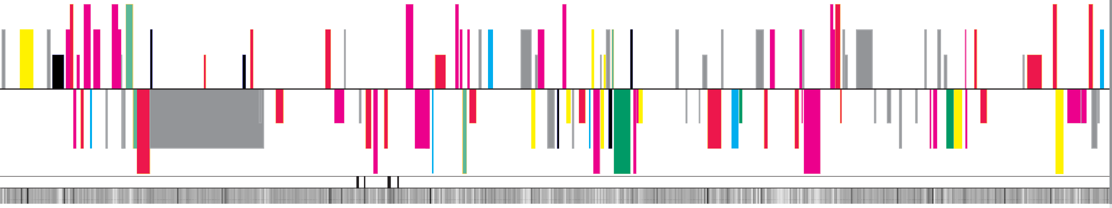


7

# CONTROL OF GENE EXPRESSION

An organism's DNA encodes all of the RNA and protein molecules required to construct its cells. Yet a complete description of the DNA sequence of an organism—be it the few million nucleotides of a bacterium or the few billion nucleotides of a human—no more enables us to reconstruct the organism than a list of English words enables us to reconstruct a play by Shakespeare. In both cases the problem is to know how the elements in the DNA sequence or the words on the list are used. Under what conditions is each gene product made, and, once made, what does it do?

In this chapter we discuss the first half of this problem—the rules and mechanisms by which a subset of the genes is selectively expressed in each cell. The mechanisms that control the expression of genes operate at many levels, and we discuss the different levels in turn. At the end of the chapter, we examine how modern-day genomes and their systems of regulation have been shaped by evolutionary processes. We begin with an overview of some basic principles of gene control in multicellular organisms.

## AN OVERVIEW OF GENE CONTROL

The different cell types in a multicellular organism differ dramatically in both structure and function. If we compare a mammalian neuron with a lymphocyte, for example, the differences are so extreme that it is difficult to imagine that the two cells contain the same genome (Figure 7–1). For this reason, and because cell differentiation is often irreversible, biologists originally suspected that genes might be selectively lost when a cell differentiates. We now know, however, that cell differentiation generally depends on changes in gene expression rather than on any changes in the nucleotide sequence of the cell's genome.

### The Different Cell Types of a Multicellular Organism Contain the Same DNA

The cell types in a multicellular organism become different from one another because they synthesize and accumulate different sets of RNA and protein molecules. They generally do this without altering the sequence of their DNA. Evidence for the preservation of the genome during cell differentiation comes from a classic set of experiments in frogs. When the nucleus of a fully differentiated frog cell is injected into a frog egg whose nucleus has been removed, the injected donor nucleus is capable of directing the recipient egg to produce a normal tadpole (Figure 7–2A). Because the tadpole contains a full range of differentiated cells that derived their DNA sequences from the nucleus of the original donor cell, it follows that the differentiated donor cell cannot have lost any important DNA sequences. A similar conclusion has been reached in experiments performed with various plants. Here differentiated pieces of plant tissue are placed in culture and then dissociated into single cells. Often, one of these individual cells can regenerate an entire adult plant (Figure 7–2B). Finally, this same principle has been recently demonstrated in mammals, including sheep, cattle, pigs, goats, and mice by introducing nuclei from somatic cells into enucleated eggs; when placed into surrogate mothers, some of these eggs (called reconstructed zygotes) develop into healthy animals (Figure 7–2C).

Further evidence that large blocks of DNA are not lost or rearranged during vertebrate development comes from comparing the detailed banding patterns detectable in condensed chromosomes at mitosis (see Figure 4–11). By this criterion the chromosome sets of all differentiated cells in the human body appear to be identical. Moreover, comparisons of the genomes of different cells based on recombinant DNA technology have shown, as a general rule, that the changes in gene expression that underlie the development of multicellular organisms are not accompanied by changes in the DNA sequences of the corresponding genes. There are, however, a few cases where DNA rearrangements of the genome take place during the development of an organism—most notably, in generating the diversity of the immune system of mammals (discussed in Chapter 24).

## Different Cell Types Synthesize Different Sets of Proteins

As a first step in understanding cell differentiation, we would like to know how many differences there are between any one cell type and another. Although we still do not know the answer to this fundamental question, we can make certain general statements.

1. Many processes are common to all cells, and any two cells in a single organism therefore have many proteins in common. These include the structural proteins of chromosomes, RNA polymerases, DNA repair enzymes, ribosomal proteins, enzymes involved in the central reactions of metabolism, and many of the proteins that form the cytoskeleton.
2. Some proteins are abundant in the specialized cells in which they function and cannot be detected elsewhere, even by sensitive tests. Hemoglobin, for example, can be detected only in red blood cells.
3. Studies of the number of different mRNAs suggest that, at any one time, a typical human cell expresses approximately 10,000–20,000 of its approximately 30,000 genes. When the patterns of mRNAs in a series of different human cell lines are compared, it is found that the level of expression of almost every active gene varies from one cell type to another. A few of these differences are striking, like that of hemoglobin noted above but most are much more subtle. The patterns of mRNA abundance (determined using DNA microarrays, discussed in Chapter 8) are so characteristic of cell type that they can be used to type human cancer cells of uncertain tissue origin (Figure 7–3).
4. Although the differences in mRNAs among specialized cell types are striking, they nonetheless underestimate the full range of differences in the pattern of protein production. As we shall see in this chapter, there are many steps after transcription at which gene expression can be regulated. In addition, alternative splicing can produce a whole family of proteins from a single gene. Finally, proteins can be covalently modified after they are synthesized. Therefore a better way of appreciating the radical differences in gene expression between cell types is through the use of two-dimensional gel electrophoresis, where protein levels are directly measured and some of the most common posttranslational modifications are displayed (Figure 7–4).

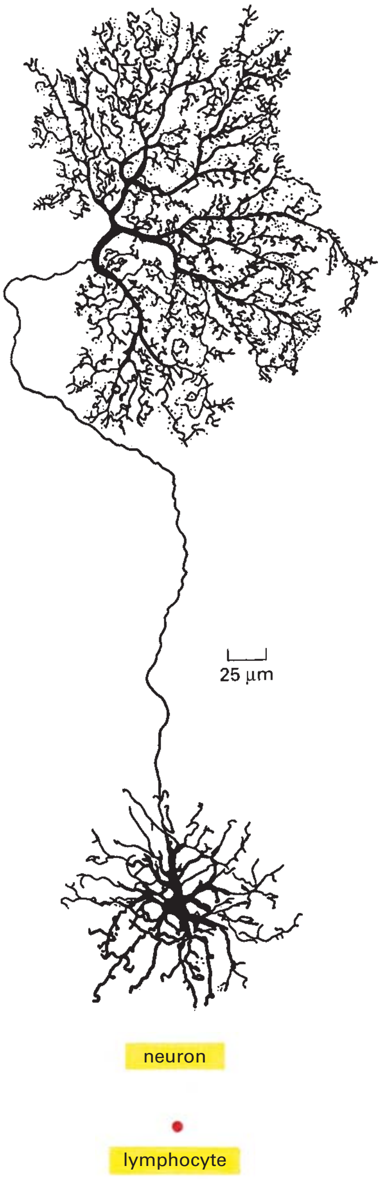
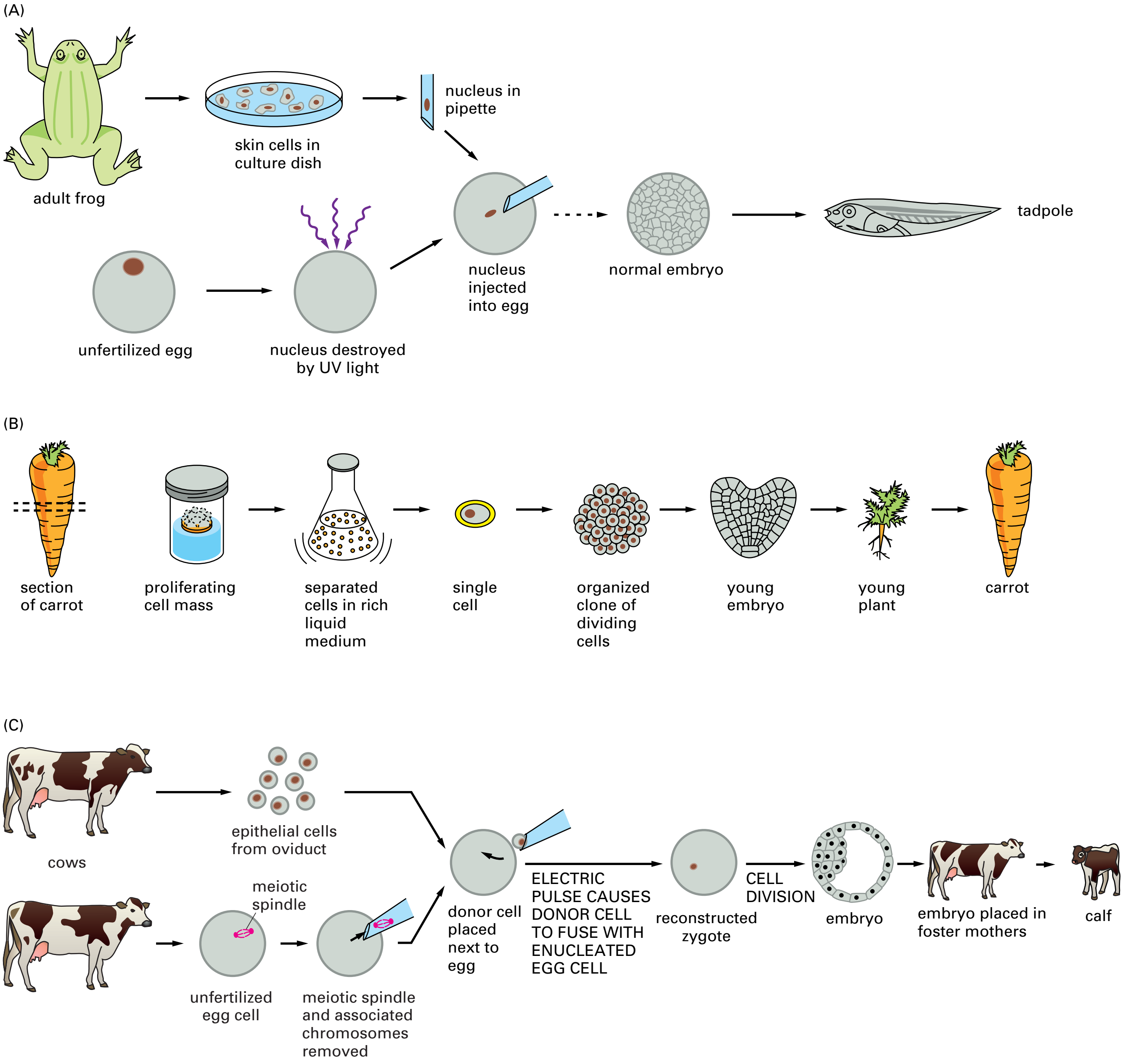


**Figure 7–1 A mammalian neuron and a lymphocyte.** The long branches of this neuron from the retina enable it to receive electrical signals from many cells and carry those signals to many neighboring cells. The lymphocyte is a white blood cell involved in the immune response to infection and moves freely through the body. Both of these cells contain the same genome, but they express different RNAs and proteins. (From B.B. Boycott, Essays on the Nervous System [R. Bellairs and E.G. Gray, eds.]. Oxford, UK: Clarendon Press, 1974.)

## A Cell Can Change the Expression of Its Genes in Response to External Signals

Most of the specialized cells in a multicellular organism are capable of altering their patterns of gene expression in response to extracellular cues. If a liver cell is exposed to a glucocorticoid hormone, for example, the production of several specific proteins is dramatically increased. Glucocorticoids are released in the body during periods of starvation or intense exercise and signal the liver to increase the production of glucose from amino acids and other small molecules;

**Figure 7–2 Evidence that a differentiated cell contains all the genetic instructions necessary to direct the formation of a complete organism.** (A) The nucleus of a skin cell from an adult frog transplanted into an enucleated egg can give rise to an entire tadpole. The *broken arrow* indicates that, to give the transplanted genome time to adjust to an embryonic environment, a further transfer step is required in which one of the nuclei is taken from the early embryo that begins to develop and is put back into a second enucleated egg. (B) In many types of plants, differentiated cells retain the ability to "dedifferentiate," so that a single cell can form a clone of progeny cells that later give rise to an entire plant. (C) A differentiated cell from an adult cow introduced into an enucleated egg from a different cow can give rise to a calf. Different calves produced from the same differentiated cell donor are genetically identical and are therefore clones of one another. (A, modified from J.B. Gurdon, *Sci. Am.* 219(6):24–35, 1968.)


(A)
adult frog
skin cells in culture dish
nucleus in pipette
unfertilized egg
nucleus destroyed by UV light
nucleus injected into egg
normal embryo
tadpole
(B)
section of carrot
proliferating cell mass
separated cells in rich liquid medium
single cell
organized clone of dividing cells
young embryo
young plant
carrot
(C)
cows
epithelial cells from oviduct
meiotic spindle
unfertilized egg cell
meiotic spindle and associated chromosomes removed
donor cell placed next to egg
ELECTRIC PULSE CAUSES DONOR CELL TO FUSE WITH ENUCLEATED EGG CELL
reconstructed zygote
CELL DIVISION
embryo
embryo placed in foster mothers
calf

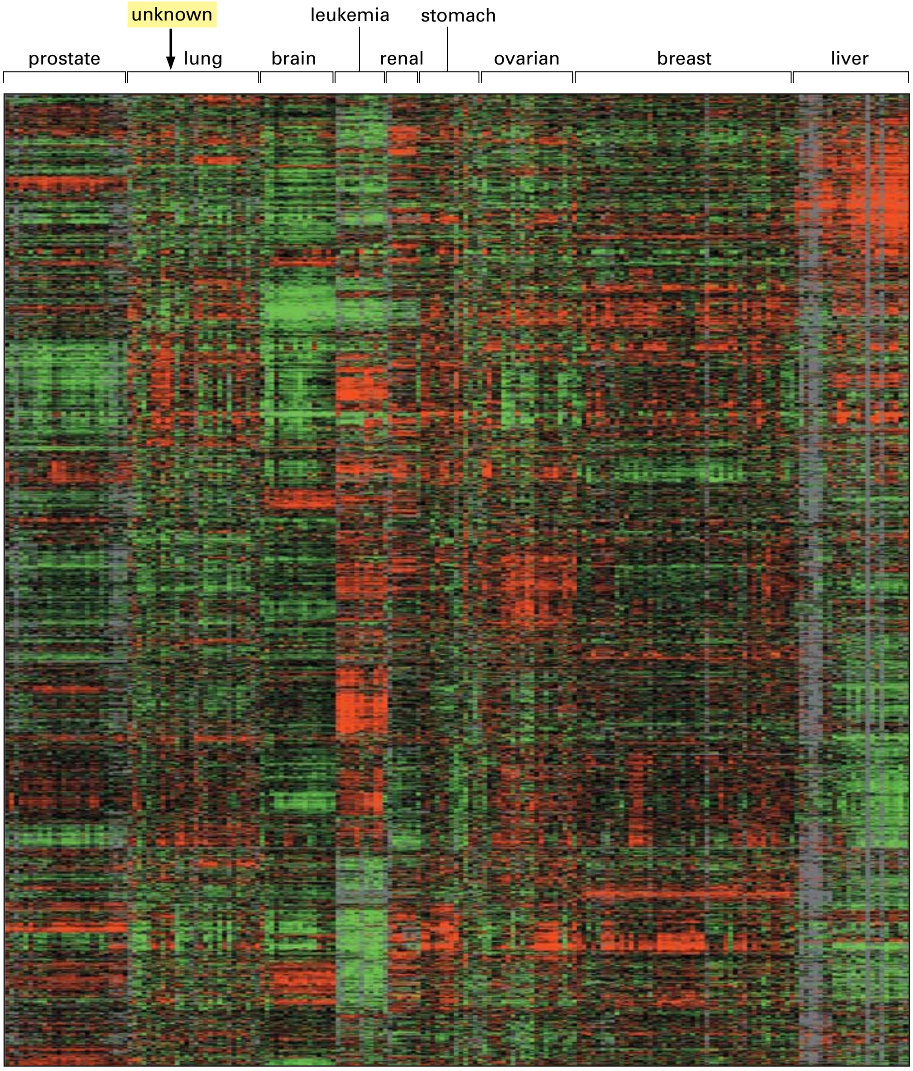


**Figure 7–3 Differences in mRNA expression patterns among different types of human cancer cells.** This figure summarizes a very large set of measurements in which the mRNA levels of 1800 selected genes (arranged *top* to *bottom)* were determined for 142 different human tumors (arranged *left* to *right),* each from a different patient. Each small *red* bar indicates that the given gene in the given tumor is transcribed at a level significantly higher than the average across all the cell lines. Each small *green* bar indicates a less-than-average expression level, and each *black* bar denotes an expression level that is close to average across the different tumors. The procedure used to generate these data—mRNA isolation followed by hybridization to DNA microarrays—is described in Chapter 8 (see pp. 533–535). The figure shows that the relative expression levels of each of the 1800 genes analyzed vary among the different tumors (seen by following a given gene *left* to *right* across the figure). This analysis also shows that each type of tumor has a characteristic gene expression pattern. This information can be used to "type" cancer cells of unknown tissue origin by matching the gene expression profiles to those of known tumors. For example, the unknown sample in the figure has been identified as a lung cancer. (Courtesy of Patrick O. Brown, David Botstein, and the Stanford Expression Collaboration.)

the set of proteins whose production is induced includes enzymes such as tyrosine aminotransferase, which helps to convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins drops to its normal level.

Other cell types respond to glucocorticoids differently. In fat cells, for example, the production of tyrosine aminotransferase is reduced, while some other cell types do not respond to glucocorticoids at all. These examples illustrate a general feature of cell specialization: different cell types often respond in different ways to the same extracellular signal. Underlying such adjustments that occur in response to extracellular signals, there are features of the gene expression pattern that do not change and give each cell type its permanently distinctive character.

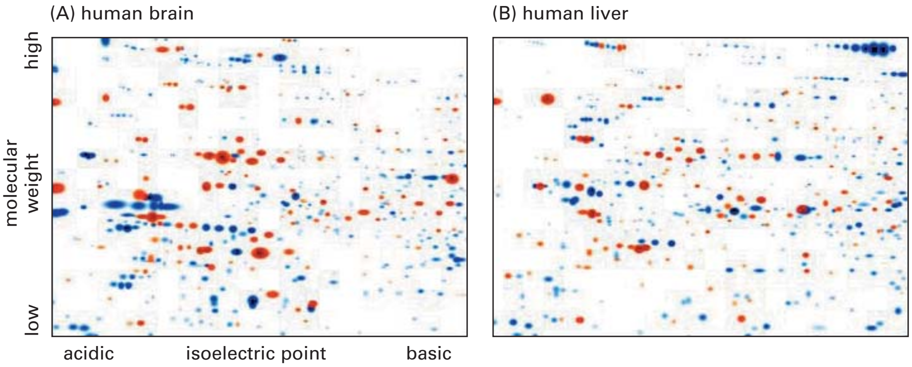


**Figure 7–4 Differences in the proteins expressed by two human tissues.** In each panel, the proteins have been displayed using two-dimensional polyacrylamide gel electrophoresis (see pp. 485–487). The proteins have been separated by molecular weight *(top* to *bottom)* and isoelectric point, the pH at which the protein has no net charge *(right* to *left).* The protein spots artificially colored *red* are common to both samples; those in *blue* are specific to one of the two tissues. The differences between the two tissue samples vastly outweigh their similarities: even for proteins that are shared between the two tissues, their relative abundance is usually different. Note that this technique separates proteins both by size and charge; therefore a protein that has, for example, several different phosphorylation states will appear as a series of *horizontal spots* (see *upper right-hand* portion of *right* panel). Only a small portion of the complete protein spectrum is shown for each sample. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

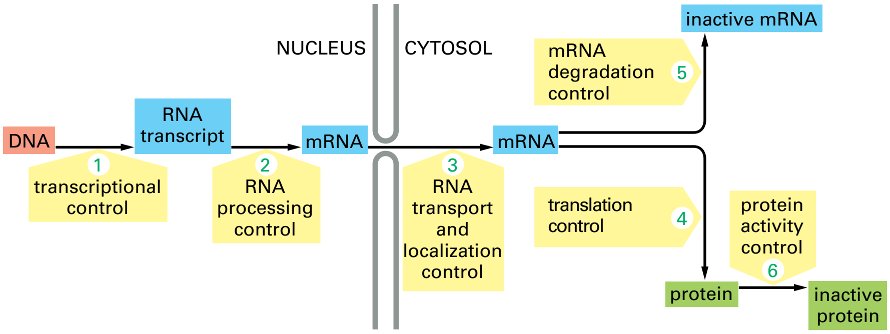


**Figure 7–5 Six steps at which eucaryotic gene expression can be controlled.** Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, includes reversible activation or inactivation by protein phosphorylation (discussed in Chapter 3) as well as irreversible inactivation by proteolytic degradation (discussed in Chapter 6).

### Gene Expression Can Be Regulated at Many of the Steps in the Pathway from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that the cells express, at what level is the control of gene expression exercised? As we saw in the last chapter, there are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. Thus a cell can control the proteins it makes by (1) controlling when and how often a given gene is transcribed (**transcriptional control**), (2) controlling how the RNA transcript is spliced or otherwise processed (**RNA processing control**), (3) selecting which completed mRNAs in the cell nucleus are exported to the cytosol and determining where in the cytosol they are localized (**RNA transport and localization control**), (4) selecting which mRNAs in the cytoplasm are translated by ribosomes (**translational control**), (5) selectively destabilizing certain mRNA molecules in the cytoplasm (**mRNA degradation control**), or (6) selectively activating, inactivating, degrading, or compartmentalizing specific protein molecules after they have been made (**protein activity control**) (Figure 7–5).

For most genes transcriptional controls are paramount. This makes sense because, of all the possible control points illustrated in Figure 7–5, only transcriptional control ensures that the cell will not synthesize superfluous intermediates. In the following sections we discuss the DNA and protein components that perform this function by regulating the initiation of gene transcription. We shall return at the end of the chapter to the additional ways of regulating gene expression.

### Summary

***The genome of a cell contains in its DNA sequence the information to make many thousands of different protein and RNA molecules. A cell typically expresses only a fraction of its genes, and the different types of cells in multicellular organisms arise because different sets of genes are expressed. Moreover, cells can change the pattern of genes they express in response to changes in their environment, such as signals from other cells. Although all of the steps involved in expressing a gene can in principle be regulated, for most genes the initiation of RNA transcription is the most important point of control.***

## DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS

How does a cell determine which of its thousands of genes to transcribe? As mentioned briefly in Chapters 4 and 6, the transcription of each gene is controlled by a regulatory region of DNA relatively near the site where transcription begins. Some regulatory regions are simple and act as switches that are thrown by a single signal. Many others are complex and act as tiny microprocessors, responding to a variety of signals that they interpret and integrate to switch the neighboring gene on or off. Whether complex or simple, these switching devices contain two types of fundamental components: (1) short stretches of DNA of defined sequence and (2) *gene regulatory proteins* that recognize and bind to them.

