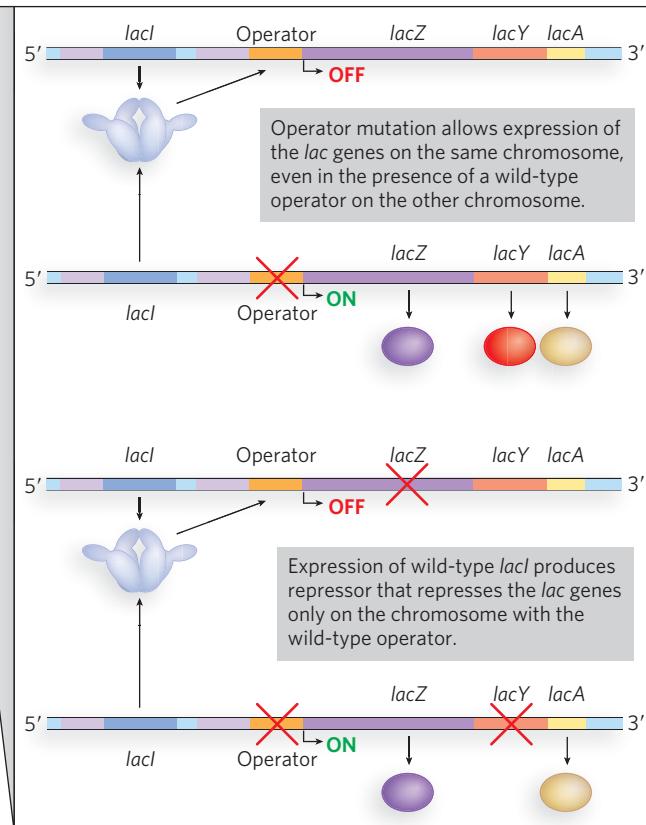
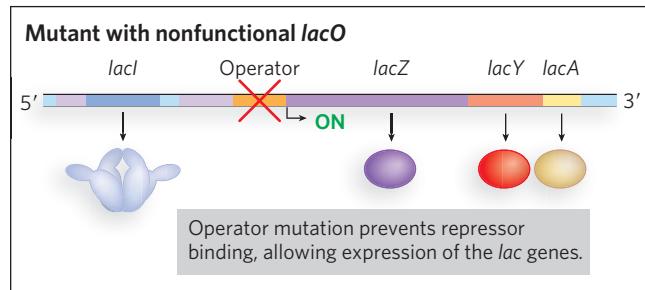
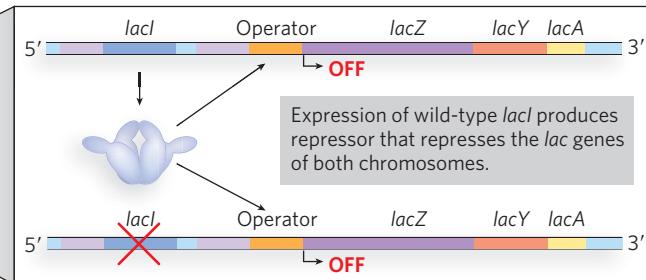


**FIGURE 20-2** Lactose metabolism in *E. coli*. Galactoside permease, encoded by *lacY*, is a membrane protein that permits entry of lactose into the cell. β-Galactosidase, encoded by *lacZ*, converts lactose to galactose and glucose, and also converts a small amount of lactose to allolactose, the *lac* operon inducer.

## (a) Wild type

(b) Mutations in normal (haploid) *E. coli*

## (c) Merodiploid analysis



**FIGURE 20-3** Jacob and Monod's merodiploid analysis of the *lac* operon. (a) The wild-type *lac* operon. (b) Jacob and Monod isolated two mutant strains of otherwise normal (haploid) bacterial cells with mutations resulting in constitutive expression from the *lac* operon. These strains had mutations in either *lacI* (the repressor gene) or *lacO* (the operator region) that rendered the gene or region nonfunctional. (c) Results from Jacob and Monod's analysis of merodiploid (partial diploid) strains carrying both a mutant and a wild-type *lac* operon suggested that the product of the *lacI* gene acts in trans (i.e., is diffusible; top) and the *lacO* region functions in cis (i.e., does not produce a diffusible product; middle). A double mutant analysis confirmed this hypothesis (bottom).

mutant. However, a wild-type copy of the *lacO* regulatory region was *not* capable of rescuing the defect in a *lacO* mutant; these partial diploids still constitutively expressed the *lac* genes. Jacob and Monod hypothesized that *lacO* did not produce a diffusible substance: the *lac* operon DNA from the *lacO* mutant could not be cor-

rectly regulated, even in the presence of a wild-type copy of *lacO*.

Jacob had served in the military, and he likened the observations on the *lac* operon to the communication link between a bomber aircraft and a ground-based radio transmitter. If the transmitter on the ground were

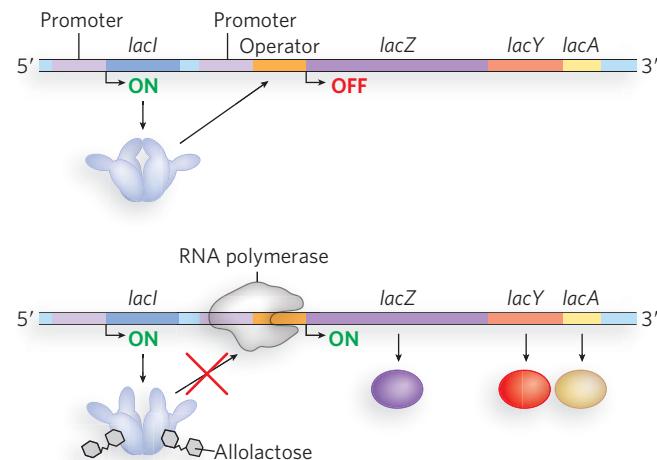
knocked out, a second transmitter could be used to direct the actions of the bomber. But if the receiver in the bomber were knocked out, neither a second transmitter nor a new bomber could direct the actions of the first bomber. Jacob and Monod hypothesized that *lacI* functioned like the transmitter; if knocked out, it could be replaced by a second transmitter. But *lacO* functioned like the receiver in the bomber: if a message could not be received, the action (transcription of the *lac* genes) could not be controlled.

With this hypothesis, the researchers tested a prediction (Figure 20-3c, bottom). In a further merodiploid analysis they combined an operon carrying a *lacZ* mutation with an operon carrying mutations in *lacY* and *lacA*. The mutations in genes encoding different enzymes (*lacZ* and *lacY*) effectively “marked” the operons and allowed the researchers to determine from which operon a gene product was coming. They predicted that the *lacI* gene product (no matter which allele produced it) would not be capable of repressing the *lacZ* gene in an operon containing a *lacO* mutation. In the absence of lactose, this diploid construction produced β-galactosidase (the *lacZ* gene product)—exactly the result they predicted! These findings confirmed that the *lacI* gene encodes a diffusible molecule (i.e., acts in trans) that represses *lac* gene expression (whether on the same or, experimentally, on a different DNA), whereas *lacO* controls only the expression of *lac* operon genes to which it is connected (i.e., acts in cis).

#### KEY CONVENTION

In genetics, genes or gene products that operate “in cis” are those that must be physically linked to have an effect. Genes or gene products that operate “in trans” can function even when not physically associated with one another (a diffusible product is involved). Note that these definitions are distinct from the conventions governing cis and trans terminology in chemistry, where these terms refer to the orientation of covalently attached functional groups with respect to each other (see Chapter 4).

From the experiments performed by Jacob and Monod, we know how the *lac* operon functions. Operon control consists of two main elements: a protein repressor (the Lac repressor, encoded by the *lacI* gene) and a DNA sequence called the operator (*lacO*) to which the Lac repressor binds. In the absence of lactose, the *lac* genes are not transcribed, because the Lac repressor binds to the operator sequence. The *lacI* gene is located near the *lac* operon, but it is transcribed from its own promoter, independent of the *lac* genes. The operator is



**FIGURE 20-4** Negative regulation of the *lac* operon by the Lac repressor. In the absence of lactose (top), the Lac repressor binds the operator region and thus prevents RNA polymerase from leaving the promoter site and transcribing the operon. In the presence of lactose (bottom), its metabolite allolactose binds the Lac repressor, resulting in a conformational change that causes the repressor to dissociate from the operator. RNA polymerase can then initiate transcription. Note: Allolactose is enlarged for clarity; it is actually much smaller relative to the Lac repressor.

adjacent to the *lac* operon promoter, and repressor binding to the operator prevents RNA polymerase from initiating transcription of the DNA (Figure 20-4). When lactose is present, a small amount of allolactose, an isomer of lactose, is produced (see Figure 20-2). Allolactose is a small effector molecule that functions as an **inducer** of the *lac* operon, binding to the Lac repressor and causing the repressor to lose affinity for and dissociate from the operator. On dissociation of the repressor, the operon becomes active, because RNA polymerase is able to initiate transcription and synthesize the polycistronic mRNA encoding the *lac* genes. It is important to note that Jacob and Monod’s initial experiments examined *E. coli* grown solely in the presence of lactose, with no other sugars available for metabolism. As we’ll see shortly, bacteria preferentially metabolize some sugars over others and impose additional levels of regulation on the *lac* operon to shut down its transcription if a more highly preferred sugar source (such as glucose) is also available.

Although it may seem simple today, the regulatory circuitry of the *lac* operon, the first to be discovered, was revealed only through powerful insight, prediction, intuitive reasoning, and creative thinking. Prior studies had focused on the fact that DNA encoded enzymes. But research on lactose metabolism revealed that some

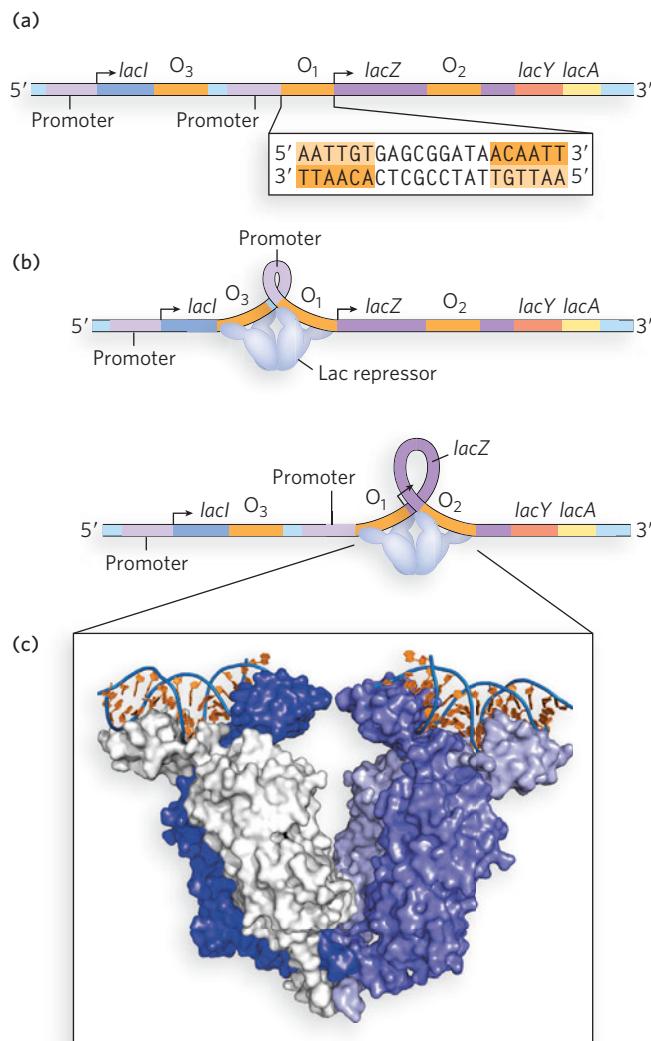
genes encode other kinds of proteins, such as DNA-binding proteins (e.g., the Lac repressor), and that some DNA sequences do not code for a gene product at all, but instead form genetic loci that affect cell function (e.g., *lacO*, the operator region).

The *lac* operon has been the subject of intense scrutiny by many laboratories since Jacob and Monod made their first observations. As research has progressed, new details of the regulatory machinery have come to light. We now know, for instance, that the operator region is more complex than suggested in Figure 20-1; in fact, there are three operator sequences.

The O<sub>1</sub> operator, to which the Lac repressor binds most tightly, abuts the *lac* operon's transcription start site (Figure 20-5a), but the operon also has two secondary binding sites for the repressor. One (O<sub>2</sub>) is about 400 bp downstream, within the gene encoding β-galactosidase (*lacZ*); the other (O<sub>3</sub>) is about 100 bp upstream, at the end of the *lacI* gene. We learned in Chapter 19 that most repressor proteins function as dimers, with each subunit binding to one half of an inverted repeat. The Lac repressor is unusual in that it functions as a tetramer of identical subunits, with two dimers tethered together at the end distant from the DNA-binding sites (Figure 20-5b, c). DNA recognition occurs in the major groove by means of a helix-turn-helix motif near the N-terminus of the repressor protein. An adjacent helix known as the hinge is important for positioning the helix-turn-helix and ensuring high-affinity DNA binding. As shown by crystal structures of the repressor, both alone and bound to inducers similar to allolactose, inducer binding causes disordering of the hinge and a resulting increase in flexibility of the DNA-binding region. DNA-binding affinity of inducer-bound Lac repressor is decreased, in a classic example of allosteric regulation (see Chapter 5).

When lactose levels are low, an *E. coli* cell contains about 20 tetramers of the Lac repressor. Each tethered dimer of the repressor separately binds to one of the three inverted-repeat operator sequences (see Figure 20-5b). To repress the operon, one dimer binds to the O<sub>1</sub> operator and the other dimer binds simultaneously to one of the two secondary sites, O<sub>2</sub> or O<sub>3</sub>. The symmetry of the O<sub>1</sub> sequence corresponds to the twofold axis of symmetry of two paired Lac repressor subunits. The tetrameric Lac repressor binds to its operator sequences *in vivo* with very high affinity, with an estimated dissociation constant of about  $10^{-10}$  M. The repressor discriminates between the operator and nonoperator sequences by a factor of about  $10^6$ , so binding to just these few base pairs among the  $4.6 \times 10^6$  bp of the *E. coli* chromosome is highly specific.

The simultaneous binding of the Lac repressor tetramer to O<sub>1</sub> and to O<sub>2</sub> or O<sub>3</sub> most likely results in a looped DNA structure, providing an effective steric block to



**FIGURE 20-5 Interaction of the Lac repressor and the operator region.** (a) The *lac* operon contains three operator sequences to which Lac repressor can bind. O<sub>1</sub> is adjacent to the *lac* operon promoter. The inverted repeat of the O<sub>1</sub> site is shown (sequence repeats shaded orange). (b) The Lac repressor tetramer can bind to O<sub>1</sub> and O<sub>2</sub> or to O<sub>1</sub> and O<sub>3</sub>. The intervening DNA is looped out. (c) Molecular model of the Lac repressor, a tetramer formed from two homodimers. Each homodimer can bind one operator sequence. [Source: (c) PDB ID 1LBG and PDB ID 2PE5.]

transcription initiation by RNA polymerase. Because each dimer of the repressor binds to a separate region of DNA, and DNA looping enables two regulatory sites to be bound at the same time, the sensitivity of the system is enhanced by the cooperative nature of the binding. In other words, the affinity of one dimer for DNA is affected by the conformation (DNA-bound or not) of the other dimer. The process of working out how the Lac repressor functions required the development of

techniques for detecting when and how proteins bind to specific sites in DNA (Highlight 20-1). These methods are still widely used to analyze the properties of DNA-binding proteins in a variety of systems.

Despite formation of an elaborate complex, transcriptional repression of the *lac* operon by the Lac repressor is not absolute. Binding of the repressor reduces the rate of transcription initiation by a factor of  $10^3$ . If the O<sub>2</sub> and O<sub>3</sub> sites are eliminated by deletion or mutation, the binding of repressor to O<sub>1</sub> alone reduces transcription by a factor of about  $10^2$ . Even in the repressed state, each cell has a few molecules of  $\beta$ -galactosidase and galactoside permease, presumably synthesized on the rare occasions when the repressor transiently dissociates from the operators. This basal level of transcription is essential to operon regulation.

When cells are provided with lactose, the *lac* operon is induced. The few existing molecules of galactoside permease enable lactose from the medium to enter the cell, where it is converted by  $\beta$ -galactosidase to allolactose, a lactose isomer (see Figure 20-2). Allolactose binds to a specific site on the Lac repressor, causing a conformational change that results in dissociation of the repressor from the operator. Release of the operator, triggered as the repressor binds to the inducer allolactose, allows expression of the *lac* genes and a  $10^3$ -fold increase in the concentration of  $\beta$ -galactosidase.

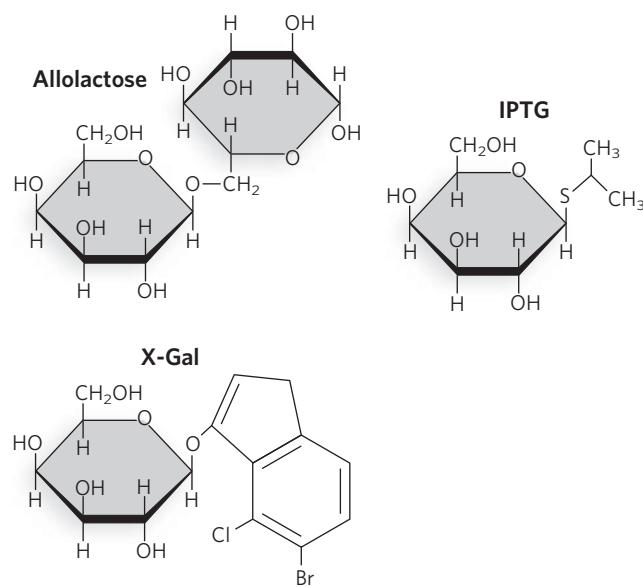
Several  $\beta$ -galactosides structurally related to allolactose are inducers of the *lac* operon but are not substrates for  $\beta$ -galactosidase; others are substrates but not inducers. One very effective and nonmetabolizable inducer of the *lac* operon often used experimentally is isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) (Figure 20-6). An inducer that cannot be metabolized lets researchers study the regulation of the *lac* operon without concern about the inducer being depleted. Equally useful as a tool in molecular biology is the noninducer substrate X-gal (5-bromo-4-chloro-3-indolyl- $\beta$ -D-galactopyranoside), which consists of galactose linked to a substituted indole.  $\beta$ -Galactosidase cleaves X-gal to produce galactose and 5-bromo-4-chloro-3-hydroxyindole, which is oxidized to an insoluble blue compound, 5,5'-dibromo-4,4'-dichloro-indigo. Bacterial colonies grown on an agar medium containing X-gal and an inducer of  $\beta$ -galactosidase (usually IPTG) turn blue if they contain a functional *lacZ* gene, a useful marker in molecular cloning (see Chapter 5, How We Know).

### The *lac* Operon Also Undergoes Positive Regulation

The operator-repressor-inducer interactions affecting the *lac* operon provide an intuitively satisfying model for an on/off switch in the regulation of gene expression.

However, operon regulation is rarely that simple. A bacterium's environment is too complex for its genes to be controlled by one signal. Other factors besides lactose affect the expression of the *lac* genes, such as the availability of glucose. Glucose, metabolized directly by glycolysis, is *E. coli*'s preferred energy source. Other sugars can serve as the main or sole nutrient, but extra steps are required to prepare them for entry into glycolysis, necessitating the synthesis of additional enzymes. Clearly, expressing the genes for proteins that metabolize other sugars, such as lactose or arabinose, is wasteful when glucose is abundant. Only in the absence of glucose is it in the cell's best interest to increase the expression of genes that allow the use of alternative food sources.

What happens to the expression of the *lac* operon when both glucose and lactose are available? A regulatory mechanism known as **catabolite repression** restricts expression of the genes required for metabolizing lactose, arabinose, or other sugars in the presence of glucose, even when these secondary sugars are also present. At first glance this may seem like another example of negative regulation, but it is a form of positive regulation for the *lac* operon. As we'll see, the *lac* operon is activated for gene expression when glucose is absent.



**FIGURE 20-6** Chemical structures of some small-molecule effectors of the *lac* operon. Like allolactose, IPTG (isopropyl  $\beta$ -D-1-thiogalactopyranoside) can bind the Lac repressor and cause its dissociation from the operator, inducing transcription of the *lac* operon. However, IPTG is not a substrate for  $\beta$ -galactosidase. The  $\beta$ -galactoside X-gal (5-bromo-4-chloro-3-indolyl- $\beta$ -D-galactopyranoside) does not induce expression of the *lac* operon, but it does serve as an experimentally useful substrate for  $\beta$ -galactosidase, producing a blue color when metabolized.

## HIGHLIGHT 20-1 TECHNOLOGY

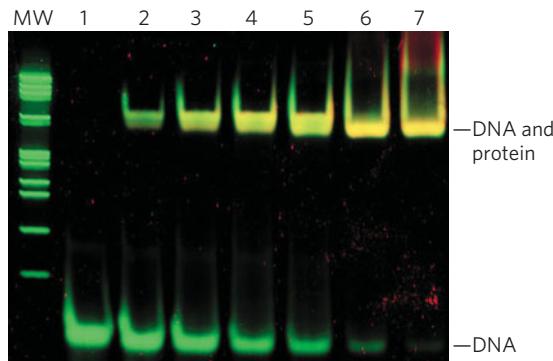
### Classical Techniques in the Analysis of Gene Regulation

Many proteins affect gene expression, and some of them do so in part by interacting directly with DNA. Researchers often want to find out whether regulatory proteins bind DNA and, if so, what sequence or structure they recognize and how tight the interaction is. The electrophoretic mobility shift assay (EMSA) and DNA footprinting experiments (in various forms) are widely used to address these questions.

Both of these techniques test the ability of a protein to interact directly with DNA. In EMSA, fragments of DNA are incubated with the protein of interest and then analyzed on a nondenaturing polyacrylamide or agarose gel. The DNA used in the experiment is visualized either by staining with a dye or by covalently attaching a radioactive phosphate group at one end. Free DNA fragments migrate more quickly through the gel than DNA bound by protein. Thus, a shift in migration of DNA from fast to slow in the presence of protein indicates a direct binding interaction between the protein and DNA (Figure 1).

Once a direct DNA-binding interaction has been established, DNA footprinting can be used to map the exact nucleotide bases in contact with the bound protein (see Figure 15-12). Nucleases are incubated with DNA-protein complexes to cleave the DNA—often radiolabeled so that it can be readily visualized; cleavage occurs at sites exposed to solvent, but not at sites that are physically protected by the presence of bound protein. Conditions are carefully controlled such that each DNA fragment is cleaved only once, generating a set of fragments that represent all possible cleavage products, with the exception of DNA protected by the bound protein. The resulting fragments are separated by denaturing gel electrophoresis and detected by exposing the gel to film or to a phosphorimager (which detects radioactive emissions from the  $^{32}\text{P}$ -labeled bands of DNA fragments). The gap in the cleavage sites where the protein associates with the DNA produces a “footprint” that indicates the boundaries of the protein-binding site. The footprint can be identified by analyzing the sites of nuclease cleavage in the DNA before and after adding the protein (Figure 2).

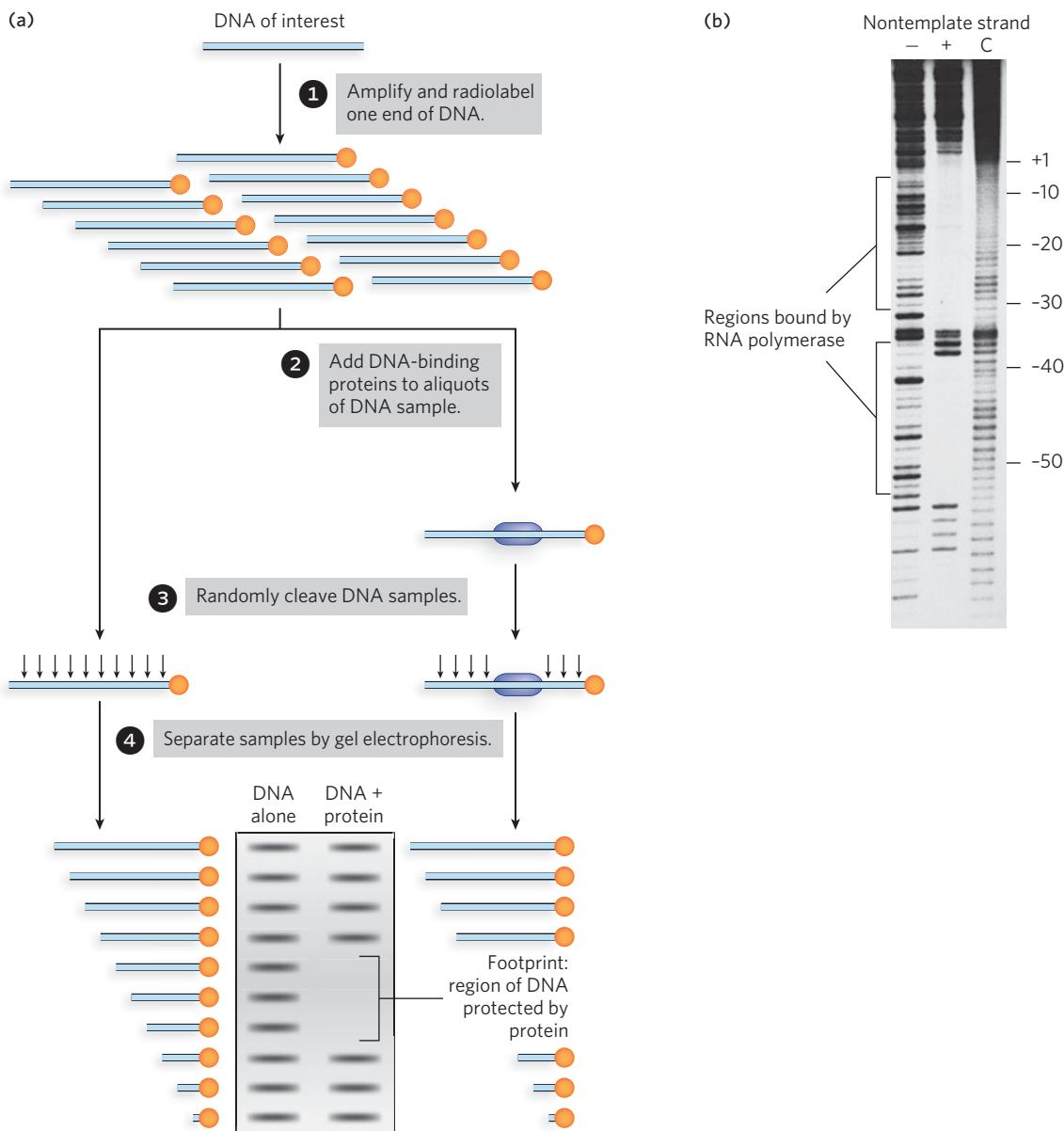
A related approach, called chemical protection footprinting, uses chemical reagents such as dimethyl sulfate to covalently modify DNA at nucleotide bases not protected by bound protein. This results in DNA



**FIGURE 1** Results of an electrophoretic mobility shift assay demonstrate the shift in migration of Lac operator DNA (lane 1) with increasing concentrations of Lac repressor protein (lanes 2 through 7). The nondenaturing polyacrylamide gel preserves the Lac repressor in its folded state so that it can interact with the operator. The gel was stained with fluorophores that bind DNA (green) and protein (red). Yellow bands indicate protein-DNA complexes. (MW lane shows molecular weight markers.) [Source: Courtesy of Life Technologies Corporation.]

containing methyl groups at sites outside the protein-binding site; the position of the protein creates a “footprint” marked by an absence of methylation sites. To locate the footprint, modified sites are detected using a DNA polymerase to copy the methylated template DNA by extending an annealed primer oligonucleotide that binds to a region just outside the DNA segment to be analyzed. The elongation activity of the DNA polymerase is blocked by the presence of a methylated base in the DNA, leading to a stop in the reaction. The products of the elongation reaction thus terminate at sites of methylated bases. These products are identified by analysis on a denaturing polyacrylamide gel, alongside primer extension reactions conducted in parallel on unmodified DNA in the presence of dNTPs plus a small amount of A, C, G, or T dideoxynucleotide, which are chain-terminating nucleotide analogs. The resulting primer-extension products in these control reactions correspond to the positions of T, G, C, or A in the sequence, enabling exact determination of the sites of the protein footprint.

Chemical modification interference, a related approach, involves first reacting the DNA with a limiting amount of a chemical reagent to introduce nucleotide modifications randomly, at just one or a few sites. The resulting pool of modified DNA molecules, each containing a modification at a different site or sites, is then



**FIGURE 2** (a) DNA footprinting analysis reveals a protein's binding site on a DNA fragment. (b) In this example, the binding site of RNA polymerase at the Lac promoter is determined using DNase to digest the lac DNA wherever the polymerase is not directly binding it and protecting it. The lanes show no polymerase added (−), polymerase added (+), and the control reaction with no DNase added (C). [Source: (b) Carol Gross, Department of Stomatology, University of California, San Francisco.]

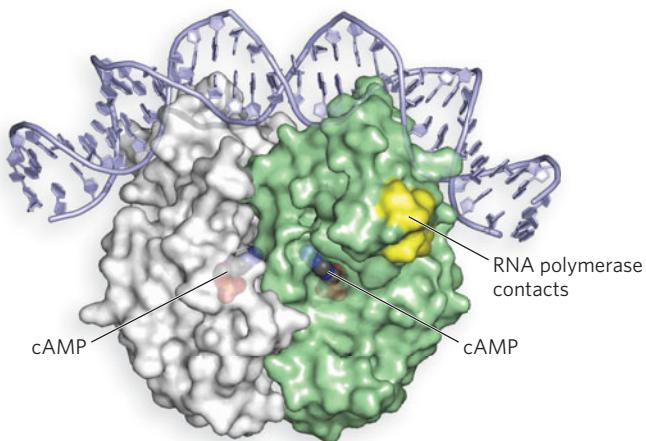
allowed to bind to protein, and free DNA is separated from DNA-protein complexes by nondenaturing gel electrophoresis, as in EMSA. Free and protein-bound DNA can then be excised from the gel, eluted from the gel matrix, and analyzed by the primer extension method. The DNA in the protein-bound sample will

contain chemical modifications only at sites that do not interfere with protein binding. DNA molecules in the unbound sample—which, not binding protein, migrated differently in the gel than the protein-bound DNA—contain chemical modifications primarily at sites that interfere with protein recognition.

The effect of glucose on expression of the *lac* operon is mediated by cyclic AMP (cAMP), a small-molecule effector, and by the activator cAMP receptor protein, or CRP, a homodimer with binding sites for DNA and cAMP (Figure 20-7). DNA binding is mediated by a helix-turn-helix motif within the protein's DNA-binding domain. When glucose is absent, CRP-cAMP binds to a site near the Lac promoter and stimulates RNA transcription fiftyfold. CRP-cAMP is therefore a positive regulatory factor responsive to glucose levels, whereas the Lac repressor is a negative regulatory factor responsive to lactose. The two act in concert.

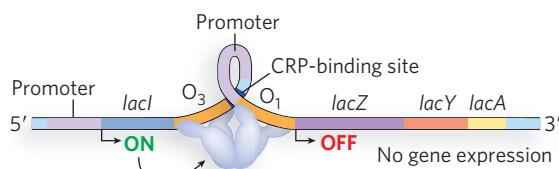
CRP-cAMP has little effect on the *lac* operon when the Lac repressor is blocking transcription (Figure 20-8a, b), and dissociation of the repressor from the operator has little effect unless CRP-cAMP is present to facilitate transcription (Figure 20-8c). When CRP-cAMP is not bound, the wild-type Lac promoter is a relatively weak promoter. The open RNA polymerase-promoter complex does not form readily unless CRP-cAMP is present. CRP interacts directly with RNA polymerase (at the region shown in Figure 20-7) through the polymerase's  $\alpha$  subunit. Binding of CRP to the  $\alpha$  subunit stimulates polymerase binding to the Lac promoter, triggering formation of the open polymerase-promoter complex.

The effect of glucose on CRP is mediated by the cAMP interaction. CRP binds to DNA most avidly when cAMP concentrations are high. When glucose is transported into the cell, the synthesis of cAMP is inhibited and efflux of cAMP from the cell is stimulated. As the cAMP concentration declines, CRP binding to DNA declines, thereby decreasing expression of the *lac*

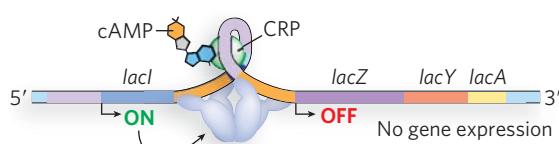


**FIGURE 20-7 Structure of the cAMP receptor protein (CRP) homodimer bound to DNA.** In the absence of glucose, the CRP-cAMP complex binds near the Lac promoter and associates with RNA polymerase to induce transcription. [Source: PDB ID 1RUN.]

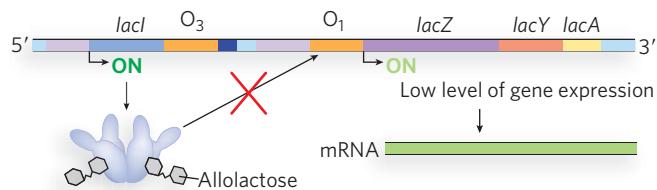
**(a) Glucose high, cAMP low, lactose absent**



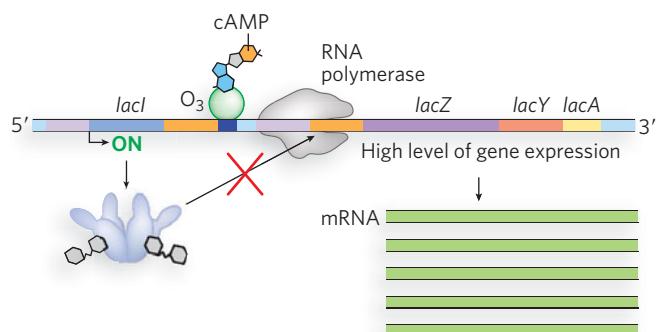
**(b) Glucose low, cAMP high, lactose absent**



**(c) Glucose high, cAMP low, lactose present**



**(d) Glucose low, cAMP high, lactose present**



**FIGURE 20-8 Positive regulation of the *lac* operon by CRP.**

The combined effects of glucose and lactose availability on *lac* operon expression are shown. (a), (b) When lactose is absent, the repressor binds the operator, blocking RNA polymerase and preventing transcription of the *lac* genes. It does not matter whether glucose is present or absent (and thus whether or not CRP-cAMP binds the operon). (c) When lactose is available, the repressor dissociates from the operator. However, if glucose is also available, low cAMP levels prevent CRP-cAMP formation and DNA binding. RNA polymerase may weakly bind the promoter and occasionally initiate transcription, leading to a very low level of *lac* operon expression. (d) Only when glucose levels are low, causing cAMP levels to rise and CRP-cAMP to bind the operon, and when lactose is present, causing repressor to dissociate, does the polymerase robustly bind and transcription proceed. Note: cAMP is enlarged for clarity; it is actually much smaller relative to CRP (see Figure 20-7).

operon. Strong induction of the *lac* operon therefore requires both lactose (to inactivate the Lac repressor) and a lowered concentration of glucose (to trigger an increase in cAMP concentration and increased binding of cAMP to CRP) (Figure 20-8d).

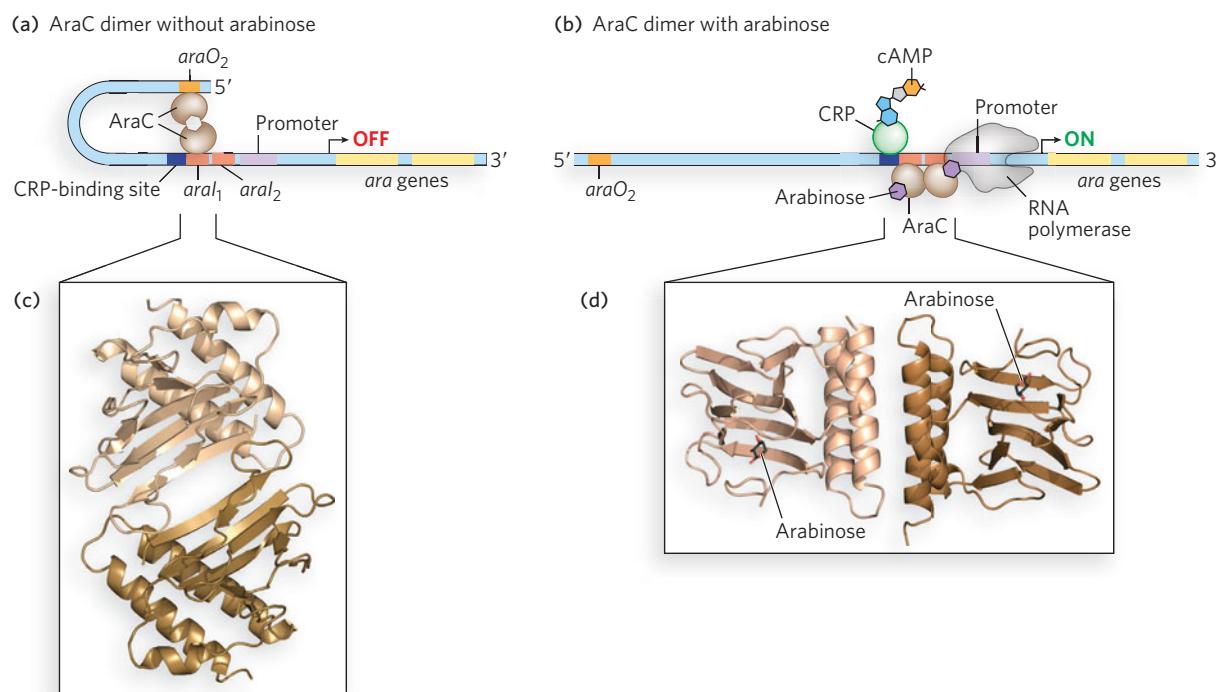
CRP and cAMP are involved in the coordinated regulation of many operons, primarily those that encode enzymes for the metabolism of secondary sugars such as lactose and arabinose. Recall from Chapter 19 that a network of operons with a common regulator is known as a regulon. Other bacterial regulons include the heat shock gene system that responds to changes in temperature (see Chapter 15) and the genes induced in *E. coli* as part of the SOS response to DNA damage (described later in this section).

### CRP Functions with Activators or Repressors to Control Gene Transcription

Other secondary sugars also trigger expression of their metabolic enzymes when present in the environment, and again, CRP provides a mechanism for activation

only in the absence of the preferred sugar, glucose. For example, arabinose metabolism is regulated by CRP and the protein AraC, which acts as either an activator or a repressor of the arabinose (*ara*) operon, depending on whether arabinose is present. When arabinose is absent, AraC forms a dimeric structure in which one AraC monomer binds to the *ara* operon gene *araI<sub>1</sub>* and the other binds a separate site much farther upstream called *araO<sub>2</sub>* (Figure 20-9a). Similar to the effect of the Lac repressor, this mode of DNA binding causes the DNA to loop into a configuration that inhibits polymerase binding. When arabinose is present, it binds to AraC, causing AraC to adopt a different dimeric conformation that allows binding to two adjacent DNA half-sites, *araI<sub>1</sub>* and *araI<sub>2</sub>* (Figure 20-9b). This positions one monomer of AraC close to the promoter, where it can recruit RNA polymerase to activate transcription.

In determining the crystal structures of the AraC arabinose-binding and dimerization domains in the presence and absence of L-arabinose, Cynthia Wolberger found that arabinose binding changes the structure of



**FIGURE 20-9 Regulation of the *ara* operon.** (a) When arabinose is absent, AraC forms a dimer in which one monomer binds to *araO<sub>2</sub>* and the other to *araI<sub>1</sub>*, preventing RNA polymerase binding and transcription of the operon. (b) Activation of the *ara* operon occurs when AraC binds arabinose (its small-molecule effector; purple) and CRP (formed in the absence of glucose). The AraC

dimer changes conformation such that one monomer binds *araI<sub>1</sub>* and the other binds *araI<sub>2</sub>*. The interaction with *araI<sub>2</sub>* recruits RNA polymerase to the promoter and activates transcription of the *ara* operon.

(c), (d) Molecular models showing the AraC dimerization domain in the absence and presence of arabinose. [Source: (c) PDB ID 2ARA. (d) PDB ID 2ARC.]

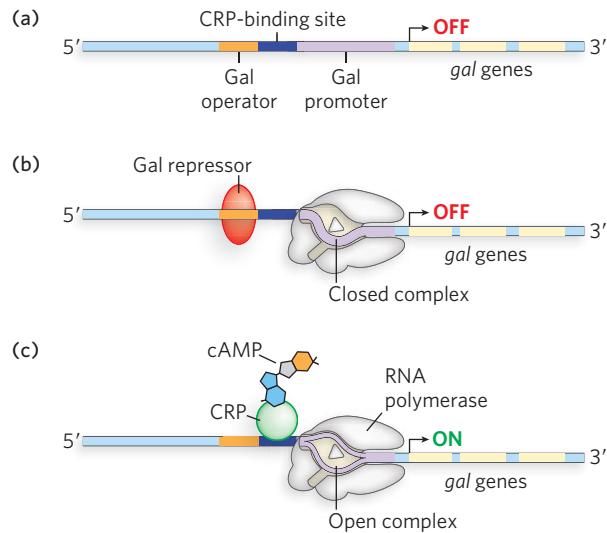


**Cynthia Wolberger** [Source: Courtesy of Cynthia Wolberger.]

In the case of the galactose (*gal*) operon, the Gal repressor inhibits transcription of the operon in the absence of galactose, and CRP-cAMP serves as the activator in the absence of glucose. The Gal repressor works differently from the Lac repressor in that it does not prevent RNA polymerase from binding to the Gal promoter. Instead, the Gal repressor probably prevents transition of the polymerase-promoter complex from the closed to the open form, thereby blocking formation of the elongation-competent form of RNA polymerase (Figure 20-10).

### Transcription Attenuation Often Controls Amino Acid Biosynthesis

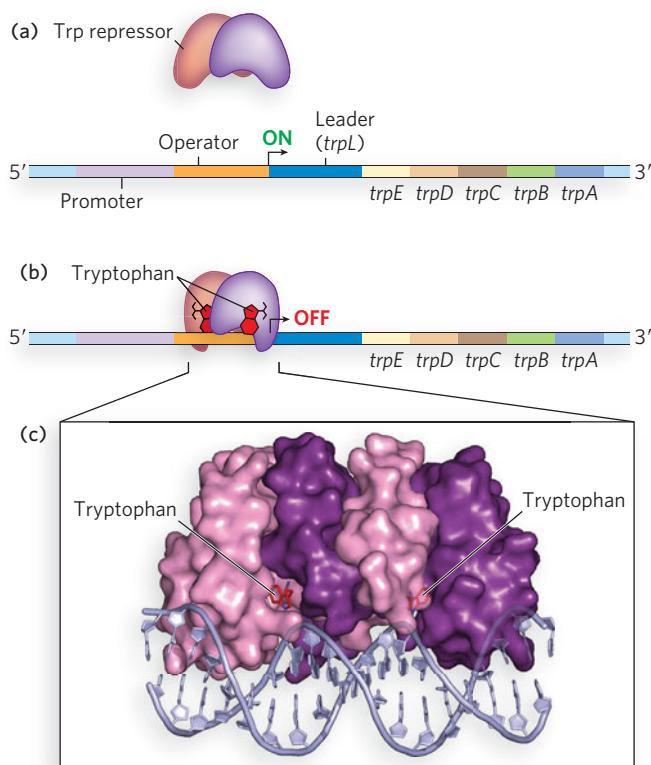
Other important small molecules, besides sugars, also help regulate expression of the genes involved in their metabolism. *E. coli* can produce all 20 of the



**FIGURE 20-10 Regulation of the *gal* operon.** (a) Structure of the *gal* operon. (b) The Gal repressor does not prevent RNA polymerase from binding the promoter; rather, it prevents formation of the open promoter-polymerase complex that is required for transcription initiation. (c) Like the *lac* and *ara* operons, the *gal* operon is transcribed only when glucose is absent, thus CRP-cAMP binding is required.

common amino acids required for protein synthesis, but biosynthesis of an amino acid is necessary only when the intracellular concentration of that amino acid is low. The genes encoding the enzymes for synthesizing an amino acid generally cluster in an operon that is repressed whenever existing supplies of that amino acid are adequate for cellular requirements. When more of the amino acid is needed, the operon is actively transcribed and the biosynthetic enzymes are expressed.

The *E. coli* tryptophan (*trp*) operon provides a classic example of the kind of regulation that enables fine-tuning of gene expression levels to suit the needs of the cell (Figure 20-11a, b). The *trp* operon includes five genes encoding the enzymes required to synthesize tryptophan. The short half-life (~3 minutes) of the mRNA transcribed from the *trp* operon allows the cell to respond



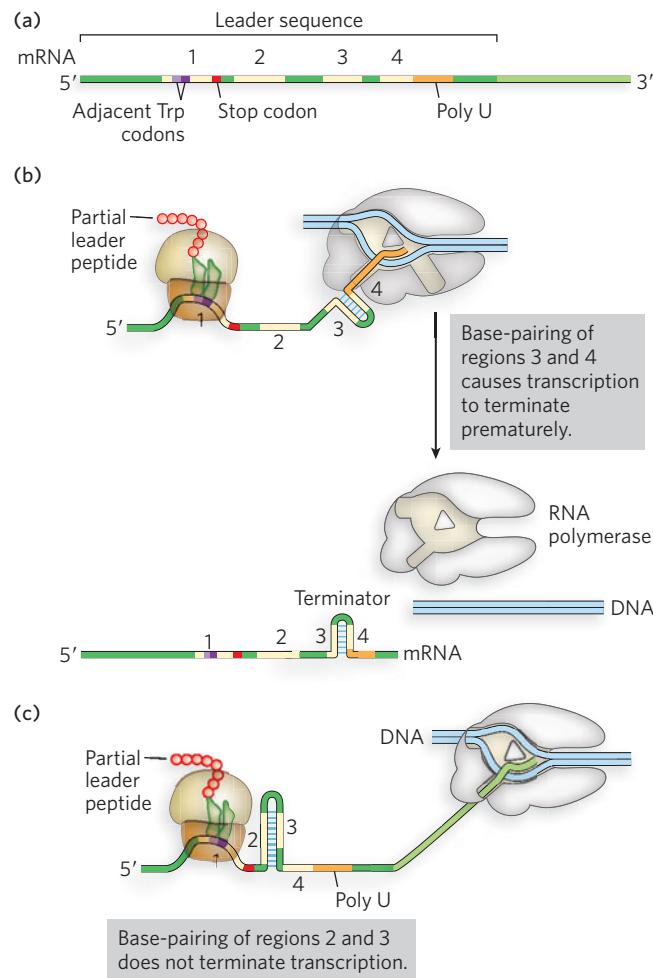
**FIGURE 20-11 Regulation of the *trp* operon.** (a) In the absence of tryptophan, the Trp repressor cannot bind the operator, and transcription of the *trp* operon is initiated. (b) When tryptophan is abundant, the protein products from the *trp* operon are no longer needed. Tryptophan serves as the effector molecule for the Trp repressor; their association causes the Trp repressor to bind the operator, blocking transcription. Notice the presence of the leader sequence; this is required for a second level of transcriptional control (see Figure 20-12). (c) Molecular model of the homodimeric *trp* repressor bound to DNA. [Source: (c) PDB ID 1TRO.]

rapidly to changing needs for tryptophan. A homodimeric repressor protein, the Trp repressor, regulates the operon. Tryptophan acts as a small-molecule effector for the Trp repressor (Figure 20-11b, c). When tryptophan is abundant, it binds the repressor and induces a conformational change that permits the repressor to bind the Trp operator and inhibit expression of the *trp* operon. The Trp operator site overlaps the promoter such that binding of the repressor blocks the binding of RNA polymerase. In this way, the *trp* operon is negatively regulated: a **corepressor** (in this case tryptophan) binds the repressor protein, rendering the repressor competent for DNA binding. This is distinct from the negative regulation of the *lac* operon, in which the Lac repressor binds the operator in the absence of inducer (allolactose), dissociating from DNA only when the small-molecule effector is present.

Once again, this simple on/off circuit mediated by a repressor protein and small effector molecule is only part of the regulatory story. Different cellular concentrations of tryptophan can alter the rate of synthesis of the biosynthetic enzymes over a 700-fold range. Repressor action accounts for only about a 70-fold difference in gene expression between the repressed and activated states of the operon. Once repression is lifted and transcription begins, the rate of transcription is modulated by a second regulatory process, **transcription attenuation**, in which transcription is initiated normally but is abruptly halted *before* the operon genes are transcribed. Attenuation provides a honing of gene expression that, when combined with repressor action, results in the 700-fold difference in expression of the tryptophan biosynthetic enzymes.

The frequency with which transcription of the *trp* operon is attenuated is regulated by the availability of tryptophan in the cell and relies on the very close coupling of transcription and translation in bacteria. This mode of regulation is necessarily unique to cells that lack a nucleus. In eukaryotic cells, where transcription and translation are physically and temporally separated, these processes cannot be coupled for the kind of attenuation described here.

The mechanism of attenuation in the *trp* operon relies on a 162-nucleotide region at the 5' end of the mRNA, called the **leader sequence**, which precedes the initiation codon of the first gene (Figure 20-12a). Within the leader sequence are four regulatory regions, sequences 1 through 4. Sequences 3 and 4 can base-pair to form a **terminator**, a G≡C-rich stem-and-loop (hairpin) structure, closely followed by a series of U residues. Formation of the terminator causes RNA polymerase to terminate transcription prematurely and dissociate from the DNA before the operon genes can be transcribed. The termination mechanism involves polymerase slowing or stalling when encountering the stable hairpin



**FIGURE 20-12** Graded control of the *trp* operon through transcription attenuation. (a) The leader sequence of the *trp* mRNA. The transcript generated from the *trp* promoter includes a leader sequence at the 5' end (containing four regulatory regions labeled 1–4). A portion of this sequence (sequence 1) is translated into the leader peptide, which has no known function other than to regulate the *trp* operon. (b) In the presence of tryptophan, the ribosome translates quickly through the Trp codons of sequence 1 and into sequence 2, allowing sequences 3 and 4 to associate to form a hairpin that stalls the RNA polymerase and terminates transcription. (c) In the absence of tryptophan, the ribosome stalls in sequence 1, allowing sequences 2 and 3 to associate. With sequence 3 unavailable to associate with sequence 4, the terminator structure is not formed and transcription can proceed. The amount of free tryptophan available for protein synthesis thus determines whether the *trp* operon is transcribed.

(terminator), then dissociating from the DNA as a result of relatively weak base pairing between the adjacent U-rich sequence and the complementary A-rich sequence in the DNA template. However, sequence 3 can also base-pair with sequence 2. When sequences 2 and 3 associate, the terminator cannot form, and uninterrupted

transcription continues into the *trp* genes. The loop formed by the pairing of sequences 2 and 3 does not block transcription.

How is hairpin choice determined? Regulatory sequence 1 and the availability of tryptophan are crucial for determining whether sequence 3 pairs with sequence 2 (letting transcription continue) or with sequence 4 (attenuating transcription). Formation of the terminator stem-and-loop structure depends on events that occur during *translation* of regulatory sequence 1. Sequence 1 encodes a **leader peptide** of 14 amino acids, two of which are Trp residues (Figure 20-12b). The leader peptide has no other known cellular function; its synthesis is simply an operon regulatory device. This peptide is translated immediately after the leader RNA is transcribed, by a ribosome on the nascent mRNA that follows closely behind RNA polymerase on the DNA as transcription proceeds.

When tryptophan concentrations are high, concentrations of Trp-tRNA<sup>Trp</sup> are also high. This allows translation to proceed rapidly past the two Trp codons of sequence 1 and into sequence 2, before sequence 3 is transcribed by RNA polymerase. In this situation, sequence 2 is covered by the ribosome and unavailable for pairing to sequence 3 when sequence 3 is synthesized; the terminator structure (sequences 3 and 4) forms instead, and transcription halts (see Figure 20-12b). However, when tryptophan concentrations are low, the ribosome stalls at the two Trp codons in sequence 1, because Trp-tRNA<sup>Trp</sup> is less available. Sequence 2 remains free while sequence 3 is transcribed, allowing these two sequences to base-pair. Sequence 3 is then unavailable for pairing with sequence 4, preventing formation of the terminator and letting transcription proceed (Figure 20-12c). In this way, the proportion of transcripts that are prematurely terminated declines as tryptophan concentration declines. Other bacteria also use multiple levels of regulation for exquisite control of tryptophan biosynthetic genes, such as the TRAP system found in *Bacillus subtilis* (see How We Know).

In *E. coli*, other amino acid biosynthetic operons use a similar attenuation strategy to fine-tune enzyme production to meet the prevailing cellular requirements. For example, the 15-residue leader peptide produced by the *phe* operon contains seven Phe residues. The *leu* operon leader peptide has four contiguous Leu residues. The *his* operon leader peptide has seven contiguous His residues. In fact, in the *his* operon and several others, attenuation is sufficiently sensitive to be the *only* regulatory mechanism.

## The SOS Response Leads to Coordinated Transcription of Many Genes

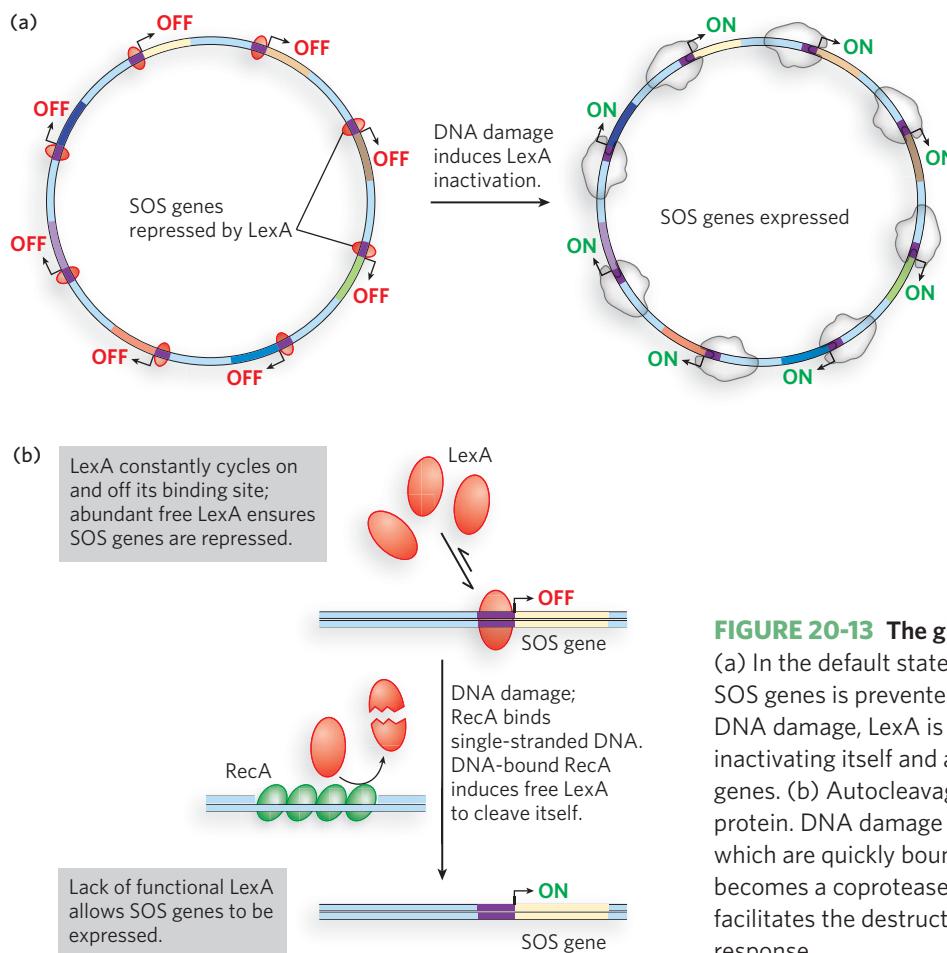
As described in Chapter 19, the many different genes that are required for a particular cell function are sometimes regulated together by a single transcription factor and/or small-molecule effector. This global regulation of transcription is an economical way for the cell to coordinate the expression of multiple genes that are needed at the same time.

An interesting example of this kind of genetic control is the cell's response to DNA damage. Extensive breakage or mutation of the bacterial chromosome triggers the expression of genes involved in DNA repair, which are located at different sites in the chromosome. This response is known as the **SOS response** and requires two key regulatory proteins: the RecA protein and the LexA repressor protein.

SOS genes encode proteins useful to cells with damaged DNA. These include Y-family polymerases, also known as translesion synthesis (TLS) polymerases, that have relaxed fidelity and can replicate DNA containing chemical lesions, such as UV-induced cross-links. The LexA repressor inhibits transcription of all the SOS genes by binding near their promoters, and induction of the SOS response requires the removal of LexA (Figure 20-13a). This is not a simple dissociation from DNA in response to the binding of a small molecule, as in the regulation of the *lac* operon. Instead, the LexA repressor inactivates itself by catalyzing self-cleavage at a specific Ala-Gly peptide bond, producing two roughly equal protein fragments. At physiological pH, this autocleavage reaction requires the RecA protein. RecA is not a protease in the classical sense, but its interaction with LexA promotes the repressor's self-cleavage reaction. This function of RecA is sometimes called a coprotease activity.

The RecA protein provides a functional link between the biological signal (DNA damage) and activation of the SOS genes. Heavy DNA damage leads to numerous single-strand gaps in the DNA, and RecA binds tightly to single-stranded DNA. Only RecA that is bound to single-stranded DNA can facilitate cleavage of the LexA repressor (Figure 20-13b). Binding of RecA at the gaps eventually activates its coprotease activity, leading to cleavage of LexA and induction of the SOS response.

Active RecA, bound to single-stranded DNA, induces cleavage of LexA molecules, a system that works in part because LexA constantly cycles on and off the DNA. Eventually, all the LexA is proteolyzed and there is no intact LexA left to repress the SOS genes.



**FIGURE 20-13** The global SOS response to DNA damage.

(a) In the default state of the *E. coli* cell, transcription of the SOS genes is prevented by the LexA repressor. In response to DNA damage, LexA is stimulated to undergo autocleavage, inactivating itself and allowing transcription of the SOS genes. (b) Autocleavage of the LexA repressor requires RecA protein. DNA damage creates sites of single-stranded DNA, which are quickly bound by RecA protein. DNA-bound RecA becomes a coprotease for LexA, and their association facilitates the destruction of LexA and induction of the SOS response.

The SOS response is an example of how a single regulatory mechanism can coordinate the expression of related sets of genes. It also provides a remarkable illustration of evolutionary adaptation. During induction of the SOS response in a severely damaged bacterial cell, RecA also facilitates cleavage of repressors that allow the propagation of certain viruses present in the cell in a dormant, lysogenic state. These repressors, like LexA, undergo self-cleavage at a specific Ala-Gly peptide bond. Induction of the SOS response permits replication of the virus and lysis of the cell, releasing new viral particles. Thus, bacteriophages have evolved to use the bacterial SOS system to their advantage, giving themselves the opportunity to make a hasty exit from a compromised bacterial host cell.

The bacterial SOS response is just one of the many ways in which cells control the expression of related genes. Another kind of mechanism involves the synthesis and detection of small molecules that can diffuse between cells in a process called **quorum sensing** (see Moment of Discovery and How We Know). Understanding how some kinds of pathogenic bacteria use quorum

sensing (and other regulatory mechanisms) to control the genes necessary for rapid growth in infected individuals could offer new avenues for therapeutic intervention.

## SECTION 20.1 SUMMARY

- Dissociation of a repressor from, or binding of an activator to, its target sequence to activate transcription can be triggered by a specific small molecule called an inducer. This was first elucidated in studies of the *lac* operon of *E. coli*. The Lac repressor dissociates from the Lac operator when the repressor binds its inducer, allolactose.
- Catabolite repression is a mechanism of positive gene regulation in bacteria in which the presence of a preferred carbon source, such as glucose, prevents the activation of operons encoding enzymes required for metabolizing secondary sugars, such as lactose and arabinose. When glucose is depleted, cAMP concentrations increase and, in turn, increase the amount of CRP-cAMP complex, which stimulates transcription of these operons.

- When arabinose is present, CRP binds to the activator protein AraC, causing AraC to dimerize and bind to two DNA sites, activating the promoter of the *ara* operon. Alternative AraC-binding sites occupied by AraC in the absence of arabinose and CRP configure the promoter in an inactive state.
- Bacterial operons that produce the enzymes of amino acid synthesis use transcription attenuation, a regulatory process that involves a transcription termination site in the mRNA. Formation of the terminator is modulated by a mechanism that couples transcription and translation while responding to small changes in amino acid concentration.
- Many biosynthetic pathway operons, such as those encoding amino acid-synthesizing enzymes, are repressed by the end product of the pathway. In this way, amino acids inhibit their own production.
- In the SOS response, multiple genes throughout the chromosome, repressed by a single repressor protein, LexA, are activated simultaneously when DNA damage triggers RecA protein-facilitated autocatalytic proteolysis of LexA.

## 20.2 Beyond Transcription: Control of Other Steps in the Gene Expression Pathway

Genetic regulation in its simplest form is the process by which cells sense their metabolic needs and modulate the levels of certain gene products in response to those needs. Transcription, particularly at the initiation step, was for a long time the focus of research efforts to understand how bacteria change gene expression in response to various signals. But, as eventually became apparent, several key points after transcription have equally interesting mechanisms for controlling levels of active protein product. These points include mRNA stability, protein synthesis, and protein modification and degradation (see Figure 19-1). As we've noted, some of the regulatory steps in Figure 19-1, such as certain aspects of mRNA processing and protein transport, are not relevant in bacteria, given the absence of a nucleus and lack of pre-mRNA splicing in these cells.

The control of gene expression at steps after transcription enables a rapid up- or down-regulation of protein synthesis in response to molecular signals. Such changes can occur more quickly at posttranscriptional levels because the cell need not wait for mRNA levels to change through increased or decreased transcription. We discuss here some of the ways in which bacteria

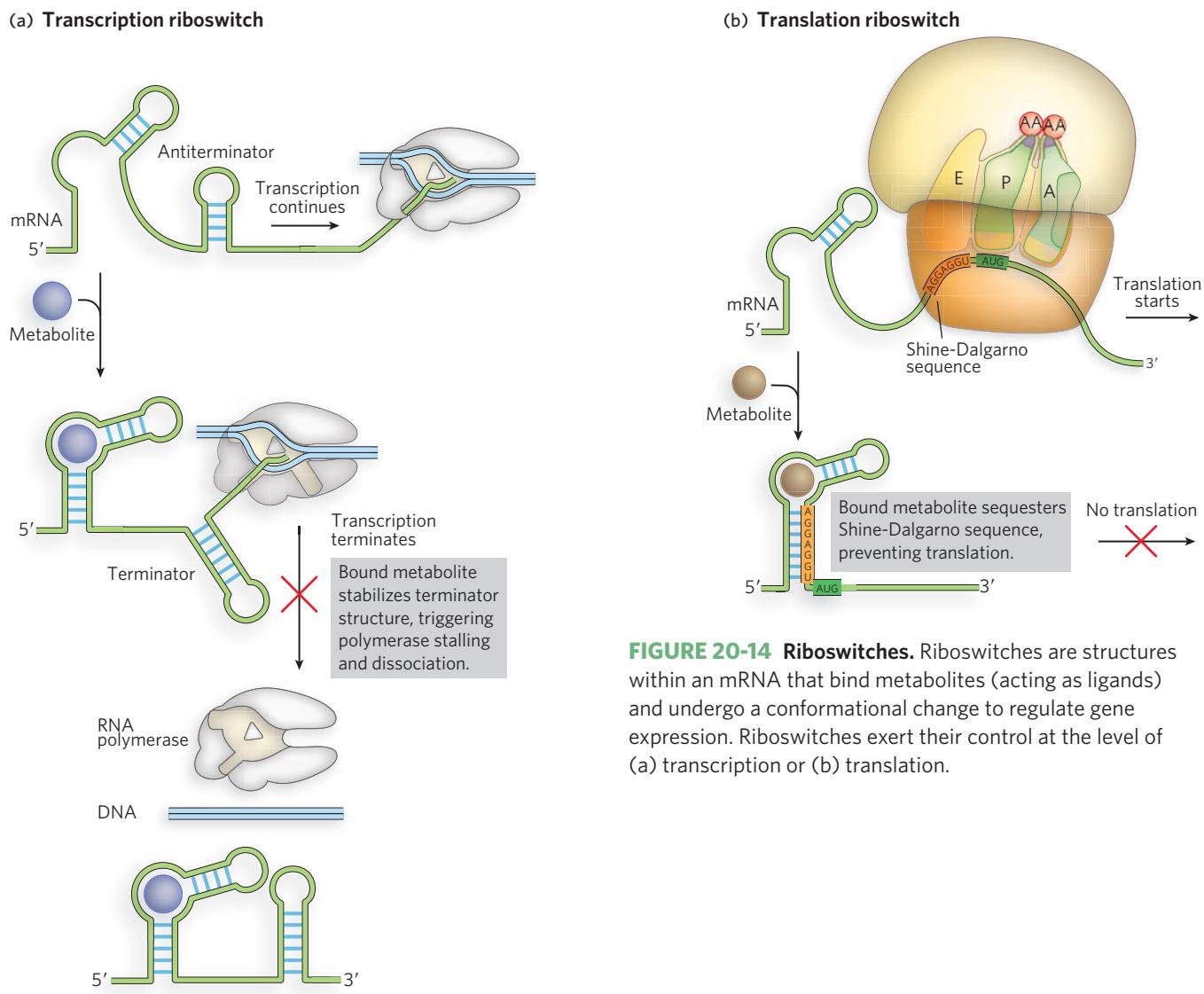
regulate gene expression beyond transcription, as a means of altering the amount of protein produced from a given mRNA. As we'll see, RNA transcripts themselves can be central players in these types of regulatory pathways.

### RNA Sequences or Structures Can Control Gene Expression Levels

The most direct way of controlling gene expression at the level of mRNA is for the RNA itself to have regulatory properties. These regulatory regions of mRNAs, called **riboswitches**, are structures that exist within the 5' untranslated region (5'UTR) of the RNA; they bind to small-molecule metabolites with the affinity and specificity required for the precise regulation of gene expression. Riboswitches consist of a small molecule-binding element connected to a regulatory region; binding of the small molecule (the ligand) triggers a conformational change in the regulatory region, such that the entire RNA molecule changes shape.

Different riboswitches have different downstream effects on gene expression; in most cases, ligand binding affects either transcription or translation (Figure 20-14). Depending on how well the ligand binds to the riboswitch, the regulation may be very sensitive to the presence of low ligand concentrations or may occur only when ligand concentrations rise to a significant level. Thus far, more than a dozen distinct classes of riboswitches have been identified, each class with common sequence and structural features, as well as distinct ligand-binding specificity, as shown in Table 20-1. There may be more riboswitch families to be discovered—these are just the ones identified so far!

When riboswitches were first discovered, researchers wondered how they could translate a binding event into a change in gene expression level. Ronald Breaker and his colleagues at Yale University carried out elegant experiments, *in vitro* and *in vivo*, to analyze what happens to an mRNA containing a riboswitch on exposure to a ligand. Researchers used chemicals and enzymes that cleave single-stranded, but not double-stranded, regions of RNA to digest riboswitch-containing transcripts (referred to simply as “riboswitch RNA”) in the presence or absence of ligand. By comparing the different patterns of RNA fragments generated with and without bound ligand, they found that riboswitches undergo conformational changes on binding to a favored ligand. Genetic experiments showed that mutations in these regulatory regions prevented changes in gene expression in response to changing



**FIGURE 20-14 Riboswitches.** Riboswitches are structures within an mRNA that bind metabolites (acting as ligands) and undergo a conformational change to regulate gene expression. Riboswitches exert their control at the level of (a) transcription or (b) translation.

**Table 20-1 Types of Riboswitches**

Riboswitch Class	Ligand	Function of Regulated Genes
FMN	Flavin mononucleotide (FMN)	Riboflavin biosynthesis and transport
THI box	Thiamine pyrophosphate (TPP)	Thiamine biosynthesis and transport
B12	Adenosylcobalamin	Vitamin B <sub>12</sub> biosynthesis and transport
S box (SAM-I)	S-Adenosylmethionine (adoMet)	Methionine and adoMet biosynthesis and transport
SAM-II	S-Adenosylmethionine (adoMet)	Methionine and adoMet biosynthesis and transport
S <sub>MK</sub> box (SAM-III)	S-Adenosylmethionine (adoMet)	Methionine and adoMet biosynthesis and transport
SAH	S-Adenosylhomocysteine (adoHcy)	Recycling of adoHcy, a metabolite of adoMet
L box	Lysine	Lysine metabolism and transport
Glycine	Glycine	Glycine metabolism
Purine	Guanine/adenine	Purine metabolism and transport
dG	2'-Deoxyguanosine	Deoxyribonucleotide biosynthesis
Cyclic di-GMP	Cyclic di-GMP	Virulence, motility, and biofilm formation
glmS	Glucosamine 6-phosphate	Glucosamine 6-phosphate biosynthesis
preQ1	7-Aminoethyl-7-deazaguanine (preQ1)	Synthesis of queuosine, a modified nucleotide in wobble position of some tRNAs
Mg <sup>2+</sup>	Magnesium	Mg <sup>2+</sup> transport
T box	Uncharged tRNAs	Aminoacyl-tRNA and amino acid biosynthesis

Source: Adapted from T. Henkin, *Genes Dev.* 22:3383–3390, 2008.



**Ronald Breaker** [Source:  
Courtesy of Ronald Breaker.]

levels of ligand. This led to the idea that riboswitch RNAs undergo structural rearrangements on binding to their target ligands, resulting in either transcription termination or sequestration of a Shine-Dalgarno sequence and thus termination of translation. Although these mechanisms are established for some riboswitches, other riboswitches may have different kinds of downstream effects stemming from their ability to undergo conformational change on binding a specific ligand.

More detailed insights have come from structural studies of the riboswitch RNAs, particularly by x-ray crystallography. To use this approach, experimenters must purify large amounts of homogeneous RNA corresponding to the riboswitch and concentrate it slowly in the presence of salts that favor crystallization. Several riboswitch structures determined in this way, in the presence or absence of a bound ligand, have revealed the nature and magnitude of the structural changes that occur on ligand binding.

For example, the thiamine pyrophosphate (TPP)-binding riboswitch, located in the 5'UTR region of mRNAs involved in vitamin B<sub>1</sub> (thiamine) biosynthesis, controls gene expression by inhibiting translation in the presence of abundant vitamin B<sub>1</sub>. This regulatory RNA region, also known as the THI box or THI element, is one of the few examples of a riboswitch found in all organisms—archaea, eukaryotes, and bacteria. Like other riboswitches, the THI box adopts a globular structure that encircles the TPP ligand (Figure 20-15a). In the absence of TPP, this structure is not energetically favored. Evidence for this conclusion comes from experiments in which the TPP riboswitch RNA is incubated in a buffered solution with or without TPP, then subjected to ribonuclease cleavage. When TPP is present, ribonuclease cleaves the riboswitch RNA at only a few positions in the nucleotide sequence, but in the absence of TPP, most of the RNA is susceptible to cleavage (Figure 20-16). This result indicated that the three-dimensional structure of the riboswitch is stable only in the presence of the TPP ligand.

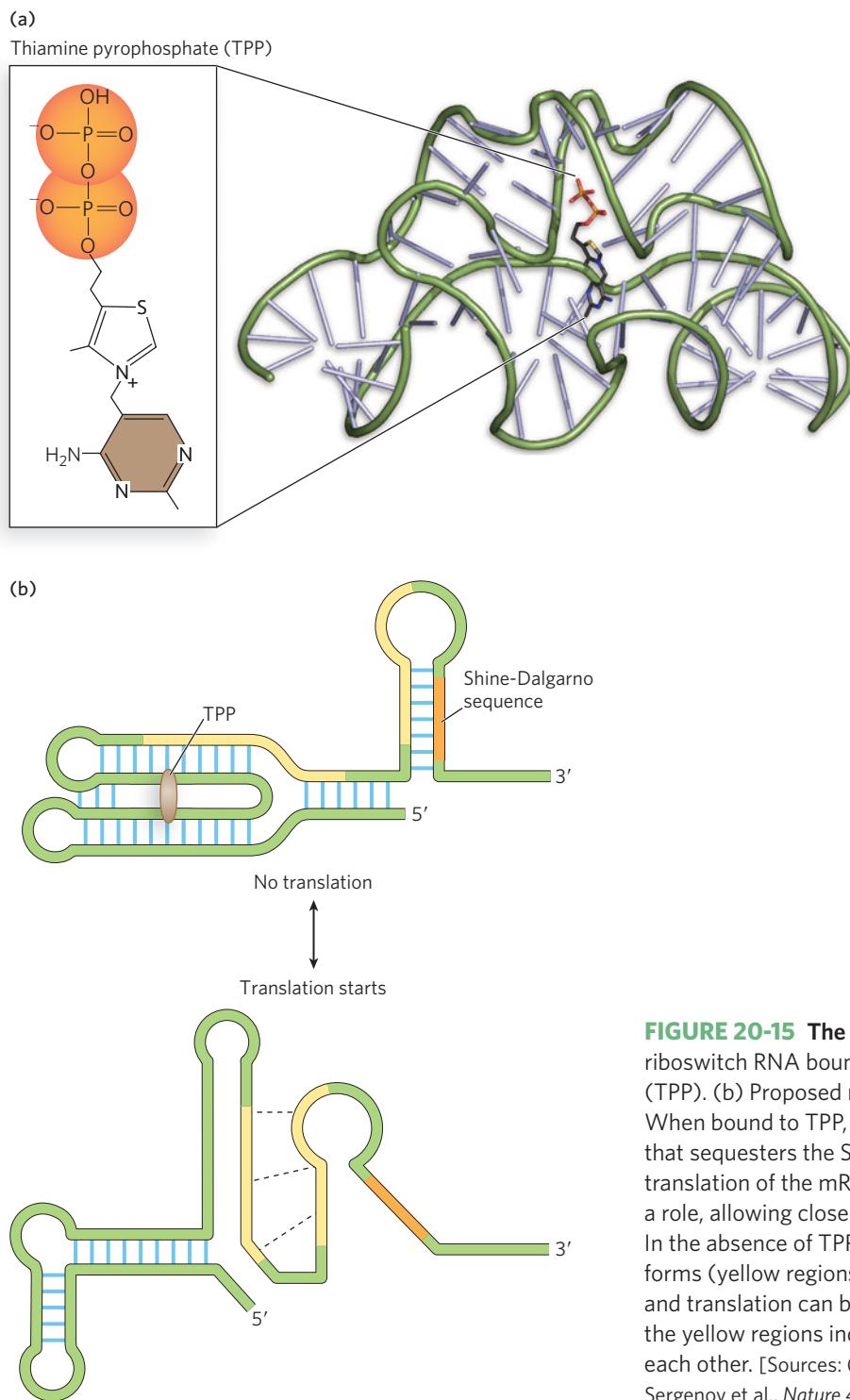
This discovery immediately suggested a mechanism by which the TPP riboswitch could regulate gene expression. Inspection of sequences containing the THI box showed that it occurs near the gene's Shine-Dalgarno sequence. When no TPP is present, the riboswitch structure does not form, or at least is

unstable, and the Shine-Dalgarno sequence is accessible for ribosome binding. However, when TPP is abundant and the cell no longer needs to synthesize the enzymes required for vitamin B<sub>1</sub> production, the riboswitch structure forms, and the Shine-Dalgarno sequence is no longer available to the ribosome. Examination of the TPP-bound form of the riboswitch shows how this works: TPP binds at the surface between two stem-and-loop structures in the RNA, creating a network of contacts between the small-molecule ligand (TPP) and the two RNA helices (Figure 20-15b). Nearby Mg<sup>2+</sup> ions help neutralize the negative charges on the pyrophosphate group of the ligand, as well as the charges on the RNA phosphodiester backbone, allowing the TPP and the RNA helices to pack close together. This mode of binding, in which the structure of the ligand-recognition cavity forms only in the presence of its cognate ligand, is an example of induced fit (see Chapter 5).

To be useful as gene regulators, riboswitches must be able to distinguish between chemically related small molecules. Riboswitch RNAs prepared by *in vitro* transcription were found to bind much better to their natural ligand than to small molecules with similar but distinct chemical structures. Riboswitch RNAs can distinguish between molecules based on atomic charge, stereochemistry, and the presence or absence of particular functional groups.

An example of the exquisite specificity of riboswitches can be seen in Breaker and colleagues' analysis of an interesting bacterial riboswitch controlling the gene *glmS*, which encodes an enzyme that catalyzes the conversion of fructose 6-phosphate and glutamine to glucosamine 6-phosphate—a metabolite that down-regulates *glmS* expression. The riboswitch, in the 5'UTR of the *glmS* mRNA, binds glucosamine 6-phosphate to form a catalytically active structure—a ribozyme—that cleaves the *glmS* mRNA and thus leads to its degradation (Figure 20-17a). Glucosamine 6-phosphate is a co-factor for the catalytic reaction and participates directly in the reaction chemistry. Experiments with different, structurally related sugars showed that sugars with even single-atom differences relative to glucosamine 6-phosphate could not support *glmS* ribozyme cleavage (Figure 20-17b, c).

Because the *glmS* ribozyme is active only when bound to glucosamine 6-phosphate, this system affords a simple but effective mechanism for reducing *glmS* transcript levels when the gene product is not needed, thereby preventing translation. Although this is the only known example of a ribozyme that uses a small-molecule cofactor as part of its catalytic mechanism, it suggests that ribozymes are inherently capable

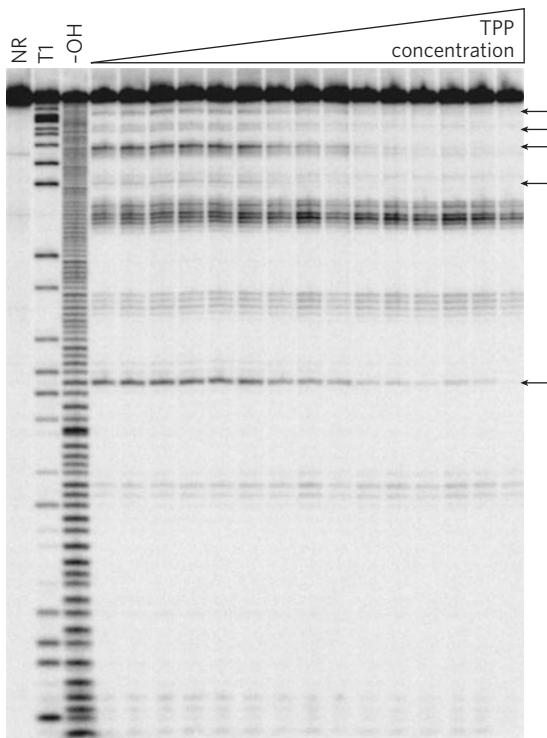


**FIGURE 20-15 The TPP riboswitch.** (a) Structure of the riboswitch RNA bound to its ligand, thiamine pyrophosphate (TPP). (b) Proposed mechanism of TPP riboswitch function. When bound to TPP, the riboswitch assumes a conformation that sequesters the Shine-Dalgarno sequence, preventing translation of the mRNA.  $Mg^{2+}$  ions (not shown) also play a role, allowing closer packing of the RNA helices and TPP. In the absence of TPP, an alternative secondary structure forms (yellow regions are complementary to each other), and translation can be initiated. Black dashed lines between the yellow regions indicate affinity of those sequences for each other. [Sources: (a) PDB ID 2GDI. (b) Adapted from A. Sergenov et al., *Nature* 441:1167–1171, 2006, Fig. 4.]

of such chemical collaborations and hence might be capable of more complex functions than have been discovered so far.

Several crystal structures are known for the *glmS* ribozyme, in the precleaved or postcleaved state or bound to ligand. Unlike the TPP riboswitch, the *glmS*

riboswitch was found to have essentially the same structure in all states. Preformed catalytic and cofactor-binding sites may favor glucosamine 6-phosphate binding, enabling greater sensitivity to glucosamine 6-phosphate levels and hence to the metabolic state of the cell. Another notable feature of the *glmS* riboswitch



**FIGURE 20-16** The effect of ligand binding on RNA susceptibility to ribonucleases. In this experiment, a transcript containing just the TPP riboswitch sequence was incubated with ribonuclease T1 in the presence or absence of TPP. The leftmost lane (NR) of the agarose gel contains untreated RNA; the dark band indicates the size of the full-length RNA. The next two lanes show the fragments generated by partial digestion of the RNA with either ribonuclease T1 (T1) or alkali ( $-OH$ ) in the absence of TPP. Many sites on the RNA are susceptible to cleavage when TPP is not present, indicating that the RNA does not form a stable three-dimensional structure. As the remaining lanes show, when the RNA is treated with ribonuclease T1 in the presence of TPP, fewer fragments are generated, suggesting that TPP promotes formation of secondary structure that protects certain residues from digestion. As the TPP concentration increases, more sites are protected, as evidenced by the disappearance of most bands (arrows). [Source: R. Welz and R. Breaker, *RNA* 13:573–582, 2007, Fig. 4a.]

structure is that the cofactor binds in an open, accessible pocket, perhaps further favoring ligand binding and ribozyme activation (see Figure 20-17a). How does glucosamine 6-phosphate participate in the ribozyme reaction chemistry? Although not yet confirmed, the crystal structures suggest that the amine group of the ligand may help to activate a specific 2'-hydroxyl group in the RNA backbone for nucleophilic attack.

In some cases, riboswitches occur in similar types of genes and share common regulatory features. For

example, levels of aminoacylated tRNAs in certain bacteria are controlled by RNA structures that respond to the concentration of specific amino acids in the cell (Highlight 20-2).

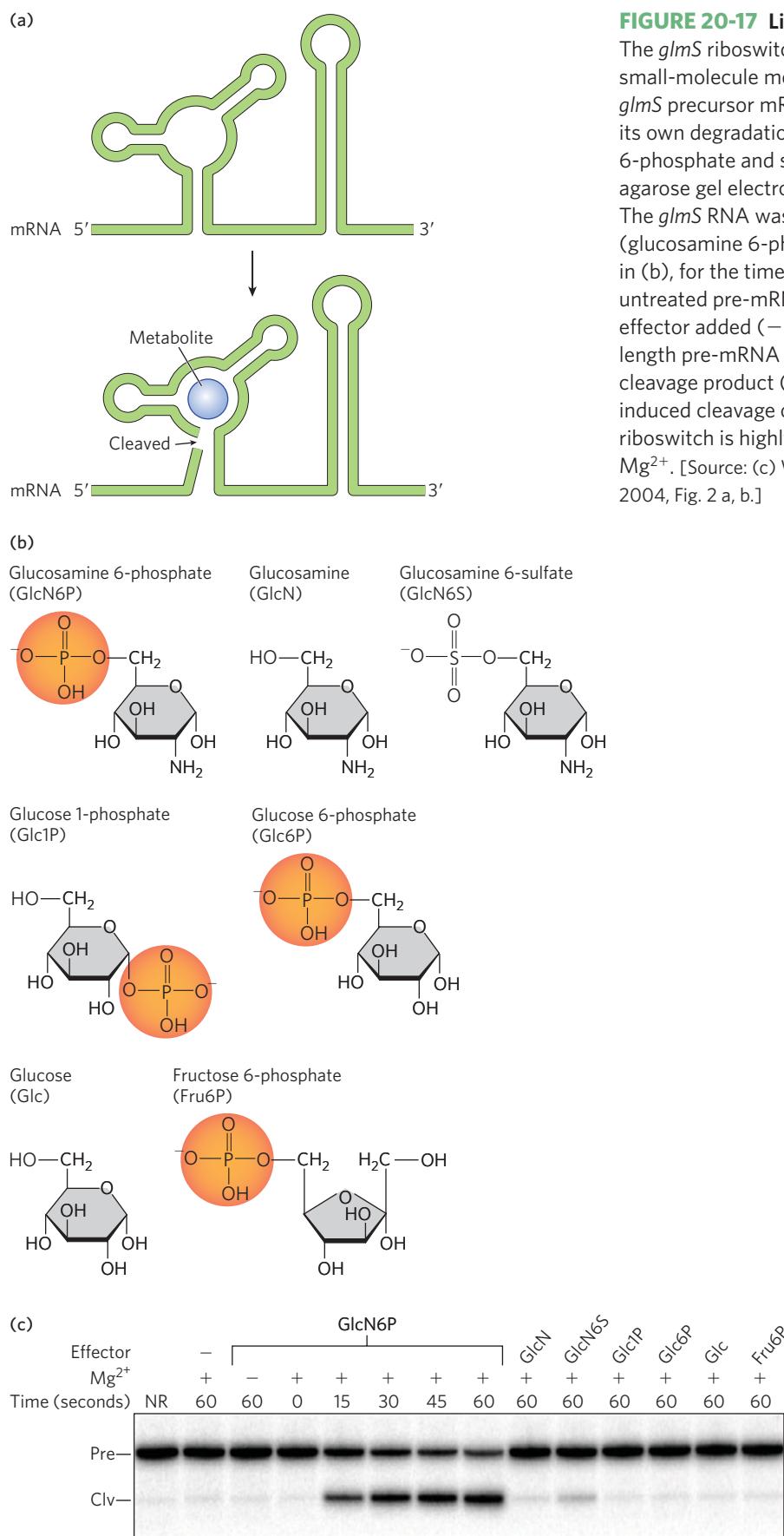
Although examples of riboswitches have been identified in all three domains of life—bacteria, archaea, and eukaryotes—they seem to be most common in the bacterial world, where they are employed in the regulation of genes involved in the biosynthesis of vitamins, enzyme cofactors, and various amino acids. What might be the reason for this greater frequency of riboswitches in bacterial cells? The use of RNA as a metabolite detector affords cells the opportunity to regulate gene expression without requiring the presence or synthesis of additional regulatory proteins. For microbes that must rapidly adjust to changing environmental conditions and availability of vitamins, enzyme cofactors, amino acids, and other small molecules, riboswitch mechanisms seem to be a highly suitable mode of regulation. However, these mechanisms may lack the multiple layers of regulatory control that are typically found in eukaryotic organisms, and they may not be sufficient to enable the kind of coordination and fine-tuning of gene expression required for such complex processes as development and organ differentiation (see Chapter 22).

### Translation of Ribosomal Proteins Is Coordinated with rRNA Synthesis

In bacteria, an increased cellular demand for protein synthesis is met by increasing the number of ribosomes, rather than increasing the activity of individual ribosomes. In general, the number of ribosomes rises as the cellular growth rate increases. At high growth rates, ribosomes make up about 45% of the bacterial cell's dry weight. The proportion of cellular resources devoted to making ribosomes is so large, and the function of ribosomes so important, that bacteria must coordinate the synthesis of the ribosomal components: the ribosomal proteins (r-proteins) and RNAs (rRNAs). This regulation occurs largely at the level of synthesis of r-proteins.

The 52 genes that encode the r-proteins occur in at least 20 operons, each containing 1 to 11 genes. Some of these operons also contain the genes for the subunits of DNA primase, RNA polymerase, and the elongation factors required for protein synthesis—revealing the close coupling of replication, transcription, and protein synthesis during bacterial cell growth.

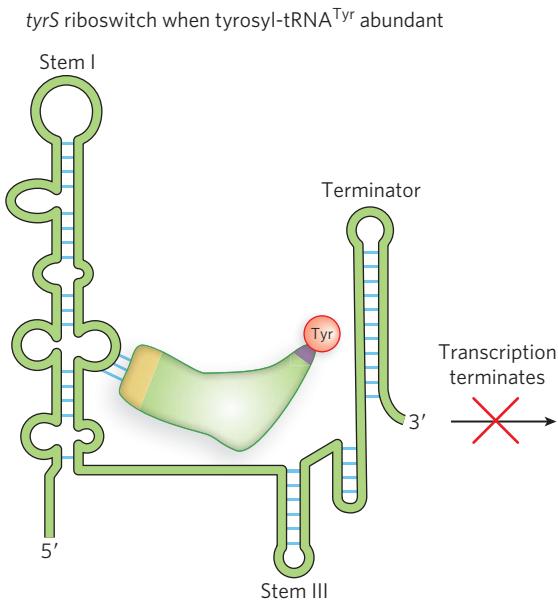
The r-protein operons are regulated primarily through a translational feedback mechanism. One r-protein encoded by each operon also functions as a **translational repressor**, which binds the mRNA transcribed from that operon and blocks translation of all



## HIGHLIGHT 20-2 A CLOSER LOOK

### T-Box Riboswitches

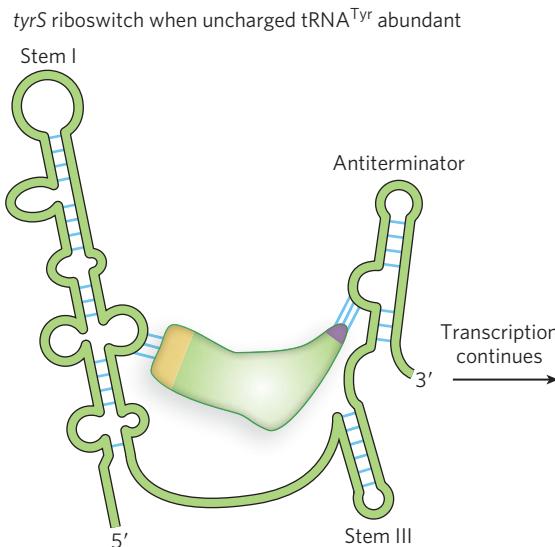
RNA molecules rather than proteins are sometimes employed by bacterial cells to detect the presence of small molecules. While studying how the soil bacterium *Bacillus subtilis* responds to environmental changes, Tina Henkin and her colleagues at the Ohio State University discovered that the *tyrS* gene transcript, encoding tyrosyl-tRNA synthetase, includes an RNA structure in the noncoding leader region that monitors levels of tyrosine in the cell. Rather than the *tyrS* mRNA leader region binding directly to tyrosine, the limitation of tyrosine availability in the cell is detected by interaction of the leader region with uncharged tRNA<sup>Tyr</sup>. The anti-codon of the tRNA<sup>Tyr</sup> base-pairs with a single detector “codon” in the mRNA leader, and the amino acid arm of the tRNA<sup>Tyr</sup> makes a second interaction with the leader, promoting RNA polymerase read-through of a structure that would otherwise cause transcription termination (Figure 1). In this way, the depletion of tyrosine, leading to increased levels of uncharged tRNA<sup>Tyr</sup>, enhances transcription of mRNA encoding tyrosyl-tRNA synthetase, which in turn charges more tRNA<sup>Tyr</sup> to Tyr-tRNA<sup>Tyr</sup>. This mechanism is not unique to tyrosine: the leader regions of at least 18 aminoacyl-tRNA synthetase and amino acid biosynthesis gene transcripts in *B. subtilis* and related gram-positive bacteria have conserved structural features similar to those of the *tyrS* gene, including the anti-TRAP protein that helps regulate tryptophan levels (see How We Know).



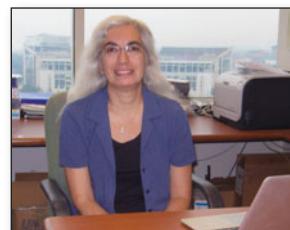
**FIGURE 1** The *tyrS* riboswitch senses the cellular level of uncharged tRNA<sup>Tyr</sup>. Both charged and uncharged tRNA<sup>Tyr</sup> interact with the leader region of the mRNA at the detector codon, but only uncharged tRNA<sup>Tyr</sup> can make a second interaction between its amino acid arm and the

Collectively, these RNAs are known as the T-box family of regulators, or T box riboswitches.

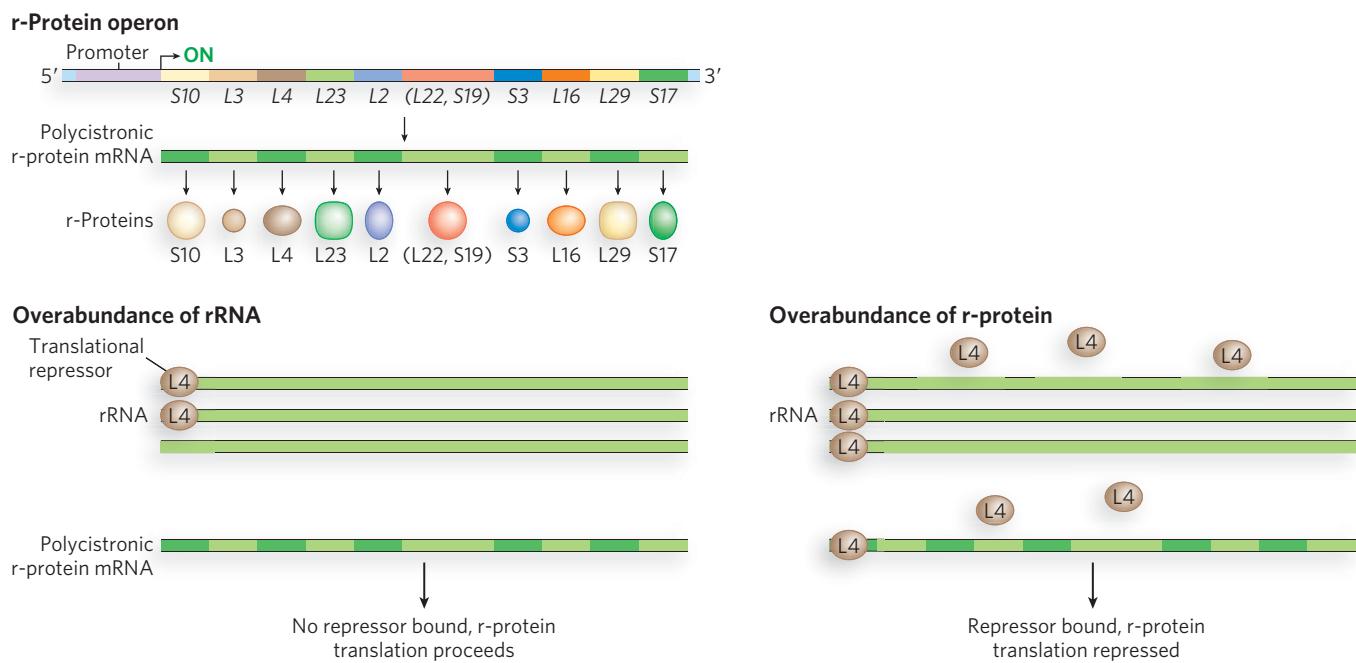
Using genetic methods, Henkin and her colleagues found that mutations in these regulatory regions of the mRNA prevent or change the ability of the genes to respond to specific amino acid levels. For example, a change in the detector codon of the *tyrS* mRNA leader from UAC (which base-pairs with tRNA<sup>Tyr</sup>) to UUC (which base-pairs with tRNA<sup>Phe</sup>) resulted in loss of transcription induction by low tyrosine levels and a switch to induction by low phenylalanine levels. In contrast, insertion of an extra nucleotide immediately before the detector codon did not affect regulation. Because such an insertion would change the reading frame of the ribosome, this finding indicates that the regulation results directly from RNA-mediated structural changes, rather than requiring production of a protein. Ultimately, Henkin demonstrated that *tyrS* anti-termination can occur in a purified transcription system with no additional cellular factors, indicating that the leader RNA is sufficient for specific recognition of the cognate tRNA. These findings were some of the first hints that RNA molecules play broader roles in gene regulation than previously expected.



antiterminator. This second interaction stabilizes the antiterminator and promotes continued transcription of the tyrosyl-tRNA synthetase gene. [Source: Adapted from T. M. Henkin and F. J. Grundy, *Cold Spring Harb. Symp. Quant. Biol.* 71:231-237, 2006, Fig. 2b.]



**Tina Henkin** [Source: Courtesy of Tina Henkin.]



**FIGURE 20-18** Regulation of r-protein operons through a translational feedback loop. In many r-protein operons, such as the one shown here, one of the r-proteins produced by the operon (in this case, protein L4) also functions as a translational repressor in a mechanism that involves sensing the relative levels of r-protein and rRNAs. L4 protein has a higher affinity for rRNA and will bind it preferentially over the

the genes encoded by the mRNA (Figure 20-18). In general, the r-protein that plays the role of repressor also binds directly to an rRNA. Each translational repressor r-protein binds with higher affinity to the appropriate rRNA than to its mRNA, so the mRNA is bound and translation is repressed only when the level of the r-protein exceeds that of the rRNA. This ensures that translation of the mRNAs encoding r-proteins is repressed only when synthesis of these r-proteins exceeds the level needed to make functional ribosomes. In this way, the rate of r-protein synthesis is kept in balance with rRNA availability.

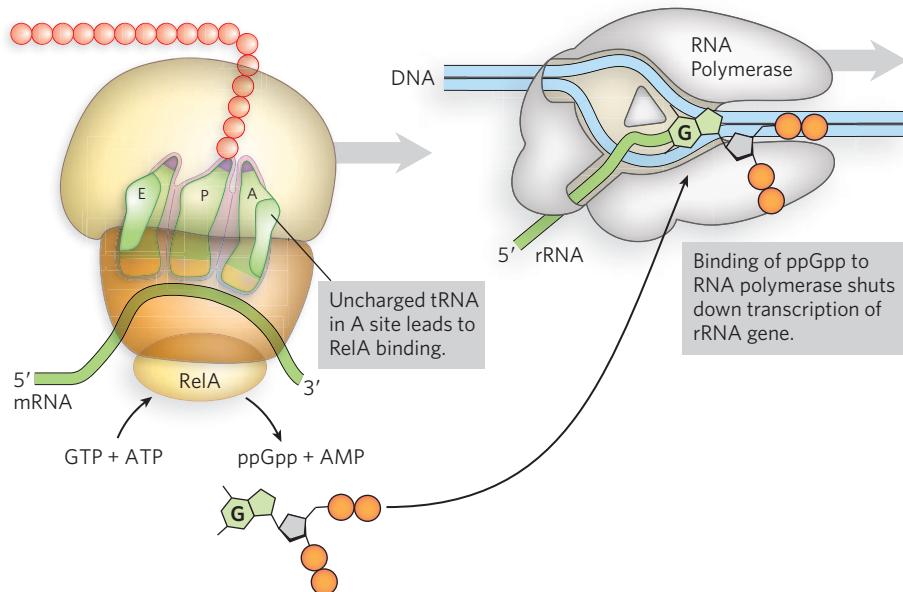
The mRNA binding site for the translational repressor is near the start site of one gene in the operon, usually the first gene (see Figure 20-18). In other operons this would affect only that one gene, because in polycistronic mRNAs, most genes have independent translation signals. In r-protein operons, however, the translation of one gene depends on the translation of all the others. The mechanism of this translational coupling is not yet understood in detail. In some cases, the translation of multiple genes seems to be blocked by folding of the mRNA into an elaborate three-dimensional structure that is stabilized by both internal base pairing and binding of the translational repressor. When the repressor is absent, ribosome binding and translation of

r-protein mRNA. When levels of r-protein are low relative to rRNA, few L4 protein molecules are available to bind the mRNA and translation proceeds, making more r-proteins. But when r-protein concentrations build up in excess of rRNA levels, the L4 protein binds to the mRNA generated from the operon and blocks production of additional r-proteins by preventing efficient initiation of translation.

one or more genes disrupts the folded structure of the mRNA, allowing all the genes to be translated.

Because the synthesis of r-proteins is coordinated with the availability of rRNA, ribosome production reflects the regulation of rRNA synthesis. In *E. coli*, rRNA synthesis from the seven rRNA operons responds to cellular growth rate and to changes in the availability of crucial nutrients, particularly amino acids. The regulation that is coordinated with amino acid concentrations is known as the **stringent response** (Figure 20-19). It enables cells to direct resources away from growth and division and toward the biosynthesis of amino acids, to ensure survival until amino acid availability increases. The effects include down-regulation of rRNA and tRNA transcription. When amino acid concentrations are low, rRNA synthesis is halted. Amino acid starvation leads to the binding of uncharged tRNAs to the ribosomal A site, triggering a sequence of events that begins with the binding of an enzyme called **stringent factor** (RelA protein) to the ribosome. When bound to the ribosome, stringent factor catalyzes formation of the unusual nucleotide guanosine tetraphosphate (ppGpp). The enzyme adds pyrophosphate to the 3' position of GTP, in the reaction:



**FIGURE 20-19** The stringent response in *E. coli*.

Synthesis of rRNA from the seven rRNA operons is regulated by amino acid concentrations. When amino acid supplies are low, uncharged tRNAs can enter the

ribosomal A site, triggering the stringent response and repressing rRNA synthesis. The formation of ppGpp actually occurs in two steps; the details of the reaction are described in the text.

A phosphohydrolase then cleaves off one phosphate to form ppGpp. The abrupt rise in ppGpp concentration in response to amino acid starvation results in greatly reduced rRNA synthesis; this is mediated, at least in part, by the binding of ppGpp to RNA polymerase, blocking transcription of the rRNA genes.

Like cAMP, ppGpp belongs to a class of modified nucleotides that act as cellular second messengers (see Chapter 6). In *E. coli*, ppGpp and cAMP serve as starvation signals; they cause large changes in cellular metabolism by increasing or decreasing the transcription of hundreds of genes. In eukaryotic cells, similar nucleotide second messengers also have multiple regulatory functions. The coordination of cellular metabolism with cell growth is highly complex, and further regulatory mechanisms undoubtedly remain to be discovered.

r-protein encoded by each operon functions as a translational repressor by binding to the mRNA transcribed from that operon and blocking translation of all the genes it encodes.

- In the stringent response, a starvation-induced pathway, stalled ribosomes synthesize the small signaling molecule ppGpp, which binds to RNA polymerase and reduces transcription of rRNA genes (and thus the number of ribosomes) and other genes needed for rapid growth.

## 20.3 Control of Gene Expression in Bacteriophages

Our discussion of gene regulatory pathways up to this point has focused on bacteria. Historically, however, much of the early research on gene regulation was done on viruses—specifically, bacteriophages (phages), viruses that infect bacterial cells. Bacteriophage genomes are small, yet the genes they contain must be carefully regulated to enable efficient infection and viral reproduction. In addition, bacteriophage genes are frequently transferred to or from host cell chromosomes, thereby providing a critical means of transferring genetic information between organisms (i.e., from one host to another). Such horizontal gene transfer plays a significant role in driving

### SECTION 20.2 SUMMARY

- Riboswitches regulate gene expression without the need for a separate regulatory protein, such as a repressor or an activator, to respond to signaling molecules. Direct binding of a riboswitch to a small-molecule ligand triggers a conformational change in the adjacent regulatory region that alters mRNA stability or translation efficiency.
- Ribosomal protein operons are regulated primarily through a translational feedback mechanism. One

the evolution of new bacterial traits, including resistance to drugs and toxins.

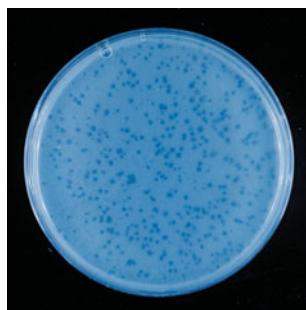
We focus here on the bacteriophage  $\lambda$  ( $\lambda$  phage) system, one of the best-studied systems of gene regulation in all of biology. Gene regulation in  $\lambda$  phage is intimately coupled to the state of the host cell, and we discuss some examples illustrating the principles involved. As we'll see,  $\lambda$  phage gene regulation involves a series of integrated pathways that exemplify the coordinated gene expression occurring in cells: some of the mechanisms used by  $\lambda$  phage have been found to govern gene expression in other systems. For example, animal cells take advantage of differential binding affinities and cooperative interactions of regulators to turn genes and gene networks on and off during development. Thus, insights from the  $\lambda$  phage

system continue to guide our understanding of more complex organisms.

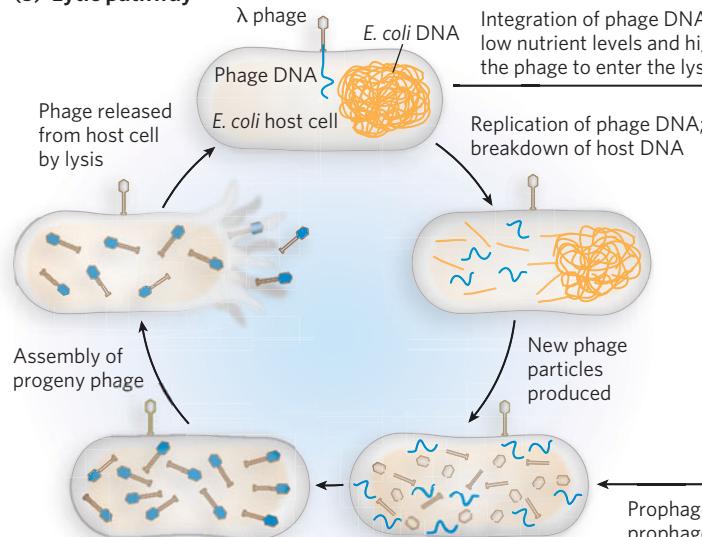
### Bacteriophage Propagation Can Take One of Two Forms

Most organisms are susceptible to infection by viruses, which usurp the host cell machinery to produce more viral particles. One well-studied class of bacterial viruses, the lysogenic bacteriophages, uses two kinds of replication mechanism to ensure viral propagation and transmission. As we noted in Chapter 14, after introduction of its DNA into a host cell, the phage has the potential to enter one of two pathways for propagation (Figure 20-20). Most of the time, viruses use the **lytic pathway**, in which phage DNA is immediately replicated and viral proteins

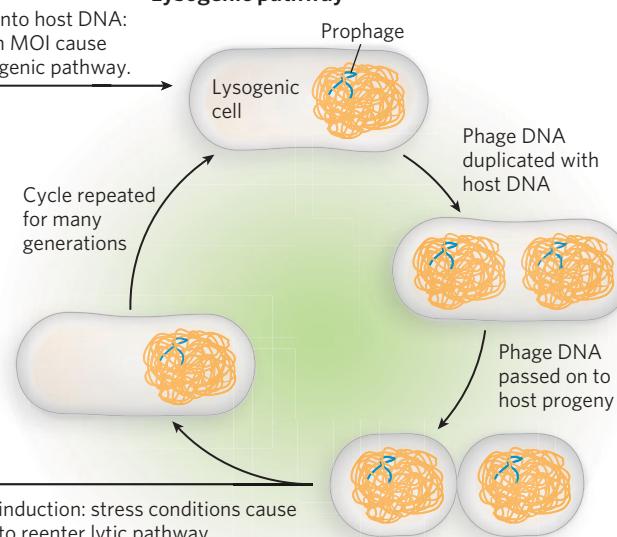
(a) Bacteriophage  $\lambda$



(b) Lytic pathway



Lysogenic pathway



**FIGURE 20-20** The growth and life cycle phases of bacteriophage  $\lambda$ . (a) Bacteriophage  $\lambda$  infecting a lawn of *E. coli*. The viruses eventually kill their host cells, leaving cleared spots, or plaques, in the bacterial lawn. (b) The lytic and lysogenic pathways. The decision to enter the lytic or

lysogenic pathway is based on cellular nutrients and the viral multiplicity of infection (MOI). Under conditions of cellular stress,  $\lambda$  phages can exit the lysogenic pathway and enter the lytic pathway, in the process of prophage induction. [Source: (a) Dr. Edward Chan/Visuals Unlimited.]

are synthesized to construct the viral coating around the DNA. Once many viral particles have been made, they burst from the cell, killing it, and enter the extracellular environment; they can now infect other cells. Occasionally, viruses enter the **lysogenic pathway**, whereby phage DNA integrates into the host cell chromosome and is replicated along with the chromosome as the cell divides. In this state, the bacteriophage is called a **prophage**, and the cell carrying it is a **lysogen**. Under normal conditions the prophage is stable within the lysogen, but if the host cell is stressed by DNA-damaging agents or other conditions that threaten survival, the prophage can rapidly excise from the chromosome and enter the lytic pathway. The switch from lysogenic to lytic growth is referred to as **prophage induction**.

Numerous bacteriophages are capable of this kind of genetic switch, but the underlying mechanisms are best understood for  $\lambda$  phage, a virus of *E. coli* that consists of a double-stranded linear DNA encapsulated in a head structure and attached tail region made of virally encoded proteins. Bacteriophage  $\lambda$  has been used extensively as a model system for studying integrated gene regulatory networks, and has also provided many tools useful in molecular biology research. For example, engineered versions of  $\lambda$  phage have been co-opted to transfer and express genes in *E. coli* and to inactivate host genes.

Two regulatory proteins, the  $\lambda$  repressor (also called cI) and Cro, govern the growth pathway of the virus. When the  $\lambda$  repressor protein is predominant,  $\lambda$  phage enters the lysogenic pathway; its DNA integrates into the chromosome, and only the  $\lambda$  repressor itself is expressed. When Cro predominates, however,  $\lambda$  phage enters the lytic pathway; most of the  $\lambda$  genes are expressed, and viral replication and packaging ensue. Eventually, the host cell is broken open by cell **lysis**, a process that releases the progeny phages.

The decision between lytic and lysogenic growth occurs early in a  $\lambda$  phage infection and depends largely on two other  $\lambda$  proteins: cII and cIII. The cII protein is a transcription activator that enhances transcription of the  $\lambda$  repressor gene and hence stimulates production of the repressor protein (cI). When cII is abundant and active, infection proceeds through the lysogenic path-

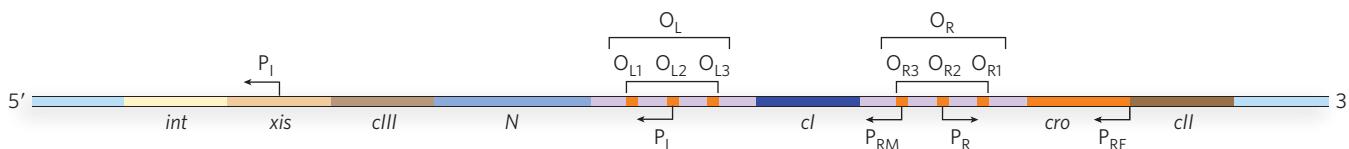
way. But because cII is susceptible to degradation by bacterial proteases, it often gets destroyed before it can trigger substantial production of the  $\lambda$  repressor. This tends to occur under nutrient-rich conditions, when proteases are present in high concentrations. It makes sense for the phage to enter the lytic pathway, triggered by low cII levels, because plenty of materials are available for viral reproduction and particle assembly.

The cIII protein stabilizes cII, probably by acting as a decoy, or alternative substrate, for protease molecules that would otherwise degrade cII. In this way, elevated cIII levels can trigger the switch to lysogen formation by enhancing cII concentrations, which increases production of the  $\lambda$  repressor.

Usually,  $\lambda$  phage infection leads to lytic growth. In addition to nutrient levels, the **multiplicity of infection (MOI)** also affects this growth pathway. This is because infections typically occur at a low MOI, in which there are many more host cells than viral particles. In this situation, it is advantageous for viruses to grow lytically so that more progeny can be made and released to infect the available host cells. However, once there is an abundance of viral particles relative to host cells, multiple viruses begin to infect each host. In this circumstance, the high MOI leads to more frequent lysogen formation. The virus propagates silently within the chromosome and awaits future opportunities to enter the lytic pathway when host cells are again abundant. The mechanisms underlying this fascinating genetic switch have been elucidated over many years, using numerous genetic and biochemical methods.

### Differential Activation of Promoters Regulates Bacteriophage $\lambda$ Infection

Most of the ~50 genes of the  $\lambda$  phage genome are required for viral replication and packaging; a relatively small region of the genome is critical for the gene regulation necessary to induce lytic versus lysogenic growth (Figure 20-21). On initial infection of a host cell by  $\lambda$  DNA, viral transcription begins at two opposing promoters,  $P_L$  (leftward promoter) and  $P_R$  (rightward promoter), to produce the “immediate early” transcripts.

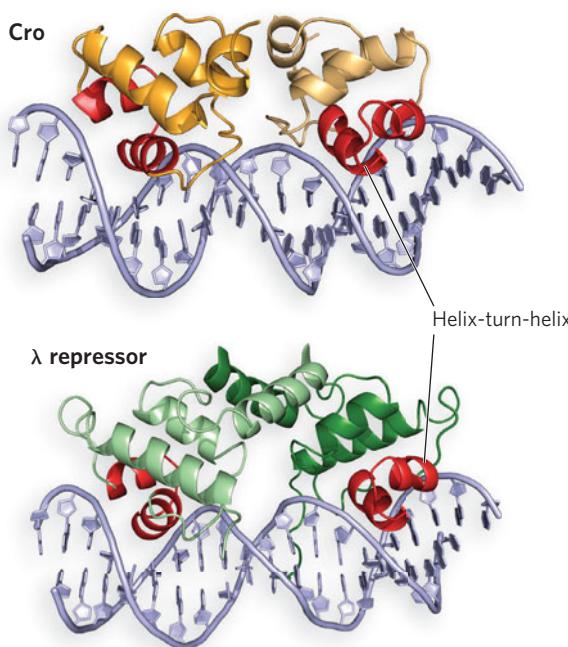


**FIGURE 20-21 A partial map of bacteriophage  $\lambda$ .** The genes and regulatory sites involved in establishing the lytic and lysogenic pathways are shown. Transcription begins at two opposing promoters,  $P_L$  and  $P_R$ , to produce the “immediate early” transcripts, which encode the N and Cro

proteins. Two weak promoters,  $P_{RM}$  and  $P_{RE}$ , drive transcription of the  $cl$  gene, which encodes the  $\lambda$  repressor. Overlapping the  $P_L$ ,  $P_{RM}$ , and  $P_R$  promoters are two operator sites,  $O_L$  and  $O_R$ , each containing three binding sites for the  $\lambda$  repressor and Cro.

This leads to production of two proteins, N and Cro, which begin to increase in concentration. Two additional promoters,  $P_{RM}$  (promoter for repressor maintenance) and  $P_{RE}$  (promoter for repressor establishment), drive transcription of the  $cI$  gene, which encodes the  $\lambda$  repressor. However, unlike the strongly constitutive  $P_L$  and  $P_R$ ,  $P_{RM}$  and  $P_{RE}$  are weak promoters and require activators to recruit RNA polymerase. Thus, at first, the  $cI$  gene is not transcribed and no  $\lambda$  repressor is produced.

Overlapping the  $P_L$ ,  $P_{RM}$ , and  $P_R$  promoters are two operator sites,  $O_L$  and  $O_R$ , each of which contains three binding sites for the  $\lambda$  repressor and Cro. Like the Lac repressor, both the  $\lambda$  repressor and Cro form homodimers in which two DNA-binding domains in the protein recognize DNA through a helix-turn-helix motif. In x-ray crystallographic structures of these proteins bound to DNA, researchers observed that each DNA-binding domain in the dimer binds to half of the 17 bp inverted repeat of the operator sequence (Figure 20-22). Each of the six operator binding sites, three sites in  $O_R$  and three in  $O_L$  (see Figure 20-21), can bind a  $\lambda$  repressor dimer or a Cro dimer. However, these sites have different affinities for the regulatory proteins, so they are not all occupied at once, or at



**FIGURE 20-22** The structures of Cro and  $\lambda$  repressor proteins. Both Cro and  $\lambda$  repressor form homodimers that bind the  $O_L$  and  $O_R$  operator regions. Both use helix-turn-helix motifs to associate with the DNA. [Sources: PDB ID 3CRO and PDB ID 1LMB.]

random. In addition, operator-protein binding at one site influences binding at the other sites, an example of cooperative interaction.

During initial viral infection, once appreciable levels of Cro are expressed, this protein binds the  $O_{R3}$  site; because  $O_{R3}$  overlaps  $P_{RM}$ , Cro blocks the access of RNA polymerase to  $P_{RM}$ , and the  $cI$  gene is not transcribed. Cro does not bind as well to  $O_{R1}$  and  $O_{R2}$ , and thus these operator sites are not occupied by Cro. If this situation continues, lytic growth ensues (Figure 20-23a).

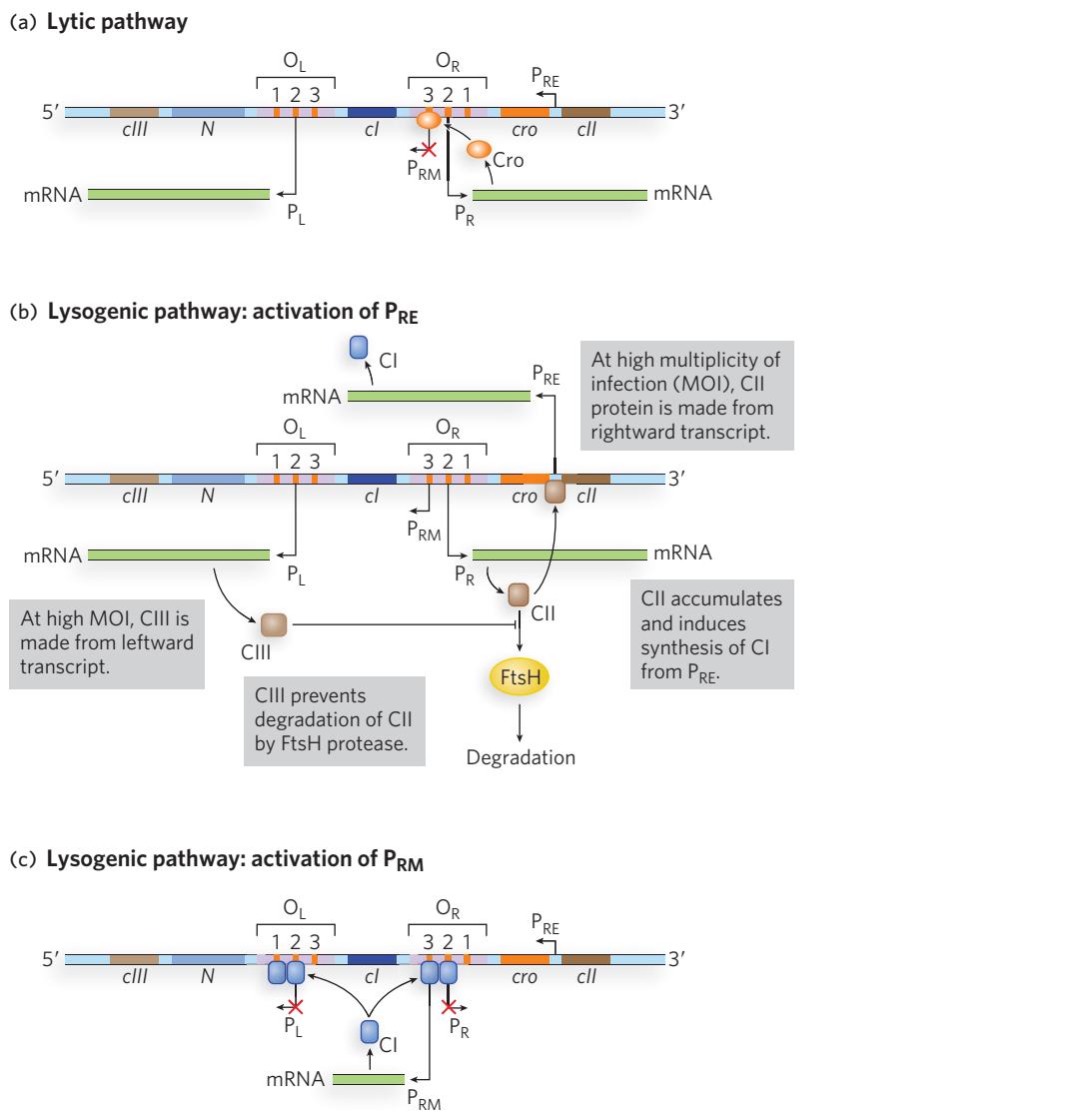
For lysogenic growth,  $P_{RE}$  is required (Figure 20-23b). Initial transcription of the  $cI$  gene requires activation of  $P_{RE}$  by cII, another gene product expressed early in the infection process. The cII protein can activate the weak  $P_{RE}$  promoter by binding to a site upstream from the transcription start site (the -35 region; see Chapter 15) and recruiting RNA polymerase to  $P_{RE}$ . However, cII is a substrate of the host cell protease FtsH, which cleaves and inactivates cII. At a low MOI, in which one viral particle at most has infected any one cell, the concentration of cII does not reach a level sufficient to activate  $P_{RE}$ . But under conditions of high MOI, in which multiple viral particles have infected the cell, more cII is produced because of the multiple copies of the  $cII$  gene present in the cell. Furthermore, the phage protein cIII (also expressed early in infection) helps stabilize cII by competing as a substrate for FtsH. In this situation, FtsH cannot keep up with cII production, and the increased levels of cII lead to activation of  $P_{RE}$ .

As a result of the activation of  $P_{RE}$ , the  $\lambda$  repressor protein is produced. The repressor activates  $P_{RM}$  by binding to  $O_{R1}$  and  $O_{R2}$ , leading to stable expression of the repressor (Figure 20-23c). The  $P_{RM}$  promoter is necessary to maintain ongoing production of the  $\lambda$  repressor because, when the repressor binds operator sites  $O_{R1}$  and  $O_{R2}$ ,  $P_R$  is silenced and production of cII stops, leading to loss of the activator (cII) for  $P_{RE}$ . Under these conditions, lysogenic growth is favored.

Note that if Cro were to bind  $O_{R3}$  before cI (the repressor) bound  $O_{R1}$  and  $O_{R2}$ , the phage would have trouble establishing lysogeny. This interference by Cro probably never occurs, because when cII is highly active, cI production is sufficient to shut off the  $cro$  gene before enough Cro is made to turn off  $P_{RM}$ .

### The $\lambda$ Repressor Functions as Both an Activator and a Repressor

The  $\lambda$  repressor is capable of complex regulation, in part because it is a two-domain protein that forms a functional dimer. The N-terminal region of the protein contains the helix-turn-helix DNA-binding motif, and



**FIGURE 20-23** The bacteriophage  $\lambda$  genetic switch between lytic and lysogenic growth. A small region of the  $\lambda$  phage genome controls the genetic switch between lysis and lysogeny. Two critical proteins, Cro and  $\lambda$  repressor (CI), regulate the switch. See text for details.

the C-terminal domain is responsible for dimerization. As we've seen, the  $\lambda$  repressor can bind to any of six operator binding sites in the  $O_L$  and  $O_R$  regions (see Figure 20-21). Despite its name, repressor bound at  $O_{R2}$  actually *activates* transcription from  $P_{RM}$ , and this activation is critical to the lysogenic switch. However, the repressor has highest binding affinity for  $O_{R1}$ , and because binding is cooperative,  $O_{R1}$  recognition increases the affinity of the bound  $\lambda$  repressor for  $O_{R2}$ . Repressor bound cooperatively at  $O_{R1}$  and  $O_{R2}$  blocks RNA polymerase binding at  $P_R$ , repressing transcription from that promoter. Similarly, repressor bound cooperatively at  $O_{L1}$  and  $O_{L2}$  prevents transcription from  $P_L$ . Thus, the repressor can simultaneously activate transcription of

its own gene from  $P_{RM}$  and repress transcription of the immediate early genes necessary for lytic growth from  $P_L$  and  $P_R$ .

Once the lytic-to-lysogenic switch has occurred, the phage DNA integrates into the host chromosome by a mechanism described below. The integrated viral DNA can be maintained and replicated stably as part of the host cell chromosome. But, as we've seen, when the host cell is exposed to agents or conditions that damage DNA, the prophage is rapidly excised, and lytic growth begins. This lysogenic-to-lytic switch comes about because the  $\lambda$  repressor protein resembles the bacterial protein LexA, which undergoes self-cleavage when stimulated by a second bacterial protein, RecA,

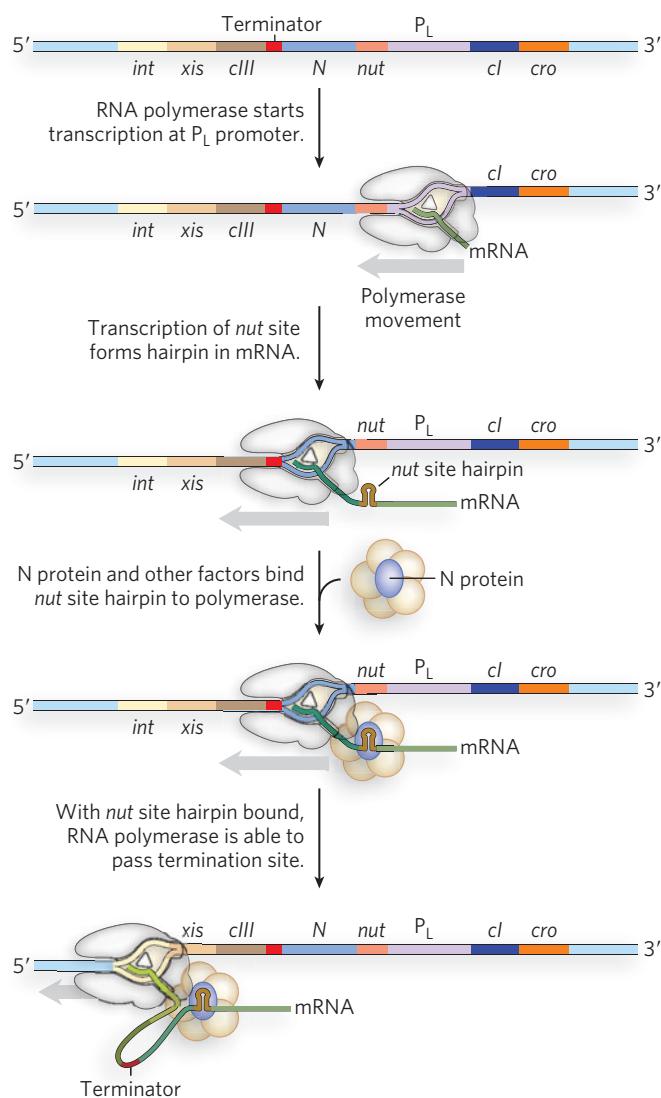
that is a sensor of DNA damage in bacteria (see Figure 20-13). When RecA is activated in response to DNA damage (the SOS response), the  $\lambda$  repressor protein, if present in the cell, also undergoes self-cleavage. This reaction clips off the C-terminal region of the repressor, removing its dimerization domain and thereby destroying its cooperative binding to  $O_L$  and  $O_R$ , sites 1 and 2. As a result, the cleaved  $\lambda$  repressor dissociates from the operator DNA, allowing transcription from  $P_R$  and  $P_L$ , and thus lytic growth.

### More Regulation Levels Are Invoked during the Bacteriophage $\lambda$ Life Cycle

Switching between the activation and repression of promoters enables  $\lambda$  phage to alternate between lytic and lysogenic growth, as dictated by environmental conditions. But additional levels of regulation further control the expression of genes required to establish and maintain early steps in the infection cycle. Two  $\lambda$  proteins, the N and Q proteins, are known as antiterminators because they prevent RNA polymerase from prematurely stopping transcription of the genes they regulate. N protein binds to specific regions of the viral transcript called *nut* (*N* utilization) sites (Figure 20-24). The resulting RNA-protein complexes assemble with additional proteins produced from the host cell genes *nusA*, *nusB*, *nusE*, and *nusG*. Although the functions of these proteins are not known in detail, they somehow help the phage-produced N protein bind to RNA polymerase and enhance its ability to bypass terminator structures downstream from the genes for N and Cro. In this way, elongation of the viral transcript is favored once initial infection is established.

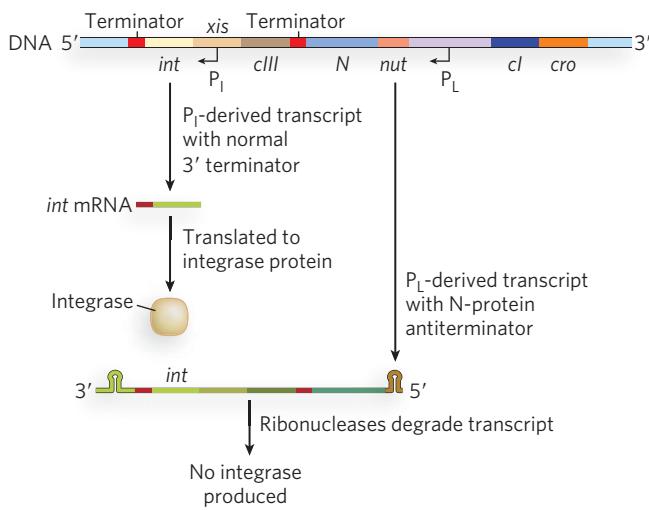
One target of the N protein antitermination mechanism is the viral gene encoding the Q antiterminator protein. Unlike N, Q binds DNA sequences between the -10 and -35 regions of  $P_R'$ , the promoter for genes that are active late in the infection process. In the absence of Q, RNA polymerase pauses shortly after initiation of the transcript, then falls off the template when encountering a terminator hairpin structure ~200 bp further on. When Q is present, it binds first in the promoter region, then transfers to the polymerase when the enzyme pauses after initiation. With Q bound, the polymerase can transcribe through the terminator hairpin, thereby producing full-length transcripts; the mechanism of this antitermination process is not yet clear.

To establish a lysogenic infection, the virus must produce an integrase, an enzyme responsible for integrating phage DNA into the host cell chromosome (see Chapter 14). The *int* gene, which encodes the integrase, is transcribed from two promoters,  $P_L$  and  $P_I$ . The cII protein activates  $P_L$  in addition to  $P_{RE}$  (as described



**FIGURE 20-24** **N** protein-mediated antitermination of transcription. N protein is transcribed early in the infection process from the  $P_L$  promoter, but early transcription stops at the terminator just downstream from the *N* gene. N protein binds newly transcribed *nut* sites, and through interaction with several other factors, causes a change in RNA polymerase that allows it to read through terminator structures, generating longer mRNA transcripts.

above); thus, when cII is abundant and the lysogenic state is favored, both repressor and integrase are expressed. Although the *int* gene is also transcribed from  $P_L$ , the two different transcripts have inherently different stabilities due to different RNA structures at their 3' ends. The  $P_I$ -derived transcript ends at a terminator hairpin structure downstream from the integrase-coding sequence, whereas the  $P_L$ -derived transcript extends past the terminator—because the transcript is made by a polymerase modified by the N protein (see



**FIGURE 20-25 Regulation of *int* gene expression.**

Transcripts of the integrase gene (*int*) are generated from two promoters,  $P_L$  and  $P_I$ . The  $P_I$ -derived transcript is short, ending at a terminator structure that follows the coding sequence of the *int* gene.  $P_L$ -derived transcripts are longer and are generated as a result of N-protein antitermination. The 3' end of the  $P_L$ -derived transcript extends beyond the terminator. A structure at the 3' end makes the transcript a substrate for cellular ribonucleases, and the transcript is degraded. Thus, more integrase protein is generated from the  $P_I$ -derived transcript than from the  $P_L$ -derived transcript.