We begin our discussion of gene regulatory proteins by describing how these proteins were discovered.

## Gene Regulatory Proteins Were Discovered Using Bacterial Genetics

Genetic analyses in bacteria carried out in the 1950s provided the first evidence for the existence of **gene regulatory proteins** that turn specific sets of genes on or off. One of these regulators, the *lambda repressor*, is encoded by a bacterial virus, *bacteriophage lambda.* The repressor shuts off the viral genes that code for the protein components of new virus particles and thereby enables the viral genome to remain a silent passenger in the bacterial chromosome, multiplying with the bacterium when conditions are favorable for bacterial growth (see Figure 5–81). The lambda repressor was among the first gene regulatory proteins to be characterized, and it remains one of the best understood, as we discuss later. Other bacterial regulators respond to nutritional conditions by shutting off genes encoding specific sets of metabolic enzymes when they are not needed. The *lac repressor*, the first of these bacterial proteins to be recognized, turns off the production of the proteins responsible for lactose metabolism when this sugar is absent from the medium.

The first step toward understanding gene regulation was the isolation of mutant strains of bacteria and bacteriophage lambda that were unable to shut off specific sets of genes. It was proposed at the time, and later proven, that most of these mutants were deficient in proteins acting as specific repressors for these sets of genes. Because these proteins, like most gene regulatory proteins, are present in small quantities, it was difficult and time-consuming to isolate them. They were eventually purified by fractionating cell extracts. Once isolated, the proteins were shown to bind to specific DNA sequences close to the genes that they regulate. The precise DNA sequences that they recognized were then determined by a combination of classical genetics, DNA sequencing, and DNA-footprinting experiments (discussed in Chapter 8).

## The Outside of the DNA Helix Can Be Read by Proteins

As discussed in Chapter 4, the DNA in a chromosome consists of a very long double helix (Figure 7–6). Gene regulatory proteins must recognize specific nucleotide sequences embedded within this structure. It was originally thought that these proteins might require direct access to the hydrogen bonds between base pairs in the interior of the double helix to distinguish between one DNA sequence and another. It is now clear, however, that the outside of the double helix is studded with DNA sequence information that gene regulatory proteins can recognize without having to open the double helix. The edge of each base pair is exposed at the surface of the double helix, presenting a distinctive pattern of hydrogen bond donors, hydrogen bond acceptors, and hydrophobic patches for proteins to recognize in both the major and minor groove (Figure 7–7). But only in the major groove are the patterns markedly different for each of the four base-pair arrangements (Figure 7–8). For this reason, gene regulatory proteins generally bind to the major groove—as we shall see.

Although the patterns of hydrogen bond donor and acceptor groups are the most important features recognized by gene regulatory proteins, they are not the only ones: the nucleotide sequence also determines the overall geometry of the double helix, creating distortions of the "idealized" helix that can also be recognized.

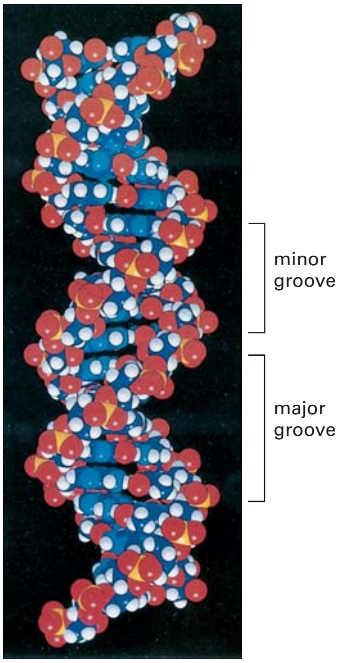


**Figure 7–6 Double-helical structure of DNA.** The major and minor grooves on the outside of the double helix are indicated. The atoms are colored as follows: carbon, *dark blue;* nitrogen, *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.*

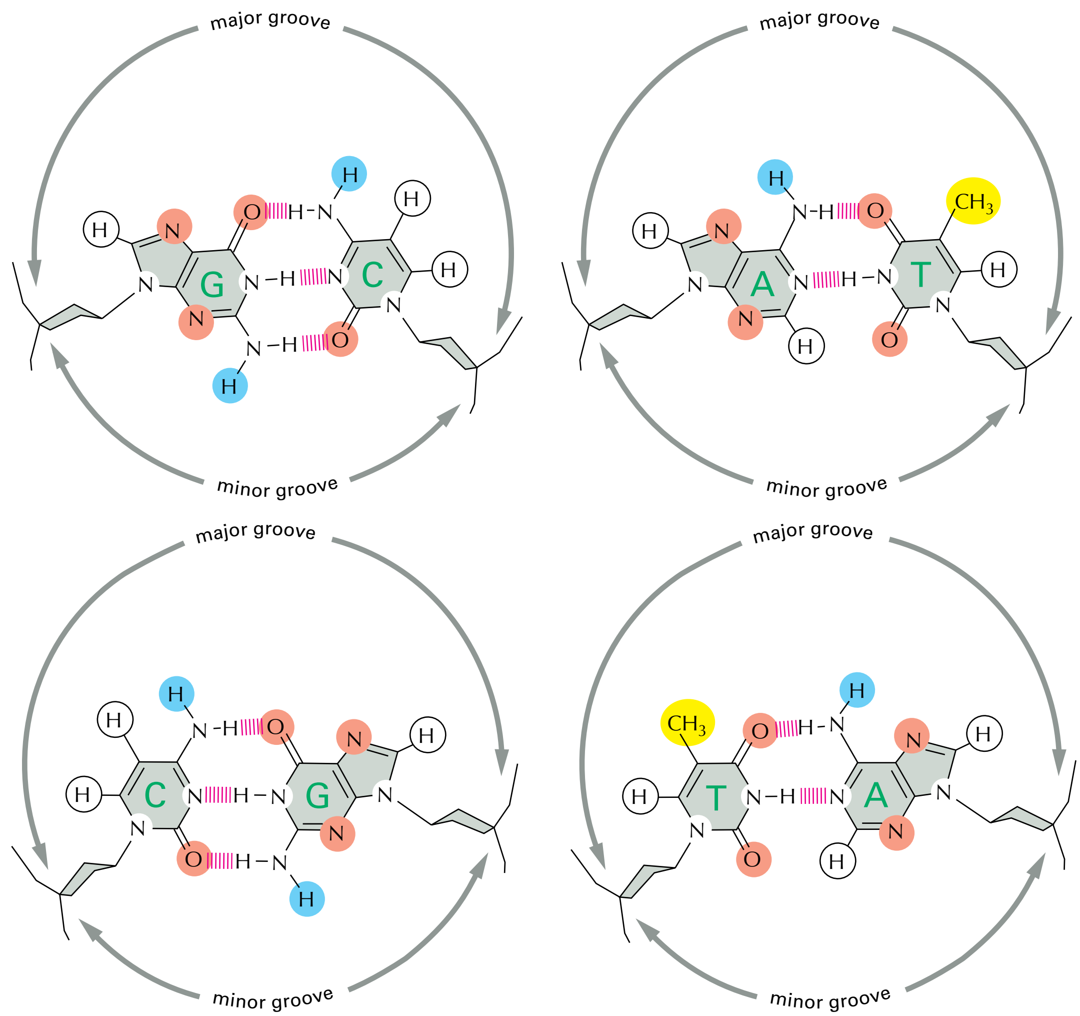


**Figure 7–7 How the different base pairs in DNA can be recognized from their edges without the need to open the double helix.** The four possible configurations of base pairs are shown, with potential hydrogen bond donors indicated in *blue,* potential hydrogen bond acceptors in *red,* and hydrogen bonds of the base pairs themselves as a series of short parallel *red* lines. Methyl groups, which form hydrophobic protuberances, are shown in *yellow,* and hydrogen atoms that are attached to carbons, and are therefore unavailable for hydrogen bonding, are *white.* (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd edn. New York: Garland Publishing, 1999.)

## The Geometry of the DNA Double Helix Depends on the Nucleotide Sequence

For 20 years after the discovery of the DNA double helix in 1953, DNA was thought to have the same monotonous structure, with exactly 36° of helical twist between its adjacent nucleotide pairs (10 nucleotide pairs per helical turn) and a uniform helix geometry. This view was based on structural studies of heterogeneous mixtures of DNA molecules, however, and it changed once the three-dimensional structures of short DNA molecules of defined nucleotide sequence

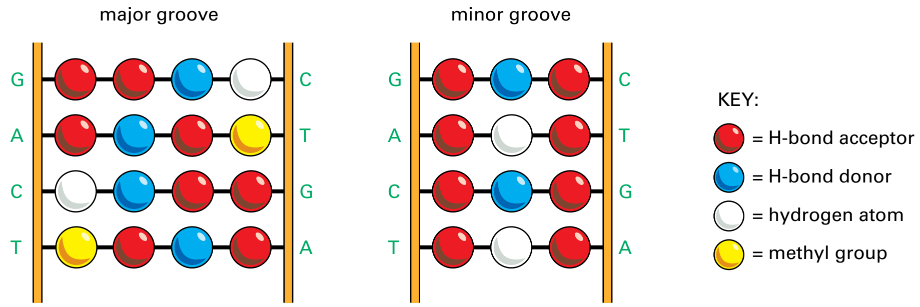


**Figure 7–8 A DNA recognition code.** The edge of each base pair, seen here looking directly at the major or minor groove, contains a distinctive pattern of hydrogen bond donors, hydrogen bond acceptors, and methyl groups. From the major groove, each of the four base-pair configurations projects a unique pattern of features. From the minor groove, however, the patterns are similar for G–C and C–G as well as for A–T and T–A. The color code is the same as that in Figure 7–7. (From C. Branden and J. Tooze, Introduction to Protein Structure, 2nd edn. New York: Garland Publishing, 1999.)

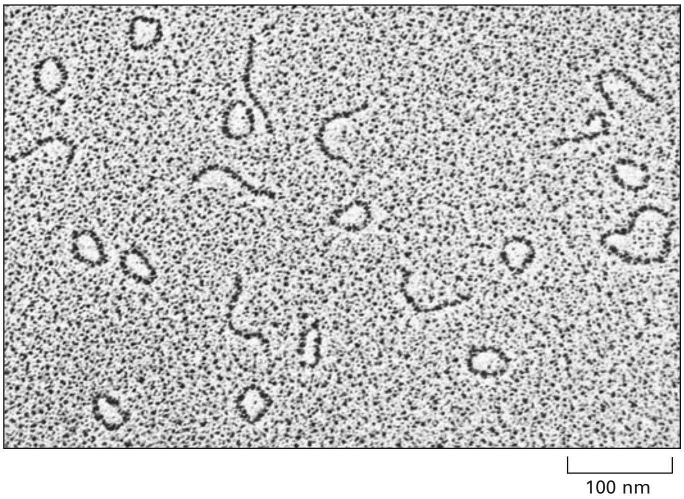


**Figure 7–9 Electron micrograph of fragments of a highly bent segment of DNA double helix.** The DNA fragments are derived from the small, circular mitochondrial DNA molecules of a trypanosome. Although the fragments are only about 200 nucleotide pairs long, many of them have bent to form a complete circle. On average, a normal DNA helix of this length would bend only enough to produce one-fourth of a circle (one smooth right-angle turn). (From J. Griffith, M. Bleyman, C.A. Raugh, P.A. Kitchin, and P.T. Englund, *Cell* 46:717–724, 1986. © Elsevier.)

were determined using x-ray crystallography and NMR spectroscopy. Whereas the earlier studies provided a picture of an average, idealized DNA molecule, the later studies showed that any given nucleotide sequence had local irregularities, such as tilted nucleotide pairs or a helical twist angle larger or smaller than 36°. These unique features can be recognized by specific DNA-binding proteins.

An especially striking departure from the average structure occurs in nucleotide sequences that cause the DNA double helix to bend. Some sequences (for example, AAAANNN, where N can be any base except A) form a double helix with a pronounced irregularity that causes a slight bend; if this sequence is repeated at 10-nucleotide-pair intervals in a long DNA molecule, the small bends add together so that the DNA molecule appears unusually curved when viewed in the electron microscope (Figure 7–9).

A related and equally important variable feature of DNA structure is the extent to which the double helix is deformable. For a protein to recognize and bind to a specific DNA sequence, there must be a tight fit between the DNA and the protein, and often the normal DNA conformation must be distorted to maximize this fit (Figure 7–10). The energetic cost of such distortion depends on the local nucleotide sequence. We encountered an example of this in the discussion of nucleosome assembly in Chapter 4: some DNA sequences can accommodate the tight DNA wrapping required for nucleosome formation better than others. Similarly, a few gene regulatory proteins induce a striking bend in the DNA when they bind to it (Figure 7–11). In general, these proteins recognize DNA sequences that are easily bent.

## Short DNA Sequences Are Fundamental Components of Genetic Switches

We have seen how a specific nucleotide sequence can be detected as a pattern of structural features on the surface of the DNA double helix. Particular nucleotide sequences, each typically less than 20 nucleotide pairs in length, function as fundamental components of genetic switches by serving as recognition sites for the binding of specific gene regulatory proteins. Thousands of such DNA sequences have been identified, each recognized by a different gene regulatory protein (or by a set of related gene regulatory proteins). Some of the gene regulatory proteins that are discussed in the course of this chapter are listed in Table 7–1, along with the DNA sequences that they recognize.

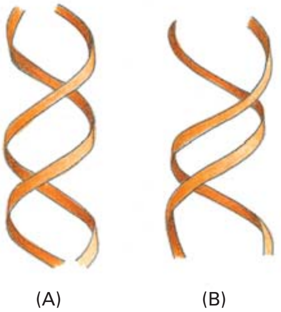


**Figure 7–10 DNA deformation induced by protein binding.** The figure shows the changes of DNA structure, from the conventional double-helix (A) to a distorted form (B) observed when a well-studied gene regulatory protein (the bacteriophage 434 repressor, a close relative of the lambda repressor) binds to specific sequences of DNA. The ease with which a DNA sequence can be deformed often affects the affinity of protein binding.

We now turn to the gene regulatory proteins themselves, the second fundamental component of genetic switches. We begin with the structural features that allows these proteins to recognize short, specific DNA sequences contained in a much longer double helix.

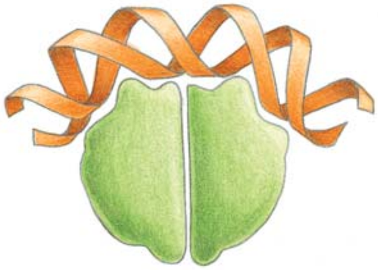


**Figure 7–11 The bending of DNA induced by the binding of the catabolite activator protein (CAP).** CAP is a gene regulatory protein from *E. coli*. In the absence of the bound protein, this DNA helix is straight.

## Gene Regulatory Proteins Contain Structural Motifs That Can Read DNA Sequences

Molecular recognition in biology generally relies on an exact fit between the surfaces of two molecules, and the study of gene regulatory proteins has provided some of the clearest examples of this principle. A gene regulatory protein recognizes a specific DNA sequence because the surface of the protein is extensively complementary to the special surface features of the double helix in that region. In most cases the protein makes a large number of contacts with the DNA, involving hydrogen bonds, ionic bonds, and hydrophobic interactions. Although each individual contact is weak, the 20 or so contacts that are typically formed at the protein–DNA interface add together to ensure that the interaction is both highly specific and very strong (Figure 7–12). In fact, DNA–protein interactions include some of the tightest and most specific molecular interactions known in biology.

Although each example of protein–DNA recognition is unique in detail, x-ray crystallographic and NMR spectroscopic studies of several hundred gene regulatory proteins have revealed that many of the proteins contain one or another of a small set of DNA-binding structural motifs. These motifs generally use

**TABLE 7–1 Some Gene Regulatory Proteins and the DNA Sequences That They Recognize**

| | NAME | DNA SEQUENCE RECOGNIZED* |
|---|---|---|
| Bacteria | lac repressor | 5′ AATTGTGAGCGGATAACAATT<br>3′ TTAACACTCGCCTATTGTTAA |
| | CAP | TGTGAGTTAGCTCACT<br>ACACTCAATCGAGTGA |
| | lambda repressor | TATCACCGCCAGAGGTA<br>ATAGTGGCGGTCTCCAT |
| Yeast | Gal4 | CGGAGGACTGTCCTCCG<br>GCCTCCTGACAGGAGGC |
| | Matα2 | CATGTAATT<br>GTACATTAA |
| | Gcn4 | ATGACTCAT<br>TACTGAGTA |
| *Drosophila* | Kruppel | AACGGGTTAA<br>TTGCCCAATT |
| | Bicoid | GGGATTAGA<br>CCCTAATCT |
| Mammals | Sp1 | GGGCGG<br>CCCGCC |
| | Oct-1 Pou domain | ATGCAAAT<br>TACGTTTA |
| | GATA-1 | TGATAG<br>ACTATC |
| | MyoD | CAAATG<br>GTTTAC |
| | p53 | GGGCAAGTCT<br>CCCGTTCAGA |

*Each protein in this table can recognize a set of closely related DNA sequences (see Figure 6–12); for convenience, only one recognition sequence, rather than a consensus sequence, is given for each protein.

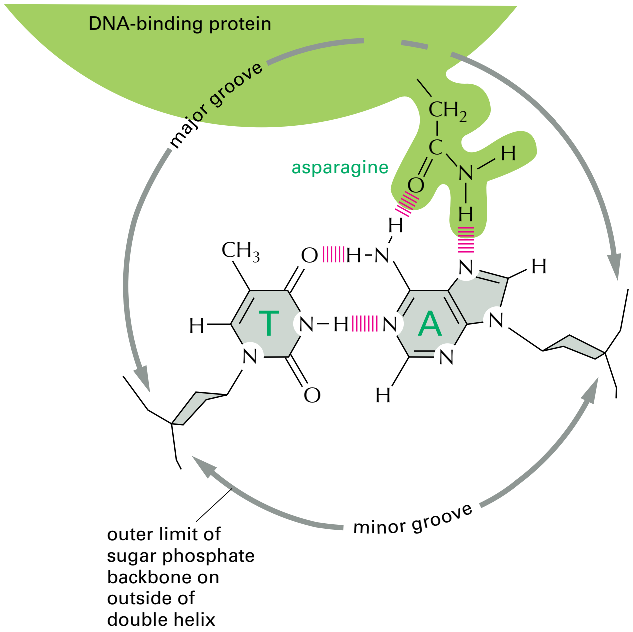


**Figure 7–12 The binding of a gene regulatory protein to the major groove of DNA.** Only a single contact is shown. Typically, the protein-DNA interface would consist of 10 to 20 such contacts, involving different amino acids, each contributing to the strength of the protein–DNA interaction.

either α helices or β sheets to bind to the major groove of DNA; this groove, as we have seen, contains sufficient information to distinguish one DNA sequence from any other. The fit is so good that it has been suggested that the dimensions of the basic structural units of nucleic acids and proteins evolved together to permit these molecules to interlock.

## The Helix–Turn–Helix Motif Is One of the Simplest and Most Common DNA-binding Motifs

The first DNA-binding protein motif to be recognized was the **helix–turn–helix**. Originally identified in bacterial proteins, this motif has since been found in hundreds of DNA-binding proteins from both eucaryotes and procaryotes. It is constructed from two α helices connected by a short extended chain of amino acids, which constitutes the "turn" (Figure 7–13). The two helices are held at a fixed angle, primarily through interactions between the two helices. The more C-terminal helix is called the *recognition helix* because it fits into the major groove of DNA; its amino acid side chains, which differ from protein to protein, play an important part in recognizing the specific DNA sequence to which the protein binds.

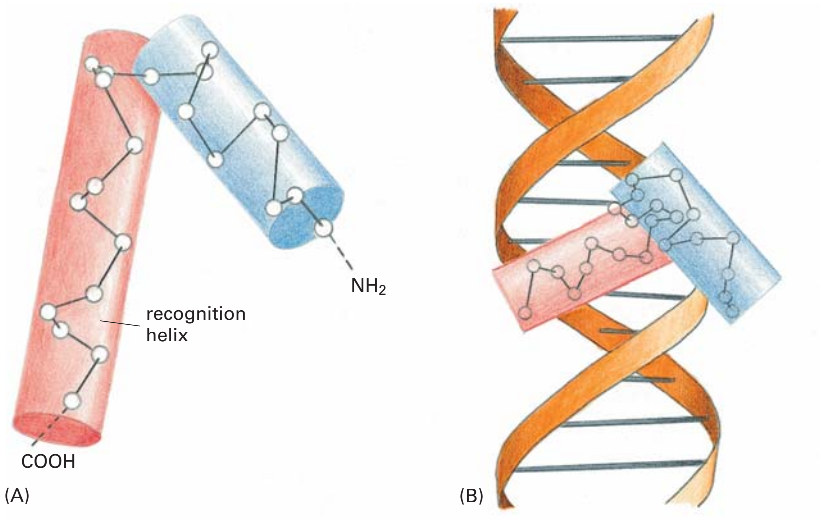


**Figure 7–13 The DNA-binding helix–turn–helix motif.** The motif is shown in (A), where each *white* circle denotes the central carbon of an amino acid. The C-terminal α helix *(red)* is called the recognition helix because it participates in sequence-specific recognition of DNA. As shown in (B), this helix fits into the major groove of DNA, where it contacts the edges of the base pairs (see also Figure 7–7). The N-terminal α-helix *(blue)* functions primarily as a structural component that helps to position the recognition helix.

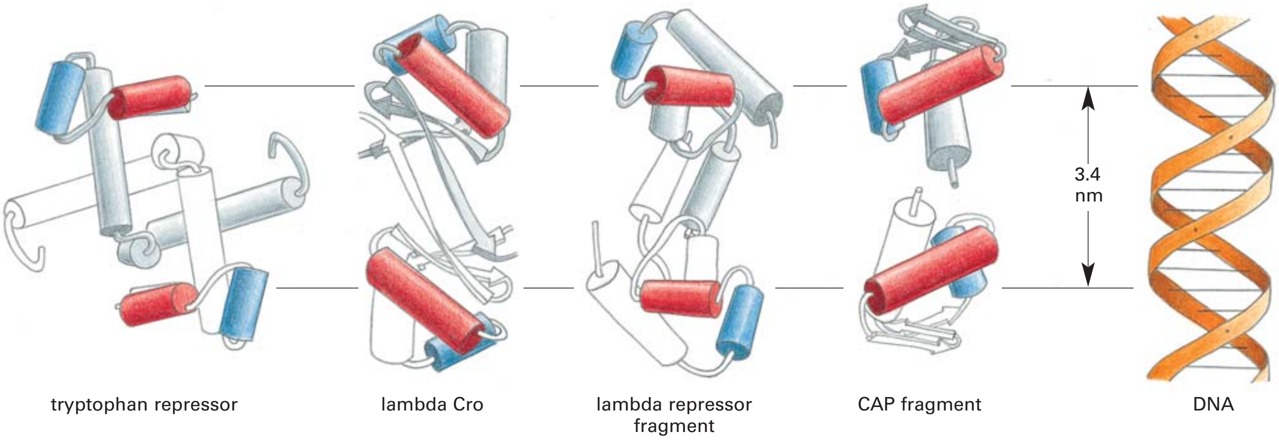


**Figure 7–14 Some helix–turn–helix DNA-binding proteins.** All of the proteins bind DNA as dimers in which the two copies of the recognition helix *(red cylinder)* are separated by exactly one turn of the DNA helix (3.4 nm). The other helix of the helix–turn–helix motif is colored *blue*, as in Figure 7–13. The lambda repressor and Cro proteins control bacteriophage lambda gene expression, and the tryptophan repressor and the catabolite activator protein (CAP) control the expression of sets of *E. coli* genes.