Figure 20-24). This longer transcript forms an alternative hairpin structure that is a substrate for cellular ribonucleases. As a result, only the  $P_I$ -derived transcript is maintained, and it can be translated into integrase protein (Figure 20-25).

In a nifty twist on this regulatory mechanism, the  $\lambda$  DNA integrates into the host chromosome such that a small portion of the phage DNA is removed—the portion that, when transcribed, becomes the destabilizing RNA sequence at the end of the  $P_L$ -derived integrase transcript. Thus, when  $\lambda$  phage is in the lysogenic state and integrated into the host chromosome, integrase can be produced from transcripts derived from either  $P_I$  or  $P_L$ . Because integrase is also required to excise the prophage from the host genome, the ability to use either type of *int* gene transcript ensures that this step is not dependent on cII protein concentration.

## SECTION 20.3 SUMMARY

- Bacteriophage gene activities can be controlled by regulatory networks that integrate multiple signals into a common gene regulatory response.

- Infection of a bacterial cell by  $\lambda$  phage can lead to either (1) a lytic pathway in which host cells are lysed as new viral particles are assembled and released, or (2) a lysogenic pathway in which the viral DNA integrates into the host chromosome and is propagated through chromosome replication and cell division, without immediate production and release of new viruses.
- In  $\lambda$  phage, two regulators, Cro and the  $\lambda$  repressor, control whether the phage propagates through the lytic pathway (when Cro dominates) or the lysogenic pathway (when the  $\lambda$  repressor dominates). Both repressors bind to similar phage DNA sequences, but with different binding affinities. Cro blocks  $\lambda$  repressor synthesis while allowing expression of other genes needed for lytic growth; the  $\lambda$  repressor blocks transcription of all phage genes except its own.
- In addition to Cro, the N protein, expressed early in the infection process, favors  $\lambda$  gene expression by binding *nut* sites in the viral transcripts. The resulting RNA-protein complexes favor transcription by assembling with host proteins that aid binding of N protein to RNA polymerase, enhancing the enzyme's ability to bypass terminator structures downstream from the N and Cro genes.
- N protein-mediated antitermination favors production of the Q protein, another antiterminator that binds DNA sites near the promoter of genes expressed later in the infection cycle. Binding of the Q protein allows it to transfer to paused RNA polymerase molecules and enhance their ability to traverse terminator structures in the regulated genes.
- To establish a lysogenic infection, the virus produces an integrase that integrates the phage DNA into the host chromosome. Integrase is also required to excise the prophage from the host genome. Both processes are highly regulated in response to environmental conditions.

## Unanswered Questions

The study of gene regulation continues to expand as new regulatory mechanisms are uncovered. RNA has emerged as a major player in controlling levels of protein production, and many fascinating aspects of its involvement remain to be deciphered.

- How widespread are RNA-based gene regulatory mechanisms?** Early work on the regulation of gene expression focused on proteins with the sole role of controlling when and how much of a protein is

synthesized. More recent research has revealed an increasing number of instances in which RNA plays this regulatory role. In addition to the use of riboswitches, bacteria encode many small noncoding RNAs that may be important for modulating gene expression. Small RNAs are now known to be extensively involved in regulating gene expression in plants and animals (see Chapter 22), and whether bacteria use their small RNAs in analogous ways is currently unknown and will be a fascinating area of research.

2. **How does a bacteriophage compete for a host cell's gene expression machinery?** Studies of bacteriophage  $\lambda$  have provided an exciting introduction to the world of phage-host cell competition. But there are many different mechanisms by which viruses might take over a host cell's gene expression machinery and thus regulate the propagation of new viral particles. Some researchers estimate that Earth is home to more phage particles than cells! Bacteriophages thus provide an enormous pool of genes that are readily exchanged and introduced into new hosts, driving the evolution of new traits and viral defense

mechanisms. A broader understanding of gene regulatory pathways in phages will offer new insights into bacterial gene regulation, and perhaps into the relationships between viral propagation and gene transfer.

### 3. How are gene regulatory networks integrated?

Much of the research on gene regulation in bacteria has focused on individual genes or operons, giving the impression that just one or a few changes occur in response to signaling molecules. Studies using DNA microarrays, however, show that changes in gene expression in response to stresses or altered nutrient levels occur in hundreds of different genes. How these changes are coordinated and how different gene regulatory pathways integrate multiple signals at once are the subjects of active research. Investigators are using traditional genetic and biochemical methods, as well as more recently developed approaches such as microarray analysis and bioinformatics. This area of research is referred to as systems biology, to indicate that gene expression operates not in isolation but as part of a system, as defined by the cell or organism.

# How We Know

## TRAPped RNA Inhibits Expression of Tryptophan Biosynthetic Genes in *Bacillus subtilis*

**Babitzke, P., and P. Gollnick. 2001.** Posttranscription initiation control of tryptophan metabolism in *Bacillus subtilis* by the *trp* RNA-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. *J. Bacteriol.* 183:5795–5802.

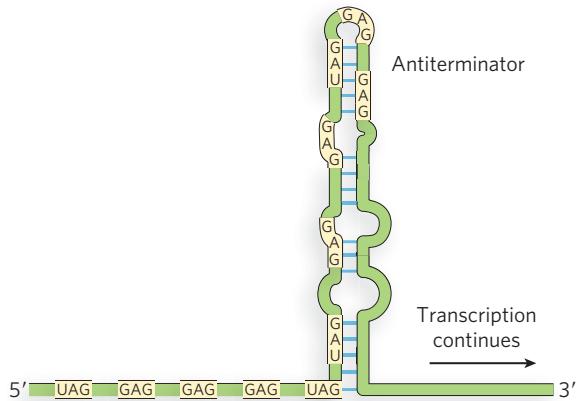
The TRAP system of tryptophan regulation in *B. subtilis* beautifully illustrates how the mechanistic details of multi-level gene regulation were eventually worked out using a combination of experimental methods. As in other bacteria, the *trp* genes of *B. subtilis*, contained in the *trpEDCFBA* operon (each letter after *trp* standing for a gene in the operon), are transcribed only when tryptophan is in short supply in the cell. Genetic experiments revealed that transcription of the *trpEDCFBA* operon requires a regulatory protein called TRAP (*trp* RNA-binding attenuation protein). Using purified TRAP, RNA polymerase, nucleoside triphosphates, and a plasmid DNA with an inserted *trpEDCFBA* operon, researchers found that adding L-tryptophan to the mix caused transcription to stop in the leader sequence of the operon, upstream from the coding sequences. In the presence of L-tryptophan, TRAP bound the newly synthesized leader RNA and prevented formation of an antiterminator structure (Figure 1a). As a result, a competing RNA structure—a terminator—could form, blocking passage of the polymerase and causing premature transcription termination (see Chapter 15), as shown in Figure 1b.

How does TRAP respond to L-tryptophan? TRAP is composed of 11 subunits, each of 6 to 8 kDa. TRAP binds to 11 triplet repeats, primarily GAG and UAG, in the leader sequence that are separated from each other by two or three nonconserved nucleotides. The crystal structure of TRAP bound to a 53-nucleotide single-stranded *trp* leader RNA looks like a molecular spool in which the RNA wraps around the outer surface of the protein core (see Figure 1b). Each GAG or UAG triplet tucks into a binding pocket formed by one of the TRAP subunits, and Trp residues are positioned between the subunits, where they presumably stabilize interactions required for high-affinity protein-RNA binding.

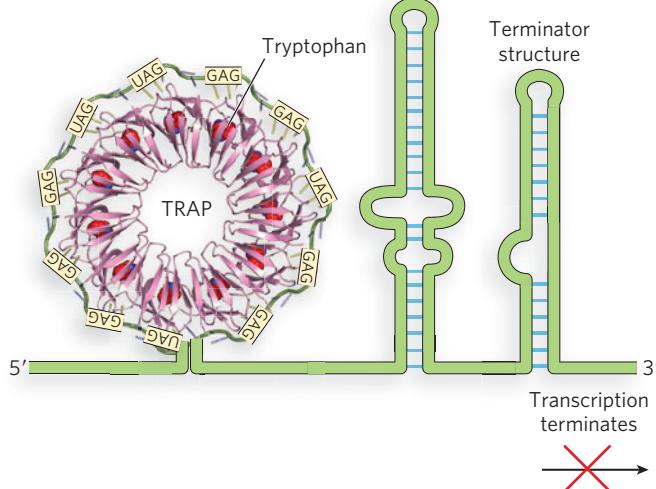
Subsequent experiments showed that TRAP can also bind its target sequence in the leader region of mature mRNAs, blocking access of the ribosome to the Shine-Dalgarno sequence, preventing efficient translation initiation. Furthermore, a second protein, called anti-TRAP (AT), induced by uncharged tRNA<sup>Trp</sup>, can bind TRAP and prevent its binding to the *trp* leader RNA, allowing transcription of the *trp* operon to proceed. Through TRAP and anti-TRAP, *B. subtilis* senses the levels of both tryptophan and uncharged tRNA<sup>Trp</sup> in order to regulate

tryptophan biosynthesis by changing the accessibility of the RNA to both RNA polymerase and the ribosome.

(a) Tryptophan absent



(b) Tryptophan present



**FIGURE 1** Regulation of the *B. subtilis* *trp* operon. (a) In the absence of tryptophan, the structure of the *trp* leader RNA allows continued transcription of the *trp* operon. (b) In the presence of tryptophan, the TRAP protein binds tryptophan and associates with the *trp* operon leader through interaction with 11 GAG and UAG triplets (yellow). This leads to formation of a terminator structure in the RNA, which halts transcription. [Sources: (a) Adapted from P. Babitzke and P. Gollnick, *J. Bacteriol.* 183:5795–5802, 2001, Fig. 2. (b) PDB ID 1C9S.]

## Autoinducer Analysis Reveals Possibilities for Blocking Cholera Infection

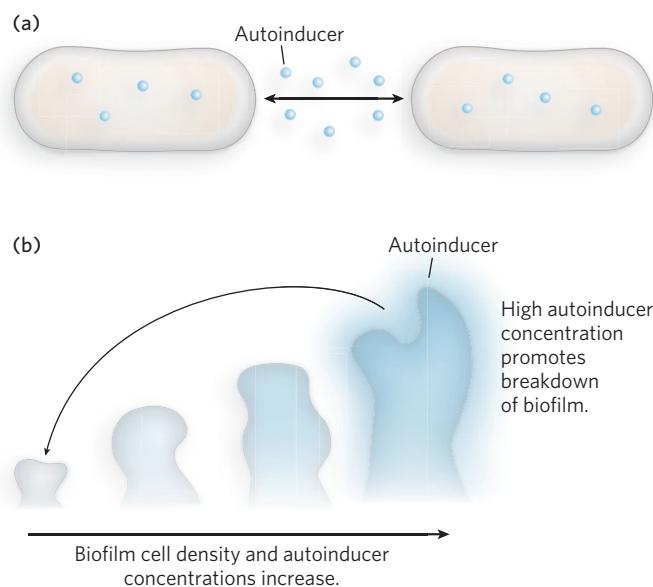
Higgins, D.A., M.E. Pomianek, C.M. Kraml, R.K. Taylor, M.F. Semmelhack, and B.L. Bassler. 2007. The major *Vibrio cholerae* autoinducer and its role in virulence factor production. *Nature* 450:883–886.

Many single-celled organisms and multicellular tissues use cell-to-cell signaling to communicate information about population density, allowing cells to change gene expression levels in response to the group environment. In bacteria, such signaling molecules are secreted from one cell and detected or imported by neighboring cells, where the signal triggers complex changes in gene expression (Figure 2a). The ability to sense and respond to high population density, a type of cell-to-cell signaling referred to as quorum sensing, frequently involves secreted peptides known as autoinducers. In the pathogenic bacterium *Vibrio cholerae*, which causes cholera, autoinducers terminate rather than promote virulence, and activation of quorum sensing by introducing an autoinducer could form the basis of a treatment for cholera.

Bonnie Bassler's laboratory at Princeton has investigated the molecular details of quorum sensing in *V. cholerae*. The bacterium uses quorum sensing to control the production of virulence factors and its ability to grow as a biofilm, a contiguous layer of cells. At low cell density, *V. cholerae* expresses virulence factors and forms a biofilm. As cell density increases, two autoinducers increase in concentration until they repress both virulence factor expression and biofilm formation (Figure 2b). Synthesized by autoinducer synthase enzymes, the small-molecule autoinducers are secreted from the bacterial cells and bind to receptors for import into neighboring cells.

Information from both types of autoinducers is transmitted through the protein LuxO, which in turn controls the level of HapR, a transcription factor that controls the expression of many other genes. At low cell density, in the absence of autoinducers, HapR is not produced, virulence factors are expressed, and biofilms form. Because HapR is required for the expression of genes that cause the cells to luminesce, the cells are not bioluminescent. This provides an easy way for experimentalists to determine whether HapR is turned on. At high cell density, autoinducers increase and bind to LuxO, leading to the production of HapR. HapR represses the genes for virulence factor production and biofilm formation, while activating the expression of the bioluminescence genes. The end result is that the production of autoinducers and quorum-sensing activation terminate virulence in *V. cholerae* cells growing in crowded conditions.

Bassler and colleagues cloned the autoinducer synthase genes and introduced them into *E. coli*. Because *E. coli* is not sensitive to the *V. cholerae* autoinducers, the signaling molecules were produced and secreted in large amounts without interfering with cell growth. The experimenters could then purify the autoinducers and determine their chemical structure by nuclear magnetic resonance spectroscopy (NMR). The NMR method enables exact determination of the chemical structure of the autoinducers, because it provides information about the chemical environment of each proton in the molecule. Bassler's group has also developed a method for chemically synthesizing these autoinducers—an exciting development that might allow researchers to "trick" cells into the quorum-sensing response even at low cell density. This could provide a clever way to expropriate the biology of gene regulation to prevent cholera infection, a strategy that could also be developed for other kinds of pathogenic bacteria.



**FIGURE 2** (a) In quorum sensing, bacterial cells sense and respond to high population density by sending and receiving small-molecule signals called autoinducers, which freely diffuse through the bacterial membranes. (b) In *V. cholerae*, autoinducers stimulate the breakdown of biofilms, reducing biofilm size as autoinducer concentration increases.

## Key Terms

inducer, p. 701	SOS response, p. 710	lysogenic pathway, p. 722
catabolite repression, p. 703	quorum sensing, p. 711	prophage, p. 722
corepressor, p. 709	riboswitch, p. 712	lysogen, p. 722
transcription attenuation, p. 709	translational repressor, p. 716	prophage induction, p. 722
leader sequence, p. 709	stringent response, p. 719	lysis, p. 722
terminator, p. 709	stringent factor, p. 719	multiplicity of infection (MOI),
leader peptide, p. 710	lytic pathway, p. 721	p. 722

## Problems

1. A researcher engineers a *lac* operon on a plasmid, but inactivates all parts of the Lac operator (*lacO*) and the Lac promoter, replacing them with the binding site for the LexA repressor (which acts in the SOS response) and a promoter regulated by LexA. The plasmid is introduced into *E. coli* cells that have a *lac* operon with an inactive *lacZ* gene. Under what conditions will these transformed cells produce β-galactosidase?
2. Describe the probable effects on the expression of *lac* genes of mutations that (a) relocate the Lac operator so that it is on the other side of the operon, (b) inactivate the binding site for CRP, and (c) alter the promoter sequence around position -10.
3. In the *ara* operon, the AraC protein can act as either an activator or a repressor. If AraC remains bound to the DNA in the absence of arabinose, why does the protein not always function as an activator?
4. *E. coli* cells are growing in a medium with glucose as the sole carbon source. Tryptophan is suddenly added. The cells continue to grow, and they divide every 30 minutes. Describe (qualitatively) how the levels of tryptophan synthase (an enzyme produced by the *trp* operon) change with time under the following conditions:
- (a) The *trp* mRNA is stable (degraded slowly over many hours).
  - (b) The *trp* mRNA is degraded rapidly, but tryptophan synthase is stable.
  - (c) The *trp* mRNA and tryptophan synthase are both degraded rapidly.
5. How would the SOS response in *E. coli* be affected by mutations in the *lexA* gene that (a) prevented autocatalytic cleavage of the LexA protein or (b) weakened the interaction of LexA with its normal binding site?
6. A typical bacterial repressor protein discriminates between its specific DNA binding site (operator) and nonspecific DNA by a factor of  $10^4$  to  $10^6$ . About 10 molecules of repressor per cell are sufficient to ensure a high level of repression. Assume that a very similar repressor existed in a human cell, with a similar specificity for its binding site. How many copies of the repressor would be required to elicit a level of repression similar to that in the bacterial cell? (Hint: The *E. coli* genome contains about  $4.6 \times 10^6$  bp; the human haploid genome has about  $3.2 \times 10^9$  bp.)
7. The dissociation constant for a particular repressor-operator complex is very low, about  $10^{-13}$  M. An *E. coli* cell (volume  $2 \times 10^{-12}$  mL) contains 10 copies of the repressor. Calculate the cellular concentration of the repressor protein. How does this value compare with the dissociation constant of the repressor-operator complex? What is the significance of this answer?
8. *E. coli* cells are growing in a medium containing lactose but no glucose. Indicate whether each of the following changes or conditions would increase, decrease, or not change expression of the *lac* operon. It may be helpful to draw a model depicting what is happening in each situation.
- (a) Addition of a high concentration of glucose
  - (b) A mutation that prevents Lac repressor binding to the operator
  - (c) A mutation that completely inactivates β-galactosidase
  - (d) A mutation that completely inactivates galactoside permease
  - (e) A mutation that prevents binding of CRP to its binding site near the Lac promoter
9. How would transcription of the *E. coli trp* operon be affected by the following manipulations of the leader region of the *trp* mRNA?
- (a) Increasing the distance (number of bases) between the leader peptide gene and sequence 2
  - (b) Increasing the distance between sequences 2 and 3
  - (c) Removing sequence 4
  - (d) Changing the two Trp codons in the leader peptide gene to His codons
  - (e) Eliminating the ribosome-binding site for the gene that encodes the leader peptide
  - (f) Changing several nucleotides in sequence 3 so that it can base-pair with sequence 4 but not with sequence 2

- 10.** Many riboswitches have been characterized in bacteria, including one that binds to thiamine pyrophosphate (TPP) and another that binds to glucosamine 6-phosphate. Compare and contrast the mechanisms by which these two riboswitches inhibit translation of their RNAs.
- 11.** A mutation is found in the gene encoding the translational repressor of an r-protein operon. The mutation increases the affinity of the repressor protein for mRNA and decreases its affinity for rRNA. What is the likely effect of such a mutation?
- 12.** A bacteriophage  $\lambda$  lysogen (an *E. coli* cell with a  $\lambda$  prophage integrated into its genome) is largely immune to lysis by  $\lambda$  phages introduced into the cell later. Explain.
- 13.** Mutant versions of CRP have been isolated that bind DNA normally but do not activate transcription. What does the existence of these mutants indicate about the mechanism of CRP-mediated transcription activation,
- and what phenotype would you expect to observe for cells expressing one of these mutant CRP alleles?
- 14.** How does the organization of related genes into operons enable the coordinated production of proteins? Suggest a way that would allow different genes in an operon to be expressed at different levels.
- 15.** Name one advantage and one disadvantage to using RNA structures such as riboswitches to regulate gene expression in response to small-molecule effectors.
- 16.** What general principle of gene regulation is illustrated by transcription attenuation?
- 17.** Would you expect the mechanism of transcription attenuation described for the *trp* operon to function similarly in eukaryotic cells?
- 18.** Name three properties of riboswitch-regulated mRNAs.

## Data Analysis Problem

### Oxender, D.L., G. Zurawski, and C. Yanofsky. 1979.

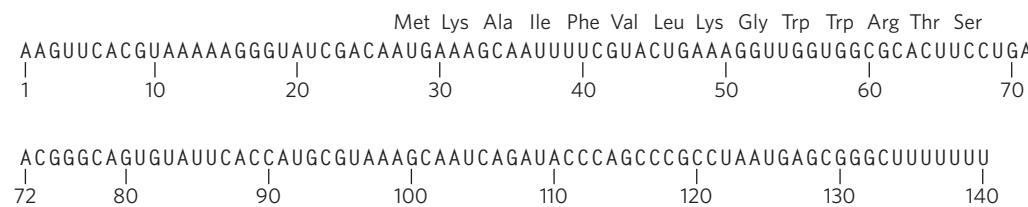
Attenuation in the *Escherichia coli* tryptophan operon—role of RNA secondary structure involving the tryptophan codon region. *Proc. Natl. Acad. Sci USA* 76:5524–5528.

**19.** Unraveling of the complicated regulatory mechanism of the *trp* operon proceeded in stages over more than a decade. The work was carried out mainly in the lab of Charles Yanofsky at Stanford University. Beginning with a study of tryptophan synthase in the early 1970s, Yanofsky's group gradually focused on regulation of the *trp* operon. They initially thought that regulation could be explained in terms of a *lac* operon-like repressor-operator interaction. However, they found a long leader region between the promoter and the first *trp* operon gene. Certain deletions in that leader region increased the expression of tryptophan synthase even when the repressor was present. The leader region was sequenced (laboriously, using the methods in existence before invention of the Sanger method), revealing a short open reading frame. When RNA was labeled in the cell, the researchers detected a large amount of truncated *trp* mRNA, terminated before the *trp* genes were transcribed.

The sequence of this truncated mRNA is shown in Figure 1. The plot thickened.

In the study published in 1979, Yanofsky and coworkers explored the secondary structure in the leader region of the *trp* mRNA, in work that defined the outlines of the overall regulatory system. They used the enzyme RNase T1, a ribonuclease that cuts single-stranded RNA (unpaired linear regions or the unpaired loop ends of hairpins) much faster than double-stranded RNA. A partial digest was carried out on the labeled, attenuated RNA (so that the RNAs in the population were not cut at every single-stranded region). This produced the RNA species shown in the polyacrylamide gel (run under nondenaturing conditions) in Figure 2. The RNAs in the three bands, A, B, and C, were separately isolated. When these were run on denaturing gels that would separate any paired RNAs, the A and C bands separated into multiple species (Figure 3).

(a) The band A RNA (Figure 2) is approximately 140 nucleotides long. What species is this, and what can be said about the three bands that appear when it is separated on the denaturing gel (Figure 3)?



**FIGURE 1**

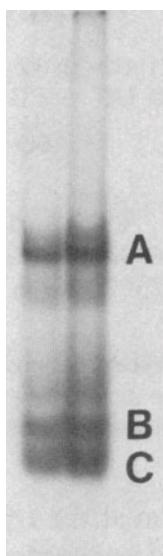


FIGURE 2

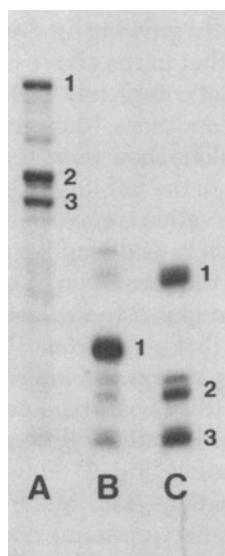


FIGURE 3

**(b)** Bands B and C are smaller than band A. What does this tell you?

**(c)** The band C RNA also separates into three species on the denaturing gel, but the band B RNA does not. What can you conclude from this?

The authors went on to identify the RNA species on the gels. Band B consisted of sequences from approximately position 108 to the 3' end of the RNA. Band C1 was the segment from approximately position 51 to position 95, with C2 and C3 arising from a cleavage at around position 70.

**(d)** From what you know about the attenuation mechanism, what regions of the RNA are present in band B?

**(e)** What regions are present in band C?

**(f)** Are there any missing elements of secondary structure that are important to the regulatory mechanism? If so, what are they, and why might they be missing from the isolated RNA?

## Additional Reading

### General

- Bassler, B.L., and R. Losick. 2006.** Bacterially speaking. *Cell* 125:237–246.
- Fang, F.C. 2005.** Sigma cascades in prokaryotic regulatory networks. *Proc. Natl. Acad. Sci. USA* 102:4933–4934.
- Gollnick, P., P. Babitzke, A. Antson, and C. Yanofsky. 2005.** Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu. Rev. Genet.* 39:47–68.
- Lewis, M. 2005.** The lac repressor. *Crit. Rev. Biol.* 328:521–548.
- von Hippel, P.H. 2007.** From “simple” DNA-protein interactions to the macromolecular machines of gene expression. *Annu. Rev. Biophys. Biomol. Struct.* 36:79–105.
- Wakeman, C.A., W.C. Winkler, and C.E. Dann 3rd. 2007.** Structural features of metabolite-sensing riboswitches. *Trends Biochem. Sci.* 32:415–424.

### Transcriptional Regulation

- Jacob, F., and J. Monod. 1961.** Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.
- Kolb, A., S. Busby, H. Buc, S. Garges, and S. Adhya. 1993.** Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.* 62:749–795.
- Lawson, C.L., D. Swigon, K.S. Murakami, S.A. Darst, H.M. Berman, and R.H. Ebright. 2004.** Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* 14:10–20.
- Yanofsky, C., K.V. Konan, and J.P. Sarsero. 1996.** Some novel transcription attenuation mechanisms used by bacteria. *Biochimie* 78:1017–1024.

### Beyond Transcription: Control of Other Steps in the Gene Expression Pathway

- Barrick, J.E., and R.R. Breaker. 2007.** The power of riboswitches: Discovering relics from a lost world run by RNA molecules may lead to modern tools for fighting disease. *Sci. Am.* 296(1):50–57.
- Coppins, R.I., K.B. Hall, and E.A. Groisman. 2007.** The intricate world of riboswitches. *Curr. Opin. Microbiol.* 10:176–181.
- Gao, R., and A.M. Stock. 2010.** Molecular strategies for phosphorylation-mediated regulation of response regulator activity. *Curr. Opin. Microbiol.* 13:160–167.
- Magasanik, B. 2000.** Global regulation of gene expression. *Proc. Natl. Acad. Sci. USA* 97:14,044–14,045.
- Shapiro, L., H.H. McAdams, and R. Losick. 2009.** Why and how bacteria localize proteins. *Science* 326:1225–1228.

### Control of Gene Expression in Bacteriophages

- Gottesmann, M., and R. Weisberg. 2004.** Little lambda, who made thee? *Microbiol. Mol. Biol. Rev.* 68:796–813.
- Hochschild, A. 2002.** The switch: *cI* closes the gap in auto-regulation. *Curr. Biol.* 12:R87–89.
- Murray, N.E., and A. Gann. 2007.** What has phage lambda ever done for us? *Curr. Biol.* 17:R305–312.
- Oppenheim, A.B., O. Oren Kobiler, J. Stavans, D.J. Court, and S. Adhya. 2005.** Switches in bacteriophage lambda development. *Annu. Rev. Genet.* 39:409–429.
- Ptashne, M. 1992.** *A Genetic Switch*. Cambridge, MA: Cell Press.

# The Transcriptional Regulation of Gene Expression in Eukaryotes



**Tracy Johnson** [Source: Courtesy of Tracy Johnson.]

mutations in a second gene that would cause cell death, an effect referred to as synthetic lethal. For a long time we found absolutely nothing interesting. But we kept working, and at last discovered that a deletion of the gene encoding the Gcn5 protein was synthetic lethal when combined with mutations in genes encoding parts of the U2 snRNP component of spliceosomes. Gcn5 is a histone acetyltransferase (HAT), a well-characterized enzyme that adds acetyl groups to histone proteins within nucleosomes, but it had no known connection to pre-mRNA splicing.

The real moment of surprise came when we found that when the GCNS gene is deleted, cotranscriptional splicing is completely messed up! This is because the U2 snRNP is no longer recruited to pre-mRNA splice sites. The splicing defect is specific to this HAT and requires the enzyme's catalytic activity, which is targeted toward promoter-bound histones. I never imagined there would be a link between chromatin structure and pre-mRNA splicing, which is implied by this finding. We envision that a specific pattern of histone acetylation leads to physical recruitment of proteins to acetylated histones within chromatin, which in turn recruits spliceosomes to newly transcribed pre-mRNAs. Because the HAT is very well conserved in mammals, it could be a general mechanism that affects which splice sites are chosen in pre-mRNAs, depending on the acetylation state of the histones associated with the parent gene.

—Tracy Johnson, on discovering that pre-mRNA splicing requires specific histone acetylation

- 21.1 Basic Mechanisms of Eukaryotic Transcriptional Activation 734**
- 21.2 Combinatorial Control of Gene Expression 743**
- 21.3 Transcriptional Regulation Mechanisms Unique to Eukaryotes 751**

Eukaryotic cells, like bacteria, express only a subset of their genes at any given time. We learned in Chapter 20 that through gene regulation, bacteria are able to adapt to environmental changes and respond to signaling molecules and viral assaults. Eukaryotes, too, must respond to their environment and external stimuli. But in addition, multicellular eukaryotes must manage complex pathways of cell division and differentiation that give rise to the multitude of cell types required for organismal development. Developmental programs are extremely precise—it is critical that each protein influencing cellular differentiation is active at the right time and in the right place—and any deviation from the program can have drastic consequences. Many of the genes needed for development are so critical that if mutation renders them nonfunctional, the embryo dies before the organism is fully formed. Yet, even though the needs of a eukaryote are more complex than those of a bacterium, basic principles of gene regulation are still the key to all of these processes.

Recall that many bacterial genes and operons are regulated at the level of transcription initiation. This is true in eukaryotes as well, and as we'll see, many eukaryotic regulatory mechanisms build on those used in bacteria. However, there is a fundamental difference in bacterial and eukaryotic regulation of transcription. The **transcriptional ground state**, the inherent activity of promoters and transcription machinery *in vivo* in the absence of regulatory mechanisms, is not the same in bacteria and eukaryotes. In bacteria, the transcriptional ground state is nonrestrictive. In other words, RNA polymerase generally has access to every promoter and can bind and initiate transcription at some level of efficiency in the absence of activators or repressors. In contrast, eukaryotic genes contain strong promoters that are generally inactive in the absence of regulatory proteins; that is, the transcriptional ground state in eukaryotes is restrictive.

Crucial differences in DNA packaging and cell structure give rise to at least four important distinguishing features of regulation of gene expression in eukaryotes. First, access to eukaryotic promoters is restricted by the structure of chromatin, and transcriptional activation is associated with many changes in chromatin structure in the transcribed region. Second, although eukaryotic cells have both positive and negative regulatory mechanisms, positive mechanisms predominate in all systems investigated so far; given that the transcriptional ground state is restrictive, virtually every eukaryotic gene requires activation. Third, eukaryotic cells have larger, more complex, multiprotein regulatory networks than bacteria. And finally,

transcription in the nucleus is separated from translation in the cytoplasm, in both space and time. As a result, posttranscriptional control plays a larger role in controlling gene expression in eukaryotes, as we'll see in Chapter 22.

The complexity of regulatory circuits in eukaryotic cells is extraordinary, and we won't attempt comprehensive coverage of all aspects. This chapter and the next cover some of the guiding principles of eukaryotic gene regulation, drawing parallels to the mechanisms discussed for bacteria in Chapter 20 wherever applicable. The need to control the multitude of genes in a higher eukaryote requires an array of regulatory proteins for every single gene. We explain the basic logic of gene activation as used in essentially all eukaryotes. We also take a brief look at experiments that first revealed the modular architecture of gene activators and highlight some of the regulatory networks that govern gene expression, from the simple system in yeast to the complex developmental controls typical in a multicellular eukaryote. The chapter concludes with a discussion of transcriptional control processes unique to eukaryotic gene expression, some of which are still far from understood.

## 21.1 Basic Mechanisms of Eukaryotic Transcriptional Activation

As in bacteria, the basal level of eukaryotic transcription is determined by the effect of regulatory sequences on the function of RNA polymerase and its associated transcription factors. As we learned in Chapter 19, the nature of the eukaryotic genome lends itself to different regulatory strategies from those used in bacteria. The eukaryotic genome is packaged in chromatin, which presents a physical block to the RNA polymerases, and therefore the majority of eukaryotic genes are repressed in their default (ground) state and require protein activators to stimulate expression. Because of the large size of eukaryotic genomes and the need to guard against nonspecific protein-DNA interactions, the binding of multiple protein regulators is required to activate each gene. As a result, eukaryotic promoters are more complex than their bacterial counterparts and contain many more regulatory protein-binding sites. In reality, though, the additional complexity in eukaryotes is handled in strategic ways that are not as complicated as we might expect, given the overwhelming difference between a bacterium and an animal.

## Eukaryotic Transcription Is Regulated by Chromatin Structure

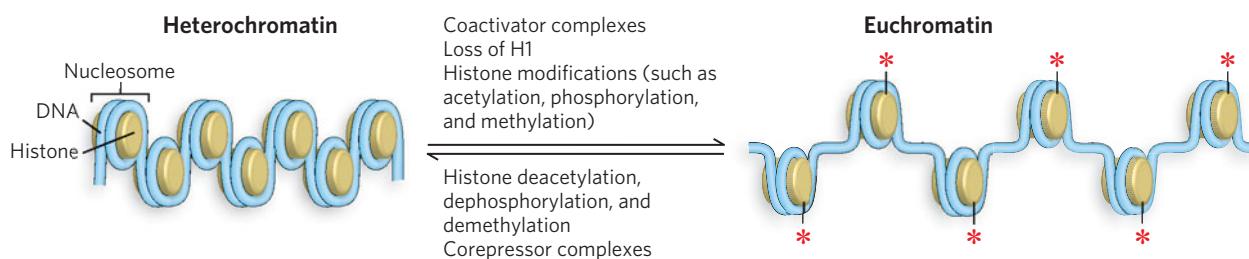
The genomic DNA of eukaryotes wraps around small basic proteins called histones to form nucleosomes, the building blocks of chromatin (see Figure 10-4). The transcription machinery must necessarily deal with chromatin structure in order to access particular genes. As a result, eukaryotic genes are generally expressed at low levels—or not at all—in the absence of regulatory proteins.

Chromatin structure is controlled and altered by at least three interrelated mechanisms: ATP-dependent changes in nucleosome positioning on the DNA, post-translational chemical modifications of histone proteins, and substitution of specialized histone variants into chromatin. These mechanisms were discussed in detail in Chapter 10, and we recap briefly here. Nucleosome remodeling complexes use ATP to shift nucleosomes along the DNA. Active promoters contain open regions with nucleosomes positioned away from the promoter region, allowing access to transcription factors. Some posttranslational modifications of histones, including acetylation by histone acetyltransferases (HATs), result in a decondensing of chromatin and provide access to DNA-binding factors; proteins containing a bromodomain bind acetylated histones and facilitate opening of the chromatin structure. Alternatively, histone modifications cause chromatin to become tightly closed to transcription. For example, methylated histones are bound by proteins containing chromodomains, and these proteins help condense the chromatin. Chromatin structure is also modulated by several histone variants. These proteins are homologous to the common histones and can take their place in nucleosomes, but they also contain amino acid extensions that have a variety of functional consequences.

In the eukaryotic cell cycle, interphase chromosomes appear to be dispersed and amorphous. However, chromosomes are not uniform structures, and several different forms of chromatin can be found along each chromosome. About 10% of the chromatin in a typical eukaryotic cell is in a much more condensed form than the rest of the chromatin. This form, **heterochromatin**, is transcriptionally inactive. Heterochromatin is often associated with particular chromosome structures, including centromeres and telomeres. The remaining, less condensed chromatin is called **euchromatin** (Figure 21-1).

Transcription of a eukaryotic gene is strongly repressed when its DNA is condensed within heterochromatin, but in euchromatin some of the DNA is transcriptionally active. Regions of transcriptionally active DNA can be detected based on their increased sensitivity to nuclease-mediated degradation. Nucleases such as DNase I tend to cleave the DNA of carefully isolated chromatin into fragments of multiples of about 200 bp, reflecting the regular repeating structure of the nucleosome (see Figure 10-1). However, in actively transcribed regions, the fragments produced by nuclease activity are even smaller and more heterogeneous in size. Actively transcribed regions contain **hypersensitive sites**, sequences especially sensitive to DNase I, which are typically found in noncoding regions within 1,000 bp of the 5' ends of transcribed genes. In some genes, hypersensitive sites are found farther from the 5' end, or near the 3' end, or even within the gene itself. The presence of hypersensitive sites suggests that DNA in that region is not packaged in the regular repeating nucleosomal structure.

Many hypersensitive sites correspond to binding sites for known regulatory proteins, and the relative absence of nucleosomes in these regions may allow the binding of these proteins. Nucleosomes are



**FIGURE 21-1 Heterochromatin and euchromatin.**

Nucleosomes in heterochromatin are tightly packed together, and the DNA is transcriptionally silent. Nucleosomes in euchromatin are spaced farther apart, and the DNA can be

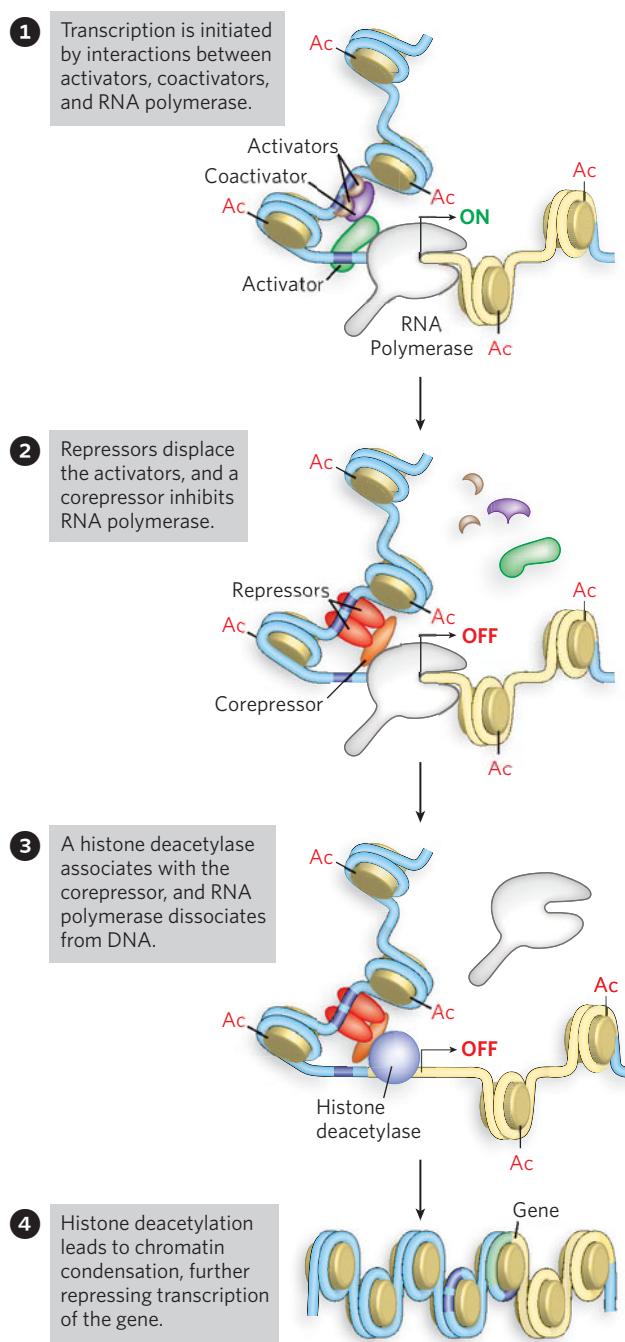
decondensed so that it becomes accessible to the transcription machinery. The two chromatin states are regulated by histone modifications (represented by red asterisks) and the binding of other factors (see Chapter 10).

entirely absent in some regions that are very active in transcription, such as the rRNA genes. Transcriptionally active chromatin also tends to be deficient in histone H1, which binds the linker DNA between nucleosome particles.

Histones within transcriptionally active chromatin and heterochromatin also differ in their patterns of covalent modification. The C-terminal tails of the core histones are modified by the acetylation and methylation of Lys and Arg residues, phosphorylation of Ser or Thr residues, and ubiquitination or sumoylation (see Chapter 22). In particular, the acetylation-deacetylation of histones figures prominently in the processes that activate chromatin for transcription. The HAT-mediated acetylation of multiple Lys residues in the N-terminal domains of histones H3 and H4 can reduce the affinity of the entire nucleosome for DNA. Acetylation may also prevent or promote interactions with other proteins involved in regulating transcription. When transcription of a gene is no longer required, acetylation of nucleosomes in that vicinity is reduced by the activity of histone deacetylases (HDACs), resulting in condensation of the chromatin to reduce or inactivate gene transcription. HDACs often function through protein-protein interactions, such as by binding corepressors or as components of chromatin remodeling complexes.

A general model for deacetylation and gene inactivation is shown in **Figure 21-2**. In addition to the removal of certain acetyl groups, new covalent modifications of histones mark chromatin as transcriptionally inactive. For example, the Lys residue at position 9 in histone H3 is often methylated in heterochromatin.

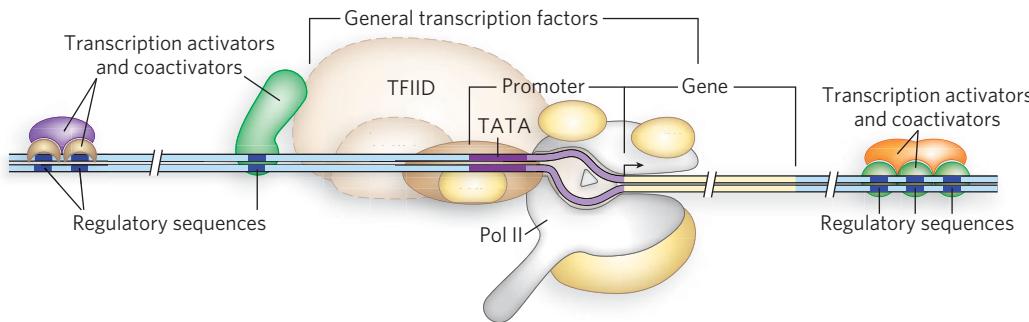
Gene regulation through histone modifications is typically achieved through activation or inhibition of transcription initiation. However, an exciting finding has demonstrated that chromatin structure is also involved in the control of mRNA splicing for some genes (see Moment of Discovery). This may seem surprising, given that histones do not bind RNA. Yet Gcn5, a well-studied transcription factor with HAT activity, is now known to also affect RNA processing: loss of Gcn5 HAT activity prevents proper pre-mRNA splicing in yeast, because components of the splicing machinery fail to properly bind the pre-mRNA splice sites (Highlight 21-1). This example shows how different levels of regulation (in this case, transcription and mRNA processing) can be interrelated. It seems likely that processes previously thought to be isolated and separable events may be interwoven in complex regulatory networks in the living cell.



**FIGURE 21-2** Gene inactivation by histone deacetylation.  
[Source: Adapted from T. M. Malavé and S. Y. R. Dent, *Biochem. Cell Biol.* 84:437–443, 2006, Fig. 1.]

## Positive Regulation of Eukaryotic Promoters Involves Multiple Protein Activators

Each of the three eukaryotic RNA polymerases has little or no intrinsic affinity for its promoters. Instead, initiation of transcription almost always requires activator



**FIGURE 21-3 A typical eukaryotic promoter.** General transcription factors and RNA polymerase II bind the promoter, assisted by transcription activators. Activator-binding sites (regulatory sequences) can be distant from the promoter and located either before or after the gene.

Activators bind regulatory sequences in DNA directly, whereas coactivators bind activators instead of DNA. Activation of Pol II is mediated by coactivator binding to core subunits of the polymerase through DNA looping.

proteins. An important reason for the apparent predominance of positive regulation is clear from the earlier discussion: chromatin structure effectively renders most promoters inaccessible, so genes are normally silent in the absence of other regulation. The structure of chromatin affects access to some promoters more than others, but repressor binding to DNA to block access of RNA polymerase (negative regulation) would often be simply redundant. Other factors are also at play in the use of positive regulation, however, and speculation generally centers around two: the large size of eukaryotic genomes and the greater efficiency of positive regulation.

Because eukaryotes have much larger genomes than bacteria, there is an increased likelihood that a specific binding sequence for a regulatory protein will occur randomly in other regions of the DNA. Recall that a sequence of  $n$  nucleotides will occur randomly every  $4^n$  bp. Thus, any single regulatory protein with a small binding site will probably bind nonspecifically to multiple places in the eukaryotic genome (see How We Know). Specific transcriptional activation of a gene through the binding of one regulatory protein to a small binding site, as often occurs in bacteria, would be an ineffective strategy in eukaryotes. In theory, specificity of regulator binding would be increased if the DNA sequences recognized by the proteins were longer. Yet eukaryotes did not evolve in this way: eukaryotic regulators do not bind longer DNA sequences than their bacterial counterparts. Instead, to achieve specific transcriptional activation of a gene, eukaryotes employ *multiple* regulatory proteins, or transcription factors, each of which binds a short sequence; successful gene activation occurs only when all the factors are bound at their individual sites. This “combination” of factors for

activating one gene is used in combinatorial control (see Section 21.2).

To accommodate the binding of multiple transcription factors, eukaryotic promoters are necessarily more complicated than are their bacterial counterparts (Figure 21-3). Take, for example, a typical promoter recognized by RNA polymerase II (Pol II), the enzyme responsible for mRNA synthesis. Many (but not all) Pol II promoters include the TATA box and Inr (initiator) sequences, with their standard spacing (see Figure 15-19). These sequences comprise the core promoter.

Eukaryotic genes also include regulatory sequences called **enhancers** in higher eukaryotes and **upstream activator sequences (UASs)** in yeast, to which transcription activators bind. These sequences cannot all be positioned adjacent to the promoter—there is simply not enough room to accommodate the binding of so many regulatory proteins. The binding sites for multiple transcription factors must be able to act at a distance. In fact, they can be surprisingly far from the promoter. A typical enhancer may be hundreds or even thousands of base pairs upstream from the transcription start site, or downstream from the gene, or even within the gene itself. When bound by the appropriate regulatory proteins, an enhancer increases transcription at nearby promoters regardless of its orientation in the DNA. Yeast UASs function in a similar way, although generally they must be positioned upstream and within a few hundred base pairs of the transcription start site. An average Pol II promoter may be affected by half a dozen regulatory sequences of this type, and even more-complex promoters are common. In contrast, bacteria have very few genes that use a distantly bound transcription activator, and when they do, the distance from the promoter is usually quite short.

## HIGHLIGHT 21-1 A CLOSER LOOK

### The Intertwining of Transcription and mRNA Splicing

Initiation is the most highly regulated step in transcription, an intricate process that requires, in eukaryotes, the coordinated action of numerous proteins. Transcription generates a pre-mRNA needing many modifications before it can be transported to the cytoplasm for translation. One of the most complex modifications en route to active mRNA is the removal of introns. The splicing machinery requires more than 100 proteins and five different splicing RNAs with complicated three-dimensional structures. Splicing is generally regarded as a separate step occurring after transcription initiation, or even after generation of the entire pre-mRNA, partly because of the complexity of the transcription and splicing processes. So it was surprising to discover that these two complicated processes—transcription and splicing—can happen simultaneously for some genes: transcription seems to deposit the U2 snRNP component of the spliceosome at specific sites in the pre-mRNA as it is synthesized. These sites correspond to branch points, the sites containing the 2'-OH nucleophile that initiates intron splicing and results in a branched, lariat-type structure when the intron RNA is excised (see Chapter 16). To explain why transcription and splicing would coordinate in this fashion, researchers have proposed that the rate of transcription elongation may be regulated by the spliceosome to help pick and choose alternative splice sites, thereby controlling the relative levels of different mRNAs produced from a pre-mRNA.

The true picture of what is going on is even more complicated, however, as revealed by recent work in Tracy Johnson's laboratory (see Moment of Discovery). Johnson made the fascinating observation that Gcn5, a histone acetyltransferase (HAT), is an integral component in the coregulation of transcription and pre-mRNA splicing. The HAT activity of Gcn5, like other HATs, can

The more complex the eukaryotic organism, the more complex its promoters are likely to be. For example, yeast promoters are generally much simpler than those of mammals (Figure 21-4).

The requirement for binding of several transcription activators to several specific DNA sequences vastly reduces the probability of the random occurrence of a functional juxtaposition of all the necessary binding sites. In principle, a similar strategy could be used by multiple

alter chromatin structure, which is thought to be important in regulating transcription initiation. But results from Johnson's lab demonstrate that accurate mRNA splicing, too, requires the Gcn5 HAT activity.

How does a HAT help splicing? After all, RNA is not bound by histones, so what role does Gcn5 play in the splicing process? One possibility is that Gcn5 alters the speed at which Pol II moves through the chromatin structure, thus influencing alternative splicing by allotting different amounts of time for the transcription machinery to deposit splicing factors at splice sites before completion of the transcript. But Johnson's work argues against this simplistic idea: deletion of Gcn5 does not affect the transcription elongation rate.

Johnson found that the coordination between transcription and splicing occurs even before the pre-mRNA is fully synthesized. The first experiment hinting at this conclusion came from genetic experiments showing that deletion of the gene encoding Gcn5 (and not other yeast lysine acetyltransferases that target histones) is lethal in yeast cells that also lack either of the genes encoding the U2 snRNP proteins Lea1 and Msl1. Neither Lea1 nor Msl1 is an essential protein in yeast, except when Gcn5 is missing.

Next, using the technique of chromatin immunoprecipitation (ChIP), Johnson's group showed that spliceosomal proteins are recruited directly to an intron branch point within the well-characterized *DBP2* gene. In the ChIP experiment, individual snRNP particles are formaldehyde cross-linked to the transcription complex or to the nascent RNA and immunoprecipitated (see Figure 10-21). When the associated DNA is amplified using specific PCR primer sets, the signal is enriched in regions of the gene where the snRNPs associate with the corresponding pre-mRNA. Johnson's results revealed that antibodies to Lea1, a component of the spliceosome, immunoprecipitated a relatively large amount of DNA corresponding to the branch point of the *DBP2*

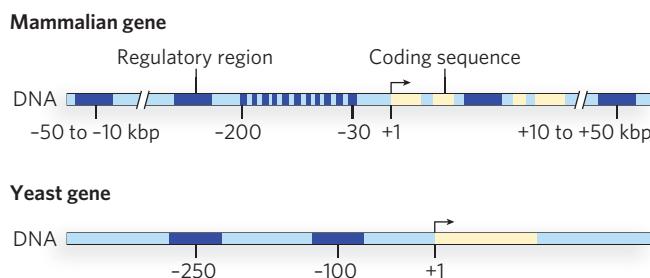
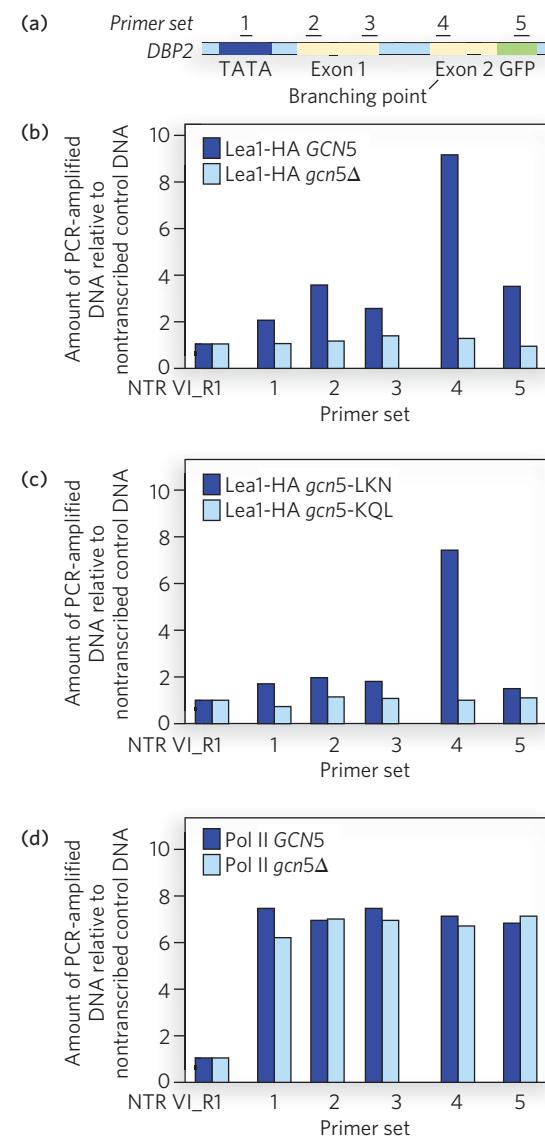
negative-regulatory elements. However, positive regulation is simply more efficient. From an energy standpoint, it makes more sense for the cell to synthesize several activators to promote transcription of the subset of genes needed at that time, rather than constantly synthesize one or more repressors for every gene in the genome to keep them turned off until needed. Positive regulation of transcription predominates in eukaryotes, although as we'll see, there are some examples of negative regulation.

pre-mRNA (Figure 1a, b). Recruitment of Lea1 to the branch point depended on the presence and catalytic activity of Gcn5 (Figure 1c). A control experiment showed that the occupancy by Pol II of these regions of the *DPB2* gene was unaltered, whether or not Gcn5 was active (Figure 1d). Thus, the data indicate that Gcn5 sets the stage for the recruitment of spliceosomal components before the splice site junctions are even transcribed. Further experiments (not shown) have demonstrated the same results for other genes.

Johnson also proposes other explanations for these observations. One possibility is that Gcn5 acetylates nonhistone proteins, perhaps even spliceosomal subunits. Another possibility is that hyperacetylation of histones at the promoter may facilitate recruitment of the splicing apparatus. Understanding the

**FIGURE 1** Gcn5 activity helps recruit spliceosomal components to *DPB2* pre-mRNA. (a) Numbers represent regions of DNA in the *DPB2* gene that are amplified in the ChIP analysis. (b) ChIP analysis of yeast cells expressing an engineered version of Lea1 tagged with a hemagglutinin (HA) peptide. Lea1-HA was immunoprecipitated with anti-HA antibodies, and Lea1 occupancy in the indicated regions of *DPB2* was compared with that of a nontranscribed region of DNA (NTR VI\_R1). Sets of PCR primers corresponding to the regions indicated in (a) were used to amplify specific segments of chromatin after Lea1-HA immunoprecipitation. Dark blue bars are data for cells with wild-type Gcn (*GCN5*); light blue bars are data for cells with a Gcn5 deletion (*gcn5Δ*). (c) ChIP analysis as in (b), except that the dark blue bars are results for cells with a point mutation in a nonessential region of Gcn5 (*gcn5-LKN*) and the light blue bars are results for cells with a point mutation in the Gcn5 active site (*gcn5-KQL*). (d) ChIP analysis as in (b), except that this control experiment uses Pol II instead of Lea1. [Source: Adapted from F. Q. Gunderson and T. L. Johnson, *PLoS Genet.* 5:1–12, 2009.]

full details of coordinated regulation of transcription initiation, histone acetylation, and recruitment of the spliceosomal machinery will take considerably more time and work.



**FIGURE 21-4 A comparison of mammalian and yeast promoter regions.** The promoter regions of multicellular organisms contain more control elements than those of unicellular eukaryotes, such as yeast. This reflects the need in higher eukaryotes for changes in gene expression during development and for intercellular communication. All regulatory regions are shown in dark blue, coding regions in yellow.

To further conserve resources, differently regulated eukaryotic promoters often use some of the same protein activators, so diverse promoters can have some of the same binding sequences. However, only a specific combination of regulatory factors can unlock a given promoter and activate transcription of that gene. With this mechanism, the cell can achieve specificity of gene regulation with a smaller number of transcription activators than if each gene were regulated by a set of unique proteins (see Figure 19-14). Some regulatory proteins facilitate transcription at hundreds of promoters, whereas others are specific for only a few promoters. In addition, many transcription activators are sensitive to the binding of effector signal molecules, providing the capacity to activate or deactivate transcription in response to a changing cellular environment.

### Transcription Activators and Coactivators Help Assemble General Transcription Factors

Successful binding of active Pol II holoenzyme at one of its promoters usually requires the action of three types of regulatory proteins: **general (basal) transcription factors**, which are required at every Pol II promoter; **DNA-binding transcription activators**, also called **DNA-binding transactivators**, which bind to enhancers or UASs to facilitate transcription; and **coactivators**, which act indirectly—by binding other proteins rather than DNA—and are required for essential communication between the DNA-binding transactivators and the complex composed of Pol II and the general transcription factors (Figure 21-5a). Sometimes, a variety of repressor proteins can interfere with communication between Pol II and the DNA-binding transactivators, resulting in repression of transcription (Figure 21-5b). In fact, some proteins act as an activator or coactivator at one promoter and a repressor or corepressor at another promoter. Here we focus on the protein complexes shown in Figure 21-5a and how they interact to activate transcription.

For transcription to begin, the Pol II holoenzyme must be recruited to the promoter to form a preinitiation complex with the general transcription factors. Assembly of a preinitiation complex at a typical Pol II promoter begins with the binding of **TATA-binding protein (TBP)** to the TATA box. TBP, which is part of the larger transcription factor complex called TFIID, then recruits additional general transcription factors and Pol II (see Figure 15-23). The minimal preinitiation complex, however, is often insufficient for the initiation of transcription, and generally does not form at all if the promoter is buried in chromatin. Positive regulation

by transcription activators and coactivators is required. We now know that the basal Pol II machinery is not as uniform as originally thought; the individual components can vary with cell type. A well-documented example is muscle cells (see How We Know). Thus, different combinations of general transcription factors form a complex at a promoter and are acted on by specific activator and coactivator binding proteins, adding a further level of control to the regulation of that gene.

As noted above, binding sites for transcription activators are often located far from the promoters they regulate. Recall from Chapter 19 that the intervening DNA is looped so that the various protein complexes can interact, directly or indirectly. DNA looping is promoted by certain nonhistone proteins that are abundant in chromatin and bind nonspecifically to DNA. These **high-mobility group (HMG) proteins** play an important structural role in chromatin remodeling and transcriptional activation. (“High mobility” refers to their rapid electrophoretic mobility in polyacrylamide gels.) A structure formed by an HMG-box domain in HMG proteins can bind directly to nucleosomes, leading to altered local chromatin structure. Figure 21-6 shows the high degree to which DNA is bent by the HMG-box domain of the protein HMG-D of *Drosophila melanogaster*, one of many DNA-interactive protein structures determined in the laboratory of Mair Churchill.

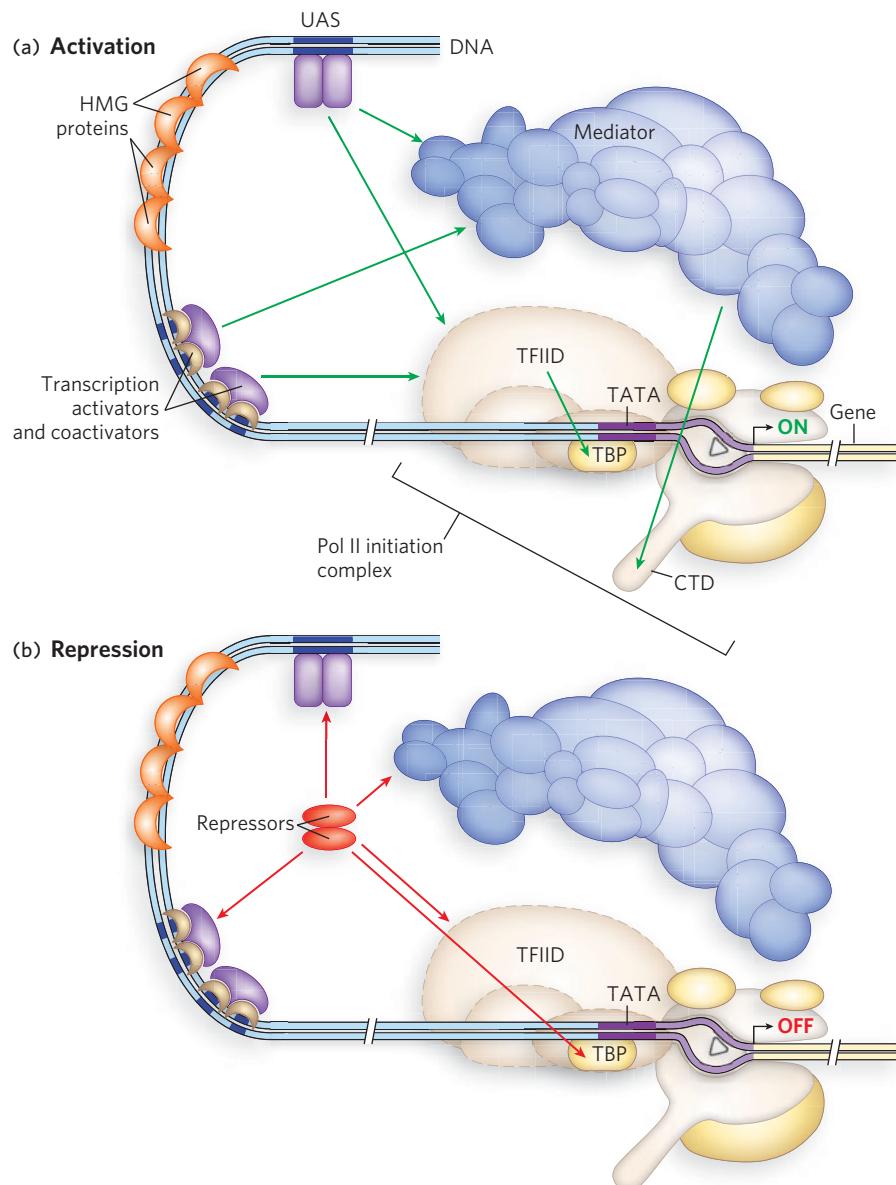
In addition to transcription activators, most transcription requires coactivator protein complexes. Some major regulatory protein complexes that interact with Pol II have been defined both genetically and biochemically. They act as intermediaries between the DNA-binding transactivators and the Pol II complex (Pol II and the general transcription factors). The best-characterized coactivator is TFIID (see Chapter 15). In eukaryotes, TFIID includes TBP and 10 or more TBP-associated factors (TAFs). Some TAFs resemble histones and may play a role in competing with and thus displacing nucleosomes during the activation of transcription. Many DNA-binding transactivators aid in transcription initiation by interacting with one or more TAFs.

Another important coactivator is the **Mediator complex** (see Figure 15-24), which consists of 20 core polypeptides that are highly conserved, from fungi to humans. Mediator binds tightly to the C-terminal domain (CTD) of the largest Pol II subunit. The Mediator



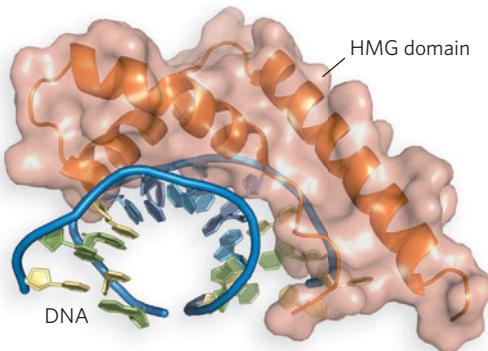
**Mair Churchill**

[Source: Courtesy of Mair Churchill.]

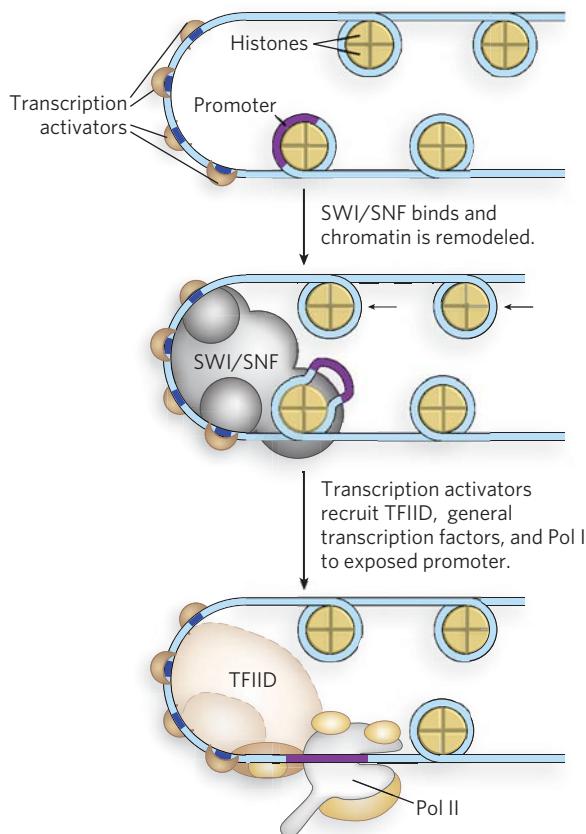


**FIGURE 21-5 Mechanisms of activation and repression of eukaryotic gene expression.** (a) Transcription activators and coactivators bound to distant regulatory sites (enhancers and UASs) recruit components of the Pol II general (basal) transcription machinery to the promoter. Coactivators

such as Mediator and TFIID are required at essentially all promoters. They function as a bridge between activators and the polymerase, and do not interact with DNA directly. (b) Repression is mediated by proteins that disrupt or prevent essential contacts between Pol II and activators or coactivators.



**FIGURE 21-6 DNA looping facilitated by HMG proteins.** HMG proteins bend DNA, helping form loops between enhancer and promoter elements. Binding is nonspecific. Shown here is the HMG-box DNA-binding domain of the protein HMG-D of *Drosophila*, bound to DNA. [Source: PDB ID 1QRV.]



**FIGURE 21-7** **Transcription activator-mediated chromatin remodeling.** Transcription activators can remodel chromatin structure by mobilizing nucleosomes; nucleosome repositioning is influenced by histone modifications. Some transcription activators have HAT activity or recruit enzyme complexes such as SWI/SNF, accelerating the remodeling of chromatin by relocating nucleosomes near a promoter. This leads to recruitment of the transcription machinery to newly exposed promoters, stimulating transcription.

complex is required for both basal and regulated transcription at Pol II promoters, and it also stimulates phosphorylation of the Pol II CTD by the general transcription factor TFIH. Phosphorylation of the CTD enhances Pol II efficiency. As with TFIID, some DNA-binding transactivators interact with one or more components of the Mediator complex. Some promoters require both Mediator and TFIID coactivators. The coactivator complexes function at or near the promoter's TATA box.

We can now begin to piece together the sequence of transcriptional activation events at a typical Pol II promoter. Crucial remodeling of the chromatin takes place in stages. Some DNA-binding transactivators have significant affinity for their binding sites even when the sites are within condensed chromatin. Binding of one transactivator may facilitate the binding of others, gradually displacing some of the nucleosomes that previously obscured the relevant DNA.

The bound transcription activators may have HAT activity or may recruit HATs or enzyme complexes such as SWI/SNF, accelerating the remodeling of surrounding chromatin (Figure 21-7). In this way, transcription activator binding can lead to the stepwise assembly of components necessary for further chromatin remodeling, to permit the transcription of specific genes. The bound transactivators, acting through complexes such as TFIID or Mediator (or both), stabilize the binding of Pol II and its associated general transcription factors, greatly facilitating formation of the preinitiation complex. Complexity in these regulatory circuits is the rule rather than the exception, with multiple DNA-bound transactivators promoting transcription.

The script can change from one promoter to another, but most promoters seem to require a precisely ordered assembling of components to initiate transcription. The assembly process is not always fast. At some genes it may take minutes; at certain genes in higher eukaryotes, the process can take days.

Although rarer, some eukaryotic regulatory proteins that bind Pol II promoters can act as repressors, inhibiting the formation of active preinitiation complexes. Some transcription activators can adopt different conformations, enabling them to serve as activators or repressors. For example, some steroid hormone receptors function in the nucleus as DNA-binding transactivators, stimulating transcription of certain genes when a particular steroid hormone signal is present (see Section 21.3). When the hormone is absent, the receptor proteins revert to a repressor conformation, preventing formation of preinitiation complexes. In some cases this repression involves interaction with HDACs and other proteins that help restore the surrounding chromatin to its transcriptionally inactive state.

## SECTION 21.1 SUMMARY

- Most eukaryotic genes are inactive in their ground state, as histones cover the DNA, and are under positive control; they require multiple activator proteins to stimulate transcription.
- The eukaryotic RNA polymerases require activator binding to promoter sequences to activate gene expression. The cell produces only the activator proteins necessary for transcription of the subset of genes needed at that time.
- Many Pol II promoters include the TATA box and Inr sequences, as well as other sequences located far from the promoter. When bound by the appropriate regulatory proteins, these distant regulatory sequences—enhancers in higher eukaryotes and upstream activator sequences in

yeast—function at the promoter through DNA looping, increasing transcription regardless of their orientation in the DNA. The DNA bending is facilitated by HMG proteins.

- Transcription is stimulated by interactions between RNA polymerase core subunits and transcription activators (transactivators) bound to enhancer sequences. Often, coactivator complexes such as TFIID or Mediator act as bridges between the core transcription machinery and transactivators.

## 21.2 Combinatorial Control of Gene Expression

The expression of eukaryotic genes is modulated by combinations of transcription factors, and when some of these factors are common to the regulation of multiple genes, the regulation is called **combinatorial control**. We learned in Chapter 20 that different bacterial genes driving sugar metabolism use a common transcription activator, cAMP receptor protein (CRP). CRP is employed in regulation of the lactose and galactose operons, as well as operons involved in the metabolism of other sugars. This is an example of combinatorial control.

Eukaryotes make much more extensive use of combinatorial control than do bacteria. First of all, as we've seen, eukaryotes generally require many regulatory proteins at any given promoter, increasing the combinatorial possibilities severalfold. Indeed, analysis of genome sequences reveals the use of greater numbers of transcription factors as genome size and complexity increase. For example, yeast are thought to use about 300 transcription factors, *Caenorhabditis elegans* and *D. melanogaster* more than 1,000, and humans more than 3,000. Although the number of transcription factors increases with the number of genes, there are still many fewer factors than there are genes to be regulated. Somehow, different genes must use the same transcription factors, but in different ways, to achieve activation. Given the increasing complexity of promoter sequences in more complex genomes and the greater number of transcription factors, combinatorial control allows higher eukaryotes to achieve exquisite specificity in gene regulation.

We begin with the relatively simple combinatorial control system that regulates the yeast *GAL* genes, driving the metabolism of galactose. The mechanism behind galactose metabolism is one of the best-understood systems (Highlight 21-2). We then describe some increasingly complex mechanisms of combinatorial gene regulation.

### Combinatorial Control of the Yeast *GAL* Genes Involves Positive and Negative Regulation

The enzymes required for importing and metabolizing galactose in yeast are encoded by *GAL* genes scattered over several chromosomes. Yeast cells have no operons like those in bacteria, and each of the *GAL* genes is transcribed separately. However, all the *GAL* genes have similar promoters and are regulated coordinately by a common set of proteins. The promoters for the *GAL* genes consist of the TATA box and an upstream activator sequence, which for each *GAL* gene is composed of one or more sequences denoted UAS<sub>GAL</sub>. Each UAS<sub>GAL</sub> site is recognized by a DNA-binding transactivator, the Gal4 protein (Gal4p). For example, the UAS of the gene *GAL1* is 118 bp long and contains four Gal4p-binding sites of 17 bp each (Figure 21-8).

Like the bacterial *lac* operon, the yeast *GAL* genes require more than just one protein (Gal4p) for activation. Control of gene expression by galactose depends on three proteins—the transcription activator Gal4p, the inhibitor Gal80p, and the ligand sensor Gal3p (Figure 21-9). Gal4p binds the 17mer UAS<sub>GAL</sub> sites and, left to its own devices, would activate gene expression at *GAL* promoters. However, at low galactose concentrations, Gal80p binds to Gal4p and blocks its transcription-activating region. When galactose is present, it binds Gal3p; Gal3p also binds ATP, and the Gal3p-galactose-ATP complex then interacts with Gal80p. This interaction causes a conformational change that relieves the inhibition of Gal4p and allows it to function as a transactivator at *GAL* promoters.

Glucose is the preferred carbon source for yeast, as it is for bacteria. When glucose is present, most of the *GAL* genes are repressed—whether galactose is available or not. The *GAL* gene regulatory system described above is effectively overridden by a global catabolite repression system. Global repression is achieved by the protein Mig1, which binds near the *GAL* promoter. Repression of the *GAL* genes also requires Tup1, a corepressor that binds Mig1 (Figure 21-10). Mig1 is regulated in a way that is not possible in bacteria—namely,



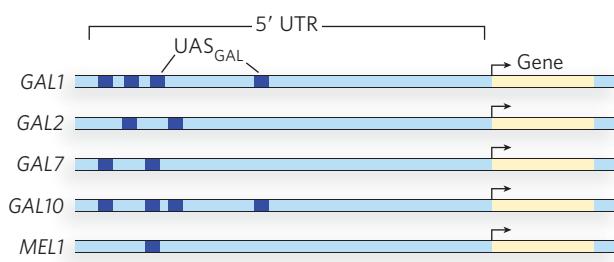
**FIGURE 21-8** The *GAL1* promoter. The promoters of the *GAL* genes of yeast each contain an upstream activator sequence (UAS), composed of one or more UAS<sub>GAL</sub> sites. Each 17 bp UAS<sub>GAL</sub> sequence is a binding site for transcription activator Gal4p. The UAS of the *GAL1* gene has four UAS<sub>GAL</sub> sites.

## HIGHLIGHT 21-2 TECHNOLOGY

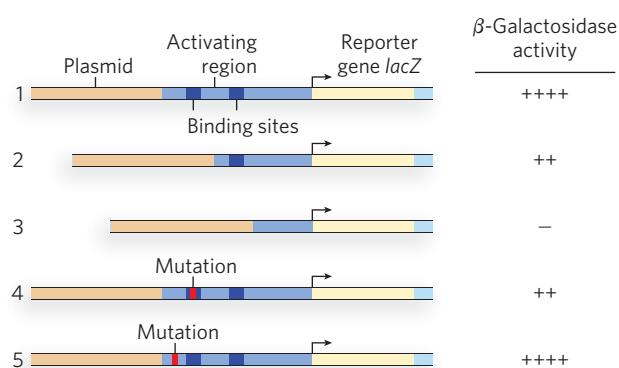
### Discovering and Analyzing DNA-Binding Proteins

Regulatory DNA sequences, such as the binding site for Gal4p in yeast, can be identified by sequence comparisons of genes that code for proteins of the same metabolic pathway. The Gal4p-binding site was one of the first eukaryotic activator-binding sites to be recognized. Genetic studies identified several genes in the pathway of galactose metabolism in yeast. In the presence of galactose, expression of the *GAL* genes increases as much as 1,000-fold. Clones containing the regulated *GAL* genes were sequenced, and comparison of the regions upstream from the TATA boxes revealed a common sequence, designated UAS<sub>GAL</sub> (Figure 1). The UAS<sub>GAL</sub> sequence is a 17mer, CGG(N)<sub>11</sub>CCG, with a twofold axis of symmetry, indicating that the protein that binds it probably functions as a dimer. In vivo, mutation of UAS<sub>GAL</sub> sequences upstream from the *GAL* genes eliminated the usual activation in response to galactose. In a reporter gene assay in which the UAS<sub>GAL</sub> sequence was cloned into the upstream region of *lacZ*,  $\beta$ -galactosidase (the *lacZ* gene product) expression was induced by addition of galactose. Furthermore, expression levels of  $\beta$ -galactosidase depended on the number and sequence of UAS<sub>GAL</sub> sites, confirming their importance in transcription activation (Figure 2).

We now know that the *GAL* genes are activated by the protein Gal4p, which recognizes UAS<sub>GAL</sub>. Early experiments demonstrated that Gal4p binds UAS<sub>GAL</sub> and functions as a transcription activator. Genetic



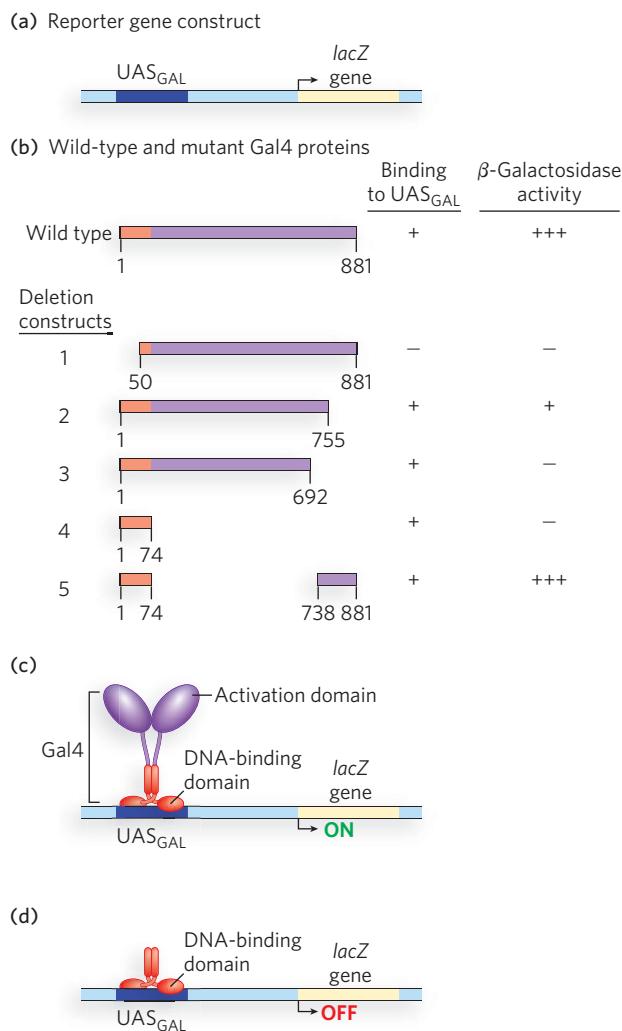
**FIGURE 1** A comparison of the upstream sequences of the yeast *GAL* genes showed that they have common sequences, the UAS<sub>GAL</sub> sites, each 17 bp long (blue). (One *GAL* gene, *GAL5* (*PGM2*), is missing here because it is regulated differently and does not have a UAS<sub>GAL</sub> sequence.)



**FIGURE 2** The function of UAS<sub>GAL</sub> sequences was confirmed in reporter gene assays in which promoter activity was determined by the activity of  $\beta$ -galactosidase (produced by the bacterial *lacZ* gene). As shown in these five assays (1 is the wild-type), deletion or mutation of UAS<sub>GAL</sub> elements, but not other areas close to the promoter, resulted in decreased promoter activity ( $\beta$ -galactosidase level).

studies revealed that a single gene, when mutated, results in loss of activation of all *GAL* genes. These results suggested that this single gene, *GAL4*, was a master regulator, much like the bacterial CRP protein. *GAL4* was isolated by transforming a yeast genomic library into *GAL4*-mutant cells and selecting for colonies in which the *GAL* genes were again activated in the presence of galactose. *GAL4* was then cloned into an *E. coli* expression vector, and Gal4p was purified (see Chapter 7 for these cloning methods).

The technique of **deletion analysis** revealed the modular architecture of Gal4p, which is now known to be common among many bacterial and eukaryotic transcription activators. In deletion analysis, nucleases or restriction enzymes are used to selectively delete pieces of DNA from a specified gene. The truncated protein product of this gene can be purified and tested for activity in vitro, or tested for function in vivo using a reporter assay. Studies such as these were performed with deletion constructs of Gal4p. DNA binding of the truncated proteins was measured in vitro with electrophoretic mobility shift assays, and the ability of the truncated proteins to activate transcription was tested in vivo with a reporter gene assay. In the reporter assay, deletion constructs of *GAL4* were transferred into *GAL4*-mutant yeast cells containing a plasmid with the bacterial *lacZ* reporter gene, driven by a typical *GAL* promoter with a



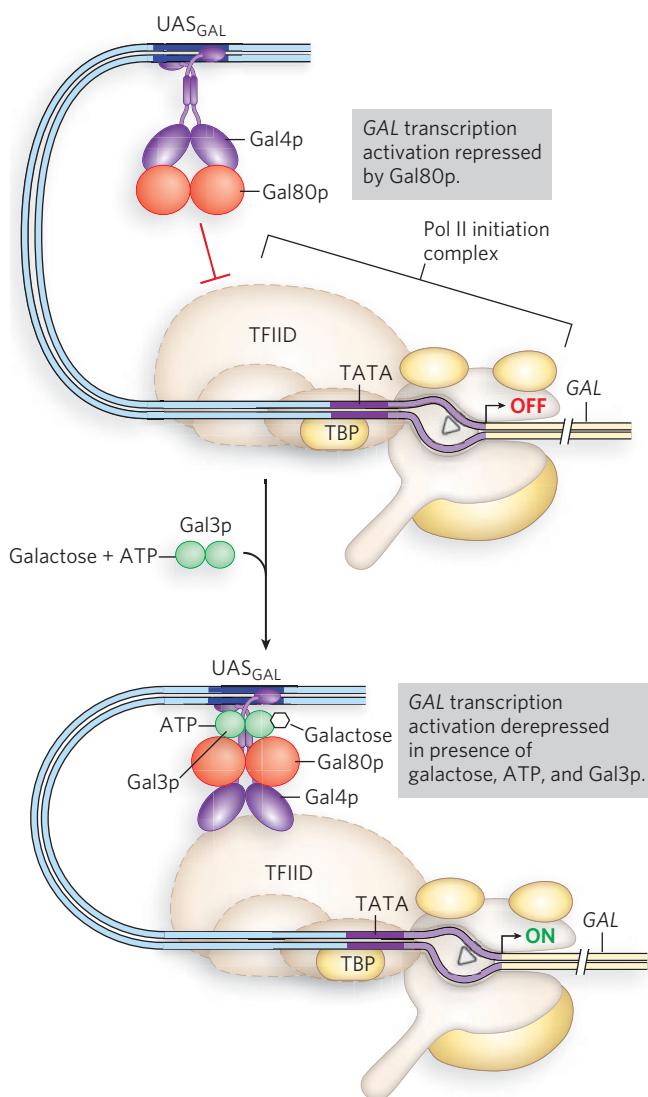
**FIGURE 3** (a) The reporter gene construct used for deletion analysis of Gal4p. Only constructs with functional Gal4p will bind  $UAS_{GAL}$  and drive expression of the reporter gene ( $lacZ$ ). (b) Deletion analysis of Gal4p. Two activities were measured: in vitro DNA binding (indicated by + or – in the first column on the right) and in vivo transcriptional activation of the reporter gene construct (second column). (c) In this model of the Gal4 protein, derived from the deletion analysis, Gal4p has separable DNA-binding and transcription-activation domains joined by a flexible linker. (d) The DNA-binding domain, expressed alone, will bind DNA but will not activate transcription.

$UAS_{GAL}$  sequence (Figure 3a). The ability of each Gal4p-deletion construct to activate transcription of the  $lacZ$  gene was determined by measuring the activity of  $\beta$ -galactosidase (Figure 3b).

The in vitro DNA-binding activity of Gal4p was destroyed by a small deletion at the protein's N-terminus, but was not affected by small or large C-terminal deletions. Only the N-terminal 74 amino acid residues were needed for DNA-binding activity. Transcriptional activation required the DNA-binding region, as one would expect. Deletion of 60 residues from the C-terminus of Gal4p had little effect on gene activation. But deletion of 126 C-terminal residues reduced activation substantially, and a 191-residue C-terminal deletion completely eliminated activation. A large segment between these N- and C-terminal regions could be deleted without interfering with either activity.

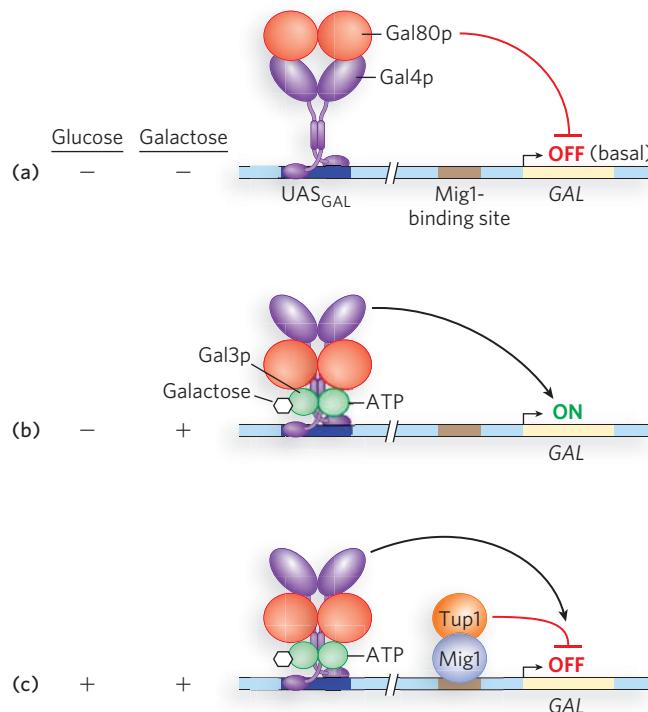
The findings suggested that the two activities inherent in Gal4p require 260 or fewer residues: 74 at the N-terminus and 191 at the C-terminus. This result was surprising, given that the entire Gal4 protein is 881 amino acids long. To confirm the result, the researchers spliced together DNA for the 74-residue N-terminal DNA-binding domain and various lengths of the C-terminal transcription-activation region. They found that a 217-residue protein, missing 664 amino acids between the two regions, restored full activity in both DNA-binding and transcription-activation assays!

Clearly, the ability of Gal4p to activate transcription is the result of two distinct and separable domains. Similar results were obtained with other transcription activators from several, diverse eukaryotes. Furthermore, examination of some transcription activators showed that the region between the two functional domains is highly sensitive to proteases, suggesting that the two domains are linked by sections of polypeptide that are open and flexible. These experiments gave rise to a model for some transcription activators, with two functional domains joined by a flexible linker (Figure 3c, d). The flexible region may help loosen the geometric constraints imposed by the DNA loop that forms between the transcription activator at an upstream binding site and the proteins it binds at the distant promoter. That the DNA-binding and transcription-activation domains of regulatory proteins can act independently has been demonstrated by “domain-swapping” experiments (see Figure 19-23).



**FIGURE 21-9** Regulation of *GAL* genes by the proteins Gal3p, Gal4p, and Gal80p. (a) Gal4p binds UAS<sub>GAL</sub>, but Gal80p binds Gal4p and prevents its activation of Pol II and the general transcription factors. (b) Galactose is a small-molecule effector for Gal3p, causing it to bind Gal80p and alter Gal80p conformation, which frees Gal4p to activate transcription.

through intracellular localization, which is regulated by phosphorylation. In the absence of glucose, Mig1 is phosphorylated and cannot enter the nucleus. Relegated to the cytoplasmic compartment, it is unable to bind DNA and repress the *GAL* genes. But when glucose is present, phosphorylation of Mig1 is blocked and it enters the nucleus, where it can bind DNA and associate with Tup1. Tup1 represses *GAL* gene expression by blocking transcription initiation, and possibly also by stimulating histone deacetylation at neighboring nucleosomes.

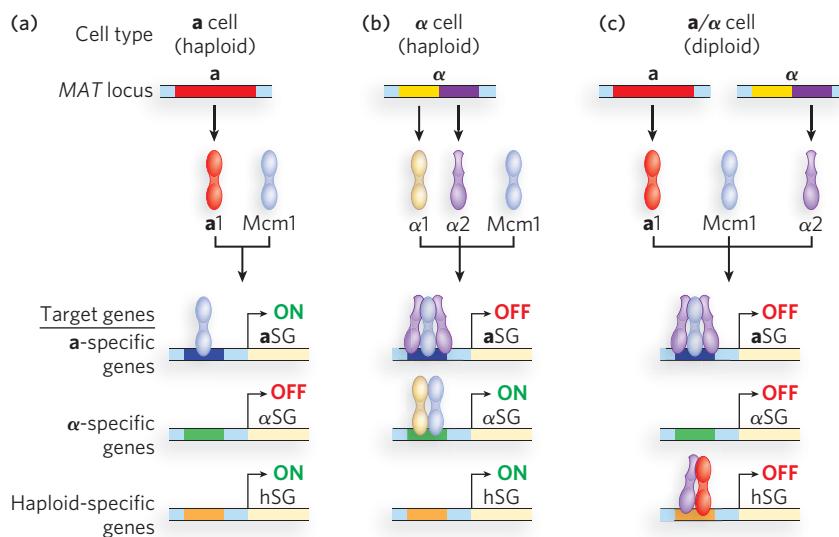


**FIGURE 21-10** Combinatorial control in global repression of yeast *GAL* genes. Expression levels of a *GAL* gene are shown under three different growth conditions, with (+) or without (–) glucose or galactose. (a) In the absence of glucose and galactose, Gal4p occupies UAS<sub>GAL</sub>, but the *GAL* gene is repressed by Gal80p. (b) In the presence of galactose and absence of glucose, Gal4p activates transcription of the *GAL* gene because Gal80p repression is relieved by binding of Gal3p. (c) In the presence of both glucose and galactose, glucose is the preferred carbon source; there is no transcription of the *GAL* gene because the Mig1-Tup1 complex represses its expression below basal levels.

### Yeast Mating-Type Switches Result from Combinatorial Control of Transcription

*Saccharomyces cerevisiae* (baker's yeast) can grow as either diploid or haploid cells, both of which reproduce by mitosis (see Model Organisms Appendix). The diploid cells contain two copies of each of the four yeast chromosomes, whereas haploid cells contain just one copy of each. When stressed by starvation, diploid cells can undergo meiosis to produce four haploid spores, two each of the mating types **a** and **α**. Haploid cells of the **a** mating type (**a** cells) can only mate with **α** haploids (**α** cells), and vice versa; thus, haploid cells display a simple sexual differentiation that is readily distinguishable when tested for mating ability.

Mating type is determined by the allele present at a single genetic locus, *MAT*. The identity of the allele at the *MAT* locus can switch as often as every cell division



**FIGURE 21-11 Combinatorial control of the yeast mating-type switch.** In all *S. cerevisiae* cells, haploid and diploid, McM1 is expressed and is used in combinatorial control. (a) The haploid **a** cell expresses protein **a1**, but this protein is used only in diploid cells. McM1 alone turns on **a**-specific genes (**aSG**). Other haploid-specific genes (**hSG**) are also expressed. (b) The haploid  $\alpha$  cell expresses the  $\alpha 1$  activator and  $\alpha 2$  repressor;  $\alpha 2$  associates with McM1 to turn off

**a**-specific genes, and  $\alpha 1$  binds McM1 to turn on  $\alpha$ -specific genes ( $\alpha SG$ ). Other haploid-specific genes are expressed. (c) Diploid cells express both  $a1$  and  $\alpha 2$ . Each, in conjunction with McM1, represses transcription of a set of genes:  $a1$ -McM1 represses haploid-specific genes, and  $\alpha 2$ -McM1 represses **a**-specific genes. Because  $\alpha 1$  is not expressed,  $\alpha$ -specific genes are also not expressed.

cycle. The mating-type switch occurs through site-specific recombination (see Chapter 14), to express either the *MAT $\alpha$*  allele or the *MAT $\alpha$*  allele. The *MAT $\alpha$*  allele encodes the  $\alpha 1$  protein, which directs transcription of  $\alpha$ -specific genes, and the *MAT $\alpha$*  allele encodes the  $\alpha 1$  and  $\alpha 2$  proteins, which stimulate transcription of  $\alpha$ -specific genes (Figure 21-11). After mating, the resulting diploid cells contain two *MAT* loci, one with the *MAT $\alpha$*  allele and the other with the *MAT $\alpha$*  allele, and the presence of both the  $\alpha$  and  $\alpha$  gene products directs the diploid-specific transcriptional program, while haploid-specific gene expression is turned off.

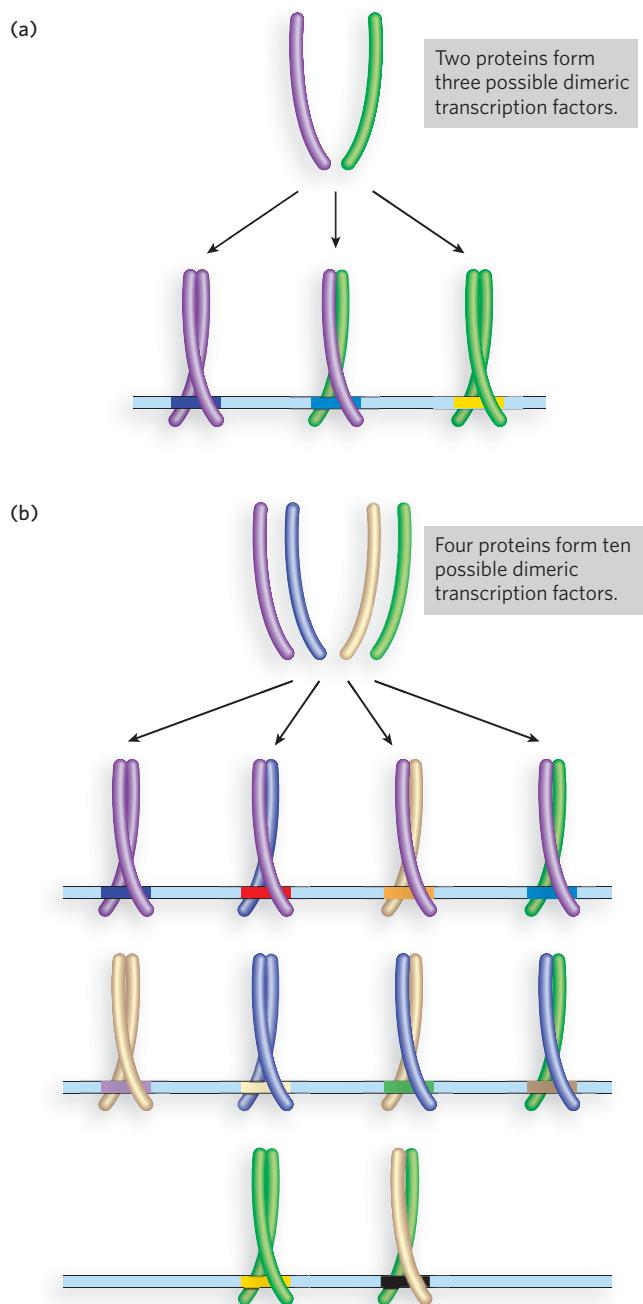
The transcriptional activation and repression of genes in each mating type is an example of combinatorial control, because control is achieved by combinations of regulators, at least one of which is common to the different cell types. In addition to the presence or absence of the  $\alpha 1$ ,  $\alpha 1$ , and  $\alpha 2$  proteins, specific activation and repression also involves McM1, expressed by both haploid cell types, as well as by diploid cells. In  $\alpha$  cells, McM1 binds the promoters of  $\alpha$ -specific genes and activates transcription. The genes specific to  $\alpha$  cells are turned off in  $\alpha$  cells because the  $\alpha 1$  activator is not present (Figure 21-11a). In  $\alpha$  cells, McM1 and  $\alpha 1$  interact to activate  $\alpha$ -specific gene transcription, while  $\alpha 2$  (in association with McM1) represses transcription of  $\alpha$ -specific genes (Figure 21-11b).

There are also genes specific to both haploid states, but on mating to produce a diploid cell, the haploid-specific genes are turned off. Repression of genes specific to all haploid cells is made possible by the interactions of  $\alpha 1$  and  $\alpha 2$  repressor proteins, which are only ever expressed together in diploid cells (Figure 21-11c).

### Combinatorial Mixtures of Heterodimers Regulate Transcription

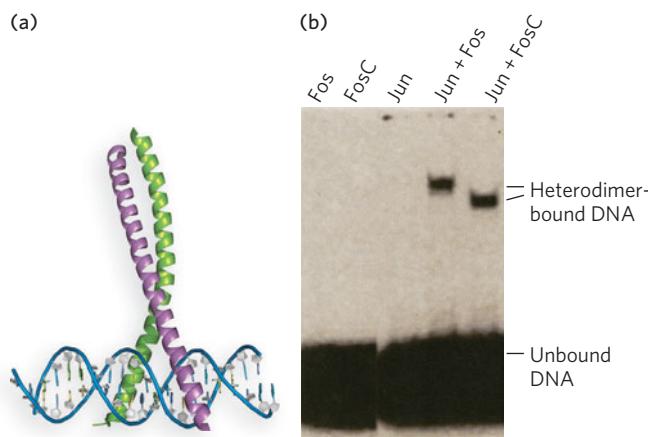
Like their bacterial counterparts, most eukaryotic transcription factors bind to DNA as homodimers. However, several types of eukaryotic transcription factors can form heterodimers of two different members of a family of similarly structured proteins, creating a larger number of functional transcription factors from a smaller number of individual proteins. For example, three possible dimers can form from just two similarly structured proteins: two homodimers and one heterodimer (Figure 21-12a). A hypothetical family of four different but structurally related proteins could form up to 10 different dimeric species (Figure 21-12b).

An example of proteins that behave in this fashion are the mammalian AP-1 transcription activators. AP-1 activators can be either homodimers or heterodimers, formed from subunits that belong to the family of



**FIGURE 21-12** Combinatorial control by heterodimer formation. (a) Two regulatory proteins that form homodimers and a heterodimer could form 3 different structures, which could bind 3 different regulatory sites. (b) Four proteins have the potential to form 10 different structures and bind 10 regulatory sites. The possible combinations increase dramatically as the number of potential dimerization partners increases.

proteins that includes Fos, Jun, and ATF. Gene regulation by AP-1 homodimers and heterodimers occurs in response to a variety of external stimuli, including growth factors, cytokines, and factors involved in stress



**FIGURE 21-13** AP-1 transcription factors. (a) Structural model of the AP-1 heterodimer of Fos (purple) and Jun (green), bound to DNA. (b) Gel from an electrophoretic mobility shift assay using a  $^{32}\text{P}$ -end-labeled DNA fragment containing the AP-1-binding site sequence. The DNA was mixed with Fos, or FosC (a fragment of Fos), or Jun (all of which would form a homodimer), or with a mixture of Fos and Jun or a mixture of FosC and Jun (both of which would form the two types of homodimer and the heterodimer). Reactions were analyzed by polyacrylamide gel electrophoresis, followed by autoradiography. Binding of protein dimers to the DNA causes the complex to migrate more slowly through the gel, resulting in distinct bands. The radioactive signal at the bottom of the gel is unbound DNA. [Sources: (a) PDB ID 1FOS. (b) T. Kouzarides and E. Ziff, *Nature* 336:649–651, 1988.]

and infection. Thus, AP-1 transcription factors control such important processes as cell proliferation, differentiation, and programmed cell death. Indeed, some members of the Fos and Jun protein family are encoded by proto-oncogenes, which are genes that promote tumor formation when overexpressed. In other words, alterations in one or more of the subunits that make up AP-1 can be fatal for the cell, or even the entire organism.

The protein-dimerization and DNA-binding regions of AP-1 family members are of the basic leucine zipper type. The crystal structure of the dimerization and DNA-binding portions of a Fos-Jun heterodimer bound to DNA is shown in Figure 21-13a. AP-1 dimers activate genes containing an AP-1-binding site. AP-1 variants bind to AP-1-binding sites with different affinities and activate gene transcription to different extents, depending on the composition of that AP-1. Figure 21-13b shows an electrophoretic mobility shift assay that examined DNA-binding affinity of Jun-Jun or Fos-Fos homodimers, as well as an AP-1 Fos-Jun heterodimer. A short DNA fragment containing an AP-1-binding site was end-labeled with  $^{32}\text{P}$ , then mixed

with either Fos, Jun, or the Fos-Jun heterodimer. The experiment also examined the binding affinity of a subfragment of Fos (FosC) that contains the DNA-binding and dimerization elements. The result shows that Fos, FosC and Jun do not bind appreciably to the AP-1-binding site on their own. However, the Fos-Jun or FosC-Jun heterodimers bind the AP-1 site much more tightly, such that they could be detected in this experiment.

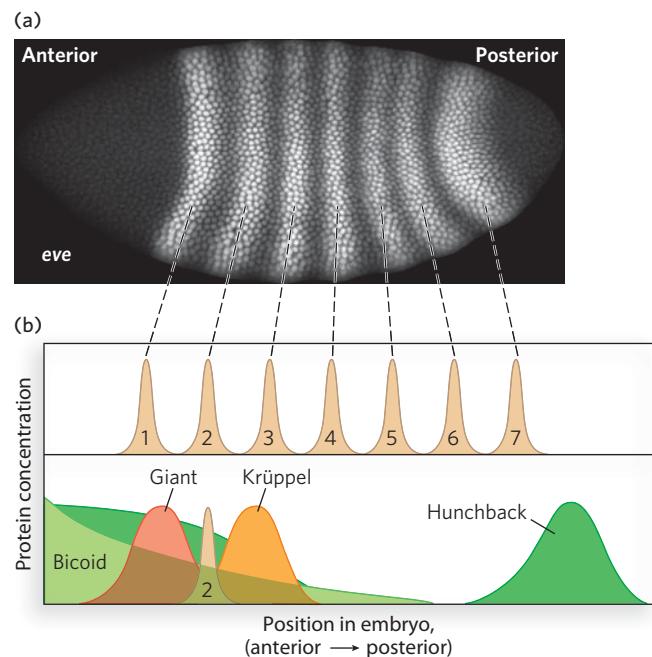
This differential DNA binding, depending on the composition of the AP-1 transcriptional control complex, represents another example of combinatorial control. Although many AP-1 variants contain transcription-activation domains, some lack them and instead function as transcription inhibitors. Thus, the effect of AP-1 can be varied by its composition, depending on the needs of the cell.

### Differentiation Requires Extensive Use of Combinatorial Control

A more complicated example of combinatorial control can be seen in body plan development in the fruit fly, *D. melanogaster*. Before it is released to become fertilized, the developing oocyte is surrounded by cells called nurse cells. The nurse cells secrete mRNAs encoding various transcription factors into the egg at specific locations, establishing concentration gradients of mRNA for the different transcription factors within the egg. During early embryonic development the nuclei divide quickly, producing 3,000 to 6,000 nuclei before plasma membranes form to delineate individual cells. When plasma membranes do form, the newly formed cells trap the specific mRNAs present at that particular position in the embryo. Each new cell thus produces a unique complement of transcription factors that act in a combinatorial fashion to express different proteins in the early embryo.

An example of combinatorial control by these unevenly distributed transcription factors is regulation of the *eve* gene, which produces a protein called even-skipped. Even-skipped is expressed only in specific cells of the embryo, generating a pattern of seven stripes that are visualized using a fluorescent antibody to even-skipped (Figure 21-14a). The *eve* gene is essential to development; the even-skipped product is a transcription activator that promotes further differentiation in the cells where it is expressed.

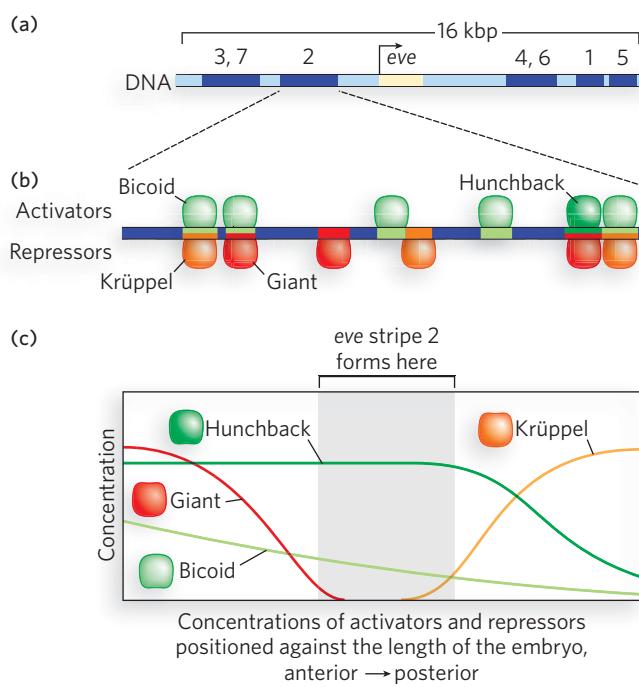
Expression of *eve* is controlled by the concentrations of four proteins translated from the original mRNAs deposited in the developing oocyte by the nurse cells. Two of these four proteins, Bicoid and Hunchback, are activators; the other two, Giant and Krüppel, are repressors. Different gradients of the mRNAs for these activators and repressors, established by the nurse cells,



**FIGURE 21-14** Combinatorial control of *eve* gene expression in fruit fly development. (a) This *Drosophila* embryo was stained with fluorescent antibodies that recognize the protein even-skipped (product of *eve*), showing its striped pattern of expression. (b) The graphs represent the relative levels and positions along the length of the embryo of even-skipped (top) and four transcription factors that regulate its expression (bottom). Specific combinations of transcription factors activate the *eve* gene. [Source: (a) Photo from A. V. Spirov and D. M. Holloway, *In Silico Biol.* 3:0009, 2003, Fig. 1. © 2002, Bioinformation Systems e.V.]

result in unique ratios of the four regulatory proteins in nearly every cell of the embryo. Expression of even-skipped occurs only in cells that have the proper ratio of the four proteins to activate *eve* (Figure 21-14b). But if *eve* were activated by only one particular ratio of protein concentrations, *eve* would be expressed in only one place in the embryo. How can the *eve* gene be expressed in seven different stripes? The striped pattern of *eve* expression is made possible by combinatorial control.

The *eve* gene has five different enhancers, each with a complex array of binding sites for transcription activators and repressors (Figure 21-15a). Only one enhancer needs to be active for *eve* to be expressed in a given cell. But if *eve* is to be expressed normally, all five enhancers need to be active (albeit in different cells). Each enhancer is activated by a different combination of transcription factors. Some activator and repressor sites overlap and are controlled by competition, whereas some repressor sites are distinct from activator sites and repress the gene at a distance (Figure 21-15b).



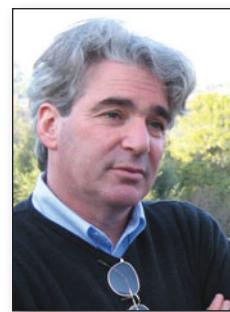
**FIGURE 21-15** Five independently acting enhancers of the *eve* gene producing seven stripes of *eve* expression in the early embryo. (a) The *eve* gene and its upstream and downstream enhancers, any one of which can activate *eve* expression if bound by the correct combination of transcription factors. Numbers 1 through 7 indicate the stripe(s) activated by each enhancer. (b) The binding sites in the stripe 2 enhancer for the Bicoid and Hunchback activators and the Krüppel and Giant repressors. (c) Changes in concentration of the four transcription factors along the length of the embryo, in the region that expresses *eve* stripe 2.

Seven stripes of even-skipped expression, each four cells wide, are formed because the local concentration of each activator and repressor is just right for activation of one of the five enhancers in particular cells across the length of the embryo. The same four transcription factors are used by the five different enhancers in different ways. The expression of the *eve* gene is an example of exquisite, complex combinatorial control!

The enhancer that activates *eve* expression in stripe 2 has been extensively studied in Michael Levine's laboratory. This enhancer is 500 bp long and contains binding sites for both repressors and activators (see Figure 21-15b). Both activators, Bicoid and Hunchback, must bind to their sites for gene expression to occur. Some binding sites for these activators overlap repressor-binding sites, and other activators bind DNA but are inactivated by repressors that bind within about 100 bp of the activators' binding sites. Increasing the distance between activator- and repressor-binding sites

of this type prevents repressor function. Although the mechanism of repression is unclear, it might occur through covalent modification of the activator. The region of the embryo that expresses *eve* in stripe 2 is largely deficient in both repressors, yet contains both activators (Figure 21-15c). Hence, the particular cells that express *eve* in stripe 2 do so because this is the only location in the embryo where the condition for activator binding in the absence of repressors is met. In all other cells, the stripe 2 enhancer is silent. Combinatorial control also governs formation of the other stripes expressing *eve*. The other *eve* enhancers contain different combinations and arrangements of the repressor- and activator-binding sites, such that each enhancer is only active in a narrow region of the embryo.

These examples of combinatorial control of transcription illustrate a central mechanism by which eukaryotic cells govern gene expression. Through the use of a relatively small number of regulatory proteins in each case, many different genes can be regulated either in concert or differentially, depending on the immediate needs of the cell. In this way, cells can respond quickly and appropriately to changing environmental conditions or to developmental changes, within the context of a tissue or an entire organism.



**Michael Levine**

[Source: Courtesy of Michael Levine.]

## SECTION 21.2 SUMMARY

- Eukaryotic transcription activators such as Gal4p have DNA-binding and transcription-activation domains.
- Eukaryotes make greater use of combinatorial control of gene expression than do bacteria. In combinatorial control, the same transcription factor is used in the regulation of more than one gene.
- Combinatorial control can be achieved in a variety of ways. Some transcription factors are formed from combinations of two different subunits that form heterodimers, each of which has different strengths as an activator. Or a gene has several enhancers, each of which uses a different combination of transcription factors.
- Mating-type switching in yeast is a classic example of combinatorial control. Unique sets of genes are expressed specifically in the **a** and **α** haploid states,

due to activation by specific regulatory factors. On mating to produce a diploid cell, the haploid-specific genes are repressed by the interactions of  $\alpha 1$  and  $\alpha 2$  repressor proteins, which are expressed together only in diploid cells.

- Body plan organization in *D. melanogaster* uses gradients of mRNAs for different transcription factors in the developing embryo. Different concentrations of transcription activators and repressors control where the gene *eve* is activated, to produce seven stripes that influence cell differentiation.

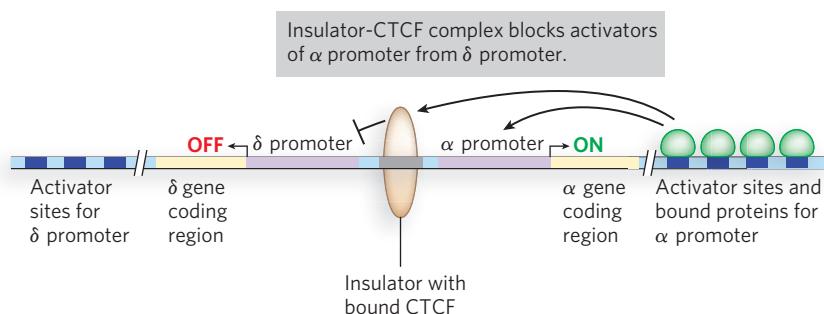
## 21.3 Transcriptional Regulation Mechanisms Unique to Eukaryotes

Gene regulation is necessarily more complex in eukaryotes than in bacteria, as a consequence of some of the key differences between these two domains of organisms. The larger eukaryotic genomes mean there will be more nonspecific DNA binding and more genes to regulate. And the multicellular nature of most eukaryotes requires mechanisms for development and intercellular communication, the formation and function of intracellular compartments, and the speedy control of gene expression as cells grow and change. We now turn to a discussion of some regulatory processes that are necessary to deal with the complexity of the eukaryotic genome. Such gene control mechanisms provide for situations that arise only in eukaryotes, such as the need to regulate gene dosage in diploid cells.

### Insulators Separate Adjacent Genes in a Chromosome

The use of multiple transcription factors to control eukaryotic genes requires dispersed binding sites. Some enhancers are located well over 1,000 bp from the promoters they regulate, or they can be within the gene or at the noncoding 3' end of the gene. This is quite different from the situation in bacteria, where control elements are almost always located close to, or overlap with, the promoter. As discussed earlier, DNA looping accommodates the large distances between enhancer and promoter elements in eukaryotes. Indeed, the large size of DNA loops provides the flexibility that enables enhancers to function in either orientation. However, this raises a new question: what prevents the enhancer for one gene forming a loop to interact with the promoter of another gene, thereby activating the wrong gene? In part, misregulation of this type is prevented in eukaryotes by **insulators** (sometimes referred to as boundary elements), DNA sequences that form boundaries between genes or groups of genes.

Insulators are relatively short sequences, sometimes fewer than 50 bp. Exactly how insulators function is still unknown, and it is likely that they have a variety of functions. An example of an insulator in T cells (T lymphocytes, white blood cells involved in the immune response) is shown in Figure 21-16. T cells must express either the  $\alpha$  chain or the  $\delta$  chain of the T-cell receptor, but not both. The promoters for these genes are adjacent in human cells, but an insulator sequence between them prevents the enhancer region for one gene from activating transcription from the promoter of the other gene. Insulators can also prevent packaging of a gene into heterochromatin. When a gene is experimentally inserted into a heterochromatic region of a



**FIGURE 21-16** Insulator regulation of the expression of T-cell receptors. With DNA looping over long distances, binding at enhancers could activate the wrong promoter. The insulator confines the action of enhancers to their matching promoter. In the regulatory regions of the genes shown here,

for the  $\alpha$  and  $\delta$  chains of the T-cell receptor, the insulator prevents activation of the  $\alpha$  promoter by the  $\delta$  enhancer, and of the  $\delta$  promoter by the  $\alpha$  enhancer. CTC-binding factor (CTCF) binds to insulators, although how it functions is still uncertain.

chromosome, it is typically repressed. But if the gene also contains an adjacent insulator, the gene remains active even after insertion into a heterochromatic region. An insulator can have a repressive effect if located between the promoter and enhancer of a gene.

All insulator sequences in higher eukaryotes require for their functioning CTC-binding factor (CTCF) (see Figure 21-16). CTCF was first identified as a protein that binds to a site containing a 5'-CCCTC sequence, from which the protein derives its name. The binding site for CTCF is much longer than this sequence, but the flanking sequences are divergent and not as easy to recognize. CTCF contains 11 zinc fingers and binds a diverse set of DNA sequences up to 50 bp long. All insulators seem to require CTCF binding, although the mechanism of the insulators' function and the role of CTCF are currently unknown. CTCF might recruit other proteins to the insulator.

### Some Activators Assemble into Enhanceosomes

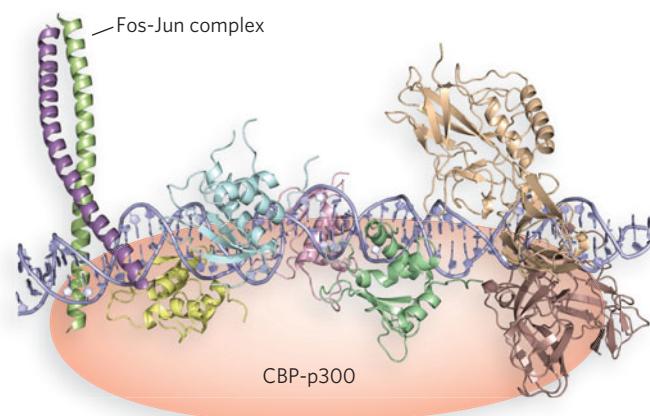
Exquisite transcriptional regulation can be achieved through the interaction of multiple activators at a single gene. In some cases, cooperating activators form a stable, tightly folded nucleoprotein complex called an **enhanceosome**, which integrates regulatory information from multiple signaling cascades and generates a single transcriptional outcome at the target promoter. An example of an enhanceosome can be seen in the regulation of the gene for interferon beta (IFN $\beta$ ). Interferons

are produced in response to a viral infection and lead to programmed cell death, thereby halting further production of viral particles and infection of surrounding cells. At the IFN $\beta$  promoter, multiple activators can present their activation domains together to simultaneously interact with the cofactor protein complex CBP-p300 (Figure 21-17). Recruitment of this cofactor is most efficient only when all of the activators in the enhanceosome are present together. The placement of each activator-binding site in the DNA is critical, because of the three-dimensional structure required for enhanceosome function. For example, the experimental insertion of 5 bp (i.e., a half-turn of the helix) between regulatory elements inactivates the function of the IFN $\beta$  enhanceosome.

Similar activator clusters can also function together to repress transcription, and it is possible that an enhanceosome can switch to a repressor function under different cellular conditions. Enhanceosomes tend to form at genes that need to be tightly regulated in pathways important to the organism's defense system, such as wound healing and antipathogen mechanisms.

### Gene Silencing Can Inactivate Large Regions of Chromosomes

Thus far we have focused on the activation or repression of gene expression through the action of activator or repressor proteins at single promoters or enhancers. In some cases, however, the position of a gene within the chromatin, or its location on a particular



**FIGURE 21-17 Hypothetical structure of the IFN $\beta$  enhanceosome.** The IFN $\beta$  enhancer is referred to as an enhanceosome because, unlike other enhancers, it requires accurate spacing and helical phasing of the DNA between several protein-binding sites. This requirement indicates that a specific tertiary structure is formed with the various regulatory

proteins. HMG proteins, while not part of the completed complex, facilitate enhanceosome formation by helping bend the DNA (see Figure 21-6). The Fos-Jun complex is shown on the left, bound to DNA; individual regulatory proteins in the complex are indicated by different colors. [Source: Adapted from PDB ID 2O6G, 2O6I, and 1T2K.]

chromosome, leads to almost complete repression of gene expression. This **gene silencing**, a powerful mode of regulation in eukaryotes, is the absence of gene expression due to the location of the gene, rather than its response to the presence or absence of a regulatory factor or complex. As a result, silencing can encompass relatively large segments of a chromosome, or an entire chromosome, thus controlling the expression of many genes at once.

Recall that chromatin is organized into heterochromatin and euchromatin. The loosely packed euchromatin is often transcriptionally active, whereas the densely packed heterochromatin is transcriptionally silent. Heterochromatin is often found at centromeres and telomeres, as well as other inactive parts of the genome. Experiments have shown that genes normally active in a region of euchromatin become transcriptionally silent when moved into heterochromatic regions. These observations led to the conclusion that a primary function of heterochromatin is to maintain certain parts of the genome in a transcriptionally inactive state by preventing access of the transcription machinery to these chromatin regions. Indeed, burying a gene within heterochromatin may be a preferred mechanism for long-term silencing.

The formation of heterochromatin requires many different factors, depending on the specific region of the chromosome. For example, a mechanism known as gene dosage compensation, occurring in the cells of female mammals, involves the formation of heterochromatin over an entire X chromosome, inactivating it (as we discuss in more detail below). Recent studies of heterochromatin in other regions of a chromosome show that small nuclear RNAs (snRNAs) are required for heterochromatin formation, along with certain proteins and histone modifications. Highlight 21-3 describes studies of heterochromatin in the centromere region of yeast chromosomes that reveal a role of the silencing machinery mediated by small RNAs.

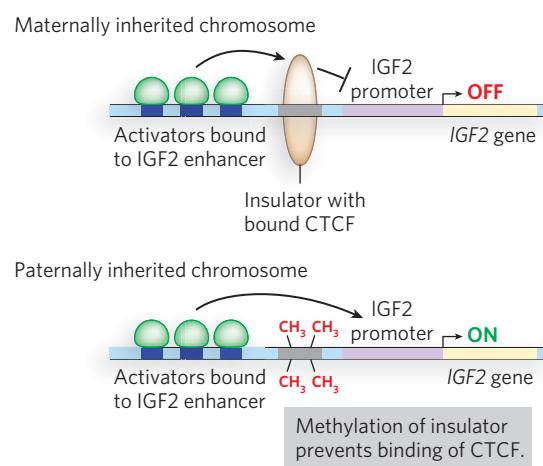
### Imprinting Enables Selective Gene Expression from One Allele Only

In most diploid cells, both homologous genes (i.e., both alleles of a gene) are expressed equally. One allele may be dominant over the other, as Mendel found in his work on the garden pea, or the two alleles may deviate from Mendelian behavior and both may contribute to the phenotype. For example, we learned in Chapter 2 that both alleles of the genes responsible for human blood type, when expressed together, can lead to type AB (see Figure 2-5). But regardless of the phenotype, both alleles of a gene are usually expressed in the

diploid cell. Some higher eukaryotes, however, have mechanisms to completely shut down the expression of an allele derived from one parent, in a process called **imprinting**. Because only one parental allele of an imprinted gene is repressed, the usual rule of equal expression of each allele in a diploid cell is violated.

Imprinting is typically restricted to mammals, although some examples have been found in flowering plants. There are currently about 80 genes in the human genome that are known to be imprinted. Imprinting occurs during development of the gametes. A set of genes is imprinted during oocyte development, and a different set of genes is imprinted during sperm cell development. Nearly every cell of the offspring has the gene expression pattern dictated by the imprinted genes from the egg and sperm. The sole exception is in cells that will give rise to gametes. All imprinting information from the parents is lost during development of the germ cells. Developing gametes adopt the imprinting pattern specific to the sex of that individual organism.

Imprinting of a gene is not based on the DNA sequence; it is an epigenetic process (see Chapter 10). Epigenetic marks are created by nucleosome modification patterns and DNA methylation. Figure 21-18 shows imprinting based on DNA methylation for the mammalian gene for insulin growth factor-2 (IGF-2). In this case, DNA near the paternally inherited *IGF2* allele becomes methylated (imprinted) and is active, but DNA near the maternally inherited allele is unmethylated



**FIGURE 21-18 Imprinting of the mammalian *IGF2* gene.**

When CTC-binding factor (CTCF) binds the insulator of *IGF2* (top), this enables insulator function and turns off *IGF2*, because the insulator is located between the promoter and the enhancer. When the insulator sequence is methylated (bottom), CTCF can no longer bind the DNA, and *IGF2* can be activated by its enhancer.

## HIGHLIGHT 21-3 A CLOSER LOOK

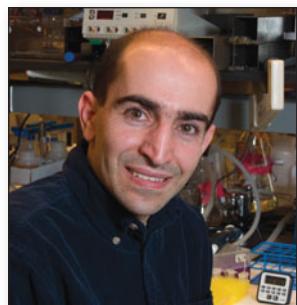
### Gene Silencing by Small RNAs

Large tracts of DNA are encased in heterochromatin, a form of DNA that is so compact that its genes are silenced by the exclusion of RNA polymerase. The nucleosomes in heterochromatin have a distinctive histone modification pattern of methylation and hypoacetylation compared with actively transcribed regions of DNA (euchromatin). The epigenetic marks (marks that are inheritable but do not occur in the DNA) result in stable inheritance of the silent heterochromatin state.

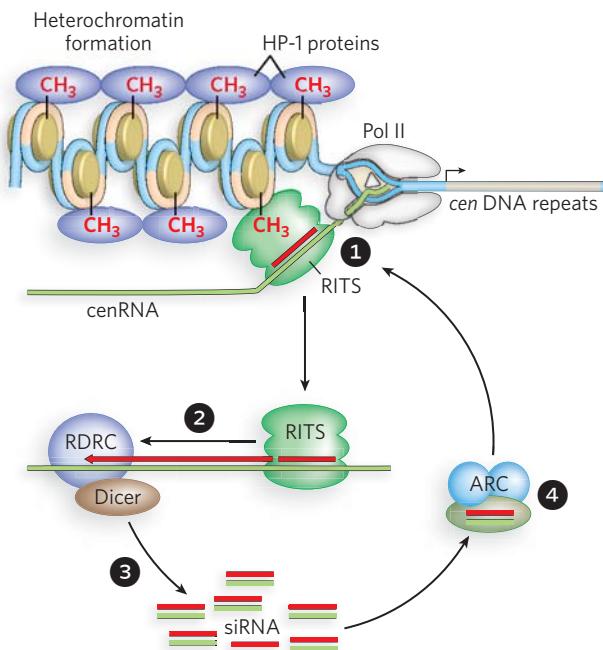
Research from Danesh Moazed's laboratory shows that the formation of heterochromatin in fission yeast (*Schizosaccharomyces pombe*) requires the RNA silencing machinery, which localizes to heterochromatin nucleation sites. To study the process of heterochromatin formation, the researchers purified the RNA silencing complex of *S. pombe* called RITS (RNA-induced transcriptional silencing), and also identified a second complex that interacts and functions with RITS. The second complex is referred to as RDRC (RNA-directed RNA polymerase complex), because it contains an RNA-directed RNA polymerase, Rdp1. RDRC also contains a helicase known as Hrr1, and a putative poly(A) polymerase known as Cid12. A hypothesis for the process of heterochromatin formation at the centromere, based on the possible functions of these complexes, is shown in Figure 1.

The RITS complex contains siRNAs, derived from heterochromatic regions of the chromatin, and the protein Argonaute, belonging to a family of proteins implicated in RNA-induced silencing pathways. RITS matches the siRNAs to complementary RNA sites known as cenRNA sequences, which are noncoding transcripts produced at repeat sequences (*cen* DNA) near the centromeres (see Figure 1, step 1). The RITS complex also contains a chromodomain protein, which

binds methylated histones of heterochromatin and probably helps target RITS to heterochromatin. When RITS associates with a growing cenRNA transcript, Rdp1 within the RDRC complex uses the siRNA-cenRNA as a primer to synthesize double-stranded RNA (step 2). Dicer then chops the double-stranded RNA into multiple siRNAs



**Danesh Moazed** [Source: Steve Gilbert, Studioflex Productions.]



**FIGURE 1** Proposed model for heterochromatin formation at the centromere in *S. pombe*; the steps are described in the text. [Source: Adapted from M. Buhler and D. Moazed, *Nat. Struct. Mol. Biol.* 14:1041–1048, 2007 (online, November 5, 2007), Fig. 3.]

(step 3). Moazed and colleagues identified another complex, called ARC (Argonaute chaperone), that contains Argonaute and two other protein components. ARC binds individual siRNAs and is thought to chaperone them into the RITS complex (step 4). During this step, the siRNA base-pairs to complementary sequences in heterochromatin, recruiting RITS and completing a self-propagating cycle that expands the heterochromatic region.

Methylation of histone H3 at Lys<sup>9</sup>, initiated by RNA silencing complexes, can help expand the heterochromatin. The methylation is associated with recruitment of HP-1 protein, which is required for heterochromatin formation. The poly(A) polymerase Cid12 belongs to a family of proteins that target RNAs to the exosome, a large ribonuclease complex that catalyzes RNA degradation. Hence the Cid12 component of RDRC is thought to add another layer of RNA surveillance to ensure that heterochromatic regions are completely silenced.

Although this system was elucidated in a single-celled eukaryote, we know that the *S. pombe* proteins identified in the processes described here are conserved in higher eukaryotes. Perhaps a similar mechanism of siRNA-mediated heterochromatin formation occurs in mammals.

and is not expressed in the offspring. Imprinting of this gene is important, because the expression of both alleles tends to result in cancer. The mechanism of imprinting in this case involves an insulator sequence. The *IGF2* gene contains an insulator sequence between its promoter and an enhancer, rather than between the gene and an adjacent gene. Thus when the insulator is active, *IGF2* is repressed because the activating effect of the enhancer is insulated from the promoter.

The biochemical explanation for imprinting in *IGF2* is thought to lie in 5-methylation of C residues of CpG sequences (see Figure 6-34a). Cytosine methylation in eukaryotic DNA is generally associated with gene repression, whereas transcriptionally active regions of DNA tend to be undermethylated. In the case of *IGF2*, cytosines in CpG sequences recognized by CTCF in the insulator sequence that regulates the paternal *IGF2* gene become methylated when the gene is imprinted, thereby inactivating *IGF2*. When CpG sites are methylated, CTCF can no longer bind the insulator. Hence, when the insulator in *IGF2* is methylated during sperm development, the paternally inherited copy of *IGF2* is expressed because CTCF no longer binds the insulator, and the enhancer activates the promoter. The insulator site is not methylated in the egg, so CTCF binds the insulator and represses the maternally inherited copy of *IGF2*.

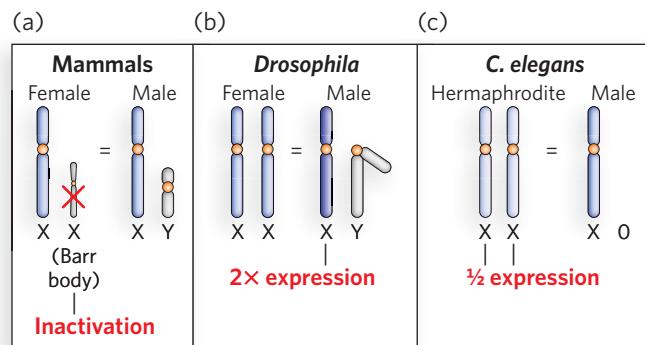
Imprinting is essential for development in mammals. Studies in mice have shown that a genetically engineered egg containing two complete sets of maternally inherited chromosomes will not develop past the blastula stage. The same is observed for eggs containing two complete sets of chromosomes from sperm. In either of these situations, the alleles of the genes that are normally differentially marked will have identical imprinting patterns in the developing egg—leading to either no expression or double expression of those genes. For example, if an embryo were to develop from a fully diploid cell in which the two chromosome sets were derived from a female, both *IGF2* alleles would be inactive. Imprinting therefore underlies why mammals are not capable of **parthenogenesis**—the development of an embryo with a diploid genome that is entirely maternally derived.

It is possible that imprinting evolved by natural selection to increase the fitness of the organism. Imprinting in mammals is thought to confer certain behavioral traits that enhance fitness. One hypothesis proposes that imprinting reflects the different interests of the mother and father in the growth and development of offspring. The father is more interested in seeing that the offspring grow rapidly, regardless of whether this occurs at the expense of the mother. The mother is more interested in balancing her own survival with

sufficient nourishment of the offspring, and thus tends toward growth-limiting measures that conserve resources. In support of this “parental conflict” hypothesis, male-expressed imprinted genes tend to promote growth, and female-expressed imprinted genes tend to limit growth. The model is also supported by the lack of imprinting in animals, such as birds, that have lower requirements for raising offspring.

### Dosage Compensation Balances Gene Expression from Sex Chromosomes

Diploid organisms carry two copies of each autosomal chromosome, but the sex chromosomes are unequal in copy number in females and males. In mammals, females have two copies of the X chromosome, and males have one X chromosome and a Y chromosome. The X chromosome carries genes that are required by both males and females, and the gene products are required in the same amounts in both sexes. **Dosage compensation** mechanisms have evolved to control the level of gene expression from these chromosomes so that the levels are similar in males and females. We can imagine three different ways that gene dosage compensation could occur. (1) Total inactivation of one X chromosome in the female would make gene expression equal to that from the single X chromosome in the male. (2) Expression of the single X chromosome in the male could be doubled. (3) Expression of the two X chromosomes in the female could be halved. All three of these mechanisms are employed in one species or another (Figure 21-19).

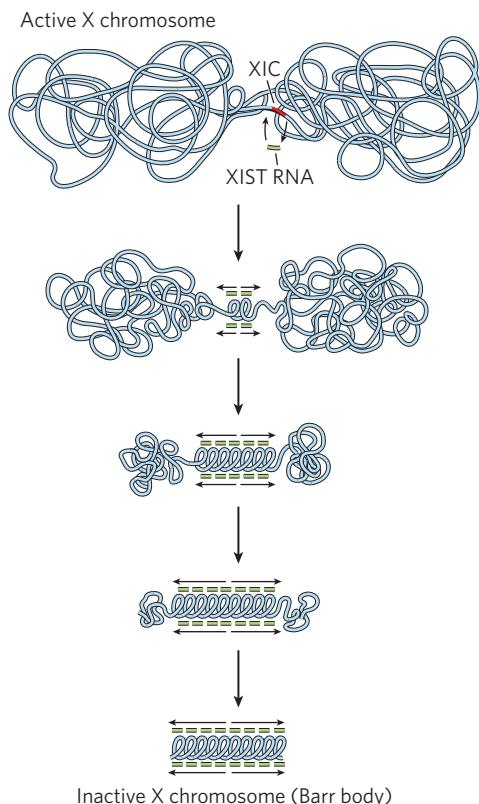


**FIGURE 21-19** Three mechanisms of gene dosage

**compensation.** Different dosage compensation mechanisms balance the level of gene expression from the X chromosomes in male and female cells. (a) In mammals, one X chromosome of the female (XX) is inactivated, forming a compact structure called a Barr body. (b) In *Drosophila*, the single X chromosome of the male (XY) is transcribed at twice the level of the X chromosomes in the female (XX). (c) In *C. elegans*, the two X chromosomes of the hermaphrodite (XX) are transcribed at half the level of the single X chromosome of the male (XO).

In mammals, one X chromosome of female cells is inactivated, compacted into a tightly condensed structure called a **Barr body**. This process of **X chromosome inactivation** starts at the **X inactivation center (XIC)**, a region of about  $10^6$  bp near the middle of the X chromosome, which condenses into heterochromatin. Condensation spreads from this nucleation point in both directions until the entire X chromosome is compacted (Figure 21-20). The process involves XIST, an RNA produced from the XIC DNA in the inactivated chromosome. XIST is not translated into protein, but instead coats the chromosome non-sequence-specifically at the XIST locus, and then spreads in both directions. In addition to XIST, X chromosome inactivation in mammals also involves a histone variant, macroH2A (see Highlight 10-1).

In humans, inactivation occurs on only one of the two X chromosomes in each cell and thus seems to be a



**FIGURE 21-20 X chromosome inactivation in mammals.** Inactivation of one X chromosome in each cell of the female mammal begins at the XIC locus, which encodes XIST (a non-protein-coding RNA), and also requires the histone variant macroH2A and other chromatin regulatory factors. XIST coats the X chromosome starting at the nucleation site, then spreads in both directions until the entire chromosome is coated, resulting in repression of nearly all the genes on the chromosome.

type of imprinting. However, the selection of which X chromosome becomes inactive is random; either the maternal or the paternal X chromosome is inactivated in any given cell. Therefore, this X chromosome inactivation is not strictly imprinting. There are some tissues in which X chromosome inactivation is a true example of imprinting in humans. This occurs in the umbilical cord and placental tissue, where only the paternal chromosome is inactivated.

In *Drosophila*, male cells contain one X chromosome and solve the gene dosage problem in the opposite way: genes on the male's X chromosome are transcribed at twice the level of genes on the female's two X chromosomes. This overactivation arises from coating of the chromosome with an X-encoded RNA-protein complex called the **dosage compensation complex (DCC)**, which contains five different proteins and two noncoding RNAs. Two of the proteins have HAT and phosphokinase activities. The DCC may enhance the transcription of genes on the male's X chromosome by modifying chromatin through these two enzymatic activities.

The two sexes in the nematode *C. elegans* are the hermaphrodite (XX) and the male (X0). Expression from each of the two X chromosomes in the hermaphrodite is down-regulated by half relative to gene expression from the single X chromosome of the male. Like *Drosophila*, *C. elegans* has a dosage compensation complex, but here its function is to repress gene transcription instead of activating it. The DCC is expressed only in cells with two X chromosomes (i.e., in the hermaphrodite), where it coats both X chromosomes. The *C. elegans* DCC is composed of proteins that resemble the condensing complex of mitotic and meiotic chromosomes. Thus, evolution has modified and recruited these chromosome-condensing proteins to the task of gene dosage compensation. The DCC partially condenses both X chromosomes such that transcription is reduced by half, balancing the gene dosage with that of the male cell with its single X chromosome. Targeting of the *C. elegans* gene dosage complex to X chromosomes is accomplished by DNA sequence elements dispersed along the X chromosome that nucleate the complex. Nucleation is followed by spreading of the complex across the entire chromosome, maintaining the repressed epigenetic state throughout the life of the hermaphrodite.

## Steroid Hormones Bind Nuclear Receptors That Regulate Gene Expression

Regulatory changes in response to hormones are an important aspect of eukaryotic transcriptional regulation. Intercellular communication is essential in

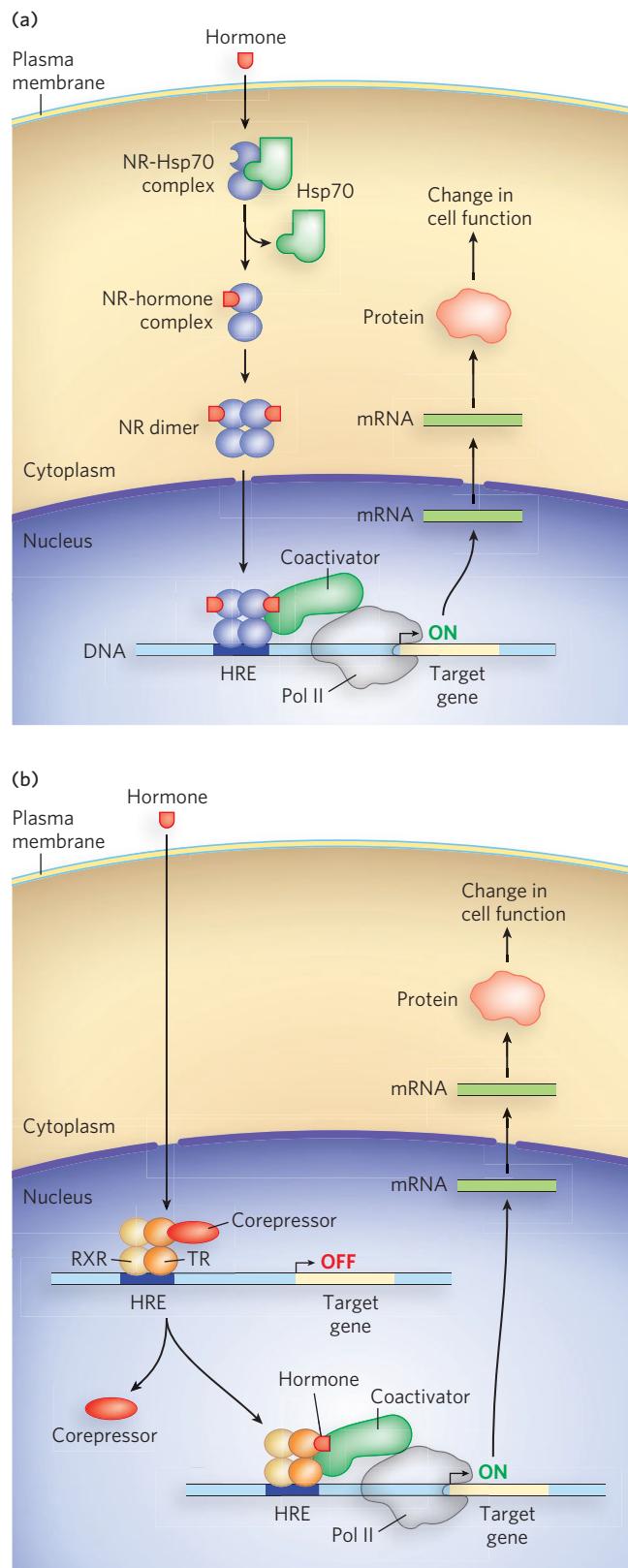
multicellular organisms; tissues, organs, and organ systems need to work together and must be able to respond to external signals. One group of molecular signals is the steroid hormones, which operate in the nucleus to activate transcription of particular genes in response to tissue or system requirements.

The effects of steroid hormones (and thyroid and retinoid hormones, which have the same mode of action) provide well-studied examples of the modulation of eukaryotic regulatory proteins by direct interaction with molecular signals. Steroid hormones too hydrophobic to dissolve readily in the blood (e.g., estrogen, progesterone, and cortisol) travel on specific carrier proteins from their point of release to their target tissues, where the hormone enters cells by simple diffusion and binds to its specific nuclear receptor protein, which is a transcription activator.

There are two major types of steroid-binding nuclear receptors: those initially located in the cytoplasm (type I) and those always located in the nucleus, bound to DNA (type II). Examples of steroid hormones that bind type I nuclear receptors are estrogen, progesterone, androgens, and glucocorticoids. The action of type I nuclear receptors is shown in **Figure 21-21a**. The receptor is initially bound to a heat shock protein (Hsp70) in the cytoplasm, keeping the receptor in its monomeric state. On binding the steroid hormone, Hsp70 dissociates, the receptor dimerizes and exposes a nuclear import signal, and the receptor-hormone complex migrates into the nucleus, where it acts as a transcription factor.

Type II receptors also require binding of the hormone before they activate transcription, but these receptors are already bound to the DNA, whether their molecular signal (steroid hormone) is present or not. In addition, type II receptors typically bind DNA as a heterodimer. Thyroid hormone receptor (TR) is an example

of a type II nuclear receptor. TR forms a heterodimer with a protein known as the retinoid X receptor (RXR, another type II receptor) and, in the absence of thyroid



**FIGURE 21-21** Steroid hormone receptor action. Steroid hormones diffuse across the plasma membrane and associate with a type I or type II nuclear receptor. (a) The type I nuclear receptor (NR), located in the cytoplasm, is complexed with a heat shock protein (Hsp70). Hormone binding releases Hsp70, and the NR dimerizes and exposes a nuclear import signal sequence. The NR-hormone complex then enters the nucleus and binds to a hormone response element (HRE) to activate transcription. (b) Type II nuclear receptors are bound to DNA whether or not the hormone signal is present. For example, the thyroid hormone receptor (TR) forms a heterodimer with the protein RXR to bind the HRE, but it is inactive without thyroid hormone. When the hormone enters the cell and the nucleus and binds the complex at the HRE site, it activates gene transcription.

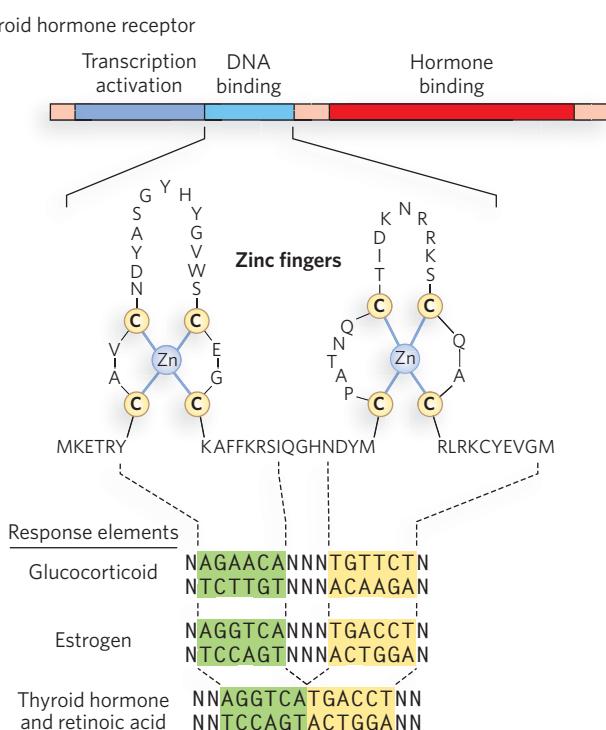
hormone, also binds a third protein, a corepressor. This complex does not activate transcription. When thyroid hormone enters the nucleus and binds its receptor, this releases the corepressor and promotes binding of a co-activator, resulting in recruitment of Pol II. Expression of the thyroid hormone-induced genes produces proteins involved in metabolism and regulation of heart rate.

Steroid hormone–nuclear receptor complexes act by binding to highly specific DNA sequences known as **hormone response elements (HREs)**. The bound hormone-receptor complex can either enhance or suppress the expression of adjacent genes. The HREs for the various steroid hormones are similar in length and organization in the genome, but differ in sequence. Each receptor has a consensus HRE sequence to which the hormone-receptor complex binds well (**Figure 21-22**). The consensus sequence consists of two six-nucleotide sequences, either contiguous or separated by three nucleotides, in tandem or inverted with

respect to each other. The steroid hormone receptors have a highly conserved DNA-binding domain with two zinc fingers. The hormone-receptor complex binds to the DNA as a dimer, and the zinc finger domains of each monomer recognize the six-nucleotide HRE sequences. The ability of a given hormone to act through its hormone-receptor complex to alter the expression of a specific gene depends on the exact sequence of the HRE, its position relative to the gene, and the number of HREs associated with the gene.

The ligand-binding region of the steroid hormone receptor protein—always at the C-terminus—is specific to the particular receptor. For example, the ligand-binding region of the glucocorticoid receptor shares only 30% sequence similarity with the estrogen receptor, and only 17% similarity with the thyroid hormone receptor. The size of the ligand-binding region also varies dramatically; the vitamin D receptor has only 25 residues, whereas the mineralocorticoid receptor has 603 residues in this region. Mutations that change one amino acid in the ligand-binding region can result in loss of responsiveness to a specific hormone. In humans, medical conditions resulting from the inability to respond to cortisol, testosterone, vitamin D, or thyroxine are caused by mutations of this type.

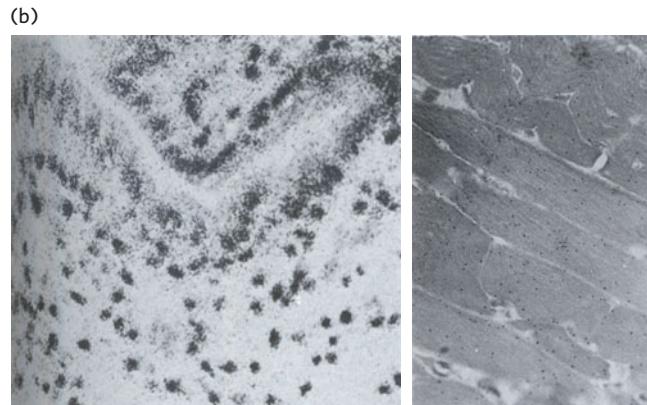
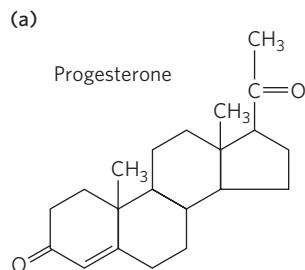
Responsiveness to a steroid hormone is tissue-specific. The specificity is due in part to transcriptional regulation of the gene encoding the hormone receptor. Cells that are not responsive to a particular steroid hormone do not seem to express its receptor. This can be seen experimentally by using radioactive steroid hormone to examine different tissues for accumulation of the hormone in the nucleus. For example, radioactive progesterone accumulates in the nuclei of endometrial cells, which prepare the uterus for pregnancy, but not in the nuclei of other tissues such as muscle (**Figure 21-23**).



**FIGURE 21-22 Structural organization of steroid hormone receptors and hormone response elements.** Nuclear receptors are multidomain proteins (top, showing the three domains) that bind steroid hormones and DNA to activate transcription. As shown in the enlarged structure of the DNA-binding region (middle), two adjacent zinc fingers bind to the HRE in the DNA (bottom; the binding regions are indicated by dashed lines). The receptors bind the DNA as dimers. The HREs of several steroid hormone receptors are indirect repeats (highlighted in yellow and green).

### Nonsteroid Hormones Control Gene Expression by Triggering Protein Phosphorylation

Nonsteroid hormones, grouped together as chemically distinct from steroid hormones, cannot cross the plasma membrane. Instead, they deliver their regulatory message via a cell surface receptor. We saw in Chapter 19 how the effects of insulin on gene expression are mediated by a series of steps leading ultimately to activation of a protein kinase that phosphorylates specific DNA-binding proteins. Phosphorylation alters the ability of the proteins to act as transcription factors (see Highlight 19-1). This general mechanism mediates the effects of many nonsteroid hormones on gene regulation.



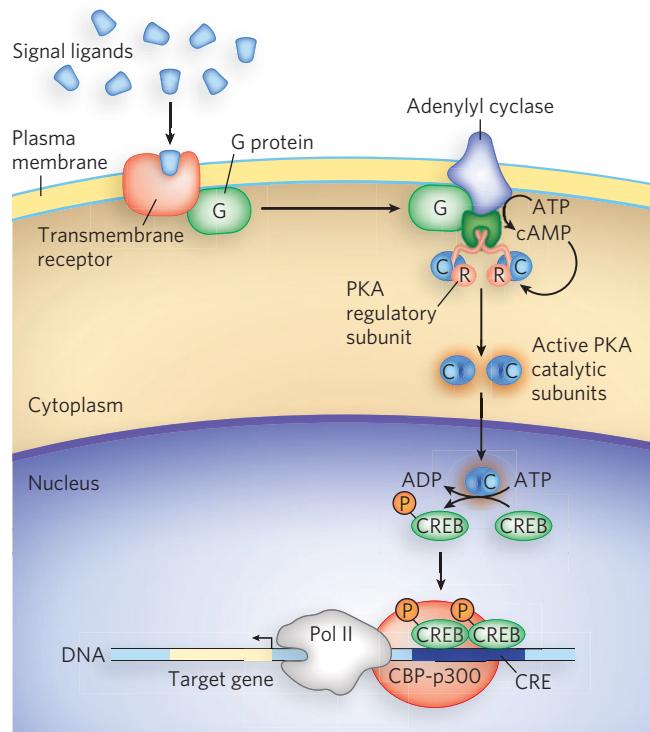
**FIGURE 21-23 Tissue-specific response to the steroid hormone progesterone.** (a) The chemical structure of progesterone. (b) Autoradiograph of endometrial cells of a guinea pig uterus (left), 15 minutes after the animal was injected with radioactive progesterone; dark spots indicate concentrations of radioactive progesterone in nuclei. Other tissues isolated from the animal (muscle cells of the diaphragm; right) do not show evidence of radioactive progesterone in cell nuclei. [Source: (b) M. Sar and W. E. Stumpf, *Endocrinology*, 94(4) : 1116–1125, 1974. Courtesy of Madhabananda Sar.]

A widely used mechanism of signal transduction for many nonsteroid hormones and other ligands involves the action of **G protein-coupled receptors (GPCRs)** that span the plasma membrane. In this pathway, a transmembrane GPCR binds the signal molecule on the outside of the cell, and binding activates a guanine nucleotide-binding protein (G protein) on the cytoplasmic side of the membrane. G proteins function as molecular switches: when bound to GTP they are active; on hydrolysis of the GTP to GDP they are inactive. When activated, G proteins promote the phosphorylation of proteins that in turn activate gene transcription.

The many different types of nonsteroid ligands that function through GPCRs include olfactory molecules such as odorants and pheromones; peptide hormones such as insulin, calcitonin, follicle-stimulating hormone, gonadotropin-releasing hormone, neurokinin,

thyrotropin-releasing hormone, and oxytocin; neurotransmitters such as dopamine, epinephrine (adrenaline), norepinephrine (noradrenaline), serotonin, and acetylcholine; glucagon; prostaglandins; and leukotrienes. Ligand binding to GPCRs triggers a wide variety of physiological processes. For example, serotonin and dopamine act through GPCRs in the mammalian brain to regulate mood and behavior. Glucagon and prostaglandins bind to GPCRs to trigger changes in metabolism and contraction of smooth muscle. The malfunctioning of GPCRs is associated with a range of human diseases, and they are the target of more than 25% of pharmaceuticals used in medicine.

A G protein-coupled pathway is shown in **Figure 21-24**. First, a signal ligand binds to and activates a surface receptor—the GPCR—that spans the plasma membrane. The signal is then transduced through a



**FIGURE 21-24 Gene expression regulated by protein phosphorylation and cAMP.** Cyclic AMP-dependent protein kinase A (PKA) is repressed by a regulatory subunit of the adenylyl cyclase holoenzyme and becomes active only on binding of cAMP to this subunit. The cAMP is produced when a signal molecule binds a transmembrane receptor and induces it to activate adenylyl cyclase. (Several steps are omitted here.) Once active, PKA catalytic subunits enter the nucleus and phosphorylate various target proteins, such as CREB, which then recruits RNA polymerase to DNA.

G protein to activate adenylyl cyclase, the enzyme that converts ATP to cyclic AMP (cAMP), leading to elevated levels of cytosolic cAMP. Recall from Chapter 20 that bacteria also use cAMP as a signal molecule; the cAMP binds directly to a transcription activator such as CRP. Eukaryotic cells use cAMP in a very different way. Instead of cAMP binding directly to a transcription factor, eukaryotes use cAMP as a **second messenger** that carries a message received from outside the cell (from the first messenger) to proteins inside the cell. The target of cAMP is a kinase called cyclic AMP-dependent protein kinase A (PKA). It is bound in the cytoplasm by a regulatory subunit of the adenylyl cyclase holoenzyme that inhibits its kinase activity. When cAMP binds to the regulatory protein, the protein dissociates, releasing active PKA.

Protein kinase A has many different target proteins that can lead to the activation or repression of various sets of genes. Figure 21-24 shows the activation of CREB (cAMP-responsive element-binding protein) by phosphorylation. CREB is a transcription activator that is inactive when unphosphorylated, but when phosphorylated and activated by PKA, CREB binds its CRE (cAMP-response element) site in the DNA. It then activates transcription through a coactivator, CBP (CREB-binding protein) of the CBP-p300 complex. CBP is a coactivator for numerous genes, including genes encoding other transcription activators, and it functions in many organs. Most of its effects are still unknown. The most widely studied CREB functions are related to the brain, where CREB is implicated in the formation of long-term memories.

### SECTION 21.3 SUMMARY

- Insulators are DNA sequences that prevent transcription factors bound at distant enhancers from activating the wrong promoters.
- Enhanceosomes are stable, tightly folded nucleoprotein complexes in which cooperating activators integrate regulatory information from multiple signals to produce a single transcriptional outcome at the target promoter.
- Some genes are blocked from active transcription within regions of densely packed heterochromatin. The formation of heterochromatin requires small RNAs, as well as proteins that condense the DNA.
- In imprinting, which occurs in the genes of some higher eukaryotes, the expression of an allele derived from one parent is shut down. Imprinting is an epigenetic process based on nucleosome modification patterns and DNA methylation.
- Gene dosage compensation, required because of the different number of X chromosomes in males and

females, is achieved in one of three ways, depending on the organism. A protein-RNA complex covers the X chromosome(s) to inactivate one female X chromosome, or double the expression of the single male X chromosome, or down-regulate each female X chromosome 50%.

- Steroid hormones control the transcription of specific genes by interacting with intracellular receptors that are transcription activators. Hormone binding triggers interaction of receptor proteins with additional transcription factors. Hormone-receptor complexes bind hormone response elements in the DNA, altering gene expression.
- Nonsteroid hormones and other signal molecules regulate genes through binding to cell surface receptors, triggering protein phosphorylation that leads to modulation of gene expression.

### Unanswered Questions

Eukaryotic cells contain more DNA and more genes than do bacteria, in keeping with their larger size, their intracellular compartmentation, and their cooperation within multicellular organisms. Another way in which they differ from bacteria is that eukaryotes have nucleosomes, which compact the DNA and form different types of chromatin structure, depending on epigenetic alterations of the DNA and histone subunits. All of these differences necessitate greater complexity in gene regulation in eukaryotes. Although research has taught us much about eukaryotic gene expression, numerous questions have yet to be answered.

1. **Why do eukaryotic genes need so many different regulatory protein-binding sites?** Given their greater genomic complexity, we might expect eukaryotes to need more gene-regulatory elements than bacteria. But some eukaryotic genes have so many regulatory sites that it is hard to understand what they all do. Some coactivators even have enzymatic activity that modifies proteins, such as RNA polymerase (see Chapter 16) or histones (see Chapter 10). For genes regulated by enhanceosomes, why are so many proteins required to come together to activate a single gene? An exciting area of study will be to gain a deeper understanding of how transcription modulators function.
2. **Do different gene-regulatory processes intertwine?** Transcription, mRNA processing, replication, recombination, and repair all occur in the nucleus. It seems possible that additional levels of gene regulation might be achieved by interconnections

among these different processes. There is also evidence that transcription and mRNA splicing are coordinated. Future studies are likely to reveal increasingly complex regulatory networks among these diverse processes.

### 3. How is heterochromatin assembled and regulated?

The role of RNA-mediated silencing machinery in assembling heterochromatic regions of a chromosome is a fascinating topic. Recent studies suggest that different areas of heterochromatin may have unique mechanisms of formation, depending

on their location along the chromosome. Indeed, heterochromatin formation in X chromosome inactivation occurs by a different process than heterochromatin formation at a centromere. Understanding the generation of this important epigenetic silencing mechanism, and how heterochromatin formation is regulated during differentiation, are important avenues of future research.

# How We Know

## Transcription Factors Bind Thousands of Sites in the Fruit Fly Genome

Li, X., S. MacArthur, R. Bourgon, D. Nix, D.A. Pollard, V.N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C.L. Luengo Hendriks, et al. 2008. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* 6(2):e27, doi: 10.1371/journal.pbio.0060027.

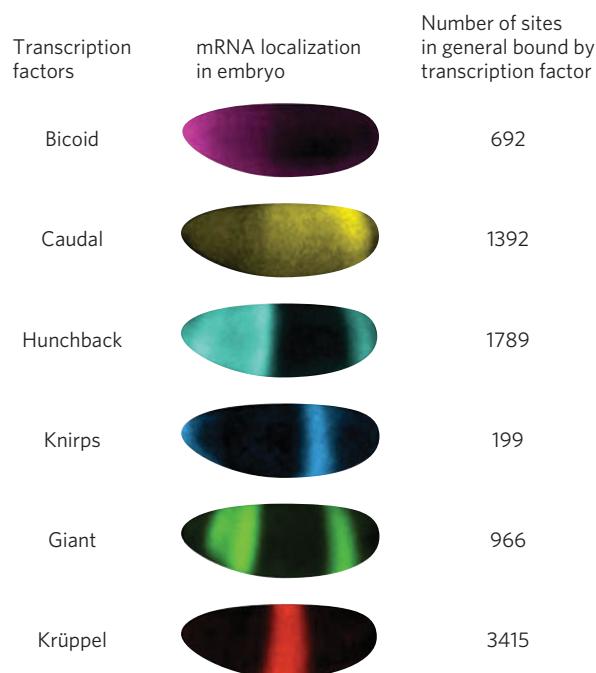


**Mark Biggin** [Source: Courtesy of Mark Biggin.]

Until recently, much of the research on transcription factor binding to DNA focused on experiments with purified proteins and short DNA sequences in vitro. Mark Biggin, of the Lawrence Berkeley National Laboratory, wondered how transcription factors might interact with DNA in living cells. Using *Drosophila* embryos (undergoing a transcriptionally controlled program of anterior-posterior segmentation) and the ChIP-Chip method (see Figure 10-21), Biggin and his colleagues set out to identify the binding sites for six transcription factors known to be active at this stage of fruit fly development. In the ChIP (chromatin immunoprecipitation) part of the experiment, chromatin from the embryos was chemically cross-linked to bound proteins, then purified by precipitation with antibodies to the six transcription factors. Next, in the Chip (DNA microarray chip analysis) part, the DNA in the immunoprecipitated samples was identified using microarray chips containing short DNA segments corresponding to every sequence in the fruit fly genome.

The results were surprising (Figure 1). The six transcription factors were bound to several thousand DNA segments located near half of all the protein-coding genes in the *Drosophila* genome! These binding sites corresponded to many more sequences, and many more genes, than the transcription factors were thought to regulate based on DNA-binding preferences determined in vitro. However, only some of the in vivo binding sites showed up repeatedly in the data analysis, indicating that these sites are frequently occupied by the transcription factors. These high-occupancy sites correspond to DNA targets that are almost certainly regulatory, given their proximity to genes that are activated during fruit fly development. The remainder of the in vivo binding sites are less frequently bound, indicat-

ing that they may not be used to regulate transcription. Instead, these may represent sites where transcription factors can bind nonproductively, perhaps as part of their search for higher-affinity sites along the chromatin. Biggin and coworkers' study should open the way for further investigation of the binding and regulatory roles of transcription factors in the context of chromatin-packaged DNA, and of the myriad other proteins and regulatory factors found in vivo.



**FIGURE 1** The patterns of mRNA expression for six transcription factors in the *Drosophila* embryo show that each factor is expressed in a unique subset of cells. The fruit fly embryos are shown with the anterior end to the left and the dorsal surface at the top. The number of DNA sites bound by each transcription factor is shown on the right. [Source: Adapted from X. Li et al., *PLoS Biol.* 6(2):e27, doi: 10.1371/journal.pbio.0060027, 2008, Table 1 and Fig. 1.]

## Muscle Tissue Differentiation Reveals Surprising Plasticity in the Basal Transcription Machinery

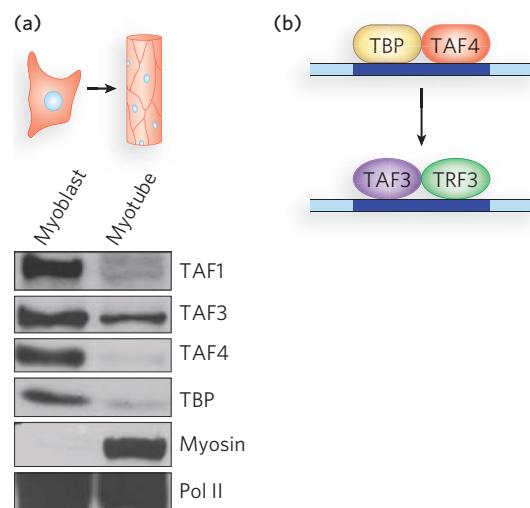
**Deato, M.D.E., and R. Tjian. 2007.** Switching of the core transcription machinery during myogenesis. *Genes Dev.* 21:2137–2149.

**Hu, P., K.G. Geles, J.H. Paik, R.A. DePinho, and R. Tjian. 2008.** Codependent activators direct myoblast-specific MyoD transcription. *Dev. Cell* 15:534–546.

Like many tissue-development processes, muscle differentiation begins with the development of progenitor cells into cells with more specialized functions. In mammalian muscle, myoblasts, the precursor cells, differentiate into myotubes, which subsequently form the muscle fibers of skeletal muscle tissue. The transformation of myoblasts to myotubes involves both selective gene silencing and gene-activation pathways. Transcriptional regulation in these cells has long been known to require cell type-specific basic helix-loop-helix activator proteins, and many researchers suspected that these activators somehow modify the function of the basal transcription machinery in developing muscle.

To test this idea directly, Robert Tjian and his colleagues examined mouse myoblasts to determine which transcription factors are important for the differentiation process. The researchers used the Western blot method (see Figure 7-22), treating cell extracts with antibodies that recognize specific transcription factors, including the TATA-associated factors (TAFs) and TBP components of the TFIID general transcription factor complex, as well as TAFs present only in certain cell types. Because TFIID is part of the basal transcription machinery thought to be common to all cells, Tjian and coworkers expected to find it in muscle cells harvested at all stages of differentiation, along with variable levels of muscle-specific TAFs.

What they discovered instead was that an alternative form of general transcription factor complex, containing the activator proteins TAF3 and TRF3 in place of TFIID, initially coexists with the TFIID-containing core complex in myoblasts. As the cells differentiate into myotubes, however, TFIID decreases to undetectable levels, while TAF3 and TRF3 levels are maintained and eventually become dominant (Figure 2a). When Tjian and colleagues used short interfering RNAs (siRNAs) to reduce the amount of either TAF3 or TRF3 in myoblasts (in experiments not shown here), the expression of the muscle-specific protein MyoD also dropped, and muscle differentia-



**FIGURE 2** (a) Gels resulting from Western blot analysis of TFIID components (TAFs and TBP) involved in differentiation of myoblasts to myotubes in mouse muscle tissue. TFIID is represented by its component TBP; the TAF3-TRF3 complex is represented by TAF3. (b) A proposed model for cell differentiation from myoblast to myotube cells. A core transcription initiation complex including TAF3 and TRF3 functionally replaces the canonical TFIID complex in myotube cells, switching on the unique transcription pattern required during cell type-specific terminal differentiation. [Source: (a) Adapted from M. D. E. Deato and R. Tjian, *Genes Dev.* 21:2137–2149, 2007, Fig. 1b. © Cold Spring Harbor Laboratory Press.]

tion was compromised. These effects could be reversed by supplying fresh TAF3 and TRF3 to depleted myoblast cells.

These findings implicate TAF3-TRF3 complexes in the transcription of proteins central to the muscle cell differentiation pathway (Figure 2b). More importantly, they suggest that previously unexpected changes in the basal transcription machinery are required for the widespread changes in transcription patterns responsible for cellular differentiation in higher eukaryotes.

## Key Terms

transcriptional ground state, p. 734	TATA-binding protein (TBP), p. 740	Barr body, p. 756
heterochromatin, p. 735	high-mobility group (HMG) protein, p. 740	X chromosome inactivation, p. 756
euchromatin, p. 735	Mediator complex, p. 740	X inactivation center (XIC), p. 756
hypersensitive site, p. 735	combinatorial control, p. 743	dosage compensation complex (DCC), p. 756
enhancer, p. 737	insulator, p. 751	hormone response element (HRE), p. 758
upstream activator sequence (UAS), p. 737	enhanceosome, p. 752	G protein-coupled receptor (GPCR), p. 759
general (basal) transcription factor, p. 740	gene silencing, p. 753	second messenger, p. 760
DNA-binding transactivator, p. 740	imprinting, p. 753	
coactivator, p. 740	parthenogenesis, p. 755	
	dosage compensation, p. 755	

## Problems

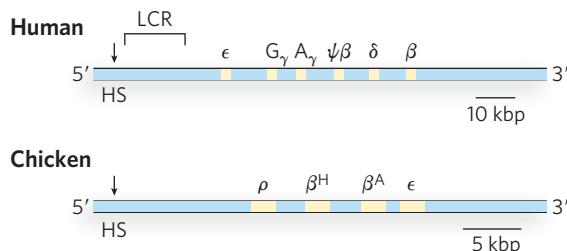
- In eukaryotes, most genes are normally turned off, and RNA polymerases do not function without activation. In bacteria, RNA polymerase can transcribe almost any gene, in the absence of bound inhibitors. Suggest a few reasons for this difference between bacteria and eukaryotes.
- The regulatory proteins in eukaryotes bind to DNA sequences of about the same length as those bound by bacterial regulatory proteins. However, the genomes of eukaryotes are generally orders of magnitude larger than those of bacteria. What effect does this have on the strategy of eukaryotes for regulating a particular gene?
- Optimal activation of transcription of the *GAL* genes in yeast requires the function of the Gal4 and Gal11 proteins (Gal4p and Gal11p). Gal11p has not been mentioned in this chapter. Elimination of either protein decreases activation of the *GAL* promoters. However, inactivation of Gal11p has the additional and dramatic effect of cell lethality. Suggest why elimination of Gal11p might have a greater effect than elimination of Gal4p.
- What is the phosphorylation state of the yeast protein Mig1 when: (a) glucose and galactose are absent; (b) galactose is present and glucose is absent; (c) glucose is present and galactose is absent; and (d) both glucose and galactose are present?
- Perhaps 3,000 or more transcription factors participate in the activation of human genes. However, this is far fewer than the number of genes in the human genome ( $\sim 20,000$  to 25,000). Explain how specific gene activation is achieved when there are tenfold fewer gene activators than there are genes.
- If mice are engineered with a homozygous gene knockout (inactivation) for the gene encoding CTC-binding factor (CTCF), they exhibit an embryonic lethal phenotype. Explain.
- Enhanceosomes consist of multiple transcription factors that activate transcription at particular genes. The enhanceosomes also often include HMG proteins (see Figure 21-6). Suggest a function for the HMG proteins.
- Housekeeping genes are those that must be expressed at all times, providing a protein or RNA that is essential for general cellular metabolism. They are often expressed at a low but constant level. If an essential housekeeping gene were experimentally moved from euchromatin to a region of heterochromatin, what would be the likely effect on the cell?
- A scientist is studying the function of a type of nuclear steroid receptor protein in mouse cells. She introduces various mutations into the gene encoding the receptor protein and transfers the genes into mice. If mutations are introduced that (a) eliminate the nuclear import signal in the receptor protein or (b) alter the receptor protein surface so that the receptor can no longer interact with Hsp70 protein, how will the molecular pathway of hormone-receptor interaction be altered?

## Data Analysis Problem

**Chung, J.H., M. Whiteley, and G. Felsenfeld. 1993.** A 5' element of the chicken  $\beta$ -globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* 74:505–514.

- 10.** By the early 1990s, a few examples of insulator sequences had been discovered and characterized in eukaryotic cells.

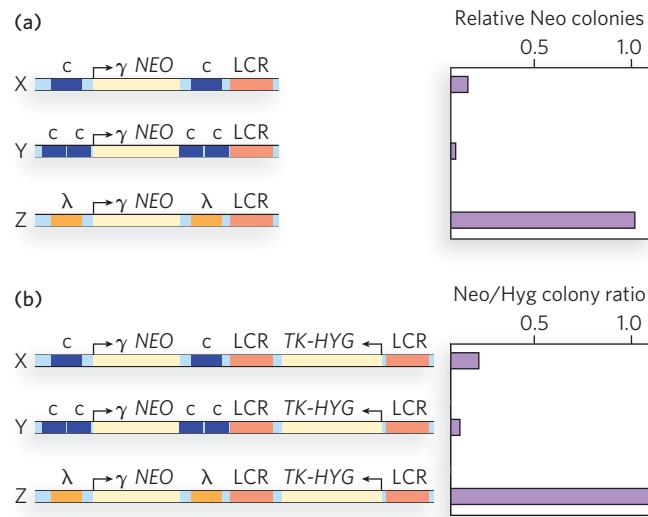
These insulators were mostly found in *Drosophila*. To extend the work to vertebrates, Chung, Whitley, and Felsenfeld focused on features of the  $\beta$ -globin gene cluster that were conserved in chicken, mouse, and human. They noted that locus control regions (LCRs) near the 5' end of the  $\beta$ -globin gene cluster served to attract enzymes that

**FIGURE 1**

opened up the chromatin from the 5' to the 3' end of the cluster, a distance encompassing a few hundred thousand base pairs of the human genome. In principle, the LCR can serve that function—attracting enzymes to open up the chromatin to prepare it for transcription—in either direction. However, at the 5' end of the cluster and beyond, the chromatin remained condensed. Something was blocking the chromatin remodeling in that direction. The investigators focused on a prominent and constitutive nuclelease-hypersensitive site (HS) at the 5' end of the cluster. The location of that site is shown for the chicken and human systems by the black arrows in **Figure 1**, which shows the genes of the  $\beta$ -globin gene cluster in each organism.

- (a)** What is a nuclease-hypersensitive site in chromatin, and what is its significance?

The investigators constructed a series of vectors in which the constitutive hypersensitive region (denoted C in **Figure 2**) was placed at sites on either side of a gene conferring resistance to the antibiotic neomycin ( $\gamma$ -NEO in Figure 2). For eukaryotes, researchers generally use the related antibiotic geneticin, or G418, to kill cells. The  $\gamma$ -NEO gene confers resistance. The constructs are labeled X and Y in Figure 2. As a control, the investigators replaced the constitutive hypersensitive site (C) with a fairly random DNA segment of comparable length derived from  $\lambda$  phage (construct Z in Figure 2). They transfected human erythroleukemia cells with these constructs, isolated stably transfected cell lines (in which the construct had integrated at some random site in the genome), and counted the number of colonies produced when the cells were suspended in semisolid agar medium containing G418. As shown in the graph in **Figure 2a**, the number of G418-resistant colonies decreased when one or two of the chicken hypersensitive sites were inserted between the neomycin-resistance gene and the LCR.

**FIGURE 2**

- (b)** The LCR generally controls genes to its 3' side (to the right as shown in the figures). Why would it affect a neomycin-resistance gene placed on the 5' side?  
**(c)** Why was it necessary to insert the hypersensitive sites on both sides of the neomycin-resistance gene?  
**(d)** Was the chicken hypersensitive site effective in isolating the neomycin-resistance gene?

The investigators made a second series of constructs, shown in **Figure 2b**. Here, they added a gene for resistance to the antibiotic hygromycin (TK-HYG). This was set up in each construct so that the gene would be expressed constitutively. The constructs were again transfected into the same human cell line, and the number of colonies growing in media containing either hygromycin or G418 were tallied. The ratio again revealed that the number of neomycin-resistant colonies was greatly reduced when the neomycin-resistance gene was flanked by the chicken hypersensitive site.

- (e)** Why was this control experiment necessary?  
**(f)** Given the results presented above, can you conclude that the constitutive hypersensitive site from the chicken is an insulator that affects gene transcription? Justify your answer.  
**(g)** What other experiments would be needed to demonstrate or confirm that transcription was affected?

## Additional Reading

### General

D'Alessio, J.A., K.J. Wright, and R. Tjian. 2009. Shifting players and paradigms in cell-specific transcription. *Mol. Cell* 36:924–931.

Michel, D. 2010. How transcription factors can adjust the gene expression floodgates. *Prog. Biophys. Mol. Biol.* 102:16–37.

### Basic Mechanisms of Eukaryotic Transcriptional Activation

- Arnosti, D.N., and M.M. Kulkarni.** 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biol.* 94:890–898.
- Bernstein, E., and S.B. Hake.** 2006. The nucleosome: A little variation goes a long way. *Biochem. Cell Biol.* 84:505–517.
- Björklund, S., and C.M. Gustafsson.** 2005. The yeast Mediator complex and its regulation. *Trends Biochem. Sci.* 30:240–244.
- Segal, E., and J. Widom.** 2009. What controls nucleosome positions? *Trends Genet.* 25:335–343.

### Combinatorial Control of Gene Expression

- Campbell, R.N., M.K. Leverentz, L.A. Ryan, and R.J. Reece.** 2008. Metabolic control of transcription: Paradigms and lessons from *Saccharomyces cerevisiae*. *Biochem. J.* 414:177–187.

**Davidson, E.H., and M.S. Levine.** 2008. Properties of developmental gene regulatory networks. *Proc. Natl. Acad. Sci. USA* 105:20,063–20,066.

**Juven-Gershon, T., J.Y. Hsu, J.W. Theisen, and J.T. Kadonaga.** 2008. The RNA polymerase II core promoter—the gateway to transcription. *Curr. Opin. Cell Biol.* 20:253–259.

### Transcriptional Regulation Mechanisms Unique to Eukaryotes

- Gaszner, M., and G. Felsenfeld.** 2006. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* 7:703–713.
- Werner, M.H., and S.K. Burley.** 1997. Architectural transcription factors: Proteins that remodel DNA. *Cell* 88:733–736.
- Wood, A.J., and R.J. Oakey.** 2006. Transcriptional control: Imprinting insulation. *Curr. Biol.* 10:R463–465.

# The Posttranscriptional Regulation of Gene Expression in Eukaryotes



**Judith Kimble** [Source: Courtesy of Judith Kimble.]

ing only on incubation temperature. Julie Ahringer, a student in my lab at the time, began to sequence the gene mutants, but found no molecular changes in the gene's open reading frame! This was really puzzling, but she pushed on into noncoding regions and discovered a single base pair change in the part of the gene corresponding to the 3'UTR of the mRNA.

When Julie tested the effect of introducing wild-type or mutant 3'UTRs back into the worm, she confirmed the effect of that single base pair change. She soon sequenced the same region in nine independently isolated mutations of the same class, and they all carried 3'UTR mutations falling within a five base pair region!

This was a truly exhilarating discovery, because at the time, no one expected the noncoding bits of an mRNA to be so important for cell-fate regulation. That breakthrough paved the way for many subsequent studies of 3'UTR regulation, which we now know is a fundamental and conserved mechanism of gene control.

—**Judith Kimble**, on discovering that noncoding regions of mRNA regulate cell fate

- 22.1 Posttranscriptional Control inside the Nucleus** 768
- 22.2 Translational Control in the Cytoplasm** 773
- 22.3 The Large-Scale Regulation of Groups of Genes** 779
- 22.4 RNA Interference** 782
- 22.5 Putting It All Together: Gene Regulation in Development** 790
- 22.6 Finale: Molecular Biology, Developmental Biology, and Evolution** 801

**H**ow do different cell types arise during the development of a multicellular organism? It is one of the most complex and fascinating processes in molecular biology. For example, the adult human body contains about fifty trillion ( $50 \times 10^{12}$ ) cells that originated from a single fertilized egg cell. Almost all these cells contain the same DNA, yet the cells of each organ, and even those within an organ, have vastly different shapes and functions. The differences must reflect gene regulation.

In eukaryotic cells, gene transcription and pre-mRNA processing—splicing, 5' capping, and 3' polyadenylation—occur in the nucleus. Only after export to the cytoplasm are mature mRNAs recognized by ribosomes for translation into proteins. This physical and temporal separation of transcription and translation, distinct from the situation in bacterial cells, requires additional steps in the information pathways of eukaryotes. Although these extra steps take time (which is not always available), they provide unique opportunities to impose control.

The early research on eukaryotic gene regulation focused on transcription, and particularly transcription initiation. It made practical sense that cells would control gene expression by regulating the first step, avoiding energy expenditure on unneeded transcripts. However, the experimental data have increasingly pointed to an abundance of regulatory mechanisms that occur *after* transcription. In humans and other multicellular organisms, for many genes, transcripts and even proteins are routinely produced that are not immediately used. Instead, the mRNAs and proteins are stored and used later, bypassing the time-consuming transcription and transport steps and thus allowing a more rapid response to cellular needs or metabolic signals.

To a large degree, the importance of posttranscriptional regulation parallels the complexity of the cellular processes that are regulated. Signal transmission in the brain, color patterns in flower petals, and that most complex of all biological processes, the development of a multicellular organism, are all governed by regulatory processes that take place after transcription. In this chapter, we discuss some of the predominant ways that cells select which mRNAs are to be translated into protein, and how much protein is to be made.

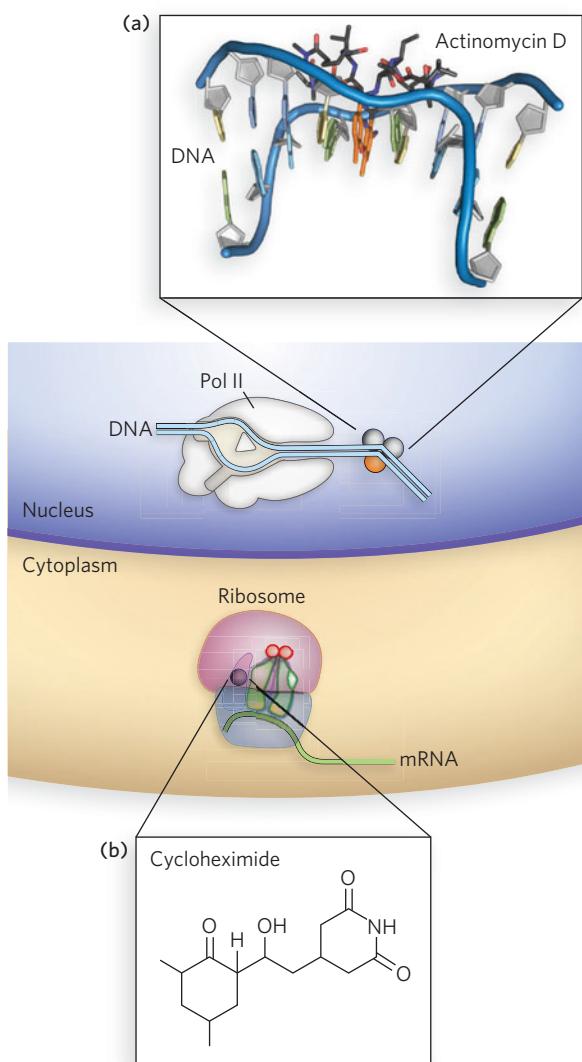
We begin with overviews of mechanisms that provide exquisite posttranscriptional control of gene expression levels in the nucleus and in the cytoplasm. We then turn to pathways for regulating groups of genes, including the exciting discovery of small RNAs that alter gene expression through the process of RNA interference (RNAi). Next, we discuss embryonic development,

a process in which almost all the transcriptional and posttranscriptional regulatory mechanisms described in the last several chapters come together. Most of the regulatory mechanisms that guide development are highly conserved in eukaryotes, from nematodes to humans. In addition to exemplifying mechanisms of gene regulation, elucidation of developmental pathways has taught us much about evolution and how it generates alterations in function and appearance in organisms. Thus, we end the book where we started—with a discussion of molecular biology from the perspective of evolution.

## 22.1 Posttranscriptional Control inside the Nucleus

The multitude of posttranscriptional regulatory mechanisms in eukaryotes is usefully divided into those that occur in the nucleus and those that occur in the cytoplasm. In the nucleus, mRNA is modified in many ways and prepared for transport to the cytoplasm. In the cytoplasm, the focus shifts to translation. Control mechanisms determine which transcripts are translated, which are stored (and where they are stored), and how long each transcript is present in the cell. Additional controls are layered onto the translation process itself and the fate of the proteins thus produced (see Chapter 18).

Experiments to test which step in gene expression—transcription or translation—is regulated in cells often take advantage of molecular inhibitors that are specific for one step or the other. Two natural antibiotics made by bacteria of the genus *Streptomyces* have been particularly useful. Actinomycin D is a polypeptide that specifically blocks eukaryotic gene transcription by binding to DNA within the transcription initiation complex, preventing transcript elongation by RNA polymerase II (Figure 22-1a). Cycloheximide is a small molecule that interferes with the movement of tRNAs during polypeptide elongation on the ribosome, thus blocking protein synthesis (Figure 22-1b). These compounds are useful tools for the preliminary dissection of mechanisms of eukaryotic gene regulation. For example, Mark Ashe and Alan Sachs discovered that yeast cells that are starved for sugar rapidly shut down their protein synthesis. To determine whether this regulation is based in transcription or translation, yeast cells growing in a nutrient-rich medium were isolated by centrifugation and transferred to a medium containing no sugar. Protein production in these cells, as measured by the abundance of large ribosome-mRNA complexes, decreased to almost zero within a few minutes. On addition of sugar to the growth medium, protein synthesis was restored.



**FIGURE 22-1** Compounds that selectively block transcription or translation. (a) Actinomycin D blocks transcription elongation by RNA polymerase II, by inserting itself into DNA. (See Figure 15-8 for the chemical structure of actinomycin D.) (b) Cycloheximide blocks translation elongation by the ribosome. [Source: (a) PDB ID 1DSC.]

Actinomycin D had no effect on either the inhibition or the reactivation of protein production, showing that no new transcription was needed for this kind of regulation.

In this section, we embark on detailed descriptions of the processes by which mRNAs are altered in the nucleus to control the temporal expression and structure of the final gene products. Regulation by the alteration of mRNA sequences can allow a single gene to yield more than one protein product, thus optimizing the efficiency of information storage in a genome. We focus on some examples of mRNA alterations that

illustrate this potential, including mRNA splicing, 3'-end cleavage, and mRNA transport. These processes were introduced in Chapter 16 and are expanded on here. RNA editing, another process that makes important contributions to the coding potential of some eukaryotic mRNAs, is also described in Chapter 16.

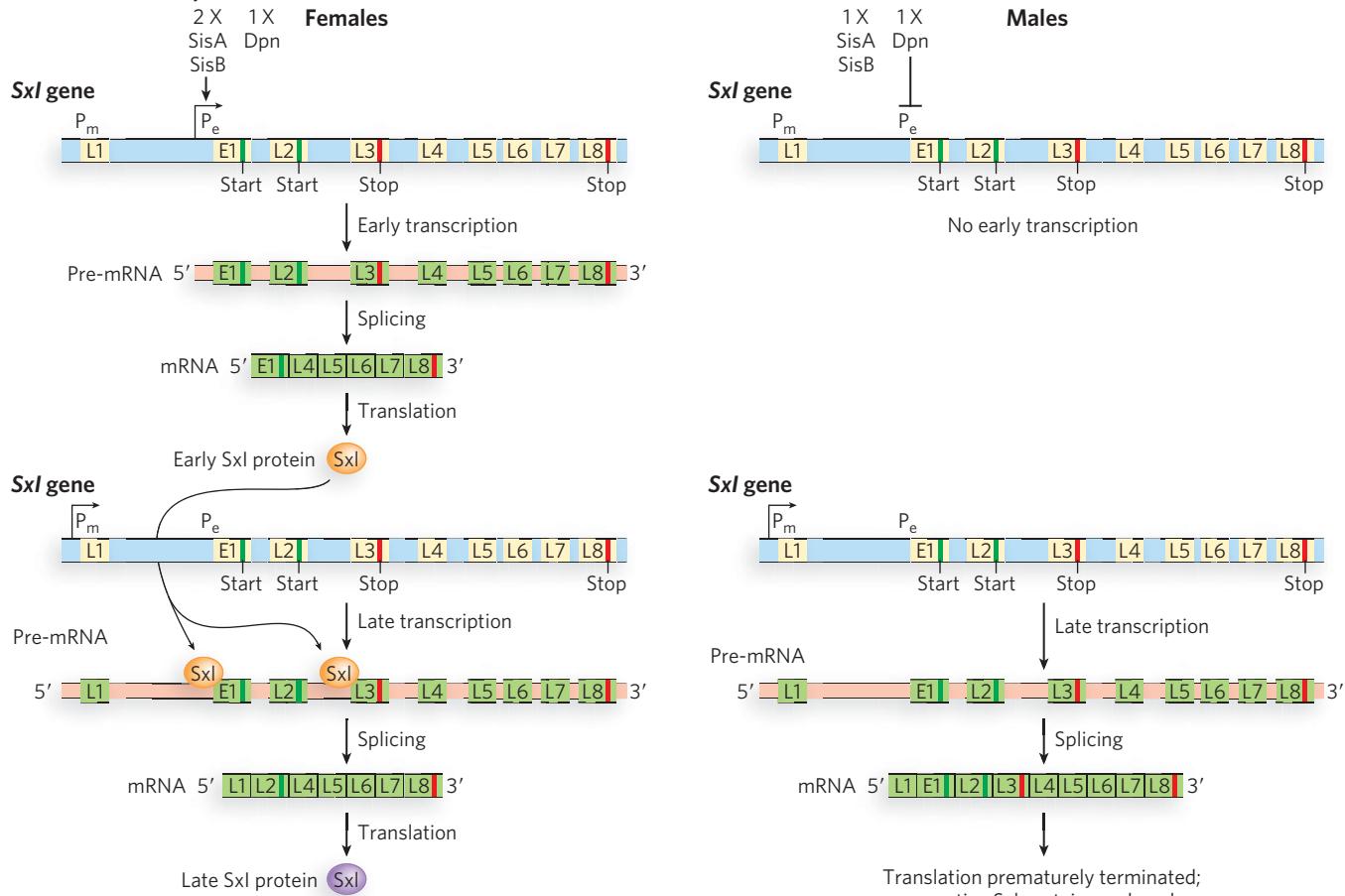
### Alternative Splicing Controls Sex Determination in Fruit Flies

The splicing of a new mRNA transcript does not always lead to the exclusive excision of all the introns. Instead, for some genes, one or more exons may also be deleted during maturation of a subset of the mRNAs. This process, called **alternative splicing**, creates new forms of the mature transcript that encode versions of the protein lacking one or more peptide segments. In this way, two or more different proteins can be produced from a single coding sequence.

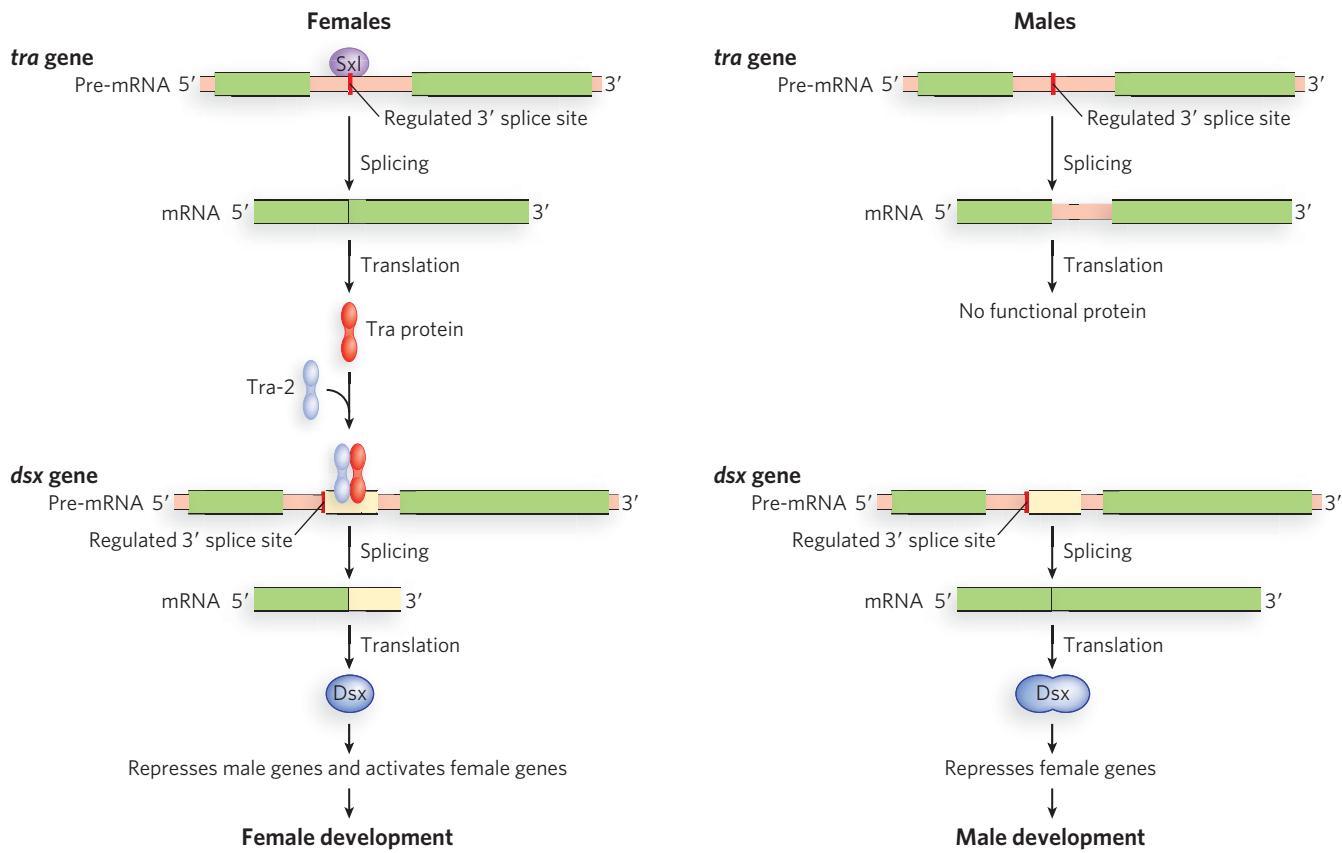
In *Drosophila*, an interesting example of alternative pre-mRNA splicing dictates sexual fate during development. In both male and female flies, each cell contains two copies of the autosomes; female cells also have two X chromosomes, whereas male cells contain just one X chromosome. The resulting difference in the ratio of X to autosomes leads to different levels of expression of the X-chromosomally encoded SisA (*sisterless*) and SisB transcription activators. These regulatory proteins, along with an autosomally encoded repressor called Deadpan (Dpn), act on the gene *Sex-lethal* (*Sxl*). In females, with SisA and SisB in twice the amounts found in males, *Sxl* is activated; in males, *Sxl* is repressed.

Two promoters govern *Sxl* expression:  $P_e$  (establishment) and  $P_m$  (maintenance) (Figure 22-2a). The *Sxl* gene has one exon, L3, that contains an in-reading-frame UGA stop codon, and this exon must be removed if a full-length protein is to be produced. When *Sxl* is expressed from  $P_e$ , the primary transcript does not include the L1, L2, or L3 exons. Due to either RNA folding or the action of a protein not yet identified, the default splicing pathway of this transcript connects the exon E1 (encoding the N-terminal 20 amino acid residues of the protein) directly to L4. The splice junctions at the ends of exons L2 and L3 are not recognized by the splicing machinery, and thus are deleted along with the neighboring introns. The mature transcript encodes an active *Sxl* protein with the N-terminal residues encoded by exon E1. When *Sxl* is expressed from  $P_m$ , the default splicing pathway eliminates all introns precisely; translation is initiated in L2 and halted at the UGA stop codon in L3, and inactive *Sxl* protein results.

(a) Differential expression of *Sex-lethal*



(b) Differential expression of downstream genes



**FIGURE 22-2 Alternative splicing of Sxl mRNA in male and female fruit flies.** (a) The structure of the Sxl gene, showing the exons, introns, promoters, start codons, and stop codons, and its splicing patterns. In females, transcription from promoter P<sub>e</sub> results in a characteristic splicing pattern and the production of a burst of active Sxl protein (top left). Later transcription from P<sub>m</sub>, in the presence of Sxl, continues to produce active Sxl protein (bottom left). In males, transcription from P<sub>e</sub> is repressed (top right). The absence of Sxl leads to a different splicing pattern of P<sub>m</sub> transcripts and premature termination of Sxl (bottom right). (b) Sxl protein mediates differential expression of genes further downstream in the developmental pathway. In females, active Sxl protein produced from P<sub>m</sub> leads to production of proteins that facilitate female development—first Tra protein, which then, with Tra-2, facilitates splicing of the mRNA for active Dsx protein (left). In males, the absence of Sxl leads to a different splicing pattern of tra transcripts; a different version of Dsx is produced that leads to male development (right).

In females, the higher levels of SisA and SisB overcome Dpn repression and activate P<sub>e</sub> early in development, generating active Sxl protein. P<sub>m</sub> is switched on slightly later, and transcription from P<sub>m</sub> is not dependent on SisA, SisB, or Dpn. The Sxl protein, already present at this point in development, is a splicing repressor that binds to the primary P<sub>m</sub>-dependent transcript and prevents the splicing machinery from recognizing the splice junctions on either side of the E1 or L3 exon. L1 is spliced directly to L2, and L2 to L4. In this way, the L3 exon is spliced out of the transcript as part of a larger segment effectively containing two introns, and production of functional Sxl protein continues (now with N-terminal residues encoded by L2).

The Sxl protein does not simply facilitate its own expression, however; it is also needed to regulate the expression of several additional genes downstream in the female developmental pathway (Figure 22-2b, left). Sxl regulates the splicing of additional transcripts, including that from the *transformer* (*tra*) gene. Sxl-mediated alternative splicing produces transcripts that encode functional Tra protein. In turn, the Tra protein activates splicing of pre-mRNA from the gene *double sex* (*dsx*), such that the truncated protein expressed from this spliced form represses the expression of male-specific genes.

In males, P<sub>e</sub> is repressed by Dpn and never activated, due to the lower levels of SisA and SisB than in females. There is no early production of the Sxl protein. When, later, the expression of Sxl occurs from P<sub>m</sub>, the lack of presynthesized Sxl protein leads to splicing of the P<sub>m</sub>-produced transcript by its default pathway, with production of inactive Sxl fragments (see Figure 22-2a). Without functional Sxl protein to regulate splicing of

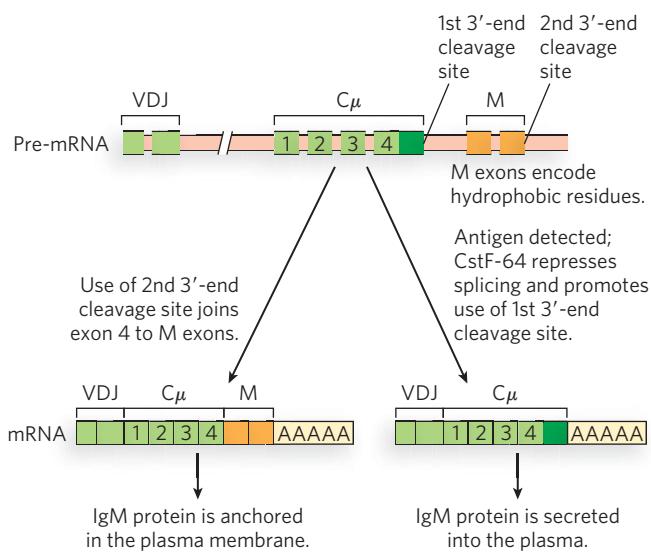
the *tra* gene, no functional Tra protein is produced. In the absence of Tra protein, splicing of the *dsx* transcript produces mRNAs encoding an extended form of Dsx protein that represses transcription of female-specific genes. In the absence of genes enforcing the female developmental pathway, a male-specific pathway is initiated (Figure 22-2b, right).

This interwoven regulatory cascade, with many steps involving alternative splicing, provides exquisite control over the sexual development of fruit flies by ensuring that sex-specific transcription patterns are maintained.

## Multiple mRNA Cleavage Sites Allow the Production of Multiple Proteins

The **3' cleavage** and polyadenylation of eukaryotic mRNAs is carried out by a protein complex recognizing a site defined, in part, by the sequence AAUAAA. This is a key event in the maturation of mRNAs (see Chapter 16). Many genes have multiple sites for 3'-end cleavage. By regulating which site is used, cells can produce different proteins from a single gene. Sequencing of the human genome has revealed that as many as 60% of the genes may have multiple alternative 3'-end cleavage sites.

A well-studied example is the gene encoding immunoglobulin M (IgM) heavy chains. As the B cells (B lymphocytes) of the immune system mature, they enter a quiescent state in which each cell expresses a unique IgM on its surface. The many different B cells express immunoglobulins with different binding specificities, enabling the immune system to respond to a wide array of antigens. The membrane-bound IgM is generated by a splicing event in the primary transcript that eliminates one of two 3'-end cleavage sites and attaches two exons (M exons) that encode a series of hydrophobic amino acid residues at the C-terminus of the IgM; this hydrophobic sequence serves to anchor the protein in the membrane (Figure 22-3). When an antigen appears in the extracellular environment that is recognized by the IgM of a particular B cell, this triggers cellular signaling that leads to rapid proliferation (sometimes called clonal expansion) of that B cell. As the proliferation proceeds, a change occurs in the processing of the IgM mRNAs. Increased concentration of the protein CstF-64 (cleavage stimulation factor) leads to predominant use of the first 3'-end cleavage site of the IgM transcript rather than its deletion by splicing. The resulting IgM molecules no longer have the membrane anchor and are secreted into the surrounding plasma to help neutralize the threat posed by the antigen.



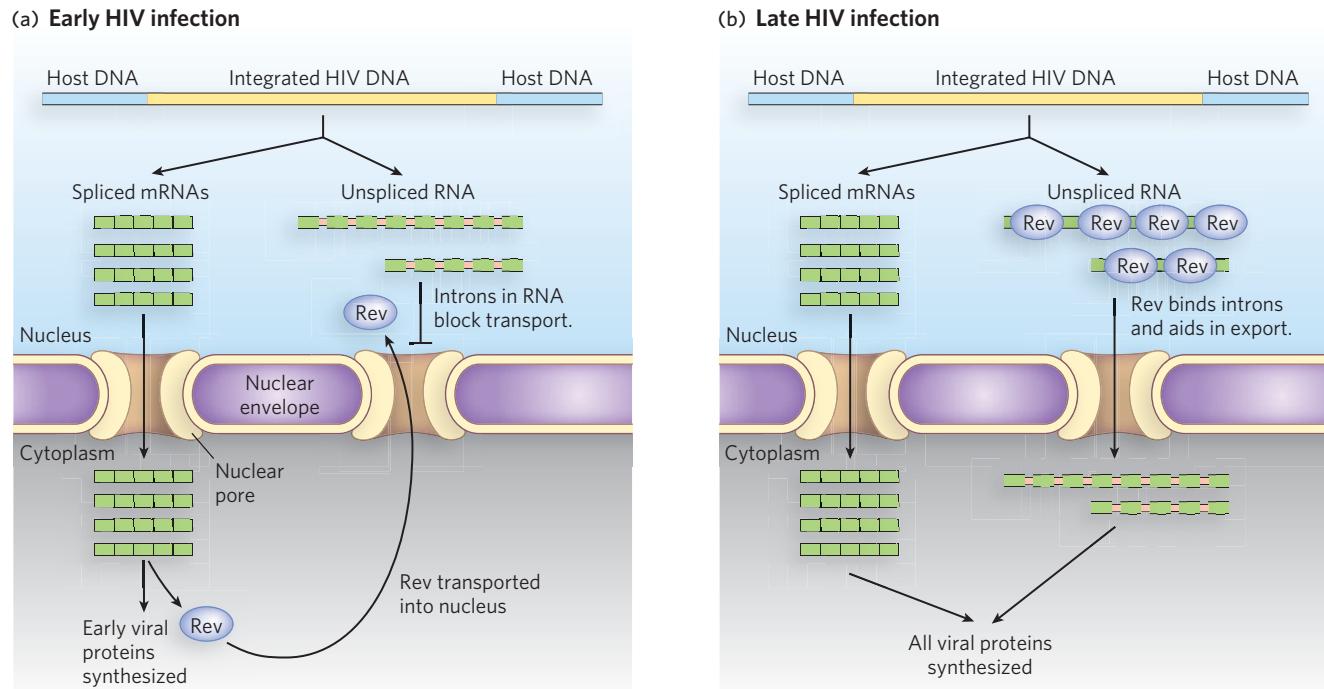
**FIGURE 22-3 Alternative 3'-end cleavage sites and the fate of IgM proteins.** In the absence of antigen, the first cleavage site is spliced out during mRNA processing, leaving the second site and the two M exons encoding hydrophobic C-terminal sequences that anchor the IgM in the membrane. When the cell encounters antigen, the splicing reaction that eliminates the first cleavage site is suppressed through the action of CstF-64; the mature mRNA produces an IgM that is secreted. C<sub>μ</sub> exons encode the constant region of the IgM heavy chain. VDJ exons encode the variable, diversity, and joining segments of the IgM heavy chain. [Source: Adapted from J. Zhao, L. Hyman, and C. Moore, *Microbiol. Mol. Biol. Rev.* 63:405–445, 1999.]

## Nuclear Transport Regulates Which mRNAs Are Selected for Translation

Incompletely processed mRNAs, such as those that still contain introns, are not exported from the nucleus and are degraded by the nuclear exosome. If this mechanism could be suppressed, various mRNAs made from

the same gene, some containing one or more introns, could be exported from the nucleus. Translation would then produce different proteins from these mRNAs.

A regulatory mechanism that overcomes the nuclear surveillance of introns in mRNA exists, and it is used strategically by the human immunodeficiency virus. HIV makes one very long pre-mRNA, which is



**FIGURE 22-4 Regulation of nuclear export by the HIV Rev protein.** (a) Early in HIV infection, only fully spliced viral mRNAs, containing the coding sequences for Rev and other proteins, are exported from the host cell nucleus and translated. (b) Once sufficient Rev protein has accumulated and been transported into the nucleus, unspliced viral RNAs can be exported into the cytoplasm. Many of these RNAs are translated into proteins. These and all other viral proteins are packaged into new viral particles.

spliced in a variety of ways to make more than 30 mature mRNAs. Some of these mRNAs contain introns that are spliced out in other mRNAs, but become part of the coding sequence in those where they are retained. These intron-containing mRNAs must be exported from the nucleus to be translated, and that export would normally be blocked by the splicing signals in the introns. To subvert the cell's nuclear surveillance process, HIV encodes a protein called Rev. After synthesis in the cytoplasm, Rev is transported into the nucleus and binds to introns in HIV transcripts (Figure 22-4). Rev also binds transport receptors, and thus escorts intron-containing viral RNAs out of the nucleus.

### SECTION 22.1 SUMMARY

- Eukaryotic cells, and multicellular organisms in particular, express just a subset of proteins, depending on cell type and in response to various chemical signals. To maximize versatility, even at the expense of producing unneeded transcripts, cells sometimes regulate gene expression posttranscriptionally.
- Two natural antibiotics, actinomycin D and cycloheximide, block eukaryotic transcription and translation, respectively. These are useful tools for determining which step of gene expression is regulated in response to a particular stimulus.
- In fruit flies, alternative splicing of the pre-mRNA encoding the Sex-lethal (Sxl) protein determines sexual fate during development.
- Alternative selection of sites for 3'-end cleavage determines the fate of IgM molecules produced by B cells.
- HIV produces a protein (Rev) that suppresses cellular intron surveillance and allows export of intron-containing viral transcripts from the nucleus.

## 22.2 Translational Control in the Cytoplasm

Regulation at the level of translation assumes a much more prominent role in eukaryotes than in bacteria, and is observed in a range of cellular situations. In contrast to the tight coupling of transcription and translation in bacteria, transcripts generated in a eukaryotic nucleus must be processed and transported to the cytoplasm before translation. This can impose a significant delay on the appearance of a protein. When a rapid increase in protein production is needed, a translationally

repressed mRNA already in the cytoplasm can be activated for translation without delay.

Translational control is responsible for activating the expression of proteins necessary for cell fate decisions. Translational regulation may play an especially important role in controlling the expression of certain very long eukaryotic genes (a few are measured in the millions of base pairs!), for which transcription and mRNA processing can require many hours. Some genes are regulated at both the transcriptional and translational stages, with the latter playing a role in fine-tuning of cellular protein levels. In some non-nucleated cells, such as reticulocytes (immature red blood cells), transcriptional control is unavailable and translational control of stored mRNAs becomes essential. Translational controls are essential during development, when the regulated translation of pre-positioned mRNAs creates a local gradient of the protein product (see Section 22.5).

Eukaryotes have at least three main mechanisms for regulating translation. First, various initiation factors are subject to phosphorylation by protein kinases. The phosphorylated forms are often less active and generally depress translation in the cell. Second, some proteins bind directly to mRNA and act as translational repressors. Many bind at specific sites in the 3'UTR and interact with other initiation factors bound to the mRNA, or interact with the 40S ribosomal subunit to prevent initiation. Third, binding proteins can disrupt the interaction between eIF4E and eIF4G. (Recall from Chapter 18 that interaction between these initiation factors is required for proper assembly of the ribosome-mRNA complex.) Such binding proteins are present in eukaryotes from yeast to mammals, and the mammalian versions are known as 4E-BPs (eIF4E binding proteins). When cell growth is slow, these proteins limit translation by binding to the site on eIF4E that normally interacts with eIF4G. When cell growth resumes or increases in response to growth factors or other stimuli, the binding proteins are inactivated by protein kinase-dependent phosphorylation.

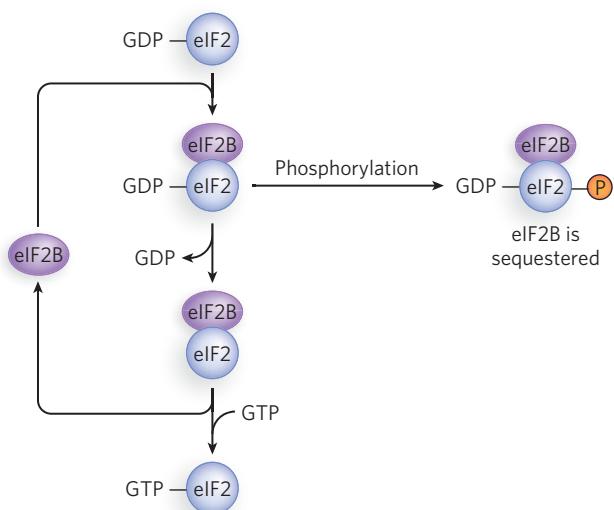
The variety of translational regulation mechanisms in eukaryotes provides flexibility, allowing focused repression of a few mRNAs or global regulation of all cellular translation. Here we explore several examples in which these mechanisms come into play.

### Initiation Can Be Down-Regulated by Phosphorylation of eIF2

Translation initiation in eukaryotes is a complex process involving multiple initiation factors that recruit ribosomes to mRNAs. Reversible phosphorylation of

initiation factors plays a central role in regulating initiation. The phosphorylation is triggered by a wide range of cellular conditions, depending on cell type and function. One of the main phosphorylation pathways involves eIF2. In all eukaryotes, eIF2 is composed of three polypeptide subunits—eIF2 $\alpha$ , eIF2 $\beta$ , and eIF2 $\gamma$ —that together bind the initiator tRNA and GTP. When Met-tRNAA<sup>Met</sup> has bound to the peptidyl (P) site on the 40S ribosomal subunit, GTP is hydrolyzed, and eIF2-GDP dissociates from the initiator tRNA. Recycling of eIF2-GDP to eIF2-GTP requires eIF2B (Figure 22-5). If eIF2 is phosphorylated, eIF2B binding to eIF2 is nearly irreversible and the GDP is not dislodged. Because the cell has less eIF2B than eIF2, only a little phosphorylated eIF2 is needed to sequester all the eIF2B, thereby shutting down protein synthesis.

This process has been well studied in mammalian reticulocytes. The maturation of reticulocytes into red blood cells includes destruction of the cell nucleus, leaving behind a hemoglobin-packed cell. Messenger RNAs deposited in the cytoplasm before loss of the nucleus allow for the replacement of hemoglobin. When reticulocytes become deficient in iron or heme, the translation of globin mRNAs is repressed. A protein kinase called HCR (*hemin-controlled repressor*) is activated, catalyzing phosphorylation of eIF2 $\alpha$ , the smallest subunit of eIF2. When its eIF2 $\alpha$  subunit is phosphorylated, eIF2 forms a stable complex with eIF2B, blocking dissociation of GDP after GTP hydrolysis and thus making



**FIGURE 22-5** Halting the recycling of eIF2 by phosphorylation. The recycling of used eIF2 (eIF2-GDP), after it has served its role in translation initiation, is facilitated by a guanine nucleotide exchange factor, eIF2B. Phosphorylation of the eIF2 protein closes down this cycle, and thus controls translation rates, by tying up eIF2B.

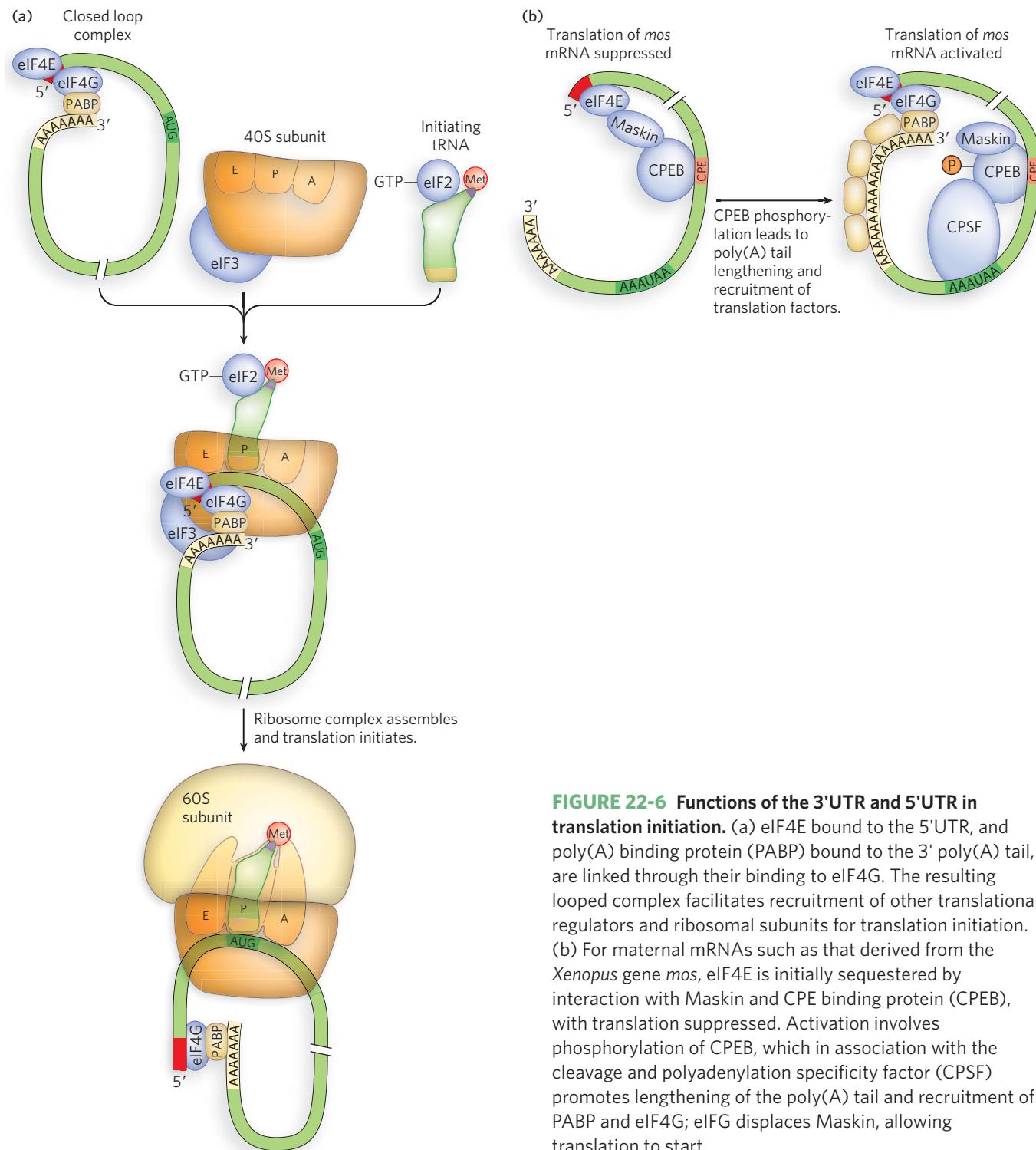
these initiation factors unavailable for further rounds of translation. In this way, the reticulocyte coordinates the synthesis of globin with the availability of heme.

Phosphorylation of eIF2 $\alpha$  regulates translation in other systems as well. For example, a double-stranded RNA-dependent protein kinase (PKR) phosphorylates eIF2 $\alpha$  in some cells in response to viral infection. This helps block the translation of viral mRNAs and interferes with the viral life cycle. In yeast, activation of the kinase Gcn2 by nitrogen starvation leads to eIF2 $\alpha$  phosphorylation and repression of most translation, until more nitrogen (and the amino acids that incorporate it) becomes available. In an interesting mechanistic twist, eIF2 $\alpha$  phosphorylation in yeast induces translation of the transcription factor Gcn4 (which we discuss below).

# The 3'UTR of Some mRNAs Controls Translational Efficiency

The 3' untranslated region of an mRNA communicates with the 5' end through protein-protein interactions between factors that bind specifically to the ends of a fully processed mRNA. This communication ensures that the mRNA is fully processed before translation begins. Circularization occurs when eIF4E bound at the 5' terminus and poly(A) binding protein (PABP) bound at the 3' poly(A) tail both interact with eIF4G (**Figure 22-6a**). Initiation of translation requires the recruitment of eIF4G by eIF4E, through a conserved motif in eIF4G that allows interaction with eIF4E. The same motif is used by other proteins to repress translation of certain mRNAs.

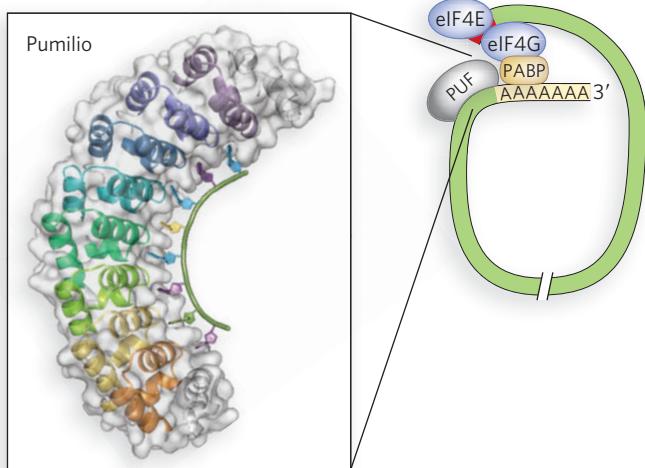
Regulation of translation by protein binding to the 3'UTR is especially important during the development of early embryos. Many maternal RNAs (deposited in the egg cytoplasm during oogenesis) have relatively short poly(A) tails (20 to 40 A residues), and their translation is suppressed until it is needed. Activation requires two sequences in the 3'UTR: the nuclear AAUAAA polyadenylation sequence and the cytoplasmic polyadenylation element (CPE, with consensus sequence UUUUUAUU). CPEs are bound by the protein CPEB (CPE binding protein), which helps establish translational masking of maternal mRNAs by interacting with a protein known, appropriately, as Maskin. Maskin also interacts with eIF4E, functioning much like 4E-BPs in preventing binding of eIF4E to eIF4G. An example of this mechanism is the regulation of a maternal mRNA in *Xenopus* embryos, transcribed from a gene called *mos* (Figure 22-6b). To activate the *mos* mRNA, CPEB is first phosphorylated. This stimulates interaction between CPEB and the AAUAAA-binding protein CPSF (cleavage and polyadenylation specificity factor).



**FIGURE 22-6 Functions of the 3'UTR and 5'UTR in translation initiation.** (a) eIF4E bound to the 5'UTR, and poly(A) binding protein (PABP) bound to the 3' poly(A) tail, are linked through their binding to eIF4G. The resulting looped complex facilitates recruitment of other translational regulators and ribosomal subunits for translation initiation. (b) For maternal mRNAs such as that derived from the *Xenopus* gene *mos*, eIF4E is initially sequestered by interaction with Maskin and CPE binding protein (CPEB), with translation suppressed. Activation involves phosphorylation of CPEB, which in association with the cleavage and polyadenylation specificity factor (CPSF) promotes lengthening of the poly(A) tail and recruitment of PABP and eIF4G; eIF4G displaces Maskin, allowing translation to start.

In turn, CPSF recruits the cytoplasmic polyadenylation enzymes that lengthen the poly(A) tail on the mRNA. The longer poly(A) tails enable the binding of multiple copies of PABP. PABP recruits eIF4G, and eIF4G displaces Maskin and interacts with eIF4E. Translation then begins.

In addition to CPEBs, other types of proteins bind the 3'UTR of mRNA to regulate translation, including proteins of the **PUF family** (named for *Pumilio* and *FBF*, the first two proteins of this type to be discovered). PUF proteins are a highly conserved family of RNA-binding proteins associated with translational control



**FIGURE 22-7 Structure and function of the PUF family protein Pumilio.** The protein is crescent shaped, with ten repeated  $\alpha$ -helical segments, each consisting of about 40 amino acids. The eight repeats in the middle (colored) are involved in mRNA binding. The two repeats on the ends (white) are important to Pumilio function, but do not directly participate in mRNA binding. Binding of a PUF family protein to the 3'UTR of an mRNA, as shown, suppresses protein production through interference with translation initiation by blocking assembly of translational factors, or through promotion of mRNA degradation by recruiting RNA degradation enzymes. [Source: PDB ID 1M8Y.]

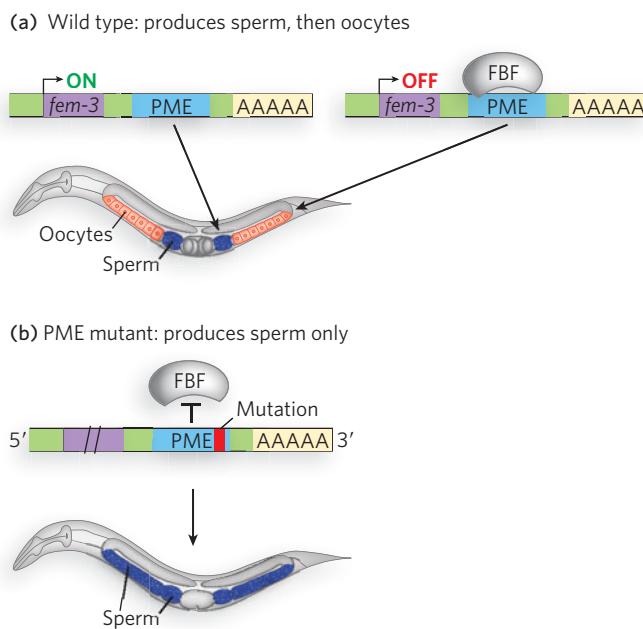
in a variety of organisms, from yeast to mammals. PUF proteins typically include eight consecutive 40-residue repeat sequences, each of which contains characteristic aromatic and basic amino acids. The crescent-shaped structure of PUF proteins reveals two extended surfaces (Figure 22-7). Based on the location and effects of mutations, one surface probably binds to mRNA and the other to other regulatory proteins.

Each PUF protein is thought to regulate multiple mRNAs, because experiments have demonstrated the proteins' ability to bind multiple targets. Researchers have engineered fruit flies to express a molecularly tagged version of Pumilio. The tagged protein was purified with its bound RNA partners, and the identities of the bound RNAs were determined using microarray technology (see Chapter 7). The study showed that many of the bound mRNAs shared a short, characteristic sequence in their 3'UTRs. Furthermore, the mRNAs tended to encode functionally related proteins, suggesting that Pumilio acts by binding to sets of mRNAs. Many of the proteins regulated by Pumilio function in developmental pathways considered in Section 22.5.

Once bound by the PUF protein, the targeted mRNAs are typically blocked from efficient translation

on the ribosome. Although the mechanism of repression is not completely known, PUF proteins seem to block initiation. In some cases, PUF proteins may also increase the rate of mRNA degradation.

In the nematode *Caenorhabditis elegans*, the hermaphrodite produces sperm and oocytes, successively, during development of the germ line. This cell fate decision is controlled by the PUF family protein FBF (*fem-3* binding factor), which binds to a site in the 3'UTR of the *fem-3* gene called the PME (point mutation element) (see How We Know). This binding site was defined by gene mutants first isolated in the laboratory of Judith Kimble (see Moment of Discovery). When FBF is not present in the germ-line cells, translation of *fem-3* transcripts proceeds and sperm cells are produced (Figure 22-8). When FBF protein is present, it binds the 3'UTR and blocks translation of *fem-3* transcripts, and oocytes are produced.



**FIGURE 22-8 The regulation of germ-line development in *C. elegans* hermaphrodites.** (a) Hermaphrodites produce first sperm cells, then oocytes. The transition is regulated by interaction of the FBF protein with the transcript of the *fem-3* gene. When FBF is absent, *fem-3* transcripts are translated and sperm are produced. When FBF is present, *fem-3* translation is blocked and oocytes are produced. (b) Mutations in the sequence PME (point mutation element) eliminate FBF binding. The mutant worms produce only sperm. The FBF protein is not drawn to scale; it spans a region much wider than the PME. [Source: Adapted from B. Zhang et al., *Nature* 390, 477–484, 1997.]

## Upstream Open Reading Frames Control the Translation of *GCN4* mRNA

Some eukaryotic genes are controlled by short open reading frames located upstream from the gene's authentic start codon. These **upstream open reading frames (uORFs)** do not produce functional protein. Instead, they are a gene regulatory mechanism that generally decreases translation by diverting ribosomes, often making them halt and dissociate before reaching the AUG start codon. It might seem that uORFs, rather than regulating levels of gene expression, would simply down-regulate genes under all conditions. However, the effectiveness of uORFs in terminating ribosome activity is altered by phosphorylation of eIF2 $\alpha$ . An instance of this type of regulation is observed for the yeast *GCN4* gene, encoding a transcription activator that regulates many other genes.

Although the phosphorylation of eIF2 $\alpha$  typically down-regulates translation initiation, in *S. cerevisiae*, low-level eIF2 $\alpha$  phosphorylation in response to amino acid starvation induces expression of the transcription factor Gcn4. Because Gcn4 activates transcription of at least 40 genes encoding amino acid biosynthetic enzymes, its induction alleviates nutrient limitation that could otherwise trigger more extensive eIF2 $\alpha$  phosphorylation and more general translational repression.

The mechanism of activation involves four short uORFs preceding the Gcn4-coding sequence of the mRNA. Located between 150 and 360 nucleotides upstream from the AUG start codon, the uORFs prevent ribosomes from initiating translation at the *GCN4* start site when nutrients are abundant. This occurs because ribosomes initiate efficiently at these "decoy" open reading frames instead of at the true protein-coding start site farther downstream (Figure 22.9). When eIF2 $\alpha$  is phosphorylated, however, ribosomes are much less likely to initiate translation in general, and thus have a greater propensity to continue scanning along a bound mRNA without forming an initiation complex. This circumstance favors initiation at the downstream *GCN4* start site, leading to expression of the Gcn4 protein.

## Translational Efficiency Can Be Controlled by mRNA Degradation Rates

Another important way in which translation is regulated in eukaryotic cells is by the degradation of mRNAs. Translation and degradation of particular mRNAs often show an inverse relationship. Those mRNAs that are efficiently translated tend to be stable in the cytoplasm, whereas those that are poorly translated tend to be degraded more quickly. The discovery of specific

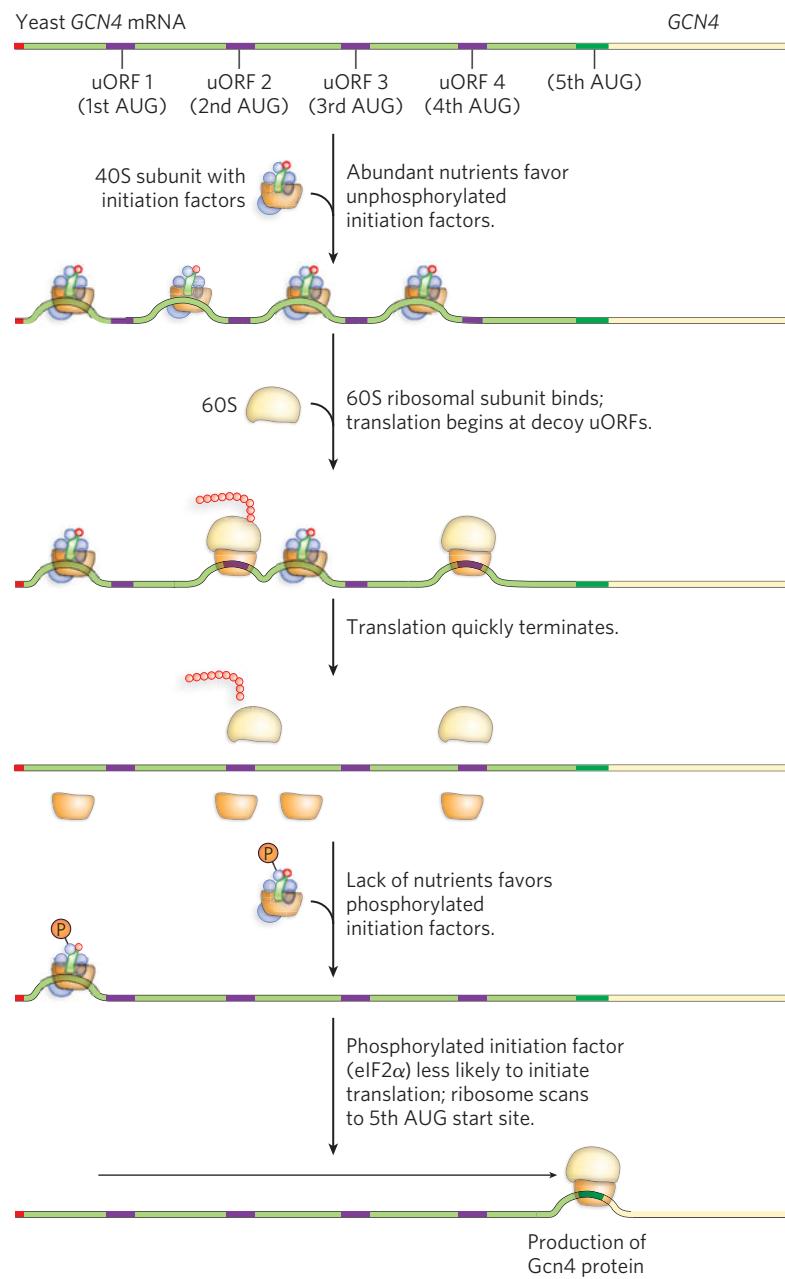
pathways of mRNA degradation has led to a better understanding of the close correlation between mRNA translation and turnover. Decay of mRNAs begins with removal of the 3' poly(A) tail (deadenylation), followed by removal of the 5' 7-meG cap (Figure 22.10). The mRNAs are then efficiently degraded by exonucleases, some in the 5' $\rightarrow$ 3' direction, others in the 3' $\rightarrow$ 5' direction.

Removal of the 5' cap involves the formation of macromolecular assemblies of translationally repressed mRNAs bound to decapping enzymes. Known as **processing bodies (P bodies)** (see Figure 16.26), these large cytoplasmic structures are easily seen by light microscopy in cells that have been starved or otherwise stressed to induce general translational repression.

Studies of the composition and formation of P bodies suggest that the rates of mRNA translation and degradation are influenced by the relative stability of the mRNA bound in polyribosomes and in P bodies. Regulatory proteins that bind sequences or structures within related groups of mRNAs serve either to recruit those mRNAs to P bodies or to stabilize their interaction with ribosomes. This is an exciting area of active research. P bodies are part of the conserved translational control machinery in eukaryotic cells, influencing patterns of gene expression in cells as diverse as oocytes and neurons.

### SECTION 22.2 SUMMARY

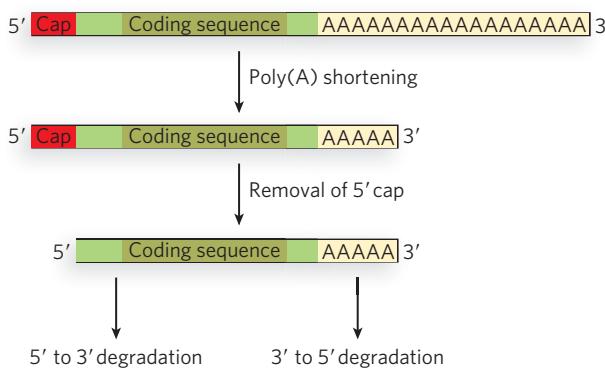
- Eukaryotes have at least three mechanisms of translational regulation: phosphorylation-dependent repression of initiation factors, direct binding of mRNAs by translational repressors, and disruption of the required interaction between eIF4E and eIF4G.
- Reversible phosphorylation of initiation factors plays a central role in regulating translation initiation in response to cellular conditions. When its eIF2 $\alpha$  subunit is phosphorylated, eIF2 forms a stable complex with eIF2B, making these factors unavailable for further rounds of translation initiation.
- The CPE binding protein binds sequences in the 3'UTR of maternal mRNAs and contributes both to suppressing translation (in conjunction with Maskin, a 4E-BP-like protein) and activating translation (through interaction with CPSF).
- All PUF proteins have characteristic sequence and structural features that enable their binding to the 3'UTRs of mRNAs. Binding leads to translational repression and/or degradation of the transcript, reducing expression levels.



**FIGURE 22-9 Translational regulation of *GCN4* in yeast by upstream ORFs.** Unphosphorylated eIF2, present when nutrients are abundant, leads to frequent translation initiation at upstream ORFs (uORFs) and little initiation at the *GCN4* gene (first three steps). When eIF2 $\alpha$  is

phosphorylated, translation initiation in general, including at the uORFs, is reduced; initiation at *GCN4* is now more likely (last two steps). [Source: Adapted from M. Holcik and N. Sonenberg, *Nat. Rev. Mol. Cell Biol.* 6:318–327, 2005.]

- In yeast cells, upstream open reading frames (uORFs) in the *GCN4* mRNA, preceding the Gcn4-coding sequence, act as ribosome decoys. They prevent ribosomes from initiating translation at the *GCN4* start site when nutrients are abundant, allowing ribosomes to initiate efficiently at the uORFs instead of at the authentic start site farther downstream.
- Efficiently translated mRNAs tend to be stable in the cytoplasm; those that are poorly translated tend to be rapidly degraded. The decay of an mRNA involves removal of the 3' poly(A) tail, decapping of the 5' end, then exonuclease-catalyzed degradation.
- Cytoplasmic foci, known as P bodies in yeast, are sites of mRNA decapping and degradation.



**FIGURE 22-10** Eukaryotic mRNA decay. The initial steps involve removal of first the 3' poly(A) tail, then the 5' cap. Exonuclease-catalyzed degradation of the remaining RNA is then rapid.

## 22.3 The Large-Scale Regulation of Groups of Genes

Although early research on gene regulation often focused on the mechanism by which one specific gene could be controlled, it has become increasingly clear that cells regulate large numbers of their genes together, to bring about changes in cell fate. How this works is an area of active research and discovery. In fact, the new field of systems biology is devoted to understanding how whole networks of genes are controlled and how they function coordinately in cells and organisms. We present here three specific examples of multigene control in which posttranscriptional regulation is a key element.

### Some Sets of Genes Are Regulated by Pre-mRNA Splicing in the Nucleus

Most eukaryotic genes, unlike bacterial genes, contain introns. Recent experiments conducted on *Saccharomyces cerevisiae* (budding yeast; referred to by nonscientists as baker's yeast) suggest that pre-mRNA splicing is regulated in response to alterations in nutrient availability. Yeast cells grown in a nutrient-rich medium then shifted to a medium lacking essential amino acids rapidly reduce their level of splicing of transcripts for ribosomal proteins. This was detected by isolating total RNA from the cells at different time points during growth and starvation, and using DNA microarrays to determine changes in the mRNA population. Notably, only the r-protein gene transcripts were affected by the nutrient shift; the splicing of other transcripts was maintained, or even enhanced (Highlight 22-1).

Although the mechanism of such targeted splicing regulation has yet to be worked out, it seems that, like transcription, splicing provides eukaryotic cells with a means of rapidly adjusting expression levels of sets of genes in response to environmental triggers. This may be an important reason that eukaryotic cells have acquired and maintained introns over evolutionary time.

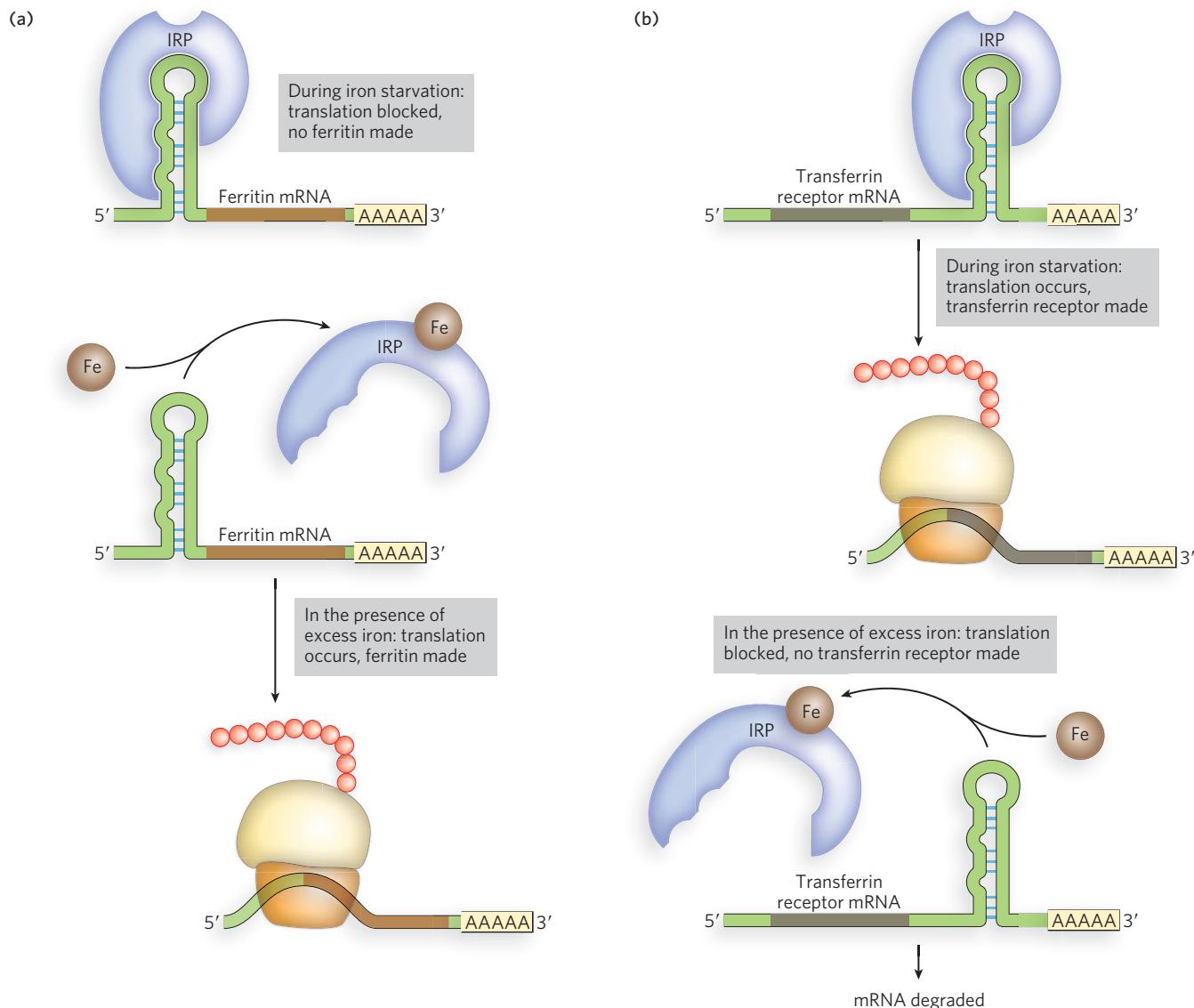
### 5'UTRs and 3'UTRs Coordinate the Translation of Multiple mRNAs

Sequence or structural elements in the untranslated regions of mRNAs provide a mechanism by which sets of transcripts can be controlled. One of the best-studied examples involves the regulation of iron concentration in mammalian cells. Although an essential element for cellular function, iron is also highly toxic, and its intracellular concentration must be carefully regulated.

A steady-state level of iron, or **iron homeostasis**, is achieved through the coordinated expression of proteins responsible for iron uptake, storage, utilization, and export. **Iron response proteins (IRPs)** bind to a hairpin structure called the **iron response element (IRE)** in the 5'UTR or 3'UTR of the mRNAs encoding the proteins for iron uptake (the transferrin receptor), storage (ferritin), utilization (aconitase), and export (Fpn1). IRP-IRE complexes formed in the 5'UTR of an mRNA inhibit translation, probably by physically blocking access to ribosomes. In contrast, IRP-IRE complexes formed in the 3'UTR, as is the case for the transferrin receptor, prevent mRNA degradation (Figure 22-11). The IRE-binding activity of IRPs is high in iron-deficient cells and low in iron-rich cells. This switch in binding affinity results from iron-sulfur clusters that assemble in the IRPs and thereby block IRE binding only when iron is abundant. In this way, groups of genes in the iron-response pathway can be controlled together to bring about rapid changes in the relevant protein levels as cellular iron concentrations change.

### Conserved AU-Rich Elements in 3'UTRs Control Global mRNA Stability for Some Genes

Large sets of genes can also be regulated by elements that affect the stability of RNA transcripts. Unlike bacterial mRNAs, many eukaryotic mRNAs are stable for hours or days. However, certain eukaryotic genes produce mRNAs with very rapid degradation rates. These genes are among those affected by a system for controlling mRNA stability that involves specific sequences in the 3'UTRs.



**FIGURE 22-11** The function of iron response elements (IREs). In mammalian cells, IREs are bound by iron response proteins (IRPs) when iron levels are low, blocking the production of proteins that use and store iron (including ferritin), while promoting the production of proteins that facilitate iron uptake (the transferrin receptor). (a) IRP

binding to the 5'UTR of ferritin mRNA inhibits translation (top). When iron is in excess (bottom), ferritin is produced. (b) IRP binding to the 3'UTR of the transferrin receptor mRNA stimulates translation (top); no transferrin receptor is made when iron is in excess (bottom).

In mammalian cells, mRNA sequence elements rich in A and U nucleotides, called **AU-rich elements (AREs)**, target those mRNAs for rapid degradation. The most common ARE motif is AUUUA, and it is often part of a larger AU-rich region such as WWWUAUUUAUUUW, where W is A or U.

ARE-containing mRNAs typically encode proteins that regulate either cell growth or an organism's response to infection, inflammation, and environmental stimuli. For example, mRNAs transcribed from proto-oncogenes

(often encoding proteins that promote cell division) have a very short half-life in the cell. In resting, or unstimulated, cells, ARE-dependent degradation ensures there is only low-level expression of these potent proteins. When cells respond to particular signals, ARE-containing mRNAs are bound by certain RNA-binding proteins, especially those in the ELAV (*embryonic lethal abnormal visual*) family, first discovered in *Drosophila*. Thus bound, the mRNAs become more stable, and the expression levels of the proteins they encode increase. Notably, in cancerous

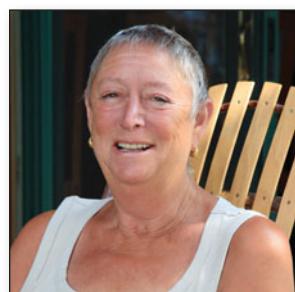
**HIGHLIGHT 22-1 EVOLUTION****Regulation of Splicing in Response to Stress**

The removal of introns from eukaryotic pre-mRNAs must occur before the mRNAs can be translated into protein. In higher eukaryotes, most genes have introns, and splicing regulates both when and where proteins are made. Furthermore, alternative splicing can produce different mRNAs and thus distinct proteins from a single initial transcript. In *S. cerevisiae*, however, only 5% of genes contain introns, and with few exceptions these genes have just one intron. All splice site sequences in yeast introns are very similar, and alternative splicing is rare. These observations led Christine Guthrie and her colleagues at the University of California, San Francisco, to wonder about the role of introns in the yeast genome: are introns maintained for a reason, perhaps to help cells respond to stressful situations?

To investigate this possibility, Guthrie and her students first examined the yeast genes that contain introns. They noticed that this set of genes includes many that code for metabolic regulators, such as enzymes that control the uptake and use of nutrients, and is also high in ribosomal protein genes (RPGs). In fact, of the 139 RPGs encoded in the yeast genome, 102 are interrupted by at least one intron. The RPGs are thus the largest functional category of intron-containing yeast genes.

Might these introns be playing a regulatory role in the expression of r-proteins? There were some

hints that this might be the case. Previous research had shown that when yeast is starved for amino acids, which are essential for synthesizing new proteins, RPG transcription is suppressed. In addition, rRNA production and overall protein synthesis are reduced, whereas



**Christine Guthrie** [Source: Courtesy of Christine Guthrie.]

the production of enzymes required for amino acid biosynthesis is increased. Guthrie hypothesized that the splicing of RPG transcripts might be regulated in response to amino acid starvation.

To test this idea, Guthrie's research team used DNA microarrays to examine the transcript-specific splicing changes resulting from exposure to two unrelated but environmentally relevant stresses: amino acid starvation and ethanol toxicity. RNA was purified from yeast cells at different times after exposure to each of these stresses and hybridized to microarrays containing DNA fragments representing the yeast genome. This analysis showed that splicing of the majority of RPGs is inhibited within minutes of induced amino acid starvation. By contrast, exposure to high levels of ethanol, which is not known to induce a global repression of translation, has little effect on RPG transcript splicing. Instead, in response to ethanol stress, the splicing of a different set of transcripts is reduced, whereas the splicing efficiency of a third group of transcripts is enhanced. The specificity of these responses and the speed of their onset—within minutes of stress induction—imply that splicing provides an important means of regulating gene expression in response to environmental stresses. This capacity for transcription-independent regulation could explain the evolutionary retention of introns in these yeast genes.

Since this initial study, several genome-wide surveys using microarray technology have shown distinct patterns of alternative splicing in various fruit fly and mammalian tissues. Furthermore, these experiments have identified RNA sequences that correspond to specific differential splicing events. These observations suggest that proteins controlling specific splicing events rely on an RNA code in their target transcripts, and that this RNA code governs the correct set of differentially spliced mRNAs. Future research will focus on discovering the identities of these splicing regulators and understanding how they work together to bring about changes in pre-mRNA splicing in response to the cell's needs.

or chronically inflamed human cells, this ARE-dependent regulation goes awry. This can be caused by increased levels of ELAV family proteins, such as HuR ("Hu" refers to an antibody, sometimes expressed in tumor-associated

neurological disorders, which binds ELAV proteins). The increased mRNA stability conferred by the binding of HuR and related proteins allows increased expression of proteins that normally should not be present in cells,

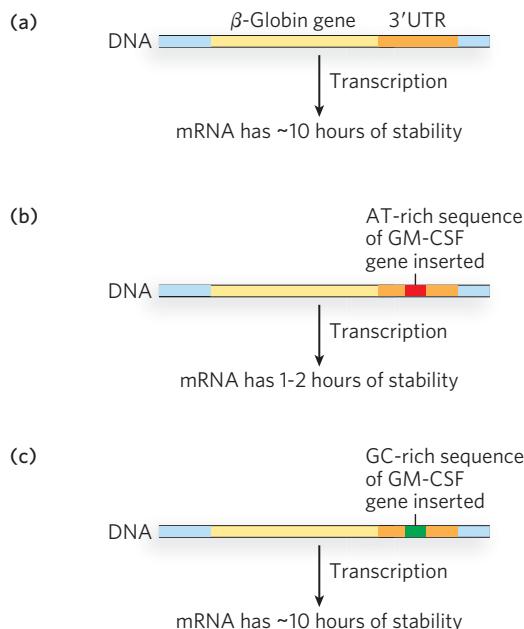
contributing to abnormal cell physiology and/or suppression of pathways that control cell growth.

One of the first demonstrations that an ARE sequence controls mRNA stability came from a study of genes for cytokines, cell proliferation factors produced principally by immune system cells that have very unstable mRNAs. In a comparison of the human and mouse gene sequences for a lymphokine (a type of cytokine) known as granulocyte-monocyte colony stimulating factor (GM-CSF), the most conserved sequence was outside the protein-coding region, in the 3'UTR. The protein-coding sequence was only 65% conserved, but a 51-nucleotide AT-rich sequence in the 3'UTR was 93% conserved, suggesting that the 3'UTR sequence plays a functional role. The sequence is also conserved in many other cytokine genes and proto-oncogenes. To test these 3'UTR sequence for function in mRNA stability, the AT-rich sequence in the 3'UTR of the GM-CSF gene was cloned into the 3'UTR of the human  $\beta$ -globin gene. The stability of the resultant mRNA was compared with that of the wild-type globin gene and that of a globin gene into which a GC-rich sequence (from the GM-CSF gene) was inserted. The AT-rich element in the 3'UTR caused a marked loss of mRNA stability (Figure 22-12).

With the human genome sequenced, we now know that 5% to 8% of human genes encode ARE-regulated mRNAs. Numerous ARE-binding proteins have been identified, in addition to the ELAV family proteins, and as in the IRP-IRE system described above, these ARE-binding proteins can bind multiple mRNAs. In contrast to ELAV family proteins, which seem to increase mRNA stability, some ARE-binding proteins recruit bound mRNAs to the cell's 3'-deadenylation and 5'-decapping machinery, leading to rapid degradation of the transcript. When cells require increased amounts of the proteins encoded by ARE-containing mRNAs, the levels of destabilizing ARE-binding proteins decline and binding efficiencies drop. A concomitant increase in ELAV family proteins may occur, thereby stabilizing ARE-containing mRNAs and ensuring more translation from those transcripts.

### SECTION 22.3 SUMMARY

- Cells often regulate large sets of genes together to bring about changes in cell fate.
- Pre-mRNA splicing provides one mechanism by which cells control the expression of sets of genes. In yeast, splicing of many r-protein RNA transcripts is regulated in response to changes in nutrient availability.
- Both mRNA stability and translational efficiency are controlled by elements in the untranslated regions of some mRNAs.
- In mammalian cells, the expression of proteins involved in iron homeostasis is controlled by the iron response element (IRE), a structure in the 5'UTR or 3'UTR of specific mRNAs. In the presence of iron, IREs bind iron response proteins (IRPs), triggering increased or decreased expression, depending on the location of the IRE in the mRNA transcript.
- About 5% to 8% of mammalian mRNAs contain AU-rich elements (AREs) in their untranslated regions that bind ARE-binding proteins, leading to either stabilization (if ELAV family proteins) or rapid degradation (if destabilizing ARE-binding proteins) of the mRNA. When cells require a change in ARE-regulated protein levels in response to infection or other stimuli, ARE-binding activity is altered and expression levels increase or decrease.



**FIGURE 22-12** AU-rich elements (AREs) affect mRNA

**stability.** AREs are binding sites for certain proteins that either stabilize or destabilize the bound mRNA. (a) mRNA transcribed from the  $\beta$ -globin gene is stable for about 10 hours. (b) When a single AT-rich ARE sequence is introduced into the globin gene, such as that derived from the GM-CSF gene, mRNA stability is greatly reduced. (c) When a GC-rich sequence is introduced into the gene, mRNA stability is restored.

## 22.4 RNA Interference

Thus far, we have been discussing mechanisms of gene regulation that involve regulatory proteins. These proteins bind to DNA or RNA targets and bring about changes in the efficiency of transcription or translation



**FIGURE 22-13 RNAi pathway effects: petal color selection in petunias.** In these petunia flowers, genes for pigmentation are silenced by RNAi. The flower on the left is the wild type; the other two flowers are from transgenic plants with inserted genes that produce siRNAs complementary in sequence to an endogenous gene required for the development of flower color. This gives rise to the unpigmented white areas of the petals. [Source: M. A. Matzke and A. J. M. Matzke, *PLoS Biol* 2(5):e133, 2004, doi: 10.1371/journal.pbio.0020133. PMID 15138502.]

in response to various stimuli. However, an entirely distinct mode of gene regulation became apparent in experiments initially carried out to examine petal color in petunias (Figure 22-13), as well as experiments on nematodes. These studies revealed the existence of a regulatory mechanism involving small RNA molecules—**RNA interference (RNAi)**. Researchers discovered that when exogenous RNA was introduced into a eukaryotic cell, either experimentally or naturally, by infection with an RNA virus, the RNA was processed into 21- to 27-nucleotide species known as short interfering RNAs (siRNAs); these mediated the silencing of certain genes. Although RNAi by siRNAs was the first RNA-mediated gene silencing pathway to be discovered, researchers soon found that silencing can also occur in pathways involving small RNAs called microRNAs (miRNAs), encoded by the cells themselves. The two gene silencing mechanisms are very similar—siRNAs make use of cellular machinery associated with the endogenous miRNAs—but have some important differences. In either case, though, these small RNAs function by base pairing with mRNAs, often in the 3'UTR, which results in mRNA degradation or translation inhibition. In some cases they can also repress transcription of the targeted mRNA.

RNA interference and related pathways control developmental timing in some organisms. They are also used as a mechanism to protect against invading RNA viruses (especially important in plants, which lack an immune system) and to control the activity of transposons. Small RNA molecules also play a critical, although still undefined, role in the formation of heterochromatin. In addition, RNAi has become a powerful tool for molecular biologists and has attracted attention as a potential therapeutic approach.

## MicroRNAs Encoded in Eukaryotic Genomes Target mRNAs for Gene Silencing

In the 1980s and 1990s, numerous experiments were under way, on a variety of organisms, to use DNA or RNA oligonucleotides complementary, or antisense, to an mRNA to block protein expression. The hypothesis was that base pairing between the antisense oligonucleotide and the target mRNA would prevent recognition by the translation machinery, or lead to degradation of the hybrid complex, or both. Experiments in plants, however, showed that many transgenic plants containing an artificial gene encoding an antisense RNA failed to suppress expression of the corresponding endogenous gene. Furthermore, in both plants and nematodes, “control” experiments in which the sense strand, rather than the antisense strand, of RNA was introduced into cells often showed just as much suppression of the targeted gene as did experiments using the antisense strand. Careful analysis of these phenomena in the nematode system by Craig Mello and Andrew Fire revealed a fascinating explanation for these puzzling observations. The observed RNAi in *C. elegans* resulted from the presence of small amounts of double-stranded RNA that contaminated the preparations of sense or antisense RNA injected into the worms. Researchers subsequently demonstrated in plants, worms, and other eukaryotes that the double-stranded siRNAs were much more efficient at inducing gene silencing than were single-stranded sense or antisense RNAs (Figure 22-14). Mello and Fire shared the 2006 Nobel Prize in Medicine or Physiology for this discovery.

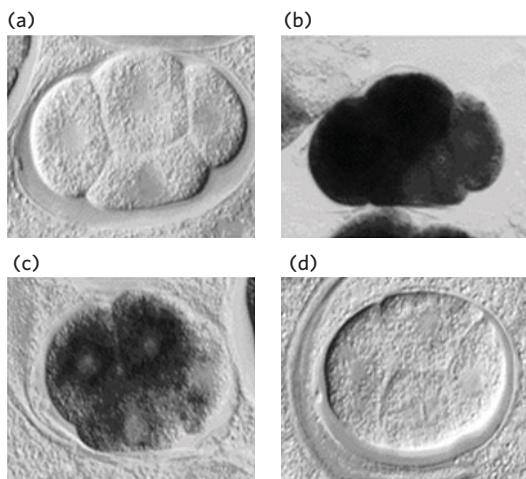
Further experiments in plants, nematodes, fruit flies, and mammals revealed many endogenous, or naturally occurring, small RNAs—**microRNAs (miRNAs)**—that correspond to sequences in cellular mRNAs. In fact, hundreds of different miRNAs have been identified in higher eukaryotes (see How We Know). They are transcribed by Pol II, or in some cases Pol III, as



**Craig Mello** [Source: Courtesy of Craig Mello.]



**Andrew Fire** [Source: Linda A. Cicero/Stanford University News Service.]



**FIGURE 22-14** Silencing of gene expression by RNAi in a nematode. Mex-3 protein is a regulator that is expressed and functions early in nematode development. Four early embryos are shown at identical stages of development, with *in situ* hybridization used to detect the presence of *mex-3* transcripts. (a) No staining is seen in a control, with no hybridization probe added. (b) Staining reveals the normal pattern of *mex-3* expression in the embryo. (c) Injecting an embryo with an RNA complementary (antisense) to *mex-3* mRNA reduces gene expression somewhat. (d) Injecting an embryo with double-stranded RNA corresponding to *mex-3* reduces gene expression dramatically. [Source: A. Fire et al., *Nature* 391, 806–811, 1998, Fig. 3.]

**primary miRNA transcripts (pri-miRNAs)** with one or more sets of internally complementary sequences that can fold to form hairpinlike structures. The pri-miRNAs are cleaved by the nuclear endonuclease Drosha, a member of the ribonuclease III family of enzymes, to produce shortened hairpins—60 to 70 nucleotides long—with a 5' phosphate and a two-nucleotide 3' overhang (Figure 22-15). These partially processed **precursor miRNAs (pre-miRNAs)** then bind to export receptor proteins and are transported from the nucleus to the cytoplasm for further processing.

Once in the cytoplasm, pre-miRNAs are cleaved by Dicer, another ribonuclease III family member, to generate a 20- to 22-nucleotide miRNA paired with its complementary sequence. Dicer is part of a larger complex that includes a protein called Argonaute. The overall complex is called the **RNA-induced silencing complex (RISC)**. After cleavage, the miRNA is unwound and the unneeded strand is discarded. The strand complementary to the target is delivered to particular mRNAs. Note that complementarity between the miRNA and the targeted mRNA is typically imperfect, with one or more mismatched or unmatched bases within the duplex. These mismatches usually occur two to eight nucleo-

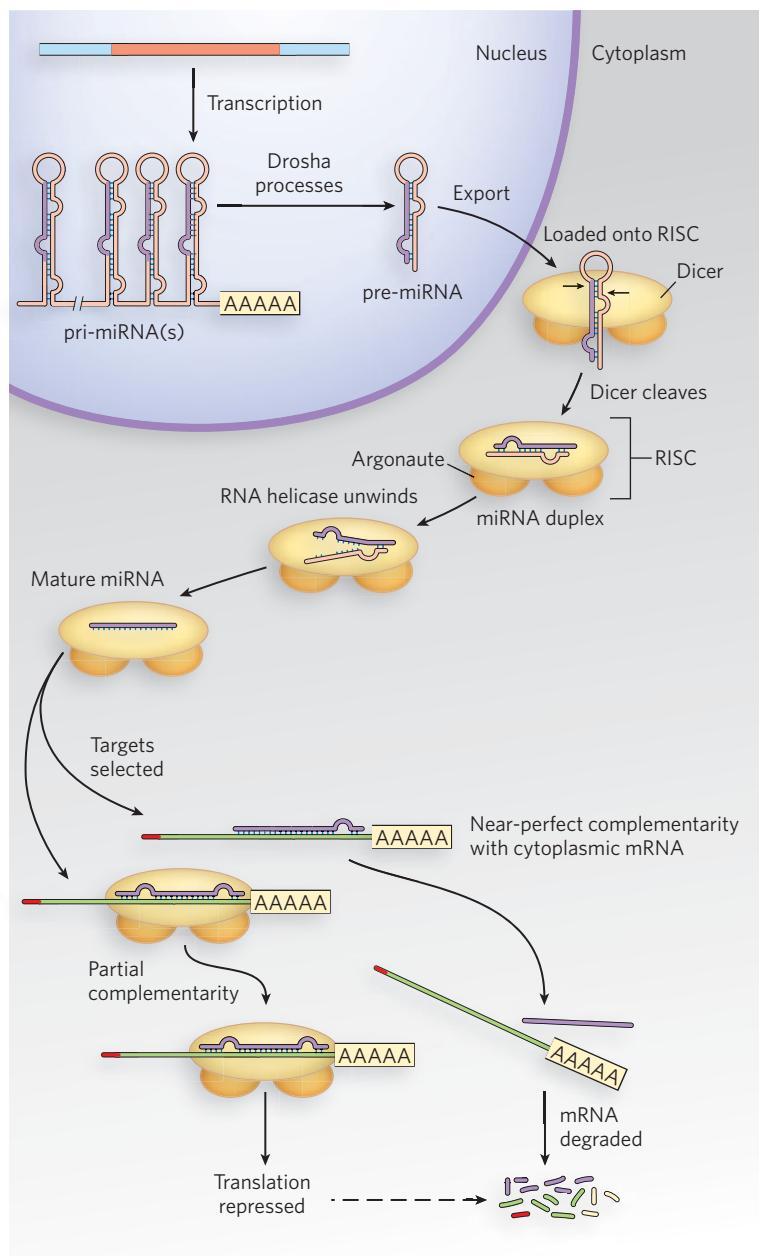
tides downstream from the 5' end of the miRNA. The nucleotides at the 5' end are called the “seed” region and must be perfectly base-paired for efficient miRNA targeting. In animals, the resulting miRNA-mRNA-protein complex somehow triggers RISC to inhibit translation of the bound mRNA, through a process that probably blocks translation initiation. In plants, miRNAs typically induce RISC-mediated cleavage of the targeted mRNA, leading to subsequent degradation.

Although the core components of RISC are conserved across organisms, some interesting differences have been noted. For example, in fruit flies and nematodes RISC activation requires ATP hydrolysis, whereas in mammalian systems it does not. Also, some of the proteins found in purified RISC are unique to a specific system, suggesting that RISC may be fine-tuned for function in different situations. Also of interest is that Dicer and Argonaute proteins sometimes occur in multiple isoforms, depending on the organism. Humans have just one Dicer enzyme and four Argonaute proteins, whereas nematode worms have two Dicers and many (more than 20) different Argonautes. Plants have at least eight distinct forms of Dicer! Although there may be some functional redundancy among these different protein family members, they may also play discrete roles in the implementation of RNAi-mediated control of gene expression.

As the astute reader will have noticed, we have not mentioned yeast in the RNAi discussion thus far. This is because *S. cerevisiae*, used so prominently in the study of many other aspects of eukaryotic gene regulation, does not seem to contain any of the enzymatic machinery required for RNAi. Furthermore, attempts to induce RNAi-mediated gene silencing in budding yeast have failed. However, other single-celled eukaryotes, such as fission yeast (*Schizosaccharomyces pombe*) and the human pathogen *Giardia intestinalis*, do express Dicer and Argonaute proteins. At least in fission yeast, small endogenous RNAs are important for silencing centromeric sequences through heterochromatin formation. Why *S. cerevisiae* lost the capacity to use RNAi as a gene regulatory mechanism, or never acquired it, remains a fascinating question.

### Short Interfering RNAs Target mRNAs for Degradation

In addition to processing pre-miRNA transcripts exported to the cytoplasm, Dicer can also recognize and cleave long double-stranded RNAs. Double-stranded RNAs can arise naturally from viral infection, or they can be introduced by experimenters. The resulting diced products, 21 to 27 bp in length (depending on the

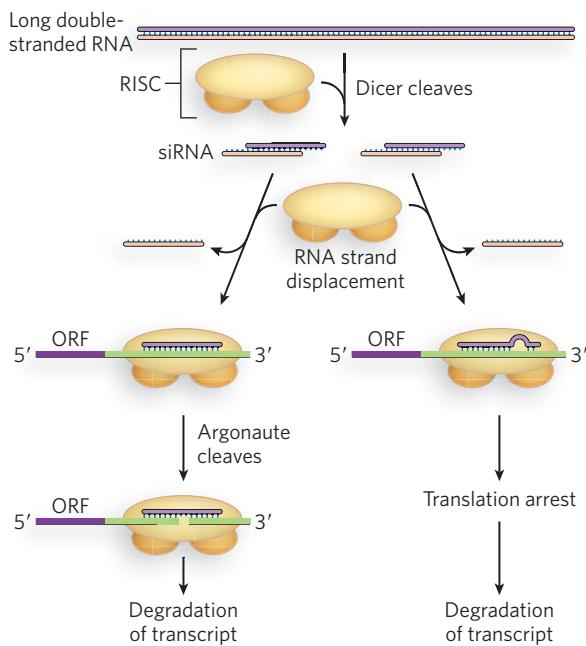


**FIGURE 22-15** **Drosha and Dicer process precursors to miRNAs.** Processing of the transcribed pri-miRNA precursor by Drosha begins in the nucleus, generating a hairpin-shaped pre-miRNA that is exported to the cytoplasm. Dicer, as part of the larger RISC complex, continues the processing in the

cytoplasm to generate a short duplex miRNA. One strand of the RNA is delivered to the target mRNA by the RISC-miRNA complex, leading to translational repression of the targeted mRNA. In plants, the targeted mRNA is often cleaved and degraded.

organism) with two-nucleotide 3' overhangs, are known as **short interfering RNAs (siRNAs)**, and they can regulate gene expression by mechanisms similar to those described for miRNAs. In the cytoplasm, siRNAs can form perfect or imperfect base pairings with targeted mRNA or viral RNA. In either case, these RNA duplexes are contained within the RISC protein machinery that includes Argonaute and multiple

additional factors. Perfect base pairing between the siRNA and its target triggers Argonaute-catalyzed cleavage of the RNA duplex, causing eventual degradation of the targeted RNA through the normal pathways, involving 3' deadenylation, 5' decapping, and subsequent exonucleolytic cleavage (Figure 22-16). Imperfect pairing leads to translation repression, as described for miRNAs, and sometimes to degradation. In some

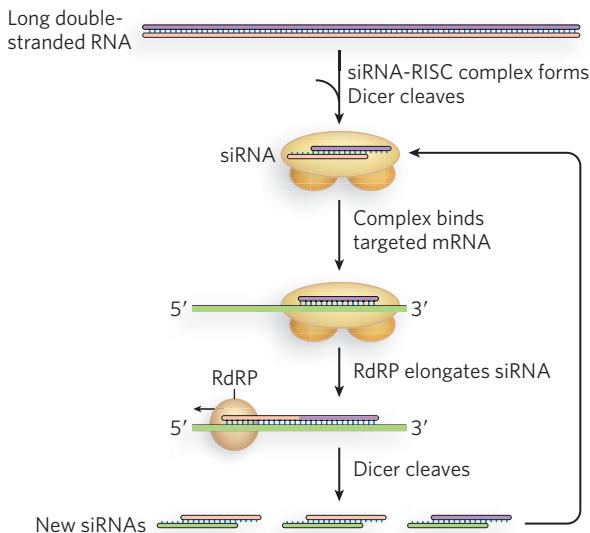


**FIGURE 22-16 Alternative fates of siRNA-targeted mRNAs.** Once the mRNA has been targeted, the degree of siRNA-mRNA base pairing leads to either Argonaute-mediated degradation of the mRNA (left) or translational repression, followed by degradation of the mRNA (right).

cases, siRNAs can enter the nucleus and induce heterochromatin formation at promoters of targeted genes, providing yet a third mechanism of gene silencing.