Outside the helix–turn–helix region, the structure of the various proteins that contain this motif can vary enormously (Figure 7–14). Thus each protein "presents" its helix–turn–helix motif to the DNA in a unique way, a feature thought to enhance the versatility of the helix–turn–helix motif by increasing the number of DNA sequences that the motif can be used to recognize. Moreover, in most of these proteins, parts of the polypeptide chain outside the helix–turn–helix domain also make important contacts with the DNA, helping to fine-tune the interaction.

The group of helix–turn–helix proteins shown in Figure 7–14 demonstrates a feature that is common to many sequence-specific DNA-binding proteins. They bind as symmetric dimers to DNA sequences that are composed of two very similar "half-sites," which are also arranged symmetrically (Figure 7–15). This arrangement allows each protein monomer to make a nearly identical set of contacts and enormously increases the binding affinity: as a first approximation, doubling the number of contacts doubles the free energy of the interaction and thereby *squares* the affinity constant.

## Homeodomain Proteins Constitute a Special Class of Helix–Turn–Helix Proteins

Not long after the first gene regulatory proteins were discovered in bacteria, genetic analyses in the fruit fly *Drosophila* led to the characterization of an important class of genes, the *homeotic selector genes*, that play a critical part in orchestrating fly development. As discussed in Chapter 21, they have since proved to have a fundamental role in the development of higher animals as well. Mutations in these genes cause one body part in the fly to be converted into another, showing that the proteins they encode control critical developmental decisions.

When the nucleotide sequences of several homeotic selector genes were determined in the early 1980s, each proved to contain an almost identical stretch of 60 amino acids that defines this class of proteins and is termed the **homeodomain**. When the three-dimensional structure of the homeodomain was determined, it was seen to contain a helix–turn–helix motif related to that of the bacterial gene regulatory proteins, providing one of the first indications that the principles of gene regulation established in bacteria are relevant to higher organisms as well. More than 60 homeodomain proteins have now been discovered in *Drosophila* alone, and homeodomain proteins have been identified in virtually all eucaryotic organisms that have been studied, from yeasts to plants to humans.

The structure of a homeodomain bound to its specific DNA sequence is shown in Figure 7–16. Whereas the helix–turn–helix motif of bacterial gene regulatory proteins is often embedded in different structural contexts, the helix–turn–helix motif of homeodomains is always surrounded by the same structure (which forms the rest of the homeodomain), suggesting that the motif is always presented to DNA in the same way. Indeed, structural studies have

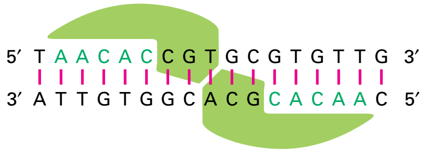

5′ T A A C A C C G T G C G T G T T G 3′
3′ A T T G T G G C A C G C A C A A C 5′

**Figure 7–15 A specific DNA sequence recognized by the bacteriophage lambda Cro protein.** The nucleotides labeled in *green* in this sequence are arranged symmetrically, allowing each half of the DNA site to be recognized in the same way by each protein monomer, also shown in *green*. See Figure 7–14 for the actual structure of the protein.

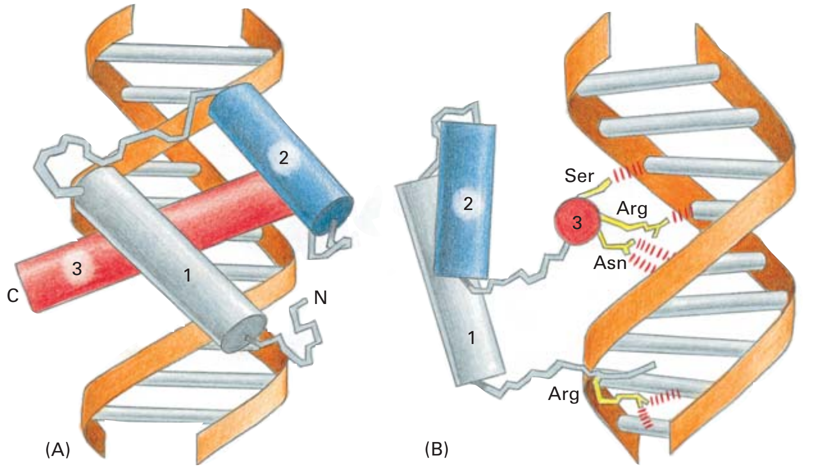


**Figure 7–16 A homeodomain bound to its specific DNA sequence.** Two different views of the same structure are shown. (A) The homeodomain is folded into three α helices, which are packed tightly together by hydrophobic interactions. The part containing helix 2 and 3 closely resembles the helix–turn–helix motif. (B) The recognition helix (helix 3, *red*) makes important contacts with the major groove of DNA. The asparagine (Asn) of helix 3, for example, contacts an adenine, as shown in Figure 7–12. Nucleotide pairs are also contacted in the minor groove by a flexible arm attached to helix 1. The homeodomain shown here is from a yeast gene regulatory protein, but it closely resembles homeodomains from many eucaryotic organisms. (Adapted from C. Wolberger et al., *Cell* 67:517–528, 1991.)

shown that a yeast homeodomain protein and a *Drosophila* homeodomain protein have very similar conformations and recognize DNA in almost exactly the same manner, although they are identical at only 17 of 60 amino acid positions (see Figure 3–15).

## There Are Several Types of DNA-binding Zinc Finger Motifs

The helix–turn–helix motif is composed solely of amino acids. A second important group of DNA-binding motifs adds one or more zinc atoms as structural components. Although all such zinc-coordinated DNA-binding motifs are called **zinc fingers**, this description refers only to their appearance in schematic drawings dating from their initial discovery (Figure 7–17A). Subsequent structural studies have shown that they fall into several distinct structural groups, two of which are considered here. The first type was initially discovered in the protein that activates the transcription of a eucaryotic ribosomal RNA gene. It is a simple structure, consisting of an α helix and a β sheet held together by the zinc (Figure 7–17B). This type of zinc finger is often found in a cluster with additional zinc fingers, arranged one after the other so that the α helix of each can contact the major groove of the DNA, forming a nearly continuous stretch of α helices along the groove. In this way, a strong and specific DNA-protein interaction is built up through a repeating basic structural unit (Figure 7–18). A particular advantage of this motif is that the strength and specificity of the DNA-protein interaction can be adjusted during evolution by changes in the number of zinc finger repeats. By contrast, it is difficult to imagine how any of the other DNA-binding motifs discussed in this chapter could be formed into repeating chains.

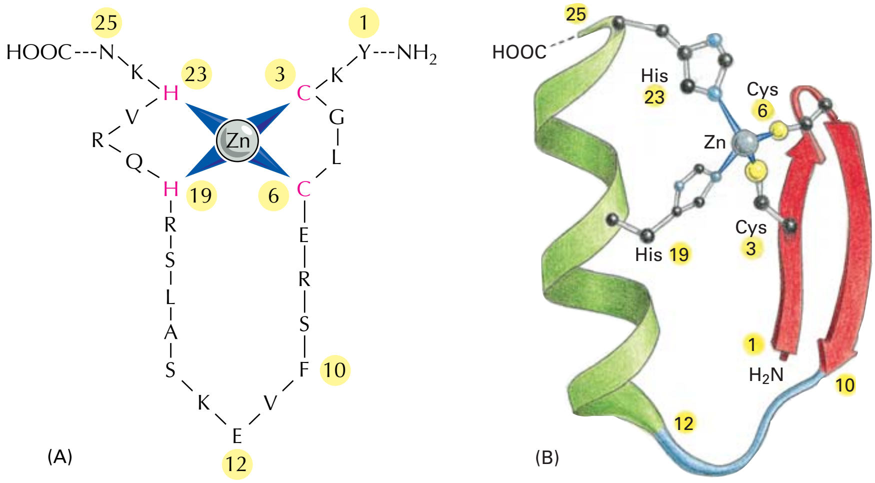


**Figure 7–17 One type of zinc finger protein.** This protein belongs to the Cys–Cys–His–His family of zinc finger proteins, named after the amino acids that grasp the zinc. (A) Schematic drawing of the amino acid sequence of a zinc finger from a frog protein of this class. (B) The three-dimensional structure of this type of zinc finger is constructed from an antiparallel β sheet (amino acids 1 to 10) followed by an α helix (amino acids 12 to 24). The four amino acids that bind the zinc (Cys 3, Cys 6, His 19, and His 23) hold one end of the α helix firmly to one end of the β sheet. (Adapted from M.S. Lee et al., *Science* 245:635–637, 1989.)

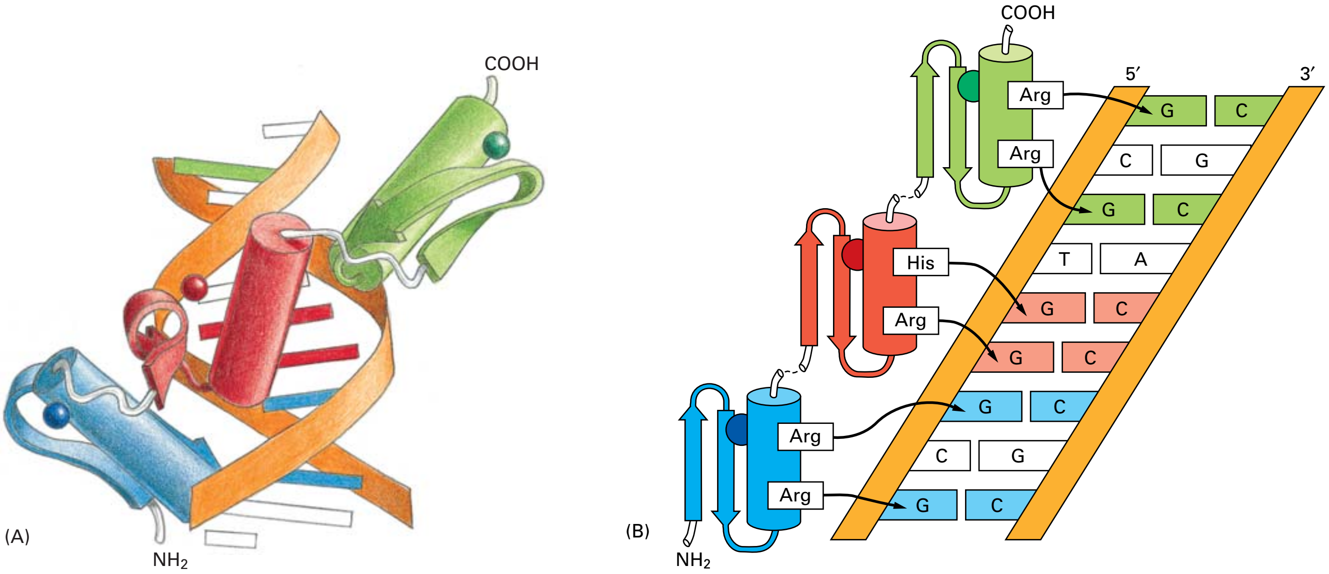


**Figure 7–18 DNA binding by a zinc finger protein.** (A) The structure of a fragment of a mouse gene regulatory protein bound to a specific DNA site. This protein recognizes DNA using three zinc fingers of the Cys–Cys–His–His type (see Figure 7–17) arranged as direct repeats. (B) The three fingers have similar amino acid sequences and contact the DNA in similar ways. In both (A) and (B) the zinc atom in each finger is represented by a small sphere. (Adapted from N. Pavletich and C. Pabo, *Science* 252:810–817, 1991.)

Another type of zinc finger is found in the large family of intracellular receptor proteins (discussed in detail in Chapter 15). It forms a different type of structure (similar in some respects to the helix–turn–helix motif) in which two α helices are packed together with zinc atoms (Figure 7–19). Like the helix–turn– helix proteins, these proteins usually form dimers that allow one of the two α helices of each subunit to interact with the major groove of the DNA (see Figure 7–14). Although the two types of zinc finger structures discussed in this section are structurally distinct, they share two important features: both use zinc as a structural element, and both use an α helix to recognize the major groove of the DNA.

## β sheets Can Also Recognize DNA

In the DNA-binding motifs discussed so far, α helices are the primary mechanism used to recognize specific DNA sequences. One group of gene regulatory proteins, however, has evolved an entirely different and no less ingenious recognition strategy. In this case the information on the surface of the major groove is read by a two-stranded β sheet, with side chains of the amino acids extending from the sheet toward the DNA as shown in Figure 7–20. As in the case of a

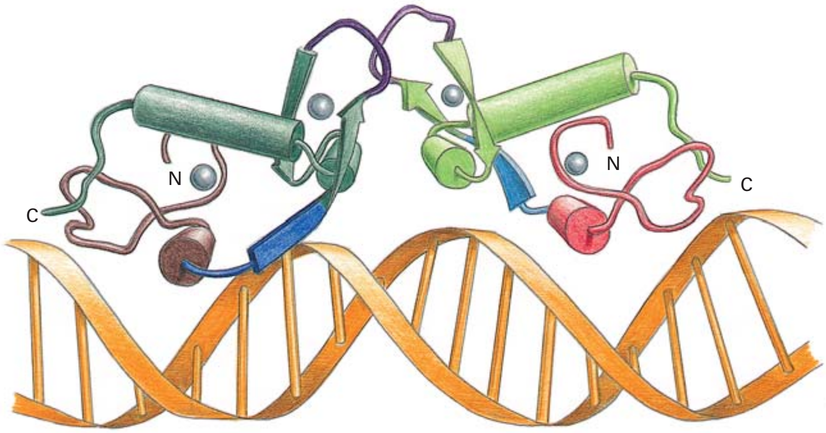


**Figure 7–19 A dimer of the zinc finger domain of the intracellular receptor family bound to its specific DNA sequence.** Each zinc finger domain contains two atoms of Zn (indicated by the small *gray spheres);* one stabilizes the DNA recognition helix (shown in *brown* in one subunit and *red* in the other), and one stabilizes a loop (shown in *purple)* involved in dimer formation. Each Zn atom is coordinated by four appropriately spaced cysteine residues. Like the helix–turn–helix proteins shown in Figure 7–14, the two recognition helices of the dimer are held apart by a distance corresponding to one turn of the DNA double helix. The specific example shown is a fragment of the glucocorticoid receptor. This is the protein through which cells detect and respond transcriptionally to the glucocorticoid hormones produced in the adrenal gland in response to stress. (Adapted from B.F. Luisi et al., *Nature* 352:497–505, 1991.)

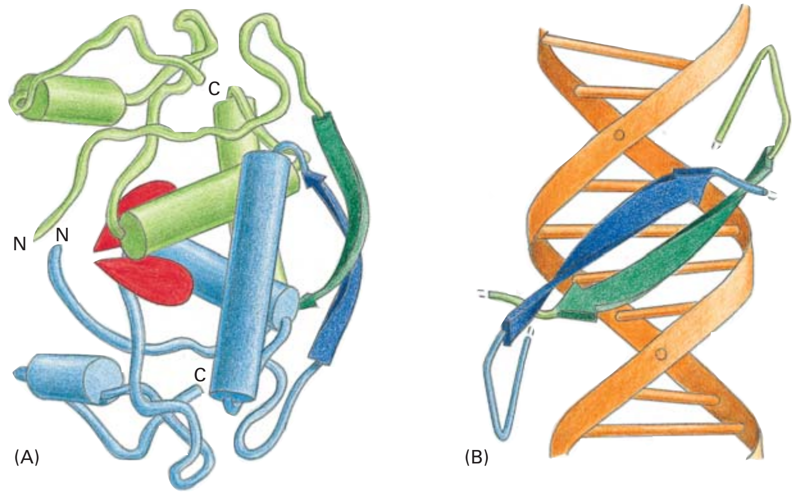


**Figure 7–20 The bacterial *met* repressor protein.** The bacterial *met* repressor regulates the genes encoding the enzymes that catalyze methionine synthesis. When this amino acid is abundant, it binds to the repressor, causing a change in the structure of the protein that enables it to bind to DNA tightly, shutting off the synthesis of the enzymes. (A) In order to bind to DNA tightly, the *met* repressor must be complexed with S-adenosyl methionine, shown in *red*. One subunit of the dimeric protein is shown in *green*, while the other is shown in *blue*. The two-stranded β sheet that binds to DNA is formed by one strand from each subunit and is shown in *dark green* and *dark blue*. (B) Simplified diagram of the *met* repressor bound to DNA, showing how the two-stranded β sheet of the repressor binds to the major groove of DNA. For clarity, the other regions of the repressor have been omitted. (A, adapted from S. Phillips, *Curr. Opin. Struct. Biol.* 1:89–98, 1991; B, adapted from W. Somers and S. Phillips, *Nature* 359:387–393, 1992.)

recognition α helix, this β-sheet motif can be used to recognize many different DNA sequences; the exact DNA sequence recognized depends on the sequence of amino acids that make up the β sheet.

## The Leucine Zipper Motif Mediates Both DNA Binding and Protein Dimerization

Many gene regulatory proteins recognize DNA as homodimers, probably because, as we have seen, this is a simple way of achieving strong specific binding (see Figure 7–15). Usually, the portion of the protein responsible for dimerization is distinct from the portion that is responsible for DNA binding (see Figure 7–14). One motif, however, combines these two functions in an elegant and economical way. It is called the **leucine zipper motif**, so named because of the way the two α helices, one from each monomer, are joined together to form a short coiled-coil (see Figure 3–11). The helices are held together by interactions between hydrophobic amino acid side chains (often on leucines) that extend from one side of each helix. Just beyond the dimerization interface the two α helices separate from each other to form a Y-shaped structure, which allows their side chains to contact the major groove of DNA. The dimer thus grips the double helix like a clothespin on a clothesline (Figure 7–21).

## Heterodimerization Expands the Repertoire of DNA Sequences Recognized by Gene Regulatory Proteins

Many of the gene regulatory proteins we have seen thus far bind DNA as homodimers, that is, dimers made up of two identical subunits. However, many gene regulatory proteins, including leucine zipper proteins, can also associate with nonidentical partners to form heterodimers composed of two different subunits. Because heterodimers typically form from two proteins with distinct DNA-binding specificities, the mixing and matching of gene regulatory proteins to form heterodimers greatly expands the repertoire of DNA-binding specificities that these proteins can display. As illustrated in Figure 7–22, three distinct DNA-binding specificities could, in principle, be generated from two types of

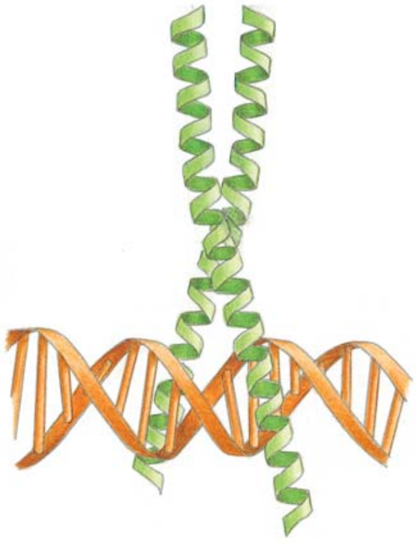


**Figure 7–21 A leucine zipper dimer bound to DNA.** Two α-helical DNA-binding domains (*bottom*) dimerize through their α-helical leucine zipper region (*top*) to form an inverted Y-shaped structure. Each arm of the Y is formed by a single α helix, one from each monomer, that mediates binding to a specific DNA sequence in the major groove of DNA. Each α helix binds to one-half of a symmetric DNA structure. The structure shown is of the yeast Gcn4 protein, which regulates transcription in response to the availability of amino acids in the environment. (Adapted from T.E. Ellenberger et al., *Cell* 71:1223–1237, 1992.)

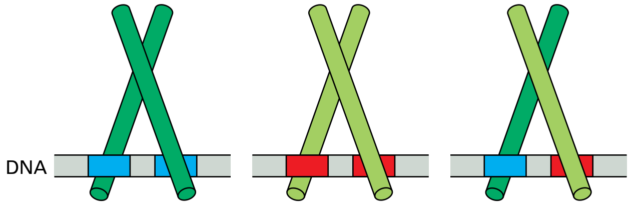


**Figure 7–22 Heterodimerization of leucine zipper proteins can alter their DNA-binding specificity.** Leucine zipper homodimers bind to symmetric DNA sequences, as shown in the left-hand and center drawings. These two proteins recognize different DNA sequences, as indicated by the *red* and *blue* regions in the DNA. The two different monomers can combine to form a heterodimer, which now recognizes a hybrid DNA sequence, composed from one *red* and one *blue* region.

leucine zipper monomer, while six could be created from three types of monomer, and so on.

There are, however, limits to this promiscuity: if all the many types of leucine zipper proteins in a typical eucaryotic cell formed heterodimers, the amount of "cross-talk" between the gene regulatory circuits of a cell would be so great as to cause chaos. Whether or not a particular heterodimer can form depends on how well the hydrophobic surfaces of the two leucine zipper α helices mesh with each other, which, in turn, depends on the exact amino acid sequences of the two zipper regions. Thus each leucine zipper protein in the cell can form dimers with only a small set of other leucine zipper proteins.

Heterodimerization is an example of **combinatorial control**, in which combinations of different proteins, rather than individual proteins, control a cellular process. Heterodimerization is one of the mechanisms used by eucaryotic cells to control gene expression in this way, and it occurs in a wide variety of different types of gene regulatory proteins (Figure 7–23). As we discuss later, however, the formation of heterodimeric gene regulatory complexes is only one of several combinatorial mechanisms for controlling gene expression.

During the evolution of gene regulatory proteins, similar combinatorial principles have produced new DNA-binding specificities by joining two distinct DNA-binding domains into a single polypeptide chain (Figure 7–24).