Much of our current understanding of siRNA function comes from experiments in nematodes, in part because of the ease and efficiency of conducting such experiments in these worms. For example, worms can be fed a diet of bacteria engineered to express specified siRNA precursors, leading to the silencing of expression from siRNA-complementary mRNAs. This technique has been used in many investigations of the siRNA sequence, target site, and phenotypic consequences of siRNA-mediated gene silencing.

Another interesting aspect of RNAi in *C. elegans* is its extraordinary efficiency. Small amounts of siRNAs are sufficient to trigger almost complete silencing of target genes. RNA-dependent RNA polymerases have been implicated in the use of siRNAs to direct the synthesis of RNAs complementary to the targeted mRNAs, with increased numbers of siRNAs then resulting from Dicer-mediated cleavage of the new duplex RNAs (Figure 22-17). This amplification mechanism has not been detected in fruit flies or mammals, but does seem to occur in fission yeast (*S. pombe*). Furthermore, siRNA-mediated gene silencing in *C. elegans* can be inherited epigenetically. This means that worms in which one or more siRNAs have silenced the expression of a gene will produce offspring in which that gene remains silenced.

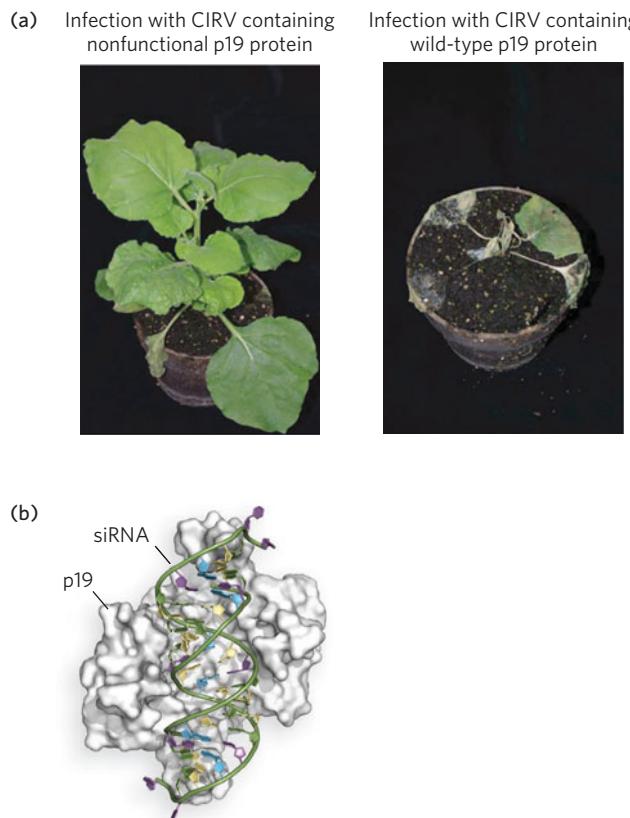


**FIGURE 22-17 Amplification of siRNA-mediated gene regulation in nematodes by RNA-dependent RNA polymerases.** The siRNAs target an mRNA by normal mechanisms. The annealed siRNA serves as a primer for RNA-dependent RNA polymerase (RdRP), which creates a longer double-stranded RNA. This is then used as a source of more siRNAs.

Genetic and biochemical experiments have shown that an RNA-specific membrane channel is responsible for the transport of siRNAs from parent to fertilized oocyte, where the siRNAs are perhaps maintained by RNA-dependent RNA polymerase-mediated amplification. So far, this ability to pass siRNAs from one generation to the next has not been observed in other organisms.

### RNAi Pathways Regulate Viral Gene Expression

Some lines of evidence suggest that RNAi originally evolved to suppress the replication of viruses and transposable elements that use double-stranded RNA as a replication intermediate. For example, plant viruses have acquired various mechanisms of suppressing the RNAi machinery in host cells. Most plant viruses have single-stranded RNA genomes that replicate through double-stranded RNA intermediates and thereby trigger RNA silencing in infected cells. On incorporation of viral-derived siRNAs into the cell's RISC machinery, the viral RNA is specifically targeted for degradation. Genetic experiments have identified numerous viral genes that can limit this effect. In tombusviruses, one of which infects carnations, a small viral protein called p19 was found to play a role in suppressing gene silencing. Traci Hall and her colleagues found that p19 binds specifically to siRNAs, based on their length and the presence of a

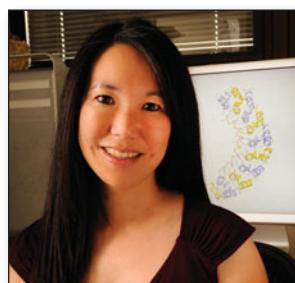


**FIGURE 22-18** The binding of tombusvirus p19 to siRNAs and blocking of RISC assembly. (a) Plants infected with CIRV (a tombusvirus) that does not (left) or does (right) express the protein p19. (b) The structure of p19 bound to siRNA. [Source: (a) J. M. Vargason et al., *Cell* 115:799–811, 2003, Fig. 4, parts a and c. Copyright 2003 by Cell Press. (b) PDB ID 1RPU.]

two-nucleotide 3' overhang. This p19-siRNA complex cannot assemble into RISCs to direct siRNA-mediated degradation of the viral RNA (Figure 22-18).

In an interesting twist on these plant viral siRNA-inhibitory mechanisms, experiments with hepatitis C virus (HCV) show that the virus takes advantage of a cellular miRNA to stimulate replication. HCV infects human liver cells, which express an endogenous miRNA

known as miR-122. HCV is unable to replicate in cells that do not express miR-122. Moreover, HCV is unable to replicate efficiently in cultured liver cells pretreated with oligonucleotides having the antisense sequence to miR-122. Such engineered oligonucleotides, known as antagomirs, have been shown to repress viral replication in HCV-infected chimpanzees, and could

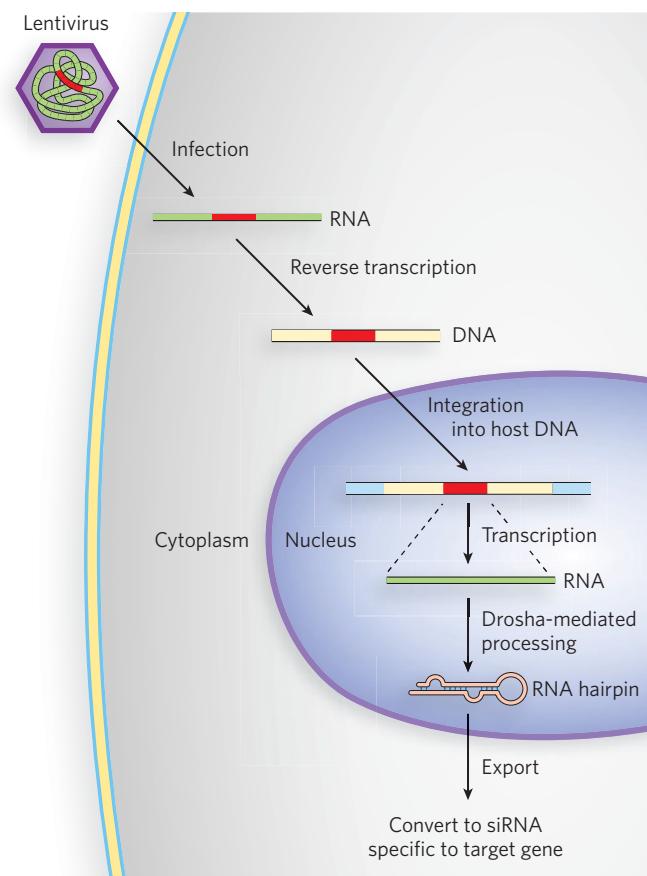


**Traci M. T. Hall** [Source: Steven R. McCaw/Image Associates. Courtesy of Traci Hall.]

potentially be developed into future antiviral therapeutics (Highlight 22-2).

### RNAi Provides a Useful Tool for Molecular Biologists

The discovery of RNA interference has an interesting and very useful practical side. If an investigator introduces into an organism a double-stranded RNA corresponding in sequence to virtually any mRNA, the Dicer endonuclease cleaves the duplex into siRNAs. These bind to the mRNA and silence it. Depending on the gene and the site(s) selected for siRNA targeting, such expression “knockdown” results in a 50% to 90% reduction in the levels of protein produced from that gene (Figure 22-19). Removal of a protein is a powerful way to investigate that protein’s



**FIGURE 22-19** The siRNA-based knockdown of gene expression in human cells. Lentiviruses, RNA viruses that infect human cells, are used as vectors. A sequence encoding a hairpin homologous to the target gene is cloned into the lentiviral vector (red). In the host cell, the viral RNA is converted to duplex DNA by reverse transcriptase, and the DNA is integrated into the cellular genome. Transcription of the integrated gene produces a hairpin RNA that is converted to siRNA specific to the target gene.

## HIGHLIGHT 22-2 MEDICINE

### Viral Takeover Using a Cell Type-Specific miRNA

Human tissues contain specific microRNAs that can base-pair with complementary sequences in mRNAs to trigger translational silencing and degradation. Even certain animal viruses that have RNA genomes encode miRNAs, indicating that viruses can use these small regulators during their life cycle. But how might viruses deal with, or even take advantage of, a host cell's miRNAs during an infection?

Many viruses infect and replicate in only certain tissue types, and the hepatitis C virus (HCV) is a case in point. This dangerous human pathogen targets liver cells selectively, where it can cause cirrhosis and sometimes cancer. Analysis of cultured human liver cells shows that the liver specifically expresses an miRNA called miR-122; in fact, miR-122 constitutes about 70% of the total miRNA population in liver cells. Stanford University virologist Peter Sarnow and his colleagues wondered how miR-122 regulates mRNA function in both normal and virally infected cells. They first investigated miR-122 expression levels in various tissue and cell types. Using Northern blotting (see Figure 6-32), in which miR-122 was detected using radiolabeled complementary DNA oligonucleotides, the investigators detected miR-122 in mouse and human liver, and in cultured mouse and human liver (Huh7) cells, but not in human cervical carcinoma-derived HeLa cells or even in human liver-derived HepG2 cells.



**Peter Sarnow** [Source: Courtesy of Peter Sarnow.]

Sarnow's research team next tried infecting these different tissues and cultured cells with HCV. Although both Huh7 and HepG2 cells are derived from human liver, HCV could replicate only in Huh7 cells. To find out whether this could be related to the presence of miR-122 in Huh7 cells, the researchers analyzed the 9,600-nucleotide HCV RNA genomic sequence for potential miR-122 binding sites that might enable a successful miRNA-target mRNA interaction. They found two sites in the noncoding regions of the viral RNA genome that are complementary to the miR-122 sequence, and are conserved in many different isolates (genotypes) of HCV.

Mutation of those potential miR-122 binding sites in the viral RNA produced HCV variants that could not replicate efficiently in Huh7 cells. These mutant viruses could be "rescued" by introducing into the host Huh7 cells miR-122 variants that restored base-pair complementarity with the mutated viral RNA. Furthermore, the use of small complementary oligonucleotides to bind and sequester miR-122 in the host Huh7 cells dramatically reduced

cellular function. Compared with traditional genetic methods, RNAi provides a much easier way to alter gene expression for experimental or therapeutic purposes. The technique has rapidly become an important tool in ongoing efforts to study gene function, because it can disrupt functionality without creating a mutant organism.

In plants, virtually any gene can be effectively shut down in this way. In nematodes, simply introducing the double-stranded RNA into the worm's diet causes effective suppression of the target gene. The procedure can be applied to human cells as well. Laboratory-produced siRNAs have already been used to block HIV and polio-

virus infections in cultured human cells for a week at a time. Although this work is in its infancy, the rapid progress makes RNA interference a field to watch for future medical advances.

### SECTION 22.4 SUMMARY

- RNA interference (RNAi) silences gene expression by means of short interfering RNAs (siRNAs), using endogenous gene silencing pathways based on genome-encoded microRNAs (miRNAs).
- Endogenous to the cell, miRNAs are encoded by many eukaryotic genomes and cause

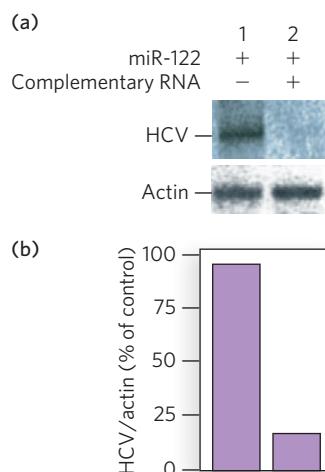


**Henrik Orum** [Source:  
Courtesy of Henrik Orum.]

synthesized a short oligonucleotide with a sequence complementary to miR-122. To ensure that this oligonucleotide would not be rapidly destroyed in the body, they introduced chemical modifications into the sequence to block the action of degradatory enzymes. When introduced into chimpanzees chronically infected with HCV, the modified “anti-miR-122” oligonucleotide, called SPC3649, produced long-lasting suppression of HCV replication, with no evidence of viral resistance or side effects in the treated animals. The prolonged protection against HCV infection afforded by SPC3649 hints at a new antiviral strategy that takes aim at the underlying mechanism used by HCV to take over human liver cells.

the amount of HCV RNA replication in these cells (**Figure 1**). Sarnow concluded that HCV requires the host cell’s miR-122 for efficient replication during infection. These findings suggest that miR-122 might present a good target for antiviral intervention.

To test this idea, Henrik Orum and his research team at Santaris Pharma in Denmark



**FIGURE 1** Hepatitis C virus replication is reduced in the absence of miR-122 RNA in the host cell. (a) When miR-122 was expressed in a cell harboring HCV, Northern blots showed production of HCV RNAs (lane 1). When an RNA complementary to miR-122 was also present, so as to sequester miR-122, HCV RNA production declined markedly (lane 2). Northern blots for a housekeeping gene, encoding actin, are also shown for comparison. (b) Quantification of the results of the same experiments. [Sources: C. L. Jopling et al., *Science* 309(5740):1577–1581, 2005, Fig. 3C.]

translational inhibition of the mRNAs to which they bind. Introduced by viral infection or experimentally, exogenous siRNAs use cellular mechanisms associated with miRNAs to silence the expression of particular genes, by suppressing translation or causing mRNA degradation.

- Hundreds of different miRNAs in higher eukaryotes are synthesized by Pol II or Pol III as primary miRNA transcripts (pri-miRNAs), about 70 to 90 nucleotides long. After initial processing in the nucleus, miRNAs are exported to the cytoplasm, where further processing leads to assembly into RNA-induced silencing complexes

(RISCs). In contrast, siRNAs are produced entirely in the cytoplasm; like miRNAs, they assemble into RISCs.

- RISCs facilitate base pairing between the small RNA and its complementary sequence in an mRNA, triggering translational arrest (when the pairing is imperfect) or mRNA degradation (when the sequences are perfectly paired).
- RNAi probably evolved as a mechanism to suppress infection by RNA viruses.
- Molecular biologists and clinicians use RNAi to alter gene expression for experimental and therapeutic purposes.

## 22.5 Putting It All Together: Gene Regulation in Development

For sheer complexity and intricacy of coordination, the patterns of gene regulation that bring about development from a zygote to a multicellular animal or plant have no peer. Development requires transitions in protein composition and in morphology that depend on tightly coordinated changes in expression of the genome.

How is a complex organism produced, with its many tissues and organs and appendages, from a single cell? Some clues can be found in that single cell—the fertilized egg. More genes are expressed during early development than at any other stage of the life cycle. For example, in the sea urchin, an oocyte (an immature egg cell) has about 18,500 *different* mRNAs, compared with about 6,000 different mRNAs in the cells of a typical differentiated tissue. The mRNAs in the oocyte give rise to a cascade of events that regulate the expression of many genes across both space and time.

The regulatory mechanisms used in development encompass all of the regulatory processes discussed in Chapter 21 and earlier in this chapter. Transcriptional regulation occurs, but posttranscriptional regulatory processes are particularly important.

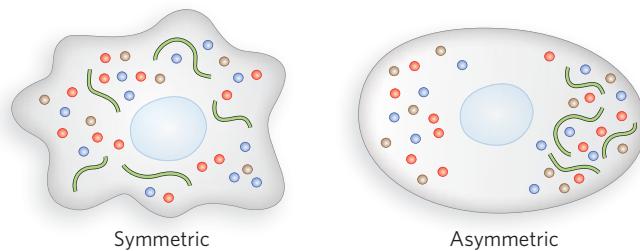
### Development Depends on Asymmetric Cell Divisions and Cell-Cell Signaling

If all cells divided to produce two identical daughter cells, multicellular organisms could never be more than a ball of identical cells. Programmed asymmetric cell divisions are required for different cell fates. Cell-cell signaling also helps guide the eventual differentiation of tissues and organs with various functions. Asymmetry in the developing embryo is thus created in several ways (Figure 22-20).

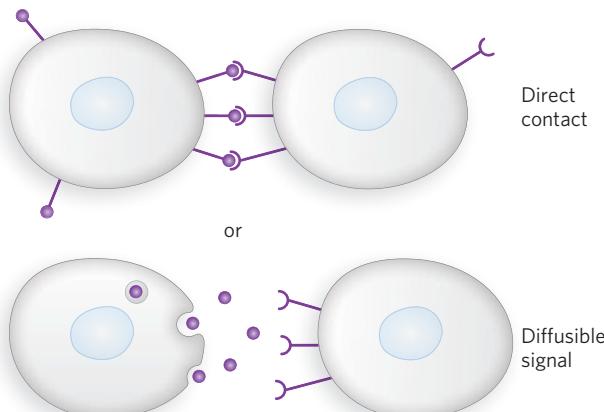
Asymmetry within the cells themselves takes the form of gradients of mRNAs and proteins that define critical axes (posterior-anterior, dorsal-ventral). In the developing oocyte, some gradients are established by deposition of mRNAs at one end or the other. Active transport in the cell can also contribute to generating a gradient. Fertilization can trigger events that create additional gradients in the fertilized egg (zygote). In many organisms, these gradients dictate different cell fates even for the daughter cells of the first cell division.

It is not enough to create a gradient in the cell, however. The mitotic spindle must also be aligned along the same axis as the gradient, so that the cell division

(a) Distribution of cellular components (mRNA and protein) creates intracellular asymmetry



(b) Cell-cell signaling creates extrinsic asymmetry



**FIGURE 22-20** Several ways to generate asymmetry in a developing embryo. (a) Intrinsic asymmetry reflects the existing distribution of cellular components, especially mRNA and protein. The asymmetries are either inherent to the developing oocyte or created during fertilization. (b) Extrinsic asymmetry is generated by cell-cell signaling. Although asymmetry is not necessarily created in any given cell, the cell-cell signals alter the fate of a cell or a group of cells in the embryo, contributing to embryonic asymmetry. The signals can involve direct cell-cell contacts or the action of a secreted, diffusible signal.

occurs on an axis perpendicular to the gradient (see Chapter 2 for a reminder of mitotic cell division). The proper alignment of the mitotic spindle in particular cell divisions is the function of some proteins critical to development.

In the developing embryo, cell-cell signaling generates additional asymmetry as development proceeds (see Figure 22-20). Direct contact between the lipids and glycoproteins on one cell surface and the receptors on another can guide changes in gene expression in the receptor-bearing cell. Some signals act at longer distances: diffusible molecules secreted by one cell or group of cells and detected by receptors on another,

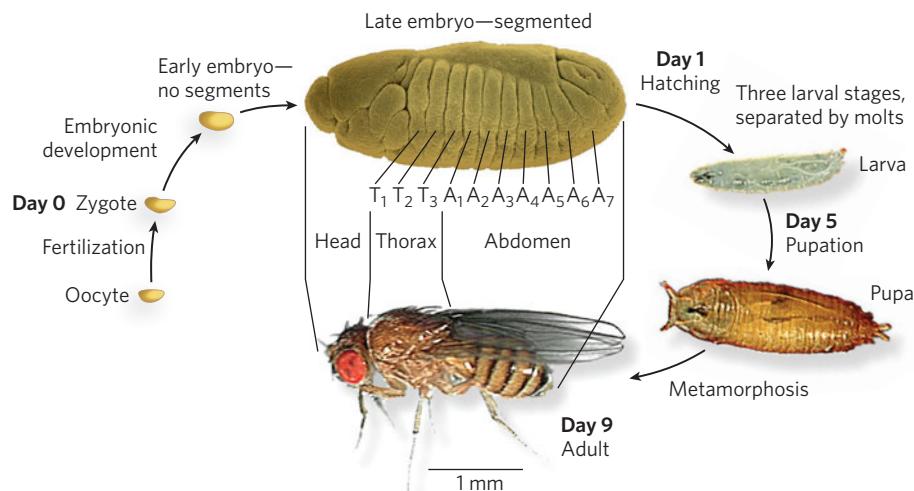
distant cell or group of cells. Ever more complex networks of signaling molecules and gene regulators are created as development proceeds.

**Characteristic Stages of Development** Several organisms have emerged as important model systems for the study of development, because they are easy to maintain in a laboratory and have relatively short generation times. These include nematodes, fruit flies, zebrafish, mice, and the plant *Arabidopsis thaliana* (see Model Organisms Appendix). The discussion here focuses on developmental pathways in the fruit fly. Our understanding of the molecular events during development of *Drosophila melanogaster* is especially well advanced and can be used to illustrate patterns and principles of general significance, and to highlight the mechanisms of gene regulation that govern this complex process.

Multicellular eukaryotes develop in a process that begins with the union of an egg and a sperm cell by fertilization, to create a zygote. The egg cell has been preprogrammed by the deposition of maternal mRNAs in gradients, such that concentrations of certain maternal mRNAs vary greatly from one end of the oocyte to the other. On fertilization, cell division begins. Early in development, the fate of particular cells is determined by the concentration of maternal mRNAs, as well as by the actions of regulatory genes. As

development proceeds, cascades of regulatory genes guide the various cell lineages as different tissue types develop. Although the regulatory genes are numerous, they generally fall into a small number of classes that are highly conserved, from nematodes to fruit flies to humans. Signaling pathways and processes are also highly conserved.

The life cycle of the fruit fly is relatively complex, and the patterns are conserved in a wide range of multicellular eukaryotes. Complete metamorphosis occurs during progression from embryo to adult fly (Figure 22-21). The final structure of the adult is forecast by features that are evident in the embryo at a very early stage. One of the most important characteristics of the embryo is its **polarity**: the anterior and posterior ends of the animal are readily distinguished, as are its dorsal and ventral surfaces. The fly embryo also exhibits the key characteristic of **metamerism**, division of the body into serially repeating segments, each with characteristic features. During development, these segments become organized into head, thorax, and abdomen. Each segment of the adult thorax has a different set of appendages. The development of this complex pattern is genetically controlled, and pattern-regulating genes—almost all with close homologs, from nematodes to humans—dramatically affect the organization of the body.



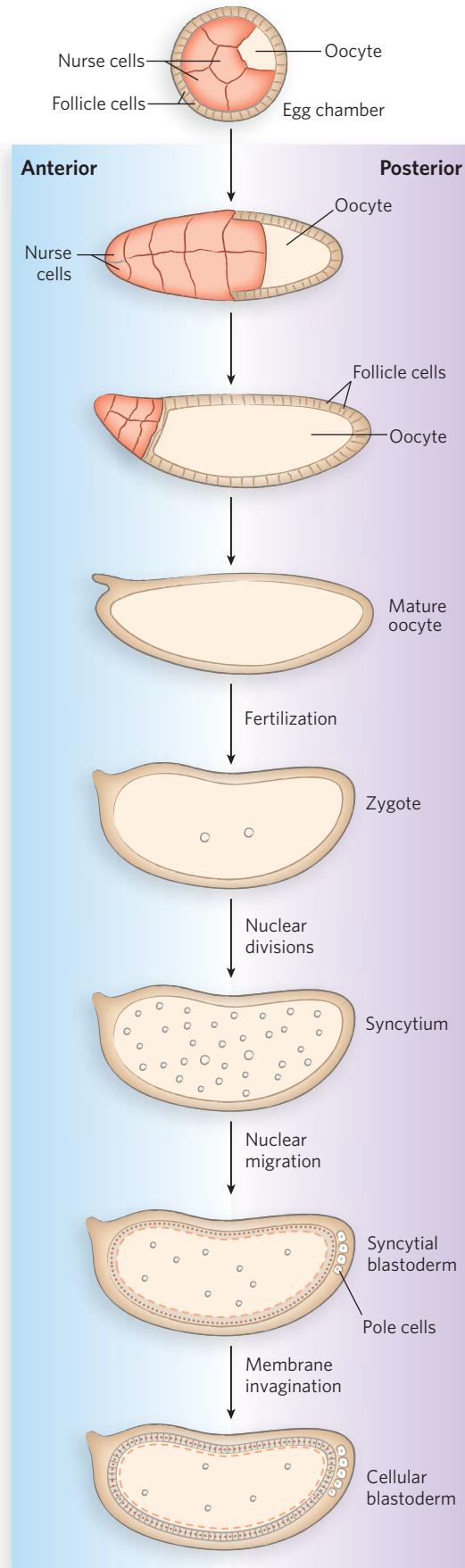
**FIGURE 22-21** The fruit fly life cycle. *Drosophila* undergoes complete metamorphosis. The adult insect is radically different in form from its immature stages, a transformation that requires extensive alterations during development. By the late embryonic stage, segments have formed, each containing specialized structures from which the various appendages

and other features of the adult fly will develop. The segmented late embryo is enlarged compared to the other stages to show detail. [Sources: Top photo from F. R. Turner, Department of Biology, University of Indiana, Bloomington. Other photos from Prof. Dr. Christian Klambt, Westfälische Wilhelms-Universität Münster, Institut für Neuro- und Verhaltensbiologie.]

The *Drosophila* egg, with its 15 nurse cells, is surrounded by a layer of follicle cells (Figure 22-22). As the oocyte matures (before fertilization), mRNAs and proteins originating in the nurse and follicle cells are deposited in the egg cell, where many play a crucial role in development. After the fertilized egg is laid, the nucleus divides and the nuclear descendants continue to divide in synchrony every 6 to 10 minutes. Plasma membranes are not formed around the nuclei, which are distributed within the egg cytoplasm, forming a syncytium. During rounds 8 to 11 of nuclear division, the nuclei migrate to the egg's outer layer, forming a monolayer surrounding the common yolk-rich cytoplasm; this is the syncytial blastoderm. After a few additional divisions (producing up to 6,000 nuclei), membrane infoldings create a layer of cells, forming the cellular blastoderm. At this stage, the mitotic cycles in the cells lose their synchrony. The developmental fate of the cells is determined by the mRNAs and proteins originally deposited in the egg by the nurse and follicle cells.

**Cascades of Regulatory Proteins in Development** The role of key genes in development is to regulate other genes. Temporal and spatial regulation is critical to the gradual maturation of cells and tissues as cell divisions continue from embryo to adult. As each successive layer of regulatory genes is activated, the embryo acquires a finer specialization of cellular function.

Several types of RNAs and proteins in the early embryo, and proteins with essential roles in later stages of development, follow patterns widely conserved in multicellular eukaryotes. As defined by Christiane Nüsslein-Volhard, Edward B. Lewis, and Eric F. Wieschaus for *Drosophila*, three major classes of pattern-regulating genes—maternal, segmentation, and homeotic genes—function in successive developmental stages to specify the basic features of the fruit fly embryo body.



**FIGURE 22-22 Early development in *Drosophila*.** During oocyte development, maternal mRNAs and proteins are deposited in the oocyte by nurse cells and follicle cells. After fertilization, nuclear divisions occur in synchrony in the common cytoplasm (syncytium), and nuclei migrate to the periphery. Membrane invaginations surround the nuclei to create a monolayer of cells at the periphery; this is the cellular blastoderm stage. During the early nuclear divisions, several nuclei at the far posterior of the embryo become pole cells, which later become the germ-line cells.



**Christiane Nüsslein-Volhard**  
(top left) [Source: Courtesy of Christiane Nüsslein-Volhard.]

**Edward B. Lewis, 1918–2004**  
(top right) [Source: Caltech Archives.]

**Eric F. Wieschaus** (bottom left) [Source: Courtesy of Eric Wieschaus.]

**Maternal genes** are expressed in the unfertilized egg, and the resulting **maternal mRNAs** remain dormant until fertilization. Maternal mRNAs provide most of the required proteins in the very early stages of development, and in fruit flies, this occurs until the cellular blastoderm forms. Some of the proteins encoded by maternal mRNAs direct the spatial organization of the developing embryo to establish its polarity. **Segmentation genes**, transcribed after fertilization, direct the formation of the proper number of body segments. In nematodes, similar genes guide the formation of specific tissues following completion of the earliest stages of embryogenesis. At least three subclasses of segmentation genes act at successive stages of *Drosophila* development. **Gap genes** divide the developing embryo into several broad regions, and **pair-rule genes**, along with **segment polarity genes**, define 14 stripes that become the 14 segments of a normal fly embryo. **Homeotic genes**, expressed at a later stage, specify the organs and appendages that will develop in particular body segments.

The many regulatory genes in these three classes direct the development of an adult organism, with a head, thorax, and abdomen, with the proper number of segments, and with the correct appendages on each segment. Although fruit fly embryogenesis takes about a day to complete, all these genes are activated during the first 4 hours. During this period, some mRNAs and proteins are present for only a few minutes at specific points. Some of the genes code for transcription factors that affect the expression of other genes in a kind of developmental cascade. Regulation at the level of translation also occurs, and many of the regulatory genes encode translational repressors, most of which bind to

the 3'UTR of mRNAs. Because many mRNAs are deposited in the egg long before their translation is required, translational repression is especially important for regulation in developmental pathways.

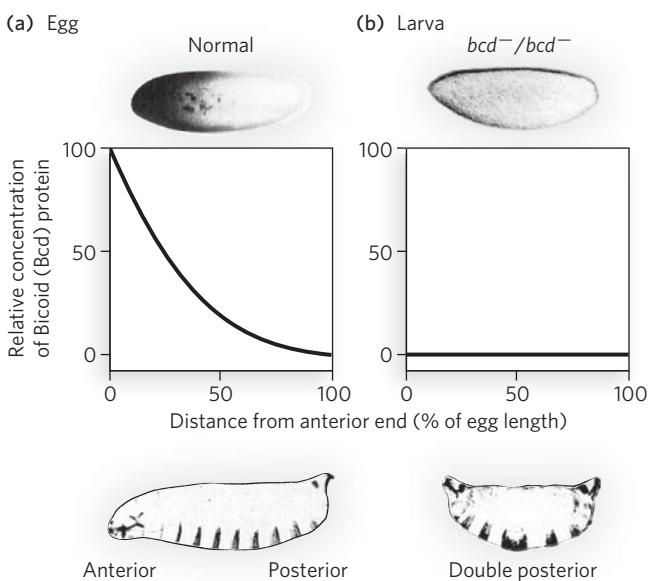
## Early Development Is Mediated by Maternal Genes

In invertebrates, a prescribed developmental path is evident from the very first embryonic cell division. The nonequivalence of the daughter cells of this first division implies a structural and functional asymmetry in the fertilized egg, and this is mediated by established gradients of mRNAs and proteins produced by maternal genes.

In *Drosophila*, some maternal genes are expressed within the nurse and follicle cells, and some in the egg itself. In the unfertilized egg, the maternal gene products establish the critical anterior-posterior and dorsal-ventral axes, thereby defining which regions of the radially symmetric egg will develop into the head and abdomen and the top and bottom of the adult fly. A key event in very early development is establishing mRNA and protein gradients along the body axes. Some maternal mRNAs have protein products that diffuse through the cytoplasm, creating an asymmetric distribution in the egg. Various cells in the cellular blastoderm therefore inherit different amounts of these proteins, setting the cells on different developmental paths. The products of the maternal mRNAs include transcription activators or repressors as well as translational repressors, all regulating the expression of other pattern-regulating genes. Thus, the resulting gene expression sequences and patterns differ among cell lineages, ultimately orchestrating the development of each adult structure.

The anterior-posterior axis in *Drosophila* is also partially defined by the transcription factors produced by the *bicoid* and *nanos* genes. The *bicoid* mRNA is synthesized by nurse cells and deposited in the unfertilized egg near its anterior pole. Nüsslein-Volhard found that this mRNA is translated soon after fertilization, and the Bicoid protein diffuses through the cell to create, by the seventh nuclear division, a concentration gradient radiating out from the anterior pole (Figure 22-23a).

Bicoid contains a homeodomain (see Chapter 19). As transcription activator, Bicoid activates the expression of several segmentation genes. It is also a translational repressor that inactivates certain mRNAs. The amounts of Bicoid in various parts of the embryo affect the subsequent expression of other genes in a threshold-dependent way. Genes are transcriptionally activated or translationally repressed only where the concentration of Bicoid exceeds the threshold. Bicoid



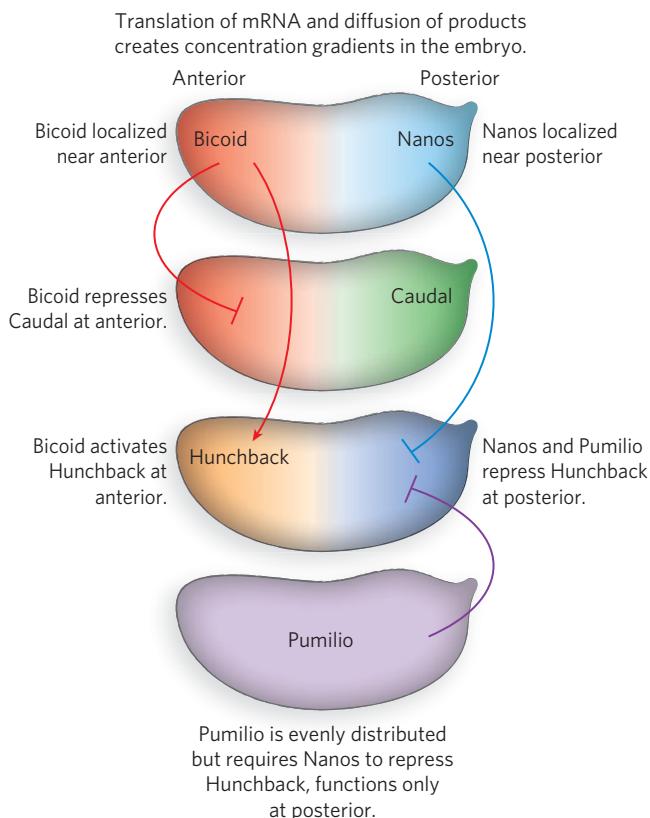
**FIGURE 22-23** Distribution of a maternal gene product in a *Drosophila* egg. (a) The micrograph of an immunologically stained egg shows the distribution of the *bicoid* (*bcd*) gene product, the Bicoid protein. The graph shows stain intensity (protein concentration) along the length of the egg. This distribution is essential for normal development of the anterior structures. (b) If the *bcd* gene is not expressed by the mother (a *bcd*<sup>-/-</sup>/*bcd*<sup>-/-</sup> mutant), no *bicoid* mRNA is deposited in the egg, resulting in lack of Bicoid protein, as seen in the micrograph. The resulting embryo has two posteriors (and soon dies). [Source: Wolfgang Driever and Christiane Nüsslein-Volhard, Max-Planck-Institut.]

plays a critical role in anterior development. The absence of Bicoid results in development of an embryo with two abdomens but no head and no thorax (Figure 22-23b). Embryos without Bicoid can develop normally if an adequate amount of *bicoid* mRNA is injected into the egg at the appropriate end.

The *nanos* gene has an analogous role, but its mRNA is deposited at the posterior end of the egg, and the anterior-posterior Nanos protein gradient peaks at the posterior pole. Nanos is a translational repressor, conserved from worms to humans.

A broader view of the effects of maternal genes in *Drosophila* reveals a more precise picture of a developmental circuit. In addition to *bicoid* and *nanos* mRNAs, deposited in the egg asymmetrically, several other maternal mRNAs are deposited uniformly throughout the egg cytoplasm. Three of them encode the Pumilio, Hunchback, and Caudal proteins—all affected by Nanos and Bicoid (Figure 22-24).

Caudal and Pumilio are involved in the development of the fruit fly's posterior end. Caudal is a



**FIGURE 22-24** Regulatory circuits of the anterior-posterior axis in a *Drosophila* egg. The *bicoid* and *nanos* mRNAs are localized near the anterior and posterior poles of the egg, respectively. The *caudal*, *hunchback*, and *pumilio* mRNAs are distributed throughout the cytoplasm. Gradients of Bicoid and Nanos proteins lead to accumulation of Hunchback protein in the egg's anterior region and Caudal protein in its posterior. Because Pumilio requires Nanos for its activity as a translational repressor of *hunchback* mRNA, Pumilio functions only at the posterior end.

transcription activator with a homeodomain; Pumilio is a translational repressor from the PUF family of proteins. Hunchback plays an important part in developing the anterior end; it is also a transcription factor for several genes, in some cases an activator and in others a repressor. Bicoid suppresses the translation of *caudal* mRNA at the anterior end and also acts as a transcription activator of *hunchback* in the cellular blastoderm. Because *hunchback* is expressed through maternal mRNAs and from genes in the developing egg, it is considered a maternal as well as a segmentation gene. The result of Bicoid's activities is an increased concentration of Hunchback at the anterior end of the egg. Nanos and Pumilio act as translational repressors of *hunchback*, suppressing synthesis of Hunchback near the posterior end of the egg. Pumilio does not function in the

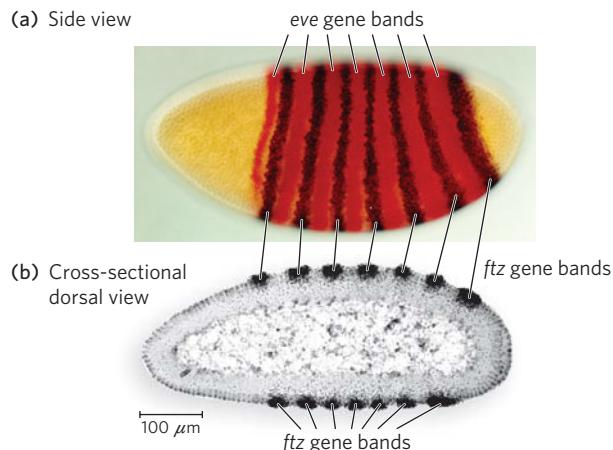
absence of Nanos, and the gradient of *nanos* expression confines the activity of both proteins to the posterior region. Translational repression of the *hunchback* gene leads to degradation of *hunchback* mRNA near the posterior end. However, a lack of Bicoid in the posterior leads to expression of *caudal*. In this way, the Hunchback and Caudal proteins become asymmetrically distributed in the egg.

### Segmentation Genes Specify the Development of Body Segments and Tissues

Segmentation genes are the zygotic genes that take over after maternal genes. Many operate at the level of transcriptional regulation. Gap genes, pair-rule genes, and segment polarity genes are activated in a cascadelike sequence at successive stages of embryonic development. The expression of gap genes is generally regulated by the products of one or more maternal genes. Gap genes activate the pair-rule genes, which in turn activate the segment polarity genes. This cascade of gene expression is accompanied by the gradual formation of 14 parasegments, then the true segments. Only a few cells (or nuclei) wide, parasegments are delimited by temporary grooves. Segments are offset from parasegments, so that each segment later encompasses the anterior part of a parasegment and the posterior part of the adjacent one. The anterior segments eventually fuse to form the head.

Pair-rule genes are expressed in alternating parasegments, and one well-characterized segmentation gene in the pair-rule subclass is *fushi tarazu* (*ftz*). When *ftz* is deleted, the embryo develops 7 segments instead of the normal 14, each segment twice the usual width. The Fushi-tarazu protein (Ftz) is a transcription activator with a homeodomain. The mRNAs and proteins derived from the *ftz* gene accumulate in a striking pattern of seven stripes that encircle the posterior two-thirds of the embryo (Figure 22-25). The stripes demarcate half of the parasegments; the development of alternating segments is compromised if *ftz* function is lost. The Ftz protein and a few similar regulatory proteins directly or indirectly regulate the expression of vast numbers of genes in the continuing developmental cascade.

In the stripes where *ftz* is repressed, repression is mediated in part by another pair-rule gene called *even-skipped* (*eve*) (see Figure 21-15). The expression of *eve* is activated by Bicoid and Hunchback, and is highest in the parasegments where *ftz* gene expression is low. This gives rise to the alternating pattern of *ftz* and *eve* expression. The expression of *eve* is repressed by two other gap genes, *krippel* and *giant*. The alternating pattern of *eve* and *ftz*

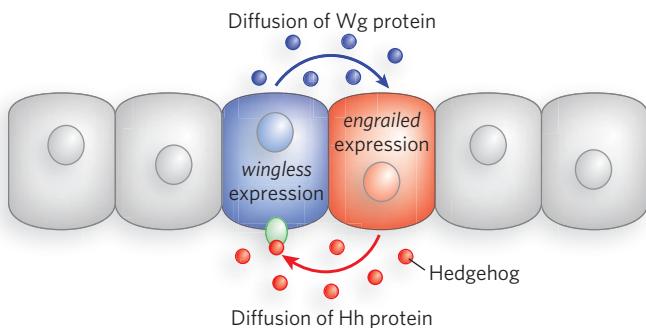


**FIGURE 22-25 Distribution of the *fushi tarazu* (*ftz*) and *even-skipped* (*eve*) gene products in early *Drosophila* embryos.** (a) The *ftz* gene product can be detected in seven bands around the circumference of the embryo (brown). These alternate with bands where the *eve* gene is expressed (reddish orange). (b) In a cross-sectional autoradiograph, the *ftz* bands appear as dark spots (generated by a radioactive label) and demarcate the anterior margins of the segments that will appear in the late embryo. [Sources: (a) Courtesy of Stephen J. Small, Department of Biology, New York University. (b) Courtesy of Phillip Ingham, Imperial Cancer Research Fund, Oxford University.]

expression is due to variations in gap gene expression from one parasegment to the next.

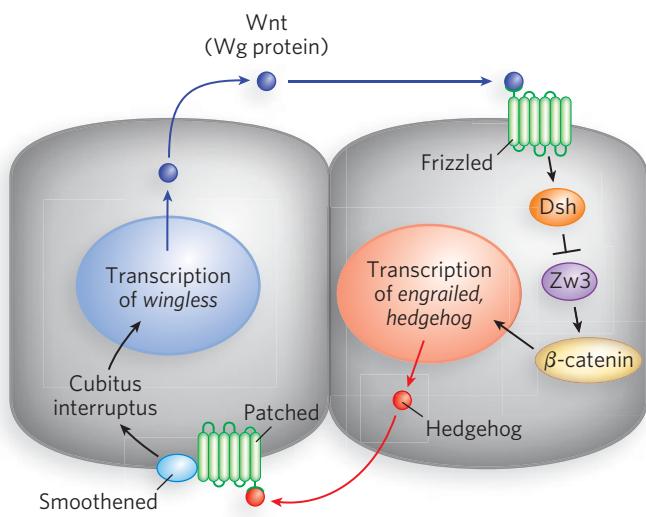
Gap and pair-rule genes operate during the first 2.5 hours of *Drosophila* development, when the embryo is still a syncytium. Virtually all these genes encode transcription factors, which have localized access to the nuclei in the syncytium. Segment polarity genes, the last group in the regulatory cascade, act at a stage when cells have formed. Some of the focus now shifts from transcription factors to cell-cell signaling pathways. The signaling helps reinforce the alternating boundaries between parasegments, then segments. Further, the function of adjacent segments is made interdependent by the pattern of segment polarity gene expression. A key example can be seen in the genes *wingless* (*wg*), *engrailed* (*en*), and *hedgehog* (*hh*). Wingless (Wg) and Engrailed (En) proteins are initially expressed in alternating cells, due to activation by pair-rule genes. However, the relevant pair-rule gene products recede after a few hours, and the continued expression of Wg and En becomes interdependent across opposite sides of parasegment boundaries (Figure 22-26).

The Wg protein is a signal of the Wnt class, and studies of the *wg* gene helped define the **Wnt-class signaling pathways**, which play key roles in development in eukaryotes, from nematodes to humans. (The name



**FIGURE 22-26 Interdependent signaling loops across segment boundaries in *Drosophila*.** Part of the maintenance of parasegment, and later segment, boundaries involves cell-cell signaling. Wnt-class signaling with the Wg protein, the wingless (*wg*) gene product, induces expression of the *engrailed* (*en*) gene in recipient cells. The En protein triggers expression of *hedgehog*, part of a non-Wnt-class signaling pathway. The Hedgehog protein (Hh) promotes expression of Wg in the original cells and completes the closed signaling loop.

“Wnt,” wingless type, originated in the study of a mouse gene homologous to the *Drosophila* gene *wingless*.) Wnt-class pathways generally consist of a secreted Wnt glycoprotein that constitutes the signal, one or more proteins involved in the secretion process, and a receptor protein in the membrane of the target cell (Figure 22-27). Additional proteins act as regulators.



**FIGURE 22-27 A Wnt-class signaling pathway in *Drosophila* development.** The Wnt signal is secreted from one cell and interacts with a receptor on another cell. The signal in this case results in gene activation, via a pathway that uses a receptor (Frizzled) and signaling proteins Dsh, Zw3, and  $\beta$ -catenin. The secreted Hedgehog protein also works through a signaling pathway, involving a different receptor and different set of signaling proteins, as shown.

Wnt proteins are highly homologous from one species to the next. They generally have a nearly invariant pattern of 23 Cys residues (some of which may form disulfide bonds needed for folding), an N-terminal signal sequence that helps guide secretion, and several N-glycosylation sites. They also have at least two lipid modifications: a palmitoyl group is added at a conserved Cys residue, and palmitoleyl group at a conserved Ser. Wnt proteins are synthesized, and the lipid modifications are made, in the endoplasmic reticulum. The proteins move through a normal secretion pathway, from ER to Golgi complex, and are then transported to the cell surface in vesicles. Some of the secreted Wnt proteins are associated with lipoprotein particles.

The Wg protein is modified with lipids by an acyltransferase in the endoplasmic reticulum that is encoded by the gene *porcupine*. In nearby cells on the opposite side of the adjacent parasegment boundary, the secreted Wg protein interacts with a receptor that is the product of the gene *frizzled* (*fz*). In the recipient cell, the interaction triggers a signaling pathway that ultimately results in expression of the En protein. En is a transcription activator of the *hedgehog* gene. The Hedgehog (Hh) protein, part of a non-Wnt-class signaling pathway, is secreted and interacts with receptors on the Wg-producing cells. The resulting signal activates more Wg protein synthesis. The entire cycle is self-sustaining and self-reinforcing. Most of the other known segment polarity genes encode proteins that are part of either the *wingless* (Wnt) or the *hedgehog* (Hh) signaling pathway. In the alternating segments, the En protein and other transcription factors activate or repress a series of additional genes that now begin to give each segment a distinctive function. Many of these targets are homeotic genes.

## Homeotic Genes Control the Development of Organs and Appendages

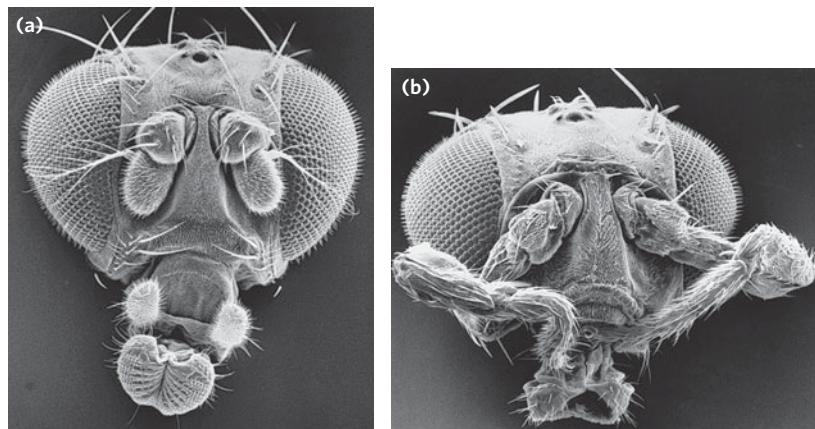
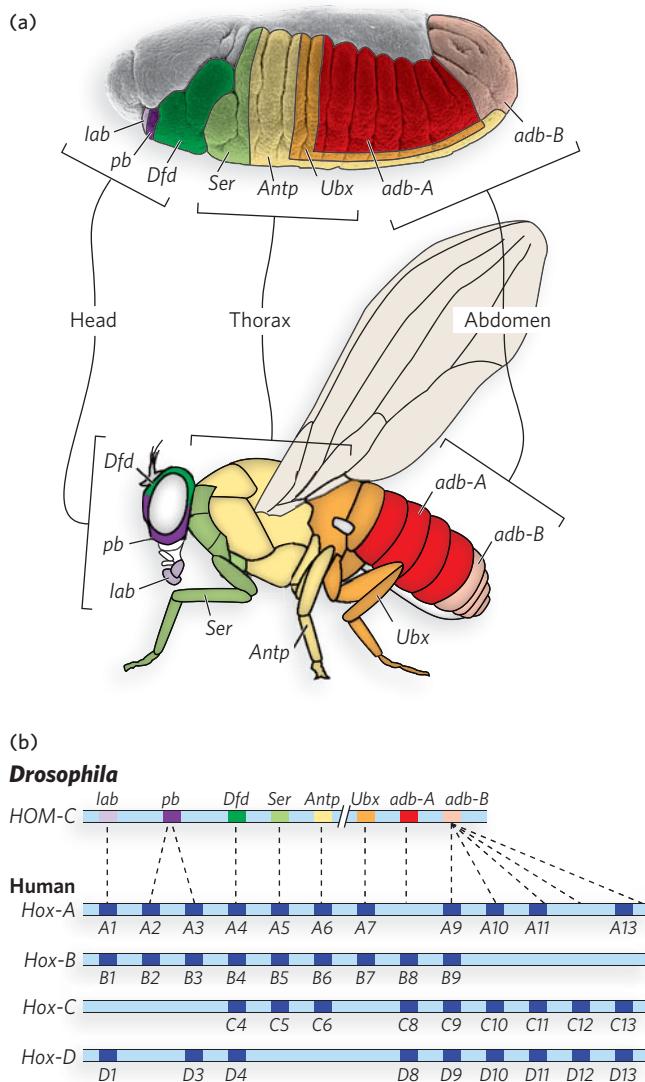
A set of 8 to 11 homeotic genes directs the formation of structures at specific locations in the body plan of most multicellular eukaryotes. Fewer homeotic genes are present in some simple eukaryotes. Even yeast has two homologs, regulators of mating-type switching (see Chapter 21). These genes are now more commonly referred to as **Hox genes** (from *homeobox*, the conserved gene sequence that encodes the homeodomain). However, these are not the only development-related proteins to include a homeodomain (e.g., as noted above, Bicoid has a homeodomain), and Hox is more a functional than a structural classification.

Hox genes are sometimes organized in genomic clusters. *Drosophila* has one such cluster, and mammals have four. The order of genes within the clusters is colinear with their targets of action, from the anterior to the posterior of the developing embryo. In *Drosophila*, each

Hox gene is expressed in a particular embryonic segment and controls the development of the corresponding part of the mature fly (Figure 22-28a). The terminology for describing Hox genes can be confusing. They have historical names in the fruit fly (e.g., *ultrabithorax*), whereas in mammals they are designated by two competing systems based on lettered (*A, B, C, D*) or numbered (1, 2, 3, 4) clusters (Figure 22-28b).

The loss of Hox genes in fruit flies, by mutation or deletion, causes the appearance of a normal appendage or body structure at an inappropriate body position. An important example is the *ultrabithorax* (*Ubx*) gene. When the *Ubx* protein function is lost, the first abdominal segment develops incorrectly, having the structure of the third thoracic segment. Other known homeotic mutations cause the formation of an extra set of wings, or two legs at the position in the head where the antennae are normally found (Figure 22-29). The Hox genes

**FIGURE 22-28 The Hox gene clusters and their effects on development.** (a) Each Hox gene in the fruit fly directs the development of structures in a defined part of the body and is expressed in defined regions of the embryo (coded by color). (b) *Drosophila* has one Hox gene cluster (*HOM-C*). The genes are color-coded to match the fly segments in (a). The human genome has four Hox gene clusters (*Hox-A* through *Hox-D*). Many Hox genes are highly conserved in animals. Evolutionary relationships between genes in the *Drosophila* Hox gene cluster and those in the mammalian Hox gene clusters are indicated by dashed lines. Similar relationships between the four sets of mammalian Hox genes are indicated by vertical alignment. [Source: (a) Photo from F. R. Turner, Department of Biology, University of Indiana, Bloomington.]



**FIGURE 22-29 Effects of Hox gene mutations in *Drosophila*.** (a) Normal head structure. (b) Homeotic mutant (*antennapedia*) in which antennae are replaced by legs. (c) Normal body structure. (d) Homeotic mutant (*bithorax*)

in which a segment has developed incorrectly to produce an extra set of wings. [Sources: (a), (b) F. R. Turner, Department of Biology, University of Indiana, Bloomington. (c), (d) E. B. Lewis, Division of Biology, California Institute of Technology.]

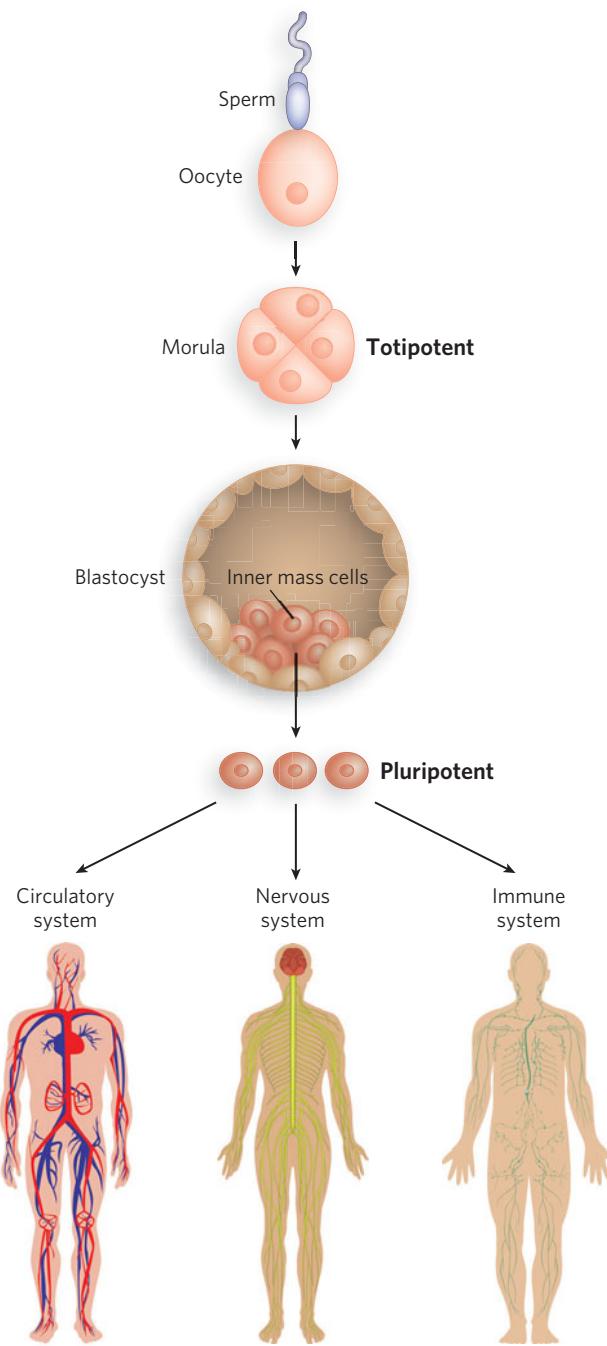
often span long regions of DNA. The *ubx* gene, for example, is 77,000 bp long. More than 73,000 bp are in introns, one of which is more than 50,000 bp long. Transcription of the *ubx* gene takes nearly an hour. The delay this imposes on *ubx* gene expression is believed to be a timing mechanism involved in the temporal regulation of subsequent steps in development. Many Hox genes are further regulated by miRNAs encoded by intergenic regions of the Hox gene clusters. All Hox gene products are themselves transcription factors that regulate the expression of an array of downstream genes.

The conservation of some Hox genes is extraordinary. For example, the products of the homeobox-containing *Hoxa-7* gene in mice and *antennapedia* gene in fruit flies differ in only one amino acid residue. Of course, although the molecular regulatory mechanisms may be similar, many of the ultimate developmental events are not conserved (humans do not have wings or antennae). The different outcomes are brought about by variance in the downstream target genes controlled by the Hox genes (see How We Know). The discovery of structural determinants with identifiable molecular functions is the first step in understanding the molecular events underlying development. As more genes and their protein products are discovered, the biochemical side of this vast puzzle will be elucidated in increasingly rich detail.

### Stem Cells Have Developmental Potential That Can Be Controlled

If we can understand development, and the mechanisms of gene regulation behind it, we can control it. An adult human has many different types of tissues. Many of the cells are terminally differentiated and no longer divide. If an organ malfunctions due to disease, or a limb is lost in an accident, the tissues are not readily replaced. Most cells, because of the regulatory processes that are in place, or even the loss of some or all genomic DNA, are not easily reprogrammed. Medical science has made organ transplants possible, but organ donors are a limited resource and organ rejection remains a major medical problem. If humans could regenerate their own organs or limbs or nervous tissue, rejection would no longer be an issue. Real cures for kidney failure or neurodegenerative disorders could become reality.

The key to tissue regeneration lies in **stem cells**—cells that have retained the capacity to differentiate into various tissues. In humans, after an egg is fertilized, the first few cell divisions create a ball of **totipotent** cells (the morula), cells that have the capacity to differentiate individually into any tissue or even into a



**FIGURE 22-30** **Totipotent and pluripotent stem cells.** Cells of the morula stage are totipotent and have the capacity to differentiate into a complete organism. The source of pluripotent embryonic stem cells is the inner mass cells of the blastocyst. Pluripotent cells give rise to many tissue types, but cannot form complete organisms. [Source: Photos from Dannyphoto/dreamstime.com.]

complete organism (Figure 22-30). Continued cell division produces a hollow ball, a blastocyst. The outer cells of the blastocyst eventually form the placenta. The inner layers form the germ layers of the developing fetus—the ectoderm, mesoderm, and endoderm. These

cells are **pluripotent**: they can give rise to cells of all three germ layers and can be differentiated into many types of tissues. However, they cannot differentiate into a complete organism. Some of these cells are **unipotent**: they can develop into only one type of cell and/or tissue. It is the pluripotent cells of the blastocyst, the **embryonic stem cells**, that are currently used in embryonic stem cell research.

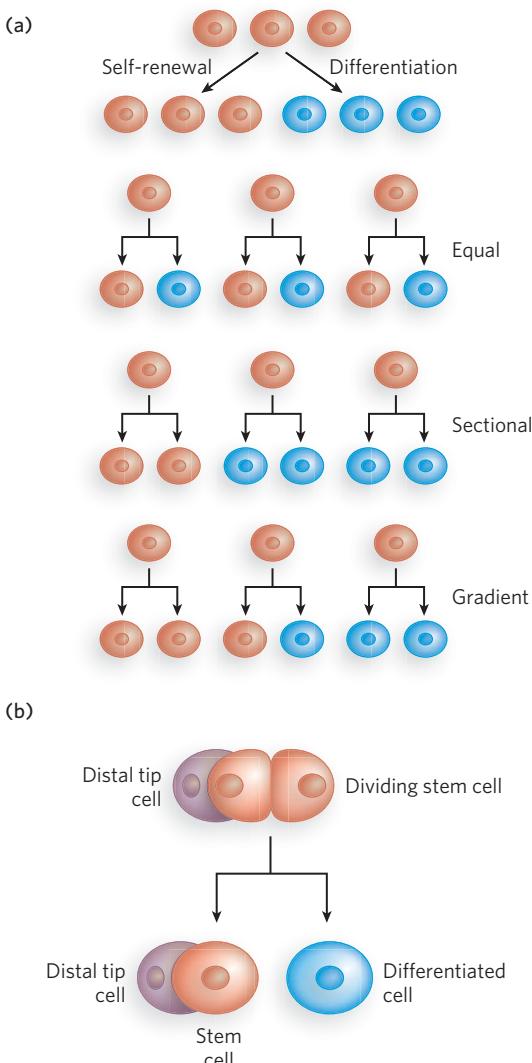
Stem cells have two functions: to replenish themselves and, at the same time, provide cells that can differentiate. These tasks are accomplished in multiple ways (**Figure 22-31a**). All or parts of the stem cell population can, in principle, be involved in replenishment, differentiation, or both.

Other types of stem cells can potentially be used for medical benefit. In the adult organism, **adult stem cells**, as products of additional differentiation, have a more limited potential for further development than

do embryonic stem cells. For example, the hematopoietic stem cells of bone marrow can give rise to many types of blood cells, and also to cells with the capacity to regenerate bone. They are referred to as **multipotent**. However, these cells cannot differentiate into a liver or kidney or neuron. Adult stem cells are often said to have a **niche**, a microenvironment that promotes stem cell maintenance while allowing differentiation of some daughter cells as replacements for cells in the tissue they serve (**Figure 22-31b**). Hematopoietic stem cells in the bone marrow occupy a niche in which signaling from neighboring cells and other cues maintain the stem cell lineage. At the same time, some daughter cells differentiate to provide needed blood cells. Understanding the niche in which stem cells operate, and the signals the niche provides, is essential in efforts to harness the potential of stem cells for tissue regeneration.

All stem cells have problems with respect to human medical applications. Adult stem cells have a limited capacity to regenerate tissues, are generally present in small numbers, and are hard to isolate from an adult human. Embryonic stem cells have much greater differentiation potential and can be cultured to generate large numbers of cells. However, their use is accompanied by ethical concerns related to the necessary destruction of human embryos. Identifying a source of plentiful and medically useful stem cells that does not raise concerns remains a major goal of medical research.

Our ability to culture stem cells (i.e., maintain them in an undifferentiated state), and to manipulate them to grow and differentiate into particular tissues, is very much a function of our understanding of developmental biology. The identification and culturing of pluripotent stem cells from human blastocysts was reported by



**FIGURE 22-31** Stem cell proliferation versus differentiation and development. Stem cells must strike a balance between self-renewal and differentiation. (a) Some possible cell division patterns that allow for replenishment of stem cells and production of some differentiated cells. Each cell may produce one stem cell and one differentiated cell, or two differentiated cells or two stem cells in defined parts of the tissue or culture. Or a gradient of growth conditions can be established, with cell fates differing from one end of the gradient to the other. (b) Establishing a developmental niche through stem cell contact with a cell or group of cells. Molecular signals provided by niche cells (in this case, for plants, a distal tip cell) help orient the mitotic spindle for stem cell division and ensure that one daughter cell retains stem cell properties.



**James Thomson** [Source: © Morgridge Institute for Research. Used with permission.]

James Thomson and colleagues in 1998. This advance led to the long-term availability of established cell lines for research.

Thus far, mouse and human embryonic stem cells have been used for most research. Although both types of stem cells are pluripotent, they require very different culture conditions, optimized to allow cell division indefinitely without differentiation. Mouse embryonic stem cells are grown on a layer of gelatin and require the presence of leukemia inhibitory factor (LIF). Human embryonic stem cells are grown on a feeder layer of mouse embryonic fibroblasts and require basic fibroblast growth factor (bFGF) or FGF-2. The use of a feeder cell layer implies that the mouse cells are providing a diffusible product or some surface signal, not yet known, that is needed by human stem cells to either promote cell division or prevent differentiation. Recent research suggests that at least one of these diffusible products may be a Wnt-class protein.

This research is in its very early stages. However, some success has been achieved in directing the differentiation of human embryonic stem cells into particular tissue types. Some of the progress in stimulating stem cell differentiation is summarized in Table 22-1.

A significant advance, reported in 2007, centers on success in reversing differentiation. In effect, skin cells—first from mice, then from humans—have been reprogrammed to take on the characteristics of pluripotent

stem cells. The reprogramming involves manipulations to get the cells to express at least four transcription factors (Oct4, Sox2, Nanog, and Lin28), all of which are known to help maintain the stem cell-like state. Gradual improvements in this technology may make the harvesting of embryonic stem cells unnecessary and provide a source of stem cells that is genetically matched to a prospective patient.

## SECTION 22.5 SUMMARY

- Development of a multicellular organism presents the most complex regulatory challenge.
- The fate of cells in the early embryo is determined in part by establishment of anterior-posterior and dorsal-ventral gradients of proteins that act as transcription activators or translational repressors, regulating the genes required for the development of structures appropriate to a particular part of the organism.
- Sets of regulatory genes operate in temporal and spatial succession, transforming given areas of an egg cell into predictable structures in the adult organism.
- The developmental fate of cell lineages during development is also shaped by cell-cell signaling pathways in which signals from one cell lineage affect the fate of others. The Wnt-class signaling pathway is one well-studied example.
- In vertebrates, stem cells retain significant developmental potential. The differentiation of stem cells into functional tissues can be controlled by extracellular signals and conditions.

**Table 22-1 Requirements for Differentiating Human Embryonic Stem Cells into Various Tissue Types**

Culture Additives/Protocol	Differentiated Tissue
Detach from mouse cell feeder layer	Neural cell types (neurons, oligodendrocytes, and glia)
Grow into neural-tube-like structures	
Grow in presence of FGF-2	
Isolate neural precursors by treatment with dispase (a protease)	
Withdraw FGF-2; differentiation into neurons	
Co-culture with another cell line, derived from human embryonic stem cells treated with retinoic acid	Cardiomyocytes
Remove mouse cell feeder layer	
Add ascorbate, $\beta$ -glycerophosphate, and dexamethasone	Some insulin-producing pancreatic beta cells
Co-culture with bone marrow cell line	Osteoblasts
Then, grow in presence of cytokines	Hematopoietic progenitors

## 22.6 Finale: Molecular Biology, Developmental Biology, and Evolution

Molecular biology is a story of biological information—the metabolism, maintenance, and transfer of that information from one generation to the next. As we go back almost unfathomable lengths of time, those generations link us to every living thing on our planet. Our genomic DNA makes our life possible. With its seemingly ragtag mix of piggy-backing transposons, integrated viruses, and genes, both borrowed and linearly evolved, our genome tells us about our past while at the same time linking us to a future rich with potential. Evolution continues.

Each topic in molecular biology has evolutionary significance. Errors or random events in DNA replication, recombination, and repair fuel genomic changes—some useful, many deleterious. Genomic changes are expressed, through transcription and translation, into the organismal phenotypes on which natural selection acts. However, there are few areas where molecular biology meets evolution more dramatically than in the regulation of organismal development. Developmental biology thus provides a fitting final topic for our exploration of molecular biology.

### The Interface of Evolutionary Biology and Developmental Biology Defines a New Field

South America has several species of seed-eating finches, commonly known as grassquits. About 3 million years ago, a small group of grassquits, of a single species, took flight from the continent's Pacific coast. Perhaps driven by a storm, they lost sight of land and traveled nearly 1,000 km. Small birds such as these might easily have perished on such a journey, but the smallest of chances brought this group to a newly formed volcanic island in an archipelago later to be known as the Galápagos. It was a virgin landscape with untapped plant and insect food sources, and the newly arrived finches survived. Over many millennia, new islands formed and were colonized by new plants and insects—and by the finches. The birds exploited the new resources on the islands, and groups of birds gradually specialized and diverged into new species. By the time Charles Darwin stepped onto the islands in 1835, there were many different finch species on the various islands of the archipelago, feeding on seeds, fruits, insects, pollen, and even blood. Some islands now have as many as 10 finch species, each adapted to a somewhat different lifestyle and food source.

The diversity of living creatures on our planet was a source of wonder for humans long before scientists sought to understand its origins. The extraordinary insight handed down to us by Darwin, inspired in part by his encounter with the Galápagos finches, provided a broad explanation for the existence of organisms with a vast array of appearances and characteristics. It also gave rise to many questions about the mechanisms underlying evolution. Answers to those questions have started to appear, first through the study of genomes and nucleic acid metabolism in the last half of the twentieth century, and more recently through an emerging field nicknamed **evo-devo**—a blend of evolutionary and developmental biology.

In its modern synthesis, the theory of evolution has two main elements: mutations in a population generate genetic diversity; natural selection then acts on this diversity to favor individuals with more useful genomic tools, and to disfavor others. Mutations occur at significant rates in every individual's genome, in every cell (see Chapter 3 and Chapter 12). Advantageous mutations in single-celled organisms or in the germ line of multicellular organisms can be inherited, and they are more likely to be inherited (i.e., passed on to greater numbers of offspring) if they confer an advantage. It is a straightforward scheme. But many have wondered whether that scheme is enough to explain, say, the many different beak shapes in the Galápagos finches, or the diversity of size and shape among mammals. Until recent decades, there were several widely held assumptions about the evolutionary process: that many mutations and new genes would be needed to bring about a new physical structure, that more-complex organisms would have larger genomes, and that very different species would have few genes in common and perhaps use very different patterns of gene regulation. All of these assumptions were wrong.

Modern genomics has revealed that the human genome contains fewer genes than expected—not many more than the fruit fly genome, and fewer than some amphibian genomes. The genomes of every mammal, from mouse to human, are surprisingly similar in the number, types, and chromosomal arrangement of genes. Meanwhile, evo-devo is telling us how complex and very different creatures can evolve within these genomic realities.

### Small Genetic Differences Can Produce Dramatic Phenotypic Changes

The kinds of mutant organisms shown in Figure 22-29 were studied by the English biologist William Bateson in the late nineteenth century. Bateson used his

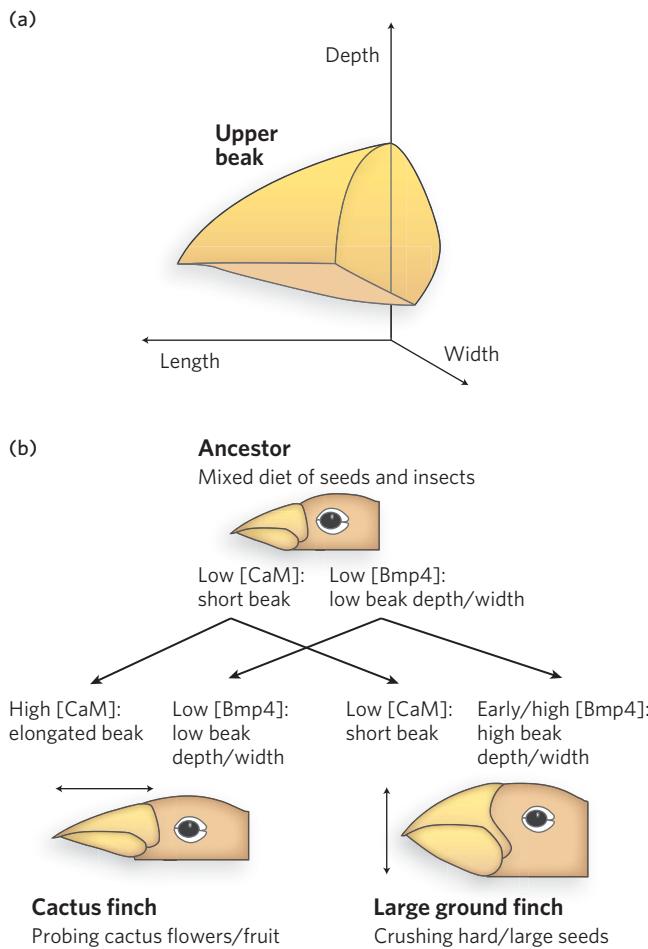
observations to challenge the Darwinian notion that evolutionary change would have to be gradual. Recent studies of the genes that control organismal development have put an exclamation point on Bateson's ideas. Subtle changes in regulatory patterns during development, reflecting just one or a few mutations, can result in startling physical changes and fuel surprisingly rapid evolution.

The Galápagos finches provide a wonderful example of the link between evolution and development. There are at least 14 species (some specialists list 15), and they are distinguished in large measure by their beak structure. The ground finches, for example, have broad, heavy beaks adapted to crushing hard, large seeds. The cactus finches have longer, slender beaks, ideal for probing cactus flowers and fruit (Figure 22-32).

Clifford Tabin and his colleagues carefully surveyed a set of genes expressed during avian craniofacial development. They identified a single gene, *Bmp4*, whose expression level correlated with the formation of the more robust beaks of the ground finches. More robust beaks were also formed in chicken embryos when high

levels of the *Bmp4* protein were artificially expressed in the appropriate tissues, confirming the importance of *Bmp4*. In a similar study, the formation of long, slender beaks was linked to the expression of the protein calmodulin in particular tissues at appropriate developmental stages. Thus, major changes in the shape and function of the beak can be brought about by subtle changes in the expression of just two genes involved in developmental regulation. Very few mutations are required, and the needed mutations affect regulation. New genes are *not* required. Note that *Bmp4* is a member of a family of signaling proteins, with roles in development similar to the Wnt and Hedgehog proteins. Like Wnt and Hedgehog, *Bmp* homologs are widely conserved in eukaryotes. And as in the other signaling pathways, alterations in *Bmp* signaling pathways can have large effects on development.

The system of regulatory genes that guides development is remarkably conserved among all vertebrates. Elevated expression of *Bmp4* in the right tissue at the right time leads to more robust jaw parts in zebra fish. The same gene plays a key role in tooth development in mammals. The development of eyes is triggered by the expression of a single gene, *Pax6*, in fruit flies and in mammals. The mouse *Pax6* gene will trigger the development of fruit fly eyes in the fruit fly, and the fruit fly *Pax6* gene will trigger the development of mouse eyes in the mouse. In each organism, these genes are part of the much larger regulatory cascade that ultimately creates the correct structures in the correct locations in each organism. The cascade is ancient; the Hox genes have been part of the developmental program of multicellular eukaryotes for more than 500 million years. Subtle changes in the cascade can have large effects on development, and thus on the ultimate appearance of the organism. These same subtle changes can promote rapid evolution. For example, the 400 to 500 described species of cichlids (spiny-finned fish) in Lake Malawi



**FIGURE 22-32** The evolution of new beak structures to exploit new food sources. Galápagos finches that feed on different, specialized food sources have different beak structures, as shown for the cactus finch and the large ground finch. (a) The beak structures can be varied along three dimensions. (b) The differences observed in the two finch species were produced largely through natural selection acting on a few mutations that altered the timing and level of expression of just two genes: those encoding *Bmp4* and calmodulin (CaM).

and Lake Victoria on the African continent are all derived from one or a few populations that colonized each lake over the past 100,000 to 200,000 years. The Galápagos finches simply followed a path of evolution and change that living creatures have been traveling for billions of years.

Our discussion of developmental regulation brings us full circle, back to a biochemical beginning—both figuratively and literally. Evolution appropriately provides a key backdrop for the first and last chapters in this book. If evolution is to generate the kind of changes in an organism that we associate with a different species, it is the developmental program that must be affected. Developmental and evolutionary processes are closely allied, each informing the other. Molecular biology ties the fields together, informs us about molecular mechanisms that underpin the changes, and provides the technology needed for new discovery. The continuing study of molecular biology has everything to do with enriching the future of humanity and understanding our origins.

## SECTION 22.6 SUMMARY

- Developmental biology and evolutionary biology are closely related. The two fields inform each other, and molecular biology is intimately intertwined with both.
- Major changes in the appearance and/or function of multicellular eukaryotes can be effected by subtle changes in an organism's developmental program, involving mutations in the regulatory genes that guide the process.

## Unanswered Questions

Our understanding of gene regulation remains incomplete in many areas. Indeed, the recent discovery of RNA interference, which has proved to be a major

mode of regulation, underscores the likelihood that more fundamentals remain to be elucidated.

- 1. How is alternative splicing regulated?** Mounting evidence suggests that alternative splicing accounts for a much greater degree of protein complexity in higher eukaryotes than would be predicted by simply counting the number of open reading frames in a genome. How such alternative splicing is regulated is not well understood, nor have mechanisms of tissue-specific splicing been worked out.
- 2. How and when do miRNAs control gene expression in human cells?** The human genome encodes several hundred miRNAs, yet the targets and functions of most of these are currently unknown. How do we harness these newly discovered regulatory mechanisms to provide new therapies for cancer and other diseases?
- 3. What other regulatory mechanisms have we just not uncovered yet?** RNA interference is a fairly recent discovery. Are there things embedded in the “noncoding” DNA between open reading frames that perhaps code for RNAs with functions not yet described?
- 4. Are transcriptional and posttranscriptional steps in gene expression coordinately regulated?** What kinds of mechanisms might enable communication between the cytoplasm and the nucleus to adjust transcription, splicing, and mRNA transport rates in response to increased or decreased translation of a particular mRNA?
- 5. What are all of the signals that guide the action of regulatory proteins and the development of specific tissues?** A much more detailed understanding is needed to completely unleash the potential of stem cell technologies. That potential includes new cancer treatments, the regeneration of lost limbs, and the replacement of diseased tissues (e.g., heart, lung, kidney) without the danger of tissue rejection.

# How We Know

## A Natural Collaboration Reveals a Binding Protein for a 3'UTR

Zhang, B., M. Gallegos, A. Puoti, E. Durkin, S. Fields, J. Kimble, and M.P. Wickens. 1997. A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390:477-484.



**Marvin Wickens** [Source: Courtesy of Marvin Wickens.]

Scientific collaborations come about in many ways, as the discovery of one gene-regulatory protein attests. The importance of the untranslated parts of an mRNA, particularly the 3'UTR, gradually became apparent over the course of the 1990s. In 1991, Judith Kimble's lab reported the discovery of a 3'UTR regulatory element in the *fem-3* gene of the nematode *C. elegans*, a sequence called PME (point mutation element) (see Moment of Discovery). Single base-pair changes in this element had a dramatic effect on germ cell fate, specifically in the switch from sperm to eggs during germ-line development in the hermaphrodite. The PME sequence had to be interacting with something, but what? An RNA-binding protein seemed a likely candidate, but what protein? How could it be identified? For Kimble, the obvious approach was unusually close at hand.

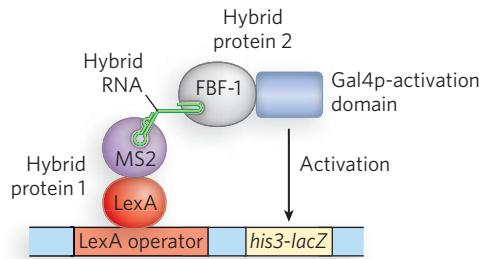
In 1996, Marvin Wickens and coworkers had reported the invention of the three-hybrid method to identify proteins that bound to particular RNA sequences (see Figure 7-26). The problem posed by the Kimble group was a perfect test of this new technology. Happily, not only was the Wickens lab quite near the Kimble lab at the University of Wisconsin, but Wickens and Kimble were husband and wife. A new kind of collaboration (for them) was soon hatched.

The investigators initiated a screen of a cDNA library in which *C. elegans* genes were fused to the gene encoding the Gal4p activation domain. The new three-hybrid method worked as advertised; one clone was found that met all criteria for the study and activated expression of the reporter gene, a *his3-lacZ* fusion. Testing a wide array of RNA binding substrates, they demonstrated that the protein encoded by the cloned nematode gene bound only to the target sequence in the *fem-3* mRNA (Figure 1). The *fbf-1* gene sequence was then used to search DNA

databases for homologs. A second gene, *fbf-2*—91% identical to *fbf-1* and encoding another protein that bound to the *fem-3* 3'UTR—was identified in the *C. elegans* genome. Perhaps more significant was that both gene products, FBF-1 and FBF-2, were identified as members of the newly described PUF family of RNA-binding proteins, a group that includes the protein Pumilio.

RNAi experiments confirmed the role of FBF proteins in germ cell fate. Studies of the FBF proteins quickly became part of a still expanding effort to characterize the function of PUF family proteins.

(a) Three-hybrid system



(b) Binding specificity

Hybrid RNA	$\beta$ -Galactosidase filter assay	$\beta$ -Galactosidase units
<i>fem-3</i> wild type UCUUG —		370
<i>fem-3 ch8</i> AGAAC —		3
<i>fem-3 cq96</i> UCUUU —		3
IRE —		6
$A_{30}$ —		5
HIV-E —		8

**FIGURE 1** (a) The protein FBF-1 was identified in a three-hybrid screen. The hybrid RNA engineered to be the link between the binding protein MS2 and the (unknown) protein X-Gal4p fusion contained two PME sequences. (b) Activation of the *his3-lacZ* reporter gene leads to production of  $\beta$ -galactosidase, which catalyzes the conversion of X-gal to a colored product. The color is seen only in cells expressing RNA containing PMEs of the proper sequence. The RNAs in other lanes of the gel provide controls that demonstrate specificity of binding. IRE is iron response element;  $A_{30}$ , a sequence of 30 A residues; HIV-E, a 573-nucleotide RNA sequence derived from HIV. [Source: Adapted from B. Zhang et al., *Nature* 390:477-484, 1997, Fig. 2.]

## Little RNAs Play a Big Role in Controlling Gene Expression

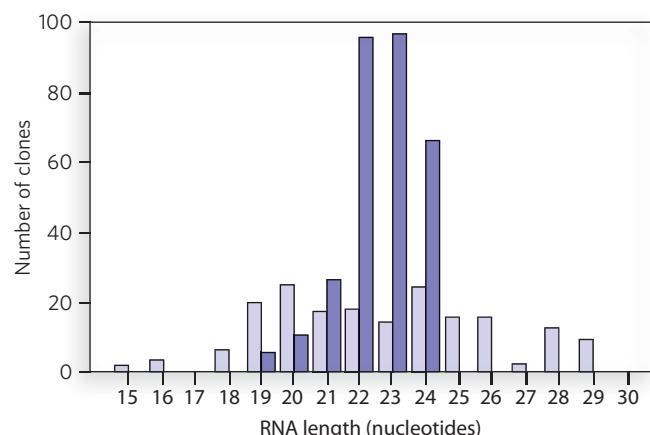
**Lagos-Quintana, M., R. Rauhut, W. Lendeckel, T. Tuschl.** 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294:853-858.

**Lau, N.C., L.P. Lim, E.G. Weinstein, and D.P. Bartel.** 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858-862.

**Lee, R.C., and V. Ambros.** 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294:862-864.

Scientific discovery has a way of occurring in bursts of insight, often with input from multiple research teams whose ideas and experiments converge on a new line of thinking. In the field of RNA interference, such a conceptual breakthrough occurred in 2001 with the finding by three different labs that small regulatory RNA molecules are abundant in eukaryotic cells. Scientists had come to suspect that small RNAs might be produced normally in cells as a means of controlling gene expression. This suspicion was based on the discovery by Craig Mello and Andrew Fire that double-stranded RNA, when fed to *C. elegans*, could silence gene expression. Research teams led by Victor Ambros, David Bartel, and Thomas Tuschl set out to find evidence of small regulatory RNAs that might be produced naturally in cells.

Each team took a similar experimental approach in which *C. elegans* or mammalian cells were grown in the laboratory, then total cellular RNA was isolated. The total RNA was fractionated by size to enable purification of RNA molecules about 20 to 30 nucleotides long, the size of the molecules used in the Mello and Fire experiments. To identify these molecules, the RNAs were covalently linked at their 3' ends to oligonucleotide sequences that provided binding sites for a complementary oligonucleotide, which could be used to prime the reverse transcription of the RNA into DNA. The complementary strand of this DNA sequence could be produced in a similar fashion, by covalently attaching it to a second oligonucleotide of defined sequence at its 3' end. Once the small RNAs had been copied into double-stranded DNA, they were cloned into plasmids using standard techniques (see Chapter 7). The plasmids could then be propagated in bacterial cell culture, purified, and sequenced. The sizes of the small RNAs identified in *C. elegans* all fell within a narrow range (19 to 24 bp), as opposed to the sizes of RNAs originating from *E. coli*, generated in a separate control experiment (**Figure 2**). These results suggested that the small RNAs cloned from *C. elegans* were indeed the product of



**FIGURE 2** One class of short RNAs derived from *C. elegans* show a characteristically narrow length distribution (dark blue bars) compared with clones of *E. coli* RNA fragments (light blue bars) produced using the same protocol in an organism that lacks miRNAs. The short but uniform lengths of the *C. elegans* RNAs provided some of the evidence that the RNAs were functional, and not degradation products. [Source: Adapted from N. C. Lau et al., *Science* 294:858-862, 2001, Fig. 2a.]

transcription of the *C. elegans* genome, and not simply short degradation products.

The sequences of these small RNAs proved very exciting, because in many cases they were complementary to sequences found in the host genome. This finding suggested that the small RNAs were produced as part of a large regulatory pathway in which small RNA molecules, dubbed microRNAs (miRNAs), could base-pair with target sequences in mRNAs. Subsequent experiments verified that this mechanism occurs widely in eukaryotes.

Why were miRNAs overlooked by molecular biologists for so long? One reason is simply size: because these RNAs are so small, they tended to be ignored or were thought to be irrelevant degradation products rather than functional RNAs produced by the cells.

## Everything Old Is New Again: Beauty at the Turn of a Developmental Switch

Carroll, S.B., J. Gates, D.N. Keys, S.W. Paddock, G.E. Panganiban, J.E. Selegue, and J.A.

Williams. 1994. Pattern formation and eyespot determination in butterfly wings *Science*

265:109–114.

As children, we may never notice the diminutive fruit fly, but butterflies rarely fail to inspire fascination. The bold colors and patterns in a butterfly wing that still catch our eye—surely these are the product of an elaborate developmental program that is distinct from the one operative in fruit flies? Or is it?

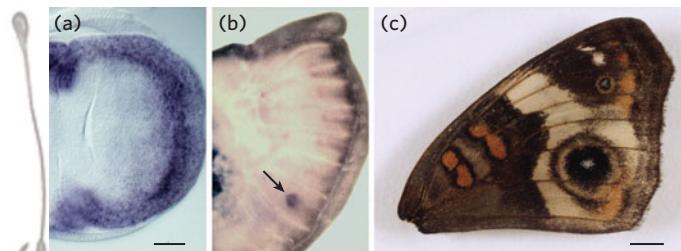
Sean Carroll's boyhood fascination with butterflies was eventually translated into research in a lab at the University of Wisconsin, where he studies insect development. In the early 1990s, it was already clear that many genes that control development are highly conserved, not just in insects, but in all higher eukaryotes. In setting out to decipher the development of butterfly wing patterns, Carroll decided that the genes known to affect the development of *Drosophila* wings was a good place to start.

Carroll's subject was the butterfly *Precis coenia*, also called the Buckeye, found each summer over much of the United States. His laboratory succeeded in cloning a series of *P. coenia* genes homologous to *Drosophila* genes known to control wing development, including the genes for signaling proteins called Wingless and Decapentaplegic, and transcription factors called Apterous, Invected, Scalloped, and Distal-less.

From the cloned genes, Carroll and his colleagues made labeled DNA probes and developed *in situ* hybridization methods to reveal the location of mRNAs in the butterfly wing at different stages of development. For months, the expression patterns looked identical to those already defined for the same genes in *Drosophila*. That changed when they



Sean Carroll [Source: Courtesy of Sean Carroll.]



**FIGURE 3** Expression of the *distal-less* gene (dark coloring) is revealed by *in situ* hybridization at three different stages of upper wing development in *P. coenia*. (a) The wing bud. (b) Partially developed wing. Expression of *distal-less* is evident at the point where an eyespot will develop (arrow). (c) Fully developed wing with eyespot. [Source: S. B. Carroll et al., *Science* 265:109–114, 1994.]

got to *distal-less*, a gene that helps control appendage development in animals from insects to humans (Figure 3). Carroll describes it as one of his most thrilling moments in science.

One day, Carroll's student, Julie Gates, called him over because she saw a pattern of genes turned on in so-called eyespots, the concentric rings of pigment in butterfly wings that look like eyes and are used in both mate recognition and predator avoidance. Carroll was stunned to be staring at the developing spotted pattern of gene expression, realizing that the gene involved, *distal-less*, had been around for at least 500 million years and was used in other organisms for appendage-patterning. In fruit flies, there is no counterpart to eyespot development, and it was suddenly clear that the ancient gene had been recruited to an entirely new function in butterflies. This turned out to be the first example of what became a major theme in understanding the evolution of animal form: old genes occasionally evolve to do something completely new.

[Background photo source: Courtesy of the Sean Carroll Laboratory.]

## Key Terms

alternative splicing, p. 769	RNA-induced silencing complex (RISC), p. 784	totipotent, p. 798
upstream open reading frame (uORF), p. 777	short interfering RNA (siRNA), p. 785	pluripotent, p. 799
iron homeostasis, p. 779	maternal gene, p. 793	unipotent, p. 799
iron response protein (IRP), p. 779	segmentation gene, p. 793	embryonic stem cell, p. 799
iron response element (IRE), p. 779	gap gene, p. 793	adult stem cell, p. 799
AU-rich element (ARE), p. 780	pair-rule gene, p. 793	multipotent, p. 799
RNA interference (RNAi), p. 783	segment polarity gene, p. 793	niche, p. 799
microRNA (miRNA), p. 783	homeotic gene, p. 793	evo-devo, p. 801
primary miRNA transcript (pri-miRNA), p. 784	Wnt-class signaling pathway, p. 795	
precursor miRNA (pre-miRNA), p. 784	Hox gene, p. 796	
	stem cell, p. 798	

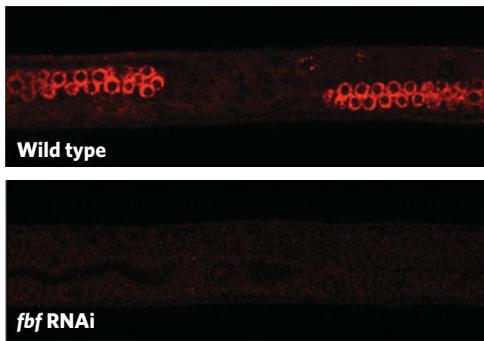
## Problems

1. An investigator monitors the production of a particular mRNA in a mouse cell line. Expression of the mRNA is induced (i.e., its concentration in the cytoplasm increases) in response to the addition of a hormone. The observed increase in the mRNA concentration is blocked by actinomycin D and by cycloheximide. What does this tell you about the requirements for increased expression of this mRNA?
2. In female fruit fly embryos, the Sxl protein initially generated by transcription from the P<sub>e</sub> promoter differs somewhat from that generated by later transcription from the P<sub>m</sub> promoter. In what part of the protein does this difference arise?
3. As an investigator, you wish to scan the sequences of all the genes of the mouse genome to determine how many genes have multiple sites for 3'-end cleavage. What sequence would you scan for? Would a scan for a single sequence find all of the sites?
4. In eukaryotes, phosphorylation of the translation initiation factor eIF2 $\alpha$  blocks translation of virtually all mRNAs. In a mammalian reticulocyte, a deficiency in iron or heme leads to eIF2 $\alpha$  phosphorylation to block the translation of globin mRNAs. The phosphorylation of eIF2 $\alpha$  does not create a problem for other cellular functions in reticulocytes. Suggest why.
5. What is the likely fate of an mRNA transcript containing the sequence (a) AAUAAA or (b) AUUUA?
6. Suggest at least three cellular mechanisms that could establish a gradient of either a protein or an mRNA during maturation of an oocyte.
7. In *C. elegans*, the Pal-1 protein specifies some developmental fates in cells where it is highly expressed. Translation of the Pal-1 mRNA is suppressed by binding of the Mex-3 protein, concentrated in anterior cells of the embryo, to the 3'UTR of the mRNA. What would be the probable effect of mutations that eliminated the binding of Mex-3 to this 3'UTR?
8. A *Drosophila* female embryo that is *bcd*<sup>-</sup>/*bcd*<sup>-</sup> may develop normally, but the adult fruit fly will not be able to produce viable offspring. Explain.
9. In the *Drosophila* ovary, a germ-line stem cell repeatedly divides. After each division, one daughter cell retains stem cell identity and the other begins to differentiate into an oocyte. The germ-line stem cell is associated with additional cells called cap cells. Describe how the asymmetric divisions might occur, and the possible role of the cap cells.
10. The stem cell genes that regulate tissue regeneration tend to be highly conserved. *Planaria* (an aquatic flatworm) has an impressive capacity to regenerate its head and other structures when they are amputated, making this a favorite subject in grade school science labs. In the wild, *Planaria* eats smaller worms and eukaryotic organisms in its environment. In the lab, it can be fed clumps of bacteria mixed with pieces of liver and agar. As a biologist, you know that tissue regeneration mechanisms are likely to be conserved. You are interested in determining which *Planaria* genes are needed to guide head regeneration. Your reading tells you that regeneration depends on certain stem cells posterior to the animal's photoreceptors and excluded from its pharynx. Using methods described in this chapter, how would you go about discovering the key genes?

## Data Analysis Problem

**Zhang, B., M. Gallegos, A. Puoti, E. Durkin, S. Fields, J. Kimble, and M.P. Wickens. 1997.** A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390:477–484.

- 11.** The collaborative effort by Judith Kimble, Marvin Wickens, and colleagues, as described in their 1997 paper (see How We Know), resulted in discovery of the FBF proteins that bind the 3'UTR of the mRNA from the *fem-3* developmental regulatory gene of nematodes. Compare the three-hybrid strategy used in this study (see How We Know, Figure 1) with the three-hybrid method presented in Chapter 7.
- (a) How was the three-hybrid method modified in the Kimble and Wickens study?



**FIGURE 1**

## Additional Reading

### General

- Carroll, S.B. 2005.** *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom.* New York: W. W. Norton & Company.
- Granneman, S., and S.J. Baserga. 2006.** Crosstalk in gene expression: Coupling and co-regulation of rDNA processing. *Curr. Opin. Cell Biol.* 17:281–286.
- Rubin, G.M., and E.T. Lewis. 2000.** A brief history of *Drosophila*'s contributions to genome research. *Science* 287:2216–2218.

### Posttranscriptional Control inside the Nucleus

- Bentley, D. 2005.** Rules of engagement: Co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* 17:251–256.
- Blencowe, B.J. 2006.** Alternative splicing: New insights from global analyses. *Cell* 126:37–47.
- Frankel, A.D., and J.A.T. Young. 1998.** HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* 67:1–25.
- Neeman, Y., D. Dahary, and K. Nishikura. 2006.** Editor meets silencer: Crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.* 7:919–931.

Six different RNA sequences were screened for FBF-1 binding. The normal PME sequence (UCUUG) gave a positive response. The other five sequences were two 5-nucleotide sequences, an iron response element, a segment containing 30 consecutive A residues, and a 573-nucleotide RNA sequence derived from HIV.

- (b) Of these controls, which would make the best case for specific binding of the PME by FBF-1?  
 (c) Why might the other controls be useful?

The investigators used immunofluorescence to detect expression of the FBF-1 protein in wild-type nematodes, as shown in the upper panel of Figure 1. All cells illuminated are in the germ line. The dark spots are cell nuclei.

- (d) What conclusion can you draw from the protein expression pattern in the upper panel?

The lower panel in Figure 1 shows results for an animal treated with RNAi directed at the gene for FBF-1 (*fbf*). The expression of FBF-1 is essentially abolished.

- (e) Given the function of FBF-1 in the germ line, what is the likely effect of this RNAi treatment on the germ line of the treated animals?

### Translational Control in the Cytoplasm

- Bashirullah, A., R.I. Cooperstock, and H.D. Lipshitz. 1998.** RNA localization in development. *Annu. Rev. Biochem.* 67:335–394.
- Fabian, M.R., N. Sonenberg, and W. Filipowicz. 2010.** Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79:351–379.
- Gingras, A.-C., B. Raught, and N. Sonenberg. 1999.** eIF4 initiation factors: Effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu. Rev. Biochem.* 68:913–963.
- Gray, N.K., and M. Wickens. 1998.** Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.* 14:399–458.
- Hartmann-Petersen, R., M. Seeger, and C. Gordon. 2003.** Transferring substrates to the 26S proteasome. *Trends Biochem. Sci.* 28:26–31.

### The Large-Scale Regulation of Groups of Genes

- Goodrich, J.A., and J.F. Kugel. 2006.** Non-coding RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* 7:612–616.

- Green, R., and J.A. Doudna.** 2006. RNAs regulate biology. *ACS Chem. Biol.* 1:335–338.
- Kapp, L.D., and J.R. Lorsch.** 2004. The molecular mechanics of eukaryotic translation. *Annu. Rev. Biochem.* 73:657–704.
- Liu, Q., and Z. Paroo.** 2010. Biochemical principles of small RNA pathways. *Annu. Rev. Biochem.* 79:295–319.

### RNA Interference

- Agami, R.** 2002. RNAi and related mechanisms and their potential use for therapy. *Curr. Opin. Chem. Biol.* 6:829–834.
- Cerutti, H.** 2003. RNA interference: Traveling in the cell and gaining functions? *Trends Genet.* 19:9–46.
- Crunkhorn, S.** 2010. RNA interference: Clinical gene-silencing success. *Nat. Rev. Drug Discov.* 9:359–359.
- Kamath, R.S., A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapink, N. Le Bot, S. Moreno, M. Sohrmann, et al.** 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421:231–237.

### Putting It All Together: Gene Regulation in Development

- Blow, N.** 2008. In search of common ground. *Nature* 451:855–858.
- Farley, B.M., and S.P. Ryder.** 2008. Regulation of maternal mRNAs in early development. *Crit. Rev. Biochem. Mol. Biol.* 43:135–162.
- Gönczy, P.** 2008. Mechanisms of asymmetric cell division: Flies and worms pave the way. *Nat. Rev. Mol. Cell Biol.* 9:355–366.

- Li, L., and T. Xie.** 2005. Stem cell niche: Structure and function. *Annu. Rev. Cell Dev. Biol.* 21:605–631.
- Morrison, S.J., and J. Kimble.** 2006. Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature* 441:1068–1074.
- Nüsslein-Volhard, C., and E. Wieschaus.** 1980. Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287:795–801.
- Passier, R., L.W. van Laake, and C.L. Mummery.** 2008. Stem-cell-based therapy and lessons from the heart. *Nature* 453:322–329.
- Pera, M.F.** 2008. Stem cells: A new year and a new era. *Nature* 451:135–136.
- Tajbakhsh, S., P. Rocheteau, and I. Le Roux.** 2010. Asymmetric cell divisions and asymmetric cell fates. *Annu. Rev. Cell Dev. Biol.* 25:671–699.
- Takai, Y., J. Miyoshi, W. Ikeda, and H. Ogita.** 2008. Necdtins and nectin-like molecules: Roles in contact inhibition of cell movement and proliferation. *Nat. Rev. Mol. Cell Biol.* 9:603–615.

### Finale: Molecular Biology, Developmental Biology, and Evolution

- Carroll, S.B.** 2005. Evolution at two levels: On genes and form. *PLoS Biol.* 3:1159–1166.
- Carroll, S.B., B. Prud'homme, and N. Gompel.** 2008. Regulating evolution. *Sci. Am.* 298(5):60–67.
- Prud'homme, B., N. Gompel, and S.B. Carroll.** 2007. Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. USA* 104:8605–8612.
- Raff, R.A.** 2000. Evo-devo: The evolution of a new discipline. *Nat. Rev. Genet.* 1:74–79.

*This page intentionally left blank*

# Appendix

---

## Model Organisms

Bacterium, *Escherichia coli*

Budding Yeast, *Saccharomyces cerevisiae*

Bread Mold, *Neurospora crassa*

Nematode, *Caenorhabditis elegans*

Mustard Weed, *Arabidopsis thaliana*

Fruit Fly, *Drosophila melanogaster*

House Mouse, *Mus musculus*

Humans are set apart from other organisms by their cognitive abilities and sense of wonder about their surroundings, and it is this curiosity that drives us to study life. What are we made of? How do we work? To understand humans, and other living creatures, scientists have explored a wide variety of organisms. These studies have revealed a great deal, including the striking universal features shared by all living things. All organisms use the same amino acids, the same nucleotides, and essentially the same genetic code.

There is more to molecular biology than satisfying our curiosity about how life works. We also strive to understand the causes of disease and to apply our understanding to medicine, agriculture, and technology. This book has pointed out numerous examples of how we have learned about human diseases—their causes, and in some cases how to treat, cure, or prevent them. Scientists have discovered antibiotics to treat most bacterial infections, developed vaccines for many types of viral infections, and now understand a great deal more about cancer and its treatments. The vast majority of these discoveries and developments came from studying model organisms.

### A Few Organisms Are Models for Understanding Common Life Processes

Scientists study a diversity of organisms. When a particular species is chosen for intensive investigation by many laboratories, it is referred to as a model organism.

This focus on one species by many labs allows the development of a large body of information that provides deep insights into that organism's living functions. The organism is considered a model because researchers assume that what they learn about it will hold true for related organisms. The particular organism selected for study depends on the questions being asked. Throughout this book, we encounter the contributions of model organisms to our knowledge of molecular biology, and several of the most frequently used organisms are reviewed here.

We should note, however, that sometimes an organism that is “off the beaten track” is studied by only a few laboratories, purely out of curiosity—and these investigations can also have a profound effect on research. For example, *Thermus aquaticus* gave us Taq polymerase for the polymerase chain reaction (PCR; see Chapter 7, How We Know). The study of *Tetrahymena thermophila* led to the discovery that RNAs can act as catalysts (i.e., ribozymes; see Chapter 16). And studies of some little known insect viruses, the baculoviruses, gave rise to a recombinant protein expression system that is now in wide use (see Chapter 7).

Focusing on a handful of different organisms is important at a practical level, as there are many more species than there are scientists. Indeed, developing an organism into a scientific tool of research is not easy. It requires many years of study to understand the organism and become familiar with its life cycle, its proper nutrition, and its optimum growth and storage conditions. Especially time-consuming is the development of

## A-2 Appendix: Model Organisms

genetic tools to manipulate the organism's genome. There are no "standard procedures" for this; genetic tools are largely specific to the organism and are often found by trial and error. This is why it is important that many laboratories work on the same organism and share their knowledge. A critical mass of interest in a model organism eventually leads to international conferences, online databases, and the formation of stock centers that maintain and distribute strains.

Of all the organisms in the world, why were certain ones chosen as models? The choices were often made with a healthy dose of serendipity. However, some common features underlie the utility of an organism as a model. Model organisms should have a rapid life cycle. They should produce many progeny, so that researchers can find and study rare genetic events. Size is important too, because large organisms and their numerous large progeny would quickly exhaust the space of a typical laboratory. Model organisms should be easily propagated using a simple and inexpensive food source. There should be a convenient method of long-term storage for accumulating strains for further study. **Table A-1** summarizes some features of the model organisms described here.

Studies on genetics and metabolic pathways in the early 1900s used complex multicellular organisms such as plants, fruit flies, rats, and mice. Later, researchers recognized that single-celled organisms are also amenable to fundamental studies of genes and cellular metabolism. In the 1940s, microbes such as *Escherichia coli*, yeast, and *Neurospora crassa* became the most useful models for understanding the basic chemistry of life. They also provided better starting material for biochemists than did animal tissues, because single-celled organisms are a uniform population of identical cells, whereas tissues are composed of different cell types.

No single model organism can answer all questions about life. Single-celled organisms continue to teach us about central aspects of fundamental life processes, such as chromosomal replication, DNA repair, recombination, gene expression, signal pathways, and control of the cell cycle. But single-celled organisms are insufficient for addressing questions about the development of multicellular organisms and most types of disease. Thus, the nematode worm and fruit fly are of enormous use in revealing how multicellular organisms are organized and the basics of how the animal body plan is determined. These organisms also provide insights about many types of disease. Similarly, the mustard weed was chosen as a model organism for plant development.

By far the most useful model of human disease is the mouse. It is, however, not the simplest of model organisms. For ease of growth and DNA manipulation,

the mouse pales in comparison with the other model organisms. Genetic strains of mice are costly and time-consuming to construct. But one of the great advantages of the mouse is that 99% of its genes have homologs in the human, including the genes associated with human disease. So despite the difficulties, the mouse is an attractive model in which to study the diseases that afflict us.

We present here a brief overview of several model organisms in use today, including how they have contributed to, and continue to further, our understanding of life. As we've noted, many other organisms have also contributed greatly to our understanding of living processes, including bacteriophages and other viruses, *Tetrahymena thermophila* (a protozoan), *Schizosaccharomyces pombe* (fission yeast), *Xenopus laevis* (frog), and *Brachydanio rerio* (zebra fish). Before we launch into details of particular model organisms, we briefly describe a few highlights of how we learn about human disease from studying model organisms, in conjunction with genomics and cell culture.

### Three Approaches Are Used to Study Human Disease

What causes heart disease, diabetes, neurodegenerative disorders, or cancer? How can these, and other diseases, be prevented, treated, or cured? The study of model organisms is usually the first step in understanding cellular processes that can be altered in human diseases. Using a homolog, we can study a human disease-causing mutation in a model organism and learn how the mutation disrupts the cellular process at the molecular, cellular, or organismal level. In addition, mouse models are often used to test treatments for diseases, as an early step in drug development. Model organisms are often the first avenue to understanding human disease, but human genomics and cell culture can provide an even deeper understanding.

The availability of the complete human genome sequence has been an enormous aid in our understanding of human disease genes at the molecular level, as well as in bioinformatics studies on human evolution and migrations (see Chapter 8). Our capacity for language and written history has played a large role in elucidating the genetics of human disease. In particular, people actively seek out medical and scientific advice for a disease, and often can recall a family pedigree stretching back generations that might provide information about how the disease is transmitted. We see examples of this throughout the book, including hemophilia in royalty (see Figure 2-27), sickle-cell anemia (see Highlight 2-1), and early-onset Alzheimer disease

**Table A-1** Basic Information on Seven Common Model Organisms

	<i>E. coli</i>	<i>S. cerevisiae</i>	<i>N. crassa</i>	<i>C. elegans</i>	<i>A. thaliana</i>	<i>D. melanogaster</i>	<i>M. musculus</i>
<i>Basic Facts</i>							
Type of organism	Bacterium, single-celled	Eukaryote, single-celled ascomycete fungus	Eukaryote, multinucleate filamentous ascomycete fungus	Eukaryote, nematode worm	Eukaryote, angiosperm plant	Eukaryote, insect	Eukaryote, mammal
Natural habitat	Animal intestine	Plant surfaces	Dead vegetation	Soil	Global, temperate climate	Rotting fruit	House, fields
Size	1–2 $\mu\text{m}$	4–6 $\mu\text{m}$	Irregular	1 mm	10–20 cm	2.5 mm	17 cm
Reproduction	Asexual fission; some conjugation	Asexual, budding (mitosis); sexual, sporulation (meiosis)	Asexual, filamentation; sexual, sporulation	Sexual, self- and cross-fertilization; some asexual	Sexual, self- and cross-fertilization	Sexual, mating	Sexual, mating
Generation time	20 min	70–90 min	3–4 weeks (for sexual reproduction)	50 h	6 weeks	12 days	9 weeks
Growth conditions	Petri plates or liquid culture	Petri plates or liquid culture	Petri plates, race tubes, or liquid culture	Petri plates or liquid culture	Petri plates or small horticultural trays	Bottles or vials	Cages
Food source	Yeast extract and tryptone broth; defined media possible	Yeast extract, peptone; defined media possible	Complete or defined media (sugar, inorganic salts, biotin, and nitrogen)	Bacteria ( <i>E. coli</i> )	Light, water, source of nitrogen and other minerals (i.e., fertilizer)	Yeast and molasses, fruit	Plant matter
Storage	Frozen glycerol stocks; lyophilized cells	Frozen glycerol stocks; lyophilized cells	Freeze-dried spores	Frozen at $-80^\circ\text{C}$	Seeds	Live propagation	Frozen embryos

(continued)

**Table A-1** Basic Information on Seven Common Model Organisms (*continued*)

	<i>E. coli</i>	<i>S. cerevisiae</i>	<i>N. crassa</i>	<i>C. elegans</i>	<i>A. thaliana</i>	<i>D. melanogaster</i>	<i>M. musculus</i>
<i>Genetic Statistics</i>							
Genome size	4,639,675 bp	12,070,898 bp	42,900,000 bp	100,269,917 bp	119,186,997 bp	166,600,000 bp	2,729,273,687 bp
Chromosome number	1 (haploid)	16 (haploid); 32 (diploid)	7 (haploid); 14 (diploid)	11 (diploid, male), 12 (diploid, hermaphrodite)	10 (diploid)	8 (diploid)	40 (diploid)
Gene number	4,410	6,000	11,000	21,035	25,540	14,065	29,083
Genes similar to human	8%	30%	6%	25%	18%	50%	99%
Gene naming convention	<i>dnaA</i>	<i>GCN4</i>	<i>arg</i>	<i>fem-3</i>	AAO	<i>ry</i>	<i>Cdc20</i>
Protein naming convention	DnaA	Gcn4	arg	FEM-3	AAO	RY	Cdc20
Genome website	<a href="http://www.ecocyc.org">www.ecocyc.org</a>	<a href="http://www.yeastgenome.org">www.yeastgenome.org</a>	<a href="http://www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html">www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html</a>	<a href="http://www.wormbook.org">www.wormbook.org</a>	<a href="http://www.arabidopsis.org">www.arabidopsis.org</a>	<a href="http://www.flybase.org">www.flybase.org</a>	<a href="http://www.jax.org">www.jax.org</a>

(see Figure 8-11). Identifying human disease genes is a difficult but important task and is being accomplished at a rapid and accelerating rate. Knowledge of the genetics involved in transmitting a disease may help couples plan their families and cope with the possible maladies that may be passed on to their sons and daughters. For scientists, knowledge of a disease gene can help devise a treatment or cure.

A third way we study ourselves is by culturing individual human cells *in vitro*. Primary cells taken directly from the body and then grown in culture typically die within 40 (or fewer) generations. But cells taken from cancer tissue have altered growth control and can often be grown through countless generations; they are referred to as “immortalized.” Cells can sometimes even be removed from normal tissue and then immortalized in tissue culture by

infection with particular viruses. Through these means and others, many different types of human tissue cells are grown and maintained in culture, including hepatocytes (liver), renal cells (kidney), fibroblasts (skin), glial cells (nerve), lymphocytes (blood), and myocytes (muscle). By investigating cancer cells and transformation agents, we have also learned a great deal about the genes involved in cancer (see Highlight 12-1). Studies of human and other primate cells in tissue culture have provided important information about surface receptors, protein trafficking, viral entry, and cellular reproduction. Human tissue cells can even be grown in quantities suitable for biochemical studies (see Chapter 7). Recent advances in stem cell research hold promise for the treatment of many diseases and for developing replacement tissue (see Chapter 22).



## Bacterium, *Escherichia coli*

*Escherichia coli* is a single-celled bacterium. It is a natural occupant of the animal gut and is typically harmless, although unusual toxic varieties exist. *E. coli* is small (microscopic), reproduces rapidly by fission to form clonal colonies on agar plates, with tens of millions of cells produced within a day, and can also be grown in liquid media to produce astounding numbers of progeny—an important characteristic for studies of rare genetic events. *E. coli* contains a single circular chromosome, and studying mutants is much easier in a haploid organism than in a diploid organism, as there are no dominant genes to mask a mutation. Common mutant phenotypes are resistance to bacteriophages, ability to grow on certain carbon sources, antibiotic resistance, and colony size and shape. *E. coli* also offers a plentiful, uniform source of starting material for biochemical studies. Thus, the study of *E. coli* combines the power of genetics with the power of biochemistry to understand life processes at the molecular level.

*E. coli* is not a eukaryote and thus has limited homology to humans. This restricts its usefulness in studying all but the most basic cellular processes. For example, *E. coli* lacks nucleosomes, and most of its proteins are unmodified. Eukaryotic protein expression in transformed *E. coli* sometimes doesn't work, because it requires posttranslational modifications that the bacterium cannot carry out.

### Early Studies of *E. coli* as a Model Organism

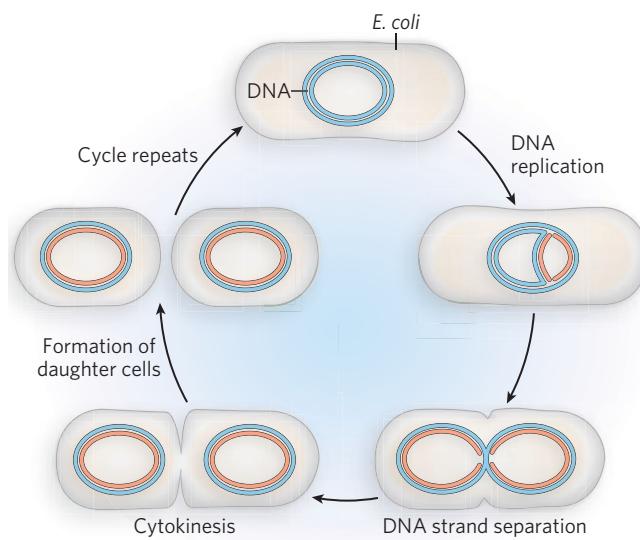
The marriage of biochemistry and genetics made *E. coli* a rich source for the discovery of new information. The famous "fluctuation analysis" by Salvadore Luria and Max Delbrück in 1943 demonstrated that *E. coli* mutates spontaneously to become resistant to bacteriophage  $\lambda$  and passes the phage-resistance trait to new progeny—thus establishing *E. coli* as a model system for understanding the nature and function of genes. The discovery by Joshua Lederberg and Edward Tatum that *E. coli* undergoes a type of mating and crossing over (DNA conjugation mediated by the F plasmid)

made possible DNA exchange and complementation tests, firmly rooting *E. coli* as a model for genetics. In 1952, Alfred Hershey and Martha Chase used *E. coli* to identify DNA as the genetic material of T2 phage (see Chapter 2, How We Know). In 1958, Matthew Meselson and Frank Stahl selected *E. coli* for their studies to demonstrate that DNA replication occurs by a semi-conservative mechanism (see Figure 11-1). In the same year, Arthur Kornberg and colleagues published their results on the purification and characterization of the first-discovered DNA polymerase (see Chapter 11, How We Know).

Francis Crick and Sydney Brenner used *E. coli* genetics to establish the triplet nature of the genetic code in 1961. By 1966, the genetic code had been cracked by Marshall Nirenberg, Heinrich Matthaei, Gobind Khorana, and their colleagues, using synthetic oligonucleotides and *E. coli* extracts (see Chapter 17). A completely new era of gene regulation research was opened up by the studies of François Jacob and Jacques Monod on the detailed workings of the *lac* operon (see Chapter 20; see also Chapter 5, How We Know). The identification of plasmids in *E. coli*, and methods to handle them, ushered in the era of recombinant DNA technology. Plasmids continue to be wonderful biotech tools, still valuable beyond compare (see Chapter 7).

### Life Cycle

Like most bacteria, *E. coli* reproduces by binary fission (Figure A-1). As the cell grows, it senses when it has reached the correct size to begin replication. The duplicate chromosomes partition to opposite poles of the cell. Cytokinesis splits the cell in half to produce two daughter cells, each containing an identical chromosome. The entire process requires only about 20 minutes at 37°C, the average mammalian body temperature. Thus, a large, visible colony of millions of identical bacteria form within only 24 hours on a Petri plate of rich media. Huge quantities of *E. coli* can also be grown in liquid culture within a day.



**FIGURE A-1** The life cycle of *Escherichia coli*.

## Genetic Techniques

For a complete description of many of these techniques, see Chapter 7.

**Mutagenesis** Mutation frequency in *E. coli* can be enhanced by chemicals or UV light. Mutant cells can be selected by phenotype or screened by replica plating on different media.

**Plasmids** *E. coli* has several natural plasmids that can harbor foreign DNA in the cell. These plasmids are useful to researchers for protein expression, as shuttle vectors, or for building and sequencing large genomes.

**Introduction of DNA** *E. coli* can be made to accept foreign DNA by chemical transformation or electroporation. Cells transformed with plasmid DNA containing an antibiotic-resistance gene can be selected and maintained by adding that antibiotic to the growth medium.

**Gene Knockouts** Plasmids without a replication origin but containing an antibiotic-resistance gene can be selected for integration into the bacterial chromosome. Integration is typically achieved by homologous recombination, convenient for gene deletion, or knockout.

**Complementation** This technique is often used to prove that a mutation is in a particular gene by adding a wild-type gene to restore function. Complementation analysis of mutant cells can be performed with plasmids

that contain a wild-type copy of the gene of interest. Complementation analysis of unknown mutations often makes use of a genomic plasmid library. Positive clones (cells that survive under conditions in which the mutant cannot live) can then be sequenced to identify the complementing gene.

**Recombinant Protein Expression** Rapid and inexpensive growth of *E. coli* make it the most widely used vehicle for the expression and purification of foreign proteins. Foreign proteins are typically expressed from genes located on high-copy-number plasmids. Expression is induced using regulatory elements derived from well-studied *E. coli* and phage promoters.

## *E. coli* as a Model Organism Today

**Multiprotein Machines** Biochemical and structural studies are now illuminating the atomic details of structures and processes of information transfer that were once unknown. These include the structure and function of the DNA replication apparatus and the DNA repair and recombination proteins, regulation of RNA polymerase, and the structure and chemistry of the ribosome. These multiprotein structures and processes occur in bacteria and eukaryotes alike, but they are streamlined in the relatively simple *E. coli* cell.

**Cytologic Studies** Processes can be visualized in living *E. coli* cells through the use of fluorescent fusion proteins. For example, recent studies of this sort have allowed detection of the proteins involved in replication of the chromosome and their movements along the chromosome during replication.

**Recombinant DNA Technology** *E. coli* holds the central position in recombinant DNA technology. Its plasmids are widely used as shuttle vectors to amplify and maintain the DNA libraries of other organisms. *E. coli* is also still one of the most widely used vehicles for the expression and purification of foreign proteins.

**Systems Biology** Its relatively small genome makes *E. coli* an attractive system in which to pioneer approaches to systems biology. DNA chips have been constructed to examine the response of every transcript to a wide variety of conditions. The functions of about 35% of *E. coli* genes are still unknown, and a genome-wide collection of *E. coli* gene knockouts is helping to assign function to these genes, and to identify new gene interaction networks.



## Budding Yeast, *Saccharomyces cerevisiae*

*Saccharomyces cerevisiae* is a microbial ascomycete fungus, used by humans for centuries to make bread, beer, and wine. It is generally referred to by scientists as budding yeast, but is also called baker's yeast, brewer's yeast, and simply yeast. Yeast has several features that make it an attractive model organism. It is a single-celled haploid organism, reproduces rapidly in liquid media, forms clonal colonies on agar plates, and can be stored frozen. As is the case for *E. coli*, these features make yeast an excellent model organism for studying the genetics and biochemistry of fundamental life processes, such as mechanisms of replication, transcription, and translation. However, unlike *E. coli*, yeast is a eukaryote with a full complement of intracellular organelles, including mitochondria and a nucleus, with DNA packaged into chromatin, and with a cytoskeletal structure. *S. cerevisiae* is thus an apt model for mechanisms unique to eukaryotes, such as cell cycle control, regulation by protein phosphorylation, chromosome structure, and mitochondrial function.

The *S. cerevisiae* genome is only ~12 Mbp long, with ~6,000 genes. Yet it contains homologs to several human disease genes, making it a model for understanding the basic function of gene products involved in some forms of disease, as well as indicating that some human diseases result from the disruption of relatively simple and fundamental life processes. However, the single-celled yeast does not serve as a model of development, and the many human diseases arising from mutations that cause developmental abnormalities cannot be studied in yeast, limiting its usefulness in medical research.

The haploid nature of yeast allows convenient mutational studies. Yeast is also easily transformed with exogenous DNA. This DNA frequently integrates into the chromosome by homologous recombination, making possible the construction of gene knockout strains. The ease of genetics, combined with the ease of growth, makes *S. cerevisiae*, like *E. coli*, a model organism that beautifully combines the power of genetics with that of biochemical studies.

### Early Studies of Yeast as a Model Organism

Yeast has been a model organism for more than 100 years, and it essentially spawned the field of modern biochemis-

try. Louis Pasteur proved that yeast ferments sugar to alcohol and CO<sub>2</sub>; he declared fermentation a vital life force that could not be separated from the living organism. In 1897, Eduard Buchner accidentally observed fermentation in a yeast cell extract, and called the fermenting substance "zymase." Later studies showed that zymase was actually a mixture of many different enzymes. The suffix -ase is used for many enzyme names to this day.

### Life Cycle

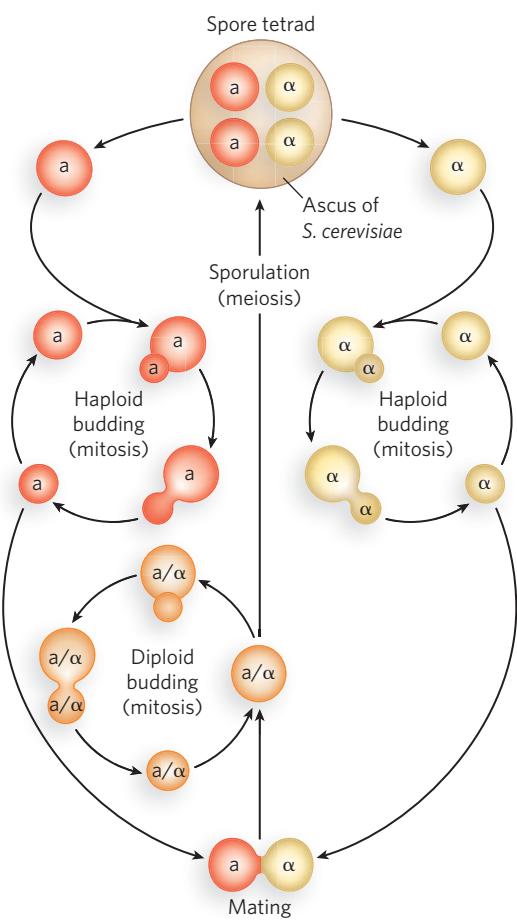
Budding yeast can exist in either the haploid or the diploid state (Figure A-2). Both haploid and diploid cells can reproduce asexually by budding. Sex in *S. cerevisiae* is determined by the gene products of the mating type (*MAT*) locus, for which there are two alleles, **a** and **α** (see Figure 13-22). Haploids of different mating types can be mated to form diploids by plating them together. Diploid **a/α** yeast can be converted to the haploid state in media that induce them to sporulate, undergoing meiosis to produce four haploid spores contained in an ascus. The spore tetrad can be dissected with a micromanipulator, and each haploid phenotype can be analyzed after growing the spores into colonies.

### Genetic Techniques

**Mutagenesis** Mutational studies in *S. cerevisiae* are simplified by its ability to grow as a haploid. Yeast can be replica-plated to study mutant phenotypes for growth on different carbon sources and to screen for conditional lethal genes (i.e., genes that, when mutated, allow the organism to grow only at a particular temperature).

**Complementation** Haploid yeast can be mated to produce diploid cells for complementation analysis of mutant haploids. Mutations in essential genes can be maintained by harboring them in the diploid state. On sporulation of a diploid with a mutation in an essential gene, two of the four spores will not be viable. Isolation of all four products of meiosis in an ascus is useful for the analysis of recombination during meiosis.

**Introduction of DNA** DNA can be introduced into *S. cerevisiae* by chemical or physical abrasion of the thick



**FIGURE A-2** The life cycle of *Saccharomyces cerevisiae*.

cell wall. Linear DNA, with a selectable marker and ends homologous to the chromosome, readily integrates into the genome by homologous recombination. This configuration can permit the ends to recombine with their homologous sequence in the yeast chromosome, replacing the gene of interest with the selectable marker and effectively knocking out the gene.

**Selectable Markers** Selection in yeast is mostly performed with auxotrophs, strains that lack a functional gene for an enzyme in an amino acid biosynthetic pathway and therefore require that amino acid to grow. Cells are grown on defined media lacking the now essential amino acid. Only cells that acquire the marker and thus are capable of synthesizing the amino acid from raw materials will survive. Drug-resistance markers can also be used for selection.

**Plasmids** DNA can also be maintained in yeast as a circular plasmid, using one of the many chromosomal origins known as an ARS (autonomously replicating

sequence). YACs (yeast artificial chromosomes) can maintain very large DNA inserts (300 kbp to 1 Mbp), which are useful in genome sequencing projects (see Figure 7-7). *S. cerevisiae* also contains a natural 2 micron ( $2\mu$ ) plasmid, which can be used to maintain exogenously derived DNA inside the cell.

## Yeast as a Model Organism Today

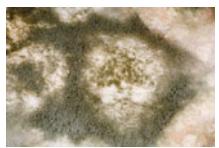
**Fundamental Mechanisms of Information Flow** The simple, single-celled structure, the ability to grow in large quantities for biochemical studies, and the ease of genetic manipulation—all make yeast ideal for studying detailed mechanisms of fundamental processes, including replication, gene regulation, DNA repair, transcription, translation, and recombination.

**Cell Division Cycle** When yeast divides, the bud is visible in the microscope, along with size changes at different phases of the cell cycle. Accumulation of mutant cells stuck in a budded state has been used as a phenotype by which to identify numerous cell division cycle (*cdc*) mutants, useful for genetic and biochemical dissection of the cell cycle. Humans have homologs to many of the cell cycle genes discovered in yeast.

**Signaling Pathways** Yeast is a model for signal transduction pathways, such as those involved in mating responses to pheromones produced by cells of opposite mating types.

**Meiotic Recombination** Because *S. cerevisiae* produces an ascus with four spores derived from a single cell, researchers can follow crossover events that occur during meiosis. This has resulted in detailed molecular models of meiotic recombination.

**Systems Biology** It is now within the scope of systems biology to understand most of the genetic, physical, and molecular interactions of the 6,000 yeast genes. Pairwise examinations of genetic interactions between these gene knockouts are leading to the identification of new pathways. Proteomic approaches have been pioneered in yeast. Intracellular copy numbers of almost all yeast proteins, genome-wide, have been determined. Protein-protein and protein-RNA interactions have also been assessed on a global scale, using yeast two-hybrid and three-hybrid techniques (see Chapter 7) and mass spectroscopy of protein complexes. Cell biological studies are facilitated by a library of fluorescent protein tags for every open reading frame in the yeast genome. In addition, yeast is used as a protein expression system for the study of foreign genes.



## Bread Mold, *Neurospora crassa*

The bread mold, *Neurospora crassa*, is a filamentous ascomycete fungus. It was one of the first eukaryotic microorganisms adopted for genetic studies. About 75% of all fungi are ascomycetes, and the remainder are basidiomycetes (mushrooms). About 90% of ascomycetes are filamentous (meaning multinucleate, as further explained below), and the rest are single-celled yeasts. The genome of *N. crassa* (often referred to simply as *Neurospora*) has ~43 Mbp and 11,000 genes, making it comparable to the fruit fly in genetic complexity.

The war between filamentous fungi and bacteria has yielded important antibiotics, including penicillin. These complex organic molecules are found nowhere else in nature, and accordingly, more than 40% of *Neurospora* genes have no identifiable counterpart in any other organism studied. The ecological niche filled by filamentous fungi includes the decay of many types of biological material. This may explain why *Neurospora* can grow on highly unusual carbon and nitrogen sources—which is probably reflected in other uncommon genes.

A unique feature of *Neurospora* is the arrangement in the spore case (ascus) of the spores produced from meiosis of a single diploid nucleus. The spores are linearly arranged in the order in which they are produced by the meiotic divisions. With the ability to precisely separate spores and study them individually, researchers have an ideal model for analysis of meiotic recombination. *Neurospora* is also attractive for studying several gene silencing mechanisms, some of which also exist in multicellular eukaryotes.

### Early Studies of *Neurospora* as a Model Organism

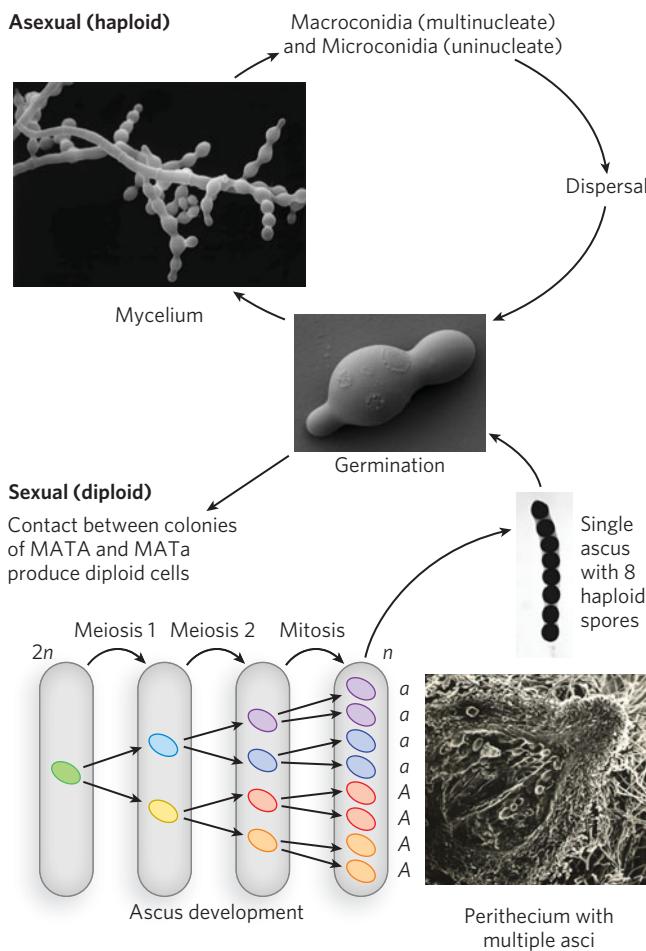
*Neurospora* achieved its early standing as a model organism because of its haploid state, rapid growth, production of millions of spores per colony, and ease of culture on simple, defined media. The capacity to synthesize all its essential biomolecules from media with known ingredients propelled *Neurospora* into the spotlight in the 1940s as an ideal model in which to study

the genetics of metabolic pathways that are similar in all cells. Of particular historical note was the use of *Neurospora* by George Beadle and Edward Tatum in studies that formed the basis for the “one gene, one polypeptide” hypothesis. This work ushered in a new era of molecular genetics and the use of microorganisms in genetic studies (see Figure 2-23).

### Life Cycle

The asexual life cycle of *Neurospora* starts with haploid spores released from microconidia and macroconidia (Figure A-3, top). On germination of the haploid spore, a tubular projection elongates by tip growth to form a mycelium, which contains branching hyphae. Growth occurs through the replication and division of haploid nuclei, unaccompanied by the formation of cross walls. Growth is rapid, up to 10 cm in a single day. On Petri plates, the compact network of filaments produces a colony. The network of hyphae in a colony is essentially one large, continuous single cell with many haploid nuclei. The haploid colony sporulates by forming conidia sacs, and one colony can produce spores that number in the millions.

Two different alleles of the mating type gene, *MATA* and *MATa*, control the *Neurospora* sexual cycle. Spores contain either the *MATA* or *MATa* allele, and contact between colonies of different mating types results in cell wall fusion followed by the fusion of haploid nuclei. The resulting diploid nuclei proceed quickly through meiotic cell divisions, each diploid nucleus yielding four haploid spores in an ascus (see Figure A-3, bottom). The haploid spores undergo one further mitotic division, forming eight spores arranged according to their cell of origin and stacked in a linear row within the long thin ascus. Numerous asci are held in a structure called a perithecium. Each spore in the perithecium is haploid and can form another *Neurospora* colony. Approximately 3 to 4 weeks elapse in proceeding from spore, to haploid hyphae, to diploid nuclei, and back to the production of spores.



**FIGURE A-3** The life cycle of *Neurospora crassa*.

[Sources: (Mycelium) From *Neurospora: Contributions of a Model Organism* by Roland Davis. © 2000 by Oxford University Press, Inc. Photo courtesy of Matthew L. Springer, University of California, San Francisco. (Germination) M. G. Roca et al., *Eukaryot. Cell* 4:911–919, 2005. (Single ascus) Courtesy of Namboori B. Raju, Stanford University. (Perithecium) Courtesy of Louise Glass.]

## Genetic Techniques

**Mutagenesis** *Neurospora* can be mutagenized by treating the spores with ionizing irradiation. Irradiated spores are germinated in complete media and allowed to sporulate, forming stocks of spores. Spores can then be screened for mutations.

**Introduction of DNA** *Neurospora* readily takes up exogenous plasmid DNA by transformation. The DNA must integrate into the genome to be stably inherited. The antibiotic hygromycin can be used to select for *Neurospora* with a transgene. Unlike in yeast, DNA insertion rarely occurs through homologous

recombination in *Neurospora*; insertion is more often random and untargeted.

**Gene Knockouts** There is an unusual way of generating a specific gene knockout in *Neurospora*. In a process apparently unique to this organism, a cell with a duplicate gene, when crossed, undergoes a process called RIP (repeat-induced point mutation). RIP occurs in the haploid nuclei of premitotic cells; the repeated DNA is searched out and destroyed by littering it with numerous G≡C-to-A=T transitions, in both copies of the duplicate gene. This process essentially results in a knockout strain.

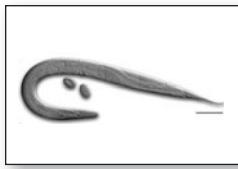
**Sporulation** As *Neurospora* chromosomes segregate during meiosis, the daughter cells become arranged linearly in a long axis. The frequency of crossing over can be used to map the distance of a mutant allele locus from the centromere.

## Neurospora as a Model Organism Today

**Meiotic Recombination** As we've seen, *Neurospora* packages its spores from a single meiotic event in an arrangement (in the ascus) that reflects the order of chromosomal segregation during meiosis. This feature makes *Neurospora* an attractive system for studies of crossover recombination during meiosis.

**Circadian Rhythms** *Neurospora* has a daily circadian rhythm of conidial spore formation. Conidial spores are conspicuous to the naked eye, so the phenotype of rhythmic behavior can be examined on Petri plates or in race tubes—long tubes in which the rapid growth of hyphae can be measured. Why this rhythm exists in *Neurospora* is not clear, but it serves as a convenient model for understanding the molecular basis of rhythmic behavior.

**Silencing Mechanisms** *Neurospora* contains at least three different silencing mechanisms that are thought to have evolved as genomic defenses against invading DNA, such as transposons and viruses. Meiotic silencing is a process in which any unpaired gene that is detected during meiosis is silenced. Quelling occurs in haploid cells and detects duplicate sequences. RIP detects duplicated sequences just before meiosis, resulting in G≡C-to-A=T mutations in both copies of a duplicated sequence. Meiotic silencing and RIP seem to be unique to *Neurospora*, whereas quelling seems to be related to RNA interference.



## Nematode, *Caenorhabditis elegans*

Small single-celled microbes such as *E. coli* and yeast are overwhelmingly successful models for studying the fundamental chemistry of life processes at the cellular level, but multicellular models are needed to investigate the complexities of development and the nervous system. In the 1960s, Sidney Brenner recognized that it was time to ask questions about the nervous system and how a multicellular organism develops from a single cell. Brenner decided on the nematode worm, a small, translucent metazoan, as a model for pursuing these new questions.

*Caenorhabditis elegans* is a member of the *Rhabditidae* family of nematodes, but unlike other family members, it is not pathogenic or parasitic to other animals. *C. elegans* is an appropriate choice as a model for animal development because it grows rapidly and, depending on temperature, can develop from an egg to a sexually mature adult in 2½ to 3½ days. Easy to grow on agar plates with *E. coli* as a food source, the worms can even be frozen for long-term storage. *C. elegans* hermaphrodites can self-fertilize, a property that allows quick recovery of homozygous mutants from a population of mutagenized worms. Males, formed at a lower frequency, can mate with hermaphrodites to create strains with new mutant combinations, an essential aspect of any genetic model organism. In addition, *C. elegans* is transparent, allowing every cell to be visualized during development. Although the hermaphrodite contains only 959 somatic cells, the nematode is highly complex, with a nervous system, muscles, and digestive and reproductive systems, and exhibits a variety of behaviors for neurological genetic studies. Each worm forms hundreds of progeny, allowing mutant screens for all types of anatomic and behavioral changes.

### Early Studies of *C. elegans* as a Model Organism

To lay a foundation for studies on how a multicellular organism develops, Brenner and others mapped the location of all 959 cells in the hermaphrodite worm by image reconstruction of tissue slices. During development, each migratory cell travels to a precise position in the animal. Further studies led researchers to the discovery that about 12% of cells consistently die during development and that cell death is genetically programmed. We now know that programmed cell death is important to the

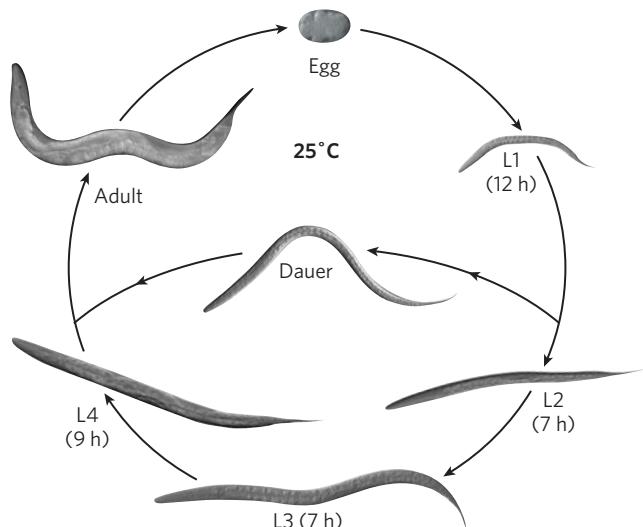
development of higher organisms, including humans, and that defects in this process can lead to cancer.

A prime example of *C. elegans* as a model for development is studies on the vulva. Development of this 22-cell organ has been studied intensively, as its individual cells are easily visualized in the microscope. Screens to identify defects in vulval development exploit the fact that fertilized eggs in the uterus must be deposited through the vulva to hatch outside the body. With developmental defects in the vulva, eggs cannot be deposited and hatch internally. The microscope reveals many worms inside the mother (sometimes referred to as a “bag of worms”). Using this obvious mutant phenotype, researchers have identified numerous genes that control vulval development, including genes in a conserved phosphorylation pathway that controls cell growth. Mammalian homologs of some of these genes encode tumor suppressors and oncogenes.

Programmed cell death in *C. elegans* development results in the death of one of the two progeny cells following cell division during various stages of development. This mechanism is essential at various stages of development for the proper developmental path of the surviving progeny cell. Many examples of programmed cell death are documented in *C. elegans*, several of which occur during development of the nervous system. Programmed cell death is also important in human embryonic development, such as in the removal of tissue between the fingers.

### Life Cycle

There are two sexes in *C. elegans*, hermaphrodite and male. Hermaphrodites contain two copies of the X chromosome. A low percentage of hermaphrodite germ cells (0.05% to 1.0%) undergo nondisjunction of the X chromosomes during meiosis, to produce so-called null-X gametes, and thus X0 (male) progeny. Hermaphrodites far outnumber males, and thus most progeny are produced by self-fertilization. Self-fertilization typically produces 200 to 300 fertilized eggs that follow a multistage life cycle (Figure A-4). After embryogenesis, which takes about 12 hours at 25°C, eggs hatch to produce larvae 80 µm long. The initial larva contains 558 cells. Development proceeds through four larval stages (L1 through L4) separated



**FIGURE A-4** The life cycle of *Caenorhabditis elegans*.

[Sources: (Egg) Center for Cell Dynamics. (L1-L4 and Dauer) Reprinted with permission from Wormatlas ([www.wormatlas.org](http://www.wormatlas.org)). (Adult) Courtesy of Ian Chin-Sang, Ph.D., Queen's University, Kingston, Ontario.]

by molts. The final molt results in the sexually competent adult worm. The overall process takes about 52 hours at 25°C. Adults live for approximately 15 days. Under stressful conditions, L2 larvae may enter a dormant stage, the dauer larva, which can persist for several months, waiting for conditions to improve.

### Genetic Techniques

**Mutagenesis** Worms can be mutagenized with chemical treatment or irradiation. Mutant worms can be screened for phenotypic alterations by direct observation with the microscope. Phenotypes include aberrant behavior or development, the inability of certain cells to undergo programmed cell death, altered life span, and inability of larvae to enter the dauer stage.

**Self-Fertilization and Cross-Fertilization** Self-fertilization produces about 300 progeny from one hermaphrodite and allows rapid recovery of recessive mutant alleles from individual mutant worms. Although males are produced only rarely, they provide opportunities to construct new genetic stocks by crossing with a hermaphrodite.

**Introduction of DNA** In transformation studies, linear recombinant DNA is injected directly into the gonad before the eggs are formed. Rare integration events lead to stable inheritance of transgene arrays. Integration rarely occurs by homologous recombination.

**Gene Knockouts** Disruption by a transposon typically destroys the function of a gene. In organisms with active transposons, mutations arise by random transposon insertion events. DNA integration into a gene of interest can be screened by PCR using a primer internal to the transposon and another primer that anneals to the gene under study.

**RNA Interference** RNA interference, originally discovered in the nematode, can be used to silence gene function by introducing double-stranded RNA homologous to the gene under study (see Chapter 22).

### *C. elegans* as a Model Organism Today

**Signaling Pathways** Programmed cell death and vulval development use signaling pathways that are apt models for human signaling pathways. Studies of *C. elegans* development continue to elucidate important features of these processes.

**Human Disease** *C. elegans* has many genes that are homologous to human disease genes, including those in the insulin-signaling pathway, as well as genes involved in heart, kidney, and neurological diseases. Study of these disease genes may illuminate the basis for human diseases.

**Aging** Genetic studies of the dauer larva have identified a set of genes that, when switched on in the adult, dramatically extend the worm's life span. The presence of homologs of these genes in other animals has obvious implications for the study of aging.

**RNA-Based Control of Gene Expression** Studies of RNA interference (first discovered in the nematode) have identified miRNAs involved in gene expression. Indeed, miRNAs are now known to be involved in gene regulation in both plants and animals.

**Neurodevelopment** The *C. elegans* nervous system is the animal's largest organ, comprising over one-third of its cells (302 neurons and 56 glial cells). Unlike the highly branched neuronal connections of vertebrates, the connectivity of neurons in *C. elegans* is relatively simple, with about 5,000 chemical synapses and 2,000 neuromuscular junctions. Behavioral abnormalities are easily observed and can be mapped with precision to particular neuronal networks. Knowledge of the complete neuronal connectivity enables researchers to study how axon growth is guided and how synapses form. In addition, *C. elegans* has several different classes of neurons, enabling genetic studies of neuronal differentiation.



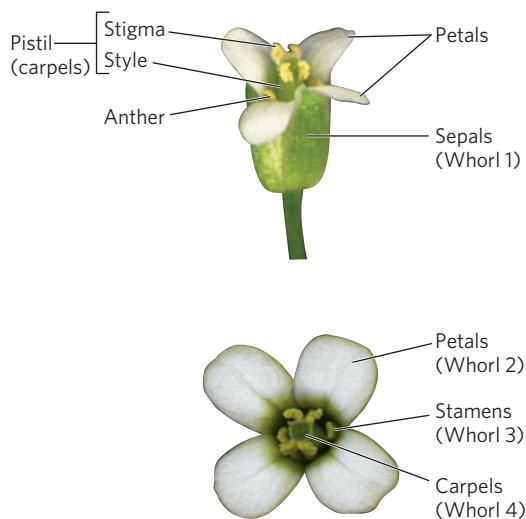
## Mustard Weed, *Arabidopsis thaliana*

*Arabidopsis thaliana* is an angiosperm, a dicot of the mustard family (*Brassicaceae*), which includes the more familiar broccoli, cabbage, and radish. *Arabidopsis* is generally regarded as a weed, but what it lacks in economic importance it makes up for in its special features as an experimental model. *Arabidopsis* is relatively small, allowing many plants to be grown in a confined space. It has a short life cycle, about 5 to 6 weeks from seed to flower, and each plant is capable of self-pollination, producing thousands of seeds for large genetic mapping studies. In addition, the *Arabidopsis* genome is less than 5% the size of the maize genome (2,500 Mbp) and less than 1% that of wheat (16,000 Mbp). Numerous labs worldwide study *Arabidopsis*, and several public stock centers maintain seed stocks of mutant lines and natural *Arabidopsis* variants called ecotypes, or accessions, as well as genomic resources.

*Arabidopsis* is a model for the physiology, development, genetics, and structure of all plants. It is also a model for how plants interact with the environment and sense day length, cold, drought, and salt, and how they respond to pathogens. We can expect many differences between plants and animals in the genetics of developmental programs. Nonetheless, studies on development of the flower whorl reveal a deep interconnection with animals. The flower whorl consists of four concentric rings (Figure A-5). The outer ring (whorl 1) consists of four sepals; whorl 2, four petals; whorl 3, six anthers; and the inner whorl, two carpels that fuse to form a pistil. Studies of plants with homeotic mutations reveal altered whorl identities. For example, in one mutant, carpels are found in whorl 1 in place of sepals. This genetic behavior is reminiscent of homeotic gene mutations in *Drosophila*, in which limbs protrude from incorrect body segments. Also as in *Drosophila*, *Arabidopsis* homeotic genes encode transcription factors that function by forming concentration gradients in the developing embryo (see Chapter 22).

### Early Studies of *Arabidopsis* as a Model Organism

*Arabidopsis* is a relatively recent addition to the select group of model organisms, only developed into a robust

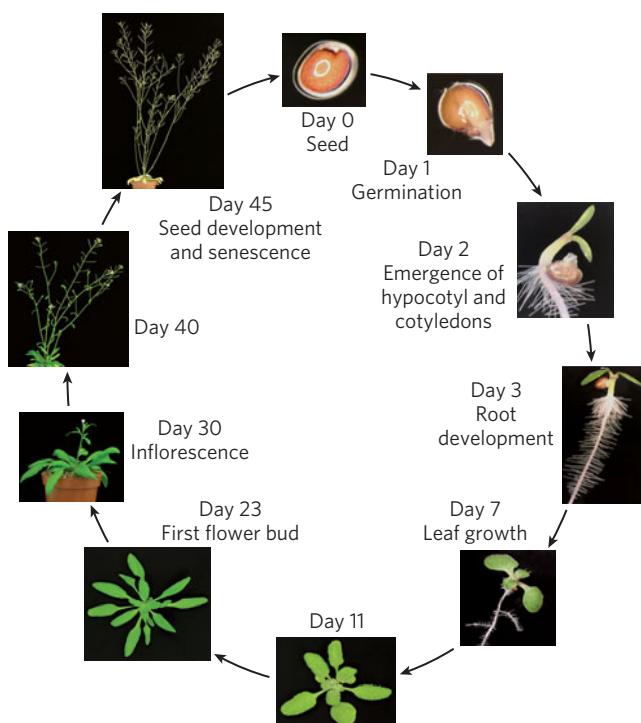


**FIGURE A-5** *Arabidopsis thaliana* flower whorl anatomy.  
[Source: Courtesy of Prof. Dr. Sabine Zachgo.]

model in the 1980s. In 1907, Friedrich Laibach identified the number of chromosomes in *Arabidopsis*, and in the 1940s he proposed the use of this plant as a model organism. With the help of Albert Kranz, Laibach collected and organized a large variety of ecotypes, which contributed to the current collection of 750 accessions of *Arabidopsis*. The beginnings of an *Arabidopsis* research community became evident with publication of a newsletter in the early 1960s, and the first International *Arabidopsis* Conference was held in 1965. By the 1980s, *Arabidopsis* was one of several plant models that also included the well-established genetic model, maize. With the development of T-DNA-mediated transformation (see below) in 1986, *Arabidopsis* rapidly became the predominant model for plant research.

### Life Cycle

*Arabidopsis* has a common plant life cycle (Figure A-6). As in many angiosperms, both male and female germ cells are present in the same flower, and self-fertilization (self-pollination) is readily accomplished. Plants can also cross-fertilize (cross-pollinate). Each plant produces



**FIGURE A-6** The life cycle of *Arabidopsis thaliana*.

[Source: C. Douglas et al., *Plant Cell* 13:1499–1510, 2001. Copyright © 2001, American Society of Plant Biologists.]

many flowers, which together can produce more than 10,000 seeds from a single plant. The life cycle from seed germination to a new crop of seeds takes 5 to 6 weeks, as roots, leaves, flowers, and finally seeds develop.

## Genetic Techniques

**Mutagenesis** Plants can be mutagenized by treating the seeds with ionizing radiation or chemical mutagens. Plants that are homozygous recessive for mutant genes are easily obtained by self-fertilization.

**Complementation** *Arabidopsis* in bloom can be self-fertilized or crossed with other plants by techniques similar to those used by Gregor Mendel, in which anthers are excised, thus permitting control over pollination and the genetic cross (see Chapter 2).

**Introduction of DNA** Transformation with recombinant DNA is mediated by *Agrobacterium tumefaciens* T-DNA (transfer DNA), which typically inserts into chromosomes randomly. *A. tumefaciens* is a gram-negative bacterium that causes tumors in plants. It contains a 200 kbp tumor-inducing (Ti) plasmid, which includes genes that facilitate replicative transfer of the DNA into a plant cell. Once inside the plant cell, the plasmid integrates a particular segment of its DNA, the T-DNA, into the plant

genome. When recombinant DNA is present in the T-DNA, this is also inserted into the genome.

**Gene Knockouts** Homologous recombination of transgenic DNA, producing a gene knockout, occurs only rarely in plants and is not typically used in plant studies. However, large collections of sequenced *Arabidopsis* mutants are available, and specific insertions in individual genes can be ordered from stock centers.

**RNA Interference** Gene function in *Arabidopsis* can be effectively knocked out using RNAi (see Chapter 22).

## Arabidopsis as a Model Organism Today

**Plant Evolution** The 750 or so different natural accessions of *Arabidopsis* vary considerably in their development and form (e.g., leaf shape, flowering time, hairiness, resistance to disease), and these are being studied to explain the evolution of traits and plant responses to the environment.

**Light Sensing** Plants have a variety of responses to light, and *Arabidopsis* is a genetic model for some of these. One such response is the switch from leaf to flower production. *Arabidopsis* flowers in response to an increase in day length and is a model for day-length sensing. Also, limiting light during seed germination results in stunted growth through a developmentally programmed switch that produces small plants with a limited root system. *Arabidopsis* is being used as a model to study the genetics of how light shapes this early development. Another light-sensing process under study in *Arabidopsis* is shade avoidance.

**Circadian Rhythms** As we might expect for organisms with an immobile lifestyle and a dependence on light, plants display strong circadian rhythms. *Arabidopsis* provides an excellent model for exploring the genetic basis of circadian rhythms and the nature of the mysterious “rhythmic oscillator.”

**Plant Resistance to Pathogens** Plants have a wide variety of strategies to survive stressful conditions, including invading pathogens, and *Arabidopsis* is a model for pathogen detection and defense. Among these defenses are antimicrobial molecules, development of physical barriers, and triggering of programmed cell death.

**Genetics of Plant Development** As a multicellular organism, *Arabidopsis* has a variety of organs and tissue types, each with its own genetics of development. Areas of developmental study include leaf growth, formation of flower whorls, seeds, and roots, development of vascular tissue, embryogenesis, and development of the flower body plan.



## Fruit Fly, *Drosophila melanogaster*

We are all familiar with the fruit fly as a cosmopolitan nuisance, but *Drosophila melanogaster* has a 100-year history as an important model organism for studying genetics and development. The fruit fly's body is divided into several segments that form three major sections: the head, thorax, and abdomen. The body sections are encased in a hard chitin cuticle, secreted by underlying epidermal cells, and is rich in anatomic details (indentations, hairs) that serve as phenotypic landmarks for genetic studies.

*Drosophila* is small (about 2.5 mm long); it is easily and inexpensively grown in the laboratory, in bottles containing a layer of cornmeal, molasses, and yeast. The fruit fly has a reasonably rapid generation time (12 days), is simple to cross by mating, and produces hundreds of progeny in each generation. The main inconvenience for researchers is its ability to fly away. Although *Drosophila* was originally used to study the basic mechanisms of transmission genetics, many new genetic tools can now scrutinize the developmental basis of embryogenesis and the body plan.

### Early Studies of *Drosophila* as a Model Organism

In 1908, Thomas Hunt Morgan looked for a suitable organism in which to study animal genetics. With little funding for science available at that time, he eventually settled on *Drosophila* because it was cheap and inexpensive to grow. In 1910, after two years of fruitless studies, Morgan found a male white-eyed spontaneous mutant (the flies normally have red eyes). Elegant and detailed studies of this single mutant revealed that genes are located on chromosomes, that each gene has two alleles, that alleles assort independently during meiosis, and that genes located on different chromosomes assort independently. These and other important findings supported Mendel's laws in the physical context of genes located on chromosomes (see Chapter 2).

The fruit fly is also the first organism for which a genetic map was constructed. Alfred Sturtevant, a student in Morgan's laboratory, mapped the relative distance between genes along chromosomes, based on their frequencies of crossover recombination. Calvin B. Bridges carried the

studies further by identifying the exact positions of genes within polytene chromosomes. These giant chromosomes, located in the fruit fly's salivary glands, consist of bundles of chromosomes packed together and have unique banding patterns when stained (Figure A-7). Bridges identified more than 5,000 bands arranged in a distinct pattern. We don't know why polytene chromosomes form, but it may be related to the job of the salivary gland in excreting the pupal casing for metamorphosis. Numerous recombination-based mutations were traced to missing bands, or to spots where chromosomes had inverted, enabling the mapping of genes along a chromosome, as well as solidifying the location of genes on chromosomes.

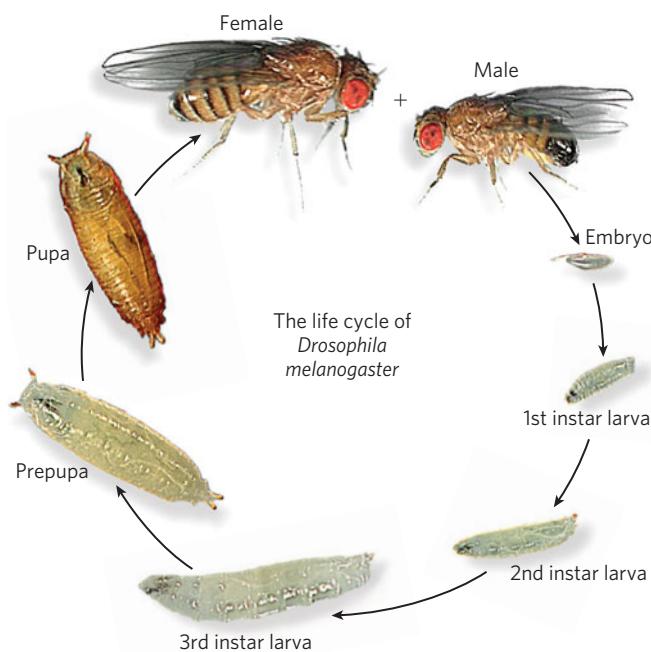
### Life Cycle

Flies have a short (12 day) diploid life cycle (Figure A-8). Sex in flies is determined by X chromosome copy number, not by the Y chromosome (XX is female, XY and the rare X0 are male), although the Y chromosome is required for the production of sperm. About a day after mating, the female begins to lay hundreds of eggs. Nuclear divisions in the embryo are the most rapid of any multicellular organism, and eggs hatch in about one day. The maggot proceeds through three larval instar stages, separated by molts, that take about 5 days. Larvae contain imaginal disks of tissue that are destined to become each of the appendages of the adult fly (e.g., eyes, antennae, legs, wings, halteres (modified wings acting as flight stabilizers), and mouthparts). The salivary glands of the



**FIGURE A-7** An insect polytene chromosome.

[Source: S. F. Werle et al., *Can. J. Zool.* 82:118–129, 2004, Fig. 2.  
© 2004, NRC Canada.]



**FIGURE A-8** The life cycle of *Drosophila melanogaster*.

[Source: Prof. Dr. Christian Klambt, Westfälische Wilhelms-Universität Münster, Institut für Neuro- und Verhaltensbiologie.]

third instar larva secrete a pupal case, within which metamorphosis occurs in 3½ to 4½ days to yield the adult fly. Flies live approximately 30 days.

### Genetic Techniques

**Mutagenesis** Flies can be mutagenized by exposure to ionizing radiation or by feeding with mutagenic chemicals. Mutations that result in changes in eye color, wing shape, or body parts can be identified by visual screening.

**Introduction of DNA** A process known as P-element transformation is used to insert recombinant DNA, typically to study the effects of protein expression on transcriptional control elements. The P-element is a 3 kbp transposon that can carry a section of recombinant DNA between its terminal repeats in place of the self-encoded transposase and repressor (see Chapter 14). The recombinant P-element DNA is injected into the fertilized egg, along with a transposase-encoding plasmid. Insertion is random, and the plasmid encoding the transposase gene is lost during cell divisions, thereby preventing reinsertion of the transposon elsewhere in the genome.

**Balancer Chromosomes** Recessive lethal alleles can be stably maintained in fruit flies when paired with a balancer chromosome—a chromosome that cannot recombine with its homologous pair, due to the presence of many internal inversions that prevent complete align-

ment during meiosis. Balancer chromosomes are homozygous lethal, so the only progeny that survive are heterozygous for the recessive lethal gene and the balancer.

**Genetic Mosaics** Genetic mosaics, patches of genetically altered tissue, can be formed in the adult fruit fly by x-ray irradiation to induce mitotic recombination. Mosaics are particularly useful in the study of lethal genes. Genetic mosaics of lethal genes can also be formed in the adult through the use of a heat-inducible yeast recombinase (called FLP) engineered into the genome. Heat induction results in a high frequency of mitotic recombination in the heated tissues to form genetic mosaics. Although the genetic alterations may be lethal to embryos, they do not necessarily kill the adult, given the localized expression in only some cells.

**Gene Knockouts** Gene knockouts are difficult to obtain in *Drosophila* because techniques for homologous recombination are not yet perfected. There are, however, two-step procedures in which a gene is inserted randomly, followed by expression of proteins that excise the gene. Once excised, the linear DNA fragment can undergo homologous recombination with the wild-type gene.

**RNA Interference** RNAi can be used to effectively knock down specific gene products in lieu of a true gene knockout.

### *Drosophila* as a Model Organism Today

**Human Disease** The ~170 Mbp *Drosophila* genome is about one-twentieth the size of the mouse and human genomes, yet it encodes nearly the same number of gene families. About 60% of the genes known to be involved in human disease have homologs in the fruit fly. For example, studies of embryonic lethal genes in *Drosophila* have helped explain the genetic basis of human birth defects. Other disease models include immunological disorders, diabetes, cancer, Huntington disease, Alzheimer disease, and Parkinson disease.

**Body Plan Development** Localization of maternal mRNA in eggs results in localized gene expression that sets up the anterior-posterior and dorsal-ventral axes in *Drosophila*. These genes control the body plan and have counterparts in more complex animals. Thus, the fly serves as a relatively simple system to understand body plan formation (see Figure 16-24; see also Chapter 21 and Chapter 22).

**Behavior** *Drosophila* provides a model for understanding the cellular and molecular basis for certain types of behavior. Fruit fly behavioral abnormalities include changes in learning and memory, foraging behavior, resting behavior, and other behaviors, and alcoholism.



## House Mouse, *Mus musculus*

The house mouse, *Mus musculus*, has been collected and bred by mouse “fanciers” for hundreds of years, and some of the purebreds developed by fanciers are used today as standards in scientific studies. The house mouse is the leading mammalian model, and it is more like a human than we may care to admit (see Figure 1-12). The mouse genome is almost the size of the human genome and encodes essentially the same number of genes, 99% of which have homologs in the human. In fact, much of the mouse’s genome is syntenic with ours, meaning that whole blocks of genes occur in the same order in both species (see Figure 8-4).

Compared with other model organisms, mice are more cumbersome to work with in every way. They are larger, of course, but they also have a generation time of about 8 to 10 weeks and produce, on average, only 6 to 8 pups per litter. These statistics are attractive when compared with other mammals, but pale in comparison with other model organisms. Colonies of mice are also costly to maintain, and they simply cannot be dealt with in the numbers needed to perform large genetic screens, as with other model organisms. However, unlike the nematode and fruit fly, mice have biological systems that have no parallel in lower animal models, such as the immune and skeletal systems, or are simply better models for studies of complex systems such as the cardiovascular system, endocrine system, and many others. The mouse is a model for human disease, including cancer, in virtually all of these systems.

### Early Studies of the Mouse as a Model Organism

Genetic studies in mice began in the early 1900s, when selection and breeding were the main methods of obtaining progeny with the desired traits. These early studies produced a general model that explained coat coloring in all other fur-bearing mammals. Mice and rats also have a long history in nutritional studies, especially in the identification of vitamins and the symptoms caused by vitamin deficiencies in the diet. Use of mice as a human disease model was pioneered by Clarence Cook Little. In the 1920s he developed an inbred mouse strain, C57BL/6 (commonly known as Black6),

which eventually became the mouse strain used to determine the genome sequence. Little also founded the Jackson Laboratory in Bar Harbor, Maine, a center for mouse genetics that also serves as a public repository of mouse models for scientific research.

### Life Cycle

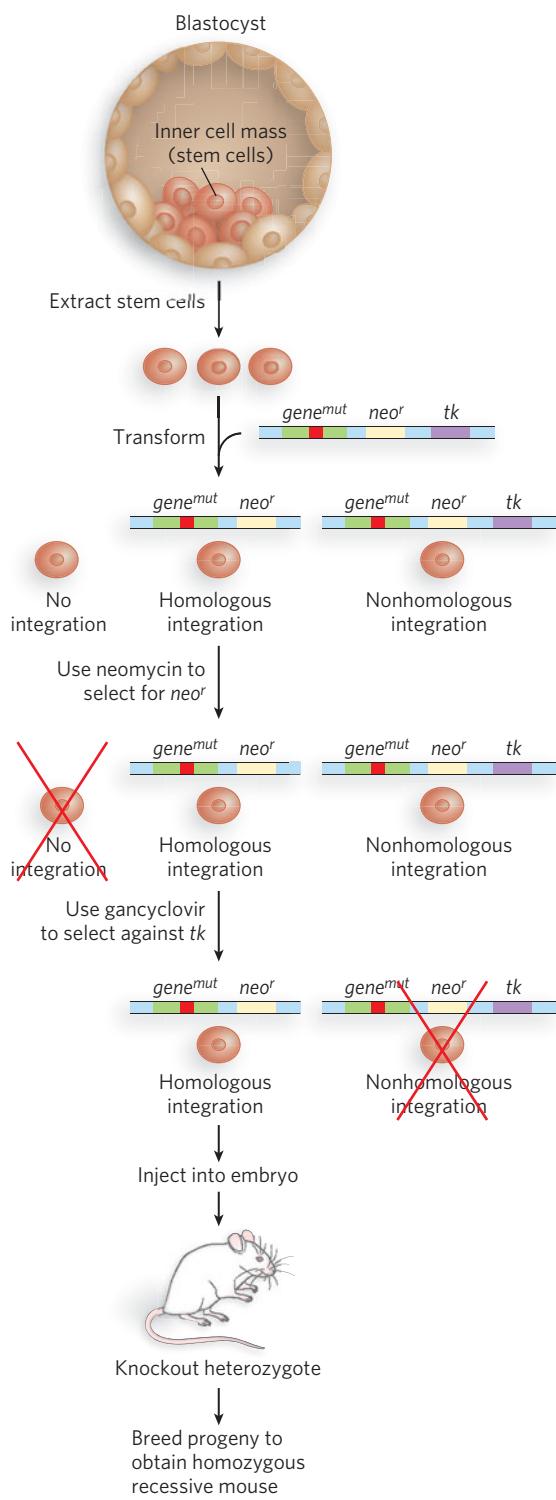
The X and Y chromosomes determine sex in mice, as in humans. Fertilization gives rise to a blastocyst containing some undifferentiated cells bunched together in the inner cell mass, the source of embryonic stem cells. Gestation is complete within 19 to 21 days and gives rise to a litter of 3 to 14 pups. Sexual maturity requires about 6 weeks for females and 8 weeks for males, but breeding can take place in as short a time as 35 days. Mice live for 1 to 2 years, and females can produce about 5 to 10 litters, mainly in their first year of life.

### Genetic Techniques

**Mutagenesis** Inbreeding over many generations has produced many useful strains of mutant mice. Adding mutagenic chemicals to the food supply also facilitates development of mutant strains.

**Introduction of DNA** Foreign DNA can be injected directly into the nucleus of fertilized eggs, followed by implantation of the eggs in the oviduct of the female recipient. Random integration occurs with high frequency. The recombinant DNA used typically has a mouse promoter that directs expression of a reporter gene, such as *lacZ* or GFP (green fluorescent protein), so that expression of the transgene can be followed during development. About half of the transgenic mice contain recombinant DNA in the germ line and therefore pass on the recombinant gene to future generations.

**Gene Knockouts** Targeted knockouts for mouse disease models are constructed in embryonic stem cells (Figure A-9). The stem cells are extracted from the inner cell mass of the blastocyst and grown in culture. Cultured stem cells are then transformed with linear DNA containing a mutated copy of the gene under study, along with genes for neomycin resistance (*neo*<sup>r</sup>)



**FIGURE A-9** The construction of a knockout mouse.

and thymidine kinase (*tk*). Homologous DNA flanks the mutated gene (*gene<sup>mut</sup>* in Figure A-9) and the *neo<sup>r</sup>* gene, such that homologous integration replaces the wild-type gene with the mutant gene plus the *neo<sup>r</sup>* gene. On the other hand, DNA that inserts randomly

results in integration of the entire DNA fragment, including *tk*.

To select for cells with the desired gene knockout, two steps are required. Selection for *neo<sup>r</sup>*, using neomycin, kills all cells that fail to integrate transformed DNA. Selection against *tk*, using the antiviral gancyclovir, kills cells with DNA integrated by nonhomologous recombination, because these cells contain *tk*, and thymidine kinase converts gancyclovir to a toxin that kills these cells. Only cells that contain gene knockouts produced by homologous recombination survive both selections. Engineered stem cells are injected into a host wild-type blastocyst-stage embryo. This results in formation of an embryo that is a chimera of host wild-type and donor engineered cells. Resulting chimeras are bred for germ-line transmission of the genetic modification. The F<sub>1</sub> (first-generation) heterozygous mice are then crossed to obtain F<sub>2</sub> wild-type, heterozygous, and homozygous offspring, in the expected Mendelian 1:2:1 ratio. Selective breeding results in homozygous knockout mice. Methods have been worked out to preserve valuable, and hard to obtain, mouse strains by cryopreservation of sperm or egg cells.

## The Mouse as a Model Organism Today

**Human Disease** The mouse is an important model for studying human diseases. Disease models can be derived by inbreeding or by producing knockouts of known disease genes. Mouse models of human disease include cancer, atherosclerosis, aging, hypertension, metabolic diseases, immune disorders, type 1 and type 2 diabetes, Down syndrome, Alzheimer disease, glaucoma, osteoporosis, obesity, epilepsy, Lou Gehrig disease (amyotrophic lateral sclerosis), Huntington disease, blood disorders, and many others.

**Mapping of Mutant Genes** Mutant genes can be identified more quickly in the mouse than can mutant genes in humans. Closely related strains of mice (e.g., *Mus musculus* and *Mus spretus*) can be crossed to produce hybrids that usually contain different sequences at polymorphic positions, enabling researchers to use linkage analysis to develop detailed genetic maps and locate a disease gene.

**Behavior** Mouse models exist for certain types of behavior, including alcoholism, drug addiction, anxiety disorders, and aggressive behavior.

**Mammalian Development** Transgenic mice are used to study the location and timing of expression of particular genes at various stages of development. In addition, models exist for studying certain human developmental disorders, including cleft lip and cleft palate.

*This page intentionally left blank*

# Glossary

---

**A site:** The site in a ribosome where the aminoacyl-tRNA binds.

**AAA+ proteins:** A family of proteins with ATPase activity that share a common structural domain called the AAA domain, which includes Walker A and Walker B motifs. AAA stands for ATPases associated with diverse cellular activities.

**abasic site:** A position in an intact DNA backbone that is missing the base. Also called an AP (apurinic or apyrimidinic) site.

**abortive initiation:** Release of an 8 to 10 base-pair RNA transcript from the bacterial RNA polymerase initiation complex before it clears the promoter and enters the elongation stage.

**accommodation:** A process of checking for appropriate codon-anticodon pairing prior to rotation of an incoming aminoacyl-tRNA into position for peptidyl transfer.

**achiral:** Refers to a molecule that can be superimposed on its mirror image.

**acid dissociation constant:** The dissociation constant of an acid, HA, describing its dissociation into its conjugate base,  $\text{A}^-$ , and a proton.  $K_a = [\text{A}^-][\text{H}_3\text{O}^+]/[\text{HA}]$ .

**acridine:** A planar heterocyclic molecule from coal tar that intercalates between successive G≡C base pairs, deforming the DNA. Acridine inhibits transcription by preventing movement of the RNA polymerase along the DNA template. Acridine is also a mutagen, causing DNA polymerase to insert an extra base during replication.

**actinomycin D:** A peptide antibiotic that inhibits transcription elongation by RNA polymerase in bacteria, eukaryotes, and cell extracts. The molecule has a planar heterocyclic region that intercalates between successive G≡C base pairs, deforming the DNA and preventing movement of the RNA polymerase along the template.

**activation energy ( $\Delta G^\ddagger$ ):** The difference in energy between the ground state of a reacting substance and the transition state.

**activation:** The positive regulation of the expression of a gene or genes.

**activator:** (1) A DNA-binding protein that positively regulates the expression of one or more genes; that is, transcription rates increase when an activator is bound to the DNA. (2) A positive modulator of an allosteric enzyme.

**active site:** The region of an enzyme surface that binds the substrate molecule and catalytically transforms it; also known as the catalytic site.

**ADAR:** See adenosine deaminase acting on RNA.

**adenine (A):** A purine base that is a component of DNA and RNA.

**adenosine 3',5'-cyclic monophosphate:** See cyclic AMP.

**adenosine deaminase acting on RNA (ADAR):** An enzyme that catalyzes the conversion of adenosine to inosine by removal of the amino group at C-6 on the adenine ring.

**adenylation step:** The first (activation) step in the attachment of an amino acid to a tRNA. The aminoacyl-tRNA synthetase reacts with the  $\alpha$ -phosphoryl group of ATP, displacing pyrophosphate and forming a 5'-aminoacyl adenylate intermediate. *Also see* tRNA-charging step.

**A-DNA (A-form DNA):** Conformation of double-stranded DNA observed under certain nonaqueous solvent conditions. The molecule assumes a right-handed helix with 11 base pairs per turn and a rise of 2.6 Å per base pair. *Compare* B-DNA and Z-DNA.

**adult stem cells:** The cells in adult mammals that retain the ability to divide and differentiate into other cell types. *Compare* embryonic stem cells.

**affinity chromatography:** A type of column chromatography in which molecules are separated based on their binding affinity for chemical groups present on the stationary phase.

**A-form DNA:** See A-DNA.

**alkylation:** The transfer of an alkyl group from one molecule to another.

**allele:** A variant form of a gene at a specific locus.

**allopatric speciation:** Geographic isolation of a group of individuals followed by evolution to form a distinct species that no longer can interbreed with the original one.

**allosteric enzyme:** A regulatory enzyme with catalytic activity modulated by the noncovalent binding of a specific metabolite at a site other than the active site.

**allosteric modulator:** A metabolite that, when bound to the allosteric site of an enzyme, alters its kinetic characteristics.

**allosteric protein:** A protein (generally with multiple subunits) with multiple ligand-binding sites, such that ligand binding at one site affects ligand binding at another.

**$\alpha$ -amanitin:** A cyclic polypeptide antibiotic that inhibits transcription in eukaryotic cells by binding Pol II and blocking its ability to translocate along the DNA template. At high concentrations it also binds and inhibits Pol III.

**$\alpha/\beta$  barrel:** Common protein domain architecture consisting of eight hydrogen-bonded  $\beta$  strands surrounded by eight  $\alpha$  helices. The domain is formed by a series of  $\beta$ - $\alpha$ - $\beta$  motifs.

**$\alpha$  carbon ( $\text{C}_\alpha$ ):** The first carbon atom attached to a functional group. In amino acids, the  $\alpha$  carbon is the central

## G-2 Glossary

carbon to which the amino, carboxyl, and R groups are bound.

**$\alpha$  helix:** A helical conformation of a polypeptide chain, usually right-handed, with maximal intrachain hydrogen bonding; one of the most common secondary structures in proteins.

**alternative splicing:** The splicing of exons from a single gene in various combinations to produce different mRNAs and thus different polypeptides.

**Ames test:** A simple bacterial test for carcinogenicity, based on the assumption that carcinogens are mutagens.

**amino acids:**  $\alpha$ -Amino-substituted carboxylic acids, the building blocks of proteins.

**amino terminus (N-terminus):** The end of a polypeptide chain with a free  $\alpha$ -amino group.

**aminoacyl-tRNA synthetases:** Enzymes that catalyze synthesis of an aminoacyl-tRNA at the expense of ATP energy.

**aminoacyl-tRNA:** An aminoacyl ester of a tRNA; the tRNA is charged with an amino acid.

**amphipathic helix:** An  $\alpha$  helix with both polar and nonpolar domains.

**analyte:** A molecule to be analyzed by mass spectrometry.

**anaphase:** The third stage of mitosis (M phase). Sister chromatid pairs held together at the centromere now separate and the two homologous chromosomes move towards opposite spindle poles.

**annealing:** Process in which single strands of nucleic acid in solution spontaneously rewind or renature with strands of complementary base sequence to form duplex structures.

**anticodon:** A specific sequence of three nucleotides in a tRNA, complementary to a codon for an amino acid in an mRNA.

**antiparallel  $\beta$  sheet:** See  $\beta$  sheet.

**antiparallel:** Describes two linear polymers that are opposite in polarity or orientation.

**AP endonucleases:** Enzymes that cleave the DNA backbone at an AP (apurinic or apyrimidinic; abasic) site as part of the base excision repair pathway.

**apoenzyme:** The protein portion of an enzyme, exclusive of any organic or inorganic cofactors or prosthetic groups that might be required for catalytic activity.

**apoprotein:** The protein portion of a protein, exclusive of any organic or inorganic cofactors or prosthetic groups that might be required for activity.

**aqueous solution:** Solution in which the solvent is water.

**archaea:** One of the three main groups of living organisms. Like bacteria, archaea are unicellular and contain no internal organelles or nucleus; however, archaea are more closely related to eukaryotes with respect to some genes and metabolic pathways. Archaea include many species that thrive in extreme environments of high ionic strength, high temperature, or low pH.

**ARE:** See AU-rich element.

**association constant ( $K_a$ ):** An equilibrium constant for the association of a complex of two or more biomolecules from its components, for example, association of a substrate with an enzyme.  $K_a$  is the reciprocal of the dissociation constant,  $K_d$ . Compare dissociation constant ( $K_d$ ).

**atomic orbital:** Mathematical function that describes the behavior of an electron in an atom.

**ATP-coupling stoichiometry:** A property of helicases and other motor proteins that describes the number of ATP molecules consumed per distance traveled or other defined work units.

**AU-rich element (ARE):** Sequences in mRNA with 5 to 13 residues of A and U, which target the mRNA for rapid degradation.

**autoinhibition:** The reduction or elimination of a molecule's activity by one of its own segments or domains.

**autosome:** Any chromosome that is not a sex chromosome. Compare sex chromosome.

**auxotrophic mutant (auxotroph):** A mutant organism defective in the synthesis of a particular biomolecule, which must therefore be supplied for the organism's growth.

**BAC:** See bacterial artificial chromosome.

**bacmid:** A large circular DNA that includes the entire baculovirus genome and sequences that allow replication of the bacmid in *Escherichia coli*; a baculovirus shuttle vector.

**bacteria:** One of the three main groups of living organisms; bacteria have a plasma membrane but no internal organelles or nucleus.

**bacterial artificial chromosome (BAC):** A plasmid designed as a cloning vector for large segments of DNA. A BAC typically includes cloning sites, one or more selectable markers, and a stable origin of replication.

**bacterial transduction:** The transfer of genetic information from one bacterial cell to another by means of a viral vector.

**Barr body:** In female cells of mammals, the inactivated X chromosome, which is compacted into a dense chromatin particle.

**basal transcription factor:** In eukaryotic cells, a protein required at every Pol II promoter. Also called a general transcription factor.

**base excision repair (BER):** A DNA repair pathway that involves excision of a damaged base by DNA glycosylase, followed by cleavage of the DNA backbone adjacent to the site by an AP endonuclease. Nick translation, DNA polymerization, and ligation complete the repair.

**base pair:** Two nucleotides in nucleic acid chains that are paired by hydrogen bonding of their bases; for example, A with T or U, and G with C.

**base stacking:** See hydrophobic stacking.

**basic helix-loop-helix:** A protein secondary structure motif typical of transcription activators. It consists of two amphipathic  $\alpha$  helices joined by a loop of variable length. Two such motifs dimerize through one pair of  $\alpha$  helices. The other  $\alpha$  helices have a series of basic amino acid residues along one side through which they bind DNA.

**basic leucine zipper:** A leucine zipper motif in which one side of the recognition helix has a series of basic residues, which facilitates DNA binding.

**B-DNA (B-form DNA):** Standard Watson-Crick conformation of double-stranded DNA. The molecule assumes a right-handed helix with 10.5 nucleotide residues per turn and a rise of 3.4 Å per base pair. *Compare* A-DNA and Z-DNA.

**BER:** *See* base excision repair.

**$\beta$ - $\alpha$ - $\beta$  motif:** A protein secondary structure in which two parallel  $\beta$  strands are connected by an  $\alpha$  helix.

**$\beta$  barrel:** A protein structural domain in which a  $\beta$  sheet of eight or more strands with one hydrophobic surface forms a cylinder in which the first  $\beta$  strand hydrogen-bonds with the last  $\beta$  strand.

**$\beta$  hairpin:** A protein structural motif in which two antiparallel  $\beta$  strands are connected, usually by a  $\beta$  or  $\gamma$  turn.

**$\beta$  sheet:** A common protein secondary structure in which a polypeptide chain assumes an extended, zigzag arrangement with extensive hydrogen bonding between adjacent segments or strands. In parallel  $\beta$  sheets, the strands are aligned with the same polarity. In antiparallel  $\beta$  sheets, adjacent strands have opposite polarity.

**$\beta$  sliding clamp:** A component of the *E. coli* DNA polymerase III holoenzyme. The ring-shaped homodimer encircles and slides along the duplex DNA ahead of the Pol III core to which it is attached, greatly enhancing the processivity of DNA synthesis.

**$\beta$  turn:** A type of protein secondary structure consisting of four amino acid residues arranged in a tight turn so that the polypeptide turns back on itself.

**B-form DNA:** *See* B-DNA.

**binding energy ( $\Delta G_B$ ):** The energy derived from noncovalent interactions between enzyme and substrate or receptor and ligand.

**binding site:** The crevice or pocket on a protein in which a ligand binds.

**biochemical standard free-energy change ( $\Delta G^\circ$ ):** The free-energy change for a reaction occurring under a set of standard conditions: temperature, 298 K; partial pressure of each gas, 1 atm or 101.3 kPa; all solutes at 1 M concentration, pH 7.0, in 55.5 M water.

**blunt ends:** The product of restriction endonuclease action on double-stranded DNA that leaves no unpaired bases at the cleavage site.

**bond angle:** The angle between two adjacent bonds to the same atom.

**Bragg's law:** Mathematical expression that relates the wavelength ( $\lambda$ ) and angle ( $\theta$ ) of incident radiation to the distance ( $d$ ) between reflecting planes in a crystal lattice;  $\lambda = 2d \sin \theta$ .

**branch migration:** Movement of the branch point in a branched DNA formed from two DNA molecules with identical sequences. *Also see* Holliday intermediate.

**branch point:** An internal A residue just upstream of the 3' splice site of an intron that attacks the phosphate at the 5' splice site, forming the loop of the intron lariat.

**BRCA2:** A vertebrate recombination mediator protein often mutated in breast cancer.

**bromodomain:** Protein structural domain that recognizes and binds to certain acetylated Lys residues in proteins.

**buffer capacity:** Quantitative measure of the ability of a buffer solution to resist changes in pH.

**buffer solution:** A system capable of resisting changes in pH, consisting of a conjugate acid-base pair in which the ratio of proton acceptor to proton donor is near unity.

**cAMP receptor protein (CRP):** In bacteria, a specific regulatory protein that controls initiation of transcription of the genes that produce the enzymes required for the cell to use some other nutrient when glucose is lacking; also called catabolite gene activator protein (CAP).

**cAMP:** *See* cyclic AMP.

**CAP:** *See* cAMP receptor protein.

**cap-binding complex (CBC):** A protein complex that recruits capped mRNAs to the ribosome to initiate translation.

**carboxyl terminus (C-terminus):** The end of a polypeptide chain with a free  $\alpha$ -carboxyl group.

**carcinogen:** A substance directly involved in causing cancer.

**catabolite repression:** The inhibition of the expression of genes required for the metabolism of other sugars in the presence of glucose.

**catalysis:** An increase in the rate of a chemical reaction by a substance that is not consumed by the reaction.

**catalyst:** a substance that increases the rate of a chemical reaction without being consumed by the reaction.

**catalytic RNA:** *See* ribozyme.

**catenane:** Two or more circular polymeric molecules interlinked by one or more noncovalent topological links, resembling the links of a chain.

**CBC:** *See* cap-binding complex.

**cDNA library:** DNA library consisting entirely of cloned cDNAs from a particular organism or cell type.

**cDNA:** *See* complementary DNA.

## G-4 Glossary

**cell cycle:** The process by which cells replicate and divide. The bacterial cell cycle involves binary fission; the eukaryotic cell cycle has four phases, including mitosis.

**cell theory:** The theory proposed by Theodor Schwann in 1839 that cells are the basic units of all living things.

**cell:** Membrane-bounded structure that is the smallest unit of life.

**cellular function (of a gene product):** The metabolic processes in which a gene product participates and the interactions of that gene product with other proteins or RNAs in the cell. *Compare* molecular function and phenotypic function.

**central dogma:** The organizing principle of molecular biology: genetic information flows from DNA to RNA to protein. The pathways of information flow are now established and no longer constituted dogma.

**centromere:** A specialized site in a chromosome, serving as the attachment point for the mitotic or meiotic spindle.

**centrosome:** Organelle that serves as the microtubule organizing center, responsible for the creation of the spindle apparatus that moves chromosomes to opposite poles of the cell during mitosis and meiosis.

**cGMP:** *See* cyclic GMP.

**chain topology diagram:** Method of illustrating in two dimensions the topology of the polypeptide chain in supersecondary structures.

**change in free energy ( $\Delta G$ ):** *See* free-energy change ( $\Delta G$ ).

**chaperone:** Any protein that interacts with partially folded or improperly folded polypeptides, facilitating the correct folding pathway or providing a microenvironment where proper folding can occur.

**chaperonin:** Class of chaperones that forms a large, barrel-like structure, inside which certain cellular proteins fold.

**Chargaff's rules:** A set of quantitative observations about DNA from many organisms and species that helped to lay the groundwork for the discovery of the structure of DNA.

**chemical bond:** An attractive force that holds atoms to each other in a molecule or crystal.

**chemical reaction:** A process that changes the structure or energy content of atoms in a molecule, but not their nuclei.

**chemical shift:** Variations of nuclear magnetic resonance frequencies relative to a standard of the same kind of nucleus caused by variations in the electron distribution within a molecule.

**chemoheterotroph:** An organism that obtains energy by metabolizing organic compounds derived from other organisms.

**chi:** The sequence 5'-GCTGGTGG-3', which alters the endonuclease activity of bound RecC in the RecBCD complex so that it preferentially degrades the 5' end of the molecule.

**chiasma:** A cross-shaped junction that represents physical recombination between chromosomes. *Plural* chiasmata.

**ChIP-Chip:** Chromatin immunoprecipitation followed by hybridization of the precipitated DNA to a genomic microarray (chip).

**ChIP-Seq:** Chromatin immunoprecipitation followed by DNA sequencing.

**chiral center:** An atom with substituents arranged so that the molecule is not superimposable on its mirror image.

**chiral:** Describes a compound that contains an asymmetric center (chiral atom or chiral center) and thus can occur in two nonsuperimposable mirror-image forms (enantiomers).

**chloramphenicol:** An antibiotic that inhibits protein synthesis by bacterial, mitochondrial, and chloroplast ribosomes by blocking peptidyl transfer.

**chromatin remodeling complex:** Protein complex with ATPase activity that translocates nucleosomes along the DNA, making certain regions of DNA more or less accessible to transcription factors.

**chromatin:** A filamentous complex of DNA, histones, and other proteins, constituting the eukaryotic chromosome.

**chromatography:** A process in which complex mixtures of molecules are separated by many repeated partitionings between a flowing (mobile) phase and a stationary phase. The stationary phase may be packed into a tube (column chromatography), or planar (thin layer chromatography).

**chromatosome:** A structure derived from nuclease-treated chromatin that consists of the five histone proteins and a segment of about 168 bp of DNA protected from digestion by its association with histones.

**chromodomain:** Protein structural motif that recognizes and binds certain methylated Lys residues in proteins.

**chromosomal scaffold:** Proteinaceous residue after extraction of histones from chromosomes, comprised mainly of SMC proteins.

**chromosome theory of inheritance:** The hypothesis proposed by Walter Sutton in 1903 that genes are located on chromosomes.

**chromosome:** A single large DNA molecule and its associated proteins, containing many genes; stores and transmits genetic information.

**clamp-loading complex:** The portion of the *E. coli* DNA polymerase III holoenzyme that assembles the  $\beta$  sliding clamps onto the DNA.

**clone:** Identical copy.

**cloning vector:** A DNA molecule known to replicate autonomously in a host cell, to which a segment of DNA may be spliced to allow its replication; for example, a plasmid or an artificial chromosome.

**cloning:** The production of large numbers of identical DNA molecules, cells, or organisms from a single ancestral DNA molecule, cell, or organism.

**closed complex:** A complex of the RNA polymerase bound to a promoter, in which the DNA is intact and double-stranded. *Compare* open complex.

**closed form:** The conformation assumed by *E. coli* DNA polymerase I when a primed template and the correct dNTP are both bound to the active site.

**closed-circular DNA:** A continuous double-stranded DNA molecule with no free 3' or 5' ends.

**CMG complex:** A complex of the proteins Cdc24, the MCM helicase, and GINS proposed to function in the eukaryotic replisome.

**coactivator:** A protein that stimulates transcription by binding both the RNA polymerase and an activator or activators, without binding the DNA directly. *Compare* corepressor and DNA-binding transcription activator.

**coalescent theory:** Retrospective analysis of population genetics data (mutation rates, selection, genetic drift, and other factors) to trace a polymorphism back to the original ancestor in which it appeared.

**coding strand:** The strand of a double-stranded DNA that has the same sequence as the RNA transcript (with T in place of U). The coding strand is complementary to the template strand and is also called the nontemplate strand.

**codominance:** Non-Mendelian behavior in which two alleles of a gene produce distinct functional products, neither of which is dominant to the other. *Compare* incomplete dominance.

**codon bias:** The use of certain codons more than others to code for a given amino acid.

**codon family:** Four codons that specify the same amino acid.

**codon:** A sequence of three adjacent nucleotides in a nucleic acid that codes for a specific amino acid.

**coenzyme:** An organic cofactor required for the action of certain enzymes; often has a vitamin component.

**cofactor:** An inorganic ion or a coenzyme required for enzyme activity.

**cohesins:** SMC proteins that link sister chromatids immediately after chromosomal replication and keep them together as the chromosomes condense to metaphase.

**coiled-coil:** Protein motif in which  $\alpha$  helices twist around each other in a left-handed supercoil, interacting through hydrophobic contacts.

**cointegrate:** An intermediate in the migration of certain DNA transposons in which the donor DNA and target DNA are covalently attached.

**collision release:** Release of the *E. coli* DNA polymerase III from the  $\beta$  sliding clamp when it collides with an Okazaki fragment on the lagging strand. *Also see* signaling release.

**column chromatography:** A process in which complex mixtures of molecules are separated by many repeated partitionings between a flowing (mobile) phase and a stationary phase packed into a column.

**combinatorial control:** The use of specific combinations of a limited number of regulatory proteins to exert fine control over gene expression.

**comparative genomics:** The study of genome structure, function, and evolution by comparison across different species.

**competitive inhibitor:** A molecule that competes with the normal substrate or ligand for a protein's binding site.

**complementary DNA (cDNA):** A DNA complementary to a specific mRNA, used in DNA cloning; usually made by reverse transcriptase.

**complementary:** Having a molecular surface with chemical groups arranged to interact specifically with chemical groups on another molecule. Because of complementarity, if the nucleotide sequence of one strand of a double-stranded nucleic acid is known, the sequence of the opposite strand can be deduced.

**complex transposon:** A viruslike transposon with a large genome including genes not required for transposition.

**composite transposon:** A transposon that consists of two insertion elements flanking one or more genes not required for transposition, such as antibiotic-resistance genes.

**condensins:** SMC proteins that facilitate chromosomal condensation.

**consensus sequence:** A DNA or amino acid sequence consisting of the residues that most commonly occur at each position in a set of similar sequences.

**constitutive gene expression:** The continual expression of a gene. *Compare* regulated gene expression.

**constructive interference:** Phenomenon in which waves in the same phase add to create waves of larger amplitude.

**contig:** A series of overlapping clones or a continuous sequence defining an uninterrupted section of a chromosome.

**cooperativity:** The characteristic of an enzyme or other protein in which binding of the first molecule of a ligand changes the affinity for the second molecule. In positive cooperativity, the affinity for the second ligand molecule increases; in negative cooperativity, it decreases.

**core histones:** The four histone proteins (H2A, H2B, H3, and H4) that form the octameric core of a nucleosome.

**core promoter:** The DNA sequence elements in eukaryotic cells common to promoters used by Pol II. The TATA box and initiator sequence (Inr) are required elements of a core promoter; a TFIIB recognition element (BRE) and downstream promoter element (DPE) may also be involved in transcription initiation from some core promoters.

**corepressor:** A protein that inhibits transcription by binding both the RNA polymerase and a repressor or repressors, without binding the DNA directly. *Compare* coactivator.

**correlation spectroscopy (COSY):** A type of two-dimensional nuclear magnetic resonance spectroscopy in

## G-6 Glossary

which atoms that are near to one another and connected through covalent bonds can be identified.

**COSY:** See correlation spectroscopy.

**covalent bond:** A chemical bond that involves sharing of electron pairs.

**covalent modification:** The addition, dissociation, or rearrangement of an atom or functional group covalently bonded in a molecule. In biological systems, common modifying groups include acetyl, adenylyl, amide, carboxyl, hydroxyl, methyl, myristoyl, palmitoyl, phosphoryl, prenyl, sulfate, and uridylyl groups.

**CpG sequence:** DNA sequence (cytosine, guanine) that is a frequent substrate for cytosine methylation.

**Cre-lox:** A bacteriophage-encoded site-specific recombination system that promotes circularization of the phage genome and aids in proper segregation at cell division of phage plasmids in the lysogenic state.

**crossing over:** The reciprocal exchange of DNA between paired homologous chromosomes during meiosis; also called recombination.

**cross-linking:** The use of a small chemical agent with two reactive groups to covalently link molecules that are in close proximity.

**crossover:** A breaking and rejoining of DNA that results in sequences of one chromosome physically contiguous with sequences from a different, usually homologous chromosome.

**CRP:** See cAMP receptor protein.

**cruciform:** Secondary structure in double-stranded RNA or DNA in which the double helix is denatured at palindromic repeat sequences in each strand, and each separated strand is paired internally to form opposing hairpin structures. *Also see hairpin.*

**C-terminus (carboxyl terminus):** See carboxyl terminus.

**cyclic AMP (cAMP):** A second messenger, adenosine 3',5'-cyclic monophosphate; its formation in a cell by adenylyl cyclase is stimulated by certain hormones or other molecular signals.

**cyclic GMP (cGMP):** A second messenger, guanosine 3',5'-cyclic monophosphate; its formation in a cell by guanylyl cyclase is stimulated by certain hormones or other molecular signals.

**cyclobutane ring:** Structure formed by condensation of two ethylene groups on adjacent thymine residues in DNA.

**cycloheximide:** An antibiotic that inhibits protein synthesis by eukaryotic ribosomes by blocking peptidyl transfer.

**cytogenetics:** The study of chromosomes and their role in heredity.

**cytokinesis:** The final separation of daughter cells following mitosis.

**cytology:** The study of cells and cellular structures.

**cytoplasmic membrane:** The exterior membrane surrounding the cytoplasm of a cell. Also called the plasma membrane.

**cytosine (C):** A pyrimidine base that is a component of DNA and RNA.

**cytotoxic:** Deadly to cells.

**Dam methylase (DNA adenine methyltransferase):** An enzyme in *E. coli* that methylates adenine residues in the palindromic sequence GATC on both strands of the DNA. Transient hemimethylation of a DNA duplex following replication distinguishes the parental strand from the daughter strand.

**DCC:** See dosage compensation complex.

**DDE motif:** Protein secondary structure in which the amino acid residues D, D, and E (two aspartate residues and a glutamine residue) form the catalytic core in the active site of phosphoryltransferase enzymes such as integrases and transposases.

**deamination:** The enzymatic removal of amino groups from biomolecules such as amino acids or nucleotides.

**degenerate code:** A code in which a single element in one language is specified by more than one element in a second language. The genetic code is degenerate because some amino acids are specified by more than one codon.

**deletion analysis:** A method for assessing the functional importance of various regions of a protein by engineering a series of constructs with different parts of the gene deleted. The proteins expressed by these constructs can then be assayed for functionality.

**deletion mutation:** A mutation resulting from the deletion of one or more nucleotides from a gene or chromosome. *Compare insertion mutation.*

**denaturation:** Partial or complete unfolding of the specific native conformation of a polypeptide chain, protein, or nucleic acid such that the function of the molecule is lost.

**deoxyribonucleic acid:** See DNA.

**deoxyribonucleotide:** A nucleotide containing 2-deoxy-D-ribose as the pentose component.

**depurination:** The enzymatic removal of a purine base from a nucleotide.

**Dicer:** An endonuclease in eukaryotic cells that catalyzes the hydrolysis of double-stranded RNAs, producing siRNAs or processing pre-miRNAs to mature miRNAs. Dicer also plays a role in the creation of RNA-induced silencing complexes.

**diffraction pattern:** The interference pattern that results when a wave or series of waves is diffracted by an object with a regular structure, such as a crystal.

**dimer:** A molecule with two subunits.

**diphtheria toxin:** A bacterial toxin that catalyzes the ADP-ribosylation of a diphthamide (a modified histidine)

residue of elongation factor eEF2, thereby inactivating it and inhibiting protein synthesis by the eukaryotic ribosome.

**diploid:** Having two sets of genetic information; describes a cell with two chromosomes of each type. *Compare* haploid.

**directionality:** The direction in which a process or enzyme proceeds along an asymmetric molecule. For example, certain endonucleases act on DNA only in a 5' to 3' direction.

**dissociation constant ( $K_d$ ):** An equilibrium constant for the dissociation of a complex of two biomolecules into its components; for example, dissociation of a substrate from an enzyme.  $K_d$  is the reciprocal of the association constant,  $K_a$ . *Compare* association constant ( $K_a$ )

**distributive synthesis:** The enzymatic synthesis of a biological polymer in which the enzyme dissociates from the substrate after the addition of each monomeric unit. *Compare* processive synthesis.

**disulfide bond:** A covalent bond involving the oxidative linkage of two Cys residues, from the same or different polypeptide chains.

**Dmc1:** A eukaryotic recombinase structurally and functionally homologous to the RecA protein of *E. coli*. *Also see* Rad51.

**DNA (deoxyribonucleic acid):** A polynucleotide with a specific sequence of deoxyribonucleotide units covalently joined through 3', 5'-phosphodiester bonds; serves as the carrier of genetic information.

**DNA adenine methyltransferase:** *See* Dam methylase.

**DNA cloning:** *See* cloning.

**DNA genotyping:** *See* genotyping.

**DNA glycosylase:** An enzyme that hydrolyzes the N-β-glycosyl bond between a nucleotide base and pentose, creating an abasic site in the DNA.

**DNA helicase:** *See* helicase.

**DNA library:** A collection of cloned DNA fragments.

**DNA ligase:** An enzyme that creates a phosphodiester bond between the 3' end of one DNA segment and the 5' end of another.

**DNA looping:** The interaction of proteins bound at distant sites on a DNA molecule so that the intervening DNA forms a loop.

**DNA microarray:** A collection of DNA sequences immobilized on a solid surface, with individual sequences laid out in patterned arrays that can be probed by hybridization; also called a DNA chip.

**DNA nuclease:** *See* nucleases.

**DNA photolyase:** A flavoprotein enzyme that becomes an electron donor when activated by visible light. DNA photolyases can repair pyrimidine dimers and other lesions caused by ultraviolet light.

**DNA polymerase α (Pol α):** A eukaryotic DNA polymerase with both primase and error-prone DNA

polymerase activities. The enzyme synthesizes an RNA primer on a DNA template and then extends it with DNA.

**DNA polymerase δ (Pol δ):** A eukaryotic chromosomal replicase with both DNA polymerase and 3'→5' exonuclease activities. It acts on the lagging strand of the replication fork.

**DNA polymerase ε (Pol ε):** A eukaryotic chromosomal replicase with both DNA polymerase and 3'→5' exonuclease activities. It acts on the leading strand of the replication fork.

**DNA polymerase:** An enzyme that catalyzes template-dependent synthesis of DNA from its deoxyribonucleoside 5'-triphosphate precursors.

**DNA replication:** Synthesis of daughter DNA molecules identical to the parental DNA.

**DNA strand invasion:** The pairing of a single-stranded extension of a DNA molecule with a homologous region of another DNA molecule, with displacement of one strand of the recipient molecule by the invading strand.

**DNA supercoiling:** The coiling of DNA upon itself, generally as a result of bending, underwinding, or overwinding of the DNA helix.

**DNA topology:** The properties of DNA that do not change under continuous deformations such as twisting, bending, stretching, or binding other molecules.

**DNA underwinding:** Condition in which a closed-circular DNA has fewer helical turns than would be expected of B-form DNA. Its linking number,  $Lk$ , is negative, and the molecule is negatively supercoiled.

**DNA-binding transcription activator:** In eukaryotic cells, a protein that binds to enhancers or UASs to facilitate transcription. *Also called* DNA-binding transactivator. *Compare* coactivator.

**domain:** A distinct structural unit of a polypeptide; domains may have separate functions and may fold as independent, compact units.

**dominant:** The allele that determines the phenotype in a heterozygous individual. *Compare* recessive.

**donor site:** The location on a chromosome of a transposon before it moves to a target site. *Compare* target site.

**dosage compensation complex (DCC):** A ribonucleoprotein complex encoded by the X chromosome in *Drosophila*. The complex coats the single X chromosome in male cells, hyperstimulating transcription from its genes to compensate for the lack of a second X chromosome.

**dosage compensation:** Control of gene expression from sex chromosomes to ensure that both male and female cells express similar levels of each gene product.

**double bond:** Bond between two elements that involves four electrons instead of two.

**double-strand break (DSB):** A break in the phosphodiester backbone of both strands of a double-stranded nucleic acid.

## G-8 Glossary

**double-strand break repair (DSBR):** A method for repairing double-strand breaks that creates two Holliday intermediates, which must be cleaved by resolvases. The genes flanking the repair site may be unchanged or may undergo a reciprocal exchange, depending on how the crossovers are resolved.

**Drosha:** An endonuclease in eukaryotic cells that cleaves the hairpin of primary miRNA transcripts to produce pre-miRNAs.

**DSB:** See double-strand break.

**DSBR:** See double-strand break repair.

**duplication mutation:** The duplication of a large tract of DNA, leading to an increased dosage of genes in the affected area.

**E site:** The site in a ribosome occupied by the tRNA molecule released after the growing polypeptide chain is transferred to the aminoacyl-tRNA. Also called the exit site.

**editosome:** A protein complex that catalyzes the insertion or deletion of nucleoside residues during the process of RNA editing.

**eEF1 $\alpha$ :** A eukaryotic protein synthesis elongation factor factor that delivers aminoacyl-tRNAs to the A site of the elongation complex with the concomitant hydrolysis of bound GTP.

**eEF1 $\beta\gamma$ :** A eukaryotic protein synthesis elongation factor that uses bound GTP to regenerate eEF1 $\alpha$ -GTP from eEF1 $\alpha$ -GDP.

**eEF2:** A eukaryotic protein synthesis elongation factor factor with GTPase activity. GTP hydrolysis provides the energy for the ribosome to translocate along the mRNA to the next codon. Also called translocase.

**effector:** A small molecule that binds a transcription activator or repressor, causing a conformational change in the regulatory protein that results in an increase or decrease in transcription from the gene.

**EF-G:** A bacterial protein synthesis elongation factor with GTPase activity. GTP hydrolysis provides the energy for the ribosome to translocate along the mRNA to the next codon. Also called translocase.

**EF-Ts:** A bacterial protein synthesis elongation factor that uses bound GTP to regenerate EF-Tu-GTP from EF-Tu-GDP.

**EF-Tu:** A bacterial protein synthesis elongation factor that delivers aminoacyl-tRNAs to the A site of the elongation complex with the concomitant hydrolysis of bound GTP.

**EJC:** See exon junction complex.

**electric dipole moment:** A measure of the electrical polarity of a bond or molecule. It is equal to the magnitude of the charge times the distance separating the charges.

**electron density map:** Three-dimensional description of the electron density in a crystal, derived from x-ray diffraction data.

**electronegative atoms:** Atoms with a tendency to gain electrons.

**electronegativity:** The propensity of an atom to attract electrons to itself.

**electrophoresis:** See gel electrophoresis.

**electroporation:** Introduction of macromolecules into cells after rendering the cells transiently permeable by the application of a high-voltage pulse.

**electropositive atoms:** Atoms with a tendency to lose electrons.

**elongation complex:** The complex of proteins required for efficient synthesis of the RNA transcript after the RNA polymerase has moved beyond the promoter.

**elongation factors:** (1) Proteins required in the elongation phase of eukaryotic transcription. (2) Proteins required in the elongation phase of protein synthesis. *Also see eEF1 $\alpha$ , eEF1 $\beta\gamma$ , eEF2, EF-G, EF-Ts, and EF-Tu.*

**elongation:** (1) The second of three stages of RNA synthesis in which ribonucleotides are added to the 3' end of the growing RNA molecule. (2) The second of three stages of protein synthesis in which amino acids are added to the C-terminal end of the growing peptide chain.

**embryonic stem cells:** The cells in a mammalian embryo that retain the ability to divide and differentiate into other cell types. *Compare adult stem cells.*

**enantiomers:** Stereoisomers that are nonsuperimposable mirror images of each other.

**end replication problem:** The inability to replicate the final segment of DNA at the 3' end of the lagging strand where there is no primer to provide a free 3'-OH group.

**endonuclease:** An enzyme that hydrolyzes the interior phosphodiester bonds of a nucleic acid; that is, it acts at bonds other than the terminal bonds.

**enhanceosome:** A nucleoprotein complex of cooperating activators, which integrates regulatory information from multiple signals and generates a single transcriptional outcome at the target promoter.

**enhancer:** A DNA sequence that facilitates the expression of a given gene; may be located a few hundred, or even thousand, base pairs away from the gene. In yeast enhancers are called upstream activator sequences (UASs).

**entropy (S):** The extent of randomness or disorder in a system.

**enzyme kinetics:** The study of the rates of reactions catalyzed by enzymes.

**enzyme:** A biomolecule, either protein or RNA, that catalyzes a specific chemical reaction. It does not affect the equilibrium of the catalyzed reaction; it enhances the rate of the reaction by providing a reaction path with lower activation energy.

**epigenetic inheritance:** Inherited characteristic acquired by means that do not involve the nucleotide sequence of the

parental chromosomes; for example, covalent modifications of histones.

**epitope tag:** A protein sequence or domain bound by some well-characterized antibody.

**equilibrium expression:** A mathematical expression for the equilibrium constant of a chemical reaction expressed as the product of the molar concentrations of each reaction product raised to its coefficient in the balanced reaction over the product of the molar concentrations of each reactant raised to its coefficient in the balanced reaction.

**EST:** See expressed sequence tag.

**euchromatin:** The regions of interphase chromosomes that stain diffusely, as opposed to the more condensed, heavily staining, heterochromatin. These are often regions in which genes are being actively expressed.

**eukaryotes:** One of the three main groups of living organisms; eukaryotes are unicellular or multicellular organisms with cells having a membrane-bounded nucleus, multiple chromosomes, and internal organelles.

**evo-devo:** The field of evolutionary development, which demonstrates that dramatic phenotypic differences between species can be accounted for by changes in the temporal expression of shared or homologous genes and regulatory networks.

**evolution:** A process in which the population of a species changes over time. Genetic variation occurs in the populations due to mutation; competitive pressures in the environment lead to the natural selection of individuals whose genetic makeup gives them a reproductive advantage. Over time, the genetic makeup of the surviving population shifts, sometimes creating new species.

**exinuclease:** An enzyme that cleaves a phosphodiester bond in the DNA on either side of a bulky lesion in DNA; also called excision endonuclease.

**exon junction complex (EJC):** A complex of proteins deposited on an mRNA by the spliceosome 20 to 24 nucleotides upstream of exon-exon junctions.

**exon:** The segment of a eukaryotic gene that encodes a portion of the final product of the gene; a segment of RNA that remains after posttranscriptional processing and is transcribed into a protein or incorporated into the structure of an RNA. *Also see intron.*

**exonuclease:** An enzyme that hydrolyzes only those phosphodiester bonds that are in the terminal positions of a nucleic acid.

**exosome:** A complex of 3'→5' exonucleases in eukaryotic cells that processes the 3' ends of rRNAs and tRNAs and is responsible for RNA degradation in higher eukaryotes.

**exothermic reaction:** A chemical reaction that releases heat (that is, for which  $\Delta H$  is negative).

**exportin:** A protein receptor responsible for transporting RNAs from the nucleus, through a nuclear pore, into the cytoplasm.

**expressed sequence tag (EST):** A specific type of sequence-tagged site in DNA representing a gene that is expressed.

**expression vector:** Cloning vector with the transcription and translation signals needed for the regulated expression of a cloned gene. *See cloning vector.*

**extrachromosomally primed (EP) retrotransposon:** A retrotransposon that moves via a double-stranded cDNA copy of its mRNA transcript. The cDNA inserts itself into the target site in a reaction catalyzed by a recombinase or integrase. *Also see target-primed (TP) retrotransposon.*

**F<sub>1</sub> generation:** First filial generation, the hybrid offspring in a genetic cross.

**F<sub>2</sub> generation:** Second filial generation, the offspring of crossing the F<sub>1</sub> generation.

**first law of thermodynamics:** The law stating that, in all processes, the total energy of the universe remains constant.

**5' cap:** A residue of 7-methylguanosine (7-meG) linked to the 5'-terminal residue of an mRNA through a 5',5'-triphosphate linkage, which protects the mRNA from exoribonucleases.

**fork regression:** Backward movement of the replication fork, which can occur when a replication fork encounters a lesion and stalls. Fork regression allows the parental strands to reanneal until the lesion is repaired.

**four-helix bundle:** Supersecondary protein structure in which four  $\alpha$  helices associate through hydrophobic interactions.

**frameshift mutation:** A mutation caused by insertion or deletion of one or more paired nucleotides, changing the reading frame of codons during protein synthesis; the polypeptide product has an altered amino acid sequence beginning at the mutated codon.

**free energy (G):** The component of the total energy of a system that can do work at constant temperature and pressure.

**free-energy change ( $\Delta G$ ):** The amount of free energy released (negative  $\Delta G$ ) or absorbed (positive  $\Delta G$ ) in a reaction at constant temperature and pressure.

**functional RNA:** An RNA molecule that is a functional end product, as distinct from messenger RNA (mRNA), which serves as a transient intermediary between DNA and a protein product it encodes.

**fusion gene:** A hybrid gene formed when chromosomal DNA is rearranged by deletion, duplication, insertion, or transposition.

**fusion protein:** The protein product of a gene created by the fusion of two distinct genes or portions of genes.

**gamete:** A reproductive cell with a haploid gene content; a sperm or egg cell.

**gap gene:** A subclass of the segmentation genes involved in dividing the developing *Drosophila* embryo into broad regions. Gap genes are expressed before the pair-rule genes.

**gap repair:** A process for repairing gaps left when the replication fork bypasses a lesion.

**gel electrophoresis:** A technique for separating mixtures of large charged molecules such as proteins or nucleic acids by causing them to move through a gel matrix in an applied electric field.

**gel-exclusion chromatography:** A type of column chromatography in which molecules are separated by size, based on the capacity of porous polymers to exclude solutes above a certain size.

**gene conversion:** A nonreciprocal transfer of genetic information as an outcome of DNA repair, especially during meiosis.

**gene silencing:** (1) Suppression of gene expression by incorporation of the gene into transcriptionally inactive heterochromatin. (2) Suppression of gene expression by short interfering RNAs, which bind mRNAs and target them for degradation.

**gene:** A chromosomal segment that codes for a single functional polypeptide chain or RNA molecule.

**general rate constant ( $k_{cat}$ ):** The limiting rate of an enzyme-catalyzed reaction. It describes the number of substrate molecules converted to product by a single molecule of enzyme at saturating levels of substrate. The constant has units of reciprocal time. *Also see turnover number.*

**general transcription factor:** In eukaryotic cells, a protein required at every Pol II promoter. Also called a basal transcription factor.

**genetic code:** The set of triplet code words in DNA (or mRNA) coding for the amino acids of proteins.

**genetic drift:** The change in frequency of an allele in a population due to random sampling, rather than selective pressure. Genetic drift is affected by such variables as the number of reproducing individuals in a population and the number of offspring generated.

**genetic engineering:** Manipulation of an organism's genome in the laboratory.

**genetics:** The science of heredity and the variation of inherited characteristics.

**genome annotation:** Information about the location and function of genes and other regulatory and functional sequences in a genome.

**genome:** One copy of all the genetic information encoded in a cell or virus. In a eukaryote, this generally constitutes one copy of all the genetic information in the nucleus. Separate genomes are found in certain organelles, particularly mitochondria and chloroplasts.

**genomic library:** A DNA library containing DNA segments that represent all (or most) of the sequences in an organism's genome.

**genomics:** Broadly, the study of genomes. Genomics embraces sequencing, mapping, and annotating genomes, organizing databases to archive genomic data, developing computational tools to analyze the data, and application of genomic data to other fields, such as medicine.

**genotoxic:** Causing damage to the genomic DNA.

**genotype:** The genetic constitution of an organism, as distinct from its physical characteristics, or phenotype.

**genotyping:** The process of determining the genetic constitution of an individual. Also called DNA fingerprinting or DNA profiling.

**GFP:** *See* green fluorescent protein.

**glycosidic bonds:** Bonds between a sugar and another molecule (typically an alcohol, purine, pyrimidine, or sugar) through an intervening oxygen.

**G<sub>1</sub> phase:** The first gap phase of the eukaryotic cell cycle in which the cell is diploid. G<sub>1</sub>, part of interphase, occurs before the S (synthesis) phase in which the DNA is replicated.

**GPCR:** *See* G protein-coupled receptor.

**G protein-coupled receptor (GPCR):** Any of a large family of membrane receptor proteins with seven transmembrane helical segments, often associating with G proteins to transduce an extracellular signal into a change in cellular metabolism.

**Greek key:** Supersecondary protein motif in which four antiparallel  $\beta$  strands combine in a pattern seen on ancient Greek pottery.

**green fluorescent protein (GFP):** A small protein that produces a bright fluorescence in the green region of the visible spectrum. Fusion proteins with GFP are commonly used to determine the subcellular location of the fused protein by fluorescence microscopy. Variants that produce other colors (e.g., red fluorescent protein (RFP) or cyan fluorescent protein (CFP)) have been produced.

**group I intron:** A large, self-splicing ribozyme that catalyzes its own excision from an mRNA, tRNA or rRNA transcript in a reaction that requires a guanosine nucleotide or nucleoside to initiate the reaction.

**group II intron:** A large, self-splicing ribozyme that catalyzes its own excision from an mRNA transcript as a lariat structure.

**G tetraplex:** Four-stranded DNA structure that can form from G-rich segments of DNA.

**G<sub>2</sub> phase:** The second gap phase of the eukaryotic cell cycle in which the cell is tetraploid. G<sub>2</sub>, part of interphase, occurs between the S (synthesis) phase and the M (mitosis) phase.

**guanine (G):** A pyrimidine base that is a component of DNA and RNA.

**guanosine 3',5'-cyclic monophosphate:** *See* cyclic GMP.

**hairpin:** Secondary structure in single-stranded RNA or DNA, in which complementary parts of a palindromic repeat fold back and are paired to form an antiparallel duplex helix that is closed at one end.

**haploid:** Having a single set of genetic information; describes a cell with one chromosome of each type. *Compare* diploid.

**haplotype:** (1) In genetics, a group of alleles on a chromosome that are nearly always inherited together. (2) In genomics, a set of single-nucleotide polymorphisms that are nearly always inherited together.

**HAT:** *See* histone acetyltransferase.

**helicase:** An enzyme that catalyzes the separation of strands in a nucleic acid molecule in a reaction coupled to the hydrolysis of ATP.

**helix-turn-helix:** Supersecondary protein motif consisting of two  $\alpha$  helices separated by a  $\beta$  turn. This motif is crucial to the interaction of many bacterial regulatory proteins with DNA.

**heterochromatin:** The condensed, heavily staining portions of chromosomes that are not transcriptionally active, including centromeres, telomeres, some repetitive DNA sequences, and mitotic chromosomes.

**heterooligomer:** Multisubunit molecule (oligomer) with nonidentical subunits.

**heterotropic:** Describes an allosteric modulator that is distinct from the normal ligand or an allosteric enzyme requiring a modulator other than its substrate.

**heterozygous:** Having different alleles at a specific genetic locus.

**hierarchical model:** Model for protein folding that proposes that local regions of secondary structure form first, followed by longer-range interactions, continuing until complete domains form and the entire polypeptide is folded. *Compare* molten globule model.

**high-mobility group (HMG) proteins:** Three families of chromosomal proteins that bind DNA nonspecifically, promoting chromatin remodeling and DNA looping for regulating DNA transcription.

**histone acetyltransferase (HAT):** Any of a family of enzymes that transfer an acetyl group from acetyl-CoA to the  $\epsilon$ -amino group of specific Lys residues on histone tails.

**histone chaperones:** Acidic proteins required for the assembly of histone octamers on DNA.

**histone code:** Hypothetical code in which successive covalent modifications of histone tails and DNA trigger chromatin remodeling and transcriptional activation events.

**histone modifying enzymes:** A class of enzymes that covalently modify the N-terminal tails of histones.

**histone octamer:** The complex of two copies of each of the four core histones that forms the histone core of the nucleosome.

**histone tails:** The flexible, disordered N-terminal ends of the histone proteins that comprise the histone core. These ends protrude from the nucleosome and contact adjacent nucleosomes.

**histone fold:** Protein structural motif formed from three  $\alpha$  helices connected by two loops. Histone-fold dimers are instrumental in the tight wrapping of the DNA helix around the histone core in nucleosomes.

**histones:** The family of basic proteins that associate tightly with DNA in the chromosomes of all eukaryotic cells.

**HMG:** *See* high-mobility group (HMG) proteins.

**Holliday intermediate:** An intermediate in genetic recombination in which two double-stranded DNA molecules are joined by a reciprocal crossover involving one strand of each molecule.

**holoenzyme:** A catalytically active enzyme, including all necessary subunits, prosthetic groups, and cofactors.

**homeodomain:** A conserved 60 amino acid sequence motif in transcription activators encoded by genes that regulate body pattern development.

**homeotic genes:** Genes that regulate development of the pattern of segments in the *Drosophila* body plan; similar genes are found in most vertebrates. Homeotic genes are expressed after the segmentation genes.

**homing endonucleases:** Intron-encoded restriction endonucleases that recognize and cleave an asymmetric sequence of 12–40 base pairs in the cellular DNA. Repair of the break results in insertion of a copy of the intron via homologous recombination.

**homologous chromosomes:** In diploid organisms, a pair of chromosomes, one inherited from each parent, which are of similar length, structure, and gene sequence. Also called homologs.

**homologous recombination:** Recombination between two DNA molecules of similar sequence, occurring in all cells; occurs during meiosis and mitosis in eukaryotes.

**homologs:** Genes or proteins with sequence similarity. Also shorthand for homologous chromosomes.

**homooligomer:** Multisubunit molecule (oligomer) with identical subunits.

**homotropic:** Describes an allosteric modulator that is identical to the normal ligand or an allosteric enzyme that uses its substrate as a modulator.

**homozygous:** Having identical alleles at a specific genetic locus.

**Hoogsteen pairing:** Non-Watson-Crick pairing of a pyrimidine base to a purine base that is already participating in a Watson-Crick base pair with another pyrimidine. The arrangement allows for the formation of triplex DNA.

**Hoogsteen position:** The atoms in a purine base that participate in Hoogsteen pairs (non-Watson-Crick hydrogen bonding) with a pyrimidine base.

**horizontal gene transfer:** Process by which an organism receives genetic information from another organism from which it is not a descendant.

**hormone response element (HRE):** A short (12 to 20 bp) DNA sequence that binds receptors for steroid, retinoid, thyroid, and vitamin D hormones, altering the expression of the contiguous genes. Each hormone has a consensus sequence preferred by the cognate receptor.

**housekeeping gene:** A gene that must be expressed continually for the cell to survive.

**Hox genes:** A major class of homeotic genes.

**HRE:** See hormone response element.

**Hsp70:** Family of heat-shock proteins with  $M_r \approx 70,000$  that constitute a class of molecular chaperones.

**hybrid duplex:** Duplex experimentally reconstituted from single-stranded DNA (or RNA) from different sources.

**hybrid:** The offspring of a cross between genetically nonidentical individuals.

**hydrogen bond:** A weak electrostatic attraction between one electronegative atom (such as oxygen or nitrogen) and a hydrogen atom covalently linked to a second electronegative atom.

**hydrolysis:** Cleavage of a bond, such as an anhydride or peptide bond, by the addition of the elements of water, yielding two or more products.

**hydrophobic interactions:** The association of nonpolar groups or compounds with each other in aqueous systems, driven by the tendency of the surrounding water molecules to seek their most stable (disordered) state.

**hydrophobic stacking:** A property of adjacent bases in a DNA strand or of base pairs in a DNA double helix that describes their orientation relative to each other. Parallel orientation of the hydrophobic planar rings minimizes their association with water and contributes to the stability of the B-form (Watson-Crick) double helix. Also known as base stacking.

**hyperchromic effect:** The large increase in light absorption at 260 nm occurring as a double-helical DNA unwinds (melts). *Also see hypochromic effect.*

**hypersensitive site:** A DNA sequence that is especially sensitive to cleavage by DNase I and other nucleases. These sites typically precede active promoters and may be binding sites for proteins regulating expression from the downstream gene.

**hypochromic effect:** The large decrease in light absorption at 260 nm occurring as single strands of DNA anneal to form double-helical DNA. *Also see hyperchromic effect.*

**hypothesis:** A proposal that provides a reasonable explanation for observations, but has not yet been substantiated by sufficient experimental evidence to stand up to rigorous critical examination.

**IF-1:** A bacterial protein synthesis initiation factor that binds the ribosomal A site and blocks tRNA binding.

**IF-2:** A bacterial protein synthesis initiation factor that directs the initiating tRNA to the P site of the 30S subunit.

When the 50S subunit binds to the complex, it hydrolyzes the GTP bound to IF2, releasing IF2 and allowing the 70S subunit to form.

**IF-3:** A bacterial protein synthesis initiation factor that prevents premature addition of the 50S ribosomal subunit to the assembling initiation complex.

**immunofluorescence:** The labeling of antibodies with a fluorescent dye to visualize or quantify an antigen in a biological, biochemical, or histological preparation.

**immunoprecipitation:** Use of antibodies against an epitope on a protein of interest (often with secondary antibodies against those primary antibodies) to precipitate the protein from a complex mixture.

**importin:** A protein receptor responsible for transporting noncoding RNAs processed in the cytoplasm into the nucleus through a nuclear pore.

**imprinting:** An epigenetic method of regulating gene expression based on the parental origin of the gene.

**incomplete dominance:** A condition in which alleles at a specific locus are neither dominant nor recessive, and the progeny express a phenotype intermediate between those of the two parents. *Compare codominance.*

**indel:** Collective term for insertion and deletion mutations.

**induced fit:** A change in the conformation of an enzyme in response to substrate binding that renders the enzyme catalytically active; also used to denote changes in the conformation of any macromolecule in response to ligand binding such that the binding site of the macromolecule better conforms to the shape of the ligand.

**inducer:** A signal molecule that, when bound to a regulatory protein, produces an increase in the expression of a given gene.

**initial model:** Protein structure derived from an electron density map before further refinements are made.

**initial velocity ( $V_0$ ):** The velocity of a reaction while the concentration of substrate is saturating and can be regarded as constant relative to the enzyme concentration.

**initiation codon:** AUG (sometimes GUG or, even more rarely, UUG in bacteria and archaea); codes for the first amino acid in a polypeptide sequence: N-formylmethionine in bacteria; methionine in archaea and eukaryotes. Also called start codon.

**initiation complex:** A complex of a ribosome with an mRNA and the initiating Met-tRNA<sub>i</sub><sup>Met</sup> or fMet-tRNA<sup>fMet</sup>, ready for the elongation steps.

**initiation factors:** Three protein factors required to assemble the ribosomal subunits and initiator tRNA in preparation for protein synthesis in bacteria. *Also see IF-1, IF-2, and IF-3.*

**initiation:** (1) The first of three stages in the synthesis of DNA, in which the DNA polymerase binds to the origin of replication. (2) The first of three stages in the synthesis of RNA, in which the RNA polymerase binds to the promoter sequence on the DNA. (3) The first of three stages in the

synthesis of a protein, in which the ribosome binds to the mRNA and initiator aminoacyl-tRNA.

**initiator protein:** Protein that binds specific sites in an origin of replication and serves as a nucleation site for the assembly of other protein complexes necessary to initiate replication; for example, DnaA in *E. coli*, and ORC in eukaryotes.

**insertion mutation:** A mutation caused by insertion of one or more extra bases between successive bases in DNA. *Compare* deletion mutation.

**insertion sequence:** Specific base sequences at either end of a transposable segment of DNA.

**insertion site:** Site within the active site of a DNA polymerase where the template nucleotide and incoming dNTP are positioned. *Compare* postinsertion site.

**insulator:** A short sequence of DNA that prevents inappropriate cross-signaling between regulatory elements for different genes. Also called a boundary element.

**integrase:** An enzyme that catalyzes the insertion of a retrovirus or retrotransposon into its target site.

**internal ribosome entry site (IRES):** A site on the 5' side of the start codon in some viral and eukaryotic mRNAs where a eukaryotic ribosome can bind in the absence of a 5' cap.

**interphase:** The portion of the cell cycle that does not include mitosis. Subdivided into three phases: G<sub>1</sub> phase, S phase, and G<sub>2</sub> phase.

**intron:** A sequence of nucleotides in a gene that is transcribed but excised before the gene is translated; also called intervening sequence. *Also see* exon.

**inversion mutation:** A mutation that results from the inversion of a large segment of DNA in a chromosome.

**inverted repeat:** A sequence that is the reversed complement of a downstream sequence.

**ion-exchange chromatography:** A type of column chromatography where molecules are separated by charge, using a stationary phase that contains fixed charged groups.

**ionic bond:** Chemical bond in which the electrons of one atom are transferred to another, creating positive and negative ions that attract each other.

**IRE:** *See* iron response element.

**IRES:** *See* internal ribosome entry site.

**iron homeostasis:** The maintenance of a dynamic steady-state concentration of cellular iron by regulatory mechanisms that compensate for changes in external circumstances.

**iron response element (IRE):** A hairpin structure in the 3' or 5' untranslated region of the mRNAs for proteins involved in iron homeostasis. Binding of the iron response protein (IRP) to a 5' IRE inhibits translation of the mRNA; binding of IRP to a 3' IRE inhibits degradation of the mRNA.

**iron response protein (IRP):** A protein that binds the iron response element (IRE) in mRNAs for proteins involved in iron homeostasis, inhibiting their translation or degradation in response to the cell's need for iron. Iron-sulfur centers required for efficient binding of IRPs to IREs only form when iron is plentiful in the cell, and therefore serve as a sensor of the cellular level of iron.

**IRP:** *See* iron response protein.

**irreversible inhibitor:** A molecule that either forms a stable noncovalent association with an enzyme or binds the enzyme covalently, destroying a functional group necessary for its catalytic activity.

**isoenergetic:** Describes a chemical reaction in which the reactants and products have the same or very similar free energy, and therefore exist at similar concentrations at equilibrium.

**isomorphous replacement:** A method for dealing with the phase problem in x-ray crystallography. A protein crystal is soaked in a heavy metal solution such that a few atoms are incorporated. The heavy metals give strong signals that can be discerned from the rest of the protein, allowing their phases to be determined.

**K<sub>a</sub>:** *See* association constant

**karyopherins:** A family of nuclear transport receptors including importins and exportins.

**K<sub>d</sub>:** *See* dissociation constant.

**kinetic proofreading:** A mechanism for error correction in complex biological processes that maximizes the speed of correct reactions while stalling and allowing incorrect reactions to reverse.

**kinetics:** The study of reaction rates.

**K<sub>m</sub>:** *See* Michaelis constant.

**Kozak sequence:** A sequence around the start codon in eukaryotic mRNA that enhances its translation. The Kozak sequence has a purine nucleotide three residues before, and a G residue immediately after, the start codon.

**lagging strand:** The DNA strand that, during replication, must be synthesized in the direction opposite to that in which the replication fork moves.

**last universal common ancestor:** *See* LUCA.

**law of independent assortment:** In the formation of gametes there is independent assortment of alleles for different genes. Also known as Mendel's second law.

**law of segregation:** In the formation of gametes there is an equal segregation of alleles. In other words, a haploid gamete contains one copy of each gene. Also known as Mendel's first law.

**leader peptide:** A short sequence near the amino terminus of a protein that has a specialized targeting or regulatory function.

**leader sequence:** A short sequence near the 5' end of an RNA that has a specialized targeting or regulatory function.

**leading strand:** The DNA strand that, during replication, is synthesized in the same direction in which the replication fork moves.

**leucine zipper:** A protein structural motif involved in protein-protein interactions in many eukaryotic regulatory proteins; consists of two interacting  $\alpha$  helices in which Leu residues in every seventh position are a prominent feature of the interacting surfaces.

**ligand:** A small molecule that binds specifically to a larger one; for example, a hormone is the ligand for its specific protein receptor.

**linkage analysis:** Use of bioinformatics to analyze the statistical association between inheritance of a gene and the presence of specific single-nucleotide polymorphisms, with the goal of mapping the gene to a specific location on a chromosome.

**linked genes:** Genes that are close together on a chromosome and whose alleles therefore assort together during meiosis, in contradiction to Mendel's second law.

**linker histone:** The histone protein H1, which binds to the linker DNA adjacent to the nucleosome.

**linker:** Synthetic DNA fragment inserted into a cloning vector, usually to provide a specific desired sequence, such as a restriction endonuclease recognition sequence.

**linking number (*Lk*):** The number of times one closed circular DNA strand is wound about another; the number of topological links holding the circles together.

***Lk*:** See linking number.

**LUCA (last universal common ancestor):** The single cell that gave rise to all life currently existing on Earth.

**lysis:** Destruction of a plasma membrane or (in bacteria) cell wall, releasing the cellular contents and killing the cell.

**lysogen:** A bacterial cell infected with a prophage.

**lysogenic pathway:** Bacteriophage infection in which the DNA is incorporated into the host chromosome or as an autonomously replicating plasmid with most of its genes repressed. *Compare* lytic pathway.

**lytic pathway:** Parasitic bacteriophage infection in which the DNA is replicated, packaged into phage heads, and the host cell is destroyed by lysis to disperse the progeny.

*Compare* lysogenic pathway.

**MAD:** See multiwavelength anomalous dispersion.

**major groove:** The wider of two grooves that wind around the outside of a DNA double helix.

**mass spectrometry:** Analytic technique for determining the mass of a molecule, thus providing a key clue to its identity, by measuring the charge-to-mass ratio of gaseous ions formed from the molecule as the ions pass through an electromagnetic field in a vacuum.

**maternal genes:** Genes expressed in the unfertilized egg that are required for development of the early embryo.

**maternal mRNAs:** Transcripts of maternal genes that are generated in the egg during oogenesis and remain dormant until fertilization.

**mating type:** In yeast, one of the two haploid forms,  $\alpha$  and  $\alpha$ , that can only mate with a haploid cell of the opposite type to form a diploid cell.

**maximum velocity:** See  $V_{\max}$ .

**MCM helicase:** A ring-shaped eukaryotic helicase that acts at the replication fork. It is composed of six homologous, but nonidentical, AAA+ proteins and interacts with two other proteins to form the CMG complex.

**Mediator complex:** A large, multiprotein complex in eukaryotic cells that serves as the mediator between the Pol II transcription complex and any upstream transcription activators or enhancers regulating Pol II-catalyzed transcription.

**meiosis:** A type of cell division in which diploid cells give rise to haploid cells destined to become gametes.

**melting point ( $T_m$ ):** The temperature at which a specific double-stranded polynucleotide separates into single strands.

**melting:** Denaturation or unwinding of a double-stranded polynucleotide to single-stranded polynucleotides.

**messenger RNA (mRNA):** A class of RNA molecules, each of which is complementary to one strand of DNA; carries the genetic message from the chromosome to the ribosomes.

**metagenomics:** Structural and functional analysis of the collective genome of an environmental population of microorganisms rather than a pure population derived from a single cultured cell.

**metamerism:** Division of the body into a series of repeating segments, as in insects, for example.

**metaphase plate:** Equatorial plane in a dividing cell along which chromosomes align during metaphase.

**metaphase:** The second stage of mitosis (M phase). The spindle apparatus directs condensed sister chromatid pairs to align along the metaphase plate.

**Michaelis constant ( $K_m$ ):** The substrate concentration at which an enzyme-catalyzed reaction proceeds at one-half its maximum velocity.

**Michaelis-Menten equation:** The equation describing the hyperbolic dependence of the initial reaction velocity,  $V_0$ , on substrate concentration, [S], in many enzyme-catalyzed reactions.

**microprocessor complex:** A nuclear complex responsible for the early stages of miRNA and siRNA processing in eukaryotic cells. It consists of a primary miRNA transcript, an miRNA recognition protein, and the endonuclease Drosha.

**microRNA (miRNA):** A class of small RNA molecules (21 to 23 nucleotides after processing is complete) involved

in gene silencing by inhibiting translation and/or promoting the degradation of particular mRNAs.

**migration:** The movement of a population to a new geographical location.

**minor groove:** The narrower of two grooves that wind around the outside of a DNA double helix.

**miRNA:** *See* microRNA.

**mirror repeat:** A segment of duplex DNA in which the base sequences exhibit symmetry on each single strand.

**mismatch repair (MMR):** An enzymatic system for repairing base mismatches (non-Watson Crick pairs) in DNA.

**missense mutation:** A single-nucleotide change in a gene that results in an amino acid change in the protein product.

**mitosis:** In eukaryotic cells, the multistep process that results in the replication of chromosomes and cell division. In mitosis, one diploid cell gives rise to two diploid cells.

**mixed inhibitor:** An inhibitor molecule that can bind to either the free enzyme or the enzyme-substrate complex (not necessarily with the same affinity).

**MMR:** *See* mismatch repair.

**modulator:** A metabolite that, when bound to the allosteric site of an enzyme, alters its kinetic characteristics.

**MOI:** *See* multiplicity of infection.

**mole:** One gram molecular weight of a compound. A mole of any compound contains  $6.02 \times 10^{23}$  molecules.

**molecular biology:** The study of essential cellular macromolecules, including DNA, RNA, and proteins, and the biological pathways between them.

**molecular function (of a gene product):** The precise biochemical activity of a protein or an RNA, such as the reactions an enzyme catalyzes, the ligands a receptor binds, or the complex formed between a specific RNA and a protein. *Compare* cellular function and phenotypic function.

**molecular genetics:** the study of the structure and function of genes at the molecular level.

**molecular orbital model:** Mathematical function describing the wavelike behavior of electrons in a molecule.

**molecular replacement:** The use of the known structure of a closely related protein to solve the phase problem in the x-ray crystallographic analysis of a protein of unknown structure.

**molten globule model:** Model for protein folding in which the hydrophobic residues of a polypeptide chain rapidly collapse into a condensed, partially ordered state, which limits the conformations available to the rest of the molecule. As subdomains with tertiary structure develop, alternative conformations become increasingly limited, and the molecule achieves its native conformation. *Compare* hierarchical model.

**motif:** Any distinct folding pattern for elements of secondary structure, observed in one or more proteins.

A motif can be simple or complex and can represent all or just a small part of a polypeptide chain. Also called a fold or supersecondary structure.

**motor protein:** A protein that uses energy (typically from the hydrolysis of ATP) to undergo a cyclic conformational change that creates a unified, directional force.

**M phase:** Phase of the eukaryotic cell cycle during which mitosis, or cell division, occurs. The M phase follows the G<sub>2</sub> phase and precedes the G<sub>1</sub> phase.

**mRNA:** *See* messenger RNA.

**multiplicity of infection (MOI):** The ratio of infectious particles to target cells.

**multipotent:** Describes stem cells that can differentiate into a number of types of closely related cells.

**multiwavelength anomalous dispersion (MAD):** An approach to solving the phase problem in x-ray crystallography using isomorphous replacement with only a single molecule. Typically, selenomethionine is used as a replacement for methionine.

**mutation rate:** The frequency of new mutations (in a gene or in an organism) in each cellular generation.

**mutation:** An inheritable change in the nucleotide sequence of a chromosome.

**natural selection:** process by which traits (phenotypes) become more prevalent in a population because those individuals best adapted to exploit the prevailing resources are the ones most likely to survive and reproduce, passing on their advantageous traits.

**negative regulation:** Decreased expression of a gene by the binding of a repressor protein. *Compare* positive regulation.

**negative supercoiling:** The twisting of a helical (coiled) molecule on itself to form a right-handed supercoil.

**NER:** *See* nucleotide excision repair.

**NHEJ:** *See* nonhomologous end joining.

**niche:** In cellular differentiation, a microenvironment that allows for maintenance of both multipotent adult stem cells and their differentiated progeny.

**nick translation:** A concerted process of 5'→3' excision and DNA polymerization that shifts a discontinuity in the phosphodiester backbone between the 3' hydroxyl of one nucleotide and the 5' phosphate of the adjacent nucleotide along a DNA strand.

**NLS:** *See* nuclear localization sequence.

**NMD:** *See* nonsense-mediated decay.

**NMR:** *See* nuclear magnetic resonance spectroscopy.

**NOESY:** *See* nuclear Overhauser effect spectroscopy.

**nondisjunction:** The failure of paired chromosomes or sister chromatids to segregate during mitotic or meiotic cell division.

**nonhomologous end joining (NHEJ):** Method for repairing double-strand breaks by joining nonhomologous DNA ends in a process that does not conserve the original sequence.

**nonpolar:** Hydrophobic; describes molecules or groups that have no effective dipole moment and are therefore poorly soluble in water.

**nonsense mutation:** A mutation that results in the premature termination of a polypeptide chain.

**nonsense-mediated decay (NMD):** A pathway for degradation of mRNA molecules with a premature stop codon, triggered by the presence of an exon junction complex on a transcript that has been translated. *Also see* non-stop decay.

**non-stop decay:** A pathway for degradation of mRNA molecules lacking a stop codon, triggered by the release of the ribosome from the 3' end of the message. *Also see* nonsense-mediated decay.

**nontemplate strand:** *See* coding strand.

**Northern blot:** A nucleic acid hybridization procedure in which one or more specific RNA fragments are detected in a larger population by hybridization to a complementary, labeled DNA probe.

**N-terminus (amino terminus):** *See* amino terminus.

**nuclear localization sequence (NLS):** An amino acid sequence that targets a protein for transport to the nucleus.

**nuclear magnetic resonance (NMR) spectroscopy:** A technique that utilizes certain quantum mechanical properties of atomic nuclei to study the structure and dynamics of the molecules of which they are a part.

**nuclear Overhauser effect spectroscopy (NOESY):** A type of two-dimensional nuclear magnetic resonance spectroscopy in which atoms that are near to one another in space, but not necessarily nearby in the primary structure, can be identified.

**nucleases:** Enzymes that hydrolyze the internucleotide (phosphodiester) linkages of nucleic acids.

**nucleic acids:** Biologically occurring polynucleotides in which the nucleotide residues are linked in a specific sequence by phosphodiester bonds; DNA and RNA.

**nucleoid:** In bacteria, the nuclear zone that contains the chromosome but has no surrounding membrane.

**nucleolytic proofreading:** Pathway for the correction of errors in an RNA transcript in which the RNA polymerase reverses direction by one or a few nucleotides on the template and its endonuclease activity hydrolyzes the phosphodiester backbone of the transcript proximal to the mismatched base.

**nucleophile:** An electron-rich group with a strong tendency to donate electrons to an electron-deficient nucleus (electrophile); the entering reactant in a bimolecular substitution reaction.

**nucleoside:** A compound consisting of a purine or pyrimidine base covalently linked to a pentose.

**nucleosome:** In eukaryotes, the structural unit for packaging chromatin; consists of a DNA strand wound around a histone core.

**nucleotide excision repair (NER):** DNA repair pathway that involves exocinucllease-catalyzed cleavage of the phosphodiester bond on either side of a bulky DNA lesion such as a pyrimidine dimer or base adduct, followed by removal of the segment containing the lesion, and DNA polymerization and ligation to fill the gap.

**nucleotide:** A nucleoside phosphorylated at one of its pentose hydroxyl groups.

**Okazaki fragment:** Short segment of DNA synthesized on the lagging strand during DNA replication.

**oligomer:** A short polymer, usually of amino acids, sugars, or nucleotides; the definition of “short” is somewhat arbitrary, but usually fewer than 50 subunits.

**oligomeric state:** The number of identical polypeptide subunits in a particular form of a protein. For example, monomer, dimer, and trimer are different oligomeric states.

**oligonucleotide:** A short polymer of nucleotides (usually fewer than 50).

**oligonucleotide-directed mutagenesis:** Method for creating a mutation in a cloned gene. Two short, complementary synthetic DNA strands, each with the desired base change, are annealed to opposite strands of the cloned gene within a suitable vector. The two annealed oligonucleotides prime DNA synthesis, creating two complementary strands with the mutation.

**oncogene:** A cancer-causing gene; any of several mutant genes that cause cells to exhibit rapid, uncontrolled proliferation.

**open complex:** (1) A complex of the RNA polymerase bound to a promoter, in which the DNA is partially unwound. Transcription initiation occurs in the open complex. *Compare* closed complex. (2) A complex assembled on the *E. coli* *oriC* origin of replication at an early stage of replication initiation. It includes an oligomer of the AAA+ protein DnaA, ATP, and the histonelike protein HU.

**open form:** The conformation assumed by *E. coli* DNA polymerase I when a primed template is bound to the active site, but the correct dNTP is not.

**open reading frame (ORF):** A group of contiguous nonoverlapping nucleotide codons in a DNA or RNA molecule that does not include a termination codon.

**operator:** A region of DNA that interacts with a repressor protein to control the expression of a group of genes organized in an operon.

**operon:** A unit of genetic expression consisting of one or more related genes and the operator and promoter sequences that regulate their transcription.

**optically active:** Able to rotate the plane of plane-polarized light.

**ORC:** *See* origin recognition complex.

**ORF:** *See* open reading frame.

**organelles:** Membrane-bounded structures found in eukaryotic cells; contain enzymes and other components required for specialized cell functions.

**ori:** *See* origin of replication.

**origin of replication (ori):** The nucleotide sequence or site in DNA where DNA replication is initiated.

**origin recognition complex (ORC):** Eukaryotic initiator protein complex that assembles at an origin to initiate replication.

**orthologs:** Genes in different organisms that possess a clear sequence and functional relationship to each other. *Compare* paralogs.

**out-group:** A taxon outside the group of interest in a phylogenetic tree.

**pair-rule gene:** A subclass of the segmentation genes expressed in alternate body segments of the developing *Drosophila* embryo after the gap genes and before the segment polarity genes.

**palindrome:** A segment of duplex DNA in which the base sequences of the two strands exhibit twofold rotational symmetry about an axis.

**parallel  $\beta$  sheet:** *See*  $\beta$  sheet.

**paralogs:** Genes within a species that possess a clear sequence and functional relationship to each other and likely arose as the result of a gene duplication. *Compare* orthologs.

**parthenogenesis:** Reproduction by the growth and development of an unfertilized egg.

**P body:** *See* processing body.

**PCNA:** Proliferating cell nuclear antigen, the eukaryotic sliding clamp protein that tethers DNA polymerase to the DNA at the replication fork.

**PCR:** *See* polymerase chain reaction.

**PDB:** *See* Protein Data Bank (PDB).

**peptide bond:** A substituted amide linkage between the  $\alpha$ -amino group of one amino acid and the  $\alpha$ -carboxyl group of another, with the elimination of the elements of water.

**peptide prolyl cis-trans isomerase:** Enzyme that catalyzes the interconversion of the cis and trans isomers of proline peptide bonds.

**peptide translocation complex:** A complex in the endoplasmic reticulum (ER) that catalyzes the translocation into the ER lumen of a growing polypeptide chain containing an N-terminal signal sequence.

**peptidyl transferase reaction:** The reaction that synthesizes the peptide bonds of proteins—nucleophilic

attack of the  $\alpha$ -amino group of the ribosomal A-site aminoacyl-tRNA on the carbonyl carbon of the ester bond linking the fMet (or the growing peptide chain) to the P-site tRNA. The reaction is catalyzed by a ribozyme, part of the rRNA of the large ribosomal subunit.

**P generation:** The parental generation in a genetic cross.

**pH:** The negative logarithm of the hydronium ion concentration of an aqueous solution.

**phase problem:** The problem of determining the phase of the reflections in an x-ray crystallography experiment.

**phase variation:** The expression of alternative primary cell-surface antigens employed by some pathogenic bacteria and parasitic protists as a means of eluding the host immune system.

**phenotype:** The observable characteristics of an organism.

**phenotypic function (of a gene product):** The effect of a gene product on the entire organism. *Compare* cellular function and molecular function.

**phosphatases:** Enzymes that hydrolyze a phosphate ester or anhydride, releasing inorganic phosphate,  $P_i$ .

**phosphodiester bond:** A chemical grouping that contains two alcohols esterified to one molecule of phosphoric acid, which thus serves as a bridge between them.

**photoreactivation:** The repair of a cyclobutane pyrimidine dimer by electron transfer from a DNA photolyase.

**phylogenetic profiling:** Bioinformatic technique used to discover structure/function relationships by searching for genes that consistently appear together across many genomes.

**phylogenetics:** The study of the evolutionary relationships among organisms.

**phylogeny:** The evolutionary relationships among organisms.

**pKa:** The negative logarithm of an acid dissociation constant.

**plasmid:** An extrachromosomal, independently replicating, small circular DNA molecule; commonly employed in genetic engineering.

**plectonemic supercoiling:** A structure in a molecular polymer that has a net twisting of strands about each other in some simple and regular way.

**pluripotent:** Describes stem cells that can differentiate into cells derived from any of the three germ layers.

**pOH:** The negative logarithm of the hydroxyl ion concentration of an aqueous solution.

**point mutation:** A mutation consisting of a single base pair change.

**Pol I:** *See* RNA polymerase I.

**Pol II:** *See* RNA polymerase II.

**Pol III core:** A complex of the  $\alpha$ ,  $\epsilon$ , and  $\theta$  subunits of the *E. coli* DNA polymerase III with polymerase and 5'→3' proofreading exonuclease activities.

**Pol III holoenzyme:** The 17-subunit *E. coli* DNA polymerase III complex, responsible for chromosomal replication. It includes two Pol III, two sliding clamps, and a clamp-loading complex. *Compare* Pol III and RNA polymerase holoenzyme.

**Pol III:** *See* RNA polymerase III.

**polar covalent:** Type of covalent bond between atoms of different electronegativities, such that the electrons are shared unequally between the atoms.

**polar:** Hydrophilic, or “water-loving”; describes molecules or groups that have a dipole moment and are therefore soluble in water.

**polarity:** In a developing embryo, the distinct areas that will become the anterior, posterior, dorsal, and ventral parts of the adult organism.

**poly(A) addition site:** The site where an mRNA is cleaved by a specific endonuclease to generate the free 3'-hydroxyl to which A residues are added. The site is marked by a highly conserved 5'-AAUAAA sequence 10 to 30 nucleotides on the 5' side, and a G- and U-rich region 20 to 40 nucleotides on the 3' side.

**poly(A) site choice:** The existence of more than one site in an mRNA that may be cleaved to generate the free 3'-hydroxyl to which A residues are added, which can generate diverse transcripts from a single gene.

**poly(A) tail:** *See* 3' poly(A) tail.

**polycistronic mRNA:** A contiguous mRNA with more than two genes that can be translated into proteins.

**polyglutamine (polyQ) disease:** A triplet expansion disease caused by the insertion of many additional glutamine codons in a gene. Fragile X syndrome and Huntington disease are polyglutamine diseases.

**polylinker:** A short, synthetic fragment of DNA containing recognition sequences for several restriction endonucleases that is inserted into a cloning vector.

**polymerase chain reaction (PCR):** A repetitive laboratory procedure that results in a geometric amplification of a specific DNA sequence.

**polynucleotide:** A covalently linked sequence of nucleotides in which the 3' hydroxyl of the pentose of one nucleotide residue is joined by a phosphodiester bond to the 5' hydroxyl of the pentose of the next residue.

**polypeptide:** A long chain of amino acids linked by peptide bonds; the molecular weight is generally less than 10,000.

**polyQ disease:** *See* polyglutamine disease.

**polyribosome:** *See* polysome.

**polysome:** A complex of an mRNA molecule and two or more ribosomes; also called polyribosome.

**positive regulation:** Increased expression of a gene by the binding of an activator protein. *Compare* negative regulation.

**positive supercoiling:** The twisting of a helical (coiled) molecule on itself to form a left-handed supercoil.

**postinsertion site:** Site within the active site of a DNA polymerase where the primer 3'-terminal base pair is positioned. *Compare* insertion site.

**posttranslational modification:** Enzymatic processing of a polypeptide chain after translation from its mRNA.

**postulate of objectivity:** Refers to the only assumption made by scientists—that basic forces and laws in the universe are not subject to change. They can thus be studied and defined by scientific inquiry. The term was coined by Jacques Monod.

**precursor miRNA (pre-miRNA):** A partially processed RNA intermediate that is transported from the nucleus to the cytoplasm for final processing by the endonuclease Dicer into a miRNA.

**preinitiation complex:** A eukaryotic nucleoprotein complex consisting of intact, double-stranded promoter DNA, Pol II, and various transcription factors. *Compare* closed complex (in bacteria).

**pre-miRNA:** *See* precursor miRNA.

**prepriming complex:** The complex of proteins assembled at *oriC* in *E. coli* at an early stage of replication fork assembly. The complex includes a DnaA oligomer bound to the DNA and DnaB helicases stabilizing the single strands of DNA in the replication “bubble.”

**preRC:** *See* prereplication complex.

**prereplication complex (preRC):** The complex of proteins, including the origin recognition complex (ORC) and MCMs, that assembles during the G1 phase of the eukaryotic cell cycle, thereby marking the origin for replication during S phase.

**preribosomal RNA (pre-rRNA):** Primary transcript of ribosomal RNAs in bacterial and eukaryotic cells, which is processed into mature ribosomal RNAs (and transfer RNAs in bacteria).

**pre-rRNA:** *See* preribosomal RNA.

**pre-steady state:** The time immediately after an enzyme is mixed with its substrate, before the free enzyme and its intermediates have reached their steady-state concentrations.

**primary miRNA transcript (pri-miRNA):** An RNA transcript that can fold into an extensive hairpin structure, which becomes a substrate for cleavage by the nuclear endonuclease Drosha into a smaller hairpin structure called pre-miRNA.

**primary structure:** A description of the covalent backbone of a polymer (macromolecule), including the sequence of monomeric subunits and any interchain and intrachain covalent bonds.

**primary transcript:** The immediate RNA product of transcription before any posttranscriptional processing reactions.

**primase:** An enzyme that catalyzes the formation of RNA oligonucleotides used as primers by DNA polymerases.

**primed template:** A template nucleic acid strand annealed to an RNA or DNA primer.

**primer strand:** A strand of nucleic acid with a free 3'-OH group to which a polymerase can add nucleotides.

**primer terminus:** The end of a primer to which monomeric subunits are added.

**pri-miRNA:** See primary miRNA transcript.

**prion:** Misfolded protein in the nervous tissue of mammals that acts as an infectious agent, causing other proteins to misfold and accumulate, leading to the development of spongiform encephalopathy.

**probe:** A labeled fragment of nucleic acid containing a nucleotide sequence complementary to a genomic sequence that one wishes to detect in a hybridization experiment.

**processing body (P body):** Area in the cytoplasm of a eukaryotic cell where mRNAs that are not being translated are sequestered, possibly for degradation.

**processive synthesis:** enzymatic synthesis of a biological polymer in which the enzyme adds multiple subunits without dissociating from the substrate. *Compare* distributive synthesis.

**processivity:** For any enzyme that catalyzes the synthesis of a biological polymer, the property of adding multiple subunits to the polymer without dissociating from the substrate.

**product:** Molecule formed in a chemical reaction.

**proenzyme:** Precursor form of an enzyme before it is cleaved into its active form.

**promoter clearance:** Movement of the transcription complex away from the promoter, which marks the beginning of the elongation stage of transcription.

**promoter:** A DNA sequence at which RNA polymerase may bind, leading to initiation of transcription.

**proofreading:** The correction of errors in the synthesis of an information-containing biopolymer by removing incorrect monomeric subunits after they have been covalently added to the growing polymer.

**prophage induction:** The process of a lysogen switching from lysogenic growth to lytic growth.

**prophage:** A bacteriophage genome incorporated into the host DNA or as an autonomously replicating plasmid, with most of its genes repressed; a lysogenized bacteriophage genome.

**prophase:** The first stage of mitosis (M phase). Chromosomes duplicated in S phase begin to condense and become visible in the light microscope.

**proprotein:** Precursor form of a protein before it is cleaved into its functional form.

**prosthetic group:** A metal ion or an organic compound (other than an amino acid) that is covalently bound to a protein and is essential to its activity.

**proteasome:** Large assembly of enzymatic complexes that function in the degradation of damaged or unneeded cellular proteins. Also called 26S proteasome in eukaryotes.

**Protein Data Bank (PDB):** An international database ([www.rcsb.org/pdb](http://www.rcsb.org/pdb)) that archives the data describing the three-dimensional structure of nearly all macromolecules for which structures have been published.

**protein disulfide isomerase:** Enzyme that catalyzes the breakage and formation of disulfide cross-links in a protein.

**protein family:** A group of evolutionarily related proteins with similar primary sequence and function.

**protein folding:** The process by which a polypeptide chain attains its biologically active conformation.

**protein kinases:** Enzymes that transfer the terminal phosphoryl group of ATP or another nucleoside triphosphate to a Ser, Thr, Tyr, Asp, or His side chain in a target protein, thereby regulating the activity or other properties of that protein.

**protein phosphatases:** Enzymes that hydrolyze a phosphate ester or anhydride on specific amino acid residues of protein, releasing inorganic phosphate, Pi.

**proteolytic cleavage:** The enzyme-catalyzed breakage of peptide bonds in proteins.

**proteome:** The full complement of proteins expressed in a given cell under a given set of conditions, or the complete complement of proteins that can be expressed by a given genome.

**proteomics:** Broadly, the study of the protein complement of a cell or organism.

**protomer:** A general term describing any repeated unit of one or more stably associated protein subunits in a larger protein structure. If a protomer has multiple subunits, the subunits may be identical or different.

**P site:** The site in a ribosome occupied by the peptidyl-tRNA.

**PUF family:** In eukaryotic cells, a family of proteins that bind the 3' untranslated region of mRNAs to suppress their translation.

**pulsed field gel electrophoresis:** Variation on the technique of gel electrophoresis in which the direction of the current passed through the gel is altered at regular intervals. This allows the separation of larger molecules of DNA than is possible with conventional gel electrophoresis. *See* gel electrophoresis.

**Punnett square:** Matrix for displaying the genes involved in a cross and the possible combinations of the alleles in the progeny. The gamete genotypes are written along the top and sides of the square; the possible combinations of the alleles are shown in the matrix.

**purebred:** Describes an individual homozygous for a given trait or set of traits.

**purine:** A nitrogenous heterocyclic base found in nucleotides and nucleic acids; contains fused pyrimidine and imidazole rings.

**puromycin:** An antibiotic that inhibits polypeptide synthesis by being incorporated into a growing polypeptide chain, causing its premature termination.

**pyrimidine:** A nitrogenous heterocyclic base found in nucleotides and nucleic acids.

**pyrimidine dimer:** A covalently joined dimer of two adjacent pyrimidine residues in DNA, induced by absorption of UV light; most commonly derived from two adjacent thymines (a thymine dimer).

**pyrophosphorolysis:** The reverse of a nucleotide polymerization reaction, in which pyrophosphate reacts with the 3'-nucleotide monophosphate of an oligonucleotide, releasing the corresponding nucleotide triphosphate.

**qPCR:** Quantitative PCR; *see* real-time PCR.

**quantitative PCR (qPCR):** *See* real-time PCR.

**quaternary structure:** The three-dimensional structure of a multisubunit protein, particularly the manner in which the subunits fit together.

**quorum sensing:** The regulation of gene expression in response to fluctuations in cell-population density, assessed by the detection of small, diffusible signaling molecules secreted by the cells.

**Rad51:** Eukaryotic recombinase structurally and functionally homologous to the RecA protein of *E. coli*. *Also see* Dmc1.

**Ramachandran plot:** A graphical representation of the allowed values of the  $\phi$  and  $\psi$  angles of the amino acid residues in a polypeptide.

**rate constant:** The proportionality constant that relates the velocity of a chemical reaction to the concentration(s) of the reactant(s).

**rate-limiting step:** (1) Generally, the step in an enzymatic reaction with the greatest activation energy or the transition state of highest free energy. (2) The slowest step in a metabolic pathway.

**RC:** *See* replication complex.

**reactant:** Starting material in a chemical reaction.

**reaction intermediate:** Any chemical species in a reaction pathway that has a finite chemical lifetime.

**reaction kinetics:** *See* kinetics.

**reaction mechanism:** The sequence of individual steps that take place during the conversion of reactants to products in a chemical reaction.

**reading frame:** A contiguous, nonoverlapping set of three-nucleotide codons in DNA or RNA.

**real-time PCR (quantitative PCR, qPCR):** Polymerase chain reaction (PCR) protocol that allows the simultaneous amplification and detection of a sequence using a fluorescent probe. *See* polymerase chain reaction.

**RecA protein:** A non-site-specific bacterial recombinase that binds single-stranded DNA and promotes homologous recombination. RecA protein also has co-protease activity in the autocatalytic activity of some transcription repressors.

**RecBCD:** A bacterial protein complex that prepares DNA at a double-strand break for repair. The complex has helicase activity to unwind the DNA, an endonuclease activity that creates 3' single-stranded overhangs, and an activity that loads RecA protein on the 3'-ending single strand.

**recessive:** Describes an allele that manifests in a homozygous individual but is masked by the dominant allele in heterozygotes. *Compare* dominant.

**RecFOR:** A bacterial recombination mediator that loads RecA protein on single strand gaps in need of repair.

**recognition helix:** The  $\alpha$  helix in a DNA regulatory protein that recognizes and binds to the DNA regulatory site.

**recognition sequence:** Specific nucleotide sequence in a double-stranded DNA molecule that is recognized by a restriction endonuclease as a substrate.

**recombinant DNA technology:** Laboratory methods used for genetic engineering.

**recombinant DNA:** DNA formed by joining DNA molecules, usually from different species, in new combinations.

**recombinase:** An enzyme that catalyzes genetic recombination by the reciprocal exchange of short pieces of DNA between longer DNA molecules.

**recombination mapping:** The process of determining the relative distance between genes on a chromosome based on the frequency of recombination of alleles during meiosis.

**recombination:** The reciprocal exchange of alleles between chromosomes; also called crossing over. At the molecular level, any enzymatic process by which the linear arrangement of nucleic acid sequences in a chromosome or plasmid is altered by cleavage and rejoining.

**recombinational DNA repair:** Recombinational processes directed at the repair of DNA strand breaks or cross-links, especially at inactivated replication forks.

**refinement:** In x-ray crystallography, an iterative stage in which the computed diffraction pattern of a predicted model of a three-dimensional structure is compared with the diffraction pattern obtained from the crystal in an experiment.

**reflection spot:** In x-ray crystallography, area on a film or x-ray detector created by the constructive interference of x rays diffracted from the atoms in a unit cell of a crystal.

**regulated gene expression:** The conditional expression of a gene based on the cellular need for the gene product, and achieved by the presence or absence of activators,

repressors, enhancers, and other regulatory factors. *Compare* constitutive gene expression.

**regulatory enzyme:** An enzyme with a regulatory function, through its capacity to undergo a change in catalytic activity by allosteric mechanisms or by covalent modification.

**regulatory sequence:** A DNA sequence involved in regulating the expression of a gene; for example, a promoter or operator. Also called a regulatory site.

**regulatory site:** *See* regulatory sequence.

**regulon:** A group of genes or operons that are coordinately regulated even though some, or all, may be spatially distant in the chromosome or genome.

**relaxed DNA:** Any DNA that exists in its most stable and unstrained structure, typically the B form under most cellular conditions.

**release factors:** Protein factors required for the release of a completed polypeptide chain from a ribosome; also known as termination factors. *Also see* RF-1, RF-2, and RF-3.

**replicase:** General term describing any polymerase enzyme that duplicates chromosomes. Also called chromosomal replicase.

**replication complex (RC):** A complex of proteins, including the replication fork machinery, that assembles at an origin during the S phase of the eukaryotic cell cycle to replicate the chromosome.

**replication factor C (RFC):** The eukaryotic clamp loading complex.

**replication fork:** The Y-shaped structure generally found at the point where DNA is being synthesized.

**replication protein A (RPA):** A eukaryotic single-stranded DNA-binding protein; its bacterial homolog is SSB.

**replicon:** The length of DNA replicated from a single origin.

**replosome:** The multiprotein complex that promotes DNA synthesis at the replication fork.

**repression:** A decrease in the expression of a gene in response to a change in the activity of a regulatory protein.

**repressor:** The protein that binds to the regulatory sequence or operator for a gene, blocking its transcription.

**resonance hybrid:** Molecule that exists in an average of two possible resonance forms. *See* resonance.

**resonance:** Conceptual view of the delocalized electrons in the bonding structure of a molecule that can only be described as the average of two or more Lewis structures.

**restriction endonucleases:** Site-specific endodeoxyribonucleases that cleave both strands of DNA at points in or near the specific site recognized by the enzyme; important tools in genetic engineering.

**retrohoming:** The integration of a mobile group II intron into DNA by the reverse splicing of its RNA transcript into a target site (catalyzed by an encoded endonuclease), followed

by DNA synthesis (catalyzed by an encoded reverse transcriptase).

**retrotransposable element:** *See* retrotransposon.

**retrotransposon (retrotransposable element):** Transposon that moves via an RNA intermediate that is converted back to DNA by reverse transcriptase.

**retrovirus:** An RNA virus containing a reverse transcriptase.

**reverse transcriptase PCR (RT-PCR):** Polymerase chain reaction (PCR) protocol for amplifying an RNA sequence by first using reverse transcriptase to create a DNA copy. *See* polymerase chain reaction and reverse transcriptase.

**reverse transcriptase:** An RNA-directed DNA polymerase in retroviruses; capable of making DNA complementary to an RNA.