## The Helix–Loop–Helix Motif Also Mediates Dimerization and DNA Binding

Another important DNA-binding motif, related to the leucine zipper, is the **helix–loop–helix (HLH) motif**, which should not be confused with the

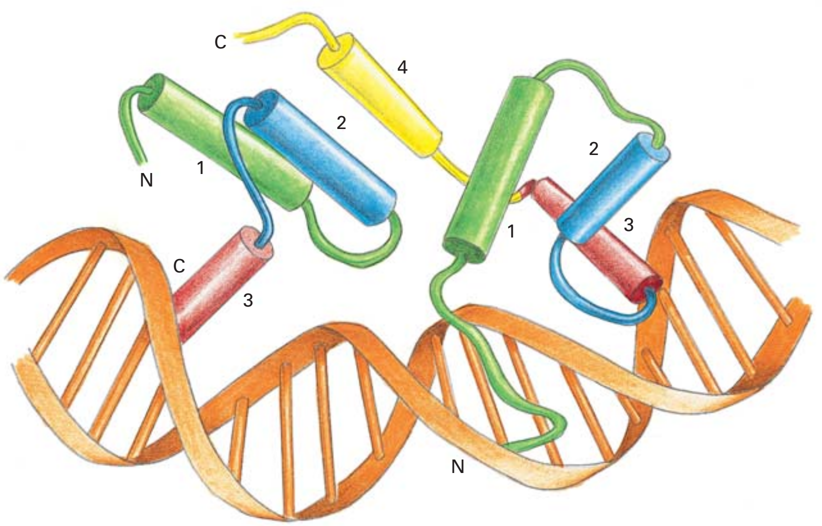


**Figure 7–23 A heterodimer composed of two homeodomain proteins bound to its DNA recognition site.** The *yellow* helix 4 of the protein on the right (Matα2) is unstructured in the absence of the protein on the left (Mata1), forming a helix only upon heterodimerization. The DNA sequence is thereby recognized jointly by both proteins; some of the protein–DNA contacts made by Matα2 were shown in Figure 7–16. These two proteins are from budding yeast, where the heterodimer specifies a particular cell type (see Figure 7–65). The helices are numbered in accordance with Figure 7–16. (Adapted from T. Li et al., *Science* 270:262–269, 1995.)

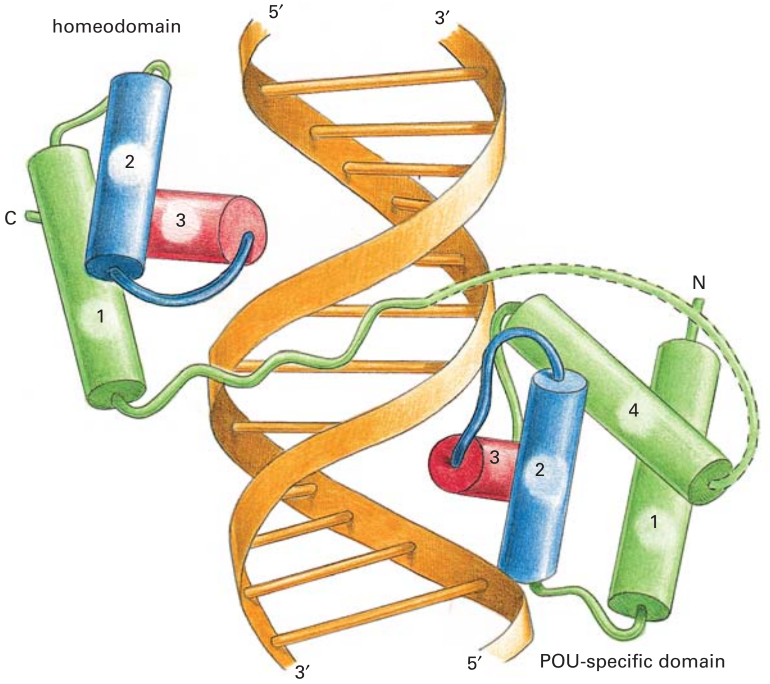


**Figure 7–24 Two DNA-binding domains covalently joined by a flexible polypeptide.** The structure shown (called a POU-domain) consists of both a homeodomain and a helix–turn–helix structure (closely related to the bacteriophage λ repressor—see Figure 7–14) joined by a flexible polypeptide "leash," indicated by the broken lines. The entire protein is encoded in a single gene and is synthesized as a continuous polypeptide chain. The covalent joining of two structures in this way results in a large increase in the affinity of the protein for its specific DNA sequence compared with the DNA affinity of either separate structure. The group of mammalian gene regulatory proteins exemplified by this structure regulate the production of growth factors, immunoglobulins, and other molecules involved in development. The particular example shown is from the Oct-1 protein. (Adapted from J.D. Klemm et al., *Cell* 77:21–32, 1994.)

helix–turn–helix motif discussed earlier. An HLH motif consists of a short α helix connected by a loop to a second, longer α helix. The flexibility of the loop allows one helix to fold back and pack against the other. As shown in Figure 7–25, this two-helix structure binds both to DNA and to the HLH motif of a second HLH protein. As with leucine zipper proteins, the second HLH protein can be the same (creating a homodimer) or different (creating a heterodimer). In either case, two α helices that extend from the dimerization interface make specific contacts with the DNA.

Several HLH proteins lack the α-helical extension responsible for binding to DNA. These truncated proteins can form heterodimers with full-length HLH proteins, but the heterodimers are unable to bind DNA tightly because they form only half of the necessary contacts. Thus, in addition to creating active dimers, heterodimerization provides a way to hold specific gene regulatory proteins in check (Figure 7–26).

## It Is Not Yet Possible to Accurately Predict the DNA Sequences Recognized by All Gene Regulatory Proteins

The various DNA-binding motifs that we have discussed provide structural frameworks from which specific amino acid side chains extend to contact specific base pairs in the DNA. It is reasonable to ask, therefore, whether there is a

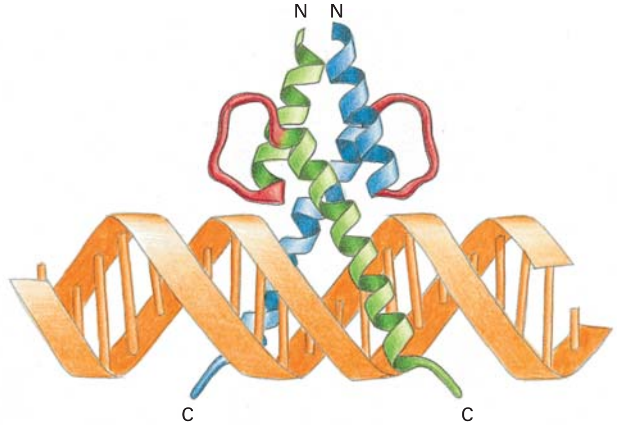


**Figure 7–25 A helix–loop–helix dimer bound to DNA.** The two monomers are held together in a four-helix bundle: each monomer contributes two α helices connected by a flexible loop of protein (*red*). A specific DNA sequence is bound by the two α helices that project from the four-helix bundle. (Adapted from A.R. Ferre-D'Amare et al., *Nature* 363:38–45, 1993.)

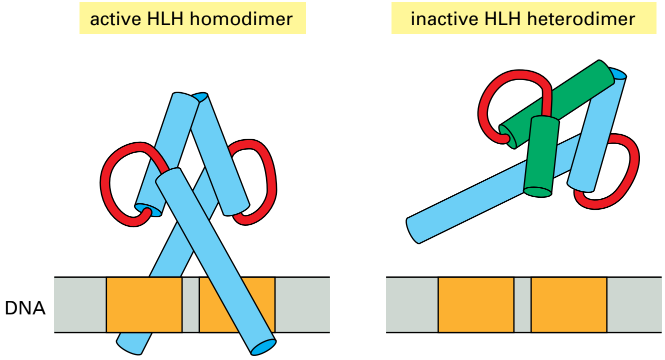


**Figure 7–26 Inhibitory regulation by truncated HLH proteins.** The HLH motif is responsible for both dimerization and DNA binding. On the *left*, an HLH homodimer recognizes a symmetric DNA sequence. On the *right*, the binding of a full-length HLH protein *(blue)* to a truncated HLH protein *(green)* that lacks the DNA-binding α helix generates a heterodimer that is unable to bind DNA tightly. If present in excess, the truncated protein molecule blocks the homodimerization of the full-length HLH protein and thereby prevents it from binding to DNA.

simple amino acid–base pair recognition code: is a G–C base pair, for example, always contacted by a particular amino acid side chain? The answer appears to be no, although certain types of amino acid-base interactions appear much more frequently than others (Figure 7–27). As we saw in Chapter 3, protein surfaces of virtually any shape and chemistry can be made from just 20 different amino acids, and a gene regulatory protein uses different combinations of these to create a surface that is precisely complementary to a particular DNA sequence. We know that the same base pair can thereby be recognized in many ways depending on its context (Figure 7–28). Nevertheless, molecular biologists are beginning to understand protein–DNA recognition well enough that we should soon be able to design proteins that will recognize any desired DNA sequence.

## A Gel-Mobility Shift Assay Allows Sequence-specific DNA-binding Proteins to Be Detected Readily

Genetic analyses, which provided a route to the gene regulatory proteins of bacteria, yeast, and *Drosophila*, is much more difficult in vertebrates. Therefore, the isolation of vertebrate gene regulatory proteins had to await the development of different approaches. Many of these approaches rely on the detection in a cell extract of a DNA-binding protein that specifically recognizes a DNA sequence known to control the expression of a particular gene. The most common way to detect sequence-specific DNA-binding proteins is to use a technique that is based on the effect of a bound protein on the migration of DNA molecules in an electric field.

A DNA molecule is highly negatively charged and will therefore move rapidly toward a positive electrode when it is subjected to an electric field. When analyzed by polyacrylamide-gel electrophoresis, DNA molecules are separated according to their size because smaller molecules are able to penetrate the fine gel meshwork more easily than large ones. Protein molecules bound to a DNA molecule will cause it to move more slowly through the gel; in general, the larger the bound protein, the greater the retardation of the DNA molecule. This phenomenon provides the basis for the **gel-mobility shift assay**, which allows even trace amounts of a sequence-specific DNA-binding protein to be readily detected. In this assay, a short DNA fragment of specific length and sequence (produced either by DNA cloning or by chemical synthesis) is radioactively labeled and mixed with a cell extract; the mixture is then loaded onto a polyacrylamide gel and subjected to electrophoresis. If the DNA fragment corresponds to a chromosomal region where, for example, several sequence-specific

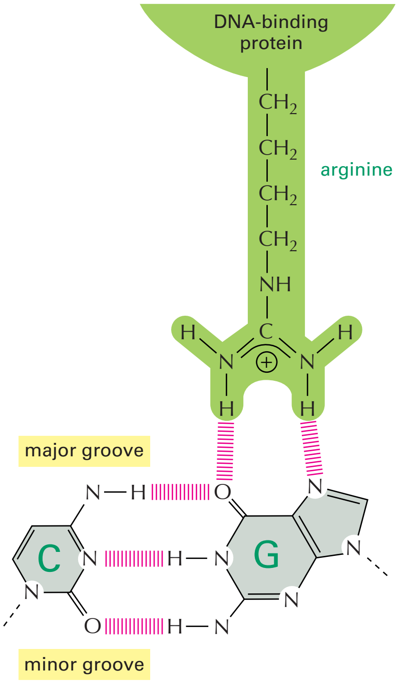


**Figure 7–27 One of the most common protein–DNA interactions.** Because of its specific geometry of hydrogen-bond acceptors (see Figure 7–7), guanine can be unambiguously recognized by the side chain of arginine. Another common protein–DNA interaction was shown in Figure 7–12.

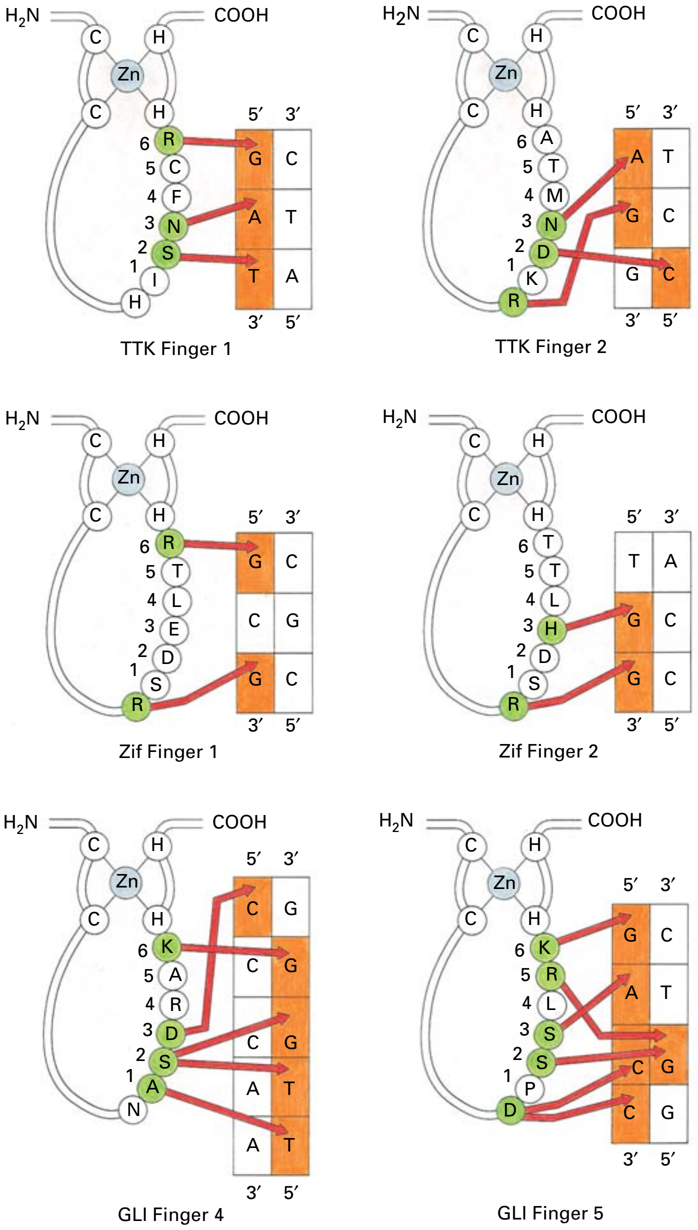


**Figure 7–28 Summary of sequence-specific interactions between different six zinc fingers and their DNA recognition sequences.** Even though all six Zn fingers have the same overall structure (see Figure 7–17), each binds to a different DNA sequence. The numbered amino acids form the $\alpha$ helix that recognizes DNA (Figures 7–17 and 7–18), and those that make sequence-specific DNA contacts are colored *green*. Bases contacted by protein are *orange*. Although arginine–guanine contacts are common (see Figure 7–27), guanine can also be recognized by serine, histidine, and lysine, as shown. Moreover, the same amino acid (serine, in this example) can recognize more than one base. Two of the Zn fingers depicted are from the TTK protein (a *Drosophila* protein that functions in development); two are from the mouse protein (Zif268) that was shown in Figure 7–18; and two are from a human protein (GLI), whose aberrant forms can cause certain types of cancers. (Adapted from C. Branden and J. Tooze, Introduction to Protein Structure, 2nd edn. New York: Garland Publishing, 1999.)

proteins bind, autoradiography will reveal a series of DNA bands, each retarded to a different extent and representing a distinct DNA–protein complex. The proteins responsible for each band on the gel can then be separated from one another by subsequent fractionations of the cell extract (Figure 7–29).

## DNA Affinity Chromatography Facilitates the Purification of Sequence-specific DNA-binding Proteins

A particularly powerful purification method called **DNA affinity chromatography** can be used once the DNA sequence that a gene regulatory protein recognizes has been determined. A double-stranded oligonucleotide of the correct sequence is synthesized by chemical methods and linked to an insoluble porous matrix such as agarose; the matrix with the oligonucleotide attached is then used to construct a column that selectively binds proteins that recognize the particular DNA sequence (Figure 7–30). Purifications as great as 10,000-fold can be achieved by this means with relatively little effort.

Although most proteins that bind to a specific DNA sequence are present in a few thousand copies per higher eucaryotic cell (and generally represent only

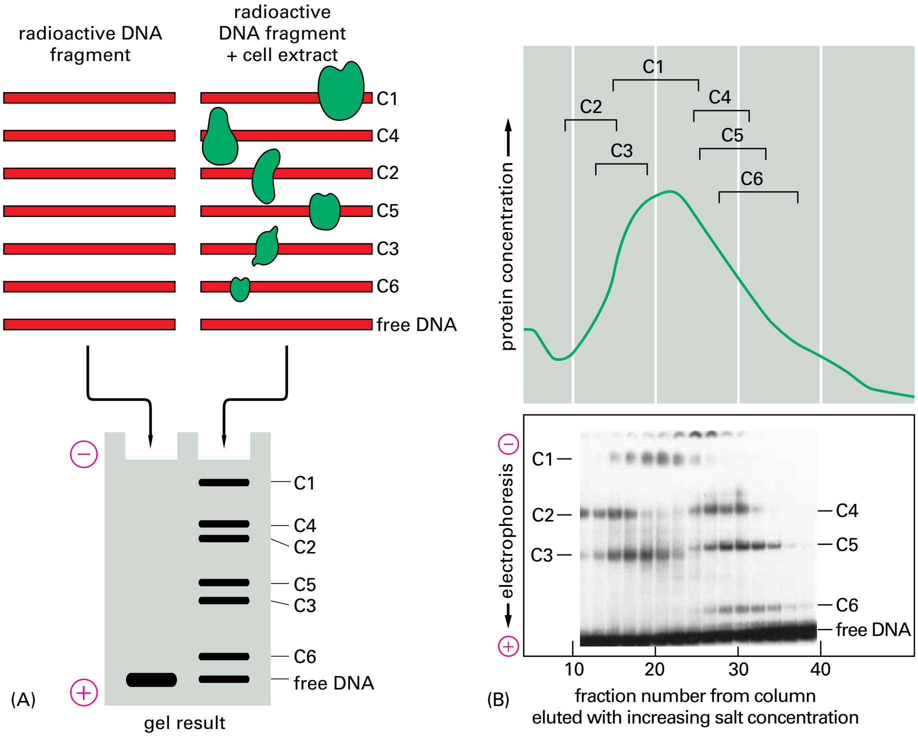


**Figure 7–29 A gel-mobility shift assay.** The principle of the assay is shown schematically in (A). In this example an extract of an antibody-producing cell line is mixed with a radioactive DNA fragment containing about 160 nucleotides of a regulatory DNA sequence from a gene encoding the light chain of the antibody made by the cell line. The effect of the proteins in the extract on the mobility of the DNA fragment is analyzed by polyacrylamide-gel electrophoresis followed by autoradiography. The free DNA fragments migrate rapidly to the bottom of the gel, while those fragments bound to proteins are retarded; the finding of six retarded bands suggests that the extract contains six different sequence-specific DNA-binding proteins (indicated as C1–C6) that bind to this DNA sequence. (For simplicity, any DNA fragments with more than one protein bound have been omitted from the figure.) In (B) the extract was fractionated by a standard chromatographic technique *(top)*, and each fraction was mixed with the radioactive DNA fragment, applied to one lane of a polyacrylamide gel, and analyzed as in (A). (B, modified from C. Scheidereit, A. Heguy, and R.G. Roeder, *Cell* 51:783–793, 1987.)

about one part in 50,000 of the total cell protein), enough pure protein can usually be isolated by affinity chromatography to obtain a partial amino acid sequence by mass spectrometry or other means (discussed in Chapter 8). If the complete genome sequence of the organism is known, the partial amino acid sequence can be used to identify the gene. The gene provides the complete amino acid sequence of the protein, and any uncertainties regarding exon and intron boundaries can be resolved by analyzing the mRNA produced by the gene, as described in Chapter 8. The gene also provides the means to produce the protein in unlimited amounts through genetic engineering techniques, as discussed in Chapter 8).

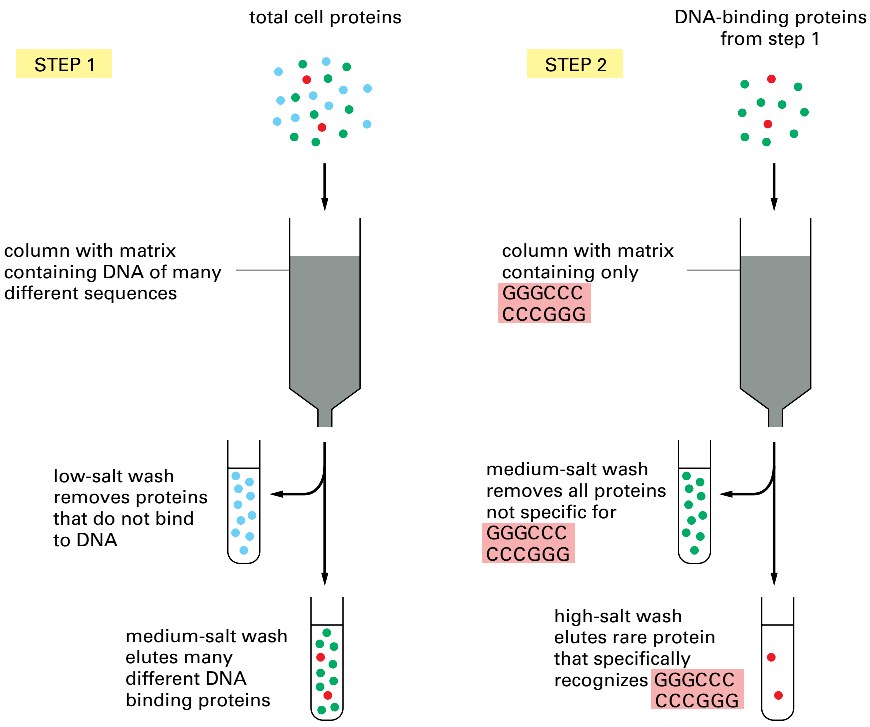

total cell proteins
STEP 1
column with matrix containing DNA of many different sequences
low-salt wash removes proteins that do not bind to DNA
medium-salt wash elutes many different DNA binding proteins
DNA-binding proteins from step 1
STEP 2
column with matrix containing only
GGGCCC
CCCGGG
medium-salt wash removes all proteins not specific for
GGGCCC
CCCGGG
high-salt wash elutes rare protein that specifically recognizes GGGCCC
CCCGGG

**Figure 7–30 DNA affinity chromatography.** In the first step, all the proteins that can bind DNA are separated from the remainder of the cellular proteins on a column containing a huge number of different DNA sequences. Most sequence-specific DNA-binding proteins have a weak (nonspecific) affinity for bulk DNA and are therefore retained on the column. This affinity is due largely to ionic attractions, and the proteins can be washed off the DNA by a solution that contains a moderate concentration of salt. In the second step, the mixture of DNA-binding proteins is passed through a column that contains only DNA of a particular sequence. Typically, all the DNA-binding proteins will stick to the column, the great majority by nonspecific interactions. These are again eluted by solutions of moderate salt concentration, leaving on the column only those proteins (typically one or only a few) that bind specifically and therefore very tightly to the particular DNA sequence. These remaining proteins can be eluted from the column by solutions containing a very high concentration of salt.