**reverse turn:** Segment of a polypeptide chain in a folded protein or domain that connects two regions of secondary structure whose N-to-C directions are reversed.

**reversible inhibition:** Inhibition by a molecule that binds reversibly to the enzyme, such that the enzyme activity returns when the inhibitor is no longer present.

**reversion mutation:** A mutation in a gene that reverses a previous mutation; also called a back mutation. A true reversion restores the original gene sequence; a second-site reversion restores the functionality (phenotype).

**R factor:** The residual error in a model of the three-dimensional structure of a molecule obtained by x-ray crystallography. The R factor is the difference between the calculated diffraction pattern based on the model and the actual diffraction pattern obtained from the crystal.

**RF-1:** A bacterial class I release factor that recognizes the stop codons UAG and UAA and induces peptidyl transferase to transfer the growing polypeptide to water.

**RF-2:** A bacterial class I release factor that recognizes the stop codons UGA and UAA and induces peptidyl transferase to transfer the growing polypeptide to water.

**RF-3:** A bacterial class II release factor with GTPase activity that catalyzes the dissociation of RF-1 and RF-2 from the ribosome.

**RFC:** *See* replication factor C.

**R group:** (1) Formally, an abbreviation denoting any alkyl group. (2) Occasionally, used in a more general sense to denote virtually any organic substituent (the R groups of amino acids, for example).

**ribonuclease:** A nuclease that catalyzes the hydrolysis of certain internucleotide linkages of RNA.

**ribonucleic acid:** *See* RNA.

**ribonucleoprotein (RNP):** A molecular complex of RNA and protein, such as the ribosome.

**ribonucleoside 2'-monophosphates:** Metabolites produced during the enzymatic or alkaline hydrolysis of RNA.

**ribonucleoside 2',3'-cyclic monophosphates:** Metabolites produced during the enzymatic or alkaline hydrolysis of RNA.

**ribonucleoside 3'-monophosphates:** Metabolites produced during the enzymatic or alkaline hydrolysis of RNA.

**ribonucleotide:** A nucleotide containing D-ribose as its pentose component.

**ribosomal protein (r-protein):** The proteins serving as components of ribosomes.

**ribosomal RNA (rRNA):** A class of RNA molecules serving as components of ribosomes.

**ribosome:** A macromolecular complex of rRNAs and r-proteins; the site of protein synthesis.

**ribosome binding site:** A sequence in an mRNA that is required for binding bacterial ribosomes. Also called the Shine-Dalgarno sequence.

**ribosome recycling:** The disassembly of a translated mRNA, deacylated tRNAs, and ribosomal subunits in preparation for new rounds of translation.

**ribosome recycling factor (RRF):** A bacterial factor involved in ribosome recycling that binds to the empty ribosomal A site and recruits EF-G to stimulate release of the deacylated tRNAs in the P and E sites.

**riboswitch:** A structured segment of an mRNA that binds to a specific ligand and affects the translation or processing of the mRNA.

**ribozyme (catalytic RNA):** Ribonucleic acid molecule with catalytic activity; RNA enzyme.

**rifampicin:** An antibiotic that acts as a bacterial transcription inhibitor by binding to the  $\beta$  subunit of bacterial RNA polymerase, preventing promoter clearance.

**RISC:** See RNA-induced silencing complex.

**RNA (ribonucleic acid):** A polyribonucleotide of a specific sequence linked by successive 3',5'-phosphodiester bonds.

**RNA degradation:** The complete hydrolysis of RNA into its component ribonucleotides, usually catalyzed by ribonucleases or the exosome.

**RNA editing:** Posttranscriptional modification of an mRNA that alters one or more codons during translation.

**RNA interference (RNAi):** Methods of gene silencing, mediated by short interfering RNAs (siRNAs) or microRNAs (miRNAs), which bind to and silence the mRNA transcript, usually by targeting it for degradation. MicroRNAs are endogenous to the cell, produced by nucleases from longer transcripts encoded in the genome. siRNAs are generated by the same cellular enzymes from exogenous double-stranded RNA, introduced into the cell by viral infection or experimental manipulation.

**RNA polymerase:** An enzyme that catalyzes the formation of RNA from ribonucleoside 5'-triphosphates, using a strand of DNA as a template. Some RNA polymerases, primarily in viruses, use RNA as a template and are called RNA-dependent RNA polymerases (RDRPs).

**RNA polymerase core:** The *E. coli* RNA polymerase complex of subunits  $\alpha_2\beta\beta'\omega$ , exclusive of the  $\sigma$  subunit. Compare Pol III core.

**RNA polymerase holoenzyme:** The complete *E. coli* RNA polymerase complex, including the  $\alpha_2\beta\beta'\omega$  core and the  $\sigma$  subunit. Compare Pol III holoenzyme.

**RNA polymerase I (Pol I):** One of the three eukaryotic RNA polymerases; Pol I transcribes genes encoding large rRNA precursors.

**RNA polymerase II (Pol II):** One of the three eukaryotic RNA polymerases; Pol II transcribes most of the protein-coding genes.

**RNA polymerase III (Pol III):** One of the three eukaryotic RNA polymerases; Pol III transcribes genes encoding tRNAs, some snRNAs, 5S ribosomal RNA, and other small functional RNAs.

**RNA secondary structure:** The local spatial arrangement of an RNA strand, including a description of any intrachain base-pairing.

**RNA splicing:** Removal of introns and joining of exons in a primary transcript.

**RNA world hypothesis:** Hypothesis that in an early stage of evolution, a living system was based on RNA. In this system, RNA enzymes could catalyze the synthesis of all the molecules required for life from simpler molecules available in the environment.

**RNAi:** See RNA interference.

**RNA-induced silencing complex (RISC):** A cytoplasmic complex, including the endonucleases Dicer and Argonaute, that incorporates an miRNA or an siRNA, delivers it to its complementary mRNA target, and then cleaves the mRNA.

**RNaseH:** An endonuclease that cleaves the 3'-O-P bond of RNA in an RNA/DNA duplex.

**RNA-Seq:** Strategy for determining the level of expression of all the genes in a cell or tissue by sequencing the transcriptome.

**RNP:** See ribonucleoprotein.

**Rossman fold:** Protein supersecondary structure with alternating  $\beta$  strands and  $\alpha$  helices ( $\beta\text{-}\alpha\text{-}\beta\text{-}\alpha\text{-}\beta$  motif) common in nucleotide-binding proteins.

**RPA:** See replication protein A.

**r-protein:** See ribosomal protein.

**RRF:** See ribosome recycling factor.

**rRNA:** See ribosomal RNA.

**RT-PCR:** See reverse transcriptase-PCR.

**Sanger sequencing method:** Method developed by Fred Sanger for determining the base sequence of a DNA molecule. Uses dideoxynucleotides to terminate synthesis; products are analyzed on a gel.

**scanning:** The process by which a partially assembled eukaryotic initiation complex slides along the mRNA until it comes to a start codon.

**SCOP:** See Structural Classification of Proteins (SCOP) database.

**screenable marker:** A gene introduced into a cell that allows any colony expressing it to be readily identifiable by its color or fluorescence. *Compare* selectable marker.

**SDSA:** See synthesis-dependent strand annealing.

**SDS-PAGE:** See sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE).

**second law of thermodynamics:** The law stating that, in any chemical or physical process, the entropy of the universe tends to increase.

**second messenger:** An effector molecule synthesized in a cell in response to an external signal (first messenger) such as a hormone.

**secondary structure:** The local spatial arrangement of the main-chain atoms in a segment of a polypeptide chain; also applied to polynucleotide structure. *See* RNA secondary structure.

**segmentation gene:** A group of genes involved in pattern formation in the *Drosophila* embryo. Segmentation genes are divided into three subclasses. *Also see* gap gene, pair-rule gene, and segment polarity gene.

**segment polarity gene:** A subclass of the segmentation genes expressed after the pair-rule genes that determine the anterior-posterior polarity of the developing *Drosophila* embryo.

**selectable marker:** A gene introduced into a cell that either permits the growth of the cell or kills it under a defined set of conditions. *Compare* screenable marker.

**semiconservative:** Mode of DNA replication in which the daughter duplex has one intact parental strand and one newly synthesized strand.

**semidiscontinuous:** Mode of DNA replication in which one strand, the leading strand, is replicated continuously, but the opposite strand, the lagging strand, is replicated in shorter, discontinuous segments.

**sequence polymorphisms:** Any alterations in genomic sequence (base-pair changes, insertions, deletions, rearrangements) that help distinguish subsets of individuals in a population or distinguish one species from another.

**sequence-tagged site (STS):** Any known sequence that has been mapped in a chromosome and/or clones derived from it.

**sex chromosome:** Chromosome that determines the male or female sex of an organism. *Compare* autosome.

**Shine-Dalgarno sequence:** A sequence in an mRNA that is required for binding bacterial ribosomes. Also called the ribosome binding site (RBS).

**short interfering RNA (siRNA):** A short (~21- to 27-nucleotide) double-stranded RNA with 3' overhangs, created from exogenous double-stranded RNA by the endonuclease Dicer, that participates in the RNAi gene silencing pathway.

**short tandem repeat (STR):** A short (typically 3 to 6 bp) DNA sequence, repeated multiple times in tandem at a particular location in a chromosome.

**shuttle vector:** A recombinant DNA vector that can be replicated in two or more different host species. *Also see* cloning vector.

**sigma factor ( $\sigma$ ):** A transient subunit of the bacterial RNA polymerase that directs the enzyme to the promoter. Different sigma factors are specific for different promoters. The core RNA polymerase plus the sigma factor constitutes the RNA polymerase holoenzyme.

**signal integration:** The control of gene expression by the net effect of multiple, sometimes conflicting, regulatory signals.

**signal recognition particle (SRP):** A protein-RNA complex that recognizes and binds the signal sequence in a nascent polypeptide, and delivers the ribosome to a peptide translocation complex in the endoplasmic reticulum.

**signal sequence:** An amino acid sequence, often at the amino terminus, that signals the cellular fate or destination of a newly synthesized protein.

**signaling release:** Release of the *E. coli* DNA polymerase III from the  $\beta$  sliding clamp signaled by priming events on the lagging strand. *Also see* collision release.

**silent mutation:** A mutation in a gene that causes no detectable change in the peptide sequence of the gene product.

**simple-sequence repeats (SSRs):** Highly repeated, nontranslated segments of DNA in eukaryotic chromosomes, often associated with the centromere and telomere, but not restricted to these regions. Its function is unknown.

**single bond:** bond between two elements that involves two electrons.

**single nucleotide polymorphism (SNP):** A genomic base-pair change that helps distinguish one species from another or one subset of individuals in a population.

**single-stranded DNA-binding protein (SSB):** A bacterial protein that binds single-stranded DNA in a sequence-independent fashion.

**siRNA:** *See* short interfering RNA.

**sister chromatid pair:** Duplicate chromosomes, attached via the centromere, produced during the S phase of the eukaryotic cell cycle. Sister chromatids separate into individual chromosomes during anaphase of mitosis or anaphase II of meiosis, when the centromere divides.

**site-directed mutagenesis:** A set of methods used to create specific alterations in the sequence of a gene.

**site-specific recombination:** A type of genetic recombination that occurs only at specific sequences.

**6-4 photoproduct:** A pyrimidine dimer in which the C-6 and C-4 atoms of adjacent pyrimidines are linked.

**small nuclear ribonucleoprotein (snRNP):** A protein and snRNA complex, found in the nucleus and a component of the spliceosome.

**small nuclear RNA (snRNA):** A class of short, noncoding RNAs, typically 100 to 200 nucleotides long, found in the nucleus and involved in the splicing of eukaryotic mRNAs.

**small nucleolar ribonucleoprotein (snoRNP):** A protein and snoRNA complex that guides the modification of rRNAs in the nucleolus.

**small nucleolar RNA (snoRNA):** A class of short, noncoding RNAs, generally 60 to 300 nucleotides long, found in the nucleolus and involved in the modification of rRNAs.

**SMC proteins:** Family of ATPases that modulate the structure and organization of chromosomes.

**snoRNA:** See small nucleolar RNA.

**snoRNP:** See small nucleolar ribonucleoprotein.

**SNP:** See single nucleotide polymorphism.

**snRNA:** See small nuclear RNA.

**snRNP:** See small nuclear ribonucleoprotein.

**sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE):** Type of electrophoresis used to separate proteins on the basis of size. Proteins in experimental samples are denatured and bound along their length by negatively charged molecules of the detergent SDS, giving them a charge proportional to their length. They move through a gel matrix of cross-linked polyacrylamide under an electric field at a rate proportional to their size. See gel electrophoresis.

**solenoid model:** Model for the arrangement of nucleosomes in the 30 nm filament in which the nucleosome array assumes a spiral shape with the flat sides of adjacent nucleosomes next to each other. Compare zigzag model.

**solenoidal supercoiling:** The wrapping of a helical molecule to form a coiled superstructure.

**somatic cells:** All cells in a multicellular organism except the germline cells.

**SOS response:** In bacteria, a coordinated induction of a variety of genes in response to high levels of DNA damage.

**Southern blot:** A DNA hybridization procedure in which one or more specific DNA fragments are detected in a larger population by hybridization to a complementary, labeled nucleic acid probe.

**S phase:** The phase of the cell cycle during which the DNA is replicated. The S phase occurs between the G<sub>1</sub> and G<sub>2</sub> phases.

**splice site:** A nucleotide sequence within an intron at the intron-exon border, where a primary mRNA transcript may be spliced.

**spliceosome:** A ribonucleoprotein complex that splices mRNAs in eukaryotic cells.

**Spo11:** A eukaryotic protein that creates double strand breaks in the DNA in meiotic prophase I.

**spongiform encephalopathies:** Transmissible, progressive neurodegenerative diseases caused by prions.

**SRP:** See signal recognition particle.

**SSB:** See single-stranded DNA-binding protein.

**SSR:** See simple-sequence repeats.

**standard (Gibbs) free-energy change ( $\Delta G^\circ$ ):** The free-energy change for a reaction occurring under a set of standard conditions: temperature, 298 K; pressure, 1 atm or 101.3 kPa; and all solutes at 1 M concentration.  $\Delta G^\circ$  denotes the standard free-energy change at pH 7.0 in 55.5 M water.

**start codon:** See initiation codon.

**steady state:** A nonequilibrium state of a system through which matter is flowing and in which all components remain at a constant concentration.

**steady-state kinetics:** The study of reaction rates under steady-state conditions. See steady state.

**stem cells:** The cells in multicellular organisms that retain the ability to divide and differentiate into other cell types.

**step size:** the average number of subunits over which a motor protein moves for each ATP molecule hydrolyzed.

**stereochemistry:** The spatial arrangement of the atoms within a molecule.

**stereoisomers:** Compounds that have the same composition and the same order of atomic connections but different molecular arrangements.

**sticky ends:** Two DNA ends in the same DNA molecule, or in different molecules, with short overhanging single-stranded segments that are complementary to one another, facilitating ligation of the ends.

**stop codons:** See termination codons.

**STR:** See short tandem repeat.

**streptomycin:** An aminoglycoside antibiotic that disrupts or inhibits bacterial protein synthesis. At low concentrations it causes misreading of the genetic code; at higher concentrations it inhibits translation initiation by preventing fMet-tRNA<sup>fMet</sup> from binding to the ribosome.

**stringent factor:** A bacterial protein (RelA) recruited to a ribosome when an uncharged tRNA binds. Stringent factor catalyzes formation of guanosine tetraphosphate (ppGpp), which binds RNA polymerase, reducing transcription from rRNA and tRNA genes and increasing transcription from biosynthetic genes.

**stringent response:** A mechanism for coordinating transcriptional activity in bacteria with the levels of amino acids available in the cell. Triggered by the binding of uncharged tRNAs to the ribosome, the stringent response directs the cellular machinery toward amino acid synthesis rather than growth and reproduction.

**Structural Classification of Proteins (SCOP)**

**database:** A database (<http://scop.berkeley.edu>) of the structural and evolutionary relationships between all proteins whose structure is known.

**STS:** See sequence-tagged site.

**substrate:** A molecule that undergoes an enzyme-catalyzed reaction.

**supercoiled DNA:** DNA that twists upon itself because it is underwound or overwound (and thereby strained) relative to B-form DNA.

**superfamily:** Structural classification that includes protein families with little sequence similarity but with the same supersecondary structural motif and functional similarities.

**superhelical density ( $\sigma$ ):** In a helical molecule such as DNA, the number of supercoils (superhelical turns) relative to the number of coils (turns) in the relaxed molecule.

**supersecondary structure:** See motif.

**suppressor tRNA:** A mutant tRNA that binds to a termination codon but carries an amino acyl residue that can be incorporated into the growing amino acid chain, suppressing the termination signal.

**surroundings:** All matter in the universe that is outside of the system being considered. See system.

**synteny:** Conserved gene order along the chromosomes of different species.

**synthesis-dependent strand annealing (SDSA):** A pathway for repairing double-strand breaks that ends with the invading strands dissociating and reannealing, and the homologous DNA molecule intact. *Also see* double-strand break repair.

**system:** An isolated collection of matter; all other matter in the universe apart from the system is called the surroundings. See surroundings.

**systems biology:** In biochemistry or molecular biology, the study of complex systems, integrating the functions of the macromolecules in a cell (RNA, DNA, proteins).

**tag:** A peptide or protein that binds a simple, stable ligand with high affinity and specificity.

**tandem affinity purification (TAP) tags:** Two tags (such as Protein A and calmodulin-binding peptide) fused to the same target protein to allow for sequential highly specific affinity purification steps of the expressed protein.

**TAP:** See tandem affinity purification tags.

**target site:** The location on a chromosome where a transposon inserts itself. *Compare* donor site.

**target-primed (TP) retrotransposon:** A retrotransposon that moves via a cDNA copy of its mRNA transcript, and is synthesized as a direct extension of the 3' end created in the target site by a retrotransposon-encoded endonuclease. *Also see* extrachromosomally primed (EP) retrotransposon.

**TATA-binding protein (TBP):** Eukaryotic transcription factor that binds all three RNA polymerases as well as the AT-rich region of many promoters known as the TATA box.

**taxon:** A grouping of organisms in a phylogenetic tree.

**TBP:** See TATA-binding protein.

**telomerase:** A ribonucleoprotein enzyme that has reverse transcriptase activity and a short RNA sequence that primes the addition of nucleotide repeats to the 3' ends of DNA. Telomerase activity ensures that no unique sequence information is lost as a result of the end replication problem.

**telomere:** Specialized nucleic acid structure found at the ends of linear eukaryotic chromosomes.

**telomere loop:** See t-loop.

**telophase:** The final stage of mitosis (M phase) in which the two sets of homologous chromosomes reach opposite spindle poles and begin to decondense. The cell physically divides in the process of cytokinesis, resulting in two daughter cells.

**template strand:** A strand of nucleic acid used by a polymerase as a template to synthesize a complementary strand.

**Ter site:** A 23 bp sequence that serves as a DNA replication termination site in *E. coli*.

**termination codons:** UAA, UAG, and UGA; in protein synthesis, these codons signal the termination of a polypeptide chain. Also known as stop codons.

**termination factors:** See release factors.

**termination sequence:** A DNA sequence, at the end of a transcriptional unit, that signals the end of transcription.

**termination:** (1) The third of three stages of RNA synthesis in which the RNA polymerase and the RNA product are released from the DNA template. (2) The third of three stages of protein synthesis in which the ribosome and the peptide product are released from the mRNA template.

**terminator:** In its broadest definition, a terminator is a place where transcription is halted. This can occur at the end of a gene (where some termination sequences are also called terminators) or in regulatory sequences preceding some operons (as occurs in transcription attenuation). See termination sequence.

**tertiary structure:** The three-dimensional conformation of a polymer in its native folded state.

**testcross:** Genetic cross of an F<sub>1</sub> hybrid with a homozygous recessive strain to determine the genotype of the F<sub>1</sub> individual. A testcross also reveals linked genes.

**tetracyclines:** A family of antibiotics that inhibit bacterial protein synthesis by occupying the ribosomal A site, thereby preventing the binding of aminoacyl-tRNAs.

**tetrad:** Structure formed in meiotic prophase I by the association of two homologous sister chromatid pairs.

**thin layer chromatography:** A process in which complex mixtures of molecules are separated by many repeated partitionings between a flowing (mobile) phase and a stationary phase coated onto a planar surface. Molecules separate based on the rate of their migration across the stationary phase as the mobile phase is drawn through it by capillary action.

**30 nm filament:** A higher-order organization of nucleosomes seen in condensed chromosomes.

**3' poly(A) tail:** A length of adenosine residues (typically 80 to 250) added to the 3' end of many mRNAs in eukaryotes (and sometimes in bacteria), which serves as a binding site for proteins that protect the mRNA from exonucleases.

**thymine (T):** A pyrimidine base that is a component of DNA, but not RNA.

**t-loop (telomere loop):** Looped structure seen in mammalian telomeres where the single-stranded 3' end of the chromosome folds back and hybridizes to a duplex portion of the telomere.

**TLS:** *See* translesion synthesis.

**T<sub>m</sub>:** *See* melting point.

**tmRNA:** A bacterial RNA that has the properties of a tRNA at its 5' end and the properties of an mRNA, including a stop codon, at its 3' end. When aminoacylated, the 5' end can bind in the A site of a ribosome stalled on a truncated mRNA, and the 3' end can serve as a template for continued translation through a termination codon that recruits the termination factors required for proper termination and ribosome recycling. Also called transfer-messenger RNA.

**topoisomerase:** An enzyme that introduces positive or negative supercoils in closed, circular duplex DNA.

**topoisomers:** Different forms of a covalently closed, circular DNA molecule that differ only in their linking number.

**totipotent:** Describes stem cells from the first few divisions of the fertilized egg that can differentiate into a complete, viable organism.

**transcription:** The enzymatic process whereby the genetic information contained in one strand of DNA is used to specify a complementary sequence of bases in an RNA strand.

**transcriptional ground state:** The inherent activity of promoters and transcription machinery *in vivo* in the absence of regulatory mechanisms.

**transcription attenuation:** A process for the regulation of expression of a bacterial operon in which transcription begins, but is halted before transcription of the operon genes.

**transcription-coupled repair:** A nucleotide-excision repair pathway in eukaryotes triggered when RNA polymerase encounters a lesion in the DNA and stalls.

**transcription factor:** In eukaryotes, a protein that affects the regulation and transcription initiation of a gene by binding to a regulatory sequence near or within the gene and interacting with RNA polymerase and/or other transcription factors.

**transcriptome:** The entire complement of RNA transcripts present in a given cell or tissue under specific conditions.

**transcriptomics:** The study of transcriptomes.

**transfection:** Incorporation of exogenous DNA into a eukaryotic cellular genome by any of several methods.

**transfer RNA (tRNA):** A class of RNA molecules ( $M_r$  25,000 to 30,000), each of which combines covalently with a specific amino acid for use in protein synthesis.

**transfer-messenger RNA:** *See* tmRNA.

**transformation:** Introduction of an exogenous DNA into a bacterial cell, causing the cell to acquire a new phenotype. Alternatively, the conversion of a cell in a multicellular eukaryote into a cancer cell.

**transition mutation:** A point mutation resulting in the exchange of one purine-pyrimidine base pair for another purine-pyrimidine pair. *See also* transversion mutation.

**transition state:** An activated form of a molecule in which the molecule has undergone a partial chemical reaction; the highest point on the reaction coordinate.

**translation:** The process in which the genetic information present in an mRNA molecule specifies the sequence of amino acids during protein synthesis.

**translational repressor:** A repressor that binds to an mRNA, blocking translation.

**translesion synthesis (TLS):** Pathway for replicating DNA across lesions that occur in unwound DNA at the replication fork. The pathway employs a TLS polymerase that lacks a proofreading exonuclease, and has a less-selective active site. Although this polymerase may introduce a mutation, it allows replication to proceed.

**translocase:** (1) An enzyme that catalyzes membrane transport. (2) An enzyme that causes movement such as the movement of a ribosome along an mRNA or movement along a double-stranded nucleic acid.

**translocation:** (1) Enzyme-catalyzed movement across a membrane. (2) Movement along a double-stranded nucleic acid without strand separation. (3) Movement of a ribosome by one codon along the mRNA.

**translocation mutation:** A mutation that results from the exchange of large segments of DNA between nonhomologous chromosomes.

**transposable element:** *See* transposon.

**transposases:** Transposon-encoded enzymes that catalyze the reactions required for the transposon to excise itself from the donor site and insert itself into the target site. These reactions typically include hydrolysis of a specific phosphodiester bond and transesterification involving attack of the liberated 3' hydroxyl on another phosphodiester bond.

**transposition:** The movement of a gene or set of genes from one site in the genome to another.

**transposon (transposable element):** A segment of DNA that can move from one position in the genome to another.

**trans-splicing:** A process seen in nematode worms in which a short leader sequence is spliced to the 5' end of a primary transcript from a separate RNA molecule.

**transversion mutation:** A point mutation resulting in the exchange of a purine-pyrimidine base pair for a pyrimidine-purine pair, or vice versa. *Also see* transition mutation.

**triplet expansion disease:** A disease caused by the insertion of many additional copies of a repeated codon triplet in a gene due to template slippage in the DNA replication process.

**triplex DNA:** DNA structure involving three polynucleotide strands bonded through non-Watson-Crick interactions called Hoogsteen pairings.

**tRNA:** *See* transfer RNA.

**tRNA-charging step:** The second step in the attachment of an amino acid to a tRNA, in which the aminoacyl-tRNA synthetase transfers the bound aminoacyl-AMP to the 2'-OH or 3'-OH of the 3'-terminal adenosine residue of the tRNA. *Also see* adenylylation step.

**trombone model:** Description of DNA replication on the lagging strand, with its repeated cycles of loop growth and disassembly, by analogy with the movement of a slide on a trombone.

**tumor suppressor gene:** One of a class of genes that encode proteins that normally suppress the division of cells. When defective, the normal gene becomes a tumor suppressor gene, and when both copies are defective, the cell is allowed to continue dividing without limitation; it becomes a tumor cell.

**tunicamycin:** An antibiotic that inhibits the *N*-glycosylation of proteins in eukaryotic cells.

**turnover number:** The number of times an enzyme molecule transforms a substrate molecule per unit time, under conditions giving maximal activity at substrate concentrations that are saturating. *Also see* general rate constant.

**twist (*Tw*):** The net number of helical turns in a DNA molecule. *Compare* writhe (*Wr*).

**two-dimensional gel electrophoresis:** Technique that separates the components of a sample on the basis of two properties, in successive steps. For example, a protein sample may be separated on the basis of isoelectric point in one dimension, followed by separation on the basis of relative molecular mass in the other dimension.

**26S proteasome:** *See* proteasome.

**two-dimensional NMR:** A type of nuclear magnetic resonance spectroscopy in which different pulses of an electromagnetic field provide two qualitatively distinct signals.

**type I topoisomerase:** Enzyme that introduces positive or negative supercoils in closed, circular duplex DNA by cleaving one of the two DNA strands, passing the intact strand through the break, and ligating the broken ends. Type I topoisomerases change *Lk* in increments of 1.

**type II restriction endonucleases:** Enzymes that cleave DNA at a specific site within a short recognition sequence with no requirement for a nucleotide triphosphate cofactor.

**type II topoisomerase:** Enzyme that introduces positive or negative supercoils in closed, circular duplex DNA by cleaving both DNA strands, passing an intact segment of DNA through the break, and religating the broken ends. Type II topoisomerases change *Lk* in increments of 2.

**UAS:** *See* upstream activator sequence.

**uncompetitive inhibitor:** Inhibitor molecule that can bind to the enzyme-substrate complex but not to the free enzyme.

**unipotent:** Describes mammalian cells that can reproduce to form more of the same kind of differentiated cell.

**unit cell:** The smallest regularly repeating unit in a crystal.

**unwinding:** The separation of paired strands of a nucleic acid.

**uORF:** *See* upstream open reading frame.

**UP:** *See* upstream promoter element.

**upstream activator sequence (UAS):** A regulatory sequence in yeast DNA to which transcription activators bind. *Also see* enhancer.

**upstream open reading frames (uORFs):** Short open reading frames upstream of a gene's start codon that serve as decoys to divert ribosomes, thereby down-regulating expression.

**upstream promoter (UP) element:** An AT-rich sequence between positions -40 and -60 in the promoters of some highly expressed bacterial genes. The sequence is bound by an  $\alpha$  subunit of RNA polymerase, improving the efficiency of transcription initiation for that gene.

**uracil (U):** A pyrimidine base that is a component of RNA but not DNA.

**$V_0$ :** *See* initial velocity.

**valence bond model:** Model that proposes that chemical bonds form when half-filled valence atomic orbitals from two atoms overlap.

**valence:** The number of covalent bonds formed by atoms of a particular element.

**van der Waals interaction:** Weak intermolecular forces between molecules as a result of each inducing polarization in the other.

**van der Waals radius:** Half the distance between two atoms of an element that are as close to each other as possible without being formally bonded. The van der Waals radius defines an imaginary sphere that represents the size of an atom in models.

**$V_{\max}$ :** The maximum velocity of an enzymatic reaction when the binding site is saturated with substrate.

**weak chemical bonds:** Chemical interactions such as van der Waals interactions, hydrophobic interactions, and hydrogen bonds that are weaker than formal ionic or covalent bonds.

**Western blotting:** A technique that employs antibodies to detect the presence of a protein in a biological sample, after the proteins from the sample have been separated in a gel. Also called immunoblotting.

**whole-genome shotgun sequencing:** Strategy for sequencing a genome in which random segments of DNA are sequenced and the segments are ordered by the computerized identification of sequence overlaps.

**wild type:** The allele or phenotype that appears with the greatest frequency in a natural population of a species.

**Wnt-class signaling pathway:** A defined type of cell-cell signaling during development that does not require cell-cell contact. Wnt-class genes and proteins are involved in the synthesis, secretion, and reception of glycoprotein signals in developing embryos.

**wobble base:** The base at the 5' end of an anticodon, which pairs loosely and can form mispairs with the base at the 3' end of the codon.

**wobble hypothesis:** Hypothesis proposed by Francis Crick in 1966 to describe how some anticodons can recognize more than one codon.

**wobble position:** The first position of the anticodon, at the 5' end, which may contain a wobble base.

**writhe (*Wr*):** The net number of supercoils in a DNA molecule. *Compare* twist (*Tw*).

**X chromosome inactivation:** A method of dosage compensation in mammals that involves inactivating one of the two X chromosomes in females by compacting it into heterochromatin.

**XIC:** *See* X inactivation center.

**X inactivation center (XIC):** A point on the mammalian X chromosome that is the nucleation center for heterochromatin formation when the chromosome is inactivated.

**x-ray crystallography:** The analysis of x-ray diffraction patterns of a crystalline compound, used to determine the molecule's three-dimensional structure.

**YAC:** *See* yeast artificial chromosome.

**yeast artificial chromosome (YAC):** Expression vector for cloning eukaryotic genes in yeast. YAC vectors have a yeast origin of replication, two selectable markers, and telomere and centromere sequences for maintaining chromosome integrity.

**yeast three-hybrid analysis:** Method for defining protein-RNA interactions *in vivo*. A series of engineered proteins and a randomized library of RNA sequences are set up such that expression of a reported gene occurs only when the RNA sequence bound by the target protein is present.

**yeast two-hybrid analysis:** Method for defining protein-protein interactions. Target protein interaction activates transcription of a reporter gene.

**Z-DNA (Z-form DNA):** Conformation of double-stranded DNA observed in solvents with a high salt concentration or with sequences rich in G≡C base pairs. The molecule assumes a left-handed helical conformation with 12 base pairs per turn and a rise of 3.7 Å per base pair. The backbone of the helix has a zigzag structure. *Compare* A-DNA and B-DNA.

**Z-form DNA:** *See* Z-DNA.

**zigzag model:** Model for the arrangement of nucleosomes in the 30 nm filament in which zigzag histone pairs stack on each other and twist about a central axis. *Compare* solenoid model.

**zinc finger:** A protein structural motif involved in DNA recognition by some DNA-binding proteins; characterized by a single atom of zinc coordinated to four Cys residues or to two His and two Cys residues.

# Appendix

---

## Solutions to Problems

### Chapter 2

1. **(a)** Plants with round seeds, 100%. **(b)** Plants with round seeds, 75%; plants with wrinkled seeds, 25%. **(c)** Plants with round seeds, 50%; plants with wrinkled seeds, 50%.
2. At least one of the parents was *RR*. The other parent was *RR*, *Rr*, or *rr*.
3. One parent was *RR*, and the other was *Rr*. In the F<sub>1</sub> generation, half of the plants were *RR* and half were *Rr*. In the F<sub>1</sub> crosses, all crosses involving *RR* plants (three-fourths of the total) produced progeny with round seeds. For the *Rr* × *Rr* crosses (one-fourth of the total), three-fourths of the progeny had round seeds and one-fourth had wrinkled seeds. The wrinkled seeds are thus found in  $0.25 \times 0.25 = 0.0625$  (~8/129), of the total plants.
4. Recall that this eye-color gene is on the X chromosome. The F<sub>1</sub> generation will have *X<sup>W</sup>Y* males (red eyes) and *X<sup>w</sup>X<sup>W</sup>* females (red eyes). The F<sub>2</sub> generation will have *X<sup>W</sup>Y* and *X<sup>w</sup>Y* males in a 50:50 mix, and *X<sup>w</sup>X<sup>W</sup>* and *X<sup>W</sup>X<sup>w</sup>* females in a 50:50 mix, so half of the males but no females will have white eyes. Some white-eyed females will appear in the F<sub>3</sub> generation, from crosses between *X<sup>W</sup>Y* males and *X<sup>w</sup>X<sup>W</sup>* females.
5. The F<sub>1</sub> generation will have *X<sup>W</sup>Y* males (all red-eyed) and *X<sup>w</sup>X<sup>W</sup>* females (all red-eyed). Half of the males in the F<sub>2</sub> generation will have white eyes (a result found by Morgan).
6. The gene producing the black color is on the Y chromosome. It is not present in any females, regardless of the generation. There are no other Y chromosomes in the population, so all males have the black trait.
7. All F<sub>1</sub> offspring are *RrLl* plants, with red flowers and large leaves. The F<sub>2</sub> generation is shown in the Punnett square.

	<b><i>RL</i></b>	<b><i>RI</i></b>	<b><i>rL</i></b>	<b><i>rl</i></b>
<b><i>RL</i></b>	<i>RRLL</i>	<i>RRll</i>	<i>RrLL</i>	<i>RrLl</i>
<b><i>RI</i></b>	<i>RRlL</i>	<i>RRll</i>	<i>RrlL</i>	<i>Rrll</i>
<b><i>rL</i></b>	<i>rRLL</i>	<i>rRll</i>	<i>rrLL</i>	<i>rrLl</i>
<b><i>rl</i></b>	<i>rRlL</i>	<i>rRll</i>	<i>rrlL</i>	<i>rrll</i>

9/16, or 56.25%, have red flowers and large leaves. 3/16, or 18.75%, have red flowers and small leaves. 3/16, or 18.75%, have white flowers and large leaves. 1/16, or 6.25%, have white flowers and small leaves.

8. 100% of the F<sub>1</sub> generation plants would have pink flowers. In the F<sub>2</sub> generation, 50% would have pink flowers.
9. 12.5% of the F<sub>2</sub> males will have red eyes and normal wings; these are marked in the Punnett square by an underline. They all have genotypes with at least one *X<sup>W</sup>* gene and one *V* gene.

	<b><i>X<sup>w</sup>V</i></b>	<b><i>X<sup>W</sup>V</i></b>	<b><i>X<sup>w</sup>v</i></b>	<b><i>X<sup>W</sup>v</i></b>
<b><i>X<sup>w</sup>V</i></b>	<i>X<sup>w</sup>X<sup>w</sup>VV</i>	<u><i>X<sup>w</sup>X<sup>W</sup>VV</i></u>	<i>X<sup>w</sup>X<sup>w</sup>Vv</i>	<u><i>X<sup>w</sup>X<sup>W</sup>Vv</i></u>
<b><i>X<sup>w</sup>v</i></b>	<i>X<sup>w</sup>X<sup>w</sup>vV</i>	<u><i>X<sup>w</sup>X<sup>W</sup>vV</i></u>	<i>X<sup>w</sup>X<sup>w</sup>vv</i>	<i>X<sup>w</sup>X<sup>W</sup>vv</i>
<b><i>YV</i></b>	<i>YX<sup>w</sup>VV</i>	<u><i>YX<sup>W</sup>VV</i></u>	<i>YX<sup>w</sup>Vv</i>	<u><i>YX<sup>W</sup>Vv</i></u>
<b><i>Yv</i></b>	<i>YX<sup>w</sup>vV</i>	<u><i>YX<sup>W</sup>vV</i></u>	<i>YX<sup>w</sup>vv</i>	<i>YX<sup>W</sup>vv</i>

10. The *G* and *S* genes are linked: they are found on the same chromosome and are sufficiently close together that crossovers between them are rare. Thus, the *G* and *S* alleles are not independently assorted.
11. In mitosis, the paired chromosomes (in the sister chromatid pair) are split up, and one chromatid passes to each daughter cell. In meiosis, the sister chromatids are not split up, but instead remain paired in the daughter cells produced by the first division.
12. Genes *A*, *B*, *C*, and *E* are on one chromosome, unlinked to genes *D* and *F*. Relative distances between genes are indicated by numbers under each chromosome.

#### Chromosome 1:

<b>A</b>	<b>C</b>	<b>B</b>	<b>E</b>
3	6	4	

#### Chromosome 2:

<b>D</b>	<b>F</b>
	16

13. M O N or N M O
14. Characteristics of rRNAs: only a few types in a cell; located in ribosomes; sequence highly conserved in different life forms. Characteristics of tRNAs: small; fold into a characteristic L shape; have many modified bases; can link to an amino acid; are functional RNAs (do not

## S-2 Appendix: Solutions to Problems

encode proteins). Characteristics of mRNAs: the most diverse of the RNAs in sequence, because each cell type makes >1,000 different mRNAs; can be very long; encode the sequence of proteins.

- 15.** Hershey and Chase could have done the work with a single batch of phage with both radioactive labels. After the blending step, the protein heads with the  $^{35}\text{S}$  label would have ended up in the supernatant, and the  $^{32}\text{P}$ -labeled DNA in the pellet containing the bacteria.

### Chapter 3

1. The O=O bond is stronger; its atoms are closer together.
2. The enantiomers have the same chemical formula, so the answer in each case is yes.
3. (d) It lowers the activation energy for a reaction.
4. pH 5
5. No. Peptides have directionality, established by the N-terminus and C-terminus. The order in which the amino acids are connected in a polypeptide confers unique properties on the polymer.
6. Activation energy is larger in the reverse direction.
7. The number of moles of NaCl remains unchanged; the *concentration* of NaCl is reduced by 50%.
8. The second law of thermodynamics, which states that, in any chemical or physical process, the entropy of the universe tends to increase.
9. A and B are equal. Resonance structures, by definition, represent the shared distribution of electrons between sets of bonding atoms, and hence the average bond lengths are equivalent.

- 10. (b)** Measuring rates.

- 11. (a)** 4.76. **(b)** 9.19. **(c)** 4.0. **(d)** 4.82.

*Details:*  $\text{pH} = -\log [\text{H}^+]$

$$\text{(a)} -\log [1.75 \times 10^{-5}] = 4.76$$

$$\text{(b)} -\log [6.50 \times 10^{-10}] = 9.19$$

$$\text{(c)} -\log [1.0 \times 10^{-4}] = 4.0$$

$$\text{(d)} -\log [1.5 \times 10^{-5}] = 4.82$$

- 12. (a)**  $1.5 \times 10^{-4}$  M. **(b)**  $3.0 \times 10^{-7}$  M.

$$\text{(c)} 7.8 \times 10^{-12}$$
 M.

*Details:*  $[\text{H}^+] = 10^{-\text{pH}}$

$$\text{(a)} [\text{H}^+] = 10^{-3.82} = 1.5 \times 10^{-4}$$
 mol/L

$$\text{(b)} [\text{H}^+] = 10^{-6.53} = 3.0 \times 10^{-7}$$
 mol/L

$$\text{(c)} [\text{H}^+] = 10^{-11.11} = 7.8 \times 10^{-12}$$
 mol/L

- 13. (a)** 5.00. **(b)** 4.22. **(c)** 5.40. **(d)** 4.70. **(e)** 3.70.

*Details:*  $\text{pH} = \text{p}K_a + \log [\text{acetate}]/[\text{acetic acid}]$

$$\text{(a)} \text{pH} = 4.70 + \log (2/1) = 5.00$$

$$\text{(b)} \text{pH} = 4.70 + \log (1/3) = 4.22$$

$$\text{(c)} \text{pH} = 4.70 + \log (5/1) = 5.40$$

$$\text{(d)} \text{pH} = 4.70 + \log 1 = 4.70$$

$$\text{(e)} \text{pH} = 4.70 + \log (1/10) = 3.70$$

- 14. (a)** 4.3. **(b)** pH decrease of 0.12. **(c)** pH decrease of 4.4.

*Details:*

$$\text{(a)} \text{pH} = \text{p}K_a + \log [\text{lactate}]/[\text{lactic acid}] = 3.6 + \log (0.05/0.01) = 3.60 + 0.69897 = 4.3$$

- (b)** Strong acids ionize completely:

$$0.005 \text{ L} \times 0.5 \text{ mol/L} \\ = 0.0025 \text{ mol of H}^+ \text{ added.}$$

The added acid converts some of the salt form (lactate) to the acid form (lactic acid):

$$[\text{H}^+] = 3.60 + \log \frac{0.05 - 0.0025}{0.01 + 0.0025} \\ = 3.6 + \log 3.8 = 3.6 + 0.58 = 4.18$$

The change in pH =  $4.30 - 4.18 = 0.12$ .

- (c)**  $\text{pH} = -\log [\text{H}^+]$

$$[\text{H}^+] = 0.0025 \text{ mol}/1.005 \text{ L} \approx 0.0025 \text{ mol/L} \\ \text{pH} = -\log 0.0025 = 2.6$$

The pH of pure water is 7.0, so change in pH =  $7.0 - 2.6 = 4.4$ .

- 15. pH 7.2**

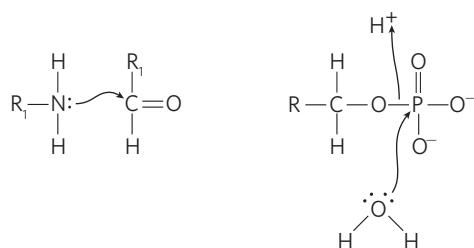
*Details:* At pH 2, 50% of the group with  $\text{p}K_a = 2.0$  is ionized ( $\text{p}K_a = \text{pH}$ ). Amount of base added =  $0.075 \text{ L} \times 0.1 \text{ mol/L} = 0.0075 \text{ mol}$ . Amount of compound =  $0.1 \text{ L} \times 0.1 \text{ mol/L} = 0.01 \text{ mol}$ . A pH increase to 6.72 completely titrates the remainder of the group with the low  $\text{p}K_a$  (because  $\text{pH} \gg \text{p}K_a$ ). So,  $50\% \times 0.01 \text{ mol}$  of compound requires 0.005 mol of base to titrate the rest of that group. Thus,  $0.0075 \text{ mol of base added} - 0.005 \text{ mol of base used} = 0.0025 \text{ mol of base remaining to titrate the second group}$ .

$$6.72 = \text{p}K_a + \log \frac{0.0025}{0.01 - 0.0025} \\ = \text{p}K_a + \log 0.333 = \text{p}K_a - 0.477 \\ 7.2 = \text{p}K_a$$

- 16.** The ring is flat/planar. The conjugated double bonds in the ring produce considerable resonance, such that all bonds in the ring have a partial double-bond character and all lie in the same plane.

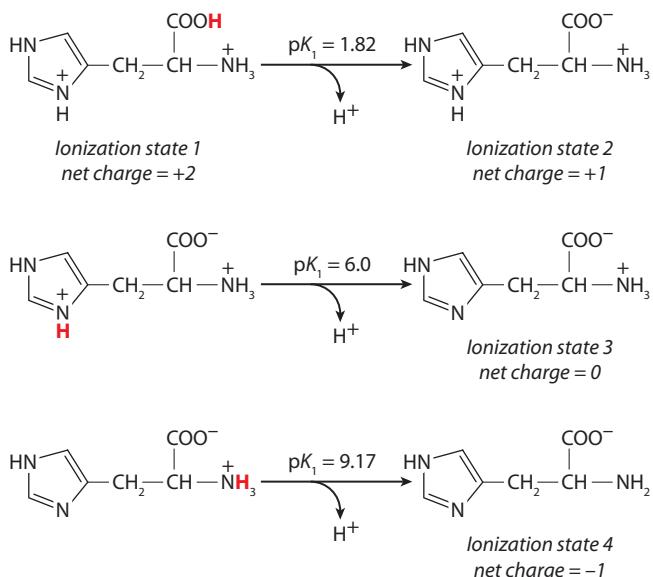
- 17.** Free rotation is restricted about the bond with angle  $\omega$ . Due to resonance between the N–C bond and the C=O bond, the N–C bond has a partial double-bond character.

- 18.**



## Chapter 4

1. (a)



(b), (c)

pH	Ionization state	Net charge	Migrates toward:
1	1	+2	Cathode
4	2	+1	Cathode
8	3	0	Does not migrate
12	4	-1	Anode

2. (a) Minimum  $M_r$  32,000. Remember that the molecular weight of a Trp residue is not the same as the molecular weight of the free amino acid. (b) 2 Trp residues.

3. (a) At pH 3, +2; at pH 8, 0; at pH 11, -2. (b) pI 7.

Details:

(a) This peptide has five ionizable groups: (1) the  $\alpha$ -amino group of E ( $pK_a$  9.67); (2) the side chain of E ( $pK_a$  4.25); (3) the side chain of H ( $pK_a$  6.0); (4) the side chain of R ( $pK_a$  12.48); and (5) the  $\alpha$ -carboxyl group of G ( $pK_a$  2.34). At pH 3, (1), (3), and (4) are protonated and positively charged, giving a net charge for these groups of +3. The pH of 3 is between the two  $pK_a$  values for (2) and (5) (one will be mostly protonated, the other will be mostly unprotonated), giving a net charge of about -1 for these two groups. This yields a net charge for the peptide of about +2. At pH 8, (2) and (5) are unprotonated, contributing a charge of -2; (3) is mostly unprotonated and neutral; (1) and (4) are mostly protonated, contributing a charge of nearly +2. The net charge is near zero. At pH 11, (2), (3), and (5) are unprotonated, with a charge of -2; (1) and (4) are mostly unprotonated and uncharged. The net charge is close to -2.

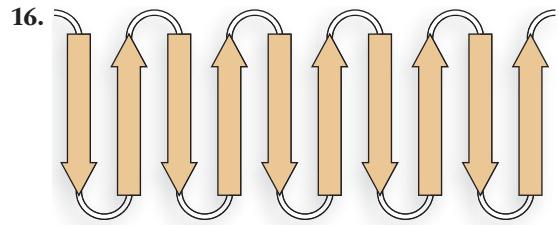
(b) The pI can be estimated by determining the pH at which the net charge is zero. In this peptide, (1), (3), and (4) can contribute + charges when protonated, and (2) and (5) can contribute - charges when unprotonated. Hence, a net charge of zero can occur at the pH where (1), (3), and (4) together contribute a net +2 charge to exactly balance the net -2 contributed by the two carboxyl groups of (2) and (5). This will occur about halfway between the  $pK_a$  values of the H and R side chains: the fraction of R side chains that are unprotonated is balanced exactly by the fraction of H side chains that are protonated. Thus, the pI is approximately 7.

4. ~~A D S E R N C Q L V I L L A W L P G V K V Q C A L L D R E T~~  
~~X X XX~~
5. The sheets are likely to be antiparallel, because the linkers that connect one  $\beta$  strand to the next are too short to connect parallel strands. The linkers are four residues long and could form  $\beta$  turns.
6. Alternating R groups are on opposite sides of the  $\beta$  sheet structure. Therefore, in a continuous layer of  $\beta$  sheet, with alternating polar and nonpolar residues in the  $\beta$  strands, the sheet is likely to fold into a  $\beta$  barrel, sequestering hydrophobic residues inside.
7. Residues 1 and 3 are in  $\beta$  sheets, residue 2 is in a right-handed  $\alpha$  helix, and residue 4 is in a left-handed helix. Residue 5 is an outlier, and the electron density map should be examined to see whether the angles are correctly assigned. If residue 5 were Gly, it might be in an acceptable position in the plot, because Gly residues are more flexible than other residues and can adopt a greater range of acceptable bond angles.
8. AIPRKKR~~E~~FICRGFAIRPNT. The P3 and P18 residues disfavor helix formation, limiting the region favorable for helix formation to the sequence between residues 4 and 17. The several positively charged residues, 4 through 7, are not likely to initiate a helix, because they would repel one another and interact unfavorably with the helix dipole, which is positively charged at the N-terminus. In contrast, E8 will interact favorably with the helix dipole. Thus, the most likely  $\alpha$ -helical region is residues 8 through 17 (boxed). Within this helix, the stabilizing interactions are: N-terminal positive dipole, stabilized by E8; C-terminal negative dipole, stabilized by R17; hydrophobic interactions between two F residues, spaced 4 residues (1 turn) apart; ion pair interaction between E8 and R12, spaced 4 residues (1 turn) apart.
9. (a) and (d). Both contain the consensus sequence for an ATP/GTP-binding site: (G/A)XXGXGK(T/S), where x is any amino acid.
10. As an  $\alpha$  helix, 210 Å ([1.5 Å/residue]  $\times$  140 residues). As a  $\beta$  strand, 490 Å ([3.5 Å/residue]  $\times$  140 residues).
11. There are many possible answers; any of the following will suffice. NMR uses magnets and radio-frequency irradiation; crystallography uses x rays. NMR is performed on

## S-4 Appendix: Solutions to Problems

proteins in solution; x-ray crystallography requires a protein crystal. NMR measures a nuclear event; x-ray crystallography measures events in the electron shell. In NMR, proteins emit radiowaves; in x-ray crystallography, proteins emit x rays. NMR makes great use of protons; protons are largely ignored in x-ray crystallography. NMR can be applied only to small proteins; x-ray crystallography can solve large proteins and complexes. Both methods irradiate the protein sample with electromagnetic radiation (photons). Both methods yield structures with atomic resolution. Both methods make heavy use of computations.

12. **(a), (b), (c)** one domain; **(d), (e)** two domains. When a protein reaches a size of about 150 to 200 residues ( $M_r \sim 20,000$ ), the polypeptide chain usually folds into two domains.
13. Completely buried residues are likely to be hydrophobic: L2, F4, I6, V8, V12, L13, L18, and L19 fit this description. Highly polar residues, or at least their polar groups, are likely to be on the surface, exposed to water: D1, K3, T5, S7, T14, R15, E16, Q17, and E20 fit this description.
14. **(a)** Destabilizes the positive dipole at the N-terminus of the helix. **(b)** Stabilizes the positive dipole at the N-terminus of the helix. **(c)** Destabilizes; the ion pair between R2 and E5, one turn apart, is eliminated. **(d)** No difference; the ion pair between the residues is maintained. **(e)** Stabilizes; the hydrophobic interaction is on the same side of the helix, one turn apart. **(f)** Destabilizes; P (a helix breaker) destabilizes the helix.
15. **(a)** The N-terminus is the lower-right end; the C-terminus is the upper-right end; the  $\beta$  turns are the U-turns at the lower left and upper right. **(b)** The more hydrophobic surface is likely to be on the right side of the  $\beta$  sheet.



### Chapter 5

1. Protein B has a higher affinity for ligand X; it is half-saturated at a much lower concentration of X than is protein A. Protein A has  $K_a = 10^6 \text{ M}^{-1}$ ; protein B has  $K_a = 10^9 \text{ M}^{-1}$ .
2. The  $K_d$  for DNA interactions increases when allolactose is bound. In other words, the affinity of the Lac repressor for its DNA binding site decreases, leading to its dissociation from the DNA.
3. DNA is a polyelectrolyte, and the many negatively charged groups in its backbone are bound by ions, primarily  $\text{Mg}^{2+}$  ions but also monovalent cations such as  $\text{K}^+$ . Binding of a protein to DNA involves displacement of some or all of the ions at the DNA site where binding occurs.

4. Nonspecific DNA binding generally involves interactions with the invariant parts of DNA structure—the phosphate and deoxyribose groups in the DNA backbone—and hydrophobic interactions with the nucleotide bases. Binding to a specific DNA sequence requires substantial interaction with groups in the bases that distinguish one nucleotide from another. These features of each base are largely accessible in the major and minor grooves of the DNA.
5. All of these situations give rise to real or apparent negative cooperativity. Apparent negative cooperativity in ligand binding can be caused by the presence of two or more different types of ligand-binding sites with different affinities for the same ligand on the same or different protein molecules in the same solution. Apparent negative cooperativity can also be observed in heterogeneous protein preparations. There are few well-documented cases of true negative cooperativity.
6.  $2.4 \times 10^{-6} \text{ M}$ .  
*Details:* The volume of a cylinder is given by  $V = \pi r^2 h$ . For a cylinder of  $r = 0.000050 \text{ cm}$  and  $h = 0.00020 \text{ cm}$ ,  $V = 1.57 \times 10^{-12} \text{ cc (mL)}$ , or  $1.57 \times 10^{-15} \text{ L}$ . The solution concentration in the cell is  $1.20 \text{ g/mL}$ , and the protein concentration is 20% of this, or  $0.24 \text{ g/mL}$ . Thus the cell contains  $3.77 \times 10^{-13} \text{ g}$  of protein. If the cell contains 1,000 proteins of the same molecular weight, it contains  $3.77 \times 10^{-16} \text{ g}$  of a given protein, or  $0.24 \text{ g}$  per liter of cytosol. Dividing  $0.24 \text{ g/L}$  by the  $100,000 \text{ g/L}$  of a  $1 \text{ M}$  solution gives a concentration of  $2.4 \times 10^{-6} \text{ M}$ . This corresponds to just over 2,200 copies of each protein in the cell.
7. **(b), (e), and (g).**
8. We now know that to catalyze a reaction, an enzyme active site must be complementary (in shape and charge) not to the substrate, but to the transition state of the reaction that is catalyzed.
9. The inactivated enzyme would no longer have a measurable  $k_{\text{cat}}$  and  $K_m$ .
10. **(a)**  $[S] = 1.7 \times 10^{-3} \text{ M}$ . **(b)**  $0.33V_{\text{max}}, 0.67V_{\text{max}}, 0.91V_{\text{max}}$ .  
**(c)** The upper (red) curve corresponds to enzyme B ( $[X] > K_m$  for this enzyme); the lower (black) curve, enzyme A.
11. **(a)**  $k_{\text{cat}} = 400 \text{ s}^{-1}$ . **(b)**  $K_m = 10 \mu\text{M}$ . **(c)**  $\alpha = 2, \alpha' = 3$ .  
**(d)** ANGER is a mixed inhibitor.
12. **(a)**  $[E_t] = 24 \text{ nM}$ . **(b)**  $[A] = 4 \mu\text{M}$  ( $V_0$  is exactly  $\frac{1}{2}V_{\text{max}}$ , so  $[A] = K_m$ ). **(c)**  $[A] = 40 \mu\text{M}$  ( $V_0$  is exactly  $\frac{1}{2}V_{\text{max}}$ , so  $[A] = 10K_m$  in the presence of inhibitor).
13.  $V_{\text{max}} \approx 140 \text{ mm min}^{-1}; K_m \approx 1 \times 10^{-5} \text{ M}$ .
14. Movement along the DNA requires ATP hydrolysis. A normal RuvB subunit can still hydrolyze ATP, even if it is adjacent to a mutant subunit in a heterohexameric complex. However, movement along DNA requires cooperation between adjacent subunits, which cannot occur if one of the subunits is mutated.
15. In principle, ATM and ATR could be enzymes that covalently modify other proteins. In fact, ATM and ATR are the most common type of such proteins: they are

protein kinases that add phosphoryl groups to hundreds of cellular protein targets.

## Chapter 6

1. An enzyme (RNA or protein) must (1) increase the rate of a chemical reaction and (2) remain unchanged on completion of a catalytic cycle.
2. N-3 and N-7.
3. Within experimental error, the number of purines ( $A + G$ ) is equal to the number of pyrimidines ( $C + T$ ); the fractional amounts of A and T are the same, and the fractional amounts of G and C are the same; the relative ratios of the bases do not vary from one tissue to another.
4. (a) 5'-TACCAGCCTTAGAATTAACTAAGGCTGTAATC-3' (note that nucleic acid sequences are always written in the 5'→3' direction, and the two strands of DNA are antiparallel). (b) 5'-CAGCCTTAG-3' and 5'-CTAAGGCTG-3' form an inverted repeat, so the strand has the potential to form a hairpin. The duplex can assume a cruciform structure.
5. Higher. RNA has greater thermal stability than DNA.
6. The DNA has a backbone with deoxyribose and will take up a B-form helix. The RNA has a backbone with ribose and will take up an A-form helix.
7. The presence of U rather than T residues in RNA is a likely mechanism by which cells monitor mutations in DNA. Uracil is produced in DNA largely by the slow, nonenzymatic hydrolytic deamination of cytosine; having thymine as the base in DNA allows the efficient detection and repair of C-to-U mutations.
8. 5'-ATTGCATCCGCGCGTGCGCGCGATCCGTTACTT-TCCG-3'
9. The double helix is the most thermodynamically stable structure. It places the hydrophobic bases in the interior of the molecule, where they interact with one another through base stacking, and the charged phosphate groups on the outside, where they can interact with water and ions.
10. The phosphate groups between the sugars (deoxyribose or ribose) in the sugar-phosphate backbone are highly acidic, giving the nucleic acids an overall negative charge.
11. Cytosine deamination to form uracil is a slow but constant reaction in all cells. In many eukaryotes, hundreds of C residues are converted to U residues every day, in every cell, creating G-U base pairs that are “seen” by the repair system as G-T pairs. Because the G is correct and the U is the damaged base, repairing G-T to G≡C restores the correct genetic information. Repair to A=T would cause a mutation.
12. No. The three-dimensional structure of a “tDNA” would almost certainly not be correct. The 2'-hydroxyl groups of ribose contribute significantly to tRNA folding, and enzymes specific for tRNAs modify certain of their bases, which is also essential to the three-dimensional structure.
13. The G and C content and the length of the DNA both influence the strength of association of the two strands in the double helix. G≡C pairs are stronger than A=T pairs, and the longer the DNA, the greater the number of base pairs and the greater the energy (i.e., the higher the temperature) required to break the hydrogen bonding between them.
14. An abundance of purines, especially adenosines, which play an important role in the three-dimensional folding of RNA; also, multiple short segments capable of base pairing with adjacent or distant regions of the RNA molecule, particularly if these short segments are conserved in related organisms.
15. The regular repeating properties of the double helix produce characteristic x-ray diffraction patterns for DNA fibers, as used in the earliest studies of DNA structure. However, these diffraction patterns result from the *averaged* properties of the DNA helices in a fiber. To determine the properties of individual DNA sequences, single crystals containing a single, homogeneous form of the DNA molecule, arranged in a three-dimensional array, were necessary. The x-ray diffraction patterns produced from single crystals could be used to determine the electron density map of the DNA in the crystal, providing an exact, rather than an averaged, image of the molecular structure.
16. (a) There is no sulfur in DNA, so proteins are uniquely labeled by  $^{35}\text{S}$ . There is little or no phosphate in proteins (at least in bacteria), so DNA is uniquely labeled by  $^{32}\text{P}$ . (b) Using  $^{14}\text{C}$  or  $^3\text{H}$  would have labeled both the DNA and the protein, permitting no differentiation. (c) The intact phages, the T2 ghosts, and the DNA are all insoluble in acid, and all are removed from solution by centrifugation. (d) The nucleotides liberated by DNase treatment are soluble in acid. Osmotic shock releases the T2 DNA into solution, where it is degraded by the DNase. The unplasmolyzed T2 phages contain DNA, but it is protected from the DNase, within the phage protein coat. (e) Both the intact viruses and the T2 ghosts adsorb to the bacteria. The components needed for attachment of T2 to the bacteria are located uniquely in the protein coat. (f) The antibodies recognize the T2 protein coat. In both the control and plasmolyzed samples, the protein coats are immunoprecipitated by the antisera, but in the plasmolyzed sample, the DNA is left behind in solution. (g) The material released by osmotic shock is entirely or almost entirely DNA. The T2 ghosts are almost entirely protein. Little or no protein is released from the phages with the DNA. The DNA does not adsorb to phage-susceptible bacteria on its own. The ghosts are protein coats that surround the DNA of the intact phage particles. These coats react with antibodies and protect the DNA within from DNase. They also are responsible for attaching phages to a bacterial host. (h) The centrifugation and resuspension remove unadsorbed phages from the solution, which otherwise would add to the background signal. (i) About 80% of the  $^{35}\text{S}$ -labeled phage heads are stripped

## S-6 Appendix: Solutions to Problems

from the cells by the blender, with only about 16% found in the supernatant without the blender treatment. The amount in the supernatant without blender treatment increases when the multiplicity of infection increases, as a result of some kind of displacement of phages by other phages attached to the same cells. (j) The bulk of the  $^{35}\text{S}$  is removed from the cells by blender treatment, whereas a relatively small amount of the  $^{32}\text{P}$  is removed. The capacity of the cells to survive and continue with the infection process is not affected by the treatment. The results indicate that the bulk of the protein remained in the protein coats at the cell surface during the infection, while the bulk of the DNA entered the cells.

### Chapter 7

#### 1. (a)

5'--- G-3'  
3'--- CTTAA-5'

5'-AATTC --- 3'  
3'-G --- 5'

#### (b)

5'--- GAATT-3'  
3'--- CTTAA-5'

5'-AATTC --- 3'  
3'-TTAAG --- 5'

#### (c)

5'--- GAATTAATTTC --- 3'  
3'--- CTTAATTAAG --- 5'

#### (d)

5'--- G-3'  
3'--- C-5'

5'-C --- 3'  
3'-G --- 5'

#### (e)

5'--- GAATTC --- 3'  
3'--- CTTAAG --- 5'

#### (f)

5'--- CAG-3'  
3'--- GTC -5'

5'-CTG --- 3'  
3'-GAC --- 5'

#### (g)

5'--- CAGAATTC --- 3' or 5'--- GAATTCTG --- 3'  
3'--- GTCTTAAG --- 5'      3'--- CTTAAGAC --- 5'

(h) *Method 1:* Cut the DNA with EcoRI as in (a). Then treat the DNA as in (b) or (d), and ligate a synthetic DNA fragment containing the BamHI recognition site

between the two resulting blunt ends. *Method 2* (more efficient): Synthesize a DNA fragment with the structure:

5'-AATTGGATCC  
3'-CCTAGGTTAA

This would ligate efficiently to the sticky ends generated by EcoRI cleavage, would introduce a BamHI site, but would not regenerate the EcoRI site.

(i) The four fragments (with N = any nucleotide), in order of discussion in the problem, are:

5'-AATTCNNNCTGCA-3'  
3'-GNNNNG - 5'

5'-AATTCNNNNTGCA-3'  
3'-GNNNNC - 5'

5'-AATTGNNNCTGCA-3'  
3'-CNNNNG - 5'

5'-AATTGNNNNTGCA-3'  
3'-CN>NNNC - 5'

#### 2.

5'-GAAAGTCCCGCTTATAGGCATG-3'  
3'-ACGTCTTCAGGCGCAATATCCGTACTTAA-5'

3. A YAC vector is not stably maintained as a yeast chromosome during mitosis unless it carries an insert of more than 100,000 bp.

4. (a) Some original pBR322 plasmids will be present, regenerated without insertion of a foreign DNA fragment. Also, two or more pBR322 plasmids might be ligated, with or without insertion of a segment of foreign DNA. All of these would retain resistance to ampicillin. (b) The clones in lanes 1 and 2 each have one DNA fragment, inserted in different orientations. The clone in lane 3 has incorporated two of the DNA fragments, ligated such that the ends closest to the EcoRI sites are joined.

5. The sequence will appear about once every 4<sup>8</sup> bp, or once every 65,536 bp. If the G + C content is greater than the A + T content (or vice versa), the frequency of occurrence of the restriction site will decrease.

6. In a large DNA molecule with a random sequence, a BamHI site will occur, on average, every 4,096 bp (assuming all four nucleotides are present in equal proportions). Cleavage of all BamHI sites in the DNA would produce fragments much smaller than the 100,000 to 300,000 bp needed for a BAC library.

#### 7.

Primer 1: CCTCGAGTCAATCGATGCTG  
Primer 2: CGCGCACATCAGACGAACCA

Recall that all DNA sequences are written in the 5'→3' direction, left to right; that the two strands of a DNA molecule are antiparallel; and that both PCR primers must target the end sequences so that their 3' ends are oriented toward the segment to be amplified.

**8.**

Primer 1: GAATTCCCTCGAGTCAATCGATGCTG  
Primer 2: GAATTCCGCGCACATCAGACGAACCA

- 9.** The test requires DNA primers, a heat-stable DNA polymerase, deoxynucleoside triphosphates, and a PCR machine. The primers are designed to amplify a DNA segment encompassing the CAG repeat. The DNA strand shown in the problem is the coding strand, oriented 5'→3', left to right. The primer targeted to the DNA to the left of the CAG repeat should be identical to any 25-nucleotide sequence in the region to the left of the repeat. The primer on the right side must be *complementary* and *antiparallel* to a 25-nucleotide sequence to the right of the CAG repeat. With these primers, an investigator would use PCR to amplify the DNA including the CAG repeat, and then determine its size by comparison with size markers on electrophoresis. The length of the DNA reflects the length of the CAG repeat, providing a simple test for the disease.
- 10.** The researcher could design PCR primers complementary to DNA in the deleted segment that will direct DNA synthesis away from each other. No PCR product will be generated unless the ends of the deleted segment are joined to create a circle.

**11.**



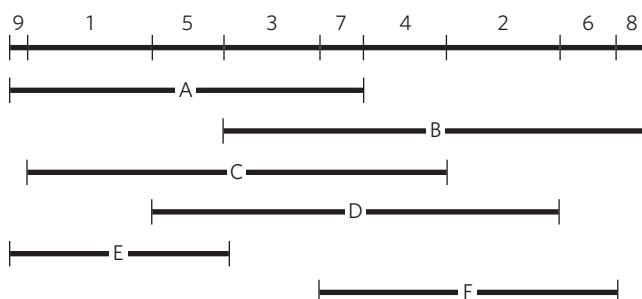
- 12.** No. The orientation of the cloned gene is very important, because the information specifying the protein is contained in only one of the two DNA strands. The promoter specifies not only where RNA polymerase binds the DNA, but also the direction in which it travels and the DNA strand that it uses as template for RNA synthesis. When the correct DNA strand is used as template, a functional protein results. If the

gene is inverted, the opposite DNA strand will become the template for synthesis of RNA, with a much different nucleotide sequence. The resulting protein will be completely different, unrelated to the normal gene product, and most likely nonfunctional.

- 13.** The bacterial *recA* gene is readily cloned by using the bacterial plasmid. For the mammalian DNA polymerase, the baculovirus system may have a better chance of generating an active protein.

- 14.** The production of labeled antibodies is difficult and expensive. The labeling of every antibody to every protein target would be impractical. By labeling one antibody preparation for binding to all antibodies of a particular class, the same labeled antibody preparation can be used in many different Western blot experiments.

**15.**



- 16.** Express the protein in yeast strain 1 as a fusion protein with one of the domains of Gal4p—say, the DNA-binding domain. Using yeast strain 2, make a library in which essentially every protein of the fungus is expressed as a fusion protein with the interaction domain of Gal4p. Mate strain 1 with the strain 2 library, and look for colonies that are colored due to expression of the reporter gene. These colonies will generally arise from mated cells containing fusion protein that interact with your target protein.

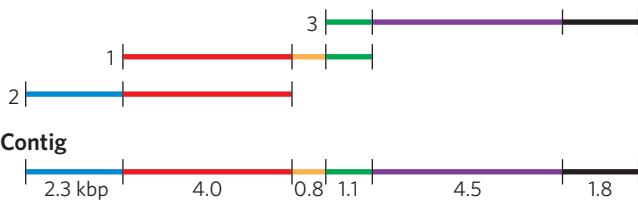
- 17.** Cover spot 4, add solution containing activated T, irradiate, wash. Result: 1. A-T; 2. G-T; 3. A-T; 4. G-C. Cover spots 2 and 4, add solution containing activated G, irradiate, wash. Result: 1. A-T-G; 2. G-T; 3. A-T-G; 4. G-C. Cover spot 3, add solution containing activated C, irradiate, wash. Result: 1. A-T-G-C; 2. G-T-C; 3. A-T-G; 4. G-C-C. Cover spots 1, 3, and 4, add solution containing activated C, irradiate, wash. Result: 1. A-T-G-C; 2. G-T-C-C; 3. A-T-G; 4. G-C-C. Cover spots 1 and 2, add solution containing activated G, irradiate, wash. Result: 1. A-T-G-C; 2. G-T-C-C; 3. A-T-G-G; 4. G-C-C-G.

- 18. (a)** R6-5, at least 11; pSC101, 1; pSC102, 3. **(b)** Each of the observed bands in a given lane represents a DNA fragment, and each fragment is present in the same concentration (total molecules). However, the fragments to the left are longer than the ones to the right and thus take up more of the fluorescent stain. **(c)** Two EcoRI

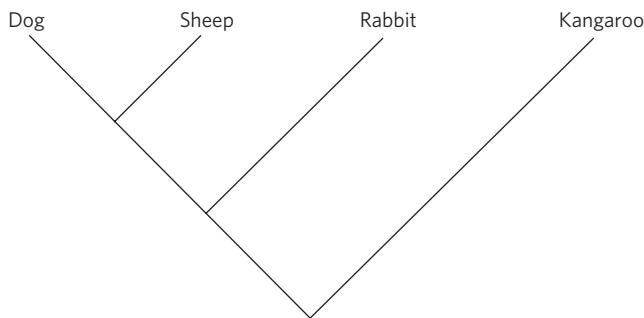
fragments derived from R6-5 are very nearly the same size, and they migrate together at this position. Thus, there are 12 fragments, derived from cleavage of R6-5 at its 12 EcoRI recognition sites. (d) The plasmid pSC102 is made up of these three EcoRI fragments from R6-5. (e) The larger fragment on the left, which comigrates with pSC101, is the only possible source of a tetracycline-resistance gene in the parent plasmids. (f) The smaller fragment on the right, which comigrates with one of the fragments of pSC102, is the only possible source of a kanamycin-resistance gene in the parent plasmids. (g) 7,000 bp. (h) 4 phosphodiester bonds; two fragments were ligated, with two new phosphodiester bonds created at each of the two ligation sites. (i) The original ligation mixture included a wide range of combinations of DNA fragments. When the mixture was used to transform *E. coli* cells, only cells that took up a combination of fragments that allowed survival on the selection media would grow. Evidently, these two fragments of pSC101 did not include a gene for resistance to tetracycline or kanamycin, the antibiotics used for selection. The pSC101 plasmid that included the tetracycline-resistance gene also had a replication origin, so no new replication origin was required. The joining of three fragments into a circle by ligation is considerably less probable than the ligation of two fragments. In effect, the selection generated the simplest possible recombinant plasmid from the available fragments.

## Chapter 8

1.



2.



3. The genomic DNA fragments are cloned into plasmid vectors. Although the sequence of the cloned DNA is not known, the plasmid sequences immediately adjacent to a DNA fragment are known. A single primer is used, targeted to the plasmid sequence near one end of the cloning site, and sequencing of each clone is initiated from that point.

4. ATSAAGWDEWEGGK**V**LHLD**G**KLQNRCALLELDIGAV
5. (a) Modern sequencing of the cDNAs generated in RNA-Seq produces many short sequence reads; each read is linked to a particular gene by its sequence, and the number of reads from each gene provides a measure of the relative numbers of RNAs derived from that gene. (b) In most cells, the majority of the RNA is rRNA, and the rRNA thus generates a very high background of sequence information, if it is not removed before the analysis.
6. The possibilities include gene duplication, horizontal gene transfer, and transposon insertion.
7. The primers can be used to probe libraries containing long genomic clones to identify contig ends that lie close to each other. If the contigs are close enough, the primers can be used in PCR to directly amplify the intervening DNA separating the contigs, which can then be cloned and sequenced.
8. If the same procedure were used in both dimensions, all the proteins would form a single diagonal line in the final gel, and much of the potential for separation would be wasted. Using different protein properties in the two electrophoresis steps effectively spreads the proteins over the entire gel.
9. The same disease condition can be caused by defects in two or more genes, which are on different chromosomes.
10. If gene X has no relationship to any other gene in species B, it may have arisen by horizontal gene transfer. If gene X is homologous to gene 2, it may have arisen by gene duplication.
11. The pattern of haplotypes in the Aleut and Eskimo populations suggests that their ancestors migrated into the American Arctic regions in a separate migration from the one that led to the populating of the rest of North and South America.
12. Yes. Mitochondrial Eve lived thousands of years before Y chromosome Adam. Given that all modern humans (male and female) have mitochondrial DNA from Eve, Adam must have had that DNA as well.
13. (a) Coronavirus genomes are composed of a single strand of RNA. At least one round of synthesis by reverse transcriptase is needed if the genome is to be amplified by PCR. (b) In all PCR experiments, researchers design primers sufficiently long and unique that they are unlikely to anneal well to other areas of a genome. In an experiment like this, one would also search for sequences that are highly conserved among similar genomes (e.g., sequences encoding parts of an enzyme that are critical to the enzyme's function), thereby maximizing the chance that the sequence will be present in an unchanged state in the target genome. (c) HEV and BCoV: 6; BCoV and SARS: 146; TGEV and SARS: 168. The coronaviruses BCoV and HEV are most closely related, with few differences between them. The differences among sequences are consistent with the phylogenetic tree.

## Chapter 9

1.  $Lk_0 = (4,200 \text{ bp})/(10.5 \text{ bp/turn}) = 400$ . From Equation 9-1,  $\Delta Lk = Lk - Lk_0 = 374 - 400 = -26$ . Substituting the values for  $\Delta Lk$  and  $Lk_0$  into Equation 9-2:  $\sigma = \Delta Lk/Lk_0 = -26/400 = -0.065$ . The superhelical density is negative, so the DNA molecule is negatively supercoiled. When the same molecule has an  $Lk$  of 412,  $\Delta Lk = 412 - 400 = 12$ , and  $\sigma = 12/400 = 0.03$ . The superhelical density is positive, so the molecule is positively supercoiled.
2. (a) The DNA has  $\sim 171,000 \text{ bp}$ ; at  $0.34 \text{ nm/bp}$ , the DNA length is  $58,140 \text{ nm}$ . The DNA is almost 600 times longer than the JS98 head. (b) 170,523 bp.
3. The content of A does not equal the content of T. The simplest explanation is that the DNA is single-stranded.
4. The DNA has a molecular weight of  $580,070 \text{ bp} \times 650/\text{bp} = 377,045,500$ . The contour length is  $197,224 \text{ nm}$ .  $Lk_0 = (580,070 \text{ bp})/(10.5 \text{ bp/turn}) = 55,245$ .  $Lk = 55,245 - (55,245 \times 0.06) = 51,930$ .
5. The DNA has  $\sim 5,250 \text{ bp}$ . (a) In the absence of strand breakage and resealing,  $Lk$  is unchanged; a positive supercoil must form elsewhere in the DNA to compensate. (b)  $Lk$  is undefined. (c)  $Lk$  decreases. (d) No change.
6.  $Lk$  remains unchanged because the topoisomerase introduces the same number of positive and negative supercoils.
7.  $Lk_0 = (13,800 \text{ bp})/(10.5 \text{ bp/turn}) = 1,314$ .  $\sigma = (Lk - Lk_0)/Lk_0 = -92/1,314 = -0.07$ . Superhelical density is the same for the cellular chromosome and the plasmid, so the probability of infection is  $>70\%$ .
8. (a)  $Lk$  undefined. (b)  $Lk = 500$ . (c) No effect. (d)  $Lk = 484$ . (e)  $Lk = 488$ . (f)  $Lk = 484$ .
9. Z-DNA is a left-handed double helix. Underwinding of the right-handed B-form helix will make a left-handed helix easier to form.
10. (a) The DNA must be unbroken and topologically constrained so that  $Lk < Lk_0$ . (b) Strand separation, formation of hairpins and cruciforms, and formation of Z-DNA are all more favorable in negatively supercoiled DNA. (c) DNA gyrase introduces negative supercoils into DNA, with the aid of ATP. (d) The mechanism involves creation of a double-strand break, passage of an unbroken DNA segment through the break, followed by strand resealing. Transient phosphotyrosyl-DNA intermediates form, and the conformational changes are coupled to hydrolysis of ATP.
11. The DNA must include origins of replication, required for DNA replication; a centromere, for proper segregation of the chromosome at cell division; and telomeres, to protect the chromosomal ends.
12. (a) The lower, faster-migrating band is negatively supercoiled plasmid DNA. The upper band is nicked, relaxed DNA. (b) DNA topoisomerase I relaxes the supercoiled DNA. The lower band will disappear, and all of the

DNA will converge on the upper band. (c) DNA ligase produces little change in the pattern. Some minor additional bands may appear near the upper band, due to the trapping of topoisomers not quite perfectly relaxed by the ligation reaction. (d) The upper band will disappear, and all of the DNA will be in the lower band. The supercoiled DNA in the lower band may become even more supercoiled, and migrate somewhat faster.

13. (a) When DNA ends are sealed to create a relaxed, closed circle, some DNA species are completely relaxed but others are trapped in slightly under- or overwound states. This gives rise to a distribution of topoisomers centered on the most relaxed species. (b) Positively supercoiled. (c) The DNA that is relaxed despite the addition of dye is DNA with one or both strands broken. DNA isolation procedures inevitably introduce small numbers of strand breaks in some of the closed-circular molecules. (d)  $\sigma \approx -0.05$ . This is determined by comparing native DNA with samples of known  $\sigma$ . In both gels, the native DNA migrates most closely with the sample of  $\sigma = -0.049$ .
14. Form I DNA was negatively supercoiled. When spread on an electron microscope grid, the DNA would tend to fold onto itself, creating DNA crossings or nodes. In form II DNA, the circles are relaxed.
15. The pattern in lane 2 is produced by DNA gyrase; that in lane 3 by DNA topoisomerase III (a type I topoisomerase).
16. (a) 25 nodes. (b) Removal of 25 of 667 DNA turns would correspond to a  $\sigma$  of  $-0.037$ . (c)  $\Delta Lk = 25/(-0.89) = -28$ ; thus,  $\sigma = -0.042$ . (d) No.

## Chapter 10

1. Histones have an unusually high concentration of positively charged amino acid residues on their surface compared with most other proteins. Although many SDS molecules bind each protein and give it an overall negative charge, the SDS does not eliminate the positive charges on a protein, it just overwhelms them. Thus, the abundance of positive charges on histones prevents the full effect that SDS normally has on the charge of a protein, and this manifests as a slower histone migration during electrophoresis compared with most other types of protein.
2. Phosphorylation and acetylation add groups that alter net charge. Methylation of lysine does not remove the positive charge of the terminal amino moiety.
3. The bacterial chromosome is divided into topologically constrained loops, defined by bound proteins at their boundaries. When the DNA in one loop is relaxed, the DNA in other loops remains supercoiled.
4. Transcription will decrease. H2A and H2B are core histones and are closely paired in the nucleosome structure. H1 is generally bound in linker regions, between the core histones, and its level can be varied independent of the core histones. An increase in H1 will lead to greater compaction of the DNA and thus decreased transcription.

5. Histone H1 is in the center of the filament, along with the linker DNA. The nucleosomes are stacked along the outside of the filament.
  6. Bacteria generally divide much more rapidly than eukaryotic cells. Stable protein-rich structures would impede the required replication and segregation of chromosomes at cell division.
  7. Transcriptionally active genes are characterized by a decrease in histone H1, an absence of bound nucleosomes at the promoter regions, the presence of specialized chromatin remodeling complexes, and the presence of histone variants such as H2AZ and H3.3.
  8. Epigenetic inheritance refers to chromatin modifications (particularly histone modifications) that are retained in the chromatin after cell division and affect gene transcription. Such modifications are not encoded in the DNA and thus are not subject to Mendelian inheritance.
  9. (c)
  10.  $62 \times 10^6$  H2A molecules. (The genome refers to the haploid genetic content of the cell; the cell is actually diploid, so the number of nucleosomes is doubled.)  $[(3.1 \times 10^9 \text{ bp})/(200 \text{ bp/nucleosome})] \times 2 \text{ H2A/nucleosome} \times 2 \text{ [for diploid cell]} = 62 \times 10^6 \text{ H2A}$ . The 62 million would double on replication.
  11. Instead of observing the eight different complexes, Kornberg would have observed five: H3, H4, H3-H3, H3-H4, and H3-H3-H4.
  12. (a) 220 bp is the approximate spacing of adjacent nucleosomes in chromatin. (b) The excess DNA sequences allowed the investigators to select for sequences that bound tightly and to eliminate the weaker binders. (c) The salt interfered with protein-DNA interactions and ensured that only the most tightly binding DNA remained bound to the nucleosomes. (d) Isolation of the DNA-nucleosome complexes reduced the total amount of DNA in each cycle. The PCR step allowed the DNA levels to be increased again. However, only the bound DNA sequences, the “winners,” were amplified; in each cycle, the solution was enriched in DNA sequences binding more tightly to the nucleosomes.
- Chapter 11**
1. The plasmid replicates unidirectionally. Molecules (c) and (d) are inverted relative to (a) and (b). Molecule (a) identifies the position of the origin relative to one end. The observation that (b), (c), and (d) have one forked end of similar size and the other forked end differing in size reveals that a single replication fork moves first through the short arm of molecule (a) and then proceeds around the circular plasmid. The order of replication time is (a), (b), (d), (c).
  2. (a) No. In the absence of any one dNTP, the polymerase would stop incorporating the other three dNTPs as soon as it encountered a template residue that should pair with the missing dNTP, and incorporation of  $^{32}\text{P}$  would be undetectable. (b) No. DNA synthesis releases the  $\beta$  and  $\gamma$  phosphates of dNTPs as pyrophosphate.
  3. Possible answers: Pol I is slow in DNA synthesis compared with the rate of replication in *E. coli*. Pol I can be mutated and the cells still survive. Pol I is not highly processive.
  4. The DNA polymerase contains a 3'  $\rightarrow$  5' exonuclease that degrades DNA to produce  $[^{32}\text{P}]$ dNMPs. The activity is not a 5'  $\rightarrow$  3' exonuclease, because addition of dNTPs inhibits  $[^{32}\text{P}]$ dNMP production: the polymerase extends radioactive 3' termini by adding nonradioactive dNTPs, protecting the radioactive portion of DNA from the 3'  $\rightarrow$  5' exonuclease. This would not protect the 5' terminus of radioactive DNA from a 5'  $\rightarrow$  3' exonuclease. Adding pyrophosphate would result in production of  $[^{32}\text{P}]$ dNTPs through reversal of the polymerase reaction.
  5. Ligase will not seal a nick in which the 5'-terminal nucleotide is a ribonucleotide. Sealing is delayed until all the RNA has been removed.
  6. (a) Either any combination of three A sites is sufficient for origin function, or three particular A sites are required. Construct four plasmids, each with a different mutant A site. Transfer the mutant plasmids into the host organism, and plate each transformed product on a medium containing the appropriate antibiotic. Use an unmutated plasmid and a plasmid without A sites as controls. If a particular A site is essential, the mutant plasmid will not form a colony. (b) Either the B sites are not essential, or one B site is needed but either one suffices. Construct a plasmid containing mutations in both B sites. If a particular B site is essential, a colony will not appear after transformation. Use an unmutated plasmid and a plasmid without B sites as controls.
  7. The preRC forms only in G<sub>1</sub>, not in other phases of the cell cycle. Cyclin kinases produced only in S phase are needed to assemble the remaining proteins to produce active replication forks. Origins do not fire a second time, because new preRC complexes cannot form until the cell completes its cycle and returns to G<sub>1</sub>.
  8. The  $\tau$  subunits link together the leading- and lagging-strand core polymerases, one  $\tau$  linked to each core, and both connected to DnaB. (a) Two. (b) Zero. The core polymerase, in conjunction with a  $\beta$  sliding clamp, is capable of processive synthesis of a new DNA strand on a single-stranded DNA template without any other subunits being present. This is analogous to leading-strand synthesis without lagging-strand synthesis.
  9. DnaA: ATP hydrolysis inactivates the DnaA for replication initiation. DnaC: ATP hydrolysis helps release DnaB helicase as it is loaded onto the DNA. Pol III  $\gamma$  and  $\tau$  subunits: ATP hydrolysis allows the  $\beta$  subunit (sliding clamp) to close around the DNA as it is loaded.
  10. The two replication forks would never meet, and part of the chromosome near the terminus would remain unreplicated.

- 11.** **(a)** As the DNA strands are linked together, the singly bonded phosphoryl groups are converted to phosphodiester bonds (doubly bonded phosphoryl groups) that are no longer susceptible to alkaline phosphatase. **(b)** The substrate for the reaction is a DNA strand break in one strand of double-stranded DNA. Ligation of single strands is not observed. **(c)** The reaction halts only because the enzyme runs out of substrate. Addition of poly(dA) creates more of the correct substrate and the reaction can continue. **(d)** The DNA ligase of *E. coli* uses NAD<sup>+</sup> as cofactor rather than ATP.

## Chapter 12

- 1.** The cross-linked pyrimidine dimer causes a distortion in the DNA that prevents base pairing in the active site of the DNA polymerase.

**2. (a)**

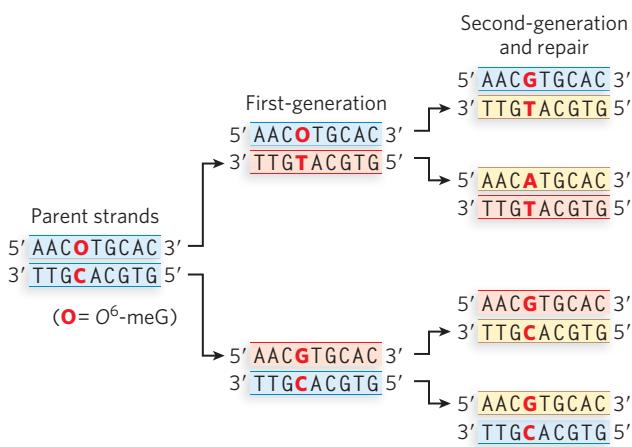


where O represents O<sup>6</sup>-meG.

**(b)**



**(c)**



- 3.** Any three of the following: defects in repair; UV light; TLS error-prone bypass; oxidative damage; spontaneous hydrolysis. For these lesions to become mutations, replication must occur (before repair), producing a naturally occurring base pair that is different from the original base pair.
- 4.** Serine auxotrophs can grow only on serine-containing medium; when plated on a serine-free medium, the cells will die—unless they have undergone a mutation that reverses the original mutation (i.e., a reversion mutation)

to restore the serine-synthesizing pathway. Treatment with a genotoxic agent, even though it kills most cells by producing mutations in vital genes, produces some of these particular reversion mutations.

- 5.** The survivors arise from spontaneous reversion mutations caused by spontaneous hydrolysis, oxidative damage, or natural sources of irradiation.
- 6.** **(a)** The colonies arose because background (spontaneous) mutations included reversion mutations that reversed the cells' histidine dependency. **(b)** 2-Aminoanthracene causes DNA damage; some of the lesions escape repair before DNA replication, thus forming a mutation. Some of these are reversion mutations in the histidine-pathway gene, allowing the cells to survive in a histidine-free medium. **(c)** 2-Aminoanthracene is a potential carcinogen, because it causes mutations in the Ames test.
- 7.** **(a)** XP mutant cells usually contain a mutation in a gene required for NER, the main UV lesion repair pathway in humans. Single-strand DNA breaks are produced during NER, accounting for the short DNA fragments observed in normal cells after irradiation. However, a defective NER system in the XPG mutant cells prevents formation of single-strand breaks. **(b)** XPG cells are defective at a step preceding the first strand incision of NER, which is needed to produce the fragmented single-stranded DNA.
- 8.** The initiation step. In global NER, the XPC protein recognizes the lesion. In TCR, RNA polymerase recognizes the lesion, by stalling at the lesion.
- 9.** Both processes cleave the phosphodiester backbone, remove the pentose phosphate, then insert a correct dNTP and ligate the nick. In BER, a specific DNA glycosylase cleaves a damaged base from the pentose to form the abasic (AP) site.
- 10.** The most common DNA lesions leading to G-T mismatches are insertion of T opposite an O<sup>6</sup>-methylguanine lesion during replication and deamination of 5-methylcytosine. In both cases, it is the T that is incorrect and potentially mutagenic if not repaired.
- 11.** Frameshift mutations would lead to alteration of many amino acid residues in the protein product and could be caused by template slippage of DNA polymerase in the region with the repeated A residues.
- 12.** Reactive oxygen species generated during normal aerobic metabolism are a major source of DNA damage, because they form free radicals that react with the DNA.
- 13.** About 1,700 phosphodiester bonds derived from dNTPs are expended in the repair: 850 in the DNA degraded between the mismatch and the GATC sequence, and another 850 in the DNA synthesis needed to fill the resulting gap. ATP is hydrolyzed by the MutL-MutS complex and by the UvrD helicase.
- 14.** NER and BER occur only in double-stranded DNA. Both processes excise the damaged base or bases from the

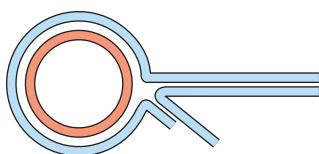
damaged strand, leaving a gap that can be filled in only if an undamaged complementary strand is present.

15. Each molecule of  $O^6$ -methylguanine methyltransferase is used only once and is degraded after the repair reaction. Thus, the reaction expends all the energy needed to synthesize the protein, along with all the energy used to mark the protein for degradation and to carry out that degradation.
16. (a) There may have been a trace contaminant of Pol III or Pol II in the preparation of UmuD' and UmuC. (b) Pol III is the main replicative DNA polymerase in the cell. A strain with a completely inactivated Pol III would be dead. The temperature sensitivity allows the cells to be grown at a permissive temperature but to be inactivated, when needed, by increasing the temperature. (c) Fraction 56 contains both Pol III and Pol V. The Pol V can replicate over the lesion, and Pol III can then extend the DNA. (d) The mutant Pol III is inactivated at  $47^\circ\text{C}$ . (e) Fraction 64 contains UmuC (and UmuD', not shown) almost exclusively. The lack of temperature dependence provides evidence that UmuC and/or UmuD' have a DNA polymerization activity.

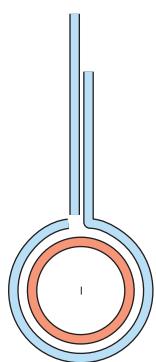
### Chapter 13

1. (1) The fork may bypass the lesion. If the DNA backbone is intact, the fork may either (2) stall or (3) leave the lesion behind in a single-strand gap. (4) If the fork encounters a break in a template strand, one arm of the fork is lost.

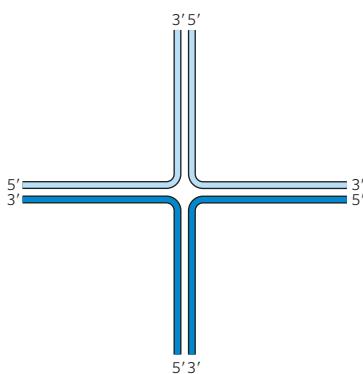
2. (a)



- (b)

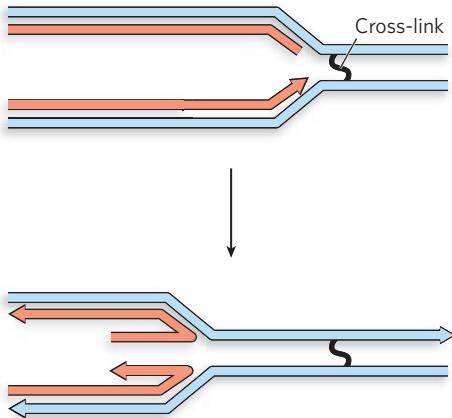


- 3.



4. (a) RecB. (b) RecB. (c) RecD. (d) RecC. (e) RecC.

- 5.

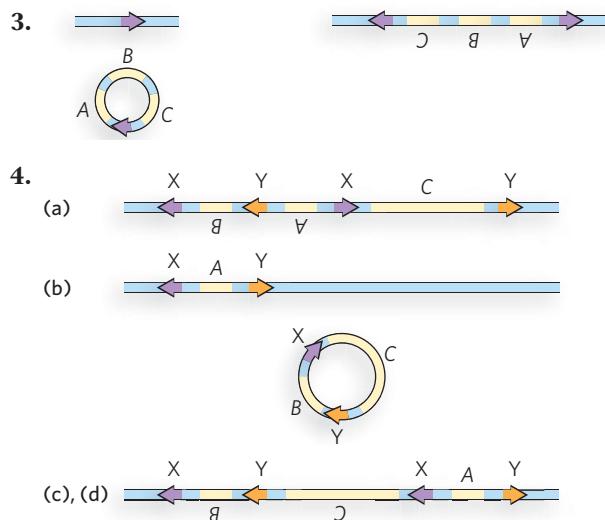


6. During normal growth, forks collapse at sites where there is a break in the template strand. The absence of RecBCD in the mutants will curtail the repair of such double-strand breaks, leading to the increase in linearized chromosomes.
7. If gene conversion occurs across the region where the sequence difference between *A* and *a* is located, this will create a heteroduplex intermediate that has an *A*-containing strand paired with an *a*-containing complement. The mismatch at this gene locus will be resolved one way or the other by mismatch repair, resulting in the loss or gain of information.
8. NHEJ involves degradation by nucleases and processing of the DNA ends, leading to some loss of base pairs.
9. The information at *HML $\alpha$*  would be subject to change. Whatever information was present in *MAT* would be transferred to *HML $\alpha$* .
10. The polyacrylamide gel separates proteins on the basis of molecular weight. For a protein covalently linked to DNA, the molecular weight is that of the combined protein and DNA. If the size of the linked DNA varies, the protein-DNA complex will have a highly variable molecular weight and will appear as a smear in the gel; detecting a single protein band would be impossible. Nuclease digestion eliminated this variability and allowed Spo11 to be detected as a discrete band.
11. (a) Points Y. (b) Points X.
12. During desiccation, DNA repair is impossible because of its requirement for metabolic energy in the form of ATP and dNTPs. Without a functioning cellular metabolism, ATP cannot form. However, DNA degradation by nucleases requires no ATP or other cofactors. Thus, chromosomes with double-strand breaks formed during desiccation are rapidly degraded, eventually destroying the genome, unless the DNA ends at the break are protected by proteins such as DdrA.
13. (a) The single-stranded DNA substrate is degraded before agarose gel electrophoresis. In addition, restriction enzymes do not cleave single-stranded DNA.

**(b)** 5'→3'. **(c)** As strand exchange is completed, fragments 1b and 1a become part of a larger fragment, fragment 1. The DNA in this larger fragment has a strand break (nick) at the position where the duplex was originally cleaved by restriction enzyme A.

## Chapter 14

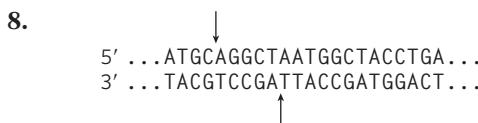
- In tyrosine-class site-specific recombination systems, the Holliday intermediate is generated in a precise set of cleavage and strand-transfer steps, all occurring at a single, unique DNA sequence. In homologous recombination, Holliday intermediates can appear at any sequence and are generated by the strand invasion and branch migration promoted by RecA recombinases.
- Homologous recombination occurs only where two DNA molecules have identical or very similar sequences over a significant region; any sequence is permitted. Site-specific recombination occurs only at particular DNA sequences that are recognized, bound, and recombined by the recombinases. Transposition, with a few exceptions, can occur at almost any sequence.



- (a)** Recombination at the X sites inverts the intervening sequence because the X sites are in opposite orientations. **(b)** Recombination at the Y sites deletes the intervening sequence because the Y sites are in the same orientation, leaving only one copy of the X site and one of the Y site on the DNA. **(c)** If the X sites react first, the orientation of one Y site changes, leading to an inversion in the later reaction with the Y sites. **(d)** If the Y sites react first, deletion occurs. Reaction of the (deleted) X site on the circle with the (remaining) X site on the original DNA yields the product shown in (c).
- (b)** The Hin-hix system works only when the two *hix* sites are in the opposite orientation and on the same supercoiled DNA molecule.
- Replicative transposition generates a cointegrate intermediate that joins the donor and target DNAs together. The cointegrate is resolved (the DNAs are separated) by

the transposon-encoded site-specific recombinase (the Tn3 resolvase).

- The λ site-specific system requires additional proteins, besides the Int recombinase, to promote the reaction. To adapt this system to eukaryotic cells, these additional proteins, as well as the recombinase, would have to be expressed. The λ *attP* site is also more complex, and is a larger segment of DNA to clone, than the FRT and lox sites.



- The tRNA, carried by the virus from one host to another, provides the primer for DNA synthesis by reverse transcriptase.
- The reverse transcriptase of TP (non-LTR) retrotransposons makes use of a 3' end derived from the target DNA to prime viral DNA synthesis. The endonuclease cleaves a phosphodiester bond in the target, exposing the needed primer terminus.
- Exons are the coding regions of genes. Insertion of a transposon of any type into an exon would almost certainly disrupt the activity of the protein encoded by the gene.
- (a)** If Cre were used for the insertion, the cassette itself could be altered during the insertion process. **(b)** When the Flp reaction is complete, intact FRT sites flank the cassette. Each FRT could be a target for a new insertion event. **(c)** The core sequence (see Figure 14-1a) must be modified to prevent recombination between the different Cre sites. **(d)** Using R, Y, and C to denote RFP, YFP, and CFP, the possible combinations are: RRR, YYY, CCC, RRY, RRC, RYY, RCC, YCC, YYC, and RYC.

## Chapter 15

- The deletion would move the -35 sequence closer to the -10 sequence by half a helical turn of the DNA, putting the two elements on opposite faces of the DNA duplex. This would dramatically reduce binding of sigma factor to the promoter, thereby decreasing transcription efficiency.
- 5'-AUGACCAUGAUUACG. A sequence reported for any gene is, by convention, that of the coding strand, and sequences are always written in the 5'→3' direction.
- At 50 to 90 nucleotides per second, the enzyme would take 34 to 61 seconds to transcribe the gene.
- The *tesB*-10 sequence deviates more from the consensus sequence, and its higher G≡C content will be more difficult to melt. Assuming these genes use similar transcription initiation modes, more *tesA* mRNA is expected relative to *tesB*.
- The number of transcripts is expected to increase, because Pol III typically generates many more transcripts than Pol II for the genes it transcribes. tRNAs, which are synthesized by Pol III, are required in much greater quantities in the cell than are most mRNAs, which are made by Pol II.

6. Initiation (including abortive initiation), the slowest step; elongation, the fastest step but punctuated by pauses; termination, requiring stalling and subsequent dissociation of the polymerase from the DNA.
7. About 8 to 10 phosphodiester bonds can be formed in the initiation phase. RNA synthesis always begins at a unique location (defined as +1) relative to the promoter sequence. Elongation can be prevented by controlling the sequence of the DNA template strand and the rNTPs added to the reaction, and the reaction products are defined by the same parameters. For example, if the first G residue in the DNA template strand does not appear until position +6, and you add ATP, UTP, and GTP, but no CTP, RNA synthesis will be limited to the first five nucleotides and the dominant reaction product is likely to be pentanucleotides. This strategy has been used in research on RNA polymerase. (See, for example, W. R. McClure, C. L. Cech, and D. E. Johnston, *J. Biol. Chem.* 253:8941–8948, 1978.)
8. Many RNAs are present in the body, synthesized before ingestion of the  $\alpha$ -amanitin. The mRNAs support cell and tissue function until they are degraded, and typically last for about two days.
9. Intercalation means that the planar portion of a small molecule inserts into the double-helical DNA between successive base pairs, deforming the DNA and preventing movement of RNA polymerase along the template. Actinomycin D and acridine act in a similar way.
10. This is difficult to do experimentally. The TATA box and the Inr sequence are often, but not always, found upstream from genes transcribed by Pol II. A significant minority of eukaryotic promoters lack a well-defined TATA box—the so-called “TATA-less” promoters—and this can confound identification by computer algorithms alone.
11. No. The two strands of a DNA molecule are antiparallel and complementary (not identical). When the promoter is inverted, it will direct RNA synthesis that uses what was originally the coding strand as the template strand. The mRNA sequence, derived from a different DNA strand and synthesized in the opposite direction, would be very different from the mRNA produced by the original gene and might not even contain an open reading frame.
12. There may be a palindromic sequence that allows formation of a hairpin structure during transcription, creating a  $\rho$ -independent terminator that is not perfectly efficient, or an imperfect rut sequence to guide the loading of the  $\rho$  protein. In both cases, the sequences would have to be imperfect so that some transcripts were elongated through gene B.
13. (a) 10–13. (b) 5–8. (c) 2–13. (d) 5–8. (e) Transcription from the P<sub>HS</sub> promoter requires a factor present in fractions 5–8. (f) Transcription from P1 and P2 uniquely requires the RNA polymerase reconstituted with the 32 kDa protein, and the designation “ $\sigma^{32}$ ” is probably warranted. (g) The use of an alternative sigma factor is an efficient way to coordinate regulation of transcription of a group of genes that are not always required by the cell.

## Chapter 16

1. Inactivation of the polymerase would lead to incomplete pre-mRNA processing, including 3' end formation, splicing, editing, and transport—and would certainly be lethal.
2. Two; one to cleave the 5' exon-intron junction and one to join the two exons, with release of the intron.
3. Group I and group II introns are generally self-splicing, with only the RNA backbone of the intron required for the reaction *in vitro*. The nucleophile in the first step, cleavage of the 5' splice site, is a guanine nucleotide or nucleoside for group I introns, and the 2'-OH of an internal A residue for group II introns. The second step uses the liberated 3'-OH at the 5' splice site as nucleophile to attack the phosphodiester bond at the 3' splice site, joining the exons. Site specificity is aided by guide sequences that are part of the intron structures. Introns removed with spliceosomes follow a mechanism similar to group II introns, but also rely on the spliceosome ribonucleoprotein complex to catalyze their excision.
4. Self-splicing catalyzes phosphodiester exchange reactions with no net loss or gain of energy. Bonds are broken and re-formed with different nucleotides, and there is no change in the number of phosphodiester bonds, just in the covalent bonding partners. Thus, there is no net change in free energy from reactants to products.
5. Incubate the total RNA in the presence of an appropriate buffer to ensure RNA folding. Add a fluorescently labeled guanine nucleotide analog to label the 5' ends of any RNA that uses this nucleotide as a nucleophile in the first step of splicing. After reaction, fractionate the RNA (by size) on a polyacrylamide gel and look for fluorescently labeled RNAs. Controls could include incubation with other labeled nucleotides or with just the fluorescent dye, in parallel reactions—which are not expected to result in labeled RNAs, based on the group I intron reaction mechanism.
6. Over many generations, the wild-type bacteria would win out and the mutant strains would disappear. The normal rRNA modifications add subtle but important thermodynamic stability to the ribosome structure, enabling more robust protein synthesis in the wild-type bacteria.
7. Sometimes. By definition, enzymes remain unchanged after catalysis, which is not the case for self-splicing and self-cleaving RNAs. They are catalysts because they enhance the rate of bond cleavage and/or joining, but in their natural form they have only one reaction turnover. However, some of these RNAs can be engineered to bind to separate substrate molecules and catalyze reactions on multiple such molecules, so they are inherently capable of functioning as enzymes. Some, such as RNase P, do so in their natural state.
8. The binding and hydrolysis of GTP by Ran ensures a cycle in which mRNA is bound in the nucleus and released in the cytoplasm, and not the reverse.
9. The lifetime would increase, because mRNAs in higher eukaryotes are degraded by 5'→3' exonuclease digestion.

- 10.** In each case, the nucleotide is modified by removing an exocyclic amine from the six-membered ring and replacing it with a keto oxygen. The change converts C to U and A to I.

**11.** The most important properties of living systems are the capacity to catalyze reactions and the capacity to store information that defines the structure of the catalysts. RNA has both of these properties.

**12.** 5'-AAUAAA and a GU-rich sequence. The 5'-AAUAAA is retained in the mature mRNA.

**13.** The 2', 3', and 5' hydroxyls of ribose. The 3'-OH and 5'-OH remain linked to the nucleotides they were bonded to before the reaction. The 2'-OH is linked to the G residue on the 5' end of the intron.

**14.** Inactivation of tRNA nucleotidyltransferase. The CCA-3' would not be added to the tRNA, so amino acids could not attach.

**15.** **(a)**  $\alpha$ -Amanitin inhibits Pol II. The rRNA precursors that the researchers wanted to isolate are synthesized by Pol I, and addition of  $\alpha$ -amanitin eliminated a lot of background RNA synthesis. **(b)** Some very stable or tightly bound protein may have remained and catalyzed the intron-splicing reaction. **(c)** The reaction requires a guanine nucleoside or nucleotide with a 2'-OH (RNA form) and a free 3'-OH. **(d)** This experiment is a compelling demonstration of self-splicing and RNA catalysis. There are no introns in bacterial rRNA. Expressing the rRNA segment with the use of bacterial enzymes eliminated the possibility that contaminating intron-splicing enzymes from *Tetrahymena* were present. The *Tetrahymena* rRNA gene is the only *Tetrahymena* macromolecule in the reaction mixture and was deproteinized before the experiment.

Chapter 17

1. There are two possible codons: AUA and UAU; only two amino acids can be incorporated into a polymer, and the only polymer produced is poly(Tyr-Ile).
  2. The codon for methionine is AUG, so an RNA that produces poly(Met) must be a repeating sequence of AUG. There are three reading frames for  $(AUG)_n$ . The  $(AUG-AUG-AUG)_n$  frame produces poly(Met); UGA is a termination codon, and no polypeptide forms; and  $(GAU-GAU-GAU)_n$  produces poly(Asp).
  3. Start and stop codons are in red.

5'-CCG-AUG-CCA-UGG-CAG-CUC-GGU-GUU-ACA-AGG-CUU-GCA-UCA-GUA-CCA-GUU-UGA-AUCC-3'  
 Met-Pro-Trp-Gln-Leu-Gly-Val-Thr-Arg-Leu-Ala-Ser-Val-Pro-Val-(stop)

4. Four possible RNA sequences can encode Met-Asn-Trp-Tyr (the variations in codons are in red):

ATG-**AAU**-UGG-**UAU**  
ATG-**AAC**-UGG-**UAU**  
ATG-**AAU**-UGG-**UAC**  
ATG-**AAC**-UGG-**UAC**

The addition of a Leu residue, which has 6 possible codons, increases the number of possible RNA sequences to 24: any of the 6 codons—UUA, UUG, CUU, CUC, CUA, or CUG—can be added to the end of each of the four sequences above.

5. Met-Tyr-Gln. These are the first three amino acids beginning at the second AUG (initiation) codon: AUG-UAU-CAG. The first reading frame, AUG-UGU-UGA, ends in a stop codon after only two amino acids.

6.

5'-AUG-GGU-CGU-GAG-UCA-UCG-UUA-AUU-GUA-GCU-GGA-GGG-GAG-GAA-UGA-3'  
 Met-Gly-Arg-Glu-Ser-Ser-Leu-Ile-Val-Ala-Gly-Gly-Glu-Glu-(stop)

Ten tRNAs are needed for these 14 amino acids (two to encode the three Gly; one to encode both Ser; one to encode the three Glu; and one for each of the other residues).

7

5'-AUG-GGU-CGU-GAG-UCA-UCG-UUA-AUU-GUA-GCU-GGA-GGG-GAG-GAA-UGA-3'  
 Met-Gly-Arg-Glu-Ser-Ser-Leu-Ile-Val-Ala-Gly-Gly-Glu-Glu-Trp

The peptide is 15 amino acids long, instead of 14; 10 tRNAs are needed, one for each different amino acid.

8

Met-Ile-Leu-Leu-Ser-Trp-Thr  
5'-AUG-AUA-UUG-CUA-UCU-UGG-ACU-3'

Transitions: C A U G C C Six positions  
 Transversions: C/U A/G A/G Three positions

9. Met-Pro-Ala-Glu-Val. A tRNA<sup>Tyr</sup> with an anticodon mutation (e.g., 5'-AUA to 5'-UUA) will suppress the first stop codon, resulting in Met-Pro-Ala-Glu-Val-Tyr-Ser-Glu-Ala.

- 10.** (a) Poly(Arg-Glu). (b) Poly(Val-Cys). (c) Poly(Glu), poly(Thr), and poly(Asn). (d) Poly(Lys), poly(Glu), and poly(Arg). (e) Poly(Leu-Leu-Thr-Tyr).

**11.** Leu:  $U_2C$  and  $UC_2$  (observed proportion 22.2; calculated proportion 20). Phe: UUU and  $U_2C$  (obs. 100; calc. 100). Pro:  $UC_2$  and CCC (obs. 5.1; calc. 4). Ser:  $U_2C$  and  $UC_2$  (obs. 23.6; calc. 20).

**12.** The anticodon contains an inosine: 5'-ICC.

**13.** The maximum distance for a four amino acid change is 11 nucleotides between the insertion (X) and deletion:

CAT-XCA-TCA-TCA-TCA(omit T)-CAT

The minimum distance is 7 nucleotides:

CAT-CAX-TCA-TCA-T(omit C)AT-CAT

14.

A-CGU-CGA-GUA-GCA-GUA-UCG-AUU-GAG-CUC-UUA-GAU-AAG-AUC-GC

The other reading frames would encode stop codons (red) in the middle of a protein:

ACG-UCG-AGU-AGC-AGU-AUC-GAU-**UGA**-GCU-CUU-AGA-**UAA**-GAU-CGC  
AC-GUC-GAG-**UAG**-CAG-UAU-CGA-UUG-AGC-UCU-**UAG**-AUA-AGA-UCG-C

15. Only one set of answers is shown.

Amino Acid	Codon	Anticodon
Phe	UUU	GAA
	UUC	
Leu	UUA	UAA
	UUG	
	CUU	GAG
	CUC	
	CUA	UAG
	CUG	
Ile	AUU	IAU
	AUC	
	AUA	
Met	AUG	CAU

16. (a) Three different codons are present in the three reading frames of this oligonucleotide, and thus a maximum of three different tRNAs might bind. (b) With limited knowledge of the genetic code, possible explanations included the presence of a four-base code rather than a triplet code, or a relaxed specificity due to reaction conditions. (c) AAG was assigned to Lys. (d) AAA had already been assigned to Lys in the earliest code-cracking experiments that used homopolymeric poly(A). (e) GAA was assigned to Glu. (f) Increased  $Mg^{2+}$  concentrations and lower temperatures relaxed the specificity of the binding. (g) AAG = Lys; GAA = Glu; AGA = Arg. (h) GAC and GAU encode Asp. The Asp-tRNA<sup>Asp</sup> may have elicited a positive signal in some experiments, because coding specificity is weakest in the third (wobble) position of the codon.

3. Using polysomes, the cell can produce several protein molecules on a single mRNA molecule. Because mRNAs have an average lifetime of just a few minutes in the cell, polysomes maximize the number of proteins that can be made per unit time.

4. Proline is incorporated during protein synthesis; post-translational processing adds the hydroxyl group.

5. Isoleucyl-tRNA synthetase sometimes catalyzes the addition of valine to tRNA<sup>Ile</sup>, an amino acid that is similar to but smaller than isoleucine and can readily fit into the synthetase active site. Histidine has no close structural analogs among the amino acids, greatly lowering the chance that tRNA<sup>His</sup> will be charged with the incorrect amino acid.

6. Yes. In polypeptide synthesis, removal of the last amino acid added (by hydrolytic cleavage of the last peptide bond to form) would sever the covalent link between the polypeptide and the tRNA in the ribosomal P site. This would terminate polypeptide synthesis.

7. IF-2: the 70S ribosome would form, but initiation factors would not be released and elongation could not start. EF-Tu: the second aminoacyl-tRNA would bind to the ribosomal A site, but no peptide bond would form. EF-G: the first peptide bond would form, but the ribosome would not move along the mRNA to vacate the A site for binding of a new EF-Tu-tRNA.

8. (a) The synthetase recognizes the G-U base pair (between G3 and U70) in the amino acid arm of tRNA<sup>Ala</sup>. (b) The mutant tRNA would insert Pro residues at codons that specify Ala. (c) A mutation in tRNA<sup>Pro</sup> that allowed it to be recognized and aminoacylated by Ala-tRNA synthetase would have similar effects. (d) Such changes would insert Pro at many inappropriate sites in polypeptides, inactivating many proteins, and thus would be lethal.

9. An unknown location, but *not* the nucleus. The protein will be bound by SRP and transported as it is synthesized into the ER lumen. Its fate from there would depend on other signals. The NLS would not be utilized, because it would not be accessible to the cytosolic proteins that normally bind it.

10. Chloramphenicol inhibits bacterial protein synthesis *and* mitochondrial protein synthesis. The effects on mitochondrial ribosomes give rise to the human toxicity.

11. (a) Poly(UG) generates polypeptides containing Phe, Leu, Val, Cys, Trp, and Gly. (b) Ala is not normally present in polypeptides synthesized in response to poly(UG). (c) Yes. In both cases, about one-third of the added label is incorporated into the product polypeptide, although the yield is slightly reduced in the Raney nickel-treated preparation. (d) This experiment was a control to confirm that Raney nickel treatment did not have a general deleterious effect on tRNA. (e) In most scientific studies, it is useful to confirm a result by two or more different methods. The extra experiment rendered the final conclusion unambiguous.

## Chapter 18

- 1,600 ( $4 \times 400$ ) phosphoanhydride bonds. Typically, four ATP/GTP molecules are hydrolyzed per amino acid incorporated into protein.
- (a) Amide or peptide. (b) Ester. (c) Phosphodiester. (d) Hydrogen bonds. (e) Noncovalent bonds, including hydrogen bonds and van der Waals forces.

- 12.** (a) The 50S subunit contains the peptidyl transferase activity of the ribosome. It includes the parts of the P and A sites that interact with the 3' ends of charged tRNAs. (b) The CCA sequence at the 3' end of the hexanucleotide is present at the 3' end of every tRNA and is required for specific binding to the 50S subunit. (c) The fMet-oligonucleotide must bind in the P site, and the puromycin in the A site. (d) The reaction is much simpler and does not rely on added protein factors such as initiation and elongation factors, which would be inactivated by the protein-removal treatments. (e) Given the capacity of *T. aquaticus* to grow at high temperatures, the rRNA in the 50S subunits might be particularly stable and able to withstand protein removal. (f) The result in Figure 2 shows that chloramphenicol and carbomycin strongly inhibit the reaction, so the reaction is indeed catalyzed by the peptidyl transferase activity of the 50S subunit. In addition, the RNA component of the subunit is essential for activity. (g) The protein extraction procedures were thorough, but there was a possibility that small amounts of protein remained associated with the 23S rRNA.
- 3.** Heterochromatin is highly condensed and transcriptionally inert, because the histone proteins make promoters inaccessible. The less-condensed euchromatin has undergone a structural remodeling, allowing some regions to be transcribed. The alterations include covalent modification (such as acetylation) of histones and displacement of nucleosomes, creating exposed regions of DNA that are probably binding sites for regulatory proteins.
- 4.** The primary transcript of an miRNA that is an stRNA is about 70 nucleotides long, with self-complementary internal sequences that form hairpin structures. The precursor is cleaved into 20- to 25-nucleotide partial duplexes, one strand of which can bind complementary stretches in cellular mRNAs. This binding can inhibit gene expression by blocking translation or facilitating mRNA degradation.
- 5.** Positive regulation. Positive and negative regulation are defined in terms of the type of protein involved in the regulation. Regulation by an activator is positive regulation; regulation by a repressor is negative regulation.
- 6.** A leucine zipper. The motif contains Leu (L) residues at every seventh position. It often functions in transcription factors to form a dimer interface by forming a coiled-coil.
- 7.** The regulon may be subject to combinatorial control, in which subsets of the regulon genes are needed in certain circumstances. The 13 genes may be subject to regulation by another regulatory protein—either another repressor that needs to be removed or an activator that needs to be present for transcription to occur.
- 8.** Most repressors with helix-turn-helix motifs function as oligomers, many as homodimers. When the plasmid-encoded mutant repressor is synthesized at high levels in the cell, most of the wild-type repressor molecules synthesized are incorporated into less functional heterodimers with a mutant repressor subunit.
- 9.** Regulatory proteins with helix-turn-helix, helix-loop-helix, or homeodomain motifs generally function as dimers and bind to sequences with inverted repeats. Proteins with zinc finger motifs can function as monomers and have no constraint to bind inverted repeats. Thus, strings of zinc finger motifs can be linked together to bind almost any sequence, whether it contains repeated elements or not.
- 10.** On binding a hormone molecule, the steroid hormone receptor dimerizes and the hormone-receptor complex is transported into the nucleus.
- 11.** There are several possible explanations. Many eukaryotic genes are regulated by more than one activator protein, and another activator may be needed. Many eukaryotic genes are encapsulated (and silenced) in heterochromatin, and remodeling of the chromatin may be required in the region where the gene is located to allow activation. The protein may need to be modified, and the modifying enzyme (e.g., kinase) may not be present in the cell. Finally, perhaps the activator cannot be transported into the nucleus.

## Chapter 19

- 1.** Screening of a genomic library to identify interacting partners of a “bait” protein is one of the common uses of the two-hybrid assay. To complete the screening system, you make a genetic library of “prey” fusion proteins by splicing a cDNA library containing random genes from the organism (the source of the bait protein) with a gene encoding a transcription-activation domain (e.g., the activation domain of Gal4p). You transfer the prey plasmid library into yeast, along with the bait fusion plasmid. Expression of β-galactosidase (as reporter gene) in the transformed yeast could be measured by the presence of blue colonies on plates containing X-gal (see Chapter 7); this will occur only in cells where the prey fusion protein interacts with the bait fusion protein. You can then sequence the prey gene to identify the interacting partner.
- 2.** Because the 18 bp site is a near palindrome, the activator probably functions as a dimer. Given that both protein A and protein B are required for activating gene X, they may form a heterodimer that has specificity for the binding site. This could be tested in an electrophoretic mobility shift assay or any other assay that measures DNA binding. Alternatively, the 18 bp site might be bound by a third, unidentified protein, and proteins A and B might bind different DNA sites. To test this, the site could be used in a functional DNA-binding assay to follow the binding protein during purification, allowing identification of the correct protein. Footprinting could be used as an assay, and the DNA sequence could be used to make an affinity chromatography resin to aid purification. A final possibility is that protein A and/or protein B interact with a different protein to bind the 18 bp site. This could be tested by purification using a functional assay such as footprinting. The purified active protein would reveal the additional protein.

- 12.** **(a)** The regulatory sequences for the chromosomal *GAL1* gene were known to respond only to Gal4p, and the DNA-binding elements of Gal4p had been removed in the fusion protein. **(b)** Given the finding that the fusion protein functioned as a repressor in *E. coli*, the researchers knew that the DNA-binding elements were properly folded and bound to the normal LexA-binding sites. **(c)** The LexA-Gal4p fusion protein is functional and stimulates gene expression to a level similar to that stimulated by Gal4p at UAS<sub>G</sub>. **(d)** LexA by itself does not activate gene expression in this system. **(e)** Positioning of the LexA operator 577 bp rather than 178 bp away from the transcription start site lowers transcription activation by about 25%. **(f)** When the upstream sequences contain UAS<sub>G</sub> or the 17mer, expression depends on the cellular Gal4p, which in turn is expressed only in the presence of galactose. **(g)** The LexA protein itself does not activate transcription; instead, a segment of Gal4p (that does not include the DNA-binding elements) is required. Thus the LexA part of the LexA-Gal4p fusion is not altering the DNA structure in any way that facilitates transcription. Instead, the Gal4p portion must be directly interacting with RNA polymerase.
- 7.** Repressor concentration is  $\sim 8 \times 10^{-9}$  M. The number of moles of repressor = 10 molecules/ $6.02 \times 10^{23}$  molecules/mol =  $1.66 \times 10^{-23}$  mol. Dividing by the volume of the cell in liters gives the concentration:  $(1.66 \times 10^{-23}$  mol)/( $2 \times 10^{-15}$  L) =  $8 \times 10^{-9}$  mol/L. This is about five orders of magnitude greater than the dissociation constant for the repressor; as a result, the operator site will almost always be bound by repressor.
- 8.** **(a), (c), (d), (e)** Operon expression decreases. **(b)** Operon expression remains unchanged, as it is already running at maximum expression.
- 9.** **(a)** Less attenuation and thus increased transcription. If the ribosome translating sequence 1 did not block sequence 2, sequences 2 and 3 would pair more often. **(b)** This could decrease the efficiency of pairing between sequences 2 and 3, leading to an increase in attenuation and thus decreased transcription. **(c)** No attenuation would occur, and transcription would proceed at the maximum level. **(d)** The attenuation system would respond more to histidine than to tryptophan concentration. **(e)** Attenuation would decrease and thus transcription would increase, because sequences 2 and 3 would pair more often. **(f)** Attenuation would increase and thus transcription would decrease, because pairing of sequences 3 and 4 would almost always occur.

## Chapter 20

- 1.** The *E. coli* cells will produce  $\beta$ -galactosidase when they are subjected to high levels of a DNA-damaging agent, such as UV light. Under such conditions, RecA binds to single-stranded chromosomal DNA and catalyzes cleavage of LexA, releasing LexA from its binding site and allowing transcription of downstream genes.
- 2.** **(a)** With the Lac operator on the other side of the *lac* operon, the *lac* genes would no longer be subject to repression by the Lac repressor. **(b)** Inactivation of the binding site for CRP would reduce or eliminate expression of the *lac* genes under all conditions. **(c)** Alterations in the promoter sequence could either increase or decrease expression of the *lac* genes, depending on the particular alteration.
- 3.** The conformation of AraC is altered by arabinose binding, and the protein binds different sites on the DNA when arabinose is absent than when arabinose is present. In the absence of arabinose, AraC binding blocks polymerase access to the promoter.
- 4.** **(a)** With high tryptophan concentration (added to the medium), tryptophan synthase levels drop due to attenuation of mRNA translation, even if the mRNA is stable. **(b)** Tryptophan synthase levels remain high for a considerable time. **(c)** Tryptophan synthase levels decline rapidly.
- 5.** **(a)** An uncleavable LexA protein would permanently repress the SOS genes and block induction of the SOS response. **(b)** Weakened LexA binding would lead to constitutive expression of the SOS response genes.
- 6.**  $\sim 7,000$  copies.  $(10 \text{ copies}/E. coli \text{ cell}) \times (3.2 \times 10^9 \text{ bp/haploid human cell})/(4.6 \times 10^6 \text{ bp}/E. coli \text{ cell}) = 7,000 \text{ copies/human cell.}$
- 10.** The TPP-binding riboswitch alters mRNA conformation when TPP is bound, making the Shine-Dalgarno sequence inaccessible to ribosome binding. The glucosamine 6-phosphate-binding riboswitch also changes the conformation of its mRNA, but in this case the change activates a ribozyme ribonuclease function that cleaves and inactivates the mRNA.
- 11.** The repressor protein would probably block translation of the mRNA even when the protein's binding sites on the ribosome were available. This would tend to slow the assembly of active ribosomes.
- 12.** The  $\lambda$  prophage produces the  $\lambda$  repressor and cII proteins, which will bind to any  $\lambda$  phages entering the cell and prevent induction of a lytic pathway.
- 13.** These mutants indicate that activation is not caused by CRP changing the local DNA structure near a promoter; the activator must do more than simply bind to DNA. Because the mutant CRP cannot activate transcription in response to low glucose levels, cells with this mutation will not grow well on lactose or other secondary sugars for which the metabolizing-enzyme genes use CRP as transcription activator.
- 14.** Because genes in an operon are transcribed together from one promoter as a polycistronic mRNA, they can be regulated by one set of activators and/or repressors. In this way, enzymes required for a common pathway can be synthesized together. To express the operon genes at different levels, translational regulation could be used: altering the ribosome-binding sites for each gene, or binding of translational repressors to one or a few of the genes.

- 15.** An advantage is that the signal sensor is contained within the mRNA itself, and regulation does not require a separate sensor molecule to be synthesized or maintained. A disadvantage is that coordinated gene expression through integration of different cellular signals is difficult.
- 16.** Conserving energy is important, but careful modulation of gene expression in response to changing growth conditions is paramount. It is preferable to “waste” energy synthesizing partial transcripts that are unused unless antitermination is triggered, so that the cell is primed to respond quickly to a sudden need for enhanced gene expression.
- 17.** No. Eukaryotic transcription occurs in the nucleus, and translation in the cytoplasm. The spatial and temporal separation prevents the coupling of transcription and translation regulation that can occur in bacterial operons such as the *trp* operon.
- 18.** (1) These mRNAs tend to have unusually long sequences upstream from the translation start site that are necessary to form the three-dimensional structure of the riboswitch. (2) The upstream sequence is conserved in the same mRNA in different bacterial species, and sometimes in archaea, plants, and fungi. (3) These RNAs do not have protein binding partners, consistent with their ability to function in directly regulating gene expression.
- 19.** (a) Band A is, in part, the undigested attenuated mRNA, and this undigested RNA accounts for one of the three bands. The two additional A bands in the denaturing gel indicate the presence of a good cleavage site for RNase T1 near the middle of the RNA. Because the two parts of the RNA separate only on the denaturing gel, sequences on either side of the T1 cleavage site must be paired and thus migrate together in the first gel. (b) Bands B and C are segments of the larger attenuated mRNA. (c) Both band B and band C have paired RNA sequences that protect the RNA from RNase T1. Because band C is sometimes cleaved in two (the gel has one band of undigested RNA and two bands derived from the cleavage), there must be a loop of significant size near the end of the paired sequences in this segment. (d) The band B RNA includes sequences 3 and 4. The loop at the end of this hairpin is small enough to limit T1 cleavage. (e) The band C RNA derives from paired regions in sequences 1 and 2. (f) The pairing of sequences 2 and 3 is not present in this analysis. This hairpin must be less stable than the pairing of sequences 3 and 4 and/or of sequences 1 and 2.
- mechanisms are needed to open up the chromatin when a gene must be expressed. Chromatin is not present in bacteria.
- 2.** Eukaryotes often use multiple regulatory proteins to activate transcription of a single gene. Given the large size of eukaryotic genomes, the chance of nonspecific binding of a given regulator to DNA sequences unrelated to that particular gene is too great for the cell to rely on a single regulatory protein.
- 3.** Gal4p is a transcription factor that largely functions as a transcription activator of the *GAL* genes. Gal11p is likely to be involved in regulation of a much larger set of genes. (In fact, Gal11p is part of the yeast Mediator coactivator complex.)
- 4.** (a) Phosphorylated. (b) Phosphorylated. (c) Unphosphorylated. (d) Unphosphorylated.
- 5.** First, multiple transcription factors (activators) act on most genes, and different (often unique) combinations of factors are used at different genes. Second, families of activators form heterodimers, such that a family of four related proteins can make a total of 10 different dimeric species that can recognize 10 different DNA sequences.
- 6.** CTFC is a key component of the gene insulator system in eukaryotes; it binds to sequences that prevent inappropriate activation of certain genes during development and in the mature organism. Its loss would be lethal.
- 7.** HMG proteins bind DNA and facilitate DNA bending. The presence of HMG protein-binding sites in the DNA that interacts with an enhanceosome may reflect HMG-mediated facilitation of DNA wrapping around the enhanceosome structure.
- 8.** Gene expression is generally silenced in regions of heterochromatin. If the gene were essential, as most housekeeping genes are, moving it into heterochromatin would be lethal for the cell.
- 9.** (a) With elimination of the nuclear import signal, the receptor can bind the hormone in the cytoplasm, but the complex will not be imported into the nucleus to activate gene expression. (b) With elimination of the interaction with Hsp70, most of the receptor molecules would be in the nucleus, where they would not have access to the hormone signal.
- 10.** (a) A hypersensitive site is a region of uncondensed (or less condensed) chromatin where the DNA is accessible to nuclelease action. The decondensed state signals a site for the binding of proteins that function in genome maintenance and/or gene regulation. (b) The LCR triggers chromatin remodeling to both sides and can help activate transcription of genes to both sides. (c) The constructs are integrated at random locations in the chromosome. Some may accidentally be integrated near an LCR that could activate transcription from a site to the 5' side. (d) Yes. Many more colonies are observed when the  $\lambda$  DNA is flanking the neomycin-resistance gene than when the chicken hypersensitive site is flanking this

## Chapter 21

- 1.** There are many more genes in a eukaryote, and most are not organized into operons. A huge number of repressors would be needed at all times if negative regulation predominated. Eukaryotic DNA is packaged into chromatin, which effectively silences most genes, and special

gene. (e) Because the constructs were integrated at random locations in the genome, they might have been subjected to the differential effects of sequences on either side of the integration sites that could affect expression of the neomycin-resistance gene. Sometimes, multiple copies of the constructs might have been integrated. If these effects were present, the hygromycin-resistance gene provided a way to normalize the results. (f) No. The effect could be at the level of posttranscriptional modification or translation. (g) To show that the effect was at the level of transcription, one would have to directly measure the production of mRNA from the genes in stably transfected cell lines (which the authors did).