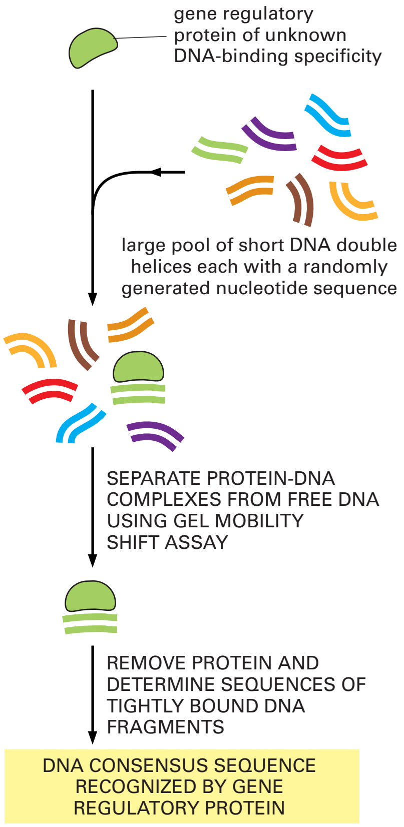

gene regulatory protein of unknown DNA-binding specificity
large pool of short DNA double helices each with a randomly generated nucleotide sequence
SEPARATE PROTEIN-DNA COMPLEXES FROM FREE DNA USING GEL MOBILITY SHIFT ASSAY
REMOVE PROTEIN AND DETERMINE SEQUENCES OF TIGHTLY BOUND DNA FRAGMENTS
DNA CONSENSUS SEQUENCE RECOGNIZED BY GENE REGULATORY PROTEIN

**Figure 7–31 A method for determining the DNA sequence recognized by a gene regulatory protein.** A purified gene regulatory protein is mixed with millions of different short DNA fragments, each with a different sequence of nucleotides. A collection of such DNA fragments can be produced by programming a DNA synthesizer, a machine that chemically synthesizes DNA of any desired sequence (discussed in Chapter 8). For example, there are $4^{11}$, or approximately 4.2 million possible sequences for a DNA fragment of 11 nucleotides. The double-stranded DNA fragments that bind tightly to the gene regulatory protein are then separated from the DNA fragments that fail to bind. One method for accomplishing this separation is through gel-mobility shifts, as described in Figure 7–29. After separation of the DNA–protein complexes from the free DNA, the DNA fragments are removed from the protein, and several additional rounds of the same selection process are carried out. The nucleotide sequences of those DNA fragments that remain through multiple rounds of selection can be determined, and a consensus DNA recognition sequence can be generated.

## The DNA Sequence Recognized by a Gene Regulatory Protein Can Be Determined

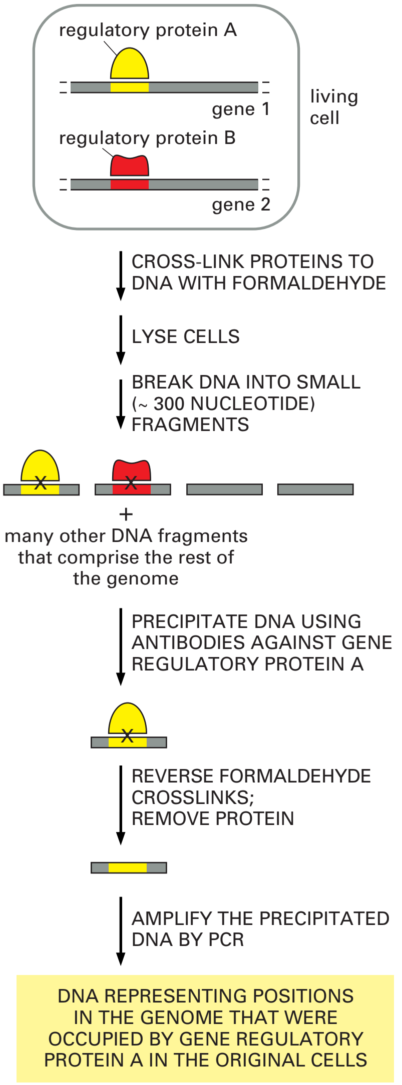
Some gene regulatory proteins were discovered before the DNA sequence to which they bound was known. For example, many of the *Drosophila* homeodomain proteins were discovered through the isolation of mutations that altered fly development. This allowed the genes encoding the proteins to be identified, and the proteins could then be over-expressed in cultured cells and easily purified. One method of determining the DNA sequences recognized by a gene regulatory protein is to use the purified protein to select out from a large pool of short nucleotides of differing sequence only those that bind tightly to it. After several rounds of selection, the nucleotide sequences of the tightly bound DNAs can be determined, and a consensus DNA recognition sequence for the gene regulatory protein can be formulated (Figure 7–31). The consensus sequence can be used to search genome sequences by computer and thereby identify candidate genes whose transcription might be regulated by the gene regulatory protein of interest. However, this strategy is not foolproof. For example, many organisms produce a set of closely related gene regulatory proteins that recognize very similar DNA sequences, and this approach cannot resolve them. In most cases, predictions of the sites of action of gene regulatory proteins obtained from searching genome sequences must be tested by more direct approaches, such as the one described in the next section.

## A Chromatin Immunoprecipitation Technique Identifies DNA Sites Occupied by Gene Regulatory Proteins in Living Cells

In general, a given gene regulatory protein does not occupy all its potential DNA-binding sites in the genome all the time. Under some conditions, the protein may simply not be synthesized, and so be absent from the cell; or, for example, it may be present but may have to form a heterodimer with another protein to bind DNA efficiently in a living cell; or it may be excluded from the nucleus until an appropriate signal is received from the cell's environment. One method for empirically determining the sites on DNA occupied by a given gene regulatory protein under a particular set of conditions is called **chromatin immunoprecipitation** (Figure 7–32). Proteins are covalently cross-linked to DNA in living cells, the cells are lysed, and the DNA is mechanically broken into small fragments. Then, antibodies directed against a given gene regulatory protein are used to purify DNA that was covalently cross-linked to the gene regulatory protein due to the protein's close proximity to that DNA at the time of cross-linking. In this way, the DNA sites occupied by the gene regulatory protein in the original cells can be determined.

This method is also routinely used to identify the positions along a genome that are packaged by the various types of modified histones (see Figure 4–35). In this case, antibodies specific to a particular histone modification are employed.

**Figure 7–32 Chromatin immunoprecipitation.** This methodology allows the identification of the sites in a genome that are occupied *in vivo* by a gene regulatory protein. The amplification of DNA by the polymerase chain reaction (PCR) is described in Chapter 8. The identities of the precipitated, amplified DNA fragments can be determined by hybridizing the mixture of fragments to DNA microarrays, described in Chapter 8.


regulatory protein A
gene 1
regulatory protein B
gene 2
living cell
CROSS-LINK PROTEINS TO DNA WITH FORMALDEHYDE
LYSE CELLS
BREAK DNA INTO SMALL (~ 300 NUCLEOTIDE) FRAGMENTS
+
many other DNA fragments that comprise the rest of the genome
PRECIPITATE DNA USING ANTIBODIES AGAINST GENE REGULATORY PROTEIN A
REVERSE FORMALDEHYDE CROSSLINKS; REMOVE PROTEIN
AMPLIFY THE PRECIPITATED DNA BY PCR
DNA REPRESENTING POSITIONS IN THE GENOME THAT WERE OCCUPIED BY GENE REGULATORY PROTEIN A IN THE ORIGINAL CELLS

## Summary

***Gene regulatory proteins recognize short stretches of double-helical DNA of defined sequence and thereby determine which of the thousands of genes in a cell will be transcribed. Thousands of gene regulatory proteins have been identified in a wide variety of organisms. Although each of these proteins has unique features, most bind to DNA as homodimers or heterodimers and recognize DNA through one of a small number of structural motifs. The common motifs include the helix–turn–helix, the homeodomain, the leucine zipper, the helix–loop–helix, and zinc fingers of several types. The precise amino acid sequence that is folded into a motif determines the particular DNA sequence that is recognized. Heterodimerization increases the range of DNA sequences that can be recognized by gene regulatory proteins. Powerful techniques are available that make use of the DNA-sequence specificity of gene regulatory proteins to identify and isolate these proteins, the genes that encode them, the DNA sequences they recognize, and the genes that they regulate.***

# HOW GENETIC SWITCHES WORK

In the previous section, we described the basic components of genetic switches—gene regulatory proteins and the specific DNA sequences that these proteins recognize. We shall now discuss how these components operate to turn genes on and off in response to a variety of signals.

Only 40 years ago the idea that genes could be switched on and off was revolutionary. This concept was a major advance, and it came originally from the study of how *E. coli* bacteria adapt to changes in the composition of their growth medium. Parallel studies on the lambda bacteriophage led to many of the same conclusions and helped to establish the underlying mechanism. Many of the same principles apply to eucaryotic cells. However, the enormous complexity of gene regulation in higher organisms, combined with the packaging of their DNA into chromatin, creates special challenges and some novel opportunities for control—as we shall see. We begin with the simplest example—an on-off switch in bacteria that responds to a single signal.

## The Tryptophan Repressor Is a Simple Switch That Turns Genes On and Off in Bacteria

The chromosome of the bacterium *E. coli*, a single-celled organism, consists of a single circular DNA molecule of about 4.6 × $10^6$ nucleotide pairs. This DNA encodes approximately 4300 proteins, although only a fraction of these are made at any one time. The expression of many of them is regulated according to the available food in the environment. This is illustrated by the five *E. coli* genes that code for enzymes that manufacture the amino acid tryptophan. These genes are arranged as a single **operon;** that is, they are adjacent to one another on the chromosome and are transcribed from a single *promoter* as one long mRNA molecule (Figure 7–33). But when tryptophan is present in the growth

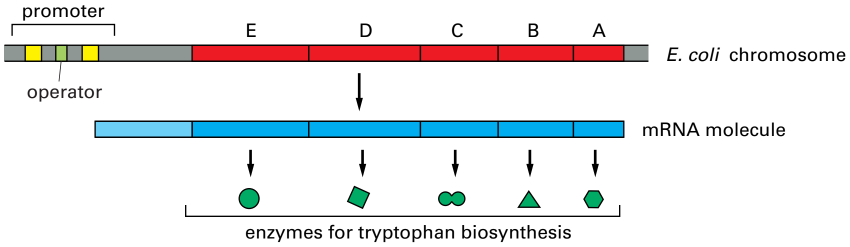
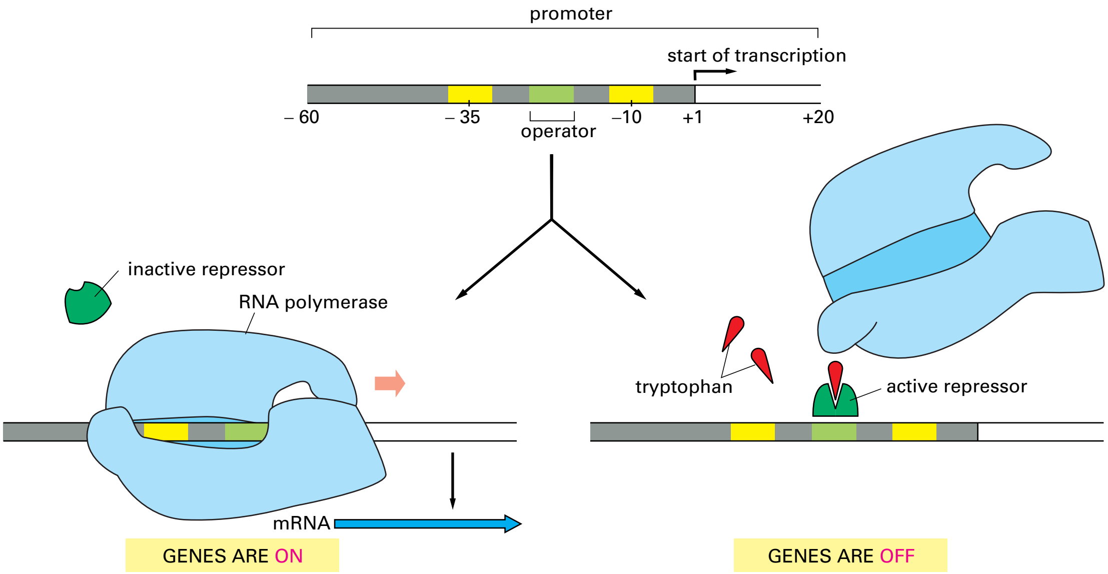


**Figure 7–33 The clustered genes in *E. coli* that code for enzymes that manufacture the amino acid tryptophan.** These five genes are transcribed as a single mRNA molecule, a feature that allows their expression to be controlled coordinately. Clusters of genes transcribed as a single mRNA molecule are common in bacteria. Each such cluster is called an operon.



medium and enters the cell (when the bacterium is in the gut of a mammal that has just eaten a meal of protein, for example), the cell no longer needs these enzymes and shuts off their production.

The molecular basis for this switch is understood in considerable detail. As described in Chapter 6, a promoter is a specific DNA sequence that directs RNA polymerase to bind to DNA, to open the DNA double helix, and to begin synthesizing an RNA molecule. Within the promoter that directs transcription of the tryptophan biosynthetic genes lies a regulating element called an **operator** (see Figure 7–33). This is simply a short region of regulatory DNA of defined nucleotide sequence that is recognized by a repressor protein, in this case the **tryptophan repressor**, a member of the helix–turn–helix family (see Figure 7–14). The promoter and operator are arranged so that when the tryptophan repressor occupies the operator, it blocks access to the promoter by RNA polymerase, thereby preventing expression of the tryptophan-producing enzymes (Figure 7–34).

**Figure 7–34 Switching the tryptophan genes on and off.** If the level of tryptophan inside the cell is low, RNA polymerase binds to the promoter and transcribes the five genes of the tryptophan *(trp)* operon. If the level of tryptophan is high, however, the tryptophan repressor is activated to bind to the operator, where it blocks the binding of RNA polymerase to the promoter. Whenever the level of intracellular tryptophan drops, the repressor releases its tryptophan and becomes inactive, allowing the polymerase to begin transcribing these genes. The promoter includes two key blocks of DNA sequence information, the –35 and –10 regions highlighted in *yellow* (see Figure 6–12).

The block to gene expression is regulated in an ingenious way: to bind to its operator DNA, the repressor protein has to have two molecules of the amino acid tryptophan bound to it. As shown in Figure 7–35, tryptophan binding tilts the helix–turn–helix motif of the repressor so that it is presented properly to the DNA major groove; without tryptophan, the motif swings inward and the protein is unable to bind to the operator. Thus the tryptophan repressor and operator form a simple device that switches production of the tryptophan biosynthetic enzymes on and off according to the availability of free tryptophan. Because the active, DNA-binding form of the protein serves to turn genes off, this mode of gene regulation is called **negative control**, and the gene regulatory proteins that function in this way are called *transcriptional repressors* or *gene repressor proteins*.

## Transcriptional Activators Turn Genes On

We saw in Chapter 6 that purified *E. coli* RNA polymerase (including the σ subunit) can bind to a promoter and initiate DNA transcription. Some bacterial promoters, however, are only marginally functional on their own, either because they are recognized poorly by RNA polymerase or because the polymerase has difficulty opening the DNA helix and beginning transcription. In either case these poorly functioning promoters can be rescued by gene regulatory proteins that bind to a nearby site on the DNA and contact the RNA polymerase in a way that dramatically increases the probability that a transcript will be initiated. Because the active, DNA-binding form of such a protein turns genes on, this mode of gene regulation is called **positive control**, and the gene regulatory proteins that function in this manner are known as *transcriptional activators* or *gene activator proteins*. In some cases, bacterial gene activator proteins aid RNA polymerase in binding to the promoter by providing an additional contact surface for the polymerase. In other cases, they facilitate the transition from the initial DNA-bound conformation of polymerase to the actively transcribing form, perhaps by stabilizing a transition state.

As in negative control by a transcriptional repressor, a transcriptional activator can operate as part of a simple on–off genetic switch. The bacterial activator protein *CAP (catabolite activator protein),* for example, activates genes that enable *E. coli* to use alternative carbon sources when glucose, its preferred carbon source, is not available. Falling levels of glucose induce an increase in the intracellular signaling molecule cyclic AMP, which binds to the CAP protein, enabling it to bind to its specific DNA sequence near target promoters and thereby turn on the appropriate genes. In this way the expression of a target gene is switched on or off, depending on whether cyclic AMP levels in the cell are high or low, respectively. Figure 7–36 summarizes the different ways that positive and negative control can be used to regulate genes.

In many respects transcriptional activators and transcriptional repressors are similar in design. The tryptophan repressor and the transcriptional activator CAP, for example, both use a helix–turn–helix motif (see Figure 7–14) and both require a small cofactor in order to bind DNA. In fact, some bacterial proteins (including CAP and the bacteriophage lambda repressor) can act as either activators or repressors, depending on the exact placement of the DNA sequence they recognize in relation to the promoter: if the binding site for the protein overlaps the promoter, the polymerase cannot bind and the protein acts as a repressor (Figure 7–37).

## A Transcriptional Activator and a Transcriptional Repressor Control the *lac* Operon

More complicated types of genetic switches combine positive and negative controls. The *lac operon* in *E. coli*, for example, unlike the *trp operon*, is under both negative and positive transcriptional controls by the lac repressor protein and CAP, respectively. The *lac* operon codes for proteins required to transport the disaccharide lactose into the cell and to break it down. CAP, as we have seen, enables bacteria to use alternative carbon sources such as lactose in the absence

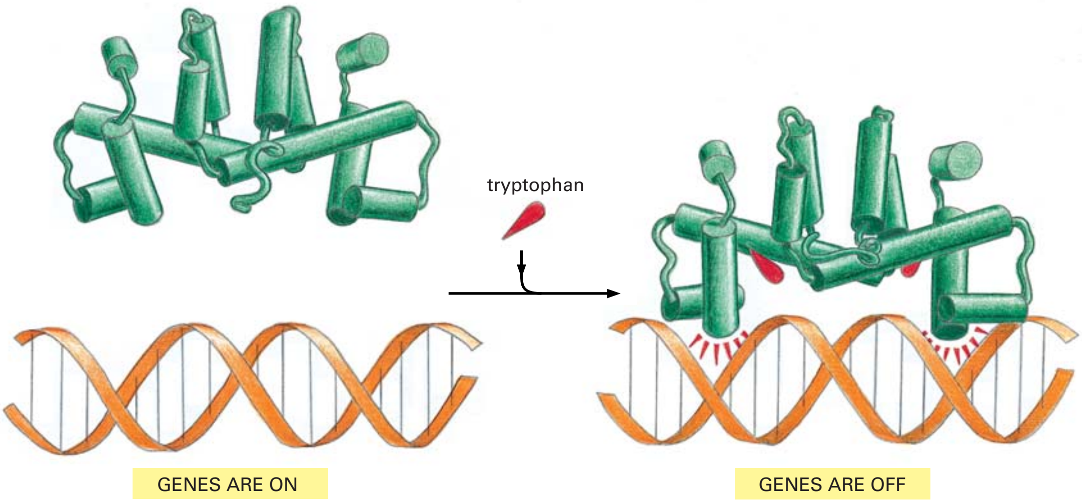


**Figure 7–35 The binding of tryptophan to the tryptophan repressor protein changes the conformation of the repressor.** The conformational change enables this gene regulatory protein to bind tightly to a specific DNA sequence (the operator), thereby blocking transcription of the genes encoding the enzymes required to produce tryptophan (the *trp* operon). The three-dimensional structure of this bacterial helix–turn–helix protein, as determined by x-ray diffraction with and without tryptophan bound, is illustrated. Tryptophan binding increases the distance between the two recognition helices in the homodimer, allowing the repressor to fit snugly on the operator. (Adapted from R. Zhang et al., *Nature* 327:591–597, 1987.)

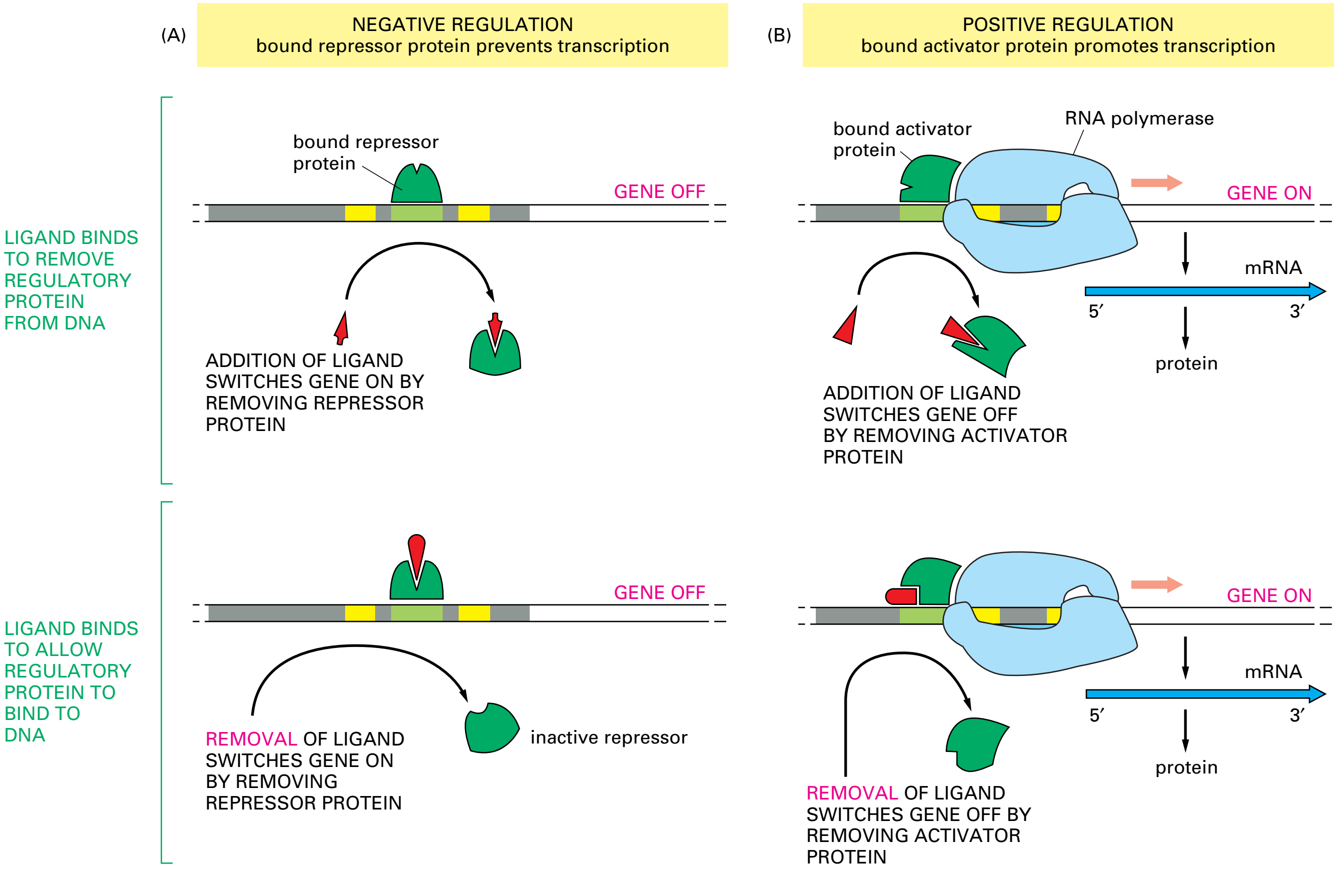

(A)
NEGATIVE REGULATION
bound repressor protein prevents transcription
(B)
POSITIVE REGULATION
bound activator protein promotes transcription
LIGAND BINDS TO REMOVE REGULATORY PROTEIN FROM DNA
bound repressor protein
GENE OFF
ADDITION OF LIGAND SWITCHES GENE ON BY REMOVING REPRESSOR PROTEIN
bound activator protein
RNA polymerase
GENE ON
mRNA
5′
3′
protein
ADDITION OF LIGAND SWITCHES GENE OFF BY REMOVING ACTIVATOR PROTEIN
LIGAND BINDS TO ALLOW REGULATORY PROTEIN TO BIND TO DNA
GENE OFF
REMOVAL OF LIGAND SWITCHES GENE ON BY REMOVING REPRESSOR PROTEIN
inactive repressor
GENE ON
mRNA
5′
3′
protein
REMOVAL OF LIGAND SWITCHES GENE OFF BY REMOVING ACTIVATOR PROTEIN

**Figure 7–36 Summary of the mechanisms by which specific gene regulatory proteins control gene transcription in procaryotes.** (A) Negative regulation; (B) positive regulation. Note that the addition of an "inducing" ligand can turn on a gene either by removing a gene repressor protein from the DNA *(upper left panel)* or by causing a gene activator protein to bind *(lower right panel)*. Likewise, the addition of an "inhibitory" ligand can turn off a gene either by removing a gene activator protein from the DNA *(upper right panel)* or by causing a gene repressor protein to bind *(lower left panel)*.

of glucose. It would be wasteful, however, for CAP to induce expression of the *lac* operon if lactose is not present, and the lac repressor ensures that the *lac* operon is shut off in the absence of lactose. This arrangement enables the control region of *lac* operon to respond to and integrate two different signals, so that the operon is highly expressed only when two conditions are met: lactose must be present and glucose must be absent. Any of the other three possible signal combinations maintain the cluster of genes in the off state (Figure 7–38).

The simple logic of this genetic switch first attracted the attention of biologists over 50 years ago. As explained above, the molecular basis of the switch was uncovered by a combination of genetics and biochemistry, providing the first insight into how gene expression is controlled. Although the same basic strategies are used to control gene expression in higher organisms, the genetic switches that are used are usually much more complex.

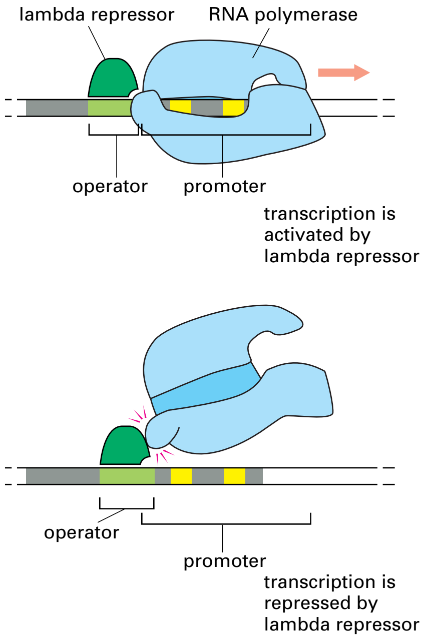
## Regulation of Transcription in Eucaryotic Cells Is Complex

The two-signal switching mechanism that regulates the *lac* operon is elegant and simple. However, it is difficult to imagine how it could grow in complexity to allow dozens of signals to regulate transcription from the operon: there is not enough room in the neighborhood of the promoter to pack in a sufficient number of regulatory DNA sequences. How then have eucaryotes overcome such limitations to create their more complex genetic switches?

The regulation of transcription in eucaryotes differs in three important ways from that typically found in bacteria.

- First, eucaryotes make use of gene regulatory proteins that can act even when they are bound to DNA thousands of nucleotide pairs away from the promoter that they influence, which means that a single promoter can be controlled by an almost unlimited number of regulatory sequences scattered along the DNA.
- Second, as we saw in the last chapter, eucaryotic RNA polymerase II, which transcribes all protein-coding genes, cannot initiate transcription on its

**Figure 7–37 Some bacterial gene regulatory proteins can act as both a transcriptional activator and a repressor, depending on the precise placement of its binding sites in DNA.** An example is the bacteriophage lambda repressor. For some genes, the protein acts as a transcriptional activator by providing a favorable contact for RNA polymerase *(top)*. At other genes *(bottom)*, the operator is located one base pair closer to the promoter, and, instead of helping polymerase, the repressor now competes with it for binding to the DNA. The lambda repressor recognizes its operator by a helix–turn–helix motif, as shown in Figure 7–14.


lambda repressor
RNA polymerase
operator
promoter
transcription is activated by lambda repressor
operator
promoter
transcription is repressed by lambda repressor

own. It requires a set of proteins called *general transcription factors*, which must be assembled at the promoter before transcription can begin. (The term "general" refers to the fact that these proteins assemble on all promoters transcribed by RNA polymerase II; in this they differ from gene regulatory proteins, which act only at particular genes.) This assembly process provides, in principle, multiple steps at which the rate of transcription initiation can be speeded up or slowed down in response to regulatory signals, and many eucaryotic gene regulatory proteins influence these steps.

- Third, the packaging of eucaryotic DNA into chromatin provides opportunities for regulation not available to bacteria.

Having discussed the general transcription factors for RNA polymerase II in Chapter 6 (see pp. 309–312), we focus here on the first and third of these features and how they are used to control eucaryotic gene expression selectively.

## Eucaryotic Gene Regulatory Proteins Control Gene Expression from a Distance

Like bacteria, eucaryotes use gene regulatory proteins (activators and repressors) to regulate the expression of their genes but in a somewhat different way. The DNA sites to which the eucaryotic gene activators bound were originally termed **enhancers**, since their presence "enhanced," or increased, the rate of transcription dramatically. It came as a surprise when, in 1979, it was discovered that these activator proteins could be bound thousands of nucleotide pairs away from the promoter. Moreover, eucaryotic activators could influence transcription

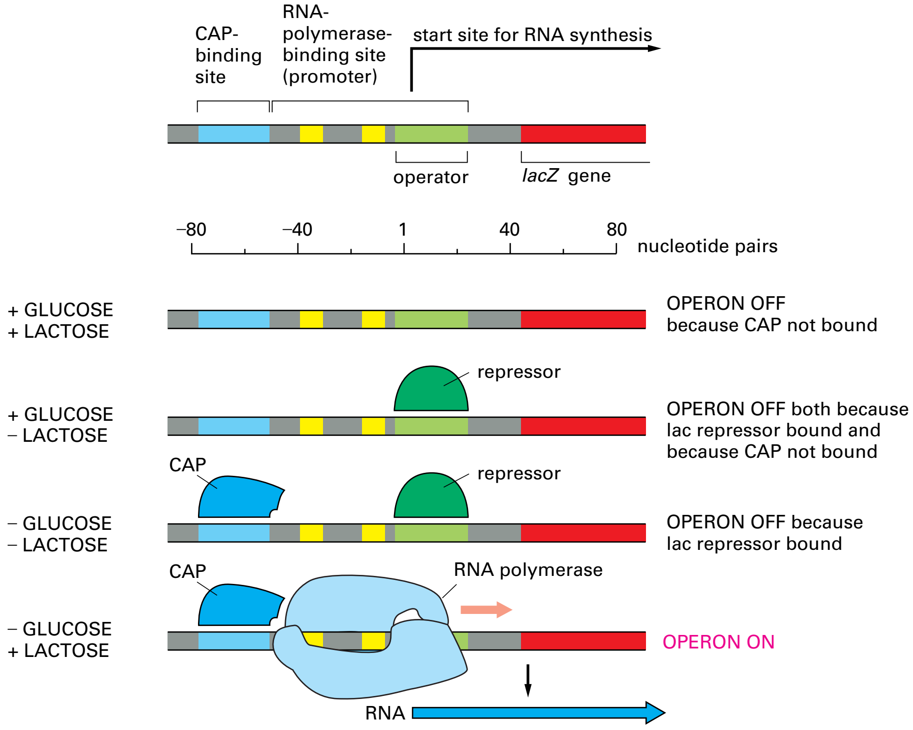

CAP-binding site
RNA-polymerase-binding site (promoter)
start site for RNA synthesis
operator
lacZ gene
–80
–40
1
40
80
nucleotide pairs
+ GLUCOSE
+ LACTOSE
OPERON OFF because CAP not bound
repressor
+ GLUCOSE
– LACTOSE
OPERON OFF both because lac repressor bound and because CAP not bound
CAP
repressor
– GLUCOSE
– LACTOSE
OPERON OFF because lac repressor bound
CAP
RNA polymerase
– GLUCOSE
+ LACTOSE
OPERON ON
RNA

**Figure 7–38 Dual control of the *lac* operon.** Glucose and lactose levels control the initiation of transcription of the *lac* operon through their effects on the lac repressor protein and CAP. Lactose addition increases the concentration of allolactose, which binds to the repressor protein and removes it from the DNA. Glucose addition decreases the concentration of cyclic AMP; because cyclic AMP no longer binds to CAP, this gene activator protein dissociates from the DNA, turning off the operon. As shown in Figure 7–11, CAP is known to induce a bend in the DNA when it binds; for simplicity, the bend is not shown here. *LacZ*, the first gene of the *lac* operon, encodes the enzyme β-galactosidase, which breaks down the disaccharide lactose to galactose and glucose.

The essential features of the *lac* operon are summarized in the figure, but in reality the situation is more complex. For one thing, there are several *lac* repressor binding sites located at different positions along the DNA. Although the one illustrated exerts the greatest effect, the others are required for full repression. In addition, expression of the *lac* operon is never completely shut down. A small amount of the enzyme β-galactosidase is required to convert lactose to allolactose thereby permitting the *lac* repressor to be inactivated when lactose is added to the growth medium.

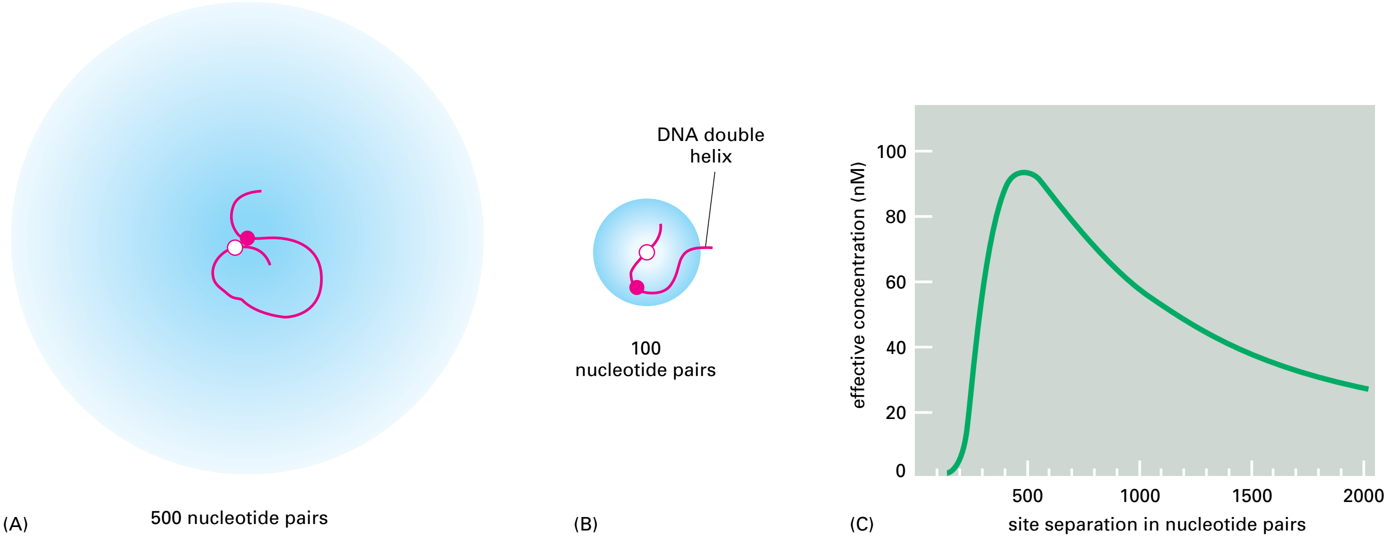


**Figure 7–39 Binding of two proteins to separate sites on the DNA double helix can greatly increase their probability of interacting.** (A) The tethering of one protein to the other via an intervening DNA loop of 500 nucleotide pairs increases their frequency of collision. The intensity of *blue* coloring reflects the probability that the *red* protein will be located at each position in space relative to the *white* protein. (B) The flexibility of DNA is such that an average sequence makes a smoothly graded 90° bend (a curved turn) about once every 200 nucleotide pairs. Thus, when two proteins are tethered by only 100 nucleotide pairs, their contact is relatively restricted. In such cases the protein interaction is facilitated when the two protein-binding sites are separated by a multiple of about 10 nucleotide pairs, which places both proteins on the same side of the DNA helix (which has about 10 nucleotides per turn) and thus on the inside of the DNA loop, where they can best reach each other. (C) The theoretical effective concentration of the *red* protein at the site where the *white* protein is bound, as a function of their separation. (C, courtesy of Gregory Bellomy, modified from M.C. Mossing and M.T. Record, *Science* 233:889–892, 1986. © AAAS.)

of a gene when bound either upstream or downstream from it. How do enhancer sequences and the proteins bound to them function over these long distances? How do they communicate with the promoter?

Many models for “action at a distance” have been proposed, but the simplest of these seems to apply in most cases. The DNA between the enhancer and the promoter loops out to allow the activator proteins bound to the enhancer to come into contact with proteins (RNA polymerase, one of the general transcription factors, or other proteins) bound to the promoter (see Figure 6–19). The DNA thus acts as a tether, helping a protein bound to an enhancer even thousands of nucleotide pairs away to interact with the complex of proteins bound to the promoter (Figure 7–39). This phenomenon also occurs in bacteria, although less commonly and over much shorter lengths of DNA (Figure 7–40).

## A Eucaryotic Gene Control Region Consists of a Promoter Plus Regulatory DNA Sequences

Because eucaryotic gene regulatory proteins can control transcription when bound to DNA far away from the promoter, the DNA sequences that control the expression of a gene are often spread over long stretches of DNA. We shall use the term **gene control region** to refer to the whole expanse of DNA involved in regulating transcription of a gene, including the **promoter**, where the general transcription factors and the polymerase assemble, and all of the **regulatory sequences** to which gene regulatory proteins bind to control the rate of the assembly processes at the promoter (Figure 7–41). In higher eucaryotes it is not unusual to find the regulatory sequences of a gene dotted over distances as great as 50,000 nucleotide pairs. Although much of this DNA serves as “spacer” sequence and is not recognized by gene regulatory proteins, this spacer DNA may facilitate transcription by providing the flexibility needed for communication between DNA-bound proteins. It is also important to keep in mind that, like other regions of eucaryotic chromosomes, much of the DNA in gene control regions is packaged into nucleosomes and higher-order forms of chromatin, thereby compacting its length.

In this chapter we generally use the term **gene** to refer only to a segment of DNA that is transcribed into RNA (see Figure 7–41). However, the classical view of a gene would include the gene control region as well. The different definitions arise from the different ways in which genes were historically identified. The discovery of alternative RNA splicing has further complicated the definition of a gene—a point we discussed briefly in Chapter 6 and will return to later in this chapter.

Although many gene regulatory proteins bind to enhancer sequences and activate gene transcription, many others function as negative regulators, as we

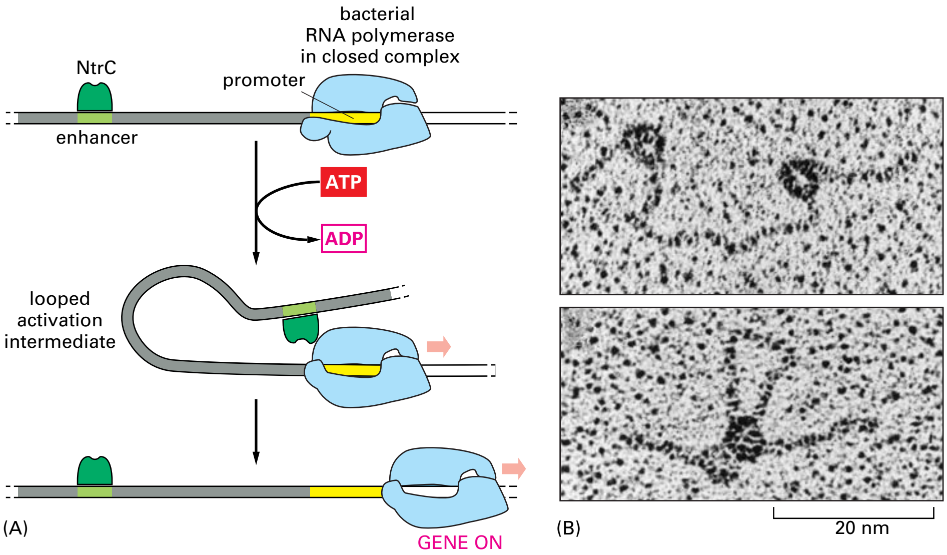


**Figure 7–40 Gene activation at a distance.** (A) NtrC is a bacterial gene regulatory protein that activates transcription by facilitating the transition between the initial binding of RNA polymerase to the promoter and the formation of an initiating complex (discussed in Chapter 6). As indicated, the transition stimulated by NtrC requires the energy produced by ATP hydrolysis, although this requirement is unusual for bacterial transcription initiation. (B) The interaction of NtrC and RNA polymerase, with the intervening DNA looped out, can be seen in the electron microscope. Although transcriptional activation by DNA looping is unusual in bacteria, it is typical of eucaryotic gene regulatory proteins. (B, courtesy of Harrison Echols and Sydney Kustu.)

see below. In contrast to the small number of general transcription factors, which are abundant proteins that assemble on the promoters of all genes transcribed by RNA polymerase II, there are thousands of different gene regulatory proteins. For example, of the roughly 30,000 human genes, an estimated 5–10% encode gene regulatory proteins. These regulatory proteins vary from one gene control region to the next, and each is usually present in very small amounts in a cell, often less than 0.01% of the total protein. Most of them recognize their specific DNA sequences using one of the DNA-binding motifs discussed previously, although as we discuss below, some do not recognize DNA directly but instead assemble on other DNA-bound proteins.

The gene regulatory proteins allow the individual genes of an organism to be turned on or off specifically. Different selections of gene regulatory proteins are present in different cell types and thereby direct the patterns of gene expression that give each cell type its unique characteristics. Each gene in a eucaryotic cell is regulated differently from nearly every other gene. Given the number of genes in eucaryotes and the complexity of their regulation, it has been difficult to formulate simple rules for gene regulation that apply in every case. We can, however, make some generalizations about how gene regulatory proteins, once bound to a gene control region on DNA, influence the rate of transcription initiation, as we now explain.

## Eucaryotic Gene Activator Proteins Promote the Assembly of RNA Polymerase and the General Transcription Factors at the Startpoint of Transcription

Most gene regulatory proteins that activate gene transcription—that is, most **gene activator proteins**—have a modular design consisting of at least two distinct domains. One domain usually contains one of the structural motifs discussed previously that recognizes a specific regulatory DNA sequence. In the simplest

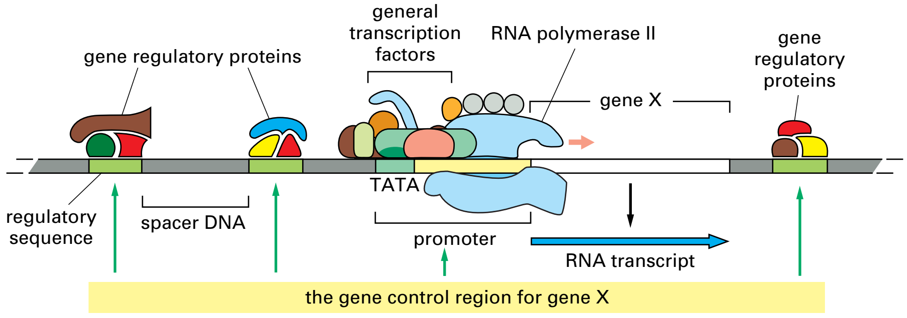


**Figure 7–41 The gene control region of a typical eucaryotic gene.** The *promoter* is the DNA sequence where the general transcription factors and the polymerase assemble (see Figure 6–16). The *regulatory sequences* serve as binding sites for gene regulatory proteins, whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or downstream of the gene. DNA looping is thought to allow gene regulatory proteins bound at any of these positions to interact with the proteins that assemble at the promoter. Whereas the general transcription factors that assemble at the promoter are similar for all polymerase II transcribed genes, the gene regulatory proteins and the locations of their binding sites relative to the promoter are different for each gene.

cases, a second domain—sometimes called an *activation domain*—accelerates the rate of transcription initiation. This type of modular design was first revealed by experiments in which genetic engineering techniques were used to create a hybrid protein containing the activation domain of one protein fused to the DNA-binding domain of a different protein (Figure 7–42).

Once bound to DNA, how do eucaryotic gene activator proteins increase the rate of transcription initiation? As we will see shortly, there are several mechanisms by which this can occur, and, in many cases, these different mechanisms work in concert at a single promoter. But, regardless of the precise biochemical pathway, the main function of activators is to attract, position, and modify the general transcription factors and RNA polymerase II at the promoter so that transcription can begin. They do this both by acting directly on the transcription machinery itself and by changing the chromatin structure around the promoter.

We consider first the ways in which activators directly influence the positioning of the general transcription factors and RNA polymerase at promoters and help kick them into action. Although the general transcription factors and RNA polymerase II assemble in a stepwise, prescribed order *in vitro* (see Figure 6–16), there are cases in living cells where some of them are brought to the promoter as a large pre-assembled complex that is sometimes called the *RNA polymerase II holoenzyme.* In addition to some of the general transcription factors and RNA polymerase, the holoenzyme typically contains a 20-subunit protein complex called the *mediator*, which was first identified biochemically as being required for activators to stimulate transcription initiation.

Many activator proteins interact with the holoenzyme complex and thereby make it more energetically favorable for it to assemble on a promoter that is linked through DNA to the site where the activator protein is bound (Figure 7–43A). In this sense, eucaryotic activators resemble those of bacteria in helping to attract and position RNA polymerase on specific sites on DNA (see Figure 7–36). One type of experiment that supports the idea that activators attract the holoenzyme complex to promoters creates an "activator bypass" (Figure 7–43B). Here, a sequence-specific DNA-binding domain is experimentally fused directly to a component of the mediator; this hybrid protein, which lacks an activation domain, strongly stimulates transcription initiation when the DNA sequence to which it binds is placed in proximity to a promoter.