## Chapter 22

- The effect of actinomycin D reflects simply a blockage of RNA synthesis. The effect of cycloheximide may reflect the requirement for a newly synthesized protein factor in the signaling pathway for induction of the gene encoding this mRNA.
- The N-terminal amino acid residues of Sxl expressed from  $P_e$  (encoded by exon E1) differ from the N-terminal residues of Sxl expressed from  $P_m$  (encoded by exon L2).
- The RNA-binding site for proteins that carry out 3'-end cleavage and polyadenylation is defined most reliably by the sequence AAUAAA. In the DNA, the sequence is AATAAA, located downstream from the final gene exon. Because AATAAA is a consensus sequence, scanning only for this sequence would not find all of the 3'-end cleavage sites; some will vary slightly from the consensus.
- Reticulocytes are the precursors of red blood cells, filled with hemoglobin and highly specialized for oxygen transport; their nucleus is destroyed during maturation. Almost all the mRNA deposited in a reticulocyte before destruction of the nucleus encodes hemoglobin. There are essentially no other translation-dependent cellular functions to disrupt.
- (a) AAUAAA is a signal for 3'-end cleavage and polyadenylation. (b) AUUUA is an ARE motif that limits mRNA stability.
- (1) mRNAs could be deposited at one end of the oocyte, such as by *Drosophila* nurse cells. (2) Proteins could be deposited at one end of the oocyte. (3) mRNAs or proteins could be actively transported from one part of the oocyte to another. (4) A set of mRNAs or proteins could be subjected to differential stability by the introduction, at one end of the oocyte, of factors leading to mRNA or protein degradation.
- Anterior cells of the embryo, where Mex-3 is concentrated and normally suppresses Pal-1 production, would take on fates similar to those of cells at the posterior end.
- The *bcd* mRNA needed for development is contributed to the egg by the mother. The fertilized egg develops normally, even if its genotype is  $bcd^-/bcd^-$ , as long as the mother has one wild-type *bcd* allele (and thus contributed the mRNA to the oocyte) and the *bcd^-* allele is recessive. However, an adult  $bcd^-/bcd^-$  female will be sterile because she cannot generate *bcd* mRNA for her oocytes.
- The cap cells may create a niche for the germ-line stem cell, providing extrinsic signals that orient the cell division to ensure that one daughter maintains the stem cell identity. In this case, the cap cells supply protein ligands that activate a Bmp-family signaling pathway in the stem cell that represses at least one gene required for differentiation.
- The ability to feed bacteria to the worms allows an RNAi approach. Given knowledge of the genome sequence and the types of genes that might be involved (e.g., Wnt-class signaling genes, homeotic genes), you could devise an RNAi screen. Short double-stranded RNAs complementary to worm genes are expressed in bacteria, using gene segments cloned between opposing promoters on a bacterial plasmid, and the bacteria are fed to worms. The heads of the fed worms are cut off, and the RNAi clones that affect regeneration are determined. Such an experiment has been done, with the finding that one homeobox gene in the *piwi* family plays a key role. (See P. W. Reddien et al., *Dev. Cell* 8:635–649, 2005; P. W. Reddien et al., *Science* 310:1327–1330, 2005.)
- (a) As described in Chapter 7, the three-hybrid method is designed to screen for RNA sequences that bind to a particular protein. The method used here was modified to screen for unknown proteins that bind to a particular RNA sequence. (b) The control shows that a single nucleotide change in UCUUU abolishes binding. (c) FBF-1 may be binding to some other sequence in the engineered RNA, and the various control sequences help eliminate that possibility. (d) The FBF-1 not only localizes to the germ line but is predominantly present in the cytoplasm. (e) The animals should (and most do) produce only sperm.

# Index

*Note:* Page numbers followed by f, t, and b indicate figures, tables, and boxed material, respectively. Page numbers preceded by A refer to the Model Organisms appendix.

A site, 621–623, 623f, 624  
**a/α** mating type, 470–471, 471f, 746–747, 747f, A-8, A-9f  
AAA+ proteins, 157–158, 379, 391–392  
Abasic site, 417  
Abegg, Richard, 68  
Abortive initiation, 527–528, 529f  
Abrin, 647b  
Accommodation, in translation, 638, 639f  
Acetylation, 163–165, 164f, 206  
  histone, 348–350, 351b, 352t, 354f, 360, 736  
  posttranslational, 654  
Achirality, 78  
Acid, pK<sub>a</sub> of, 83  
Acid dissociation constant (*K*<sub>a</sub>), 83  
Acidity, pH and, 81–84, 82f  
Acridine, 521  
Actinomycin D, 521, 522f  
  in transcription inhibition, 768–769, 769f  
Activation, gene, 669  
  site-specific recombination in, 489, 489f  
Activation energy, 84–85, 85f  
  in catalytic reactions, 147–149  
  definition of, 148  
Activators, 669–670, 670f, 745b.  
  *See also* Transcription factors  
  acting as repressors, 742  
  acting with repressors, 673  
AraC, 707, 707f  
  coactivators and, 671, 672f, 740f, 741–742  
distance from promoters, 737, 741  
DNA binding domains of, 679–684, 683f, 745b  
  vs. regulatory domains, 683–684, 684f  
DNA-binding functions of, 739–742, 740f  
effectors and, 674–675, 674f, 675f  
in enhanceosomes, 752, 752f  
in eukaryotes, 736–742, 745b  
hormone receptors as, 742  
Active sites, 146–147, 146f  
Adam, Y chromosome, 288, 290  
Adaptor hypothesis, 15, 586, 608  
ADAR (adenosine deaminase acting on RNA), 566  
Adelman, Leonard, 191b  
Adenine (A), 48, 62, 177, 177f, 178. *See also* Base(s); Base pairs/base pairing  
  deamination of, 203–204, 416–417, 417f  
  methylation of, 65–66, 67f, 205  
  nomenclature for, 64t  
  prebiotic chemistry of, 19, 19f, 89b  
  synthesis of, 19, 19f

Adenosine, 178, 179f  
  functions of, 181–182, 185f  
Adenosine 3',5'-cyclic monophosphate.  
  *See* cAMP  
Adenosine deaminase acting on RNA (ADAR), 566  
Adenosine diphosphate (ADP), synthesis of, 84, 85f, 88  
Adenosine monophosphate. *See* AMP (adenosine monophosphate)  
Adenosine triphosphate. *See* ATP (adenosine triphosphate)  
S-Adenosylmethionine (adoMet), 182, 205  
Adenylation, 163–165, 164f  
  in translation, 625, 626f  
A-DNA, 188, 189, 189f  
AdoMet (S-adenosylmethionine), 182, 205  
ADP (adenosine diphosphate), synthesis of, 84, 85f, 88  
ADP-ribosylation, 163–165, 164f  
Adriamycin (doxorubicin), 319b  
Adult stem cells, 799  
*Advice to a Young Scientist* (Medawar), 16  
Affinity chromatography, 100–101, 100f  
  tags for, 240–241, 240t, 241f  
Africa, human evolution in, 287–289  
Aging, telomerase and, 399, 401  
Agrawal, Rajendra, 622, 622f  
*Agrobacterium tumefaciens*, Ti plasmid of, 221  
Ahringer, Julie, 767  
Alanine, 99t  
  tmRNA charging with, 651  
Ala-tRNA synthetase, 651  
Alberts, Bruce, 384, 385f  
Alhazen, 13, 13f  
Aliphatic side chains, 97  
Alkaline phosphatase, in recombinant DNA technology, 217t  
Alkalinity, pH and, 81–84, 82f  
Alkaptonuria, 46  
Alkylation, mutations due to, 418  
Allele(s), 26  
  independent assortment of, 28–29, 29f, 29t  
  mutant, 38  
  segregation of, 25–28, 28f, 56  
  wild-type, 38–40, 38f  
Allele replacement, site-specific  
  recombination in, 489, 489f  
Allis, David, 360  
Allolactose, 703f  
  Lac repressor and, 701, 701f  
Allopatric speciation, 288  
Allosteric enzymes, 161–164  
Allosteric model, of transcription termination, 539, 539f  
Allosteric modulators, 161  
Alpha carbon atom (C $\alpha$ ), 64, 65f, 99–102, 102f  
  chirality and, 79–80, 80f  
Alpha helix, 104–105, 104f  
  hydrophobic residues in, 111–112, 113f  
  interaction of, 111–112, 112f  
  recognition, 679, 679f  
  in ribbon diagrams, 107, 108f  
  supersecondary structures and, 110–112, 111f, 112f  
Alpha subunit, of Pol III, 377–378, 378f, 378t  
Alpha/beta barrel, 110–111, 111f  
Alternative splicing, 555–556, 557f, 684.  
  *See also* Splicing  
  definition of, 769  
  in sex determination, 769–771, 770f  
Altman, Sidney, 6  
Alu transposon, 504  
Alzheimer's disease  
  linkage analysis in, 274–275, 276  
  PS1 gene in, 275  
 $\alpha$ -Amanitin, in transcription inhibition, 522, 523f  
Ambros, Victor, 805  
Ames test, 420, 421f  
Amino acid(s), 64–65, 65f. *See also* Protein(s)  
  abbreviations for, 97, 99t  
  in alpha helix, 104–105, 104f, 105f  
  availability of, stringent response and, 719–720, 720f  
  biosynthesis of, 708–710  
  bonds between, 69–70, 70f  
  buried, 115, 115f  
  central carbon atom of, 64, 65f, 97, 97f  
  chains of, 64, 65f  
  length of, 97  
  chemical modifications of, 65–66, 167f, 178, 205–206, 205f  
  chirality of, 79–80, 80f, 81b  
  codons specifying, 586–590.  
  *See also* Codon(s); Genetic code  
  covalent modification of, 164–167, 165t  
  hydrophobic, in alpha helix, 111–112, 113f  
  most common, 97, 99t  
  peptide bonds of. *See* Peptide bonds  
  phosphorylation of, 165–167, 165t  
  polypeptide chains and, 99–102  
  postsynthetic changes in, 66–67, 67f  
  posttranslational modification of, 654  
  properties of, 97–98, 99t  
  protein half-life and, 690, 690t

- Amino acid(s) (*continued*)  
 R groups of, 97–98, 98f  
 side chains of, 97–98, 97f, 98f  
 structure of, 64–65, 65f, 97–98, 97f, 98f  
 types of, 64  
 unnatural, incorporation into proteins, 630
- Amino acid sequences, 48, 49, 50, 97–103  
 evolutionary divergence of, 102–103  
 notation for, 99
- Amino terminus, 99
- Aminoacyl-AMP, in translation, 625
- Aminoacylation, in translation, 621–624, 623f
- Aminoacyl-tRNA, 587, 616, 621–624, 623f  
 in elongation, 638
- Aminoacyl-tRNA synthetases, 587, 588, 625–629  
 in amino acid activation, 625, 626f, 632t  
 in amino acid attachment, 587–588, 625  
 classes of, 625  
 proofreading by, 627–628, 628f  
 specificity of, 625–626, 627f  
 structure of, 625, 626f  
 substrate discrimination by, 625–626, 627f, 628, 628f
- Aminopeptidases, 632–633
- AMP (adenosine monophosphate)  
 cyclic. *See c*-AMP  
 DNA ligase and, 153–156, 154f–155f
- Amphipathic helix, 112, 113f
- Analytes, in mass spectrometry, 280
- Anaphase, 35f, 36  
 in meiosis, 35f, 36  
 in mitosis, 34, 35f
- Anemia, sickle-cell, 52b  
 malaria and, 54–55
- Aneuploidy, 466b
- Anfinsen, Christian, 115
- Angles  
 bond, 71, 71f  
 disulfide, 102, 103f
- Animals, transgenic, 238–239
- Annealing, 200–201, 200f  
 synthesis-dependent, 447–448, 449f, 450f, 469–470, 470f, 471, 471f
- Antagonists, 787
- antennapedia* gene, 797, 797f, 798
- Anterior-posterior axis, in development, 791, 793–795, 794f
- Anthracyclines, 421, 422f
- Antibiotics  
 enzyme inhibition by, 153b  
 resistance to, 302, 303b, 504  
 topoisomerase inhibition by, 318b  
 transcription inhibition by, 521–522, 522f, 768–769, 769f  
 translation inhibition by, 646–647, 648t–650t, 768, 768f
- Antibodies, 505, 506f  
 gene arrangement of, 505–507, 506f, 507f  
 in immunofluorescence, 242  
 protein chips and, 282
- van der Waals interactions and, 74–75, 75f
- Anticodon(s), 587, 588  
 definition of, 588  
 notation for, 588  
 variant, 604b–605b  
 wobble and, 589–590, 589f, 590f
- Anticodon-codon pairing, 587–588, 588f, 638
- Antigens  
 van der Waals interactions and, 74–75, 75f
- Antiparallel  $\beta$  sheet, 105f, 106
- Antiparallel helix, 186
- Antiretroviral agents, 153b  
 development of, 166b–167b
- Antitermination, transcriptional, 539, 539f, 725, 725f
- Anti-TRAP, 728
- Antiviral vaccines, 503
- AP endonuclease, in base excision repair, 431, 431f
- AP-1 transcription factors, 747–748, 748f
- APOBEC enzymes, 204, 568
- Apoenzymes, 145
- Apoproteins, 145
- Apoptosis, 401
- Aqueous solutions, pH of, 81–82
- ara* operon, 707–708, 707f
- Arabidopsis thaliana*  
 chromosomes of, 302t  
 DNA of, 302t  
 genome of, 302t  
 life cycle of, A–15  
 as model organism, A–3–A–4, A–14–A–15
- Arabinose operator, 707–708, 707f
- AraC activator, 707, 707f
- AraC repressor, 707, 707f
- Arago, Dominique, 78
- araI*<sub>1</sub>, 707, 707f
- araO*<sub>2</sub>, 707, 707f
- Arber, Werner, 218
- Archaea  
 evolution of, 9, 9f  
 genome sequencing for, 269
- Architectural regulators, 671, 672f
- Arginine, 97–98, 98f, 99t
- Argonaute, 784, 785
- A-RNA, 197  
 helix of, 197
- Arnheim, Norman, 215
- Aromatic side chains, 98
- Artemis, 474, 474f
- Artificial chromosomes, 299  
 bacterial, 222–223, 223f  
 human, 299–300  
 yeast, 223–224, 224f
- Ashe, Mark, 768
- Asparagine, 97, 98f, 99t  
 glycosylation in, 67f
- Aspartate, 98, 98f, 99t
- Association constant ( $K_a$ ), 137
- Astbury, William, 104
- Asymmetry, in development, 789–790, 790f
- AT-AC introns, 558
- A-to-I RNA editing, 567, 568f
- Atomic orbitals, 70
- Atoms  
 electronegative, 69, 69f  
 electropositive, 69  
 valence of, 70, 71f
- ATP (adenosine triphosphate)  
 chromatin remodeling complex and, 344–345, 346t  
 DNA ligase and, 153–156, 154f–155f  
 helicases and, 157–158, 157f  
 hydrolysis of, 84, 85f, 88–89, 158–160, 159f  
 motor proteins and, 158–160, 159f  
 in intracellular energy transfer, 88, 89b  
 in nucleosome translocation, 735  
 in prebiotic chemistry, 89b  
 in replication, 379–380, 380f, 381f, 383, 391–394  
 in strand exchange, 457, 459f  
 structure of, 88  
 in translation, 625, 644  
 in translocation, 158–160, 159f  
 as universal energy source, 182
- ATPases  
 AAA+, 379, 391–392  
 helicases and, 157–159
- ATP-binding proteins, sequence analysis of, 130, 130f
- ATP-coupling stoichiometry, 158
- Attenuation, transcriptional, 709–710, 709f
- AU-rich elements, 780–782, 782f
- Australopithecus*, 288, 289f
- Autographa californica* multicapsid nucleopolyhedrovirus (AcMNPV), 237–238, 237f
- Autoinducers  
 in *Vibrio cholerae*, 729  
 in *Vibrio harveyi*, 697
- Autoinhibition, 163  
 enzymatic, 163, 163f
- Autosomes, 37
- Auxotrophs, 46, 47f, 420  
 in Ames test, 420
- Avery, Oswald T., 45, 45f, 46
- Avidin, dissociation constant for, 138t
- AZT  
 for HIV infection, 502b  
 mechanism of action of, 153b
- B cells  
 differentiation of, 505–507, 506f  
 IgM and, alternative 3' end cleavage in, 771, 772f
- Bacillus subtilis*, TRAP system of, 710, 728
- Back mutations, 420, 421f
- Bacmids, 237–238
- Bacon, Francis, 13, 14
- Bacon, Roger, 13

- Bacteria  
 antibiotic-resistant, 302, 303b, 504  
 chromosomes of, 298, 301–302, 302t  
 in cloning, 236, 236f  
 DNA of, 301–302, 302f, 302t  
 packaging of, 301–302, 302f, 302t, 341–342, 343f  
 DNA replication in  
   initiation of, 391–394, 391f–393f, 405  
   termination of, 395–398  
 evolution of, 7b, 9, 9f.  
   *See also* Evolution  
 gene regulation in, 697–720  
 genes of, 302t  
 genetic code alterations and, 603–604  
 genome sequencing for, 261, 261f, 266–269, 292  
 as model organisms. *See Escherichia coli*  
 plasmids of. *See Plasmids*  
 quorum-sensing, 697, 711, 729  
 radiation-resistant, 7b  
 transcription in, 523–532  
 transformation in, 221–222  
 translation in  
   elongation in, 638–642  
   initiation of, 629, 631–634, 632t, 633f–635f  
   termination of, 643, 643f  
   virulent vs. nonvirulent, 44–46  
 Bacterial artificial chromosomes (BACs), 222–223, 223f  
 Bacterial transduction, 487  
 Bacteriophage(s)  
   DNA of, 300–301, 301t  
   gene regulation in, 720–726  
   introns in, 555  
   lytic and lysogenic pathways of, 485–487, 486f, 497, 721–722, 721f, 722f  
   switching between, 722–725, 724f  
   promoters in, 722–725, 722f, 724f  
   prophage, 722  
   prophage induction in, 722–725, 724f  
   structure of, 722f  
   transcription in, antitermination of, 725, 725f  
   viral particles in, 300–301, 301t  
 Bacteriophage λ, 170, 301t, 485, 486f  
   exonuclease, 217t  
   site-specific recombination in, 509  
 Bacteriophage lambda. *See Lambda (λ) phage*  
 Bacteriophage Mu, 510–511  
   transposons in, 497, 497f  
 Bacteriophage P1, 485, 486f  
 Bacteriophage Qβ, 300  
 Baculoviruses, in cloning, 237–238, 237f  
 Baltimore, David, 500, 500f  
*Banshee* transposons, 505t  
 Bardet-Biedl syndrome, 281  
 Barnett, Leslie, 595, 595f  
 Barr bodies, 351b, 756  
 Bartel, David, 805  
 Basal transcription factors, 739–742, 740f.  
   *See also* Transcription factors  
 Base(s), 45, 62–64, 63f  
   alkylation of, 418–419  
   deamination of, 203–204, 205f, 416–417, 417f  
   by nitrous acid, 418, 418f  
   depurination of, 417  
   hydrophobic stacking of  
    in DNA, 181, 184f  
    in RNA, 197  
   insertion/deletion of, in RNA editing, 566, 567f  
   methylation of, 65, 67f, 178, 205–206, 205f, 417  
   minor, 178  
   nomenclature for, 64t  
   oxidative damage to, 418, 419f  
    repair of, 430, 431f  
   pairing of. *See* Base pairs/base pairing  
   purine, 177, 177f  
   pyrimidine, 177, 177f  
   structure of, 180–181  
   substitution of, 66, 67f  
   tautomers of, 180–181, 183f  
   wobble, 589–590, 589f, 590f  
 Base excision repair, 430–433, 431f, 432f, 433t  
   long patch, 432, 431f  
   short patch, 432, 431f  
 Base pairs/base pairing, 4f, 47–48, 62–64, 63f, 181, 185, 518, 518f  
   accuracy in, 370–375, 370f, 375f  
   Chargaff's rules for, 185  
   definition of, 48  
   discovery of, 44–45  
   in DNA–protein binding, 141–142, 142f  
   Hoogsteen, 191, 193f  
   mistakes in  
    proofreading for, 371–372, 371f, 375, 403, 405, 425, 520  
    repair of. *See* DNA repair  
   in nucleosome binding, 336, 336f  
   in Pol I active site, 374  
   processivity and, 158, 375–376, 379–380, 380f, 381f, 404  
   in replication, 370–375, 370f, 371f, 374f, 375f, 403, 425  
   resonance structures and, 71, 72f  
   ribozymes and, 576, 577  
   in RNA, 196–197, 197f, 198f  
   in RNA interference, 783–786  
   RNA polymerases and, 518  
   in splicing, 558, 561f  
   in tetraplex DNA, 192, 193f  
   in three-dimensional DNA structures, 194b–195b  
   in transcription, 48, 49, 520–521, 616–617  
   in translation, 587–588, 624, 638  
   in triplex DNA, 191, 193f  
   wobble in, 589–590, 589f, 590f  
 Base stacking  
   in DNA, 181, 184f  
   in RNA, 197–198  
 Basic helix-loop-helix motif, 680–681, 681f  
 Basic leucine zipper motif, 680  
 Bassler, Bonnie, 697, 729  
 Bateson, William, 801–802  
*BBS5* gene, 281–282  
 B-cell lymphoma, microRNA in, 667  
 B-DNA, 188, 189, 189f  
 Beadle, George, 46–47, 46f  
 Beads-on-a-string nucleosomes, 334, 334f, 339  
 Belfort, Marlene, 495, 495f  
 Beneden, Edouard van, 34  
 Bentonite, in prebiotic evolution, 20, 20f  
 Benzer, Seymour, 662  
 Benzo[a]pyrene, as carcinogen, 419, 420f  
 Berg, Paul, 216, 216f, 241  
 Berger, James, 23, 24  
 Beta barrel, 109–110, 110f  
 Beta hairpins, 109, 110f. *See also* Hairpins  
 Beta sheet, 105–106, 105f  
   in ribbon diagrams, 107, 108f  
   supersecondary structures and, 109–111, 110f, 111f  
 Beta sliding clamp, 378, 378t, 379–380, 379f, 380f  
   clamp loaders for, 379–380, 380f, 381f  
   DNA polymerase processivity and, 379, 380f, 404  
   Pol III detachment from, 385–387, 386f–388f  
   protein binding by, 386–387  
   recycling of, 386–387, 388f  
 Beta subunit  
   of DNA polymerase, 107, 108f  
   of RNA polymerase, in bacteria, 519, 519f, 535  
 Beta turns, 106, 106f, 109, 110f  
 Beta-alpha-beta motif, 110–111, 111f  
*bicoid* gene, 793, 794  
 Bicoid protein, 570, 749, 750, 750f, 793–795, 794f  
 Biggin, Mark, 762  
 Binding energy ( $\Delta G_B$ ), 146, 148f, 149  
 Binding sites, 107, 136  
   DNA-binding protein, 139, 141, 142f  
   ribosome, 621–623, 623f, 629  
 Biochemical standard free-energy change ( $\Delta G^\circ$ ), 147  
 Biological diversity  
   independent assortment and, 469, 469f  
   metagenomic sampling of, 268b  
   sources of, 271, 288–291. *See also* Evolution  
   crossovers as, 468–469, 469f  
   mutations as, 5, 10–11, 50, 411–412, 801  
 Biological information, 2  
   flow of, 5f, 11  
    central dogma of, 47–50, 48f  
    direction of, 47–50, 48f  
 Biological literature mining, 282

- Biology  
 evo-devo concept and, 5, 801. *See also*  
 Development; Evolution  
 molecular. *See* Molecular biology  
 systems, 277
- Bioluminescence, in green fluorescent protein, 242, 243f, 253
- Biotechnology, 215–253, 278  
 affinity chromatography in, 100–101,  
 100f, 240, 240t, 241f  
 cDNA libraries in, 242–243, 244f  
 cloning in, 216–226  
 definition of, 217  
 DNA amplification in, 226–228,  
 227f, 228f  
 DNA libraries in, 224–225, 225f,  
 242–243, 244f  
 DNA microarrays in, 248–250,  
 248f–250f, 278  
 in nucleosome localization, 345–346  
 in transcriptome analysis, 278
- DNA sequencing in, 171, 228–233, 229f,  
 234b–235b, 278
- enzymes in, 217–219, 217t  
 in forensics, 230b–231b, 284b  
 fusion proteins in, 240–241,  
 240–243, 282
- gel electrophoresis in. *See* Gel  
 electrophoresis
- hybridization techniques in, 201–203,  
 202f–204f  
 oligonucleotide synthesis for, 206, 207f
- immunofluorescence in, 242, 243f
- immunoprecipitation in, 244–246, 245f,  
 282, 345
- locating genes in, 242–250
- mass spectrometry in, 280–281
- Northern blotting in, 203
- oligonucleotide-directed mutagenesis in,  
 239–240, 239f
- polymerase chain reaction in, 15,  
 219, 226–228, 227f, 228f,  
 230b–231b, 252
- protein amplification in, 233–236
- protein expression in  
 in bacteria, 236, 236f  
 in yeast, 236–237
- protein purification methods in,  
 244–246, 245f
- protein-protein interactions in, 244–248
- in proteomics, 278–281
- restriction endonucleases in, 217–219,  
 217t, 220f
- reverse transcriptase in, 501
- RNA amplification in, 228
- site-directed mutagenesis in,  
 239–240, 239f
- site-specific recombination in, 488–489,  
 489f, 490b–491b
- Southern blotting in, 203
- in transcriptome analysis, 278
- transposons in, 499b
- unnatural amino acids incorporation  
 in, 630
- Western blotting in, 243–244, 245f, 762
- yeast three-hybrid analysis in, 247–248,  
 247f, 282
- yeast two-hybrid analysis in, 246, 247f, 282
- Biotin  
 in immunofluorescence, 242
- Bird influenza virus, 129
- Blackburn, Elizabeth, 398, 398f
- BLAST algorithm, 263
- Bleomycin, 421, 422f
- Blobel, Günter, 655, 655f
- Block, Stephen, 542
- Blood, pH of, 81, 82
- Blunt ends, 218, 220f
- Bmp4* gene, 802
- Bonds, 68–73  
 angle of, 71, 71f  
 torsion (dihedral), 102, 103f  
 breaking of, 72  
 covalent, 68–70, 68f  
 double, 70–71, 71f  
 polar, 69, 72  
 single, 70–71, 71f  
 definition of, 68
- disulfide, 97, 131, 131f, 654
- in double helix, 69, 71, 74, 75, 75f, 181,  
 184f, 187
- electron distribution and, 72–73, 73f
- formation of, 72
- glycosidic, 177
- hydrogen. *See* Hydrogen bonds
- hydrolysis of, 88–89
- ionic, 68–70, 68f, 70f, 146–147
- molecular orbital model of, 70
- in nucleic acids, 178–180, 181f
- peptide. *See* Peptide bonds
- phosphodiester  
 of ATP, 88–89  
 in nucleic acids, 178–180, 181f
- phosphorus-nitrogen, 89
- phosphorus-oxygen, 89
- quantum mechanics and, 70–71
- resonance and, 71, 72f
- strength of, 69–70, 72
- strong, 68–70
- valence bond model of, 70
- weak, 73–78. *See also* Weak chemical  
 interactions
- Borate, ribose stabilization by, 585
- Boundary elements, 751–752, 751f
- Boveri, Theodor, 34, 56
- Bovine spongiform encephalopathy, 118–119
- Boyer, Herbert, 216, 216f, 251
- Brachystola magna*, chromosomes of, 36, 56
- Bragg, William Henry, 122
- Bragg, William Lawrence, 122
- Bragg's law, 122, 122f, 123
- Brainbow method, 490b–491b
- Branch migration, in DNA repair, 450, 451f,  
 456, 458f–460f, 459
- Branch points, in splicing, 557, 558f
- Branching evolution, 10, 10f
- BRCA* gene, 470
- Bread mold. *See* *Neurospora crassa*
- Breaker, Ronald, 712, 714f
- Breast cancer, 278  
*BRCA* gene in, 470  
 transcriptome analysis in, 278
- Breeding, plant, Mendel's experiments in,  
 25–31
- Brenner, Sydney, 48, 594–595, 595f
- Bridges, Calvin, 39–40
- Briggs, G.E., 149
- Brock, Thomas, 252
- 5-Bromo-4-chloro-3-indoyl-β-D-  
 galactopyranoside (X-gal), 703, 703f
- Bromodomains, 350, 352t, 353f, 354f,  
 356, 357f
- Brown, Patrick O., 325–326
- Brown, Robert, 32
- Bryant, Zev, 297
- Bubbles  
 in denatured DNA, 201, 201f  
 in replication, 392–393, 392t  
 in transcription, 520, 520f
- Budding yeast. *See* *Saccharomyces cerevisiae*;  
 Yeast
- Buffer capacity, 83
- Buffered solutions, 82–84
- Burley, Stephen, 536, 537f
- Bustamante, Carl, 297, 297f, 306
- Butterflies, wing development in, 806
- bZIP proteins, 680
- CA repeats, discovery of, 215
- CACTA transposons, 505t
- Caenorhabditis elegans*  
 chromosomes of, 302t  
 DNA of, 302t  
 dosage compensation in, 756  
 genome of, 302t  
 hermaphrodite, germ-line development  
 in, 776, 776f, 804
- life cycle of, A-12-A-13
- miRNA in, 805
- as model organism, A-3–A-4, A-12–A-13
- RNA interference in, 786
- stem cells of, 767
- Tc1/mariner transposons in, 497–498
- CAF-1 chaperone, 348, 349f, 354
- Cairns, John, 403
- Calcitonin-gene-related peptide, 556, 557f
- Calcium ions, in RNA, 198
- Calmodulin, structure of, 66f
- Calorie, 72
- C<sub>α</sub> (alpha carbon atom), 64, 65f, 97, 98f  
 chirality and, 79–80, 80f
- cAMP, 178
- ara* operon and, 707–708, 707f
- lac* operon and, 706–707, 706f
- as second messenger, 759f, 760
- as starvation signal, 720

- cAMP receptor protein (CRP), 112, 162–164  
*ara* operon and, 707  
 in combinatorial control, 742  
*lac* operon and, 675, 706–707, 706f
- Campbell, Allan, 509
- cAMP-responsive element-binding protein, in gene regulation, 759f, 760
- Camptothecin, 318b
- Cancer  
 breast  
*BRCA* gene in, 470  
 transcriptome analysis in, 278
- carcinogens in, 419  
 Ames test for, 419–421  
 chemotherapy for, 421, 422f  
 colon, 428b–429b, 429  
 defective DNA repair in, 446  
 DNA methylation in, 205  
 epigenetic control in, 355b  
 fusion genes and, 415–416  
 gene regulation in, 667, 668  
 microRNA in, 667, 685  
 microsatellite instability and, 429b  
 mismatch repair and, 428b–429b, 429  
 mutations in, 407, 412, 668  
 oncogenes and, 412  
 $p53$  in, 668  
 telomerase and, 399, 402  
 topoisomerase inhibitors for, 318b–319b  
 tumor suppressor genes and, 412
- Candida albicans*, genetic code alterations in, 603–604, 606
- Cap-binding complex, 550, 550f
- Capecchi, Mario, 490b
- Capping  
 mRNA, 548, 549–550, 550f, 552–553, 631.  
*See also* 5' cap  
 telomeric, in chromosomes, 399–401, 399f, 400f
- Carbohydrate side chains, posttranslational attachment of, 654
- Carbon, valence of, 70, 71f
- Carboxyl terminus, 99
- Carboxylation, posttranslational, 654, 655f
- Carcinogens, 419  
 Ames test for, 419–421
- Carroll, Sean, 806
- CASP competition, 116
- Catabolite gene activation protein, 112
- Catabolite repression, 703
- Catalysis, 2, 3, 87–88, 144–156. *See also* Enzyme(s)  
 activation energy and, 147–149  
 covalent interactions in, 149  
 in protein regulation, 163–165  
 definition of, 2  
 DNA ligase and, 151–156  
 enzyme-substrate interactions in, 85, 146, 149  
 hydrogen atoms in, 91  
 mechanisms of, 146–147  
 noncovalent interactions in, 148f, 149
- principles of, 146  
 reaction rate and, 147–151. *See also* Enzyme kinetics  
 Michaelis-Menten equation and, 150–151, 150f
- Catalytic RNA. *See* Ribozymes (catalytic RNA)
- Cate, Jamie, 175, 197
- Catenanes, 313, 316f
- Caudal, 794–795, 794f
- CCR5 mutation, 54
- cDNA (complementary DNA), 225, 501
- cDNA libraries, 225, 225f, 242–243
- Cech, Thomas, 6, 15, 559, 559f
- Cell(s)  
 daughter, 33, 34f, 35f  
 early studies of, 32, 32f  
 gamete (sex), 24  
 longevity of, telomerase and, 399–401  
 reprogramming of, 533b  
 somatic, 24  
 structure of, 32, 33f
- Cell cycle, 33–34, 34f  
 chromosome replication in, 394–395  
 meiosis in, 34–37, 35f, 464–469, 464f.  
*See also* Meiosis  
 mitosis in, 32–34, 34f, 35f, 36, 299.  
*See also* Mitosis
- Cell lysis, 722
- Cell membrane, 32, 33f  
 evolution of, 8–9, 8f, 602  
 protein targeting to, 657
- Cell signaling. *See* Signaling
- Cell specialization, evolution of, 9
- Cell surface receptors, 686–687
- Cell theory, 32
- Cellular function, in genome annotation, 263
- Cen proteins, 299
- CENPA, 346–347, 347f
- Centimorgan (cM), 42–43
- Central dogma, 47–50, 48f
- Centromere-binding proteins, 299
- Centromeres, 33, 298–299, 299f
- Centrosomes, 34
- CFTR mutation, 53–54, 54f
- cGMP (guanosine 3',5'-cyclic monophosphate), 172, 182
- Chain topology diagram, 110
- Chalfie, Martin, 253
- Chaperones, 119–120, 120f  
 histone, 348, 349f, 354  
 in protein folding, 654  
 in protein targeting, 657
- Chaperonins, 113, 119–120, 120f
- Chargaff's rules, 185
- Chase, Martha, 46, 58
- Chemical bonds. *See* Bonds
- Chemical modification. *See* Protein modification
- Chemical modification interference, 704b–705b
- Chemical protection footprinting, 704b
- Chemical reactions. *See* Reaction(s)
- Chemical shifts, in nuclear magnetic resonance, 125
- Chemistry. *See* Prebiotic chemistry
- Chemoheterotrophs, 9
- Chemotherapeutic agents, DNA-damaging, 421, 422f
- Chen, Irene, 1
- Chi sequences, 454–455, 455f
- Chiasmata, 41, 41f
- Chimpanzees  
 evolution of, 288, 289f  
 genetic diversity in, 290  
 genome of, 271–272
- ChIP-Chip, 345, 347f, 762
- ChIP-Seq, 345–346, 347f
- Chiral center, 78–79
- Chirality, 78–80, 79f, 80f, 81b  
 direction of, 104, 104f
- Chitin-binding domains, 240t
- Chloramphenicol, translation inhibition by, 646, 646f
- Chloroplast(s)  
 evolution of, 303–304, 621, 632  
 protein targeting to, 655–659. *See also* Protein targeting  
 ribosomes of, 619
- Chloroplast DNA (cpDNA), 303–304
- Cholera, autoinducers in, 729
- Chromatin  
 accessibility of, 344  
 condensation of, 735  
 gene regulation and, 686  
 SMC proteins in, 316–320, 317f, 320f  
 definition of, 332  
 DNA underwinding and, 308, 308f  
 epigenetic inheritance of, 352–357  
 euchromatin, 735, 735f  
 heterochromatin, 735, 735f  
 histone tails and, 336–338, 337f, 351, 353f  
 open vs. closed state of  
 in gene expression, 678  
 histone tail modification and, 351, 353f  
 histone variants and, 343–344, 735  
 nucleosomes and, 343  
 propagation of, 356–357, 357f  
 structure of, 338–343, 359  
 histone tail modifications and, 348–351, 349f, 735  
 histone variants and, 343–344, 735  
 nucleosome positioning and, 735  
 regulatory correlates of, 348–351  
 weak interactions in, 344, 345f
- Chromatin immunoprecipitation, 345
- Chromatin remodeling, 735–736, 735f, 736f.  
*See also* Chromatin, structure of  
 HMG proteins and, 741, 741f  
 regulation of, 735–736, 735f, 736f  
 in RNA splicing, 733, 736, 738b–739b
- Chromatin remodeling complexes, 343, 344–346, 344f, 346f, 346t

## I-6 Index

- Chromatography  
affinity, 100–101, 100f, 240, 240t, 241f  
column, 100–101, 100f, 101f  
gel-exclusion, 100, 100f  
ion-exchange, 100, 100f
- Chromatosomes, 338
- Chromodomains, 351, 353f, 356
- Chromosomal scaffolds, 341, 341f
- Chromosome(s), 32  
artificial, 299  
bacterial, 222–223, 223f  
human, 299–300  
yeast, 223–224, 224f
- autosomes, 37
- bacterial, 298  
artificial, 222–223, 223f
- circular, 298
- cohesin-based linkage of, 317f, 318–319, 320f, 464–465
- crossing over of, 40–41, 41f, 42f, 57
- deletion of, 412–413, 413f, 414, 415f
- dimeric, 460–461, 461f
- DNA packaging in, 341, 341f
- DNA-binding proteins and, 139–141
- duplication of, 32–34, 34f, 35f, 272, 273f, 414–415, 415f
- early studies of, 24, 32, 32f
- end extension in, 399–401, 399f, 400f
- fusion of, 272
- gene location on, 38–44  
mapping of, 41–44
- homologous, 30–31, 446, 464–465, 464f
- human artificial, 299–300
- inversions of, 272, 415, 415f
- nondisjunction of, 39–40, 39f
- recombination of, 40–41, 41f, 42f
- replication of. *See* DNA replication
- segregation of, 34–37, 35f, 56, 299, 313, 464–465, 464f  
aneuploidy and, 466b
- sex, 36–37, 37f  
X, 37, 37f  
inactivation of, 347–348, 350b, 351b
- Y, 37, 37f
- size of, 300–304, 301t, 302t
- specialized genomic sequences in, 298–304
- staining of, 32, 32f
- structure of, 298–304  
higher-order, 338–343, 342f  
regulation of, 343–358
- tandem repeats in, 271
- telomeric capping in, 399–401, 399f, 400f
- viral, 298  
yeast artificial, 223–224, 224f
- Chromosome theory of inheritance, 38–44
- Churchill, Mair, 741, 741f
- Chymotrypsin, catalytic mechanism of, 86
- cI. *See* Lambda λ repressor
- cII, in prophage induction, 722, 723
- cIII, in prophage induction, 722, 723
- Ciprofloxacin, 318b  
enzyme inhibition by, 153b
- Circular chromosomes, 298
- Circular dichroism spectroscopy, 131, 131f
- Circular DNA  
bacterial, 301
- catenanes and, 313, 316f
- closed, 307, 307f
- end replication problem and, 398
- recombination and, 487, 487f
- supercoiled, 307
- viral, 300–301
- Cis isomers, 99–102, 102f
- Cis-acting genes/gene products, 701
- Cis-acting histone modifications, 343
- Cisplatin, 421, 422f
- Cis-splicing, 565
- Clamp loaders  
in bacteria, 378, 378f, 378t, 379–380, 380f, 381f  
in eukaryotes, 389, 389f
- Clark, John A., 476, 476f
- Clay, in prebiotic evolution, 20, 20f
- Cleavage and polyadenylation specificity factor, 774–775
- Cleaver, James, 436
- Cloned genes  
alteration of, 239–240, 239f
- expression of, 233–238
- production of, 221–226. *See also* DNA cloning
- Clones, 216. *See also* DNA cloning
- Closed complex, 521, 521f, 526–527, 528f
- Closed-circular DNA, 307
- Closed-form Pol I, 374, 374f
- CMG complex, 387
- Coactivators, transcriptional, 671, 672f, 739, 740f, 741–742
- Coalescent theory, 289–290
- Cockayne syndrome, 540
- Coding strand, 519, 520f
- CODIS database, 230b–231b
- Codominance, 30, 30f
- Codon(s), 412, 413f  
amino acids specified by, 586–590.  
*See also* Genetic code
- definition of, 586
- degeneracy of, 586
- gaps between, 594–595, 595f
- nonoverlapping, 594, 594f
- nonsense. *See* Nonsense mutations
- notation for, 588
- reading frames for, 412–413, 413f, 590–591  
frameshift mutations and, 412–413, 413f, 594–595  
open, 591, 591f
- start (initiation), 591, 629  
variant, 604
- stop (termination), 591, 620, 642–643, 643f, 651
- genetic code alterations and, 603–604
- mitochondrial, 603, 603t
- premature, 547, 652–653, 653f
- variant, 604–606, 606f
- three-nucleotide structure of, 586, 588–589, 588f, 595–596
- wobble and, 589–590, 589f, 590f
- Codon bias, 606
- Codon families, 589
- Codon-anticodon pairing, 587–588, 588f, 638
- Coelenterates, bioluminescence in, 253
- Coenzymes, 145, 145f
- Cofactors, 145, 145f
- Cohen, Stanley, 216, 216f, 251
- Cohesins, 317f, 318–319, 320f, 465
- Coiled DNA, 341, 342f
- Coiled-coiled motif, 111–112, 112f
- Cointegrates, 493
- Collins, Francis, 261, 261f
- Collision release, 385
- Colon cancer, mismatch repair and, 428b–429b, 429
- Colony blot hybridization, 202–203, 203f
- Column chromatography, 100–101, 101f
- Combinatorial control, 677, 677f, 742–751  
in bacteria, 742  
definition of, 742
- in development, 748–750, 749f, 750f
- in eukaryotes, 742–751
- by heterodimerization, 747–748, 748f
- in yeast, 743–746, 743f, 746f  
in mating-type switches, 746–747, 747f
- Communication, intercellular. *See* Signaling
- Comparative genomics, 264–266  
phylogenetic profiling and, 281–282, 281f
- Competitive enzyme inhibitors, 152b–153b
- Complementary DNA (cDNA), 225, 501
- Complex transposons, 497, 497f
- Composite transposons, 496–497, 497f
- Computational biology, protein folding and, 115–121, 116f
- Computational protein design, 95, 115–116, 116f
- Concentration, notation for, 82
- Condensin, 317f, 319–320, 320f, 321f
- Consensus sequences, 165, 165t  
in bacteria, 524
- Conservative replication, 364, 365f. *See also* DNA replication
- Constitutive gene expression, 669
- Constructive interference, 122, 122f
- Contigs, 261–262, 262f
- Continuous deformations, in DNA, 306
- Cooperativity, 136–137, 139
- Copy-number amplification, 485
- Core histones, 334, 334f
- Core promoter, 534, 534f
- Core sequence, 534, 534f
- Corepressors, 671, 672f, 709
- Corey, Robert, 99, 104
- Correlation spectroscopy (COSY), 125, 126, 126f
- Correns, Carl, 32

- Coumarins, 318b  
 Coumermycin A1, 318b  
 Coupled reactions, energy transfer in, 89, 90b  
 Covalent bonds, 68–70, 68f  
   double, 70–71, 71f  
   polar, 69, 73  
   single, 70–71, 71f  
 Covalent modification, 164–167, 165t, 684, 686–692. *See also* Protein modification  
 Cozzarelli, Nicholas, 325–326  
 cpDNA, 303–304  
 CPE (cytoplasmic polyadenylation element), 774  
 CPE binding protein, 774  
 CpG sequences, 417  
 CPSF, 774–775  
 Craigie, Robert, 510–511  
 Cre recombinase, 487, 487f, 488, 490b–491b  
 CREB, in gene regulation, 759f, 760  
 Creighton, Harriet, 41, 57  
 Cre-Lox recombination system, 487, 487f, 488, 490b–491b  
 Creutzfeldt-Jakob disease, 118–119, 118f  
 Crick, Francis, 6, 14, 15, 24, 25, 47–48, 49, 176f, 594–595, 595f, 616, 621  
   adaptor hypothesis of, 15, 586, 608  
   wobble hypothesis of, 590  
 Crixivan (indinavir), 166b–167b  
 Cro, 722, 723, 723f  
 Crosses  
   testcrosses, 40  
   three-factor, 44  
 Cross-linking, 334  
 Crossovers, 40–41, 41f, 42f, 57, 446  
   double-strand breaks and, 446, 447f, 465–469, 468f  
   genetic diversity and, 468–469, 469, 469f  
   in homologous recombination  
    in bacteria, 446, 449  
    in eukaryotes, 465–469, 465f, 468f  
 CRP (c-AMP receptor protein), 112, 162–164  
   *ara* operon and, 707–708, 707f  
   in combinatorial control, 742  
   *lac* operon and, 675, 706–707, 706f  
 Cruciforms, 191, 192f, 310–311, 310f  
 Cryo-electron microscopy, 617  
 Crystals, DNA, 210  
 CTC-binding factor, insulators and, 752  
 C-terminal domain, 536, 537  
 C-terminus, 99  
 C-to-U RNA editing, 567, 568f  
 Cut-and-paste transposition, 492–493, 492f–494f  
   transposons in, 497–498  
     activation of, 499b  
 Cyclic AMP. *See* cAMP  
 Cyclin kinases, in replication initiation, 395  
 Cyclobutane ring, 422, 422f  
 Cycloheximide  
   transcription inhibition by, 768–769, 769f  
   translation inhibition by, 646, 646f  
 Cysteine, 97, 98f, 99t  
   disulfide bonds of, 97, 654  
   fatty acid modification in, 67f  
 Cystic fibrosis, 53–54, 54f  
   misfolded proteins in, 118  
 Cystic fibrosis transmembrane conductance regulator (CFTR), 118  
 Cytidine deaminases, in RNA editing, 568, 568f  
 Cytogenetics, 31–37  
 Cytokinesis, 34  
 Cytology, 31–32. *See also* Cell(s)  
 Cytoplasm  
   RNA transport to, 568–571, 569f, 570f  
   translational control in, 772–778  
 Cytoplasmic membrane, 32, 33f  
 Cytoplasmic polyadenylation element (CPE), 774  
 Cytosine (C), 48, 62, 177, 177f, 178, 179f.  
   *See also* Base(s); Base pairs/base pairing  
   deamination of, 203–204, 416  
   methylation of, 65–66, 67f, 205, 417  
   nomenclature for, 64t  
 Cytotoxic agents, 419  
   chemotherapeutic, 421, 422f  
 Dahiyat, Bassil, 95  
 Dalgarno, Lynn, 629  
 Dalton (Da), 54  
 Dam methylase, 377t, 393  
   in mismatch repair, 426, 428  
 Darwin, Charles, 2, 9, 10f, 15, 21, 24, 801  
 Databases  
   CODIS, 230b–231b  
   of disease-causing mutations, 277  
   genomic, 266–271, 277  
   of protein structure, 114, 121, 281  
   of transcriptome analysis, 278  
 Daughter cells, 33, 34f, 35f  
 DDE motif, 503, 504f  
 ddI (dideoxyinosine), for HIV infection, 502b  
 De Vries, Hugo, 32  
 Deacetylation, histone, 736, 736f  
 DEAD box, 130, 130f  
 Deadenylation, in translation, 777, 779f  
 Deadpan (Dpn) repressor, 769–771, 770f  
 Deamination, 203–204, 205f  
   mutations due to, 416–417, 417f  
   by nitrous acid, 418, 418f  
   of tRNA, 573  
   viral infections and, 204  
 Deductive reasoning, 13  
 Deformylase, 632  
 Degenerate code, 586, 588–589, 589f  
   mutations and, 591–592  
*Deinococcus radiodurans*  
   evolution of, 7b  
   radiation resistance in, 460, 462b–463b  
 Delbecco, Renato, 440  
 Delbrück, Max, 58  
 Deletion analysis, 744b–745b  
 Deletion mutations, 412–413, 413f, 414, 415f  
 Denatured DNA, 200, 200f  
 Density, superhelical, 309  
 Deoxyadenosine, 178, 179f  
 Deoxycytidine, 178, 179f  
 Deoxyguanosine, 178, 179f  
 Deoxyribonucleic acid. *See* DNA  
 Deoxyribonucleoside diphosphates (dNDPs), 371  
 Deoxyribonucleoside monophosphates (dNMPs), 370–371, 370f  
 Deoxyribonucleoside 5'-triphosphates (dNTPs), 370–371, 370f, 373–375, 374f, 375f  
 Deoxyribonucleotides, 62–64, 63f, 178, 179f. *See also* Nucleotide(s)  
 Deoxyribose, 62, 64  
 Deoxythymidine, 178, 179f  
 Dephosphorylation, in gene regulation, 687  
 Depurination, 417, 418f  
 DeRisi, Joe, 259  
 Descartes, René, 13–14  
 Development  
   anterior-posterior axis in, 791, 793–795, 794f  
   asymmetry in, 789–790, 790f  
   combinatorial control in, 748–750, 749f, 750f  
   dorsal-ventral axis in, 793–795, 794f  
   in *Drosophila melanogaster*, 791–798. *See also* *Drosophila melanogaster*, development in  
   Hox genes in, 796–798, 797f, 802  
   imprinting in, 755  
   of insect wings, 806  
   interspecies similarities in, 11f  
   muscle differentiation in, 763  
   pattern-regulating genes in, 792–798  
    homeotic, 793, 796–798, 797f  
    maternal, 793–795, 794f  
    segmentation, 793, 795–796, 796f, 797f  
   polarity in, 791, 793–795, 794f  
   segmentation in, 791, 793, 795–796, 796f, 797f  
   sex determination in, 36–37, 37f  
    alternative splicing in, 769–771, 770f  
    XO, 37  
    XY, 37  
   signaling in, 790–793, 795–796, 796f  
   stages of, 791–792, 791f  
   stem cells and, 798–880, 798f, 799f  
 Developmental biology, evolutionary biology and, 5, 801  
 Dextrorotary (D-form) enantiomers, 79, 80f, 81b  
 Dicer, 66f, 576, 685  
   in duplex DNA cleavage, 786, 786f  
   in long-stranded DNA cleavage, 784–785, 787  
 in pre-miRNA cleavage, 784, 785f  
 structure of, 66f

- Dickerson, Richard, 210  
 Dickerson dodecamer, 210, 210f  
 Dideoxy chain-termination (Sanger) sequencing, 171, 228–233, 229f  
 Dideoxyinosine (ddI), for HIV infection, 502b  
 Diffraction patterns, 122, 122f resolution and, 123, 123f  
 Dihedral angles, 102, 103f  
 Dimer(s), 107 heterodimers, 747–748, 748f histone-fold, 335, 335f homodimers, 112, 113f pyrimidine photorepair of, 429–430, 430f, 440 ultraviolet radiation and, 422  
 Dimeric chromosomes, 460–461, 461f  
 Dimerization, in gene regulation, 679  
 Dimerization motifs, 679–682. *See also* Motifs  
 DinI, 458  
 Dintzis, Howard, 596, 609, 631  
 Dipeptidyl-tRNA, fMet-tRNA conversion to, 638–640, 641f  
 Diphtheria toxin, translation inhibition by, 646–647  
 Diploids, 26  
 Discontinuous deformations, in DNA, 306  
 Diseases and disorders, 50, 51f alkaptonuria, 46 Alzheimer's disease, 275 Bardet-Biedl syndrome, 281 cancer. *See* Cancer Cockayne syndrome, 540 cystic fibrosis, 53–54, 54f misfolded proteins in, 118 databases for, 277 digenic, 276–277 Down syndrome, 466b fragile X syndrome, 52–53 gene therapy for, 300 genes causing, identification of, 274–277, 276f hemophilia, 50, 51f HIV/AIDS. *See* Human immunodeficiency virus infection Huntington disease, 50–52, 51f Kennedy disease, 52 Liddle syndrome, 691 Machado-Joseph disease, 52 polyglutamine (polyQ) diseases, 413–414 sickle-cell anemia, 52b–53b, 54–55, 412 single-gene, 275–276, 276f spongiform encephalopathies, 118–119, 118f stem cell therapy for, 799–800 transcriptome analysis in, 278 triplet expansion diseases, 413–414, 414t trisomies, 466b Werner syndrome, 412 xeroderma pigmentosum, 434, 436b, 539–540  
 Dispersive replication, 364, 365f  
 Dissociation constant ( $K_d$ ), 137 acid, 83 for protein-ligand interactions, 137–138, 138f  
*distal-less* gene, 806  
 Disulfide angles, 102, 103f  
 Disulfide bonds, 97, 131, 131f, 654  
 Diversity. *See* Biological diversity  
 D-loops, 458f  
 Dmc1, 467–468, 467f, 470  
 DNA, 23–58. *See also* Nucleotide(s) accessibility of, 344, 348–351 A-form, 188, 189, 189f annealing of, 200–201, 200f automated synthesis of, 206, 207f backbone of, 179–180, 181f, 186, 186f base pairing in, 186–187, 186f. *See also* Base(s); Base pairs/base pairing bending of, 190, 190f. *See also* DNA looping B-form, 188 blunt ends of, 218, 220f bonds in strong, 69, 71 weak, 74, 76–77 chemical modifications of, 205, 205f chirality of, 79–80, 80f chloroplast, 303–304 circular. *See* Circular DNA cleavage of, 217–219, 217t, 220f coding strand, 519, 520f compaction/packaging of, 298 in bacteria, 301–302, 302f, 302t, 341–342, 343f into chromatin, 338–342, 338f–342f chromatin condensation in, 316–320, 317f, 320f in chromosomes, 341, 341f coiled forms in, 341, 342f in eukaryotes, 301–302, 332–360. *See also* Nucleosome(s) levels of, 342f histones in, 332–339 loop forms in, 341, 341f, 345, 346f. *See also* DNA looping in nucleoids, 341–342 nucleosomes and, 332–338, 343f. *See also* Nucleosome(s) rosette forms in, 341, 342f SMC proteins in, 316–320 supercoiling in, 305–306, 306f, 307f, 309–311. *See also* DNA, supercoiling of in 30 nm filament, 339–341, 340f topoisomerases in, 312–316, 312f, 312t, 314f–317f in viruses, 300–301, 301f, 301t complementary, 225, 501 cruciform, 191, 192f, 310, 310f damage to by alkylation, 418–419, 420f by cytotoxic agents, 421, 422f by deamination, 203–204, 205f, 416–417, 417f by nitrous oxide, 418, 418f by depurination, 417, 418f by hydrolysis, 416–417, 417f, 418f by methylation, 417, 430, 431f oxidative, 418, 430, 431f by radiation, 421–423, 440 deformations in, 306 denaturation of, 200, 200f early studies of, 44–48, 176, 209 elasticity of, discovery of, 297 functions of, 4 hairpin, 109, 110f, 191, 192f helix of. *See* Double helix as heredity chemical, 44–46 histone binding of, 141, 336–338, 337f hydrolysis of, 179–180 junk, 482, 505 length of, 300–304 linker, 219 in nucleosomes, 334, 334f linking number of, 308–310, 308f melting of, 200–201, 200f methylation of, 205, 205f damage from, 417, 430, 431f in mismatch repair, 426, 426f in replication initiation, 393, 393f mitochondrial, 288, 290, 303–304 genetic code variations and, 602–604, 603t mutations in, 289–290 as motor protein ligand, 157–160 mutations in. *See* Mutations nongene, 270 non-template strand, 519, 520f nucleosome binding by, 335–336, 335f, 336f nucleotides in, 45, 62–64, 63f palindromic, 190–191, 192f pentose rings in, 177, 177f, 178, 178f, 186 plasmid, 302, 302t polarity of, 179 postsynthetic changes in, 65, 67f quadruplex, 192, 193f recognition sequences in, 217–218, 218t recombinant, 217. *See also* Biotechnology relaxed, 305, 307f renaturation of, 200–201, 200f replication fork in. *See* Replication fork resonance in, 71, 72f restriction sites in, 217–218 selfish, 482, 505 single-strand breaks in, 422–423, 423f, 425–435, 451 repair of, 422–423, 423f replication fork collapse and, 447, 448f single-stranded, 377t, 382f, 384. *See also* Single-stranded DNA-binding protein (SSB) stability of, 200

- sticky ends of, 218, 219, 219f  
 strained, 307, 307f  
 strand separation in, 200–201, 201f  
   underwinding and, 307–308  
 structure of, 4, 4f, 65, 66f. *See also* DNA,  
   compaction/packaging of  
 secondary, 185–194  
   variations in, 187–188, 188f, 192f, 193f  
 Watson-Crick model of, 185–187, 209.  
   *See also* Double helix  
 tertiary, 297–326  
 supercoiling of, 305–306, 306f, 307f  
   around histone octamer, 334–336,  
   335f, 336f  
   discovery of, 323  
 DNA gyrase in, 325–326  
   negative, 309, 309f, 313, 316f  
   in nucleosome binding, 336  
   plectonemic, 310–311, 311f  
   positive, 309, 309f  
   solenoidal, 311, 311f  
 synthesis of. *See* DNA replication  
 template strand, 48, 364, 369, 370f, 516,  
   519, 520f  
   breaks in, replication fork stall/collapse  
     and, 447, 448f  
 tetraplex, 192, 193f  
 thermal properties of, 200–201,  
   200f, 201f  
 three-dimensional structures from,  
   194b–195b  
 topology of, 297–326  
 in transcription. *See* Transcription  
 translocation and, 158, 158f, 160f  
 transposon, 482, 505  
 triplex, 191–192, 193f  
 twists in, 297, 306  
 underwinding of, 307–310, 307f, 336  
   topoisomerases in, 312–316, 312f, 312t,  
   314f–317f  
 unwinding of, 382f, 383  
   denaturation and, 200–201, 200f, 201f  
   helicases in, 157–160, 159f, 381, 382f,  
   392–393, 392f  
   nonhelicase, 158–159, 382f, 383  
   topoisomerases in, 382f, 383  
     in transcription, 520, 535, 536, 536f  
 UV light absorption by, 201, 201f  
 viral, 300–301, 301t  
 writhes in, 297, 306  
 X-ray crystallography of, 121–125,  
   122f–124f  
 Z-form, 188, 189, 189f  
**Dna2**, 467, 467f  
 DNA adenine methyltransferase (Dam  
   methylase), 377t, 393  
 DNA amplification, 226–228, 227f, 228f  
 DNA bubbles  
   in denatured DNA, 201, 201f  
   in replication, 392–393, 392t  
   in transcription, 520, 520f  
 DNA chips. *See* DNA microarrays  
 DNA cloning, 216–226  
   bacteria in, 236, 236f  
   baculoviruses in, 237, 237f  
   definition of, 217  
   discovery of, 251  
   DNA libraries for, 224–225, 225f  
   DNA ligase in, 217, 217t, 219  
   mammalian cell cultures in, 238  
   oligonucleotide-directed mutagenesis  
     and, 239–240, 239f  
   pulsed field gel electrophoresis in,  
   223–224  
   restriction endonucleases in, 217  
   site-directed mutagenesis and,  
   239–240, 239f  
   steps in, 217, 218f  
   transformation in, 221–222  
   in transgenic animals, 238, 238f  
   vectors in, 217–224  
   bacterial artificial chromosome, 222–223  
   expression, 233–236, 233f  
   mammalian virus, 238  
   pBR322, 220–221, 222f  
   plasmid, 220–222, 222f  
   preparation of, 217–219, 220f  
   shuttle, 223  
   yeast artificial chromosome,  
   223–224, 224f  
   yeast in, 236–237  
 DNA computing, 191b  
 DNA duplexes, 201  
   hybrid, 202  
 DNA fingerprinting, 230b–231b  
 DNA footprinting, 526, 527f, 541, 704b  
 DNA genotyping, 230b–231b  
 DNA glycosylase, in base excision repair,  
   430–433, 432f, 433t  
 DNA gyrase, 312t, 313, 315f, 325–326, 377t  
   discovery of, 324–326  
   in replication termination, 397–398  
   in supercoiling, 324–326  
 DNA helicases, 157–160, 157f–160f  
   in bacteria, 381, 382f, 383f  
     Pol III and, 384, 384f  
       in replication initiation, 392–393, 392f  
   in eukaryotes, 387  
 DNA hybridization, 201–203, 202f–204f  
   oligonucleotide synthesis for, 206, 207f  
 DNA libraries, 224–225, 225f  
   cDNA, 225, 225f, 242–243, 244f  
   genomic, 224–225  
 DNA ligase, 151–156, 382f, 383  
   catalytic action of, 152, 153–156, 155f  
   cofactors for, 152  
   discovery of, 14–15, 16, 363  
   in DNA cloning, 217, 217t, 219  
   in mismatch repair, 425t, 426  
   in nucleotide excision repair, 433t,  
   434, 434f  
   in recombinational repair, 449,  
   454t, 460  
   structure of, 156  
 DNA looping, 341, 341f, 345, 346f, 670–673,  
   671f–673f  
   *ara* operon and, 707  
   HMG proteins and, 741, 741f  
   *lac* operon and, 702, 702f  
   between promoters and enhancers,  
   741, 751  
   in replication, 384–386, 385f  
 DNA microarrays, 248–250, 248f–250f  
   in nucleosome localization,  
   345–346  
   in transcriptome analysis, 278  
 DNA nucleases, 371–372  
 DNA octahedrons, 194b–195b  
 DNA photolyase, 429–430, 430f, 440  
 DNA polylinkers, 219, 220f  
 DNA polymerase(s)  
   antimutator, 375  
   β subunit of, 107, 108f  
   discovery of, 14  
   in DNA repair, 373, 373t  
   functions of, 373t  
   primed templates for, 403  
   processivity of, 158, 375–376, 404  
     β clamp and, 379, 380f, 404  
   properties of, 373t  
   in replication, 368, 369–376. *See also* DNA  
     polymerase I (Pol I)  
   structure of, 373–375, 373f  
   TLS, 436–438, 437f, 437t  
   in translesion synthesis, 436–438,  
   437f, 437t  
 DNA polymerase α (Pol α), 387, 390t  
 DNA polymerase δ (Pol δ), 388, 390t, 433t  
 DNA polymerase ε (Pol ε), 433t  
 DNA polymerase I (Pol I), 403  
   accuracy of, 370–372, 370f, 373–375  
   closed form, 374, 374f  
   in DNA repair, 369, 454t  
   dNTPs and, 370–371, 370f, 373–375, 374f  
   fingers domain of, 373–374, 373f  
   5'→3' exonuclease of, 371, 371f–373f,  
   372, 373t  
   in nick translation, 372, 372f,  
   383, 386  
   in nucleotide excision repair, 433t,  
   434, 434f  
   open form, 374, 374f  
   palm domain of, 373, 373f  
   processivity number of, 376  
   in recombinant DNA technology, 217t  
   in replication, 369–376, 370f  
   structure of, 373–375, 373f  
   3'→5' exonuclease of, 371–372, 371f–373f,  
   373, 373t, 375  
   thumb domain of, 373, 373f  
 DNA polymerase II (Pol II), 372–373, 373t  
   promoter binding of, 739  
   transcription factors and, 740, 741  
   transcription machinery of,  
   739–742, 740f  
   in translesion synthesis, 437, 437t

- DNA polymerase III (Pol III), 373, 373t  
 β clamps and, 385–386, 386f–388f  
 DnaB and, 384, 384f  
 Pol III core and, 377–378, 378t  
 Pol III holoenzyme and, 378, 378f, 384, 384f  
 processivity of, 158, 375–376, 404  
 β clamp and, 379, 380f, 404  
 subunits of, 377–378, 378f, 378t
- DNA polymerase IV (Pol IV), 373, 373t  
 in translesion synthesis, 437, 437t
- DNA polymerase V (Pol V), 373, 373t  
 discovery of, 409, 441  
 in translesion synthesis, 437, 437t, 441
- DNA probes, in hybridization, 202–203, 203f
- DNA profiling, 230b–231b
- DNA repair, 409, 424–440  
 base excision, 430–433, 431f, 432f, 433t  
 direct, 429–430  
 DNA polymerases in, 373, 373t  
 of double-strand breaks, 423, 435, 445–477  
 enzymatic, 163  
 glycosylation in, 432–433, 432f, 433t  
 long patch, 432  
 mismatch, 412, 425–429, 425t, 426f, 427f.  
*See also* Mismatch repair  
 mutations affecting, 412  
 nonhomologous end joining in, 435  
 nucleotide excision, 433–435, 433t, 434f, 435f  
 overview of, 409  
 oxidation in, 433, 433t  
 photoreactivation, 429–430, 430f, 440  
 recombinational, 423, 435, 446–464. *See also* Recombinational DNA repair  
 at replication forks, 435–438, 437f, 437t  
 short patch, 431f, 432  
 of single-strand breaks, 422–423, 423f, 425–435, 451  
 of single-strand gaps, 452–453, 453f. *See also* Recombinational DNA repair  
 solar radiation in, 429–430, 430f, 440  
 SOS response in, 707, 710–711, 711f, 725  
 transcription-coupled, 434, 435f, 539–540  
 translesion synthesis in, 436–438, 437f, 437t, 441, 447, 452  
 uracil removal in, 204
- DNA replication, 47–48, 363–405. *See also under* Replication  
 accuracy of, 371–372, 371f, 375, 403, 425, 629  
 antiparallel strands in, 368  
 ATP in, 379–380, 380f, 381f, 383, 391–394  
 base pairing in, 370–375, 370f, 371f, 374f, 375f. *See also* Base pairs/  
 base pairing  
 conservative, 364, 365f  
 copy-number amplification in, 485  
 daughter strands in, 364  
 synthesis of, 368–369
- definition of, 364  
 direction of, 367f, 368, 368f, 381, 382f, 383f  
 dispersive, 364, 365f  
 DNA accessibility in, 344, 348–351  
 DNA polymerase III in, 373  
 DNA polymerases in, 368, 369–376.  
*See also* DNA polymerase(s)  
 DNA repositioning in, 375–376  
 dNTPs in, 370–371, 370f, 373–375, 374f, 375f  
 double-strand breaks in, 446  
 repair of. *See* Recombinational DNA repair  
 error rate in, 520  
 free energy in, 370–371  
 helicases in, 381, 382f, 383f, 384  
 histone modification preservation during, 354–357, 356f, 357f  
 initiation of, 368, 390–395  
 in bacteria, 391–394, 391f–393f, 405  
 in eukaryotes, 391–395, 394f, 405  
 proteins in, 390–391, 391f  
 insertion site in, 370, 370f  
 lagging strand in, 369, 384–385, 385f  
 leading strand in, 369  
 metal ions in, 375, 376f  
 nick translation in, 372, 372f, 383, 386  
 Okazaki fragments in, 383, 384–385, 385f, 386–387  
 origin of replication in, 221, 298, 367, 367f, 368  
 overview of, 364  
 parental strands in, 364, 368  
 Pol III in, 373  
 postinsertion site in, 370, 370f  
 primed template in, 369, 370f  
 primer strand in, 369, 370f  
 primer terminus in, 369  
 primers in, 368, 369, 382f, 392, 500  
 processive synthesis in, 376  
 proofreading in, 371–372, 371f, 375, 403, 425  
 proteins in  
 in bacteria, 377t, 380–384  
 in eukaryotes, 387–389, 390t  
 rate of, 379–380, 381f, 389  
 reaction mechanism in, 370–375, 370f, 371f, 374f, 375f  
 replication fork in. *See* Replication fork  
 replisome in, 384, 384f  
 restarting of, 450, 460  
 reverse transcriptase in, 500–501, 501f  
 RNA polymerase in, 394, 396–397  
 RNA-dependent, 500–501, 501f  
 rolling-circle, 485, 486f  
 semiconservative, 364–367, 365f  
 semidiscontinuous, 368–369  
 site-specific recombination and, 485  
 strand breaks in, repair of, 422–423, 423f.  
*See also* DNA repair
- template strand in, 187, 187f, 364, 369, 370f  
 termination of, 395–402  
 in bacteria, 395–398  
 in eukaryotes, 398–402  
 θ-form, 367  
 translocation in, 158–159, 381, 392–393  
 trombone model of, 384, 385f  
 vs. transcription, 516, 518
- DNA sequencing, 171, 228–233  
 454 Sequencing in, 234b–235b  
 automated, 232–233, 232f  
 high-throughput, in transcriptome analysis, 278  
 with Illumina sequencer, 235b  
 next-generation methods in, 234b–235b  
 Sanger method for, 228–233, 229f
- DNA topology, 306
- DNA translocases, 158, 159, 160f
- DnaA, in replication initiation, 377t, 391–394, 393f
- DnaB helicase, 377t, 381, 383f, 392  
 Pol III and, 384, 384f  
 reloading of, 450
- DNA-binding domains, 107, 136  
 of transcription factors, 679–684, 683f  
 vs. regulatory domains, 683–684, 684f
- DNA-binding motifs, 679–682. *See also* Motifs
- DNA-binding proteins, 135, 138–144  
 binding sites for, 139, 141, 142f  
 cooperativity and, 139  
 deletion analysis of, 744b–745b  
 functions of, 138  
 in gene expression, 141–144  
 histone, 141, 335, 335f, 338. *See also* Histone(s)  
 hydrogen bonds and, 141–142  
 Lac repressor as, 142–144, 143f  
 nonspecific, 138–141, 144, 144f  
 reporter gene assays for, 744b–745b  
 single-stranded, 135, 139–141, 140f, 141f, 377t, 382f, 384  
 specific, 138, 141–144  
 telomere, 399–401, 400f
- DNA-binding transactivators, 739
- DNA-binding transcription activators, 739
- DnaC, 377t, 392
- DnaG primase, 383
- DNase, 46
- dNDPs (deoxyribonucleoside diphosphates), 371
- dNMPs (deoxyribonucleoside monophosphates), 370–371, 370f
- dNTPs (deoxyribonucleoside 5'-triphosphates), 370–371, 370f, 373–375, 374f
- Dobzhansky, Theodosius, 4, 10
- Dominance, 25, 26f  
 codominance and, 30, 30f  
 incomplete, 29–30, 30f
- Donor site, in transposition, 489

- Dorsal-ventral axis, in development, 793–795, 794f
- Dosage compensation, 753, 755–756, 755f, 756f
- Dosage compensation complex, 756
- Double bonds, 70–71, 71f
- Double helix, 4f, 185–189
- antiparallel, 186
  - bonds in, 69, 71, 74, 75, 75f, 181, 184f, 187
  - complementary strands in, 48
  - crystallography of, 210
  - discovery of, 24, 47–48, 176, 176f, 185–186, 209
  - left-handed, 186
  - major groove of, 186
  - minor groove of, 186
  - right-handed, 185–187, 186f
  - template strand in, 187, 187f
  - underwinding of, 307–310, 307f, 382f
  - untwisting of, 382f, 383
  - variations in, 187–188, 188f
  - Watson-Crick model of, 185–187, 209
  - weak interactions in, 74, 75, 75f
- double sex (dsx) gene*, 771
- Double-helical structure, of RNA, 196–197, 197f
- Double-strand break(s), 423, 423f, 446
- causes of, 447
  - crossovers and, 446, 447f, 465–469, 468f
  - formation of, 447, 448f, 465–468, 465f, 467f
  - hot spots for, 465
  - in immunoglobulin gene rearrangement, 507, 507f
  - in mating-type switch, 471, 471f
  - in meiosis
    - in bacteria, 445, 446
    - in eukaryotes, 465–469, 465f, 467f, 468f  - in mitosis, 469–470, 470f
  - processing of, 445
  - repair of, recombinational, 423, 435, 446–464
- Double-strand break pathway, in
- recombinational repair
  - in bacteria, 448–450, 449f, 450f
  - in eukaryotes, 468, 469–470, 470f
- Double-strand break repair, 423, 435
- by nonhomologous end joining, 472–474, 473t, 474f
  - recombinational. *See also*
    - Recombinational DNA repair
    - in bacteria, 448–450, 449f, 450f
    - in eukaryotes, 469–470, 470f
- Down syndrome, 466b
- Downstream promoter element, 534, 535f
- Doxorubicin (Adriamycin), 319b, 421, 422f
- Dpn repressor, 769–771, 770f
- Dröge, Peter, 693
- Drosha, 575, 575f
- Drosophila melanogaster*
- chromosomes of, 302t
  - development in, 791–792, 791f, A-17
  - alternative splicing in, 769–771, 770f
  - anterior-posterior axis in, 791, 793–795, 794f
  - combinatorial control in, 748–750, 749f, 750f
  - dorsal-ventral axis in, 793–795, 794f
  - homeotic genes in, 793, 796–798, 797f
  - Hox genes in, 796–798, 797f, 802
  - maternal genes in, 793–795, 794f
  - metamersion in, 791
  - pattern-regulating genes in, 792–798
  - polarity in, 791, 793–795, 794f
  - segmentation genes in, 793, 795–796, 796f, 797f
  - segmentation in, 791
  - sexual, 769–771, 770f
  - signaling in, 795–796, 796f
  - of wings, 806
- DNA of, 302t
- dosage compensation in, 756
- genome of, 302t
- gypsy* elements in, 498
- life cycle of, 791–792, 791f, A-16–17, A-17f
- as model organism, 38, A-3–A-4, A-16–A-17
- Morgan's studies of, 38–40, 38f
- pattern-regulating genes in, 792–798
- homeotic, 793, 796–798, 797f
  - maternal, 793–795, 794f
  - segmentation, 793, 795–796, 796f, 797f
- RNA transport in, 570–571, 570f
- sex determination in, alternative splicing in, 769–771, 770f
- sex-linked genes in, 38–40, 38f, 39f
- transcription factors in, 762
- Drugs
- development of, 166b–167b
  - resistance to, 302, 303b, 504
- Dscam* gene, alternative splicing of, 556
- dsx* gene, 771
- Dulbecco, Renato, 323
- Dunaway, Marietta, 693
- Duplications, chromosome, 32–34, 34f, 35f, 272, 273f, 414–415, 415f
- Dynan, William, 541
- E site, 621–623, 623f
- Ecols, Harrison (Hatch), 441
- Editing, RNA, 565–568, 567f, 568f
- Editosomes, 566
- eEF2, 641
- eEF1 $\alpha$ , 641
- eEF1 $\beta\gamma$ , 641
- Effectors, 674–675, 674f, 675f
- EF-G, 638, 639f, 640f
- in translocation, 640, 641, 642f
- EF-Ts, 638, 639f, 640f
- EF-Tu, 638, 639f, 640f
- eIF1, 634, 635
- eIF2, 774
- eIF3, 634, 635
- eIF5, 634, 635
- eIF1A, 634, 635
- eIF4A, 634, 635
- eIF4B, 634, 635
- eIF5B, 634, 635
- eIF4E, 631
- eIF4E-eIF4G interactions, disruption of, 773, 774–775, 775f
- eIF4F, 634, 635
- eIF4G, 636, 773, 774–775, 775f
- 18S rRNA
- processing of, 574
  - transcription of, 534, 534f, 534t
- 80S ribosome, 635, 636f
- Electric dipole moment, 73, 73f
- Electron density maps, 123, 123f
- Electron pushing, 154
- Electronegativity, 69, 69f
- Electrophoresis. *See* Gel electrophoresis
- Electrophoretic mobility shift assay (EMSA), 704b–705b
- Electroporation, 221
- Electropositivity, 69
- Ellipticine, 319b
- Elongation complex, 521, 521f
- Elongation factors, 684
- in transcription, 537–538
  - in translation
  - in bacteria, 638–641
  - in eukaryotes, 642
- Embryonic stem cells, 799, 800, 800t
- transcription factors and, 533b
- Enantiomers, 79–80, 80f, 81b
- End replication problem, 398–400, 399f
- Endo, Y., 647b
- Endonucleases, 371
- homing, 472, 472f
- Endoplasmic reticulum, posttranslational modification in, 655–657, 656f
- Endoribonucleases, in mRNA degradation, 571
- Energy
- activation, 84–85, 85f
  - catalysis and, 147–149
  - definition of, 148
- ATP-driven transfer of, 88, 89b
- binding, 146, 148f, 149
- for bond formation, 72
- in coupled reactions, 89, 90b
- free, 87. *See also* Free-energy change ( $\Delta G$ )
- in replication, 370–371
- stored in ATP, 182
- thermodynamic laws and, 86–87

- Energy coupling, 89, 90b  
*engrailed (en)* gene, 795  
 Enhanceosomes, 752, 752f  
 Enhancers, 532, 671, 693, 737  
   definition of, 737  
   distance from promoters, 741, 751  
   for *eve* gene, 749–750, 750f  
 Ensembl website, 264, 266  
 Entropy, 86–87, 87f  
*env* gene, 500, 502f  
 Enzyme(s), 136, 144–156  
   AAA+, 157–158, 379, 391–392  
   active sites on, 146–147, 146f  
   allosteric, 161–164  
   apoenzymes and, 145  
   catalytic action of, 2, 3, 87–88, 146–149.  
     *See also* Catalysis  
   coenzymes and, 145, 145t  
   covalent modification of, 164–167, 165t, 684, 686. *See also* Protein modification  
   definition of, 3, 46  
   in DNA repair, 410. *See also* DNA repair  
   in DNA underwinding, 308  
   evolution of, 3–4, 6–7, 8, 15, 88, 548, 554, 578. *See also* Ribozymes (catalytic RNA)  
   genes and, 46–47  
   group transfers in, 146  
   heterotropic, 161–164  
   histone modifying, 343–344, 344f  
   holoenzymes, 145  
     DNA, 378, 378f, 384, 384f  
     RNA, 519, 519f  
   homotropic, 161–162  
   hydrogen tunneling and, 61, 91  
   inhibition of, 151, 152b–153b  
     autoinhibition and, 163, 163f  
     pharmacologic, 166b–167b  
   nucleosome modifying, 343–344, 348–357  
   overview of, 144–145  
   prosthetic groups and, 145  
     posttranslational attachment of, 654  
   in recombinant DNA technology, 217t  
   regulatory, 161–165, 161f–164f  
   RNA. *See* Ribozymes (catalytic RNA)  
   substrates for, 88, 146, 149  
 Enzyme cofactors, 145, 145t  
 Enzyme kinetics, 149–151  
   general rate constant and, 151  
   initial velocity in, 149, 150f  
   maximum velocity in, 149  
   Michaelis-Menten equation and, 150–151, 150f  
   pre-steady state in, 149  
   steady-state, 149–150  
   turnover number and, 151  
 Epigenetic control, in cancer, 355b  
 Epigenetic inheritance, 343–344, 410  
   in cancer, 355b  
   of chromatin, 352–357  
   definition of, 343  
   of gene silencing, 786, 786f  
   of histone modifications, 353–357  
 Epigenetic marks, 354, 356, 753  
 Epitope tags, 242, 244–246, 245f  
 Equilibrium  
   definition of, 87  
   entropy and, 86–87  
   reaction rate and, 147–148  
 Equilibrium dialysis, 171, 171f  
 Equilibrium expression, 137, 137f  
 Erythrocytes, sickled, 53b  
*Escherichia coli*. *See also* Bacteria  
   DNA polymerase β subunit of, 107, 108f  
   chromosomes of, 301–302, 302t  
   DNA of, 301–302, 302f, 302t  
     cloning of, 217–223, 236, 236f  
   genome of, 301–302, 302t  
     sequencing of, 261, 261f  
   *lac* operon of. *See lac* operon  
   Lac repressor of. *See Lac* repressor  
   λ phage of. *See Lambda (λ) phage*  
   life cycle of, A–6  
   mismatch repair in, 425–427, 425t, 427f  
   nucleotide excision repair in, 433–434, 434f  
   pBR322 plasmid of, 221, 221f, 222f  
   recombinational repair in, 453–464. *See also* Recombinational DNA repair  
   replication fork of, 377–387, 377t  
   ribosome of, 616–619, 616f, 617f  
   RNA polymerase holoenzyme of, 519, 519f  
   topoisomerases of, 312t, 313  
   transcription in, 523–532  
 Etoposide (Adriamycin), 319b  
 Euchromatin, 735, 735f. *See also* Chromatin  
 Eukaryotes  
   chromosomes of, 303–304, 304f  
   DNA of  
     packaging of, 302–304  
     size of, 302t  
   DNA replication in  
     initiation of, 391–394, 405  
     termination of, 398–402  
   evolution of, 9, 9f. *See also* Evolution  
   gene regulation in, 733–806  
     posttranscriptional, 767–806  
     transcriptional, 733–766  
   genome of, 302t  
     sequencing of, 269  
   as model organisms. *See Model* organisms  
   transcription in, 532–540  
   translation in  
     elongation in, 638–642  
     initiation of, 631–632, 632t, 634–637, 636f, 637f, 773–774, 774f  
     termination of, 643  
 Evans, Martin, 490b  
 Eve, mitochondrial, 288, 290  
*even-skipped (eve)* gene, 749–750, 749f, 750f  
 Even-skipped protein, 749–750, 749f, 750f  
 Evo-devo, 5, 801  
 Evolution, 2–12, 282–291  
   accelerated, 273–274, 275f  
   allopatric speciation and, 288  
   amino acid sequences and, 102–103  
   of bacteria, 7b, 9, 9f  
   branching, 10, 10f  
   of chloroplasts, 303–304, 621, 632  
   comparative genomics and, 264–266  
   competition and, 10  
   Darwinian, 9–12, 15, 21, 24, 801  
   early events in, 5–9, 6f  
   of enzymes, 3–4, 6–7, 8, 15, 88, 548, 554, 578. *See also* Ribozymes (catalytic RNA)  
   evo-devo and, 5, 801  
   of Galápagos finches, 801, 802, 802f  
   of genetic code, 602–606, 605b  
   genetic drift and, 288  
   genomics and, 271–274, 273f, 282–291  
   gradual nature of, 11  
   horizontal gene transfer in, 11, 11f, 286  
   human, 288, 289f  
     genomic alterations in, 271–274, 273f  
     migration in, 288–289  
     mitochondrial Eve in, 288, 290  
     out of Africa theory of, 288  
     rate of, 273–274, 275f  
     sources of variation in, 271–274, 288–291  
     species divergence in, 288, 289f  
     Y chromosome Adam in, 288, 290  
   of immune system, 507  
   of influenza virus, 129  
   of introns, 503–505, 564b, 779  
   Lamarckian, 21  
   last universal common ancestor in, 8–9, 8f, 9f, 283, 586, 602  
   migration in, 288–289  
   of mitochondria, 288, 290, 303–304, 602–604, 621, 632  
   mutations in, 5, 10–11, 50, 54–55, 289–291, 411–412, 801  
     rate of, 286, 287  
   by natural selection, 9–12, 21, 287–288, 801  
   out of Africa theory and, 288  
   out-groups in, 273–274  
   in protein world, 8–9  
   of proteins, 114  
     rate of, 5  
   of retrotransposons, 503–505  
   of retroviruses, 7b, 503–506  
   of ribozymes, 577b  
   of RNA processing, 552b–553b  
   RNA world and, 6–8, 15, 20, 88, 482, 548, 554, 578, 601–602. *See also* Ribozymes (catalytic RNA)  
   sources of variation in, 5, 10–11, 11f, 271, 288–291  
   theory of, 10–12  
   of 3' end processing, 552b–553b

- of translation, 601–606, 604b–605b. *See also* Ribozymes (catalytic RNA)
- of transposons, 503–506
- of tRNA, 601–602, 604b–605b
- of viruses, 7b
- Evolutionary biology, developmental biology and, 5, 801
- Evolutionary trees, 272, 273f, 283–286, 285f–287f
- Excinucleases, 433
- in nucleotide excision repair, 433t, 434, 434f
- Exo1, 445, 467, 467f
- Exon(s), 269, 270f, 549, 555
- definition of, 549
  - joining of, 554–555, 556f, 563–565, 565f.
- See also* Splicing
- Exon-junction complex, 570, 570f
- in nonsense-mediated decay, 571
  - in pre-mRNA splicing, 652, 653f
- Exonuclease(s), 371–372, 371f
- in mismatch repair, 425t, 426
  - in transcription termination, 539
- Exonuclease III, 217t
- Exoribonucleases, in mRNA degradation, 571
- Exosomes, 571, 571f
- Exothermic reactions, 72
- ATP in, 88, 89b
- Exportins, 569, 569f
- Expressed sequence tags (ESTs), 261
- Expression vectors, 233–236, 233f
- Extrachromosomally primed retrotransposons, 495, 496f
- F<sub>1</sub> generation, 25, 26f, 26t, 27t
- F<sub>2</sub> generation, 25, 26f, 26t, 27t
- F plasmid, 476
- FADH<sub>2</sub> in photoreactivation, 429, 430f
- fem-3 binding factor (FBF), 776, 776f, 804
- Ferris, James, 20
- Fibroblasts, reprogramming of, 533b
- 50S subunit, 8f, 616–617, 617t
- Fingerprinting, DNA, 230b–231b
- Fire, Andrew, 783, 783f, 784, 785, 805
- First law of thermodynamics, 86
- Fission yeast, gene silencing in, 754b, 784
- Fitch, Walter, 266
- 5' cap, 548, 549–550, 550f
- addition of, 548, 549–550, 550f, 552–553, 631
  - removal of, 777, 779f
  - in ribosome binding, 549–550, 551f
  - in translation, 549–550, 551f, 631, 777
- 5'→3' exonuclease, 371, 371f–373f, 372, 373t
- 5'→3' helicase, 381
- 5'-AUG initiation codon, 631
- 5'-end processing
- capping in. *See* 5' cap
  - in double-strand break repair, 449–450, 451f, 454, 455f
  - gene regulation in, 684–685
- 5S rRNA, 573, 574, 575f
- processing of, 574
  - transcription of, 534t, 535, 535f
- 5.8S rRNA
- processing of, 574
  - transcription of, 534, 534f, 534t
- 5'UTR
- in iron homeostasis, 779, 781f
  - in translation, 774, 775f
  - in gene networks, 774, 775f
- Flagellin proteins, in phase switching, 487–488, 488f
- Flavin adenine dinucleotide (FAD), 182
- Flemming, Walther, 33
- Flp recombination system, 485, 488–489, 490b–491b
- fMet-tRNA, conversion to dipeptidyl-tRNA, 638–640, 641f
- fMet-tRNA synthetase, 631, 632, 633–634
- fMet-tRNA<sup>fMet</sup>, 631, 632, 633–634
- Folding, protein. *See* Protein folding
- Footprinting
- chemical protection, 704b
  - DNA, 526, 527f, 704b
- Forensics
- DNA fingerprinting in, 230b–231b
  - phylogenetics in, 284b
- Fork, replication. *See* Replication fork
- Fork regression, 451, 452f, 457
- N-Formylmethionine, posttranslational modification of, 654
- N-Formylmethionyl-tRNA<sup>fMet</sup>, 631, 632, 633–634
- 43S preinitiation complex, 634, 635, 636f
- 45S pre-rRNA, 574
- 40S subunit, 616–617, 617t
- 454 Sequencing, 234b–235b
- Four-helix bundle, 111–112, 112f
- Fourier series, 123
- Fragile X syndrome, 52–53, 414, 414f, 414t
- Fragile X-E syndrome, 414t
- Frameshift mutations, 412–413, 413f, 594–595
- Franklin, Rosalind, 185–186, 209
- Free energy (G), 87
- in replication, 370–371
- Free-energy change ( $\Delta G$ ), 87
- biochemical standard, 147
  - in linked reactions, 89, 90f
  - standard Gibbs, 87
- Free-energy funnel model, 117–118, 117f
- Friedreich's ataxia, 414t
- frizzled ( fz ) gene*, 796
- Fruit fly. *See Drosophila melanogaster*
- FtsH, 723
- Ftz (fushi-tarazu), 795, 795f
- Functional RNAs, 49–50. *See also* rRNA (ribosomal RNA); tRNA (transfer RNA)
- Fungi. *See Yeast*
- Fushi-tarazu (Ftz), 795, 795f
- Fusion genes, 415–416
- Fusion proteins, 240–241, 282
- in gene localization, 242–243, 243f
  - in immunofluorescence, 242, 243f
- Fusions, chromosomal, 272
- G<sub>1</sub> phase, 33, 34f
- G<sub>2</sub> phase, 33, 34f
- G proteins
- G-protein-coupled receptors and, 759–760
  - in signaling, 759–760, 759f
- G tetraplex, 192, 192f
- gag gene, 500, 502f
- Gag protein, 498
- Gain of function mutations, 411
- Gal genes
- deletion analysis of, 744b–745b
  - reporter gene assays for, 744b–745b
  - in yeast combinatorial control, 743–746, 743f, 746f
- gal operon, 708, 708f
- Gal4 protein, 682, 683–684, 683f
- Gal repressor, 708, 708f
- Galactose metabolism, combinatorial control in, 743–746, 743f, 746f
- β-Galactosidase, 170, 699, 699f
- as protein tag, 240t
- Galactoside permease, 699, 699f
- β-Galactosides, as lac operon inducers, 703
- Galápagos finches, 801, 802, 802f
- Galileo Galilei, 13, 14, 14f
- Gal4p, 744b–745b
- Gamete cells, 24
- formation of, 34–37, 35f
- Gamma complex (clamp loader), 378, 378f, 378t, 379–380, 380f, 381f
- Gamma rays, 422–423, 462b
- Gamma turns, 106, 109, 110f
- γδ resolvase, discovery of, 481
- Gap genes, 793, 795
- Gap repair, 452–453, 453f, 456
- Garrod, Archibald, 46
- Gates, Julie, 806
- Gcn4 mRNA, translation of, 777
- Gcn5 subunit, 352, 733, 736, 738b–739b
- Gcn5 transcription factor, 733, 736, 738b–739b
- GDP (guanosine diphosphate)
- in RNA transport, 569, 569f
  - in translation termination, 643, 643f
- Geckos, van der Waals interactions and, 74–75, 75f
- Gel electrophoresis (SDS-PAGE), 203
- in Northern blotting, 203
  - pulsed field, 223–224
  - for replication origins, 396b–397b
- sodium dodecyl sulfate–polyacrylamide, 101, 101f
- in Southern blotting, 203
- in topoisomerase visualization, 312–313, 312f
- two-dimensional, 280–281, 280f
- in Western blotting, 244, 255f

- Gel-exclusion chromatography, 100, 100f  
 Gellert, Martin, 325–326  
 Gene(s), 26. *See also specific genes*  
 chromosomal location of, 38–44  
 cis-acting, 701  
 cloned  
     alterations in, 239–240  
     expression of, 233–238  
     production of, 221–226. *See also DNA cloning*  
 disease-causing, identification of, 274–277, 276f  
 early studies of, 24  
 enzyme coding by, 46–47  
 fusion, 415–416  
 gap, 793, 795  
 homeotic, 793, 796–798, 797f  
 homologous, 265  
 horizontal transfer of, 11, 11f, 286, 720  
 housekeeping, 669  
 Hox, 796–798, 797f, 802  
 independent assortment of, 28–29, 29f, 29t  
 linked, 30–31, 31f  
     inheritance of, 38–40, 38f, 40f  
     segregation of, 40, 40f, 56  
     unlinking of, 40–41, 41f, 42f  
 maternal, 793–795, 794f  
 Mendel’s concept of, 26  
 number in genome, 269  
 oncogenes, 412  
 orthologous, 265, 266  
 pair-rule, 793, 795  
 paralogous, 265–266  
 pattern-regulating, 792–798  
 protein-coding, accelerated evolution of, 274  
 recombination, discovery of, 476  
 reporter, 242–243, 244f, 744b–745b  
 RNA-coding, 271  
     accelerated evolution of, 274  
 segmentation, 793  
 segment-polarity, 793, 795–796, 796f, 797f  
 sex-linked, 38–40, 38f, 39f  
 trans-acting, 701  
 tumor suppressor, 412  
 universal, 8–9, 283  
 Gene activation, 669  
     site-specific recombination in, 489, 489f  
 Gene amplification, 226–228, 227f, 228f  
 Gene conversion, 468  
     mating-type switch and, 470–471, 471f  
     recombination and, 470–471, 471f  
 Gene dosage compensation, 753, 755–756, 755f, 756f  
 Gene expression. *See also Gene regulation*  
     of cloned genes, 233–238  
         in bacteria, 236, 236f  
         in baculoviruses, 236, 236f, 237–238, 237f  
         in mammalian cell cultures, 238  
         in transgenic animals, 238, 238f  
         in yeast, 236, 236f  
     constitutive, 669  
     evolution and, 274  
     histone modification and, 352–357  
     regulated, 522, 669  
 Gene function studies, transcriptome analysis in, 278, 279f  
 Gene insertion, site-specific recombination in, 489, 489f  
 Gene knockdown, 787, 790f  
 Gene mapping, recombination, 42–44, 42f, 43f  
 Gene markers, 242  
 Gene regulation, 667–693. *See also Gene expression*  
     activation in, 669  
     site-specific recombination in, 489, 489f  
 analysis of  
     chemical modification interference in, 704b–705b  
     chemical protection footprinting in, 704b  
     DNA footprinting in, 704b  
     electrophoretic mobility shift assay in, 704b–705b  
     attenuation in, 684  
     in bacteria. *See Gene regulation in bacteria (below)*  
     in bacteriophages, 720–726. *See also Bacteriophage(s)*  
     cAMP receptor protein in, 675  
     in cancer, 667, 668  
     chromatin condensation and, 686  
     chromatin structure and, 678  
     coactivators in, 671, 672f, 739, 740f, 741–742  
     combinatorial control in, 677, 677f  
     corepressors in, 671, 672f  
     covalent, 164–167, 165t, 684, 686  
     dimerization in, 679, 747–748, 748f  
     at distant sites, 670–673, 671f–673f  
     DNA looping in, 670–673, 671f–673f  
     DNA-binding proteins in, 141–144  
     effectors in, 674–675, 674f, 675f  
     efficiency vs. adaptability in, 668–669  
     enhancers in, 532, 671, 693, 737  
         distance from promoters and, 741, 751  
         for *eve* gene, 749–750, 750f  
     in eukaryotes. *See Gene regulation in eukaryotes (below)*  
     feedback loops in, 673–675, 674f, 675f  
     global, 675, 676f, 686, 686f, 779–782  
     inheritance of, 684  
     insulators in, 673, 673f, 751–752, 751f  
     by intracellular localization, 686–687  
     mRNA localization in, 571  
     negative, 670  
         effectors in, 674–675, 674f  
     nucleosomes in, 678, 686  
     overview of, 668–669  
     positive, 670  
         effectors in, 674–675, 675f  
     posttranscriptional, 684–692  
         in bacteria, 604b, 712–720  
         by covalent modification, 686–692  
     elongation in, 684  
     in eukaryotes, 767–806  
     5' end processing in, 684–685  
     mRNA splicing in, 684  
     signaling in, 686–687, 687f  
     3' end processing in, 684–685  
     translational, 685–686, 686f  
     promoters in, 669–670, 670f, 676–677, 677f. *See also Promoter(s)*  
     in bacteria, 524–525, 524f, 669–670, 670f, 676, 722–725, 722f, 724f  
         Lac, 524–525, 524f  
     in eukaryotes, 676–677, 677f  
     phage, 722–725, 722f, 724f  
 protein degradation and, 668, 668f, 690–692, 691f  
 in protein modification, 668, 668f. *See also Protein modification*  
     by phosphorylation/  
         dephosphorylation, 686, 687, 688b–689b  
     by sumoylation, 686, 687  
     by ubiquitination, 686, 690–692  
 protein targeting in, 686–689  
 in protein transport, 668, 668f  
 regulatory sites in, 670  
 repression in, 669  
 in RNA processing, 668, 668f, 684–692  
 in RNA stability, 668, 668f  
 RNA structures in, 199b  
 signaling in, 673–675, 674f, 675f, 686–687  
 sites of, 668, 668f  
     in bacteria vs. eukaryotes, 675–677  
 site-specific recombination in, 487–488  
 structural basis of, 678–682. *See also Motifs*  
 transcription rate and, 670  
 transcriptional, 668, 668f, 669–678. *See also Transcription factors*  
     in bacteria, 698–712  
     in eukaryotes, 733–766  
 translational, 668, 668f  
     in bacteria, 685–686, 686f  
     in eukaryotes, 685–686, 686f, 772–778  
 in yeast, in mating-type switches, 746–747, 747f  
 Gene regulation in bacteria, 697–720  
     combinatorial control in, 742  
     global, 675, 676f, 686  
     multilevel, 725–726, 728  
     posttranscriptional, 712–720  
         riboswitches in, 604b, 712–716. *See also Riboswitch(es)*  
         r-protein synthesis and rRNA availability, 716–720, 718b, 719f  
         stringent response in, 719–720, 720f  
         translational, 685–686  
     promoters in, 669–670, 670f, 676, 698, 722–725, 722f, 724f. *See also Promoter(s)*  
     transcriptional, 698–712  
         attenuation in, 709–710, 709f  
         negative, 698–703

- operons in, 698–710. *See also* Operons  
amino acid biosynthetic, 708–710,  
708f, 709f  
*ara*, 707–708  
*gal*, 708, 708f  
*his*, 710  
*lac*, 698–707. *See also* lac operon  
*leu*, 710  
*phe*, 710  
*trp*, 708–710, 708f, 709f  
positive, 703–707  
quorum sensing in, 697, 711, 729  
of related genes, 710–711, 711f  
SOS response in, 710–711, 711f  
vs. in eukaryotes, 734
- Gene regulation in eukaryotes, 733–806  
coactivators in, 739, 740f, 741–742  
in development, 789–800. *See also*  
Development  
in gene networks, 675–676, 676f, 686,  
779–782  
AU-rich elements in, 779–782  
5'UTRs in, 779, 780b  
RNA splicing in, 779, 780b  
3'UTRs in, 779–782, 780b  
global, 675, 676f, 686, 686f, 779–782  
posttranscriptional, 767–806  
alternative splicing in, 769–771, 770f  
5'UTRs in, 774, 775f  
gene silencing in, 782–789  
inhibitors in, 768–769, 769f  
initiation factors in, 773–774, 774f  
natural antibiotics in, 768–769, 769f  
in nucleus, 767–772  
phosphorylation in, 773–774, 774f  
repressors in, 773  
in reticulocytes, 774  
RNA interference in, 782–789  
3' end cleavage sites in, 771, 772f  
3'UTRs in, 767, 774–776, 775f  
translational, 685–686, 686f, 772–778.  
*See also* Translation, regulation in  
eukaryotes  
transcriptional, 733–766  
activators in, 739–742, 740f  
basic mechanisms of, 734–742  
chromatin remodeling and, 735–736,  
735f, 736f  
coactivators in, 740f, 741–742  
combinatorial control in, 737, 742–751.  
*See also* Combinatorial control  
dosage compensation in, 753, 755–756,  
755f, 756f  
enhancers in, 737  
gene silencing in, 752–753, 754b, 782–789  
general transcription factors in,  
739–742, 740f  
hormone receptors in, 742, 756–758,  
757f–759f  
imprinting in, 753–755, 753f  
in initiation, 736–769  
insulators in, 751–752, 751f  
nucleosomes in, 678
- phosphorylation in, 758–760, 759f  
positive, 736–742  
promoters in, 676–677, 677f, 736–742  
protein kinase A in, 759f, 760  
repressors in, 740, 740f  
RNA splicing and, 738b–739b  
signaling in, 756–760  
specificity of, 739  
upstream activator sequences in, 737  
vs. in bacteria, 734  
vs. translational regulation, 768–769  
in yeast, 743–746, 743f, 746f, 754b, 784
- Gene repression, 669  
Gene silencing, 752–753  
epigenetic inheritance of, 786, 786f  
heterochromatin in, 753, 754b  
methylation in, 754b  
microRNA in, 685, 783–784, 784f  
RNA interference in, 782–789  
siRNA in, 685, 754b, 783, 784–786, 785f  
in yeast, 784
- Gene therapy  
human artificial chromosomes in, 300  
somatic, 300
- General rate constant, 151
- General transcription factors, 739–742, 740f.  
*See also* Transcription factors
- Genetic code, 585–610, 586  
adaptor hypothesis and, 15, 586, 608  
in *Candida albicans*, 603–604  
codons of, 588–589, 588f, 589t  
codon-anticodon pairing and,  
587–588, 588f  
nonoverlapping, 594, 594f  
deciphering, 596–601  
amino acid replacement in mutant  
proteins and, 601  
defined-sequence RNA polymer assays  
in, 598–601  
protein synthesis in cell extracts and,  
597–598, 597f, 598t
- definition of, 586
- degeneracy of, 586, 588–589, 589f  
mutations and, 591–592
- evolution of, 602–604, 605b. *See also*  
Ribozymes (catalytic RNA)
- exceptions to, 601–606  
in free-living cells, 604–605  
in mitochondrial DNA, 602–604, 603t
- frameshift mutations and, 412–413, 413f,  
594–595
- gaps in, 594–595, 595f
- linearity of, 596
- mutations and, 51–593
- in *Mycoplasma capricolum*, 603
- reading frames and, 590–591
- reading of, 616
- rules for, 593–596  
exceptions to, 601–606
- specificity of, 588
- as triplet code, 586, 588–589, 588f, 595–596
- tRNA as adaptor and, 587–588, 608
- universality of, 586, 601–606, 610
- in vivo validation of, 601, 610
- wobble and, 589–590, 589f, 590f
- Genetic diseases. *See* Diseases and disorders
- Genetic diversity  
independent assortment and, 469, 469f  
metagenomic sampling of, 268b  
sources of, 271, 288–291  
crossovers as, 468–469, 469f  
mutations as, 5, 10–11, 50, 411–412, 801
- Genetic drift, 288
- Genetic engineering. *See* Biotechnology
- Genetic mutations. *See* Mutations
- Genetically modified organisms, 251
- Genetics  
definition of, 24  
historical perspective on, 24
- Mendelian, 25–31. *See also* Mendelian  
genetics  
molecular, 44–55  
terminology of, 27t
- Genome(s)  
chimpanzee, 271–272  
definition of, 216, 260  
*E. coli*, 261, 261f  
human  
components of, 269–271, 270f  
polymorphisms and, 288–291  
sequencing of, 269–271  
size of, 548  
interspecies similarities in, 801  
minimal, 283
- Neanderthal, 263, 264b–265b  
retroviral, 300, 301t, 500, 502f  
size of, 269, 548, 801  
viral, 300–301, 301t
- Genome annotation, 263
- Genome sequencing, 260–263, 261f  
applications of, 266–269  
for archaea, 269  
for bacteria, 261, 261f, 266–269  
databases for, 266–271, 277  
development of, 260–263  
for *E. coli*, 261, 261f  
for eukaryotes, 269  
genome annotation and, 263–266  
for *Haemophilus influenzae*, 261, 292  
for modern humans, 269–271, 270f  
for Neanderthals, 263, 264b–265b  
for viruses, 261, 266, 267t, 292  
whole-genome shotgun, 262
- Genomic databases, 266–271
- Genomic diversity, sources of, 271–274, 273f,  
288–291
- Genomic Eve, 288, 290
- Genomic libraries, 224–225
- Genomic polymorphisms, genetic diversity  
and, 288–291
- Genomic RNA, 195, 298, 300
- Genomics  
comparative, 264–266, 281–282, 281f  
definition of, 260  
evolution and, 282–291  
metagenomics and, 267–269

- Genotoxic agents, 419  
 Genotypes, 27  
 Germ cells, imprinting in, 353  
 Germ-line cells, meiosis in, 466b  
 Germ-line development, in hermaphrodite nematodes, 776, 776f, 804  
*giant* gene, 795  
 Giant protein, 749, 750f  
 Gibbs, Josiah Willard, 87, 87f  
 Gilbert, Walter, 171, 171f, 228  
 Giroux, Craig, 477  
 $glmS$  riboswitch, 713t, 714–716, 717f  
 Globin gene, IFN- $\beta$ , 352, 355f  
 Glucose, *lac* operon and, 703–705, 706f  
 Glucose metabolism, regulation of, 688b–689b  
 Glucose transporters, 698b  
 GLUT4, 698b  
 Glutamate, 98, 98f, 99t  
 Glutamate receptor channels, RNA editing in, 566, 567f  
 Glutamic acid, methylation of, 67f  
 Glutamine, 97, 98f, 99t  
   in triplet expansion diseases, 413–414, 414t  
 Glutathione-S-transferase tag, 240–241, 240t, 251f  
 Glycine, 97, 98f, 99t  
 Glycogen metabolism, regulation of, 689b  
 Glycoproteins, posttranslational modification of, 654  
 Glycosidic bonds, 177  
 Glycosylation, 66–67, 67f  
   in base excision repair, 432–433, 432f, 433t  
   posttranslational, 657, 686  
 Golgi complex, 657, 657f  
 Goodman, Myron, 409  
 Gosling, Raymond, 206  
 Gottesman, Max, 509  
 Gottesman, Susan, 509  
 G-protein-coupled receptors, 759–760  
 Grasshoppers, chromosomes of, 36, 56  
 Grassquits, 801  
 Greek key motif, 11f, 110  
 Green fluorescent protein (GFP), 242, 243f, 253  
 Green, Rachel, 662  
 Greider, Carol, 398, 398f  
 Griffith, Frederick, 45–46, 45f  
 Griffith, Jack, 135  
 Grindley, Nigel, 481  
 GroEL/GroES chaperonins, 113, 120, 120f  
 Gross, Carol, 525–526, 526f  
 Group I introns, 559–561, 562f, 581  
 Group II introns, 559, 561–563, 563f, 564b  
 GST tag, 240–241, 240t, 251f  
 GTP (guanosine triphosphate)  
   in 5' end capping, 549, 550f  
   in protein targeting, 658, 658f  
   in RNA transport, 569, 569f  
   in translation  
     in elongation, 638, 640–641, 640f, 649f  
   in initiation, 633–634, 634f  
   in termination, 643, 643f  
   in translocation, 640, 641, 642f  
 GTPases  
   in translation, 633–634, 634f  
   in translocation, 640, 641, 642f  
 GTP-binding proteins, sequence analysis of, 130, 130f  
 Guanine (G), 48, 62, 177, 177f, 178. *See also* Base(s); Base pairs/base pairing  
   deamination of, 416–417, 417f  
   methylation of, 65–66, 67f, 205  
   nomenclature for, 64t  
 Guanosine, 178, 179f  
 Guanosine 3',5'-cyclic monophosphate (cGMP), 172, 182  
 Guanosine diphosphate (GDP)  
   in RNA transport, 569, 569f  
   in translation termination, 643, 643f  
 Guanosine tetraphosphate (ppGpp), 183  
   as second messenger, 720  
   as starvation signal, 720  
   in stringent response, 719–720, 720f  
 Guanosine triphosphate. *See* GTP  
   (guanosine triphosphate)  
 Gutell, Robin, 211  
 Guthrie, Christine, 780b  
*gypsy* elements, 498  
 Gyrase, 312t, 313, 315f, 325–326, 377t  
   discovery of, 324–326  
   in replication termination, 397–398  
 H3.3 histone, 346, 347f  
 Haber, James, 445  
*Haemophilus influenzae*, genome sequencing for, 261, 292  
 Hairpins  
   DNA, 109, 110f, 191, 192f  
   RNA, 197, 529, 531–532, 531f  
   terminator, 709–710  
 Haldane, J.B.S., 10, 149, 150f  
 Hall, Traci M., 787, 787f  
 Hammerhead ribozymes, 577  
 Hannon, Greg, 667  
 Haploid cells  
    $\alpha/\alpha$ , 470–471, 471f  
   gamete, formation of, 34–37, 35f  
   mating types of, 470  
 Haploids, 26  
 Haplotypes, 271, 272f  
   human migrations and, 288–289  
 Harvey, Stephen, 622, 622f  
*hAT* transposons, 505t  
 H2AX histone, 347, 347f  
 H2AZ histone, 346, 347f  
 HCR (hemin-controlled repressor), 774  
 Hda, 377t, 393f, 394  
 He, Lin, 667  
 Heat, bond breakage and, 72  
 Heat shock proteins, 119–120, 707  
   steroid hormones and, 687, 690f, 757  
 Heavy chains, immunoglobulin, 505–507, 506f  
*hedgehog (hh)* gene, 795, 796  
 Hedgehog (Hh) protein, 796, 802  
 Hitler, Walter, 70  
 HeLa cells, 401  
 Helical wheel, 111, 113f  
 Helicase II (UvrD)  
   in mismatch repair, 425t, 426, 426f  
   in nucleotide excision repair, 433t, 434, 434f  
 Rec A and, 458  
   in recombinational repair, 458  
 Helicases, 157–160, 157f–160f  
   autoinhibition of, 163  
   definition of, 157  
   directionality of, 158, 158f  
 DNA  
   in bacteria, 381, 382f, 383f, 384, 392–393, 392f  
   in eukaryotes, 387  
 in DNA unwinding, 157–160, 159f, 381, 382f, 392–393, 392f  
 in mismatch repair, 425t, 426, 426f  
 in oligomeric state, 158  
 in recombinational repair, 448  
 superfamilies of, 157f, 158  
 in translocation, 158, 158f  
*Helitron* transposons, 505t  
 Helix  
   alpha, 104–105, 104f, 107, 108f  
   hydrophobic residues in, 111–112, 113f  
   interaction of, 111–112, 112f  
   recognition, 679, 679f  
   in ribbon diagrams, 107, 108f  
   supersecondary structures and, 110–112, 111f, 112f  
 amphipathic, 112, 113f  
 DNA. *See* Double helix  
 handedness of, 104, 104f  
 hinge, 702  
 RNA, 65, 66f, 196–198, 196f  
 Helix-loop-helix motif, 680–681, 681f  
 Helix-turn-helix motif, 111, 112f, 679–680, 680f  
   in Cro, 723, 723f  
   in  $\lambda$  phage, 723, 723f  
   in Lac repressor, 143–144, 702  
   in transcription factors, 680, 680f  
 Hemin-controlled repressor, 774  
 Hemoglobin  
   mutations in, 276  
   structure of, 52b, 66f, 112, 113f  
 Hemoglobin A, 52b  
 Hemoglobin S, 52b–53b  
 Hemophilia, 50, 51f  
 Henderson-Hasselbalch equation, 83–84  
 Henkins, Tina, 718b  
 Hepatitis C virus, 787, 788b–789b  
 Heptad repeats, 111–112, 113f  
 Hereditary nonpolyposis colorectal cancer, mismatch repair and, 428b–429b, 429  
 Hermaphrodite nematodes, germ-line development in, 776, 776f, 804

- Hershey, Alfred, 46, 58  
 Hershey-Chase experiment, 46, 58  
 Hertwig, Oskar, 34  
 Heterochromatin, 735, 735f. *See also* Chromatin  
   in gene silencing, 753, 754b  
   in transcription repression, 735–736, 735f, 736f  
   in X chromosome inactivation, 350b–351b  
 Heterodimerization, in combinatorial control, 747–748, 748f  
 Heterooligomers, 112, 113f  
 Heterotropic enzymes, 161–164  
 Heterozygosity, 27  
 HGM-D, 741, 741f  
 Hierarchical model, of protein folding, 117, 117f  
 High-mobility group (HMG) proteins, 741, 741f  
 Hin recombinase, 487  
 Hinge, in Lac repressor, 702  
 Hin-hix recombination, 485, 488–489, 488f  
 HIR proteins, 348  
 HIRA, 348  
*his* operon, 710  
 Histidine, 98, 98f, 99t  
 Histidine (his) tags, 240t, 241  
 Histone(s), 333–334  
   acetylation of, 348–350, 351b, 352t, 354f, 360, 736  
   in RNA splicing, 733, 737  
   assembly of, chaperones in, 348  
   cis-acting, 343  
   core, 334, 334f  
   deacetylation of, 736, 736f  
   definition of, 332  
   DNA binding by, 141, 335, 336–339, 337f  
   epigenetic changes in, 353–357  
   functions of, 348  
   histone code and, 351–352, 355f  
   linker, 338–339, 338f  
   methylation of, 350–351, 351b, 354f, 736  
   in gene silencing, 754b  
   in imprinting, 755  
   modification of, 343–344, 348–357  
   covalent, 165, 736  
   gene expression and, 352–357  
   posttranslational, 735  
   organization of, 333–334  
   phosphorylation of, 351, 354f, 736  
   in gene regulation, 758–760, 759f  
   preservation during replication, 354–357, 356f, 357f  
   propagation of, 356–357, 357f  
   properties of, 333t  
   sumoylation of, 686, 687, 736  
   30 nm filament and, 339–341, 340f  
   in transcription, 336–338, 337f  
   types of, 333t  
   ubiquitination of, 686, 690, 736  
   variant, 346–348, 348f, 349f  
     vs. wild-type, 346, 358f  
     in X chromosome inactivation, 350b–351b  
 Histone acetyltransferases (HATs), 348–350, 352t, 360, 741, 742f  
 in RNA splicing, 733, 738b–739b  
 Histone chaperones, 348, 349f, 354  
 Histone code, 351–352, 355f  
 Histone deacetylases (HDACs), 350, 352t, 736  
 Histone H1, 337–339, 338f, 339f  
 Histone H2, 346, 347–348, 347f  
 Histone H3, 346–347, 347f  
 Histone H5, 339  
 Histone modifying complexes, 356, 357f, 360  
 Histone modifying enzymes, 343–344, 344f  
 Histone octamer, 334, 334f  
   assembly of, 359  
   crystal structure of, 334–336, 334f  
   DNA supercoiling around, 334–336, 335f, 336f  
 Histone subunits, 346–347, 347f, 354, 356f  
   in histone octamer, 334–336, 359  
 Histone tails, 336–338, 337f  
   modification of, 343, 348–357, 349f. *See also* Histone(s), modification of  
   30 nm filament and, 339–341, 340f  
 Histone variants, 343–344, 735  
 Histone-fold dimer, 335, 335f  
 Histone-fold motif, 335, 335f  
 HIV. *See* Human immunodeficiency virus  
 HMG proteins, 741, 741f  
*HMR $\alpha$*  locus, 470–471, 471f  
*HMR $\alpha$*  locus, 470–471, 471f  
 H.M.S. *Beagle*, 9, 21  
 Hoagland, Mahlon, 15, 49, 586, 608  
 Hofmeister, Wilhelm, 32  
 Holliday intermediate(s)  
   dimeric chromosomes and, 460–461, 461f  
   in double-strand break repair, 449, 450, 451f  
   in fork regression, 451, 452f  
   in gap repair, 453, 453f  
   gene conversion and, 468, 468f  
   in meiotic recombination, 468  
   resolution of, 451, 452f, 460, 461f  
   RuvAB and, 458–460  
   in site-specific recombination, 484, 484f  
   translocation of, 459–460, 460f  
 Holliday intermediate resolvases, 450, 453, 454t, 460  
 Holliday, Robin, 440f, 449  
 Holoenzymes, 145  
   DNA, 378, 378f, 384, 384f  
   RNA, 519, 519f  
 Homeodomain motifs, 680, 680f  
 Homeotic genes, 793, 796–798, 797f  
 Homing endonucleases, 472, 472f  
 Hominids, evolution of, 288, 289f  
*Homo erectus*, 288, 289f  
*Homo habilis*, 288, 289f  
*Homo neanderthalis*  
   evolution of, 288, 289f  
   genome sequencing for, 263, 264b–265b  
*Homo sapiens*, 288, 289f  
*Homo* species. *See also under* Human  
   evolutionary divergence of, 288, 289f  
 Homodimers, 112, 113f  
 Homologous chromosomes (homologs), 30–31, 265, 446, 464–465, 464f  
 Homologous genes, 265  
 Homologous recombination. *See* Recombination, homologous  
 Homooligomers, 112  
 Homotrimers, 112, 113f  
 Homotropic enzymes, 161–162  
 Homozygosity, 27  
 Hoogsteen pairings, 191, 193f  
 Hoogsteen positions, 191, 193f  
 Hooke, Robert, 32  
 Horizontal gene transfer, 11, 11f, 286, 720  
 Hormone(s)  
   evolution of, 273–274, 273f  
   heat shock proteins and, 687, 690f, 757, 757f  
   in signaling, 742, 756–758, 757f–759f  
 Hormone receptors, 687, 742, 756–758, 757f–759f  
   as activators and repressors, 742  
   evolution of, 273–274, 273f  
   nonsteroid, 758–760  
   steroid. *See* Steroid hormone receptors  
   structure of, 758, 758f  
   thyroid, 757–758, 757f  
   types of, 757–760, 757f, 758f  
 Hormone response elements, 758, 758f  
 Hot spots, 465  
 House mouse. *See* *Mus musculus*  
 Housekeeping genes, 669  
 Hox genes, 796–798, 797f, 802  
*Hoxa-7*, 798  
*Hsp70*, 119–120, 120f  
   steroid hormones and, 687, 690f, 757, 757f  
 HU, 377t, 392  
 Huang, Wenhua, 20  
 Human artificial chromosomes, 299–300  
 Human evolution. *See also* Evolution  
   genomic alterations in, 271–274, 273f  
   migration in, 288–289  
   mitochondrial Eve in, 288, 290  
   out of Africa theory of, 288  
   sources of variation in, 271–274, 288–291  
   species divergence in, 288, 289f  
   Y chromosome Adam in, 288, 290  
 Human genome  
   components of, 269–271, 270f  
   evolutionary alterations in, 271–274, 273f  
   polymorphisms and, 288–291  
   sequencing of, 269–271, 270f  
   vs. chimpanzee genome, 271–272

- Human Genome Diversity Project, 288  
 Human Genome Project, 260–262, 263f  
*See also* Genome sequencing
- Human immunodeficiency virus  
 evolution of, 7b  
 genome of, 300  
 nuclear transport in, 772, 773f  
 phylogenetic analysis of, 284b  
 as retrovirus, 166b, 501–503  
 Rev and, 772, 773f  
 RNA of, 199b
- Human immunodeficiency virus infection, 501–503  
 antiretroviral agents for, 153b  
 development of, 166b–167b  
 CCR5 mutation and, 54  
 drug therapy for, 153b, 166b–167b  
 reverse transcriptase inhibitors  
     for, 502b  
 vaccine for, 503
- Human selectivity factor 1 (SL1), 534
- Hunchback, 749, 750, 750f, 794–795, 794f
- Huntington disease, 50–52, 51f, 414, 414t
- Hurwitz, Jerard, 518, 518f
- Hybrid duplexes, 201–203
- Hybridization techniques, 201–203,  
 202f–204f  
 oligonucleotide synthesis for, 206, 207f
- Hybrids, 25
- Hycamtin (topotecan), 318b–319b
- Hydrogen, valence of, 70, 71f
- Hydrogen atoms, in catalysis, 91
- Hydrogen bonds, 76, 76f, 77, 77f, 103–106  
 in  $\alpha$  helix, 104–105, 104f  
 in  $\beta$  sheet, 105–106, 105f  
 DNA-binding proteins and, 141–142  
 in double helix, 181, 184f, 187  
 in histone-fold dimer, 335, 335f  
 protein folding and, 115  
 in RNA, 198, 198f  
 in wobble base pairing, 590
- Hydrogen cyanide, adenine synthesis from, 19, 19f
- Hydrogen peroxide, as mutagen, 418
- Hydrogen tunneling, 61, 91
- Hydrolysis  
 abasic site and, 417  
 of ATP, 84, 85f, 88–89, 158–160, 159f  
 of bonds, 88–89  
 deamination by, 203–204, 205f,  
 416–417, 417f  
 depurination by, 417, 418f  
 methylation by, 163–165, 417, 418f.  
*See also* Methylation  
 mutations due to, 416–417  
 of nucleic acids, 179–180, 182f
- Hydronium ions, pH and, 82
- Hydrophobic interactions, 75, 75f, 77, 77f
- Hydrophobic residues, in alpha helix, 111–112, 113f
- Hydrophobic stacking  
 in DNA, 181, 184f  
 in RNA, 197
- Hydroxyl radicals, 418, 419f  
 as mutagens, 418
- Hypatia, 13, 13f
- Hyperchromic effect, 201
- Hypersensitive sites, 735–736
- Hypochromic effect, 201, 201f
- Hypotheses, 12–13, 14–15
- IF-1, 633–634, 635f
- IF-2, 633–634, 635f
- IF-3, 633–634, 635f
- IFN $\beta$  enhanceosome, 752, 752f
- IgM, alternative 3' end cleavage in, 771, 772f
- Ile-tRNA synthetase, proofreading by, 628
- Illumina sequencer, 235b
- Imitation switch (ISWI), 344, 345f, 346t
- Immune system  
 antibodies in, 505–507  
 in immunofluorescence, 242  
 protein chips and, 282  
 evolution of, 507  
 van der Waals interactions in, 74–75, 75f
- Immunoblots, 243–244, 245f
- Immunofluorescence, 242–243
- Immunoglobulin(s), 505–507  
 evolution of, 507
- Immunoglobulin M, alternative 3' end  
 cleavage in, 771, 772f
- Immunoprecipitation, 244–246, 245f, 282  
 chromatin, 345
- Importins, 569, 569f, 658, 658f
- Imprinting, 753–755, 753f  
 in germ cells, 353
- Incomplete dominance, 29–30, 30f
- Indels, 412–414, 413f, 414t
- Independent assortment, 28–29, 29f, 29t  
 genetic diversity and, 469, 469f
- Indinavir (Crixivan), 166b–167b
- Indirect immunofluorescence, 242, 244f
- Induced fit, 136, 714
- Induced pluripotent stem cells, 533b
- Inducers, lac operon, 701, 703, 703f
- Inductive reasoning, 13
- Influenza virus, evolution of, 129
- Information flow. *See also* Gene regulation  
 backward, 500  
 direction of, 596, 609, 631
- Inheritance  
 chromosome theory of, 38–44  
 of dominant traits, 25, 26f, 29–30, 30f  
 codominance and, 30, 30f  
 incomplete, 29–30, 30f  
 epigenetic. *See* Epigenetic inheritance  
 Mendelian, 25–29  
 of mutations, 410  
 non-Mendelian, 29–31  
 of recessive traits, 25, 26f
- Inherited diseases. *See* Diseases and disorders
- Initial model, in X-ray crystallography, 123–124
- Initial velocity, 149, 150f
- Initiation complex, 535–536, 536f, 633–634,  
 635f, 638, 639f, 640f, 737
- Initiation factors, 629, 685–686, 686f
- Initiation (start) codons, 591, 629  
 variant, 604
- Inosine, base pairing and, 590
- Insect viruses, cloning with, 237–238, 237f
- Insertion mutations, 412–414, 413f, 414t, 415f  
 three-nucleotide, 413, 413f
- Insertion sequences, 496–497, 497f
- Insertion sites, 370, 370f
- Insulators, 673, 673f, 751–7521, 751f
- Insulin regulation, by phosphorylation, 687,  
 688b–689b
- int* gene, 725–726, 726f
- Integrases, 495  
 transposases and, 503, 504f  
 in viral infections, 725–726
- Intercellular communication. *See* Signaling
- Interfering RNAs, 517b
- Internal ribosome entry site, 635–637, 637f
- International HapMap, 288
- International Human Genome Project,  
 260–262, 263f. *See also* Genome  
 sequencing
- Interphase, 35f, 36  
 in meiosis, 35f, 36  
 in mitosis, 34, 35f
- Intervening sequences. *See* Intron(s)
- Intron(s), 269–270, 270f, 549, 555  
 AT-AC, 558  
 in bacteria, 555  
 definition of, 549  
 discovery of, 554–555  
 in eukaryotes, 555  
 evolution of, 503–505, 564b, 779  
 functions of, 555–556  
 mobile, 472, 472f, 563, 564b  
 removal of, 554–555, 558. *See also*  
 Splicing
- in RNA transport, 569–570
- self-splicing, 554, 559–563, 561f–565f  
 in bacteria, 562–563, 564b  
 group I, 559–561, 562f, 581  
 group II, 559, 561–563, 563f, 564b  
 as ribozymes, 576
- splice sites on, 558
- spliceosome recognition of, 558
- surveillance of, HIV repression of,  
 772, 773f
- in yeast, 555, 780b
- Intron-retaining mRNA  
 evolution of, 779  
 nuclear transport of, 772, 773f
- Inversions, chromosomal, 272, 415, 415f
- Inverted repeats, 191, 192f, 679, 679f
- Invirase (saquinavir), 166b–167b
- Ion-exchange chromatography,  
 100, 100f
- Ionic bonds, 146–147  
 strong, 68–70, 68f, 70f  
 weak, 74, 77, 77f
- Ionizing radiation  
 bacterial resistance to, 7b, 462b–463b  
 mutations and, 422–423, 424f, 460

- Ions, 68–69  
 IPTG (isopropyl  $\beta$ -D-1-thiogalactopyranoside), 703, 703f  
 Irinotecan (Campto), 318b–319b  
 Iron homeostasis, 779, 781f  
 Iron response elements (IREs), 779, 781f  
 Iron response proteins (IRPs), 779, 781f  
 IRP-IRE complexes, 779, 781f  
 Irreversible enzyme inhibitors, 153b  
 IS elements, 496–497, 497f  
 IS50 elements, 497  
 Isoenergetic reactions, 488  
 Isoleucine, 99t  
 Isomerase, in protein folding, 121, 121f  
 Isomorphous replacement, 125  
 Isoprenyl groups, posttranslational attachment of, 654  
 Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG), 703, 703f  
 ISWI (imitation switch), 344, 345f, 346t  
 Izsvák, Zsuzanna, 499b
- J segments, immunoglobulin, 506–507, 506f  
 Jacob, François, 169f, 170, 390, 699–701  
 JAK-STAT pathway, 687, 687f  
 Janssens, F.A., 41  
 Jeffreys, Alec, 230b  
 Johnson, Tracy, 733, 733f, 738b–739b  
 Jones, Kathy, 515  
 Journals, scientific, 13, 16  
 Joyce, Gerald, 194b–195b, 578  
 Junk DNA, 482, 505  
 Jurica, Melissa, 547
- $K_a$  (acid dissociation constant), 83  
 Kadonaga, Jim, 515  
 Kaiser, Dale, 251  
 Kappa light chains, 505–507, 506f  
 Karyopherins, 569  
 $K_d$  (dissociation constant), 137 acid, 83 for protein-ligand interactions, 137–138, 138t  
 Keeney, Scott, 477  
 Kelner, Albert, 440  
 Kendrew, John, 124, 125f  
 Kennedy disease, 52  
 Khorana, Gobind, 206, 599, 599f  
 Kilocalories per mole, 72  
 Kilodalton (kDa), 54  
 Kim, Peter, 81, 81f  
 Kim, Sung-Hou, 194, 196f  
 Kimble, Judith, 767, 776, 804  
 Kinases, cyclin-dependent, in replication initiation, 395  
 Kinetic isotope effect, 91  
 Kinetic proofreading in transcription, 529, 530f in translation, 628, 628f  
 Kinetics enzyme, 149–151 reaction, 85  
 Kinetochores, 299
- Kleckner, Nancy, 477  
 Kleisins, 318  
 Klinman, Judith, 61, 91  
 Klug, Aaron, 194, 196f  
 Kolodner, Richard, 428b–429b  
 Kornberg, Arthur, 14, 16, 363, 371, 403  
 Kornberg, Roger, 334, 359, 359f, 538  
 Kornberg, Thomas, 403  
 Kossel, Albrecht, 44  
 Kozak, Marilyn, 631  
 Kozak sequence, 631, 633t  
*krüppel* gene, 749, 750f, 795  
 Krüppel protein, 749, 750f  
 Ku70-Ku80 complex, in nonhomologous end joining, 473–474, 473f, 474f  
 Kunkel, Tom, 429b
- lac* genes, 699, 699f  
 Lac operator, 142–144, 143f, 170, 701  
   Lac repressor dissociation from, 701  
   O<sub>1</sub>, 702, 702f  
   O<sub>2</sub>, 702, 702f  
   O<sub>3</sub>, 702, 702f  
   release of, 703  
*lac* operon, 142–143, 525, 698–707  
   activation of, 701  
   DNA looping in, 702, 702f  
   inducers of, 701, 703, 703f  
   Jacob-Monod experiments on, 699–701  
   merodiploid analysis of, 699–701, 700f  
   regulation of, 698–707  
     catabolite repression in, 703  
     CRP-cAMP in, 675, 706–707, 706f  
     glucose vs. lactose availability and, 703–705, 706f  
     negative, 674–675, 698–703  
     positive, 674–675, 703–707  
     structure of, 699, 699f  
     transcription in, 673–675, 675f  
   Lac promoter, 524–525, 524f  
   Lac repressor, 142–144, 143f, 673–675, 674f, 701  
   dissociation constant for, 138t  
   dissociation from Lac operator, 701  
   in DNA looping, 671f  
   early studies of, 169–171  
     helix-turn-helix motif in, 143–144  
     structure of, 702, 702f  
*lacA*, 699, 699f  
*lacI*, 699–701  
 Lacks, Henrietta, 401  
*lacO*, 699–701  
 Lactose metabolism in *E. coli*, 698 early studies of, 169–171  
 Lactose operon. *See lac* operon  
*lacY*, 699, 699f, 701  
*lacZ*, 169, 699, 699f, 701  
   Gal4p and, 744b–745b  
 Lagging strand, 369  
   in replication, 384–385, 385f
- Laipis, Philip, 323  
 Lamarck, Jean-Baptiste, 21  
 Lambda ( $\lambda$ ) phage, 170, 301t, 485, 486f, 496f  
   exonuclease, 217  
   gene regulation in, 720–726  
   integration into host cell, 724  
   life cycle of, 725–726, 728  
   lytic and lysogenic pathways of, 485–487, 486f, 497, 721–722, 721f, 722f  
   switching between, 722–725, 724f  
   as model system, 722  
   prophage induction in, 722–725, 724f  
   site-specific recombination in, 509  
   structure of, 722f  
 Lambda  $\lambda$  repressor, 722–725, 723f  
   as activator, 724–725  
   self-cleavage of, 725  
   in SOS response, 725  
   structure of, 723, 723f  
   synthesis of, 723  
 Lambda light chains, 505–507, 506f  
 Landick, Robert, 542  
 Landy, Art, 509  
 Last universal common ancestor (LUCA), 8–9, 8f, 9f, 283, 586, 602  
 Law of independent assortment, 28–29, 29f, 29t  
 Law of segregation, 25–28, 28f  
 Leader peptide, 709f, 710  
 Leader sequence, of mRNA, 709, 709f  
 Leading strand, 369  
 Lebowitz, Jacob, 323  
 Lederberg, Esther, 476  
 Lederberg, Joshua, 170, 476  
 Lehman, Robert, 14, 16, 363  
*leu* operon, 710  
 Leucine codons, 588t, 606  
   in yeast, 606  
 Leucine repeats, 111–112  
 Leucine zipper motif, 112, 680, 682, 682f  
   basic, 680, 681f  
 Levine, Michael, 750, 750f  
 Levinthal, Cyrus, 116  
 Levinthal's paradox, 116–117  
 Levorotary (L-form) enantiomers, 79, 80f, 81b  
 Lewis, Edward B., 792, 793f  
 Lewis, Gilbert Newton, 68, 68f  
 LexA, 724–725  
   in SOS response, 710–711, 711f  
 LexA-Gal4 fusion protein, 683–684, 683f  
 L-form enantiomers, 79, 80f, 81b  
 Libraries. *See DNA libraries*  
 Lichtman, Jeff, 490b  
 Liddle syndrome, 691  
 Life. *See Living systems*  
 Ligands, 136–144. *See also*  
   Protein-ligand interactions  
   definition of, 136  
 Ligase. *See DNA ligase*  
 Light chains, immunoglobulin, 505–507, 506f