Although recruitment of the holoenzyme complex to promoters provides a conceptually simple mechanism for envisioning gene activation, the effect of activators on the holoenzyme complex is probably more complicated. For example, a stepwise assembly of the general transcription factors (see Figure 6–16)

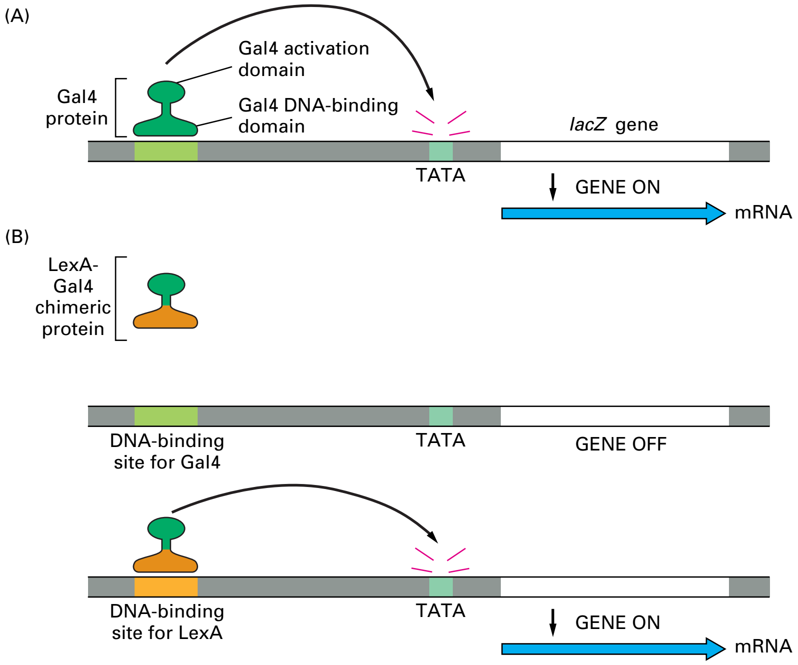


**Figure 7–42 The modular structure of a gene activator protein.** Outline of an experiment that reveals the presence of independent DNA-binding and transcription-activating domains in the yeast gene activator protein Gal4. A functional activator can be reconstituted from the C-terminal portion of the yeast Gal4 protein if it is attached to the DNA-binding domain of a bacterial gene regulatory protein (the LexA protein) by gene fusion techniques. When the resulting bacterial-yeast hybrid protein is produced in yeast cells, it will activate transcription from yeast genes provided that the specific DNA-binding site for the bacterial protein has been inserted next to them. (A) The normal activation of gene transcription produced by the Gal4 protein. (B) The chimeric gene regulatory protein requires the LexA protein DNA-binding site for its activity.

Gal4 is normally responsible for activating the transcription of yeast genes that code for the enzymes that convert galactose to glucose. In the experiments shown here, the control region for one of these genes was fused to the *E. coli lacZ* gene, which codes for the enzyme β-galactosidase (see Figure 7–38). β-galactosidase is very simple to detect biochemically and thus provides a convenient way to monitor the expression level specified by a gene control region; *lacZ* thus serves as a *reporter gene* since it "reports" the activity of a gene control region.

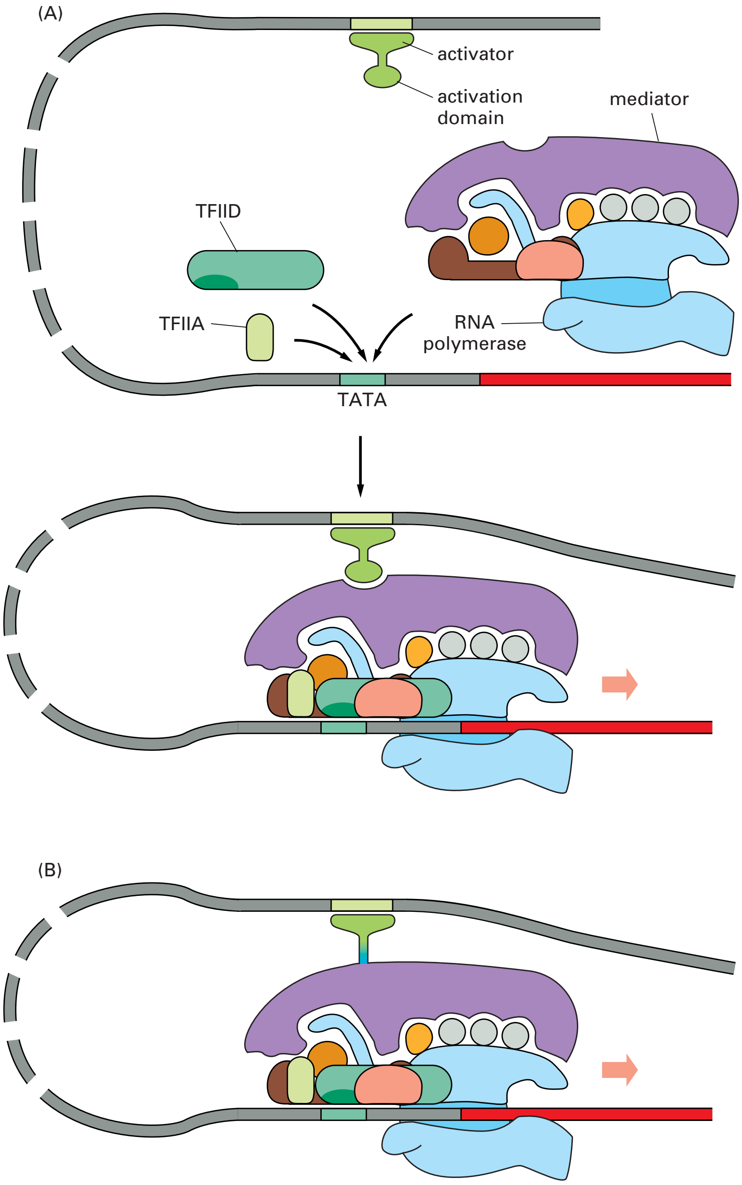


**Figure 7–43 Activation of transcription initiation in eucaryotes by recruitment of the eucaryotic RNA polymerase II holoenzyme complex.** (A) An activator protein bound in proximity to a promoter attracts the holoenzyme complex to the promoter. According to this model, the holoenzyme (which contains over 100 protein subunits) is brought to the promoter separately from the general transcription factors TFIID and TFIIA. The "broken" DNA in this and subsequent figures indicates that this portion of the DNA molecule can be very long and of variable length. (B) Diagram of an *in vivo* experiment whose outcome supports the holoenzyme recruitment model for gene activator proteins. The DNA-binding domain of a protein has been fused directly to a protein component of the mediator, a 20-subunit protein complex which is part of the holoenzyme complex, but which is easily dissociable from the remainder of the holoenzyme. When the binding site for the hybrid protein is experimentally inserted near a promoter, transcription initiation is strongly increased. In this experiment, the "activation domain" of the activator (see Figure 7–42) has been omitted, suggesting that an important function of the activation domain is simply to interact with the RNA polymerase holoenzyme complex and thereby aid in its assembly at the promoter. The ability of gene activator proteins to recruit the transcription machinery to promoters has also been demonstrated directly, using chromatin immunoprecipitation (see Figure 7–32).

DNA-bound activator proteins typically increase the rate of transcription by up to 1000-fold, which is consistent with a relatively weak and nonspecific interaction between the activator and the holoenzyme (a 1000-fold change in affinity corresponds to a change in $\Delta G$ of ~4 kcal/mole, which could be accounted for by just a few weak, noncovalent bonds).

may occur on some promoters. On others, their rearrangement, once brought to DNA as part of the holoenzyme, may be required. In addition, most forms of the holoenzyme complex lacks some of the general transcription factors (notably TFIID and TFIIA), and these must be assembled on the promoter separately (see Figure 7–43A). In principle, any of these assembly processes could be a slow step on the pathway to transcription initiation, and activators could facilitate their completion. In fact, many activators have been shown to interact with one or more of the general transcription factors, and several have been shown to directly accelerate their assembly at the promoter (Figure 7–44).

## Eucaryotic Gene Activator Proteins Modify Local Chromatin Structure

In addition to their direct actions in assembling the RNA polymerase holoenzyme and the general transcription factors on DNA, gene activator proteins also

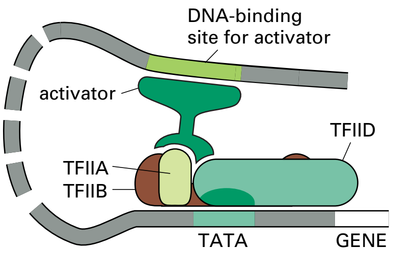


**Figure 7–44 A model for the action of some eucaryotic transcriptional activators.** The gene activator protein, bound to DNA in the rough vicinity of the promoter, facilitates the assembly of some of the general transcription factors. Although some activator proteins may be dedicated to particular steps in the pathway for transcription initiation, many seem to be capable of acting at several steps.

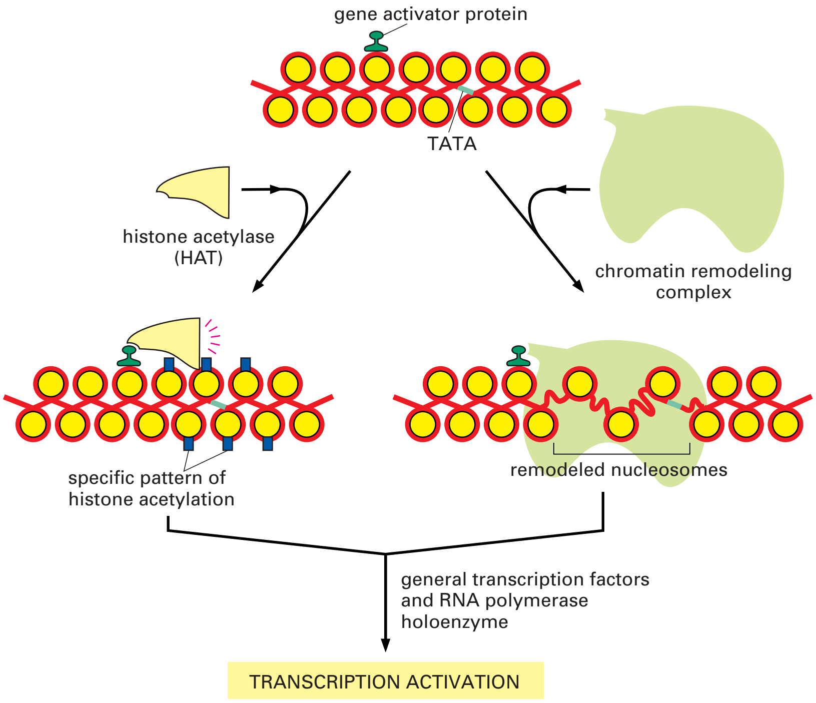

gene activator protein
TATA
histone acetylase (HAT)
chromatin remodeling complex
specific pattern of histone acetylation
remodeled nucleosomes
general transcription factors and RNA polymerase holoenzyme
TRANSCRIPTION ACTIVATION

**Figure 7–45 Local alterations in chromatin structure directed by eucaryotic gene activator proteins.** Histone acetylation and nucleosome remodeling generally render the DNA packaged in chromatin more accessible to other proteins in the cell, including those required for transcription initiation. In addition, specific patterns of histone modification directly aid in the assembly of the general transcription factors at the promoter (see Figure 7–46).

Transcription initiation and the formation of a compact chromatin structure can be regarded as competing biochemical assembly reactions. Enzymes that increase, even transiently, the accessibility of DNA in chromatin will tend to favor transcription initiation (see Figure 4–34).

promote transcription initiation by changing the chromatin structure of the regulatory sequences and promoters of genes. As we saw in Chapter 4, the two most important ways of locally altering chromatin structure are through covalent histone modifications and nucleosome remodeling (see Figures 4–34 and 4–35). Many gene activator proteins make use of both these mechanisms by binding to and thereby recruiting histone acetyl transferases (HATs), commonly known as histone acetylases, and ATP-dependent chromatin remodeling complexes (Figure 7–45) to work on nearby chromatin. In general terms, the local alterations in chromatin structure that ensue allow greater accessibility to the underlying DNA. This accessibility facilitates the assembly of the general transcription factors and the RNA polymerase holoenzyme at the promoter, and it also allows the binding of additional gene regulatory proteins to the control region of the gene (Figure 7–46A).

The general transcription factors seem unable to assemble onto a promoter that is packaged in a conventional nucleosome. In fact, such packaging may have evolved in part to ensure that leaky, or basal, transcription initiation (initiation at a promoter in the absence of gene activator protein bound upstream of it) does

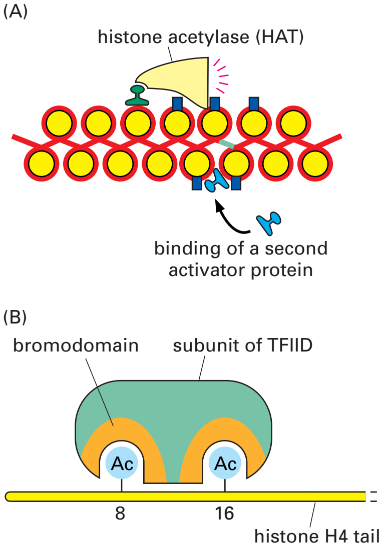


**Figure 7–46 Two specific ways that local histone acetylation can stimulate transcription initiation.** (A) Some gene activator proteins can bind directly to DNA that is packaged in unmodified chromatin. By attracting histone acetylases (and nucleosome remodeling complexes), these “pioneer” activators can facilitate the binding to DNA of additional activator proteins that cannot bind to unmodified chromatin. These additional proteins can in turn carry out additional modifications of chromatin or act directly on the transcription machinery as shown in Figures 7–43 and 7–44. (B) A subunit of the general transcription factor TFIID contains two 120-amino acid protein domains called *bromodomains*. Each bromodomain forms a binding pocket for an acetylated lysine side chain (designated Ac in the figure); in TFIID the two pockets are separated by 25 Å, which is the optimal spacing for recognizing a pair of acetylated lysines separated by six or seven amino acids on the N-terminal tail of histone H4. In addition to the pattern of acetylation shown, this subunit of TFIID also recognizes the histone H4 tail acetylated at positions 5 and 12 and the fully acetylated tail. It has no appreciable affinity for an unacetylated H4 tail and only low affinity for an H4 tail acetylated at a single lysine. As shown in Figure 4–35, certain patterns of histone H4 acetylation, including those recognized by TFIID, are associated with transcriptionally active regions of chromatin.

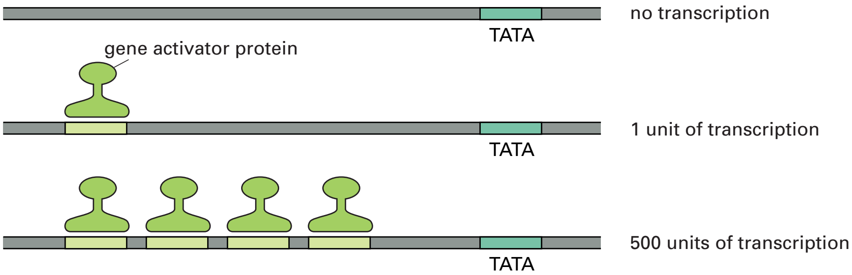


**Figure 7–47 Transcriptional synergy.** In this experiment, the rate of transcription produced by three experimentally constructed regulatory regions is compared in a eucaryotic cell. Transcriptional synergy, the greater than additive effect of the activators, is observed when several molecules of gene activator protein are bound upstream of the promoter. Synergy is also typically observed between different gene activator proteins from the same organism and even between activator proteins from widely different eucaryotic species when they are experimentally introduced into the same cell. This last observation reflects the high degree of conservation of the transcription machinery.

not occur. As well as making the DNA more generally accessible, local histone acetylation has a more specialized role in promoting transcription initiation. As discussed in Chapter 4 (see Figure 4–35), certain patterns of histone acetylation are associated with transcriptionally active chromatin, and gene activator proteins, by recruiting histone acetylases, produce these patterns. One such pattern (Figure 7–46B) is directly recognized by one of the subunits of the general transcription factor TFIID, and this recognition apparently helps the factor assemble DNA that is packaged in chromatin. Thus gene activator proteins, through the action of histone acetylases, can indirectly aid in the assembly of the general transcription factors at a promoter and thereby stimulate transcription initiation.

## Gene Activator Proteins Work Synergistically

We have seen that eucaryotic gene activator proteins can influence several different steps in transcription initiation, and this property has important consequences when different activator proteins work together. In general, where several factors work together to enhance a reaction rate, the joint effect is generally not merely the sum of the enhancements caused by each factor alone, but the product. If, for example, factor A lowers the free-energy barrier for a reaction by a certain amount and thereby speeds up the reaction 100-fold, and factor B, by acting on another aspect of the reaction, does likewise, then A and B acting in parallel will lower the barrier by a double amount and speed up the reaction 10,000-fold. Similar multiplicative effects occur if A and B speed the reaction by each helping to recruit necessary proteins to the reaction site. Thus, gene activator proteins often exhibit what is called *transcriptional synergy*, where the transcription rate produced by several activator proteins working together is much higher than that produced by any of the activators working alone (Figure 7–47). Transcriptional synergy is observed both between different gene activator proteins bound upstream of a gene and between multiple DNA-bound molecules of the same activator. It is therefore not difficult to see how multiple gene regulatory proteins, each binding to a different regulatory DNA sequence, could control the final rate of transcription of a eucaryotic gene.

Since gene activator proteins can influence many different steps on the pathway to transcriptional activation, it is worth considering whether these steps always occur in a prescribed order. For example does chromatin remodeling necessarily precede histone acetylation or vice versa? When does recruitment of the holoenzyme complex occur relative to the chromatin modifying steps? The answers to these questions appears to be different for different genes—and even for the same gene under different conditions (Figure 7–48). Whatever the precise mechanisms and the order in which they are carried out, a gene regulatory protein must be bound to DNA either directly or indirectly to influence transcription of its target promoter, and the rate of transcription of a gene ultimately depends upon the spectrum of regulatory proteins bound upstream and downstream of its transcription start site.

## Eucaryotic Gene Repressor Proteins Can Inhibit Transcription in Various Ways

Like bacteria, eucaryotes use **gene repressor proteins** in addition to activator proteins to regulate transcription of their genes. However, because of differences in the way transcription is initiated in eucaryotes and bacteria, eucaryotic

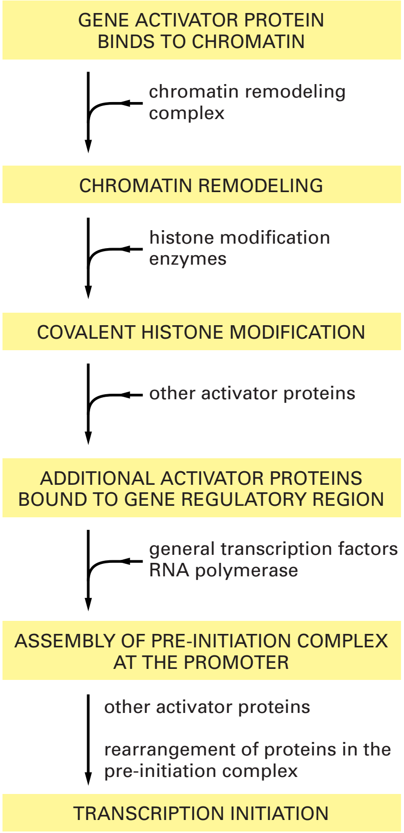

GENE ACTIVATOR PROTEIN BINDS TO CHROMATIN
chromatin remodeling complex
CHROMATIN REMODELING
histone modification enzymes
COVALENT HISTONE MODIFICATION
other activator proteins
ADDITIONAL ACTIVATOR PROTEINS BOUND TO GENE REGULATORY REGION
general transcription factors
RNA polymerase
ASSEMBLY OF PRE-INITIATION COMPLEX AT THE PROMOTER
other activator proteins
rearrangement of proteins in the pre-initiation complex
TRANSCRIPTION INITIATION

**Figure 7–48 An order of events leading to transcription initiation at a specific promoter.** The well-studied example shown is from a promoter in the budding yeast *S. cerevisiae*. The chromatin remodeling complex and histone acetylase apparently dissociate from the DNA after they sequentially act. The order of steps on the pathway to transcription initiation appears to be different for different promoters. For example, in a well-studied example from humans, histone acetylases function first, followed by RNA polymerase recruitment, followed by chromatin remodeling complex recruitment.

repressors have many more possible mechanisms of action. For example, we saw in Chapter 4 that whole regions of eucaryotic chromosomes can be packaged into *heterochromatin*, a form of chromatin that is normally resistant to transcription. We will return to this feature of eucaryotic chromosomes later in this chapter. In addition to molecules that shut down large regions of chromatin, eucaryotic cells also contain gene regulatory proteins that act only locally to repress transcription of nearby genes. Unlike bacterial repressors, most do not directly compete with the RNA polymerase for access to the DNA; rather they work by a variety of other mechanisms, some of which are illustrated in Figure 7–49. Like gene activator proteins, many eucaryotic repressor proteins act through more than one mechanism, thereby ensuring robust and efficient repression.

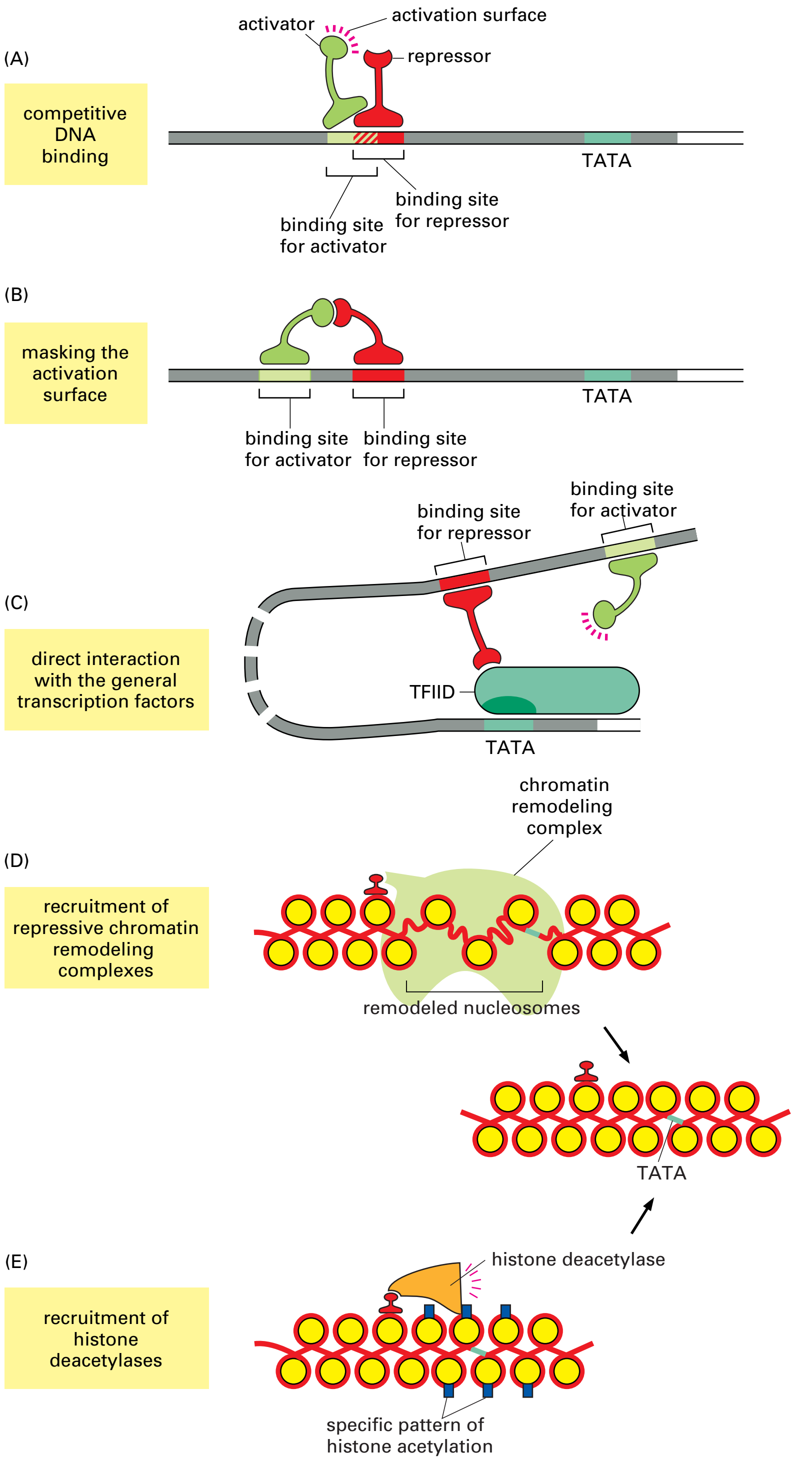

(A)
competitive DNA binding
activator
activation surface
repressor
TATA
binding site for repressor
binding site for activator
(B)
masking the activation surface
TATA
binding site for activator
binding site for repressor
(C)
direct interaction with the general transcription factors
binding site for repressor
binding site for activator
TFIID
TATA
(D)
recruitment of repressive chromatin remodeling complexes
chromatin remodeling complex
remodeled nucleosomes
TATA
(E)
recruitment of histone deacetylases
histone deacetylase
specific pattern of histone acetylation

**Figure 7–49 Five ways in which eucaryotic gene repressor proteins can operate.** (A) Gene activator proteins and gene repressor proteins compete for binding to the same regulatory DNA sequence. (B) Both proteins can bind DNA, but the repressor binds to the activation domain of the activator protein thereby preventing it from carrying out its activation functions. In a variation of this strategy, the repressor binds tightly to the activator without having to be bound to DNA directly. (C) The repressor interacts with an early stage of the assembling complex of general transcription factors, blocking further assembly. Some repressors also act at late stages in transcription initiation, for example, by preventing the release of the RNA polymerase from the general transcription factors. (D) The repressor recruits a chromatin remodeling complex which returns the nucleosomal state of the promoter region to its pre-transcriptional form. Certain types of remodeling complexes appear dedicated to restoring the repressed nucleosomal state of a promoter, whereas others (for example, those recruited by activator proteins) render DNA packaged in nucleosomes more accessible (see Figure 4–34). However the same remodeling complex could in principle be used either to activate or repress transcription: depending on the concentration of other proteins in the nucleus, either the remodeled state or the repressed state could be stabilized. According to this view, the remodeling complex simply allows chromatin structure to change. (E) The repressor attracts a histone deacetylase to the promoter. Local histone deacetylation reduces the affinity of TFIID for the promoter (see Figure 7–46) and decreases the accessibility of DNA in the affected chromatin. A sixth mechanism of negative control—inactivation of a transcriptional activator by heterodimerization—was illustrated in Figure 7–26. For simplicity, nucleosomes have been omitted from (A)–(C), and the scale of (D) and (E) has been reduced relative to (A)–(C).