- LINEs, 498, 500  
 Linkage analysis, 274–275, 276f  
 Linked genes, 30–31, 31f  
   inheritance of, 38–40, 38f, 40f  
   segregation of, 40, 40f, 56  
   unlinking of, 40–41, 41f, 42f  
 Linker DNA, in nucleosomes, 334, 334f  
 Linker histones, 338–339, 338f  
 Linkers, 219  
 Linking number (*Lk*), 308–310, 308f  
   topoisomerase-induced changes in, 312–313, 314f, 317f  
   twist and, 310, 310f  
   writhe and, 310, 310f  
 Liphardt, Jan, 297  
 Lipid-enclosed RNA systems, 1, 4f, 6, 20  
 Literature mining, 282  
 Livet, Jean, 490b  
 Living systems  
   characteristics of, 3–4, 3f  
   definition of, 2  
   evolution of. *See* Evolution  
   origin of, 5–8. *See also* Prebiotic chemistry; RNA world hypothesis  
   requirements for, 2–3  
   RNA-based, 6–7  
 lncRNA (long noncoding RNA), 517b  
 Lobban, Peter, 251  
 Lohman, Tim, 135  
 London, Fritz, 70  
 Long interspersed nuclear elements (LINEs), 498, 500  
 Long noncoding RNA (lncRNA), 517b  
 Long patch repair, 432  
 Long RNA (lRNA), 517b  
 Long terminal repeats, 493  
   in retrotransposons, 493, 495, 496f, 497–499  
 Loops  
   DNA. *See* DNA looping  
   telomere, 401, 401f  
 Loss of function mutations, 411  
 LoxP sites, 487  
 lRNA (long RNA), 517b  
 LTR retrotransposons, 493, 495, 496f, 497, 498  
 LUCA (last universal common ancestor), 8–9, 8f, 9f, 283, 586, 602  
 Lukyanov, Sergey, 253  
 Luria, Salvador E., 58  
 Lwoff, André, 169, 169f  
 Lymphocytes  
   B  
     differentiation of, 505–507, 506f  
     IgM and, 771, 772f  
   T  
     in HIV infection, 501–502  
     insulators in, 751–752, 751f  
 Lymphoma, B-cell, microRNA in, 667  
 Lyon, Mary, 350b  
 Lysine, 97, 98f, 99t  
   bromodomains and, 350, 352t, 356  
   chromodomains and, 350, 356
- Lysogen, 722  
 Lysogenic pathway, of bacteriophages, 485–487, 486f, 497, 721f, 722, 722f  
 Lysosomes, protein targeting to, 657  
 Lytic pathway, of bacteriophages, 485–487, 486f, 497, 721–722, 721f, 722f
- M phase. *See* Mitosis  
 Machado-Joseph disease, 52, 414t  
 Macro H2A, 347–348, 350b–351b  
 Mad cow disease, 118–119  
 Magnesium ions  
   in replication, 375, 376f, 518, 519  
   in RNA, 69, 70f, 77f, 198, 198f  
   RNA polymerase and, 518  
 Major groove, 186  
   of double helix, 186  
   of recognition helix, 679, 679f  
 Malaria, sickle-cell anemia and, 54–55  
 Malthus, Thomas, 21  
 Maltose-binding protein, 240t  
 Mammalian cell cultures, in cloning, 238  
 Map(s)  
   electron density, 123, 123f  
   genome, 261–262, 262f  
   recombination, 42–44, 42f, 43f  
 Map units, 42–43  
 MAPK cascade, 688b–689b  
 Marcu, Ken, 215  
 Margulies, Ann Dee, 476, 476f  
 Margulis, Lynn, 303, 303f  
 Markers. *See also* Tagging  
   screenable, 221  
   selectable, 221  
 Markov, Georgi, 647b  
 Marzluff, William, 553b  
 Maskin, 774, 775f  
 Mass, molecular, 54  
 Mass spectrometry, 280–281  
 MAT locus, 470–471, 470f, 746  
 Maternal genes, 793–795, 794f  
 Mating types, 470, 746  
 Mating-type switches, 470–472, 470f, 471f, 746–747, 747f  
 Matthaei, Heinrich, 597, 597f  
 Matthews, Brian, 131  
 Maverick transposons, 505t  
 Maxam, Allan, 228  
 Maxam-Gilbert sequencing, 171  
 Maximum velocity ( $V_{\max}$ ), 149  
 Mayo, Steve, 15, 95, 116  
 McClintock, Barbara, 41, 271, 271f, 482, 482f  
 Mcm1, 746–747, 747f  
 MCM helicase, 387  
 Mediator complex, 538–539, 538f, 741  
 Meiosis, 34–37, 35f  
   in germ-line cells, 466b  
   homologous recombination in  
     *in bacteria*, 447–464. *See also*  
       Recombination, homologous  
       crossing over in, 40–41, 41f, 42f  
       double-strand breaks in, 445, 446, 465–468, 465f, 467f
- in eukaryotes, 464–469. *See also* Meiotic recombination  
 nondisjunction in, 39–40, 39f  
 phases of, 35f, 36, 465–468  
 vs. mitosis, 35f, 36  
 Meiotic recombination, 464–469  
   chromosome segregation in, 464–465, 464f, 465f  
   crossovers in, 465, 465f, 467–468, 467f  
   double-strand breaks in, 465–468, 465f, 467f, 468f  
   gene conversion and, 468, 468f  
   initiation of, 465–468, 465f, 467f  
   Spo11 in, 465–467, 467f, 477  
   transesterification in, 466–467  
 Mello, Craig, 783, 784, 784f, 785, 805  
 Melting, of DNA, 200, 200f  
 Melting point ( $T_m$ ), of DNA, 201  
 Membrane, cell, 32, 33f  
   evolution of, 8–9, 8f, 602  
   protein targeting to, 657  
 Mendel, Gregor, 10, 24, 25f, 31  
 Mendelian genetics, 25–31  
   first law of, 25–28, 28f  
   non-Mendelian inheritance and, 29–31  
   rediscovery of, 31–32, 38  
   second law of, 28–29, 29f, 29t  
 Menten, Maud, 150, 150f  
 Merlin transposons, 505t  
 Meselson, Matthew, 48, 363, 364–366  
 Meselson-Stahl experiment, 363, 364–366, 366f  
 Messenger RNA. *See* mRNA (messenger RNA)  
 Metagenomics, 267–269  
 Metal ions  
   in active sites, 146–147  
   in replication, 375, 376f  
   in RNA, 198, 198f  
 Metals, bonding of, 69  
 Metamerism, 791  
 Metaphase, 35f, 36  
   in meiosis, 35f, 36  
   in mitosis, 34, 35f  
 Metaphase plate, 34  
 Methionine, 97, 98f, 99t  
   in posttranslational modification, 654  
 3-Methyladenine, in base excision repair, 433, 433t  
 Methylation, 163–165, 164f, 686  
   in direct repair, 430, 431f  
   DNA, 205, 205f  
     in replication initiation, 393, 393f  
   epigenetic marks and, 753  
   in 5' end capping, 549–550, 550f  
   in gene silencing, 754b  
   histone, 350–351, 351b, 354t, 736  
   in imprinting, 753–755, 753f  
   in mismatch repair, 426, 426f, 439  
   mutations due to, 417  
   nucleotide, 65–66, 67f, 178, 205–206, 205f, 417  
   posttranslational, 654, 655f  
   RNA, 206, 573

- Methylguanine, repair of, 430, 431f  
 7-Methylguanosine, in capping, 549  
 Methyltransferases, 205  
 Met-tRNA synthetase, 631, 632, 633–634  
 Mice  
     as model organisms, A-3-A-4, A-18-A-19  
     transgenic, site-specific recombination and, 490b–491b  
 Michaelis, Leonor, 150–151  
 Michaelis constant ( $K_m$ ), 150, 150f  
 Michaelis-Menten equation, 150–151, 150f  
 Microarrays, DNA, 248–250, 248f–250f  
     in nucleosome localization, 345–346  
     in transcriptome analysis, 278  
 Microprocessor complex, 575, 575f  
 MicroRNA (miRNA), 517b, 575–576  
     in *Caenorhabditis elegans*, 805  
     in cancer, 667, 685  
     definition of, 783  
     discovery of, 805  
     in gene silencing, 685, 783–784, 784f  
     precursor, 784  
     primary transcript, 784, 785f  
     processing of, 575–576, 575f  
     transient, 685  
     types of, 517b, 685  
 Microsatellite instability, in cancer, 429b  
 Microscopes, 33–34  
     cryo-electron, 617  
 Miescher, Friedrich, 44  
 Mig1, 746, 746f  
 Migration, 288–289  
 Miller, Stanley, 89b, 585  
 Mimitou, Eleni, 445  
 Minimal medium, 46  
 Minor bases, 178  
 Minor groove, 186  
     of double helix, 186  
     of recognition helix, 679, 679f  
 Mi2/NURD chromatin remodeling complex, 344–345, 346t  
 miRNA. *See* MicroRNA (miRNA)  
 Mirror repeats, 191, 192f  
 Mismatch repair, 412, 425–429.  
     *See also* DNA repair  
     colon cancer and, 428b–429b  
     definition of, 425  
     in *E. coli*, 425–427, 425t, 427f  
     in eukaryotes, 425t, 427–428, 428f  
     methylation in, 426, 426f, 439  
     proteins in, 425–429, 425t  
 Missense mutations, 411, 591, 592  
 Mitochondria  
     evolution of, 288, 290, 303–304, 602–604, 621, 632  
     protein targeting to, 655–659. *See also* Protein targeting  
         ribosomes of, 619, 621, 622b  
 Mitochondrial DNA (mtDNA), 288, 290, 303–304, 304f  
     genetic code variations and, 602–604, 603t  
     mutations in, 289–290  
 Mitochondrial Eve, 288, 290  
 Mitochondrial tRNA, genetic code  
     alterations in, 602–604, 603t  
 Mitosis, 32–34, 34f, 35f, 299  
     double-strand breaks in, 469–470, 470f  
     phases of, 34, 34f, 35f  
     vs. meiosis, 35f, 36  
 Mitotic recombination, 465, 469–470, 470f  
 Mitotic spindle, 34  
     developmental symmetry and, 790  
 Mixed enzyme inhibitors, 152b–153b  
 Mizuuchi, Kiyoshi, 509, 510–511  
 Mobile genetic elements. *See* Transposon(s)  
 Mobile introns, 472, 472f, 563, 564b  
 Model building, 15  
 Model organisms, A-1-A-19  
     *Arabidopsis thaliana*, A-3-A-4, A-14-A-15  
     *Caenorhabditis elegans*, A-3-A-4, A-12-A-13  
     chromosomes of, 302t  
     DNA of, 302t  
     *Drosophila melanogaster*, 38, A-3-A-4, A-16-A-17  
     *Escherichia coli*, 217, A-3-A-4, A-6-A-7  
     genes of, 302t  
     in medical research, A-2-A-5  
     *Mus musculus*, A-3-A-4, A-18-A-19  
     *Neurospora crassa*, 46, A-3-A-4, A-10-A-11  
     overview of, A-1-A-5  
     *Saccharomyces cerevisiae*, 394–395, 395f, A-3-A-4, A-8-A-9  
 Modrich, Paul, 429b, 439  
 Mold. *See* *Neurospora crassa*  
 Mole, 72  
 Molecular biology, 44–55  
     definition of, 2  
     historical perspective on, 24  
     information flow in, 5f  
     overview of, 2–5  
     units of measure in, 300  
 Molecular function, in genome annotation, 263  
 Molecular genetics, 44–55. *See also under* Genetic  
     Molecular mass, 54  
 Molecular motors. *See* Motor proteins  
 Molecular orbital model, 70  
 Molecular replacement, 125  
 Molecular weight, 54  
 Molten globule model, 117  
 Monod, Jacques, 48, 169, 169f, 170, 699–701  
 Monosomes, 467, 467f  
 Morgan, Thomas Hunt, 10, 24, 38–40, 38f, 42  
 Motifs, 109–112, 110f–113f, 281, 335, 335f  
      $\beta$ - $\alpha$ - $\beta$ , 110–111, 111f  
     basic helix-loop-helix, 680–681, 681f  
     bromodomains, 350, 353f  
     chromodomains, 351, 353f  
     coiled-coiled, 111–112, 112f  
     databases of, 281  
     DDE, 503, 504f  
     DNA-binding, 679–682  
     Greek key, 11f, 110  
     helix-loop-helix, 680–681, 681f  
     helix-turn-helix, 111, 112f, 679–680, 680f, 682, 682f  
     in Cro, 723, 723f  
     in  $\lambda$  phage, 723, 723f  
     in Lac repressor, 143–144, 702  
     in transcription factors, 680, 680f  
     histone-fold, 335, 335f  
     homeodomain, 680, 680f  
     leucine zipper, 112, 680, 682, 682f  
         basic, 680, 681f  
     protein dimerization, 679–682  
     RNA recognition, 558  
     transcription-activation, 682  
     zinc finger, 681–682, 682f  
 Motor proteins, 156–161  
     in ATP hydrolysis, 158–160, 159f  
     functions of, 156, 159–160  
     helicases, 157–160, 157f–160f  
     ligand binding by, 156–157  
     RuvB, 159  
     Snf, 159–160  
     translocases, 158  
 Mouse. *See* Mice; *Mus musculus*  
 Mre11, 467, 467f  
 mRNA (messenger RNA), 48–49  
     defective  
         tmRNA salvage system for, 647–651, 651f  
         translation-coupled removal of, 647–653, 651f–653f  
     definition of, 516  
     degradation of, 568, 570f–572f, 571–572, 777, 779f  
         AU-rich elements in, 780–782, 782f  
         siRNA in, 784–786  
     discovery of, 48  
     5'-end capping in, 548, 549–550, 550f, 552–553, 631. *See also* 5' cap  
     functions of, 49, 516  
     half-life of, 571  
     intron-retaining, nuclear transport of, 772, 773f  
     leader sequence of, 709, 709f  
     localization of, 570–571, 571f  
     noncoding sequences in, 554–555, 555f.  
         *See also* Intron(s)  
     non-stop, 647–653, 651f–653f  
         in bacteria, 647–651, 651f  
         degradation of, 571, 652, 652f  
         in eukaryotes, 651–653, 652f, 653f  
         stalled ribosomes on, tmRNA rescue of, 647–651, 652f, 653f  
     polyadenylation of, 539, 539f, 550–553, 631, 652–653. *See also* Poly(A) tail  
 polycistronic, 675  
 riboswitches in, 712–716  
 splicing of, 551–565. *See also* Splicing  
 stability of, 685, 685f  
 synthesis of. *See* Transcription  
 terminator of, 709–710, 709f  
 3' end of, 550–552, 552b–553b. *See also* 3' end processing

- mRNA (messenger RNA) (*continued*)  
 transcription of. *See* Transcription  
 in translation, 619–620, 619f. *See also*  
 Translation  
 truncated  
   in bacteria, 647–651, 651f  
   in eukaryotes, 651–653, 652f, 653f  
   stalled ribosomes on, tmRNA rescue of,  
 647–651, 652f, 653f
- mRNA decay  
 nonsense-mediated, 652–653, 652f  
 non-stop, 652–653, 653f
- mRNA editing, 565–568, 567f, 568f
- mRNA processing, 547–572. *See also* RNA processing  
 mRNA transport, 568–571
- MSH2/MSH6, in mismatch repair, 425t, 427
- mtDNA (mitochondrial DNA), 288, 290, 303–304, 304f  
 genetic code variations and, 602–604, 603t  
 mutations in, 289–290
- MuDR/Foldback* transposons, 505t
- Mullis, Kary, 15, 252
- Multimers, 112
- Multiplicity of infection (MOI), 722
- Multipotent stem cells, 799
- Multiwavelength anomalous dispersion (MAD), 125
- Mus musculus*  
 chromosomes of, 302t  
 DNA of, 302t  
 genome of, 302t  
 as model organism, A-3–A-4, A-18–A-19
- Muscle differentiation, 763
- Mustard weed. *See* *Arabidopsis thaliana*
- Mutagen(s), 416–424. *See also* Mutations, causes of  
 carcinogenic, 419–421  
 Ames test for, 419–421
- Mutagenesis  
 oligonucleotide-directed, 239–240, 239f  
 site-directed, 239–240, 239f
- Mutant alleles, 38
- Mutations, 50–55, 203, 409–424, 410  
 back, 420, 421f  
 beneficial, 50, 54–55, 410, 411–412, 801  
 in cancer, 407, 412, 668  
 causes of, 416–424  
   alkylation, 418–419  
   cytotoxic agents, 420, 421f  
   deamination, 416–417, 417f  
 by nitrous acid, 418, 418f  
   depurination, 417, 418f  
   hydrolysis, 416–417  
   methylation, 417  
   oxidation, 418, 419f, 430, 431f  
 radiation, 421–423, 440  
 definition of, 2, 410  
 deletion, 412–413, 413f, 414, 415f  
 detrimental, 50–54  
 dimeric chromosomes and, 461, 461f
- disease-causing, 50–55, 412. *See also*  
 Diseases and disorders  
 databases of, 277  
 identification of, 274–277, 276f
- in DNA repair genes, 412
- duplication, 32–34, 34f, 35f, 272, 273f, 414–415, 415f
- in evolution, 5, 10–11, 50, 289–291, 411–412
- frameshift, 412–413, 413f, 594–595
- functional effects of, 411
- gain of function, 411
- induced, 239–240, 239f
- inheritance of, 410
- insertion, 412–414, 413f, 414t, 415f
- inversion, 272, 415, 415f
- large-scale, 414–416, 415f
- loss of function, 411
- missense, 411, 591, 592
- in mitochondrial DNA, 289–290
- nonsense, 411, 592, 647, 652–653, 653f  
 removal of, 652, 653f
- notation for, 411
- overview of, 410–411
- point, 411–412, 411f, 412f, 776, 776f, 804
- prevention of, 410. *See also* DNA repair  
 rate of, 286, 287
- reverse transcriptase and, 501, 502
- reversion, 420, 421f
- in RNA-dependent replication, 501
- silent, 50, 411, 441, 591, 592
- as source of genetic variation, 5, 10–11, 50, 411–412, 801
- suppression of, 592–593, 592f, 593f
- transition, 411, 411f, 592
- translocation, 415–416, 415f
- transversion, 411, 411f
- types of, 410–416
- MutH, in mismatch repair, 425t, 426, 428, 439
- MutL, in mismatch repair, 425t, 426, 426f, 427, 428, 439
- MutS, in mismatch repair, 425t, 426, 426f, 427, 439
- Mycoplasma genitalium*, genome of, 283
- Myotonic dystrophy, 414t
- Myristoylation, 163–165, 164f
- N antiterminator, 725, 725f
- NAD<sup>+</sup>, DNA ligase and, 153–156, 154f–155f
- NADP<sup>+</sup>, 182
- Nägeli, Karl Wilhelm von, 32
- Nalidixic acid, 318b
- nanos* gene, 793
- Nanos protein, 793
- NAP-1 chaperone, 348, 349f, 354
- Nash, Howard, 509
- Nathans, Daniel, 218
- Natural selection, 9–12, 287–288, 801. *See also* Evolution  
 discovery of, 21
- NCBI website, 264, 266
- ncRNA (noncoding RNA), 517b
- Neanderthals  
 evolution of, 288, 289f  
 genome sequencing for, 263, 264b–265b
- Negative gene regulation, 670
- Negative supercoiling, 309, 309f, 313, 316f
- Nematodes. *See also* *Caenorhabditis elegans*  
 hermaphrodite, germ-line development  
 in, 776, 776f, 804
- as model organisms, A-3–A-4, A-12–A-13  
 trans-splicing in, 563
- Neuronal network tracing, brainbow  
 method for, 490b–491b
- Neurons, mRNA localization in, 570, 570f
- Neurospora crassa*  
 in Beadle-Tatum experiment, 46–47, 47f  
 as model organism, 46, A-3–A-4, A-10–A-11
- Niche, 799
- Nick translation, 372, 372f, 383, 386  
 in base excision repair, 431  
 in mismatch repair, 426, 426f, 427–428
- Nirenberg, Marshall, 597, 597f, 598
- Nitrogen, valence of, 70, 71f
- Nitrous acid, as mutagen, 418
- Noller, Harry, 211, 604b, 615, 617, 620
- Nomura, Masayasu, 617, 617f
- Noncanonical base pairing. *See under* Wobble
- Noncoding RNA  
 long, 517b  
 stable, 517b
- Nondisjunction, 39–40, 39f
- Nonhomologous end joining, 435, 472–474, 473t, 474f
- Nonhuman primates  
 evolution of, 288, 289f  
 genetic diversity in, 290  
 genome of, 271–272
- Non-LTR retrotransposons, 493, 495, 496f, 497, 499
- Nonpolar molecules, 73, 73f
- Nonsense codons. *See* Stop (termination) codons
- Nonsense mutations, 411, 592, 647, 652–653, 653f  
 removal of, 652, 653f  
 suppression of, 592–593, 592f
- Nonsense-mediated mRNA decay, 571, 652–653, 652f
- Non-stop mRNA  
 in bacteria, 647–651, 651f  
 degradation of, 571, 652–653, 652f, 653f  
 in eukaryotes, 651–653, 652f, 653f  
 stalled ribosomes on, tmRNA rescue of, 647–651, 652f, 653f
- Nontemplate strand, 519, 520, 520f
- Northern blotting, 203, 206
- Norvir (ritonavir), 166b–167b
- Novobiocin, 318b
- NPH-II RNA helicase, 160
- NTase, DNA ligase and, 156

- N-terminal amino acid residue, protein half-life and, 690, 690t
- N-terminal tails, 99, 686
- N-terminus, 99
- Nuclear localization sequence, 658
- Nuclear magnetic resonance (NMR), 125–127, 126f, 127f
- Nuclear Overhauser effect spectroscopy (NOESY), 125, 126–127, 126f
- Nuclear proteins, targeting of, 657–658, 658f
- Nucleases, DNA, 371–372
- Nucleic acids, 44. *See also* DNA; RNA annealing of, 200–201, 200f automated synthesis of, 206, 207f backbone of, 179–180, 181f chemical modifications of, 205–206, 205f chemical properties of, 200–208 chirality of, 79–80, 80f denaturation of, 200–201, 200f, 201f hydrolysis of, 179–180, 182f ionic bonds in, 69 length of, 180 nomenclature for, 64t notation for, 180 phosphodiester bonds in, 178–180, 181f polarity of, 179 renaturation of, 200–201, 200f, 201f resonance in, 71, 72f thermal properties of, 200–201, 200f, 201f
- Nuclein, 44
- Nucleoids, 341–342, 343f
- Nucleolytic proofreading, 529, 530f
- Nucleophiles, 84
- Nucleosides, 62, 177 nomenclature for, 64t
- Nucleosome(s) assembly of, chaperones in, 348, 349f beads-on-a-string appearance of, 334, 334f, 339 chromatin accessibility and, 344 crystal structure of, 334–335, 334f on daughter DNA strands, 354–356, 356f definition of, 332 DNA binding of, 335–336, 335f, 336f dynamic nature of, 344 functions of, 343 in gene regulation, 678, 686 histones in, 332–338. *See also* Histone(s) modifications of, 332, 343–348 inheritance of, 352–357 overview of, 332 positioning of chromatin remodeling complexes and, 343, 344–346, 344f, 346f, 346t determination of, 345–346, 347f prediction of, 331 in 30 nm filament, 339–341, 340f in transcription, 338 Nucleosome modifying enzymes, 343–344, 348–357
- Nucleotide(s), 62–64, 177–185 absorption spectra of, 181, 183f bases of. *See* Base(s); Base pairs/base pairing chemical composition of, 177–178, 177f chemical modifications of, 65–66, 167f, 178, 205–206, 205f. *See also* Protein modification definition of, 45, 177 DNA, 45, 62–64, 63f functions of, 181–183, 185f glycosidic bonds of, 177 methylation of, 65–66, 67f, 178, 205–206, 205f nomenclature for, 64t, 184 phosphodiester bonds of, 178–180, 181f postsynthetic changes in, 65–66, 67f RNA, 45, 62–64, 63f sequence of, 178–179, 180 structure of, 62, 63f, 177–180, 177–181, 177f–181f
- Nucleotide excision repair (NER), 433–435, 433t, 434f, 435f
- Nucleotide residues, 179
- Nucleus, protein targeting to, 657–658, 658f
- Nüsslein-Volhard, Christiane, 792, 793f
- nut* sites, 725, 725f
- OB fold, 140, 140f in DNA ligase, 154f, 156
- Objectivity, 12
- Ochoa, Severo, 14, 598, 598f
- Octet rule, 70
- Okazaki fragments, 369, 384–385, 385f, 386–387 RNA removal from, 372, 372f, 383 synthesis of, 384–385, 385f, 386f, 392f
- Okazaki, Reiji, 368, 369
- O<sub>L</sub>, 722f, 723, 724, 724f
- Oligomers, 112–113, 113f
- Oligonucleotide, 180
- Oligonucleotide synthesis, 206, 207f
- Oligonucleotide-directed mutagenesis, 239–240, 239f
- Oligonucleotide/oligosaccharide-binding (OB) fold, 140, 140f in DNA ligase, 154f, 156
- Omega ( $\omega$ ) protein, 324
- O<sup>6</sup>-methylguanine, repair of, 430, 431f
- On the Origin of Species* (Darwin), 9–10
- Oncogenes, 412
- One gene–one enzyme hypothesis, 47
- One gene–one polypeptide hypothesis, 47, 47f
- One-start helix model, 339
- Oocytes, mRNA localization in, 570, 570f
- Open complex, 392, 393 in replication, 392, 392f, 393, 393f in transcription, 521, 521f, 526–527, 528f
- Open reading frames (ORFs), 591, 591f upstream, 777
- Open-form Pol I, 374, 374f
- Operons, 675, 675f amino acid biosynthetic, 708–710, 708f, 709f *ara*, 707–708, 707f attenuation of, 709–710, 709f definition of, 675 *gal*, 708, 708f *his*, 710 *lac*, 142–143, 525, 673–675, 698–707. *See also* lac operon *leu*, 710 *phe*, 710 in regulons, 675, 707 regulons and, 675 *trp*, 708–710, 708f, 709f in *Bacillus subtilis*, 728
- Optical trapping, 344, 345f
- Optically active substances, 78
- O<sub>R</sub>, 722f, 723, 724, 724f
- O<sub>R1</sub>, 723
- O<sub>R2</sub>, 723
- O<sub>R3</sub>, 723
- Orbitals, atomic, 70
- Organelles, 9, 32. *See also* Chloroplast(s); Mitochondria DNA of, 288, 290, 303–304 evolution of, 621, 632 protein targeting to, 655–659. *See also* Protein targeting
- Orgel, Leslie, 6, 19
- oriC*. *See* Replication origin, in bacteria
- Origin of life. *See* Prebiotic chemistry; RNA world hypothesis
- Origin of replication. *See* Replication origin
- Origin replication complex (ORC), 390t, 395
- Oro, Juan, 19, 89b
- Orr-Weaver, Terry, 468
- Orthologs, 265, 266
- Orum, Henrik, 788b–789b
- Oryza sativa* chromosomes of, 302t DNA of, 302t genome of, 302t
- Oskar, 570
- osm* gene, 526
- Out of Africa theory, 288
- Out-groups, 273–274
- Overman, Les, 135
- Oxidation in base excision repair, 433, 433t mutations due to, 418, 430
- Oxygen, valence of, 70, 71f
- P bodies, 571–572, 572f, 777
- p53* gene, in cancer, 668
- P182* gene, in Alzheimer’s disease, 275, 276f
- P generation, 25, 26f, 26t, 27t
- P1 phage, 485–487, 487f Cre-Lox system of, 487, 487f, 488, 490b–491b
- P site, 621–623, 623f, 624
- Painter, Robert, 436
- Pair-rule genes, 793, 795

- Palade, George, 616  
 Palindromes, 190–191, 192f  
 Parallel  $\beta$  sheet, 105f, 106  
 Paralogs, 265–266  
 Pardue, Arthur, 170  
 Parthenogenesis, 755  
 Pasteur, Louis, 78  
 Pattern-regulating genes, 792–798  
   homeotic, 793, 796–798, 797f  
   maternal, 793–795, 794f  
   segmentation, 793, 795–796, 796f, 797f  
 Pauling, Linus, 69, 69f, 92, 99, 104  
 Pauling scale, 69f  
*Pax6*, 802  
 pBR322 plasmid, 221, 221f, 222f  
 PCNA, 112, 113f  
 PCNA clamp, 388–389, 389f, 390t  
   in mismatch repair, 425t, 428  
   in nucleotide excision repair, 433t, 434, 434f, 435f  
 PcrA, 160  
 PDB files, 114  
 $P_e$ , 769–771, 770f  
 Peer review, 13, 16  
 Pentoses, in nucleic acids, 177, 177f, 178, 178f, 186  
 Peptide(s), 64, 65f. *See also* Protein(s)  
   definition of, 97  
   leader, 709f, 710  
 Peptide bonds, 71, 72f, 92, 99–102, 102f  
   cis and trans isomers of, 99–102, 102f  
   formation of, 616, 621–624, 622b, 638–640, 639f  
   resonance of, 71, 72f, 102f  
   structure of, 92  
   torsion angles of, 102, 103f  
   in translation, 616, 621–624, 622b, 638–640, 639f, 644  
 Peptide prolyl cis-trans isomerase, 121, 121f  
 Peptide translocation complex, 657  
 Peptidyl transferase center, 616, 621, 621f  
 Peptidyl transferase reaction, 639–640, 641f, 644  
 Periodic table, of electronegativity, 69, 69f  
 Perutz, Max, 124, 125f  
 Pfeiffer, Richard, 292  
 pH, 81–84  
   of blood, 81, 82  
   buffers and, 82–84, 83f  
   definition of, 82  
 pH scale, 82, 82f  
 Phages. *See* Bacteriophage(s)  
 Phase problem, in X-ray crystallography, 124–125, 124f  
 Phase variation, 487–488, 488f  
*phe* operon, 710  
 Phenotypes, 26, 26t  
   primary exceptional, 39–40, 39f  
   wild-type, 38–40, 38f  
 Phenotypic function, in genome annotation, 263  
 Phenylalanine, 98, 98f, 99t  
 Phi ( $\phi$ ) angle, 102, 102f  
 Phosphate, in nucleotides, 177–178, 177f  
 Phosphodiester bonds  
   of ATP, 88–89  
   in nucleic acids, 178–180, 181f  
 Phosphorus-nitrogen bonds, 89  
 Phosphorus-oxygen bonds, 89  
 Phosphorylation, 66–67  
   in gene regulation, 686, 688b–689b, 758–760, 759f  
   histone, 351, 354f, 736  
   insulin, 687, 688b–689b  
   posttranslational, 654, 655f, 686  
   protein kinases in, 165–167, 165t  
   in signaling, 686–687, 687f, 759f, 760  
 Photolithography, for DNA microarrays, 248–249, 248f  
 Photolyase, 429–430, 430f, 440  
 Photoreactivation, 429–430, 430f, 440  
 Photosynthesis, 9  
 Phylogenetic profiling, 281, 281f  
 Phylogenetic trees, 272, 273f, 283–286, 285f–287f  
 Phylogenetics, 283–287  
   in forensics, 284b  
 Phylogeny, 283–287  
*PIF/Harbinge* transposons, 505t  
*PiggyBac* transposons, 505t  
 $pK_a$ , 83  
 $P_L$ , 722–725, 722f, 724f  
 Plants  
   breeding of, 25, 26f, 26t, 27t  
    $F_1$  generation in, 25, 26f, 26t, 27t  
    $F_2$  generation in, 25, 26f, 26t, 27t  
   in Mendel's experiments, 25–31  
    $P$  generation in, 25, 26f, 26t, 27t  
   chloroplasts of. *See* Chloroplast(s)  
 Plasma membrane, 32, 33f  
   evolution of, 8–9, 8f  
   protein targeting to, 657  
 Plasmids, 220–222  
   antibiotic resistance and, 302, 303b  
   as cloning vectors, 219–222, 222f  
   definition of, 220  
   DNA of, 302, 302f  
   functions of, 220–221, 302  
   nonbacterial, 304  
   pBR322, 221, 221f, 222f  
   size of, 302, 302f  
   Ti, 221  
 Plectonemic supercoiling, 310–311, 311f  
 P-loop, 130, 130f  
 Pluripotent stem cells, 798f, 799  
   transcription factors and, 533b  
 $P_m$ , 769–771, 770f  
 PME point mutation, 776, 776f  
 pOH, 82  
 Point mutation, 411–412, 411f, 412f  
 Point mutation element (PME), 776, 776f, 804  
 Pol  $\alpha$ , 387, 390t  
 Pol  $\beta$ , in translesion synthesis, 437, 437t  
 Pol  $\delta$ , 388, 390t  
   in nucleotide excision repair, 433t, 434, 435f  
 Pol  $\epsilon$ , in nucleotide excision repair, 433t, 434, 435f  
*pol* gene, 500, 502f  
 Pol  $\eta$ , in translesion synthesis, 437, 437t  
 Pol I  
   DNA. *See* DNA polymerase I (Pol I)  
   RNA, 519, 519t, 534t  
   promoters for, 532–534, 534f, 534t  
 Pol  $\iota$ , in translesion synthesis, 437, 437t  
 Pol II  
   DNA. *See* DNA polymerase II (Pol II)  
   RNA. *See* RNA polymerase II (Pol II)  
 Pol III  
   DNA. *See* DNA polymerase III (Pol III)  
   RNA, 519, 519t, 534t  
   promoters for, 534t, 535, 535f  
 Pol III core, 377–378, 378t  
 Pol III holoenzyme  
   DNA, 382f, 383, 392  
   in mismatch repair, 425t, 426  
   RNA, 519, 519t  
 Pol IV, 373, 373t  
   in translesion synthesis, 437, 437t  
 Pol  $\lambda$ , in translesion synthesis, 437, 437t  
 Pol V (DNA polymerase V), 373, 373t  
   discovery of, 409, 441  
   in translesion synthesis, 437, 441  
 Polar covalent bonds, 69, 72  
 Polar molecules, 73, 73f  
 Polar R groups, 97–98  
 Polarity, in development, 791, 793–795, 794f  
 Poly(A) addition site, 551  
 Poly(A) binding protein, 551, 551f, 634, 774  
 Poly(A) site choice, 556, 557f  
 Poly(A) tail, 550–553  
   addition of, 551–552  
   functions of, 551  
   in non-stop mRNA decay, 652–653, 652f  
   synthesis of, 550–551, 551f  
   gene regulation in, 684–685  
   in transcription, 539, 539f  
   in translation, 631  
 Polyacrylamide gel electrophoresis. *See* Gel electrophoresis  
 Polyadenylate polymerase, 551, 551f  
 Polyadenylation factors, 551, 551f  
 Polyadenylation signal. *See* Poly(A) tail  
 Polycistronic mRNA, 675  
 Polyglutamine (polyQ) diseases, 413–414  
 Polylinkers, 219, 220f  
 Polymerase chain reaction (PCR), 219, 226–228, 227f  
   development of, 15  
   discovery of, 252  
   in forensics, 230b–231b  
   real-time, 228, 228f  
   reverse transcriptase, 228  
 Polymers, self-replicating, 6, 8, 20  
 Polynucleotide, 180  
 Polynucleotide kinase, in recombinant DNA technology, 217t

- Polypeptide(s), 64, 65f. *See also* Protein(s)  
definition of, 47, 97
- Polypeptide backbone, 99, 102, 102f
- Polypeptide chains, 99
- PolyQ diseases, 413–4143
- Polyribosomes, 620, 620f
- Polysomes, 620, 620f
- porcupine* gene, 796
- Positive gene regulation, 670
- Positive supercoiling, 309, 309f
- Postinsertion site, 370, 370f
- Posttranscriptional processing, 65, 67f, 547–581. *See also* RNA processing
- Posttranslational protein modification, 654, 655f  
in endoplasmic reticulum, 655–657, 656f  
in gene regulation, 686  
in Golgi complex, 657, 657f
- Postulate of objectivity, 12
- Potassium ions, in RNA, 198, 198f
- PpGpp (guanosine tetraphosphate), 183  
as second messenger, 720  
as starvation signal, 720  
in stringent response, 719–720, 720f
- PP<sub>i</sub> (pyrophosphate)  
from ATP hydrolysis, 84, 85f, 88–89  
in replication, 370–371
- P<sub>R</sub>, 722–725, 722f, 724f
- Prasher, Douglas, 253
- P<sub>RE</sub>, 723, 724f
- Prebiotic chemistry, 5, 6f, 19, 19f, 20, 578  
adenine in, 19, 19f, 89b  
ATP in, 19, 19f, 89b  
genetic code evolution and, 604b–605b  
ribose in, 585  
RNA world hypothesis and, 6–8, 15, 88, 548, 554, 578, 601–602. *See also* Ribozymes (catalytic RNA)  
selenocysteine in, 604–606  
sodium montmorillonite in, 20, 20f
- Precis coenia*, wing development in, 806
- Precursor miRNA (pre-miRNA), 575–576, 575f, 784
- Preinitiation complex, 535–536, 536f, 634, 635, 636f, 740–741
- Premature stop codons, 547, 652–653, 653f
- Pre-miRNA, 575–576, 575f, 784
- Pre-mRNA, 784  
editing of, 565–568, 567f, 568f  
processing of, 549–553. *See also* RNA processing  
splicing of, 554–565, 652, 653f. *See also* Splicing
- Prepriming complex, 392
- Prereplication complex, 390t, 395
- Preribosomal RNA (pre-rRNA), 532–534  
processing of, 573–575, 575f  
transcription of, 534, 534f, 534t
- Presenilin, in Alzheimer's disease, 275, 276f, 277
- Pre-steady state kinetics, 149
- Primary exceptional phenotype, 39–40, 39f
- Primary miRNA transcripts (pri-mRNAs), 784, 785f
- Primary protein structure, 97–103. *See also* Amino acid sequences
- Primary transcripts, 548–549. *See also* RNA processing
- Primases, 368, 382f, 383, 392  
in replisome, 384
- Primates, nonhuman  
evolution of, 288, 289f  
genetic diversity in, 290  
genome of, 271–272
- Primer(s), 368, 369, 382f, 392  
lack of, in transcription, 527–528  
in replication, 368, 369, 382f, 392  
for reverse transcriptase, 500
- Primer terminus, 369
- Pri-mRNAs, 784, 785f
- Prion diseases, 118–119, 118f
- P<sub>RM</sub>, 723, 724f
- Probes, in hybridization, 202
- Processing bodies, 571–572, 572f
- Processive synthesis, 376
- Processivity, 158, 375–376, 379–380, 380f, 381f  
definition of, 158  
in replication, 375–376, 379–380, 380f, 381f  
in transcription, 521, 529
- Processivity number, 376
- Products, of reactions, 84
- Proenzymes, 167–168
- Progesterone receptor, evolution of, 273f
- Programmed cell death, 401
- Proline, 97, 98f, 99t  
in  $\beta$  sheet, 106  
in  $\alpha$  helix, 105  
as helix breaker, 105  
hydroxylation in, 67f
- Promoter(s), 517b, 519, 522, 669–670, 670f  
in bacteria, 523–526, 524f, 669–670, 670f, 698, 722–723, 722f  
in bacteriophages, 722–725, 722f, 724f  
core, 737  
distance from regulatory sites, 670–673  
enhancers for, 532, 671, 693  
in eukaryotes, 532–535, 534t, 535f, 676–677, 677f, 736–742  
accessibility of, 735, 737  
activator binding to, 736–739  
complexity of, 737–738, 738f  
distance from activators, 737, 741  
DNA looping and, 741, 741f  
enhancers and, 737, 741, 751  
for Pol I, 532–534, 534f, 534t  
for Pol II, 534–535, 534t, 535f, 536f, 541  
for Pol III, 534t, 535, 535f  
structure of, 737, 737f  
upstream activator sequences and, 737
- Lac, 524–525, 524f  
in yeast, 737, 738, 738f
- Promoter clearance, in transcription, 521, 521f
- Promoter-associated transcripts, 517b
- Proofreading, 371–372, 371f, 375, 403, 425  
kinetic  
in transcription, 529, 530f  
in translation, 628, 628f  
nucleolytic, 529, 530f  
in replication, 371–372, 371f, 375, 403, 425  
ribosomal, 652–653, 662  
in transcription, 520–521, 529, 530f  
in translation, 627–629, 628f, 638, 644
- Prophage, 485, 722. *See also* Bacteriophage(s)
- Prophage induction, 722–725
- Prophase  
in meiosis, 35f, 36, 465  
in mitosis, 34, 35f
- Proproteins, 167–168
- Prosthetic groups, 145  
posttranslational attachment of, 654
- Protease inhibitors, development of, 166b–167b
- Proteasomes, 690
- Protein(s)  
allosteric, 161–164  
alpha helix of, 104–105, 104f  
amino acids in. *See* Amino acid(s)  
binding sites on, 107, 136  
cen, 299  
chemical modification of. *See* Protein modification  
chirality of, 79–80, 80f, 81b  
conformational changes in, 136–137  
covalent modification of, 163–165, 164f.  
*See also* Protein modification
- cross-linked, 334
- definition of, 97
- DNA-binding. *See* DNA-binding proteins
- enzyme. *See* Enzyme(s)  
evolution of, 8–9, 114
- fatty acid modification in, 67f
- fibrous, 112
- fusion, 240–241, 282  
in gene localization, 242–243, 243f  
in immunofluorescence, 242, 243f
- globular, 112
- glycosylation of, 66–67, 67f
- half-life of, 690, 690t
- heterotropic, 161–164
- homotropic, 161, 162
- hydroxylation of, 67f
- localization of, 242–244  
cDNA libraries in, 242–243  
immunofluorescence in, 242
- methylation of, 163–165, 164f
- motor, 156–161
- nuclear, targeting of, 657–658, 658f
- peptide bonds of. *See* Peptide bonds
- phosphorylation of, 66–67, 67f
- polypeptide backbone of, 99, 102, 102f
- polypeptide chains of, 99
- postsynthetic changes in, 66–67, 67f

- Protein(s) (*continued*)
   
regulation of. *See* Protein function, regulation of
   
resonance in, 71, 72f
   
ribbon diagrams of, 107, 108f
   
ribosomal, 617, 617t, 618, 621
   
targeting of, 657
   
sequence homology in, 102–103, 130
   
subunits of, 112–113, 113f
   
cooperativity among, 136–137
   
synthesis of. *See* Translation
   
weak bonds in, 76–77
- Protein amplification, 233–236
- Protein chips, 282
- Protein Data Bank (PDB), 114, 121
- Protein design, computational, 95, 115–116, 116f
- Protein dimerization motifs, 679–682.
   
*See also* Motifs
- Protein disulfide isomerase, 121
- Protein domains, 107–108, 109f
- Protein families, 114
- Protein folding, 115–121, 654
   
chaperones/chaperonins in, 119–120, 120f, 654
   
defects in, 118–119
   
free-energy funnel model of, 117–118, 117f
   
hierarchical model of, 117, 117f
   
isomerases in, 121, 121f
   
mechanisms of, 116–118
   
molten globule model of, 117
   
reverse turns and, 106, 106f
   
thermodynamics of, 117–118, 117f
- Protein function, 64, 65, 135–171
   
ligands and. *See* Protein-ligand interactions
   
principles of, 136–137
   
in protein interaction networks, 282
   
regulation of, 161–168
   
by autoinhibition, 163, 163f
   
by covalent modification, 164–167, 166t, 686–692
   
enzymatic, 161–165, 161f–164f
   
by phosphorylation, 164f, 165–167, 166t
   
by proteolytic cleavage, 167–168
- Protein function studies, 242–250
   
cDNA libraries in, 242–243, 244f
   
comparative genomics in, 281–282, 281f
   
computational approaches in, 281–282
   
immunofluorescence in, 242, 243f
   
immunoprecipitation in, 244–246, 245f
   
protein-protein interactions in, 244–248
   
proteomics and, 280–281
   
purification methods in, 244–246, 245f
   
sequence relationships and, 281
   
structural relationships in, 281
   
transcriptome analysis and, 277–278
   
Western blotting in, 243–244, 245f
- Protein interaction networks, 282
- Protein isomerases, in folding, 121, 121f
   
Protein kinase A, in gene regulation, 759f, 760
   
Protein kinase cascade, in insulin regulation, 688b–689b
   
Protein kinases
   
consensus sequences for, 165, 165t
   
in phosphorylation, 165–167
   
Protein modification
   
by acetylation, 163–165, 164f, 206, 348–350, 351b, 352t, 354f, 360, 654
   
covalent, 164–167, 165t, 166t, 684, 686–692
   
by dephosphorylation, 687
   
by glycosylation, 66–67, 67f, 432–433, 432f, 433t, 657, 686
   
histone, 343–344, 348–357, 386. *See also* Histone(s), modification of
   
by methylation, 163–165, 164f, 686.
   
*See also* Methylation
   
by phosphorylation, 165–167, 351, 354f, 654, 655f, 686, 687, 688b–689b
   
posttranslational, 654–658, 655f, 686
   
in endoplasmic reticulum, 655–657, 656f
   
in gene regulation, 686
   
in Golgi complex, 657, 657f
   
by sumoylation, 686, 687, 736
   
by ubiquitination, 163–165, 164f, 686, 690–692, 691f
   
Protein phosphatases, 165
   
Protein purification, by immunoprecipitation, 244–246, 245f
   
Protein residues. *See* Amino acid residues
   
Protein sorting, in Golgi complex, 657, 657f
   
Protein structure, 65, 66f, 95–131
   
atomic, 121–127
   
computational design and, 95, 115–116, 116f
   
databases of, 114, 281
   
determination of
   
by nuclear magnetic resonance, 125–127, 126f, 127f
   
by X-ray crystallography, 121–125, 122f–124f
   
evolution of, 114
   
folding and, 115–121
   
functional correlates of, 281–283
   
motifs and, 109–112, 110f–113f, 281, 335, 335f. *See also* Motifs
   
overview of, 96–97
   
primary, 97–103. *See also* Amino acid sequences
   
quaternary, 107, 112–113, 113f
   
secondary, 103–107, 108f
   
α helix in, 104–105, 104f, 107, 108f
   
β sheet in, 105–106, 105f, 107, 108f
   
reverse turns in, 106, 106f, 107, 108f
   
ribbon diagrams for, 107, 108f
   
tertiary, 107–112
   
determination of, 121–127
   
by nuclear magnetic resonance, 125–127
   
by X-ray crystallography, 121–125
   
Protein superfamilies, 114
   
Protein tagging, 240–241, 240t, 241f, 266
   
for degradation, 690–692, 691f
   
epitope, 242, 244–246, 245f
   
in non-stop mRNA decay, 652–653, 652f
   
TAP, 246
   
by tmRNA, 651, 651f
   
by ubiquitin, 690–692, 691f
   
Protein targeting, 655–659, 656f, 686–689
   
in bacteria, 658, 659f
   
chaperones in, 657
   
endoplasmic reticulum in, 655–657, 656f
   
in eukaryotes, 655–658, 656f–658f
   
glycosylation in, 657
   
Golgi complex in, 657, 657f
   
to lysosomes, 657
   
to plasma membrane, 657
   
signal recognition particle in, 655–657, 656f
   
Protein world, 8–9
   
Protein-coding genes, accelerated evolution in, 274, 275f
   
Protein-ligand interactions, 136–144
   
association constant for, 137
   
binding energy in, 146
   
binding sites in, 136
   
conformational changes in, 136–137
   
cooperativity in, 136–137, 139
   
dissociation constants for, 137–138, 138t
   
DNA-binding proteins in, 138–144
   
in enzyme-catalyzed reactions.
   
*See* Catalysis
   
equilibrium expression for, 137, 137f
   
functional implications of, 136, 137
   
ligand affinity in, 137
   
motor proteins in, 156–157
   
overview of, 136–137
   
quantification of, 137–138
   
regulation of, 137
   
reversible, 136–137, 137f
   
Proteolytic cleavage, 163–165, 167–168
   
Proteomes, 278–279, 556
   
Proteomics, 278–281
   
computational approaches in, 281–282
   
definition of, 278
   
mass spectrometry in, 280–281
   
phylogenetic profiling in, 281–282
   
two-dimensional gel electrophoresis in, 280–281, 280f
   
Protomers, 112, 113f
   
Prusiner, Stanley, 118b
   
PS1 gene, in Alzheimer's disease, 275, 276f
   
Psi (ψ) angle, 102, 102f
   
Ptashne, Mark, 683, 683f
   
PUF family, 775–776, 776f, 804
   
Pulsed field gel electrophoresis, 223–224
   
Pumilio, 794–795, 794f
   
Punnett squares, 27, 28f, 40f
   
Purebred organisms, 25
   
Purine bases, 177, 177f. *See also* Base(s); Base pairs/base pairing
   
Purine rings, notation for, 206

- Puromycin, translation inhibition by, 646, 646f
- Pyrimidine bases, 177, 177f. *See also* Base(s); Base pairs/base pairing
- Pyrimidine dimers  
photorepair of, 429–430, 430f, 440  
ultraviolet radiation and, 422
- Pyrimidine rings, notation for, 206
- Pyrophosphatase, in replication, 370
- Pyrophosphate bond, of ATP, 88
- Pyrophosphate ( $\text{PP}_i$ )  
from ATP hydrolysis, 84, 85f, 88–89  
in replication, 370–371
- Pyrophosphorylation, in replication, 370
- Q antiterminator, 725
- Quadruplex DNA, 192, 193f
- Quantitative polymerase chain reaction, 228, 228f
- Quantitative S1 mapping, 693
- Quaternary protein structure, 107, 112–113, 113f
- Quinolones, 318b
- Quorum-sensing bacteria, 697, 711, 729
- R factor, 124
- R group, 64, 65f, 97–98, 98f
- Rad50, 467, 467f
- Rad51, 467f, 468, 470
- Rad52, 470
- Radiation  
bacterial resistance to, 7b, 462b–463b  
gamma, 422–423, 462b–463b  
solar  
double-stranded breaks and, 429–430, 430f, 440, 460  
mutations due to, 421–423, 424f, 440  
in photoreversal, 429–430, 430f, 440
- RAG1, 507, 507f
- RAG2, 507, 507f
- Ramachandran plots, 102, 103f
- Ramakrishnan, Venki, 617
- Ran  
in protein targeting, 658, 658f  
in RNA transport, 569, 569f
- Rate constant, 85
- Rate-limiting step, 85
- Rational drug design, 166b–167b
- RBP1, 535
- RBP2, 535
- RBP3, 535
- RBP11, 535
- RDRC complex, 754b
- Reactants, 84
- Reaction(s), 84–90  
activation energy in, 84–85, 85f  
in catalytic reactions, 147–151  
catalytic, 2, 3, 87–88, 91, 145–156. *See also* Catalysis; Enzyme(s)  
energy coupling in, 89, 90b  
exothermic, 72, 88–89, 89b  
free energy change in. *See* Free-energy change ( $\Delta G$ )
- isoenergetic, 488  
notation for, 84  
products of, 84  
rate of, 85, 87–88, 91  
catalytic increase in, 147–151. *See also* Enzyme kinetics  
equilibrium and, 147–148  
rate-limiting step in, 85  
reactants in, 84  
steps in, 85–86, 86f  
strong, 68–73  
thermodynamic laws and, 86–87  
transition state in, 84, 148  
weak, 73–78. *See also* Weak chemical interactions
- Reaction intermediates, 85
- Reaction kinetics, 85
- Reaction mechanisms, 85, 154–155  
in replication, 370–375, 370f, 371f, 374f, 375f  
in transcription, 518, 518f
- Reactive oxygen species, as mutagens, 418
- Reading frames, 412–413, 413f, 590–591  
frameshift mutations and, 412–413, 413f, 594–595  
open, 591, 591f
- Real-time polymerase chain reaction, 228, 228f
- recA* gene  
cloning of, 222, 236, 240  
discovery of, 476  
transcription of, 457–458
- RecA protein, 157, 456, 724–725  
in bacterial radiation resistance, 462b  
functions of, 222, 236, 240  
in recombinational repair, 454, 454t, 455–458, 456f–459f  
regulation of, 457–458  
in SOS response, 710–711, 711f, 725  
in strand exchange, 456–457, 458f  
structure of, 456, 457f
- RecBCD, 454–456, 454t, 455f, 458
- Receptors  
cell surface, 686–687  
G-protein-coupled, 759–760  
insulin, 688b–689b  
steroid hormone. *See* Steroid hormone receptors
- Recessive traits, 25, 26f
- RecFOR, 454t, 455–456, 456f, 458
- RecJ, in mismatch repair, 425t, 427f
- Recognition helix, 679, 679f
- Recognition sequences, 217–218, 219t
- Recombinant DNA, 217
- Recombinant DNA technology. *See* Biotechnology
- Recombinases, 447  
in gap repair, 452–453  
in meiotic recombination, 467–468  
in recombinational repair, in bacteria, 446, 454–459, 456f–459f  
in site-specific recombination, 482, 483–485, 487–489
- Recombination, 40–41, 41f, 42f  
errors in, mutations from, 414–416
- homologous  
in bacteria, 447–464  
branch migration in, 450, 451f, 456, 458f, 459  
crossovers and, 446, 449, 465, 465f  
definition of, 446  
DNA circularization and, 487  
in DNA repair. *See* Recombinational DNA repair  
in eukaryotes, 447–464  
meiotic, 464–469, 464f–469f. *See also* Meiotic recombination  
mitotic, 465, 469–470, 470f  
initiation of, 445  
intron movement and, 472, 472f  
in mating-type switch, 470–471, 471f  
nonrepair functions of, 446–447  
strand exchange in, 456–457, 458f–460f, 459  
strand invasion in, 446, 447–450, 449f, 451f, 456, 458f, 459f  
3' single-stranded DNA tails in, 445, 447–448, 449f
- site-specific, 482–489  
auxiliary proteins in, 488–489  
in bacteriophages, 485–487, 509  
in biotechnology, 488–489, 489f, 490b–491b  
in brainbow method, 490b–491b  
copy-number amplification and, 485  
Cre-lox system in, 487, 487f, 488–489  
DNA circularization and, 487, 487f  
DNA sequences in, 482–483, 483f  
Flp system in, 485, 488–489  
in gene expression, 487–488  
Hin-hix system in, 485, 488–489, 488f  
reactions in, 482–485, 484f, 488  
recombinases in, 482, 483–485, 487–489  
replication and, 485, 486f  
sites of, 483–485, 483f, 487  
in viral infections, 485–487, 487f  
in yeast, 485, 486f  
transposition and, 482
- Recombination genes, discovery of, 476
- Recombination mapping, 42–44, 42f, 43f
- Recombinational DNA repair, 423, 435, 446–464
- in bacteria, 453–463  
enzymes in, 453–460, 454t  
RecA protein in, 454, 454t, 455–458, 456f–459f  
RecBCD in, 454–456, 454t, 455f  
branch migration in, 450, 451f, 456, 458f–460f, 459  
chi sequences in, 454–455, 455f  
crossovers in, 446, 449  
defective, in cancer, 446  
in *Deinococcus radiodurans*, 462b–463b  
dimeric chromosomes and, 460–461, 461f

- Recombinational DNA repair (*continued*)
- double-strand break
    - in bacteria, 448–450, 449f, 450f
    - in eukaryotes, 468, 469–470, 470f
  - enzymes in, 447, 454t
  - in eukaryotes, 463–472
    - evolution of, 446
    - fork regression in, 451, 452f, 457
    - gap repair in, 452–453, 453f, 456
    - Holliday intermediates in. *See* Holliday intermediate(s)
    - initiation of, 454–455
    - in mitotic recombination, 469–470, 470f
    - overview of, 446
    - recombinases in, 447, 454t
    - steps in, 447–449, 449f
    - strand invasion in, 446, 447–450, 449f, 450f, 451f, 456, 458f, 459f
    - synthesis-dependent strand annealing in
      - in bacteria, 447–448, 449f, 450f
      - in eukaryotes, 468, 469–470, 470f, 471, 471f
    - $3'$  single-stranded DNA tails in, 445, 447–448, 449f
    - undamaged double-stranded DNA in, 447
  - RecX, 457
  - Red blood cells, sickled, 53b
  - Refinement, in X-ray crystallography, 123–124, 124f
  - Reflection spot, 122
  - Regulated gene expression, 552, 669.
    - See also* Gene regulation
  - Regulators, transcriptional. *See* Transcription factors
  - Regulatory enzymes, 161
  - Regulatory sequences, Pol II, 534–535
  - Regulatory sites, 669
  - Regulons, 675, 707
  - RelA protein, 719
  - Relative molecular mass, 54
  - Relaxed DNA, 305, 307f
  - Release factors, 643, 643f
  - Replica plating, 476
  - Replication. *See* DNA replication
  - Replication bubble, 366f, 367
  - Replication complex, 390t, 395
  - Replication enzymology, 403
  - Replication factor C (RFC), 389, 389f
  - Replication fork, 366f, 367, 367f, 377–390
    - in bacteria, 377–387, 391–393, 392f
      - advancement of, 380–384, 382f
      - assembly of, 391–393, 392f
    - $\beta$  sliding clamp and, 379–380, 379f, 380f, 385–387
    - collapsed, 422–423, 423f, 447, 448f
      - repair of, 449–450, 449f, 450f, 451f
    - collision release and, 385
    - collisions with RNA polymerase, 396–397
    - DNA loops and, 384–386, 385f
    - in eukaryotes, 387–389, 393–394, 394f
    - Okazaki fragments and, 384–387
    - proteins of, 377–384, 377t. *See also* DNA polymerase III (Pol III)
    - repair of, 446. *See also* Recombinational DNA repair
    - replisome and, 384, 384f
    - signaling release and, 385
    - stalled, 447, 447f
      - repair of, 447, 448f, 450–451, 450f, 452f
      - trombone model of, 384, 385f
    - Replication fork regression, 451, 452f, 457
    - Replication origin, 221, 298, 367–368, 367f, 368f
      - in bacteria, 390–392
        - activation of, 391–392, 392f
        - discovery of, 391, 405
        - inactivation of, 393
        - open complex and, 392, 392f, 393, 393f
        - prepriming complex and, 392, 392f
        - structure of, 391, 391f
      - in eukaryotes, 393–395, 405
      - origin replication complex and, 390t, 395
      - two-dimensional gel analysis of, 396b–397b
    - Replication protein A, 140
      - dissociation constant for, 138t
    - Replication proteins, initiator vs. replication, 390–391, 391f
    - Replicative transposition, 492f, 493, 495f
    - Replicons, 390
    - Replisomes, 384, 384f
      - DNA repair in, 435–438, 437f, 437t
    - Reporter gene assays, 744b–745b
    - Reporter genes, 242–243, 244f
    - Repression, gene, 669
    - Repressors
      - hemin-controlled, 774
      - transcriptional, 669–670, 670f. *See also* Transcription factors
        - acting with activators, 673
        - activators as, 742
        - AraC, 707, 707f
        - corepressors and, 671, 672f, 709
        - effectors and, 674–675, 674f
          - in eukaryotes, 740, 740f
          - Gal, 708, 708f
          - hormone receptors as, 742
          - Lex A, 710–711, 711f
          - Trp, 708f, 709
        - translational, 716–719, 773
          - in alternative splicing, 769
          - HCR, 774
          - $\lambda$  phage as, 722–725, 723f
      - Residues. *See* Amino acid residues
      - Resonance, 71, 72f, 102f
      - Resonance hybrids, 71
      - Restriction endonucleases
        - in cloning, 217–219
        - DNA cleavage by, 217–219, 217t, 219t
        - functions of, 218
        - recognition sequences for, 219t
        - type II, 217t, 218
      - Restriction sites, 217–218
      - Reticulocytes, translation regulation in, 774
      - Retinoid X receptor, 757
      - Retrohoming, 564b
      - Retrotransposable elements. *See* Retrotransposons
      - Retrotransposition, 270, 493–495, 496f
        - Ty elements in, 498, 498f
      - Retrotransposons, 270, 493–495, 496f, 564b. *See also* Transposon(s)
        - in eukaryotes, 408–500
        - evolution, 503–506
        - extrachromosomally primed, 495, 496f
        - LTR, 493, 495, 496f, 497, 498
        - non-LTR, 493, 495, 496f, 497, 499
        - retroviruses and, 500–503
        - target-primed, 495
      - Retroviruses, 166b, 500–503
        - definition of, 500
        - evolution of, 7b, 503–506
        - genome of, 300, 301t, 500, 502f
        - HIV as, 166b, 501–503
        - infection mechanisms of, 500, 501f
        - proteolytic cleavage in, 168
      - Rev, HIV and, 772, 773f
      - Rev1, in translesion synthesis, 437t
      - Reverse transcriptase, 153b, 494, 495, 500
        - in biotechnology, 501
        - discovery of, 500
        - in DNA synthesis, 500–501
        - LTR transposons and, 498
        - mutations and, 501, 502
        - non-LTR transposons and, 500
        - in recombinant DNA technology, 217t
      - Reverse transcriptase inhibitors, 502b
      - Reverse transcriptase polymerase chain reaction (RT-PCR), 228
      - Reverse transcription, 48, 48f
      - Reverse turns, 106, 106f
      - Reversible enzyme inhibitors, 152b–153b
      - Reversion mutations, 420, 421f
      - RF-1, 643, 643f
      - RF-2, 643, 643f, 644f
      - RF-3, 643, 643f
      - RFC
        - in mismatch repair, 425t
        - in nucleotide excision repair, 433t, 434, 435f
      - Rho factor, in transcription, 532
      - Ribbon diagrams, 107, 108f
      - Riboflavin, 182
      - Ribonuclease, 178
        - renaturation of, 115
      - Ribonucleic acid. *See* RNA
      - Ribonucleoproteins (RNPs), 197
        - small nuclear. *See* snRNPs (small nuclear ribonucleoproteins)
        - small nucleolar, 517b, 575
      - Ribonucleoside 2',3'-cyclic monophosphates, 178
      - Ribonucleoside 2'-monophosphates, 178
      - Ribonucleoside 3'-monophosphates, 178
      - Ribonucleoside 5'-triphosphates (rNTPs), 518, 518f
      - Ribonucleotides, 62–64, 63f, 178, 179f. *See also* Nucleotide(s)

- Ribose, 62, 64  
 borate stabilization of, 585  
 methylation of, 65–66
- Ribosomal proteins, 617, 617t, 618, 621  
 targeting of, 657
- Ribosomal RNA. *See* rRNA  
 (ribosomal RNA)
- Ribosome(s), 7, 8f, 48, 616–624  
 abundance of, 616  
 all-RNA, 621  
 channels in, 624, 624f  
 chloroplast, 619  
 crystallization of, 175  
 definition of, 616  
 evolution of, 604b, 621, 622b  
 internal entry site for, 635–637, 637f  
 mitochondrial, 619, 621, 622b  
 multiple, 620, 620f  
 overview of, 616  
 proofreading by, 652–653, 662  
 recycling of, 644, 645f, 651  
 as ribozymes, 615, 617, 620–621,  
     621f, 661  
 rRNA as functional core of, 615, 617,  
     620–621, 661  
 stalled, tmRNA rescue of, 647–651, 651f  
 structure of, 175, 618f  
     in bacteria, 616–619, 618f  
     in eukaryotes, 619  
 subunits of, 112–113  
     association/dissociation of, 619–620,  
         619f  
     30S, 616–617, 617t, 629  
     40S, 616–617, 617t, 631  
     50S, 8f, 616–617, 617t  
     60S, 616–617, 617t  
     in translation, 48, 588  
         in elongation, 623, 638–641, 639f  
         in initiation, 619–620
- Ribosome recycling factor, 644, 645f
- Ribosome-binding sites, 621–623, 623f, 629
- Riboswitch(es), 604b, 712–716  
 classes of, 712, 713t  
 definition of, 712  
 downstream effects of, 712, 714  
 functions of, 713t, 716  
*glmS*, 713t, 714–716, 717f  
 limitations of, 716  
 mechanism of action of, 712–716,  
     713f, 715f  
 specificity of, 714, 717f  
 T-box, 713t, 718b  
 thiamine pyrophosphate-binding, 714,  
     715f, 716f  
 TPP-binding, 714, 715f, 716f  
 transcription, 712, 713f  
 translation, 712, 713f
- Riboswitch RNA, 712
- Ribothymidine, 64t
- Ribozymes (catalytic RNA), 6–7, 15, 88, 548,  
     554, 561, 576–577, 615, 620–621,  
     621f, 661  
 base pairing and, 576, 577  
 definition of, 576  
 discovery of, 15, 559  
 diversity of, 576–577  
 evolution of, 577b, 601–602,  
     604b–605b  
 functions of, 576–577, 581  
 hammerhead, 577  
 inactivation of, 577  
 ribosomes as, 615, 617, 620–621,  
     621f, 661  
 RNA world and, 6–8, 15, 20, 88, 482, 548,  
     554, 578, 601–602  
 self-splicing introns as, 559–563, 576  
 structure of, 581  
 viral, 577b
- Rich, Alexander, 194, 196f
- Ricin, 647, 647b, 650t
- Rifampicin, 521–522
- Rio, Don, 556, 556f
- Ritonavir (Norvir), 166b–167b
- Rits complex, 754b
- RNA. *See also* Nucleotide(s)  
 acetylation of, 206  
 A-form, helix of, 197  
 amplification of, 228  
 annealing of, 200–201, 200f  
 automated synthesis of, 206, 207f  
 backbone of, 179–180, 181f  
 base pairs in, 196–197, 197f, 198f. *See also*  
     Base(s); Base pairs/base pairing  
 base stacking in, 197–198  
 base substitution in, 66, 67f  
 bonds in, weak, 76–77  
 catalytic. *See* Ribozymes (catalytic RNA)  
 chemical modifications of, 206  
 chirality of, 79–80, 80f  
 circularization of, 634, 637f  
 denaturation of, 200–201, 200f  
 early studies of, 44–46, 48, 176–177  
 folding of, 196–198, 196f  
 functional, 49–50  
 functions of, 6–8, 48, 176  
 as genome and enzyme, 576–578  
 genomic, 195, 298, 300  
 HIV-1, 199b  
 hydrolysis of, 179–180, 182f  
 long, 517b  
 long noncoding, 517b  
 messenger. *See* mRNA (messenger RNA)  
 methylation of, 206, 573  
 micro. *See* MicroRNA (miRNA)  
 modified/unusual bases in, 580  
 as motor protein ligand, 157–160  
 nuclear export of, 569–571, 569f  
 nuclear import of, 569, 569f  
 nucleotides in, 45, 62–64, 63f  
 origins of life and, 5–8  
 pentose rings in, 177, 177f,  
     178, 178f  
 polarity of, 179  
 postsynthetic changes in,  
     65–66, 67f  
 posttranscriptional modification of, 206  
 pre-miRNA, 575–576, 575f, 784  
 pre-mRNA, 784  
     editing of, 565–568, 567f, 568f  
     processing of, 549–553. *See also* RNA  
         processing  
 splicing of, 554–565, 652, 653f. *See also*  
     Splicing  
 preribosomal, 532–534  
     processing of, 573–575, 575f  
     transcription of, 534, 534f, 534t  
 preribosomal RNA, 532–534,  
     573–575, 575f  
 renaturation of, 200–201, 200f  
 resonance in, 71, 72f  
 ribosomal. *See* rRNA (ribosomal RNA)  
 riboswitch, 712  
 self-replicating. *See* Ribozymes (catalytic  
     RNA)  
 SL, 565  
 small, 517b. *See also* MicroRNA (miRNA)  
 small interfering. *See* siRNA (small  
     interfering RNA)  
 small nuclear, 517, 517b,  
     557, 559f  
     transport of, 569  
 small nucleolar, 517b  
 small temporal, 685, 685f  
 SsrA  
     in protein tagging, 651, 651f  
     in ribosome rescue, 650–651, 651f  
 stable noncoding, 517b  
 structure of, 65, 66f, 175, 176–177,  
     194–200, 211  
     in gene expression, 199b  
     helices in, 196–197, 197f  
     secondary structures and,  
         196–198, 196f  
     stabilizing forces in, 197–198, 198f,  
         576, 580  
     variations in, 196–198, 196f  
 synthesis of. *See* Transcription  
 10Sa, in ribosome rescue, 647–651, 651f  
 thermal properties of, 200–201  
 transfer. *See* tRNA  
 translocation and, 158, 158f  
 transposon, 505  
 unwinding of, 158–160, 200–201, 200f  
 weak interactions in, 197–198, 198f  
 Z-form, helix of, 197  
 RNA amplification, 228  
 RNA degradation, 568, 570f–572f,  
     571–572  
 in bacteria, 571  
 in eukaryotes, 571–572, 571f  
 nonsense-mediated decay in, 571,  
     652, 652f  
 non-stop decay in, 571, 652, 652f  
 in processing bodies, 571–572, 572f  
 RNA duplexes  
     denaturation of, 201  
     hybrid, 201, 202  
 RNA editing, 504, 565–568, 567f, 568f  
 RNA hairpins, 197, 529, 531–532, 531f

- RNA interference (RNAi), 504, 685, 782–789  
 base pairing in, 783–786  
 definition of, 783  
 experimental uses of, 787, 790f  
 functions of, 783  
 in nematodes, 786  
 siRNA in, 685, 685f, 783, 784–786  
 in viruses, 786–787, 787f, 788b–789b  
 in yeast, 784  
 RNA ligase, 152  
 RNA polymerase(s), 49, 377t, 516–522  
   architecture of, 519, 519f  
   in bacteria, 523–532  
     channels for, 526, 528f  
     inhibitors of, 521–522, 522f, 523f  
     processivity of, 521, 529  
     in promoter binding, 525–526, 525t, 527f  
     in proofreading, 529, 530f  
     sigma factors of, 519, 519f, 523–527, 524f, 525t. *See also* Sigma factors  
     structure of, 519, 519f  
     subunits of, 519, 519f, 519t, 533  
     in transcription elongation, 529, 530f  
     in transcription initiation, 527–528  
     in transcription termination, 531–532, 531f  
       vs. in eukaryotes, 532  
     definition of, 516  
     early studies of, 516–518, 518f  
     in eukaryotes, 519, 519f, 532–539  
       overview of, 516–522  
       *Pol I.* *See* RNA polymerase I (Pol I)  
       *Pol II.* *See* RNA polymerase II (Pol II)  
       *Pol III.* *See* RNA polymerase III (Pol III)  
     intrinsic speeds of, 542  
     products of, 534t  
     promoters for, 532–535, 669–670, 670f, 736–739. *See also* Promoter(s)  
     in replication  
       in initiation, 394  
       in termination, 396–397  
     structure of, 519, 519f  
     substrates for, 518  
     subunits of, 519, 519f, 519t  
     TATA-binding protein and, 534, 534f, 534t, 535  
     in transcription, 516–521  
       in elongation, 529, 536, 537–538  
       in initiation, 527–528, 534–535, 538–539, 669–678  
       in termination, 531–532, 536, 539  
     vs. in bacteria, 532  
 RNA polymerase core, 519, 519f  
 RNA polymerase holoenzymes, 519, 519f  
   sigma factor of, 519, 519f, 523–527, 524f, 525t. *See also* Sigma factors  
 RNA polymerase I (Pol I), 519, 519t  
   promoters for, 532–534, 534f, 534t  
   in transcription termination, 539  
 RNA polymerase II (Pol II), 519, 519f, 519t, 534–540, 534t, 535f–539f  
   C-terminal domain of, 536, 537  
 Mediator complex and, 538–539, 538f  
 preinitiation complex and, 535–536, 536f  
 promoters for, 534–535, 534t, 535f, 541, 736–739, 737f  
 in RNA processing, 551–553  
 structure of, 535, 536f  
 subunits of, 535, 536f  
 in transcription  
   in elongation, 536, 537–538, 537f  
   in initiation, 534–535, 535f, 538–539  
   in termination, 536, 536f  
 RNA polymerase III (Pol III), 519, 519t  
   promoters for, 534t, 535, 535f  
   in termination, 539  
 RNA primases, 382f, 383  
   in replisome, 382f, 383  
 RNA primers, 368, 369, 382f, 392  
   in replication initiation, 398  
   in replication termination, 398  
 RNA processing, 65, 67f, 547–581  
   in bacteria, vs. in eukaryotes, 548, 548f  
   capping in, 548, 549–550, 550f, 552–553, 631. *See also* 5' cap  
   coupled reactions in, 549  
   degradative, 568, 570f–572f, 571–572  
   enzymes in, 549  
   in eukaryotes, 548–553  
     vs. in bacteria, 548, 548f  
     evolution of, 552b–553b  
   gene regulation in, 668, 668f, 684–692  
   of mRNA, 549–572  
   of non-protein-coding RNAs, 572–576  
   overview of, 548–549  
   *Pol II* in, 551–553  
     *Poly(A)* tail in, 550–553. *See also* *Poly(A)* tail  
     polyadenylation in, 549, 550–553, 551f  
   primary transcripts and, 548–549, 784, 785f  
   RNA editing in, 565–568, 567f, 568f  
   RNA transport in, 568–571, 569f, 570f  
   of rRNA, 573–575, 574f, 575f  
   splicing in, 551–565. *See also* Splicing  
   transcriptional regulation of, 552–553, 554f  
     *of tRNA*, 572–573, 573f  
 RNA recognition motif, 558  
 RNA secondary structures, 196–198, 196f  
 RNA silencing, 753, 754b  
 RNA splicing. *See* Splicing  
 RNA transport, 568–571  
   mRNA localization in, 568, 570–571, 570f–572f  
   of non-protein-coding RNAs, 569  
   of rRNA, 569  
   splicing and, 569–570  
   *of tRNA*, 569  
 RNA world hypothesis, 6–8, 8, 15, 88, 548, 554, 578, 601–602. *See also* Ribozymes (catalytic RNA)  
 evidence for, 578  
 ribose stability and, 585  
 RNA-coding genes, 271  
   accelerated evolution of, 274  
 RNA-containing vesicles, 1, 4f, 6, 20  
 RNA-DNA hybrid duplexes, 201–202  
 RNAi. *See* RNA interference (RNAi)  
 RNA-induced silencing complex (RISC), 576, 784  
 RNase P, 576  
 RNaseH, 377t, 383, 494  
 RNA-Seq, 278  
 Roberts, Richard, 554, 555f, 579  
 Roderick, Thomas H., 260  
 Roeder, Robert, 537, 537f  
 Rolling-circle replication, 485, 486f  
 Rosetta stone fusions, 282  
 Rosettes, 341, 342f  
 Rossmann fold, 111, 112f, 130, 130f, 182  
 Rothstein, Rodney, 468  
 RPA  
   in mismatch repair, 425t  
   in nucleotide excision repair, 433t, 434, 435f  
 R-proteins, 617, 617t, 618, 621  
   synthesis of, rRNA availability and, 716–720, 718b, 719f  
   targeting of, 657  
   in translation repression, 718–719, 719f  
 rRNA (ribosomal RNA), 7, 48, 617, 617t  
   in archaea, 618f  
   in bacteria, 617, 617t, 618–619, 618f  
   in chloroplasts, 619  
   definition of, 516  
   in eukaryotes, 618f, 619  
   evolution of, 615, 617, 618f, 621, 622b  
   functional inactivation of, 615, 617  
   functions of, 516, 618–619  
   in mitochondria, 619, 621, 622b  
   modified/unusual bases in, 580  
   nomenclature for, 574  
   preribosomal, 573–575, 575f  
   processing of, 573–575, 574f, 575f. *See also* RNA processing  
   structure of, 211  
   synthesis of  
     *r-protein synthesis* and, 716–720, 718b, 719f  
     stringent response in, 719–720, 720f  
   transcription of, 534, 534f  
   transport of, 569  
 Rut site, 531f, 532  
 RuvA, 159, 160f  
 RuvAB, 458–460  
 RuvB, 159, 160f  
 RuvC, 460  
 S1 mapping, 693  
 S phase  
   in meiosis, 36  
   in mitosis, 33, 34f  
*Saccharomyces cerevisiae*  
   centrosomes of, 299  
   in cloning, 236–237, 236f  
   gene silencing in, 754b, 784

- genetic code alterations in, 603, 603t, 606  
 genome of, 302t  
   sequencing of, 261  
 introns in, 555, 780b  
 mating-type switch in, 470–472, 470f, 471f  
 as model organism, 394–395, 395f, A-3-A-4, A-8-A-9  
 replication in, 394–395, 395f  
 site-specific recombination in, 485, 486f  
   Ty elements in, 498, 498f  
 Sachs, Alan, 768  
 S-adenosylmethionine (adoMet), 182, 205  
 Sae2, 467, 467f  
*Salmonella typhimurium*  
   in Ames test, 420, 421f  
   phase variation in, 487–488, 488f  
 Salt bridges, 68–70, 70f  
 Sanes, Joshua, 490b  
 Sanger method, 171, 228–233, 229f  
 Santoso, Steve, 175  
 Saquinavir (Invirase), 166b–167b  
 Sarnow, Peter, 788b–789b  
 SARS virus, discovery of, 259  
 Scaffold, chromosomal, 341, 341f  
 Scaffold proteins, in DNA underwinding, 308, 308f  
 Scanning, in translation, 635  
*Schizosaccharomyces pombe*, gene silencing in, 754b, 784  
 Schultz, Peter, 630  
 Schwann, Theodore, 32  
 Scientific community, 16  
 Scientific literature, 13, 16  
 Scientific method, 12–16, 17f  
   accidental discoveries and, 15  
   classical version of, 14  
   conceptual basis of, 12–13  
   context for, 14–15  
   deductive/inductive reasoning and, 13  
   exploration and observation and, 15  
   flow chart of, 17f  
   historical perspective and, 13–14, 13f  
   hypotheses and, 12–13  
     deduction and, 15  
     testing of, 14–15  
   inspiration and, 15  
   model building and calculation and, 15  
   objectivity and, 12  
   theories and, 13  
   variations on, 14–16  
 Scientific theories, 13  
 Scientific training, 16  
 SCOP database, 114  
 Screenable markers, 221  
 SDSA pathway, in recombinational repair  
   in bacteria, 447–448, 449f, 450f  
   in eukaryotes, 469–470, 470f, 471, 471f  
 SDS-PAGE, 101, 101f  
 SECIS element, 605–606, 606f  
 Second law of thermodynamics, 86–87  
 Second messengers, 182–183, 720–721.  
   See also Signaling  
   cAMP in, 759f, 760  
   ppGpp as, 720  
 Secondary protein structure, 103–107, 108f  
    $\alpha$  helix in, 104–105, 104f, 107, 108f  
    $\beta$  sheet in, 105–106, 105f, 107, 108f  
   reverse turns in, 106, 106f, 107, 108f  
   ribbon diagrams for, 107, 108f  
 Segal, Eran, 331  
 Segment polarity genes, 793, 795–796, 796f, 797f  
 Segmentation, in development, 791, 793–795, 794f  
 Segmentation genes, 793  
 Segregation  
   allelic, 25–28, 28f, 56  
   chromosomal, 34–37, 35f, 56, 299, 313, 464–465, 464f, 465f  
 SelA, 605  
 SelB, 605  
 Selectable markers, 221  
 Selenocysteine, 604–606, 606f  
 Selfish DNA, 482, 505  
 Self-replicating polymers, 6, 8, 20, 482  
 Self-splicing, 554, 559–563, 561f–565f  
   in bacteria, 562–563  
 Semiconservative replication, 364–367, 365f.  
   See also DNA replication  
 Semidiscontinuous replication, 368–369  
 SeqA, 377t, 393, 393f  
 Sequence analysis, of viruses, 129, 129f  
 Sequence polymorphisms, 230b  
 Sequence tagged sites (STSs), 261  
 Serine, 97, 98f, 99t  
   phosphorylation of, 67f  
 Serine-class recombinases, in site-specific recombination, 484, 485  
 7SL rRNA, transcription of, 534t, 535, 535f  
 70S ribosome. See Ribosome(s)  
 Sex cells, 24  
 Sex chromosomes, 36–37  
   X, 37, 37f  
     inactivation of, 347–348, 350b, 351b  
   Y, 37, 37f  
 Sex determination, 36–37, 37f  
   alternative splicing in, 769–771, 770f  
   XO, 37  
   XY, 37  
*Sex lethal (Sxl)* gene, 769–771, 770f  
 SF2 enzymes, 160  
 Sgs1, 445, 467, 467f  
 Shapiro, James, 510  
 Sharp, Phillip, 554, 555f, 579  
 Shatkin, Aaron, 549  
 Shimomura, Osamu, 253  
 Shine, John, 629  
 Shine-Dalgarno sequences, 629, 633f, 714  
 Short interfering RNA. See siRNA (small interfering RNA)  
 Short interspersed nuclear elements (SINEs), 498, 500  
 Short tandem repeats, in DNA  
   fingerprinting, 230b–231b  
 Shotgun sequencing, 262  
 Shuttle vectors, 223  
 Sickle-cell anemia, 52b–53b, 412  
   malaria and, 54–55  
 Side chains  
   amino acid, 97–98, 97f, 98f  
   aromatic, 98, 98f  
   carbohydrate, posttranslational attachment of, 654  
 Sigley, Paul, 536, 537f  
 Sigma factors, 519, 519f, 523–526, 524f, 525t, 698  
   in closed-to-open complex conversion, 526–527, 528f  
   loss of in elongation, 528  
   in polymerase binding, 525–526, 527f  
   types of, 525, 525t  
 Sign inversion model, 325–326  
 Signal integration, 673, 674f, 676  
 Signal recognition particles, 655–657  
 Signal sequence, 655  
 Signaling  
   cAMP in, 759f, 760  
   cell surface receptors in, 686–687  
   in development, 790–793, 795–796, 796f  
   G proteins in, 759–760, 759f  
   in gene regulation, 673–675, 686–687, 690–692, 691f  
   in eukaryotes, 756–760  
   JAK-STAT pathway in, 687, 687f  
   phosphorylation in, 686–687, 687f, 759f, 760  
   in protein degradation, 690–692, 691f  
   in protein targeting, 655–659, 656f  
   quorum sensing and, 697, 711, 729  
   second messengers in, 182–183, 720–721, 759f, 760  
   cAMP in, 759f, 760  
   ppGpp as, 720  
   steroid hormones in, 742, 756–758, 757f–759f  
   Wnt-class pathways in, 795–796, 796f, 802  
 Signaling release, 385  
 Silent mutations, 50, 411, 441, 592  
 Simple-sequence repeats (SSRs), 271  
 SINEs, 498, 500  
 Singer, Maxine, 599, 599f  
 Single bonds, 70–71, 71f  
 Single nucleotide polymorphisms (SNPs), 271, 272, 272f  
 Single-gene disorders, 275–276, 276f  
 Single-strand breaks, 422–423, 423f, 425–435, 451  
   repair of, 422–423, 423f  
   replication fork collapse and, 447, 448f  
 Single-strand gap repair, 452–453, 453f  
 Single-stranded DNA (ssDNA), 377t, 382f, 384

- Single-stranded DNA-binding protein (SSB), 135, 139–141, 140f, 141f, 377t, 382f, 384
- dissociation constant for, 138t
- in mismatch repair, 425t, 426
- in recombinational repair, 455, 458
- Sinsheimer, Robert, 323
- sirNA (small interfering RNA), 517b, 575–576, 575f, 685, 685f, 785
- in gene knockdown, 787, 790f
- in gene silencing, 753, 754b, 784–786, 785f
- processing of, 575–576, 575f
- in RNA interference, 685, 685f, 783–784
- SsrA (10s), 650–651, 651f
- SisA activator, 769–771, 770f
- SisB activator, 769–771, 770f
- Sister chromatids, 33
- centromeres and, 33, 298–299, 299f
- cohesin-based linkage of, 317f, 318–319, 320f, 464–465
- formation of, 33, 34f
- in meiosis, 35f, 36
- in mitosis, 34, 34f, 35f
- segregation of, 35f, 36, 464–465, 464f, 465f
- Site-directed mutagenesis, 239–240, 239f
- Site-specific recombination, 482–489, 489f, 490b–491b. *See also* Recombination, site-specific
- 6-4 photoproduct, 422, 422f
- 16S rRNA, 573, 574, 575f
- 60S subunit, 616–617, 617t
- Sjostak, Joe, 1, 3, 6, 14, 20
- SL1 (human selectivity factor 1), 534
- SL RNA, 565
- sleeping beauty* element, 499b
- Sliding clamp
- β. *See* Beta sliding clamp
  - PCNA, 388–389, 389f, 390t
- Sm core domain, 557
- Sm proteins, 557–558, 579
- Small nuclear ribonucleoproteins (snRNPs). *See* snRNPs (small nuclear ribonucleoproteins)
- Small nuclear RNA (snRNA), 517, 517b, 557, 559f
- transport of, 569
- Small nucleolar ribonucleoproteins (snoRNPs), 517b, 575
- Small nucleolar RNA (snoRNA), 517b
- Small RNA (sRNA), 517b. *See also* MicroRNA (miRNA)
- Small temporal RNA (stRNA), 685, 685f. *See also* MicroRNA (miRNA)
- SMC proteins, 316–320, 317f, 320f
- in chromosome scaffold, 341
- Smithies, Oliver, 490b
- Snf proteins, 159–160
- snoRNPs (small nucleolar ribonucleoproteins), 517b, 575
- snRNA (small nuclear RNA), 517, 517b, 557, 559f
- transport of, 569
- snRNPs (small nuclear ribonucleoproteins), 557, 558, 559f
- discovery of, 579
- snRNA of, 557, 559f
- in spliceosomes, 557–558
- structure of, 557, 559f
- Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), 101, 101f
- Sodium ions, in RNA, 198
- Sodium montmorillonite, in prebiotic evolution, 20, 20f
- Solar radiation
- double-stranded breaks and, 429–430, 430f, 440, 460
  - mutations from, 421–423, 424f, 440, 460
  - in photoreversal, 429–430, 430f, 440
- Solenoid model, of 30 nm filament, 339, 340f
- Solenoidal supercoiling, 311, 311f
- Solutes, concentration of, notation for, 82
- Solutions
- aqueous, pH of, 81–82
  - buffered, 82–84
- Somatic cell nuclear transfer (SCNT), 533b
- Somatic cells, 24
- Somatic gene therapy, 300
- Sorcerer II*, 268b
- Sörenson, Sören, 82
- SOS response, 707, 710–711, 711f, 725
- Southern blotting, 203, 206
- SP1 (specificity protein 1)
- discovery of, 515
  - Pol II recruitment by, 541
- Speciation, allopatric, 288
- Spectroscopy
- circular dichroism, 131, 131f
  - correlation, 125, 126, 126f
  - nuclear Overhauser effect, 125, 126–127, 126f
- Spindle apparatus
- in meiosis, 36, 464
  - in mitosis, 34
  - developmental symmetry and, 790
- Spinocerebellar ataxia type 1, 414t
- Splice sites, 556
- Spliced leader, 565
- Spliceosomes, 554, 557–558, 559f
- assembly of, 557–558, 561f
  - imaging of, 547, 557
  - proteins of, 557–558
  - structure of, 557, 559f
  - in *trans*-splicing, 565
  - types of, 558
- Splicing, 551–565
- alternative, 555–556, 557f, 684
  - definition of, 769
  - in sex determination, 769–771, 770f
- base pairing in, 558, 561f
- branch points in, 557, 558f
- chromatin remodeling in, 733, 736, 738b–739b
- cis*, 565
- coordination with transcription, 733, 738b–739b
- definition of, 554
- gene regulation in, 684
- histone acetylation in, 733, 737
- nutrient availability and, 779, 780b
- overview of, 554
- poly(A) site choice in, 556, 557f
- RNA transport and, 569–570
- self-splicing and, 554, 559–563, 561f–565f
- splice sites in, 556
- selection of, 558, 684
- spliceosome in, 554, 557–558, 559f
- steps in, 558, 558f, 560f
- trans*, 563–565, 565f
- Spo11, 465–467, 467f
- discovery of, 477
- Spongiform encephalopathies, 118–119, 118f
- SR proteins, 557–558
- ssDNA (single-stranded DNA), 377t, 382f, 384. *See also* Single-stranded DNA-binding protein (SSB)
- SsrA RNA
- in protein tagging, 651, 651f
  - in ribosome rescue, 650–651, 651f
- Stable noncoding RNA (ncRNA), 517b
- Stahl, Franklin, 364–366, 468
- Standard Gibbs free energy change ( $\Delta G^\circ$ ), 87
- Start (initiation) codons, 591, 629
- variant, 604
- Steady-state kinetics, 149–150
- Steitz, Joan, 579
- Steitz, Tom, 481, 617
- Stem cells, 767, 798–880, 798f, 799f
- transcription factors and, 533b
- Stem-loop binding protein, 552b–553b
- Stem-loop structures. *See* Hairpins
- Step size, 158
- Stepping, 158
- Stereochemistry, 78–80
- Stereoisomers, 79
- Steroid hormone(s)
- evolution of, 273–274, 273f
  - heat shock proteins and, 687, 690f, 757, 757f
  - in signaling, 742, 756–758, 757f–759f
- Steroid hormone receptors, 687, 742–756–758, 756–758, 757f–759f
- as activators and repressors, 742
- evolution of, 273–274, 273f
  - structure of, 758, 758f
  - thyroid, 757–758, 757f
  - types of, 757–760, 757f, 758f
- Stevens, Nettie, 37, 37f
- Sticky ends, 218, 219, 220f
- Stone, Michael, 297

- Stop (termination) codons, 591, 620, 642–643, 643f, 651  
genetic code alterations and, 603–604  
mitochondrial, 603, 603t  
premature, 547, 652–653, 653f  
variant, 604–606, 606f
- Strained DNA, 307, 307f
- Strand exchange, RecA in, 456–457, 458f, 459f
- Strand invasion, in DNA repair, 446, 447–450, 449f, 451f
- Streptavidin, in immunofluorescence, 242
- Streptomycin, translation inhibition by, 646, 646f
- Stringent factor, 719
- Stringent response, 719–720, 720f  
ppGpp in, 719–720, 720f
- stRNA (small temporal RNA), 685, 685f
- Structural Classification of Proteins (SCOP) database, 114
- Structure-based protein design, 95, 116, 116f
- Sturtevant, Alfred, 41–42
- Substrates, 88, 146
- Subunits, cooperativity among, 136–137
- Sugar puckles, 188, 190f, 197
- Sulfur-carbon bonds, 89
- SUMO, 687
- Sumoylation, 686, 687, 736
- Sunlight  
double-stranded breaks and, 429–430, 430f, 440, 460  
mutations and, 421–423, 424f, 440, 460  
in photoreversal, 429–430, 430f, 440
- Supercoiling. *See* DNA, supercoiling of
- Superhelical density, 309
- Superoxide radicals, as mutagens, 418
- Supersecondary structures, 109–112, 110f, 111f
- Suppressor tRNA, 592–593, 592f, 593f
- Surroundings, 86
- Sutton, Walter, 36, 38, 56
- SV40 promoter, 541
- Svedberg, Theodor, 574
- Svedberg units, 574
- SWI/SNF (switch-sniff) complex, 344–345, 346t, 352, 741, 742f
- Sxl* gene, 769–771, 770f
- Symington, Lorraine, 445
- Synteny, 266, 266f
- Synthesis-dependent strand annealing (SDSA)  
in bacteria, 447–448, 449f, 450f  
in eukaryotes  
in meiotic recombination, 468  
in mitotic recombination, 469–470, 470f, 471, 471f
- Systems, thermodynamic, 86
- Systems biology, 277
- Szostak, Jack, 445, 468, 577b
- T cells  
in HIV infection, 501–502  
insulators in, 751–752, 751f
- Tabin, 802
- TAF3-TRF3 complexes, 763
- Tag single nucleotide polymorphisms, 271, 272f
- Tagging, 240–246, 240t, 241f, 266  
epitope, 242, 244–245, 244–246, 245f  
TAP, 246
- Tandem affinity purification (TAP) tags, 246
- Tandem repeats, 271
- TAP tags, 246
- Target site, in transposition, 489
- Target-primed retrotransposons, 495
- TATA box, 352, 534–535, 535f, 737, 740, 741  
TFIID binding of, 536, 537
- TATA-binding protein, 534, 536–537, 537f, 538, 740, 741
- TATA-binding protein-DNA complex, 536–537, 537f
- TATA-binding-associated factors, 534, 535, 536f
- Tatum, Edward, 36f, 46–47
- Tautomers, 180–181, 183f
- Taxa, 285
- T-box riboswitches, 713t, 718b
- Tc1/mariner transposons, 497–498, 499b, 505t
- Telomerase, 299, 398  
aging and, 399, 401  
cancer and, 399, 402  
cell immortality and, 399, 401  
end replication problem and, 398–400, 399f
- Telomere(s), 299–300, 299f  
shortening of, 399, 399f, 3398
- Telomere loops, 401, 401f
- Telomere repeats, 299, 299t
- Telomere-binding proteins, 399–401, 400f
- Telophase, 35f, 36  
in meiosis, 35f, 36  
in mitosis, 34, 35f
- Temin, Howard, 500, 500f
- Template, primed, 369, 370f
- Template strand, 48, 364, 369, 370f, 516, 519, 520f  
breaks in, replication fork stall/collapse and, 447, 448f
- 10Sa RNA, in ribosome rescue, 647–651, 651f
- Ter sites, 395, 398f
- Terminal transferase, 217t
- Termination codons. *See* Stop (termination) codons
- Termination factors, 643, 643f
- Termination sequences, 531
- Terminator, mRNA, 709–710
- Tertiary protein structure, 107–112  
determination of, 121–127  
by nuclear magnetic resonance, 125–127  
by X-ray crystallography, 121–125
- Testcrosses, 40
- Tetracyclines, translation inhibition by, 646, 646f
- Tetrads, 36  
crossing over in, 40–41, 41f
- Tetrahymena thermophila*, self-splicing  
in, 559, 561, 561f, 562f, 577, 581
- Tetranucleosomes, structure of, 339, 340f
- Tetraplex DNA, 192, 193f
- TFII transcription factors, 532, 534–540  
in elongation, 537–538  
in initiation, 536–537, 538  
in preinitiation complex, 535, 536f  
in termination, 539  
types and functions of, 536–538
- TFIIA, 537
- TFIIB, 537, 538
- TFIIB recognition element, 534, 535f
- TFIID, 352, 538, 740, 741  
TATA box binding by, 536, 537
- TFIIE, 537, 538
- TFIIF, 537, 538
- TFIIC, 537, 538  
in DNA repair, 539–540
- TFIII transcription factors, 535, 535f
- TH1 box (element), 714
- Theories, scientific, 13
- Thermodynamics, 86–87. *See also* Energy  
of coupled reactions, 89, 90b  
first law of, 86  
second law of, 86–87
- θ-form replication, 367
- Thiamine pyrophosphate-binding riboswitches, 714, 715f, 716f
- Thiogalactoside transacetylase, 699
- 30 nm filament, 339–341, 340f–342f  
solenoid model of, 339, 340f  
zigzag model of, 339–341, 340f
- 30S subunit, 616–617, 617t
- Thornton, Janet, 92, 92f
- 3' end processing  
evolution of, 552b–553b  
gene regulation in, 684–685  
in histones, 552b–553b  
poly(A) tail and, 539, 550–553, 631, 652–653  
stem-loop structures and, 552b–553b
- 3'→5' exonuclease, 371–372, 371f–373f, 373, 373t, 375
- 3' single-stranded DNA tails, in  
recombinational repair, 445, 447–448, 449–450, 449f, 451f, 454, 455f
- Three-domain theory of life, 211
- Three-factor crosses, 44
- Three-nucleotide insertion mutations, 413, 413f
- 3'UTR  
in iron homeostasis, 779, 781f  
in translation, 767, 774–776, 775f  
in gene networks, 774, 775f
- Threonine, 97, 98f, 99t
- Through-bond correlation signals, 126
- Through-space NOE signals, 126

- Thymine (T), 48, 62, 177, 177f, 178, 204. *See also* Base(s); Base pairs/base pairing  
nomenclature for, 64t
- Thyroid hormone receptor, 757–758, 757f
- Ti plasmid, 221
- Tjian, Robert, 515, 541, 763
- T-loops, 401, 401f
- TLS DNA polymerases, 436–438, 437f, 437t
- tRNA  
in protein tagging, 651, 651f  
in ribosome rescue, 650–651, 651f
- Tn5 transposon/transposase, 492–493, 494f, 497, 497f, 504
- Tombusvirus p19, RNA interference in, 787, 787f
- Topoisomerase(s), 24, 309, 482, 485  
bacterial, 312t, 313, 314f–316f  
definition of, 312  
discovery of, 324–326  
in DNA compaction, 312–316, 312f, 312t, 314f–317f  
in DNA supercoiling, 325–326  
in DNA untwisting, 382f, 383  
eukaryotic, 312t, 313–316, 316f, 317f  
structure of, 23  
type I, 312, 312f  
discovery of, 324  
type II  
discovery of, 312, 312f, 325–326  
in DNA supercoiling, 325–326  
in replication termination, 397–398, 398  
visualization of, 312–313, 312f
- Topoisomerase inhibitors, 318b
- Topoisomerase IV, 377t
- Topoisomers, 310, 312f
- Topology  
definition of, 306  
DNA, 306
- Topotecan (Hycamtin), 318b–319b
- Torpedo model, of transcription termination, 539, 539f
- Torsion angles, 102, 103f
- Totipotent stem cells, 798, 798f
- Toxins, translation inhibition by, 646, 647b, 648t–650t
- TP retrotransposons, 495
- TPP-binding riboswitches, 714, 715f, 716f
- tra* gene, 771
- Traits  
dominant, 25, 26f, 29–30, 30f  
codominance and, 30, 30f  
incomplete, 29–30, 30f  
recessive, 25, 26f  
wild-type, 38–49, 38f
- Trans isomers, 99–102, 102f
- Trans-acting genes/gene products, 701
- Transcription, 516–542  
accuracy of, 520–521, 529, 530f  
antitermination of, 539, 539f, 725, 725f  
in bacteria, 523–532  
abortive transcripts in, 527–528  
backtracking in, 529  
closed-to-open complex conversion in, 526–527, 528f  
consensus sequences in, 524  
elongation in, 529, 530f  
initiation of, 527–528, 529f  
lack of primer in, 527–528  
pauses in, 529  
promoters in, 523–526, 524f, 669–670, 670f, 698  
proofreading in, 529, 530f  
pyrophosphorylation in, 529  
rate of, 529  
sigma factors in, 523–526, 523–527, 524f, 525t, 528. *See also* Sigma factors  
termination of, 531–532, 531f  
upstream promoter elements in, 524  
base pairing in, 518, 518f, 616–617. *See also* Base pairs/base pairing  
notation for, 523  
chemical mechanism of, 518, 518f  
coding strand in, 519, 520, 520f  
definition of, 49, 516  
direction of, 518, 518f, 520  
discovery of, 48–49  
elongation in, 519, 520  
attenuation in, 684  
in bacteria, 529, 530  
in eukaryotes, 536, 536f, 537–538, 537f  
gene regulation in, 684  
in eukaryotes, 532–540  
coordination with RNA splicing, 733, 738b–739b  
in DNA repair, 539–540  
DNA unwinding in, 535, 536f  
elongation in, 536, 536f, 537–538, 537f  
enhancers in, 532  
initiation of, 535–537, 536f, 538–539, 538f  
Pol I in, 534, 534f, 534t  
Pol II in, 534–535, 534t, 535f–539f  
Pol III in, 534t, 535, 535f  
Mediator complex in, 538–539, 538f  
in mRNA processing, 540  
overview of, 532  
preinitiation complex in, 535, 536f  
promoters in, 534–535, 534t, 535f, 676–677, 677f. *See also* Promoter(s)  
rate of, 542  
RNA polymerases in, 532–539  
steps in, 535–536, 536f  
TATA-binding protein in, 534, 536–537, 537f, 538, 740, 741  
termination of, 536, 536f, 538, 539  
antitermination and, 539, 539f, 725, 725f  
exonucleases in, 539, 539f  
torpedo model of, 539, 539f  
transcription factors in, 532–540.  
*See also* Transcription factors  
Pol I, 534, 534f, 534t  
Pol II, 534–540, 534t, 535f–539f  
Pol III, 534t, 535, 535f  
inhibitors of, 521–522, 522f, 768–769, 769f  
initiation of, 519, 521  
in bacteria, 527–528, 529f  
in eukaryotes, 532–539, 536f, 736–739  
regulation of, 669–678. *See also* Gene regulation  
RNA polymerase in, 669–678  
in intergenic regions, 517  
mRNA in, 48–49, 49f  
nontemplate strand in, 519, 520, 520f  
overview of, 516, 521, 521f  
phases of, 519  
of pre-rRNA, 534, 534f  
processivity of, 521, 529  
products of, 534t  
promoters in, 517b, 519, 522. *See also* Promoter(s)  
in bacteria, 523–526, 524f, 669–670, 670f, 698  
in eukaryotes, 534–535, 534t, 535f, 676–677, 677f  
Lac, 524–525, 524f  
proofreading in, 520–521, 529, 530f.  
*See also* Proofreading  
rate of, 529, 542, 679, 709  
regulation of, 522  
by histones, 336–338, 337f, 351b  
nucleosomes in, 338  
repression of, heterochromatin in, 735–736, 735f  
reverse, 48, 48f  
rho factor in, 531f, 532  
RNA polymerases in, 49, 516–522. *See also* RNA polymerase(s)  
in bacteria, 523–532  
in eukaryotes, 532–539  
rut site in, 531f, 532  
selectivity in, 516  
steps in, 49f, 521, 521f  
template strand in, 48, 364, 369, 370f, 516, 519, 520f  
breaks in, replication fork stall/collapse and, 447, 448f  
notation for, 520  
termination of, 519  
in bacteria, 531–532, 531f  
in eukaryotes, 536, 536f, 538, 539  
antitermination and, 539, 539f, 725, 725f  
exonucleases in, 539, 539f  
torpedo model of, 539, 539f  
tRNA in, 49, 49f  
vs. replication, 516, 518  
vs. translation, 768–769  
Transcription attenuation, 709–710, 709f  
Transcription bubbles, 520, 520f  
Transcription extracts, 337  
Transcription factors, 532, 534–540, 669–678  
activation domains of, 682  
activator. *See* Activators  
architectural regulator, 671, 672f  
in bacteria, 523–532  
in cell reprogramming, 533b

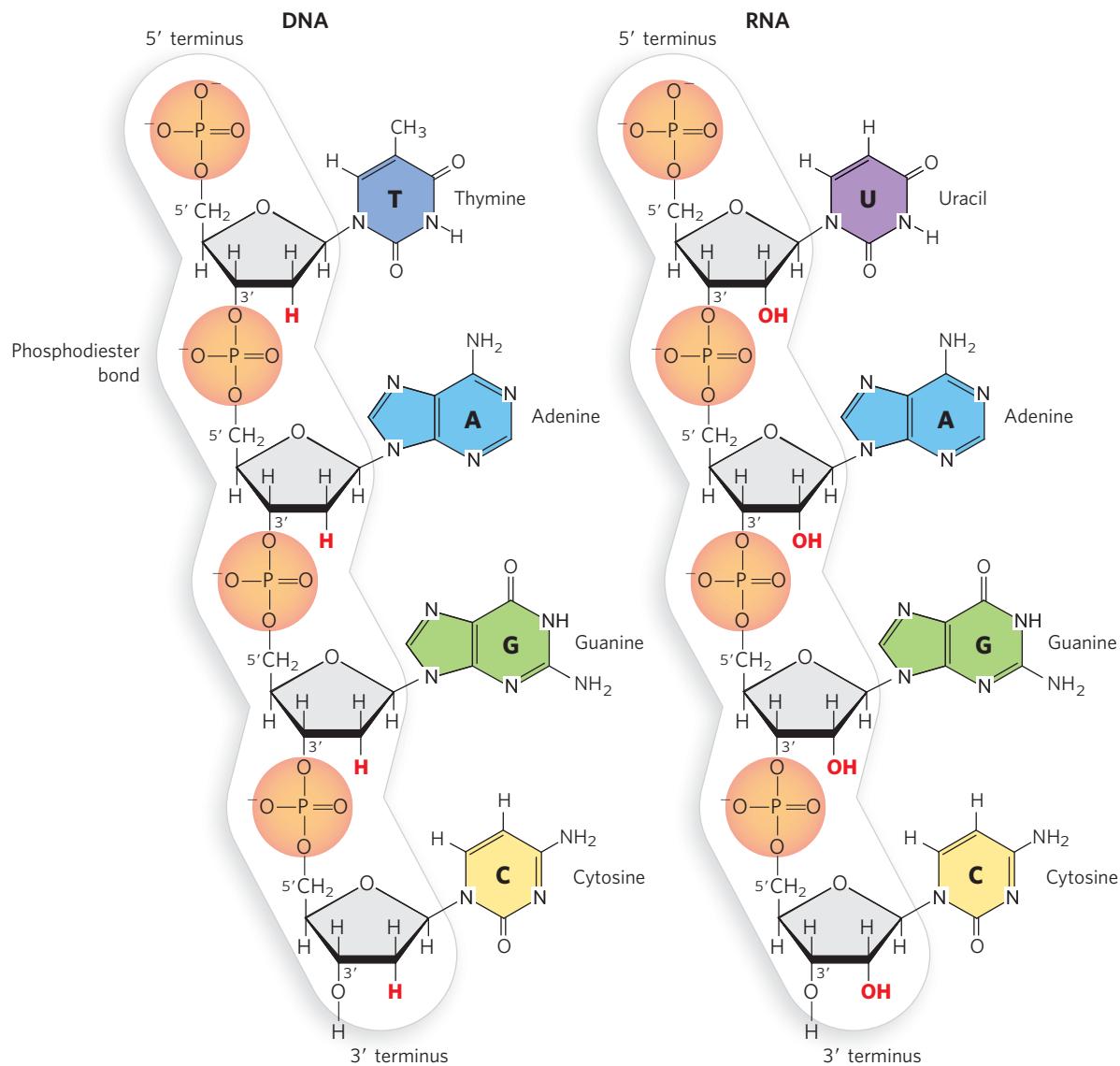
- corepressor, 671, 672f, 709  
 in *D. melanogaster*, 762  
 definition of, 669  
 discovery of, 515  
 DNA binding domains of, 679–684  
 in DNA looping, 670–673, 671f–673f  
 in eukaryotes, 532–540, 739–742  
   in combinatorial control, 748–750, 749f, 750f  
   general (basal), 739–742, 740f  
 in initiation, 536–537, 538  
 multiple, in eukaryotes, 737  
 nomenclature for, 535  
 number of, in eukaryotes vs. bacteria, 676–677, 677f, 737, 743  
 Pol I, 534, 534f, 534t  
 Pol II, 534–540, 534t, 535f–539f. *See also* RNA polymerase II (Pol II)  
 Pol III, 534t, 535, 535f  
 recruitment regions of, 682  
 repressor. *See* Repressors  
 signal integration, 673  
 structural motifs in, 678–682. *See also* Motifs  
   activation, 682  
   basic helix-loop-helix, 680–681, 681f  
   basic leucine zipper, 680, 680f  
   helix-turn-helix, 679–680, 680f, 682, 682f  
   homeodomain, 680, 680f  
   leucine zipper, 680, 682, 682f  
   zinc finger, 681–682, 682f  
 TFII, 532, 534–540  
 TFIII, 535, 535f  
 Transcription-activation motifs, 682  
 Transcriptional gene regulation, 668, 668f, 669–678. *See also* Transcription factors  
   in bacteria, 698–712  
   in eukaryotes, 733–766  
 Transcriptional ground state, 734  
 Transcription-coupled repair, 434, 435f, 539–540  
 Transcriptome, 277–278, 279f  
 Transcriptome analysis, 277–278, 278  
 Transcriptomes, 517  
 Transcriptomics, 277–278  
 Transcripts  
   primary, 548–549, 784, 785f. *See also* RNA processing  
   promoter-associated, 517b  
 Transduction, bacterial, 487  
 Transfection, 237–238  
 Transfer RNA. *See* tRNA (transfer RNA)  
 Transformation, in cloning, 221–222  
*transformer (tra)* gene, 771  
 Transgenic animals, 238–239  
   site-specific recombination and, 490b–491b  
*Transib* transposons, 505t, 507  
 Transition mutations, 411, 411f, 592  
 Transition state, 84, 148  
 Translation, 49, 50f, 615–662  
   accuracy of, 624, 625–629, 628f, 638, 644, 662  
   adaptor hypothesis and, 586–587, 586f, 608  
   adenylation in, 625, 626f  
   amino acid activation in, 625, 626f, 632t  
   amino acid attachment in, 625  
   amino acid recognition in, 625–626, 627f  
   aminoacylation in, 587–588, 621–629.  
    *See also* Aminoacyl-tRNA synthetases  
   ATP in, 625, 644  
   base pairing in, 587–588, 624, 629, 638  
   codon bias in, 606  
   in defective mRNA removal  
    in bacteria, 647–651, 651f  
    in eukaryotes, 651–653, 652f, 653f  
   definition of, 49, 588  
   direction of, 596, 609, 631  
   elongation in, 620, 621–624, 632t, 638–642  
   accommodation in, 638, 639f  
   aminoacyl-tRNA binding in, 638  
   in bacteria, 638–642  
   base pairing in, 638  
   codon-anticodon pairing in, 587–588, 588f, 638  
   cycles of, 640–641  
   in eukaryotes, 638–642  
   inhibitors of, 648t–650t  
   peptide bond formation in, 616, 621–624, 622b, 638–640, 639f, 644  
   peptidyl transferase reaction in, 639–640, 641f, 644  
   ribosome in, 621–624, 638–641, 639f  
   steps in, 638–640, 639f  
   substrate positioning in, 638–639  
   energetics of, 625, 640, 644  
   evolution of, 601–606, 604b–605b. *See also* Ribozymes (catalytic RNA)  
   5' cap in, 549–550, 551f, 631  
   folding in, 632t  
   genetic code and, 586–588. *See also* Genetic code  
   inhibitors of, 644–647, 647b, 648t–650t, 768–769, 769f  
   initiation complex in, 633–634, 635f, 638, 639f, 640f  
   initiation factors in, 629  
    in bacteria, 633–634  
    in eukaryotes, 634–635, 636f  
   initiation of, 629–638, 632t  
    in bacteria, 629, 631–634, 632t, 633–634, 633f, 634f, 635f  
    in eukaryotes, 631–632, 632t, 634–637, 636f, 637f, 773–774, 774f  
   5' cap in, 631  
   5' cap-independent, 635–637  
   5'UTR in, 774, 775f  
   inhibitors of, 648t  
   phosphorylation in, 773–774, 774f  
   PUF family in, 775–776, 776f  
   regulation of, 685–686, 686f  
   ribosomal subunit recruitment in, 629–631, 631f  
   ribosome in, 619–620  
   scanning in, 635  
   steps in, 629, 631f  
   3'UTR in, 774, 775f  
   in viruses, 635–637, 637f  
 internal ribosome entry site in, 635–637, 637f  
 methionine in, 631–633, 634f–636f  
 nick, 372, 372f, 383, 386  
   in base excision repair, 431  
   in mismatch repair, 426, 426f, 427–428  
 overview of, 619–620, 619f  
 posttranslational modification and, 632t, 654–658. *See also* Posttranslational protein modification  
 proofreading in, 627–629, 628f, 638, 644.  
   *See also* Proofreading  
 regulation in bacteria, 685–686  
 regulation in eukaryotes, 685–686, 772–778  
   in cytoplasm, 772–778  
   deadenylation in, 777, 779f  
   5'UTR in, 774, 775f  
   initiation factors in, 773–774, 774f  
   mRNA degradation rate in, 777, 779f  
   P bodies in, 777  
   phosphorylation in, 773–774, 774f  
   PUF family in, 775–776, 776f  
   repressors in, 773  
   in reticulocytes, 774  
   3'UTR in, 774–776, 775f  
   upstream open reading frames in, 777  
   vs. in transcription, 768–769  
 repressors in, 716–719, 773  
   in alternative splicing, 769  
   HCR, 774  
   λ phage as, 722–725, 723f  
 of ribosomal proteins, rRNA synthesis and, 716–720, 718b, 719f  
 ribosome in, 48, 588, 616–624, 629–631, 631f. *See also* Ribosome(s)  
   in elongation, 621–624, 623, 638–641, 639f  
   in initiation, 619–620  
   in peptide bond formation, 616, 621–624, 622b  
   recycling of, 644, 645f  
   in termination, 620  
 riboswitches in, 712, 713f. *See also* Riboswitch(es)  
 RNA circularization in, 634, 637f  
 selenocysteine in, 604–606, 606f  
 Shine-Dalgarno sequences in, 629, 633f, 714  
 steps in, 49, 50f, 619–620, 619f, 632t  
 termination of, 620, 632t, 642–647  
   in bacteria, 643, 643f  
   in eukaryotes, 643  
   release factors in, 643, 643f  
   steps in, 643, 643f  
 3'UTR in, 774–776, 775f  
 tmRNA salvage system and, 647–651, 651f

- Translation (*continued*)  
 translocation in, 640, 641, 642f  
 inhibitors of, 650t  
 tRNA-charging step in, 625  
 in viruses, initiation of, 635–637, 637f
- Translational gene regulation, 668, 668f  
 in bacteria, 685–686  
 in eukaryotes, 685–686, 686f, 772–778
- Translesion synthesis, 436–438, 437f, 437t, 441, 447, 452
- Translocases, 158, 159, 160f
- Translocation, 158, 158f–160f  
 definition of, 158  
 helicases in, 157–160, 158f, 159f, 381, 382f, 392–393, 392f  
 inhibitors of, 650t  
 in replication, 158–159, 381, 392–393.  
*See also* DNA, unwinding of  
 in translation, 640, 641, 642f  
 translocases in, 158, 159, 160f
- Translocation mutations, 415–416  
 fusion genes and, 415–416, 415f
- Transposable elements. *See* Transposon(s)
- Transposases, 492, 493, 494f
- Transposition, 482, 489–507  
 in bacteria  
   complex transposons in, 497  
   composite transposons in, 496–497  
   insertion sequences in, 496–497  
 cointegrates in, 493  
 cut-and-paste, 492–493, 492f–494f, 497–498  
 donor site in, 489  
 genomic rearrangements and, 505  
 immunoglobulins and, 505–507, 506f  
 intermediates in, 510–511  
 replicative, 492f, 493, 495f  
 retrotransposition and, 270, 493–495, 496f  
 target site in, 489  
 transposases in, 492, 493, 493f  
 Ty elements in, 498, 498f
- Transposon(s), 270–271, 482, 493–505  
 in bacteria, 496–497  
 bioengineered, 499b  
 complex, 497, 497f  
 composite, 496–497, 497f  
 cut-and-paste, 497–498  
   activation of, 499b  
   evolution of, 503–506, 504–506  
   functions of, 504–505  
   insertion sequences, 496–497, 497f  
 LINE, 498, 500  
 retrotransposons and, 493–495, 496f.  
*See also* Retrotransposons  
 silencing of, 504  
 SINE, 498, 500  
 superfamilies of, 504, 505t  
 Tc1/mariner, 497–498
- Transposon DNA, 482, 505
- Transposon RNA, 505
- Trans-splicing, 563–565, 565f
- Transversion mutations, 411, 411f
- TRAP system, 710, 728
- Tree of life, 8–9, 9f, 286f, 287, 287f
- TREX complex, 569
- Triplet code, 586, 588–589, 588f, 595–596.  
*See also* Genetic code
- Triplet expansion diseases, 413–414, 414t
- Triplex DNA, 191–192, 193f
- Trisomies, 466b
- tRNA (transfer RNA), 49, 587  
 adaptor function of, 587–588, 587f, 608  
 amino acid attachment to, 616  
 aminoacyl-tRNA and, 625–627, 627f  
 codons recognized by, 589–590  
 definition of, 516  
 discovery of, 49, 586–587  
 in DNA replication, 500–501, 501f  
 evolution of, 602, 604b–605b  
 functions of, 49, 516  
 mitochondrial, genetic code alterations in, 602–604, 603t  
 modified/unusual bases in, 580  
 notation for, 587  
 processing of, 572–573, 573f  
 release of, 643, 643f  
 serine-binding, 605–606, 606f  
 structure of, 587–588, 587f, 625–627  
 suppressor, 592–593, 592f, 593f  
 transcription of, 534t, 535, 535f  
 in translation, 619–620, 619f. *See also* Translation  
 transport of, 569
- tRNA charging, 587, 625  
 catalytic RNA and, 604b–605b  
 fidelity of, 627–629, 628f  
 self-charging and, 604b–605b
- tRNA nucleotidyltransferase, 573, 573f
- tRNA processing, 572–573, 573f. *See also* RNA processing
- tRNA<sup>fMet</sup>, 631, 639
- tRNA<sup>Met</sup>, 631
- tRNA<sup>Sec</sup>, 605–606, 606f
- Trombone model, 384, 385f
- trp operon, 708–710, 708f, 709f  
 in *Bacillus subtilis*, 728
- Trp repressor, 708f, 709
- Truncated mRNAs  
 in bacteria, 647–651, 651f  
 in eukaryotes, 651–653, 652f, 653f  
 stalled ribosomes on, tmRNA rescue of, 647–651, 652f, 653f
- Trypanosomes, RNA editing in, 566
- Tryptophan, 98, 98f, 99t  
 regulation of, 708–710  
   TRAP system in, 710, 728
- Tschermak, Erich von, 32
- Tsien, Roger, 253
- Tsugita, Akira, 610
- Tumor suppressor genes, 412
- Turnover number, 151
- Tus, 377t
- Tuschl, Thomas, 805
- Tus-Ter system, 395–397, 398f
- 23S rRNA, 573, 575f
- 25S rRNA, transcription of, 534, 534f, 534t
- 26S proteasome, 690, 691f
- Twist protein, 310, 310f
- Twists, in DNA, 297, 306  
 in histone wrapping, 335–336, 336f
- Two-dimensional gel electrophoresis, 280–281, 280f
- Two-dimensional nuclear magnetic resonance, 125–127, 126f, 127f
- Two-hybrid analysis, 246, 247f, 282
- 2μ plasmid, 485, 486f
- Two-start helix model, 339
- Ty elements, 498
- Type I topoisomerase, 312, 312f  
 discovery of, 324
- Type II restriction endonuclease, 217t, 218
- Type II topoisomerase, 312, 312f  
 discovery of, 325–326
- Tyrosine, 98, 98f, 99t
- Tyrosine kinase, in insulin regulation, 688b
- Tyrosine-class recombinases, in site-specific recombination, 484–485, 484f
- UAS<sub>GAL</sub>, 744b
- Ubiquitination, 163–165, 164f, 686, 690, 736
- ultrabithorax (ubx) gene*, 797, 797f, 798
- Ultraviolet light, in photoreversal, 429–430, 430f, 440
- Ultraviolet radiation, mutations from, 421–423, 424f, 440, 460
- Umbrella Murder, 647b
- Uncompetitive enzyme inhibitors, 152b–153b
- Underwinding, in DNA, 307–310, 307f, 336  
 topoisomerases in, 312–316, 312f, 312t, 314f–317f
- Unipotent stem cells, 799
- Unit cells, 122, 122f
- Units of measure, in molecular biology, 300
- Universal tree of life, 8–9, 9f
- Unnatural amino acids, incorporation in proteins, 630
- Untranslated regions (UTRs)  
 3'  
   in iron homeostasis, 779, 781f  
   in translation, 767, 774–776, 775f  
     in gene networks, 774, 775f  
 5'  
   in iron homeostasis, 779, 781f  
   in translation, 774, 775f  
     in gene networks, 774, 775f
- Upstream activator sequences, 737
- Upstream binding element, 534, 534f
- Upstream control element, 534, 534f, 534t
- Upstream open reading frames (uORFs), 777
- Upstream promoter element, in bacteria, 524

- Uracil (U), 64, 64t, 177, 177f, 178. *See also* Base(s); Base pairs/base pairing cytosine deamination to, 204, 204f, 416–417  
in DNA repair, 204  
nomenclature for, 64t  
Uracil DNA glycosylase, in base excision repair, 432–433, 432f, 433t  
Urey, Harold, 89b  
Uridine, 178, 179f  
UvrA, 433, 433t, 434f  
UvrB, 433, 433t, 434f  
UvrC, 433, 433t, 434f  
UvrD (helicase II)  
in mismatch repair, 425t, 426, 426f  
in nucleotide excision repair, 433t, 434, 434f  
Rec A and, 458  
in recombinational repair, 458
- V segments, immunoglobulin, 506–507, 506f  
Vaccines, antiviral, 503  
Valence, 70, 71f  
Valence bond model, 70  
Valine, 99t  
Val-tRNA synthetase, proofreading and, 628  
Van Beneden, Edouard, 34  
Van der Waals, Johannes, 74, 74f  
Van der Waals interactions, 74–75, 74f, 75f, 77, 77f  
in RNA, 198  
Van der Waals radius, 74, 74f, 107, 108f  
Variation. *See* Biological diversity  
Vectors, cloning. *See* DNA cloning, vectors in  
Velocity  
initial, 149, 150f  
maximum, 149  
Venter, J. Craig, 262, 262f, 268b, 292  
Vesicles, RNA-containing, 1, 4f, 6, 20  
*Vibrio cholerae*, autoinducers in, 729  
*Vibrio harveyi*  
autoinducers in, 697  
signaling in, 697  
Victoria, Queen of England, 50, 51f  
Vinograd, Jerome, 323  
Viral infections. *See also* Diseases and disorders  
λ phage in, 722–725. *See also* Lambda (λ) phage  
lysogenic, 725  
Viral ribozymes, 577b  
Viruses  
bacterial. *See* Bacteriophage(s)  
chromosomes of, 298, 300–301  
classes of, 267t  
deamination in, 204  
DNA of, 300–301, 301t  
evolution of, 7b, 129  
genome of, 300–301, 301t  
sequencing of, 266, 267t  
insect, cloning with, 237–238, 237f  
internal ribosome entry site in, 635–637, 637f  
proteolytic cleavage in, 168  
retroviruses, 500–503  
RNA editing in, 566–568  
RNA interference in, 786–787, 787f, 788b–789b  
sequence analysis of, 129, 129f  
site-specific recombination in, 485–487, 486f  
translation in, initiation of, 635–637, 637f  
vaccines for, 503  
Virusoids, 577  
Vitamin B<sub>2</sub>, 182  
Vitamins, coenzymes and, 145, 145t  
Vogelstein, Bert, 428b–429b  
Von Nägeli, Karl Wilhelm, 32  
Von Tschermak, Erich, 32  
Von Waldeyer, Heinrich, 32  
Vries, Hugo de, 32  
Waldeyer, Heinrich von, 32  
Walker A/B sequences, 130, 130f  
Wallace, Alfred Russel, 21  
Wang, James, 324  
Water, bound molecules of, 77  
Watson, James, 14, 15, 24, 25, 47–48, 176f  
Human Genome Project and, 261  
Weak chemical interactions, 73–78  
bound water molecules, 77, 77f  
in chromatin, 344, 345f  
cumulative effect of, 76–77, 77f  
definition of, 73–74  
hydrogen bonds. *See* Hydrogen bonds  
hydrophobic, 75, 75f, 77, 77f  
interaction among, 77–78, 77f  
ionic bonds, 74, 77, 77f, 146–147  
protein folding and, 115  
relative strength of, 76–77  
in RNA, 197–198, 198f  
structural stability and, 76–78, 77f  
van der Waals, 74–75, 74f, 75f, 77, 77f, 198  
Weeks, Kevin, 199  
Weight, molecular, 54  
Wellcome Trust Sanger Institute  
website, 264  
Werner syndrome, 412  
Wernig, Marius, 533b  
Western blotting, 243–244, 245f, 762  
wg gene, 795–796  
Whole-genome shotgun sequencing, 262.  
*See also* Genome sequencing  
Wickens, Marvin, 804  
Widom, Jonathan, 331  
Wieschaus, Eric F., 792, 793f  
Wild-type traits, 38–40, 38f  
Wilkins, Maurice, 176, 185–186, 186f, 209  
Wilson, Edmund B., 37, 37f  
Wing development, in butterflies, 806  
wingless gene, 795–796  
Winkler, Hans, 260  
Wnt-class signaling pathways, 795–796, 796f, 802  
Wobble bases, 589–590, 589f, 590f  
Wobble hypothesis, 590  
Wobble position, 590  
Woese, Carl, 6, 211, 269, 269, 615, 617  
Wolberger, Cynthia, 707f, 708  
Wollman, Elie, 170  
Wool, Ira, 647b  
Writhe, 297, 306, 310, 310f  
X chromosomes, 37, 37f  
inactivation of, 347–348, 350b–351b, 756  
X inactivation center, 756  
X rays, mutations due to, 421–423, 424f  
XerCD  
in homologous recombination, 461, 461f  
in site-specific recombination, 461, 461f  
Xeroderma pigmentosum, 433t, 434, 435f, 436b, 539–540  
nucleotide excision repair in, 433t, 434, 435f, 436b  
X-gal (5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside), 703, 703f  
XIST, 350b–351b  
XO sex determination, 37, 37f  
XP proteins, in nucleotide excision repair, 433t, 434, 435f, 436b  
XPD, in nucleotide excision repair, 433t, 434, 435f  
X-ray crystallography, 121–125, 122f–124f  
constructive interference in, 122, 122f  
diffraction patterns in, 122, 122f  
image reconstruction in, 123  
initial model in, 123–124  
isomorphous replacement in, 125  
molecular replacement in, 125  
multiwavelength anomalous dispersion in, 125  
refinement, 123–124, 124f  
X-ray diffraction, 185–186, 209  
Xrs2, 467, 467f  
XY sex determination, 37, 37f  
Y chromosomes, 37, 37f  
genomic Adam and, 288, 290  
Yang, Wei, 437, 437f, 481  
Yeast  
budding. *See* *Saccharomyces cerevisiae*  
chromosomes of, 302t  
in cloning, 236, 236f  
combinatorial control in, 743–746, 743f, 746f  
DNA of  
packaging of, 302–303  
size, 302t

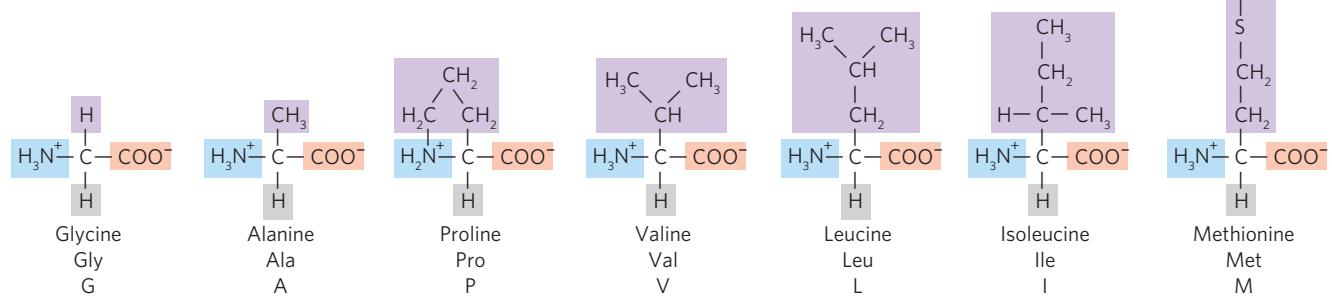
- Yeast (*continued*)  
fission, gene silencing in, 754b, 784  
galactose metabolism in, 743–746,  
743f, 746f  
gene silencing in, 754b, 784  
genetic code alterations in, 603–604,  
603t, 604–605, 606  
genome of, 302t  
introns in, 555, 780b  
mating-type switches in, 470–472, 470f,  
471f, 746–747, 747f  
Mediator complex in, 538–539, 538f, 741  
mismatch repair in, 425t, 427–428
- as model organism. *See Saccharomyces cerevisiae*  
promoters in, 737, 738, 738f  
replication in, 394–395, 395f  
site-specific recombination in, 485, 486f  
Ty elements in, 498, 498f  
Yeast artificial chromosomes,  
223–224, 224f  
Yeast three-hybrid analysis, 247–248,  
247f, 282  
Yeast two-hybrid analysis, 246, 247f, 282  
Yonath, Ada, 617
- Zaher, Hani, 662  
Zamecnik, Paul, 15, 48, 49, 586, 608, 616  
Z-DNA, 188, 189, 189f  
Zigzag model, of 30 nm filament,  
339–341, 340f  
Zinc finger motifs, 681–682, 682f  
Zinc-binding domain, 116, 116f  
Zipursky, Larry, 556, 556f  
Z-RNA, 197

## The Chemical Building Blocks of DNA and RNA

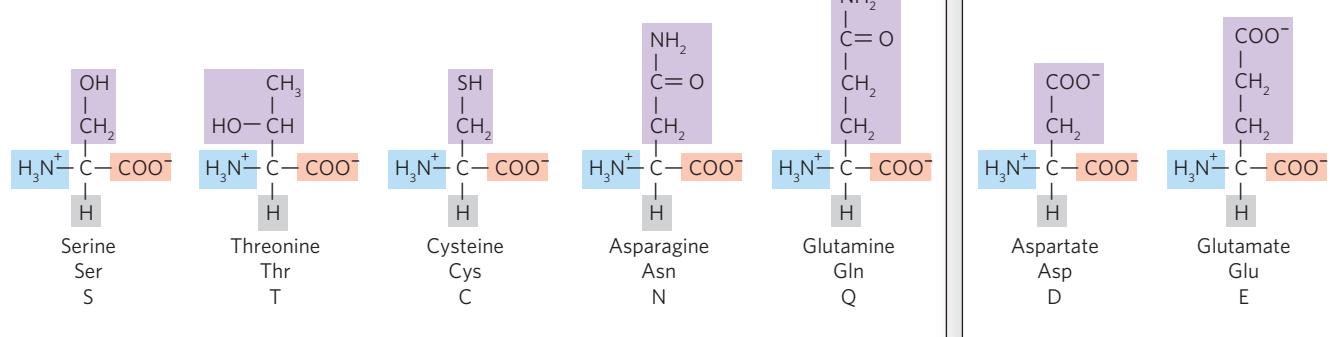


## The 20 Common Amino Acids

### Nonpolar, aliphatic R groups

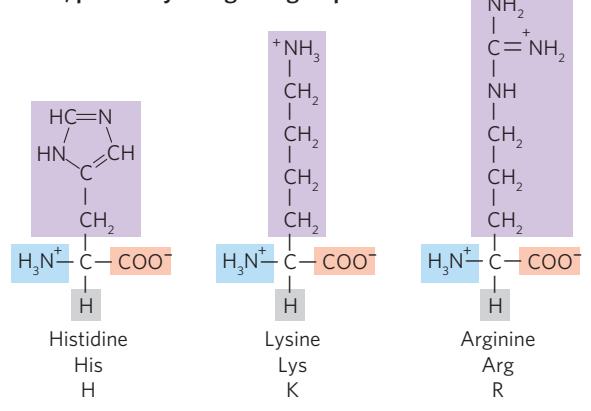


### Polar, uncharged R groups

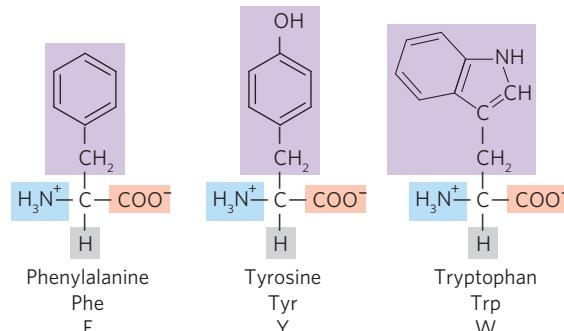


### Polar, negatively charged R groups

### Polar, positively charged R groups



### Nonpolar, aromatic R groups



## The Genetic Code

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