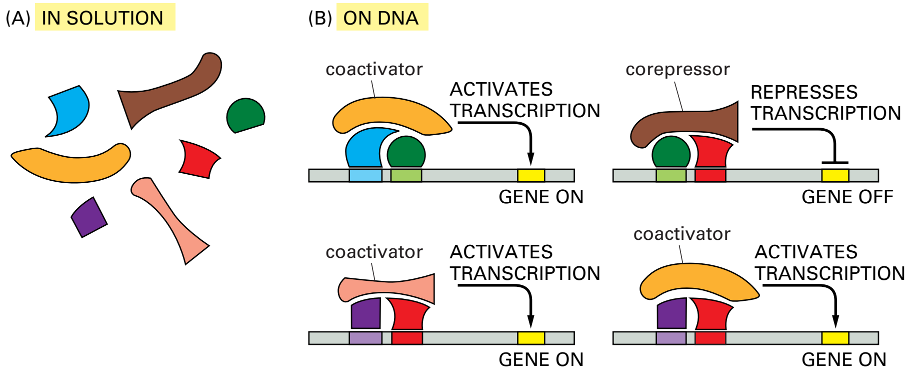


**Figure 7–50 Eucaryotic gene regulatory proteins often assemble into complexes on DNA.** Seven gene regulatory proteins are shown in (A). The nature and function of the complex they form depends on the specific DNA sequence that seeds their assembly. In (B), some assembled complexes activate gene transcription, while another represses transcription. Note that the *red* protein is shared by both activating and repressing complexes.

## Eucaryotic Gene Regulatory Proteins Often Assemble into Complexes on DNA

So far we have been discussing eucaryotic gene regulatory proteins as though they work as individual polypeptides. In reality, most act as parts of complexes composed of several (and sometimes many) polypeptides, each with a distinct function. These complexes often assemble only in the presence of the appropriate DNA sequence. In some well-studied cases, for example, two gene regulatory proteins with a weak affinity for each other cooperate to bind to a DNA sequence, neither protein having a sufficient affinity for DNA to efficiently bind to the DNA site on its own. Once bound to DNA, the protein dimer creates a distinct surface that is recognized by a third protein that carries an activator domain that stimulates transcription (Figure 7–50). This example illustrates an important general point: protein–protein interactions that are too weak to cause proteins to assemble in solution can cause the proteins to assemble on DNA; in this way the DNA sequence acts as a "crystallization" site or seed for the assembly of a protein complex.

An individual gene regulatory protein can often participate in more than one type of regulatory complex. A protein might function, for example, in one case as part of a complex that activates transcription and in another case as part of a complex that represses transcription (see Figure 7–50). Thus individual eucaryotic gene regulatory proteins are not necessarily dedicated activators or repressors; instead, they function as regulatory units that are used to generate complexes whose function depends on the final assembly of all of the individual components. This final assembly, in turn, depends both on the arrangement of control region DNA sequences and on which gene regulatory proteins are present in the cell.

Gene regulatory proteins that do not themselves bind DNA but assemble on DNA-bound gene regulatory proteins are often termed **coactivators** or **corepressors**, depending on their effect on transcription initiation. As shown in Figure 7–50, the same coactivator or corepressor can assemble on different DNA binding proteins. Coactivators and corepressors typically carry out multiple functions: they can interact with chromatin remodeling complexes, histone modifying enzymes, the RNA polymerase holoenzyme, and several of the general transcription factors.

In some cases, the precise DNA sequence to which a regulatory protein directly binds can affect the conformation of this protein and thereby influence its subsequent transcriptional activity. When bound to one type of DNA sequence, for example, a steroid hormone receptor interacts with a corepressor and ultimately turns off transcription. When bound to a slightly different DNA sequence, it assumes a different conformation and interacts with a coactivator, thereby stimulating transcription.

Typically, the assembly of a group of regulatory proteins on DNA is guided by a few relatively short stretches of nucleotide sequence (see Figure 7–50). However, in some cases, a more elaborate protein–DNA structure, termed an *enhancesome*, is formed (Figure 7–51). A hallmark of enhancesomes is the participation of *architectural proteins* that bend the DNA by a defined angle and

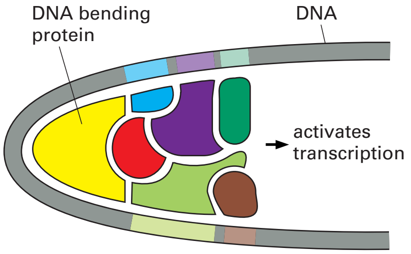


**Figure 7–51 Schematic depiction of an enhancesome.** The protein depicted in *yellow* is termed an architectural protein since its main role is to bend the DNA to allow the cooperative assembly of the other components. The protein surface of this enhancesome interacts with a coactivator which activates transcription at a nearby promoter. The enhancesome depicted here is based on that found in the control region of the gene that codes for a subunit of the T cell receptor (discussed in Chapter 24). The complete set of protein components for the enhancesome are present only in certain cells of the developing immune system, which eventually give rise to mature T cells.

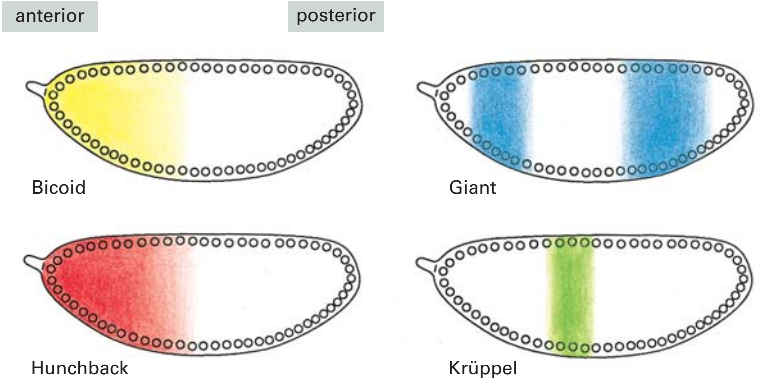


**Figure 7–52 The nonuniform distribution of four gene regulatory proteins in an early *Drosophila* embryo.** At this stage the embryo is a syncytium, with multiple nuclei in a common cytoplasm. Although not illustrated in these drawings, all of these proteins are concentrated in the nuclei.

thereby promote the assembly of the other enhancesome proteins. Since formation of the enhancesome requires the presence of many gene regulatory proteins, it provides a simple way to ensure that a gene is expressed only when the correct combination of these proteins is present in the cell. We saw earlier how the formation of gene regulatory heterodimers in solution provides a mechanism for the combinatorial control of gene expression. The assembly of larger complexes of gene regulatory proteins on DNA provides a second important mechanism for combinatorial control, offering far richer opportunities.

## Complex Genetic Switches That Regulate *Drosophila* Development Are Built Up from Smaller Modules

Given that gene regulatory proteins can be positioned at multiple sites along long stretches of DNA, that these proteins can assemble into complexes at each site, and that the complexes can influence the chromatin structure and the recruitment and assembly of the general transcription machinery at the promoter, there would seem to be almost limitless possibilities for the elaboration of control devices to regulate eucaryotic gene transcription.

A particularly striking example of a complex, multicomponent genetic switch is that controlling the transcription of the *Drosophila even-skipped* (*eve*) gene, whose expression plays an important part in the development of the *Drosophila* embryo. If this gene is inactivated by mutation, many parts of the embryo fail to form, and the embryo dies early in development. As discussed in Chapter 21, at the earliest stage of development where *eve* is expressed, the embryo is a single giant cell containing multiple nuclei in a common cytoplasm. This cytoplasm is not uniform, however: it contains a mixture of gene regulatory proteins that are distributed unevenly along the length of the embryo, thus providing positional information that distinguishes one part of the embryo from another (Figure 7–52). (The way these differences are initially set up is discussed in Chapter 21.) Although the nuclei are initially identical, they rapidly begin to express different genes because they are exposed to different gene regulatory proteins. The nuclei near the anterior end of the developing embryo, for example, are exposed to a set of gene regulatory proteins that is distinct from the set that influences nuclei at the posterior end of the embryo.

The regulatory DNA sequences of the *eve* gene are designed to read the concentrations of gene regulatory proteins at each position along the length of the embryo and to interpret this information in such a way that the *eve* gene is expressed in seven stripes, each initially five to six nuclei wide and positioned precisely along the anterior–posterior axis of the embryo (Figure 7–53). How is this remarkable feat of information processing carried out? Although the molecular details are not yet all understood, several general principles have emerged from studies of *eve* and other *Drosophila* genes that are similarly regulated.

The regulatory region of the *eve* gene is very large (approximately 20,000 nucleotide pairs). It is formed from a series of relatively simple regulatory modules, each of which contains multiple regulatory sequences and is responsible

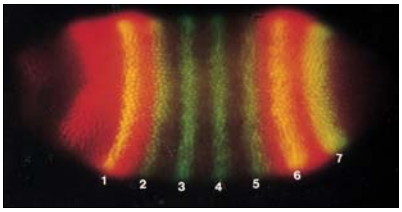


**Figure 7–53 The seven stripes of the protein encoded by the *even-skipped (eve)* gene in a developing *Drosophila* embryo.** Two and one-half hours after fertilization, the egg was fixed and stained with antibodies that recognize the Eve protein *(green)* and antibodies that recognize the Giant protein *(red)*. Where Eve and Giant proteins are both present, the staining appears *yellow*. At this stage in development, the egg contains approximately 4000 nuclei. The Eve and Giant proteins are both located in the nuclei, and the Eve stripes are about four nuclei wide. The staining pattern of the Giant protein is also shown in Figure 7–52. (Courtesy of Michael Levine.)

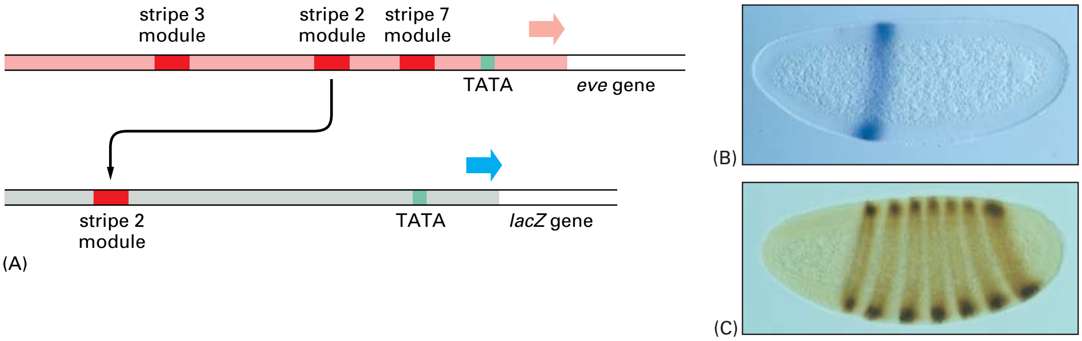


**Figure 7–54 Experiment demonstrating the modular construction of the *eve* gene regulatory region.** (A) A 480-nucleotide-pair piece of the *eve* regulatory region was removed and inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli lacZ* gene). (B) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos expressed β-galactosidase (detectable by histochemical staining) precisely in the position of the second of the seven Eve stripes (C). (B and C, courtesy of Stephen Small and Michael Levine.)

for specifying a particular stripe of *eve* expression along the embryo. This modular organization of the *eve* gene control region is revealed by experiments in which a particular regulatory module (say, that specifying stripe 2) is removed from its normal setting upstream of the *eve* gene, placed in front of a reporter gene (see Figure 7–42), and reintroduced into the *Drosophila* genome (Figure 7–54A). When developing embryos derived from flies carrying this genetic construct are examined, the reporter gene is found to be expressed in precisely the position of stripe 2 (see Figure 7–54). Similar experiments reveal the existence of other regulatory modules, each of which specifies one of the other six stripes or some part of the expression pattern that the gene displays at later stages of development.

## The *Drosophila eve* Gene Is Regulated by Combinatorial Controls

A detailed study of the stripe 2 regulatory module has provided insights into how it reads and interprets positional information. It contains recognition sequences for two gene regulatory proteins (Bicoid and Hunchback) that activate *eve* transcription and two (Krüppel and Giant) that repress it (Figure 7–55). (The gene regulatory proteins of *Drosophila* often have colorful names reflecting the phenotype that results if the gene encoding the protein is inactivated by mutation.) The relative concentrations of these four proteins determine whether protein complexes forming at the stripe 2 module turn on transcription of the *eve* gene. Figure 7–56 shows the distributions of the four gene regulatory proteins across the region of a *Drosophila* embryo where stripe 2 forms. Although the precise details are not known, it seems likely that either one of the two repressor proteins, when bound to the DNA, will turn off the stripe 2 module, whereas both Bicoid and Hunchback must bind for its maximal activation. This simple regulatory unit thereby combines these four positional signals so as to turn on the stripe 2 module (and therefore the expression of the *eve* gene) only in those nuclei that are located where the levels of both Bicoid and Hunchback are high and both Krüppel and Giant are absent. This combination of activators and repressors occurs only in one region of the early embryo; everywhere else, therefore, the stripe 2 module is silent.

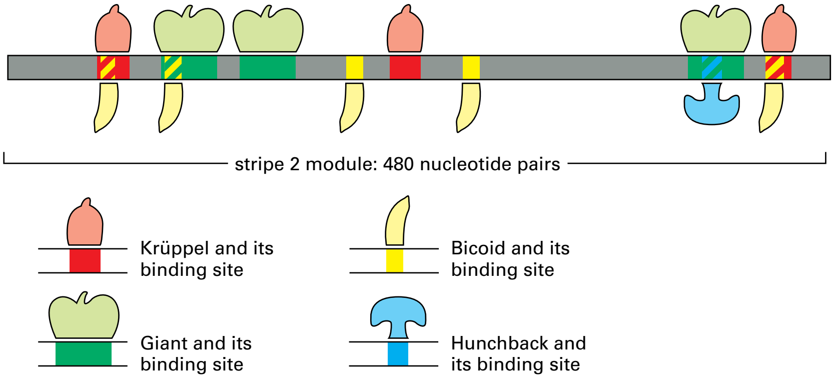


**Figure 7–55 Close-up view of the *eve* stripe 2 unit.** The segment of the *eve* gene control region identified in the previous figure contains regulatory sequences, each of which binds one or another of four gene regulatory proteins. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of *eve* in stripe 2. Flies that are deficient in the two gene activators Bicoid and Hunchback, for example, fail to efficiently express *eve in* stripe 2. In flies deficient in either of the two gene repressors, Giant and Krüppel, stripe 2 expands and covers an abnormally broad region of the embryo. The DNA-binding sites for these gene regulatory proteins were determined by cloning the genes encoding the proteins, overexpressing the proteins in *E. coli,* purifying them, and performing DNA-footprinting experiments as described in Chapter 8. The *top* diagram indicates that, in some cases, the binding sites for the gene regulatory proteins overlap and the proteins can compete for binding to the DNA. For example, binding of Krüppel and binding of Bicoid to the site at the far right are thought to be mutually exclusive.

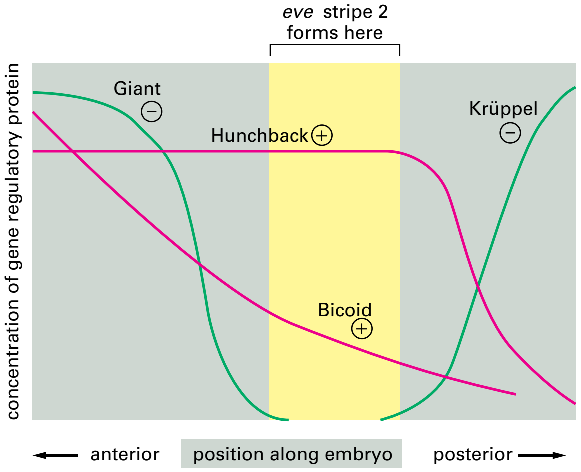


**Figure 7–56 Distribution of the gene regulatory proteins responsible for ensuring that *eve* is expressed in stripe 2.** The distributions of these proteins were visualized by staining a developing *Drosophila* embryo with antibodies directed against each of the four proteins (see Figures 7–52 and 7–53). The expression of *eve* in stripe 2 occurs only at the position where the two activators (Bicoid and Hunchback) are present and the two repressors (Giant and Krüppel) are absent. In fly embryos that lack Krüppel, for example, stripe 2 expands posteriorly. Likewise, stripe 2 expands posteriorly if the DNA-binding sites for Krüppel in the stripe 2 module (see Figure 7–55) are inactivated by mutation and this regulatory region is reintroduced into the genome.

The *eve* gene itself encodes a gene regulatory protein, which, after its pattern of expression is set up in seven stripes, regulates the expression of other *Drosophila* genes. As development proceeds, the embryo is thus subdivided into finer and finer regions that eventually give rise to the different body parts of the adult fly, as discussed in Chapter 21.

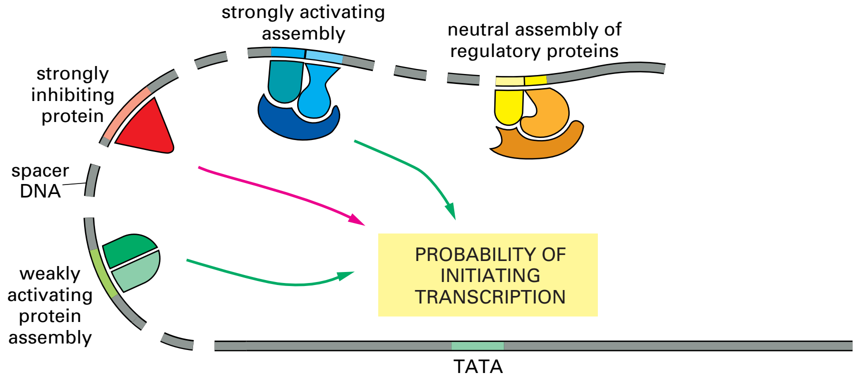
This example from *Drosophila* embryos is unusual in that the nuclei are exposed directly to positional cues in the form of concentrations of gene regulatory proteins. In embryos of most other organisms, individual nuclei are in separate cells, and extracellular positional information must either pass across the plasma membrane or, more usually, generate signals in the cytosol in order to influence the genome.

We have already discussed two mechanisms of combinatorial control of gene expression—heterodimerization of gene regulatory proteins in solution (see Figure 7–22) and the assembly of combinations of gene regulatory proteins into small complexes on DNA (see Figure 7–50). It is likely that both mechanisms participate in the complex regulation of *eve* expression. In addition, the regulation of stripe 2 just described illustrates a third type of combinatorial control. Because the individual regulatory sequences in the *eve* stripe 2 module are strung out along the DNA, many sets of gene regulatory proteins can be bound simultaneously and influence the promoter of a gene. The promoter integrates the transcriptional cues provided by all of the bound proteins (Figure 7–57).

The regulation of *eve* expression is an impressive example of combinatorial control. Seven combinations of gene regulatory proteins—one combination for each stripe—activate *eve* expression, while many other combinations (all those found in the interstripe regions of the embryo) keep the stripe elements silent. The other stripe regulatory modules are thought to be constructed along lines similar to those described for stripe 2, being designed to read positional information provided by other combinations of gene regulatory proteins. The entire gene control region, strung out over 20,000 nucleotide pairs of DNA, binds more than 20 different proteins. A large and complex control region is thereby built from a series of smaller modules, each of which consists of a unique arrangement of short DNA sequences recognized by specific gene regulatory proteins. Although the details are not yet understood, these gene regulatory proteins are thought to employ a number of the mechanisms previously described for activators and repressors. In this way, a single gene can respond to an enormous number of combinatorial inputs.

## Complex Mammalian Gene Control Regions Are Also Constructed from Simple Regulatory Modules

It has been estimated that 5–10% of the coding capacity of a mammalian genome is devoted to the synthesis of proteins that serve as regulators of gene



strongly activating assembly
neutral assembly of regulatory proteins
strongly inhibiting protein
spacer DNA
weakly activating protein assembly
PROBABILITY OF INITIATING TRANSCRIPTION
TATA

**Figure 7–57 Integration at a promoter.** Multiple sets of gene regulatory proteins can work together to influence transcription initiation at a promoter, as they do in the *eve* stripe 2 module illustrated previously in Figure 7–55. It is not yet understood in detail how the integration of multiple inputs is achieved, but it is likely that the final transcriptional activity of the gene results from a competition between activators and repressors that act by the mechanisms summarized in Figures 7–43, 7–44, 7–45, 7–46, and 7–49.

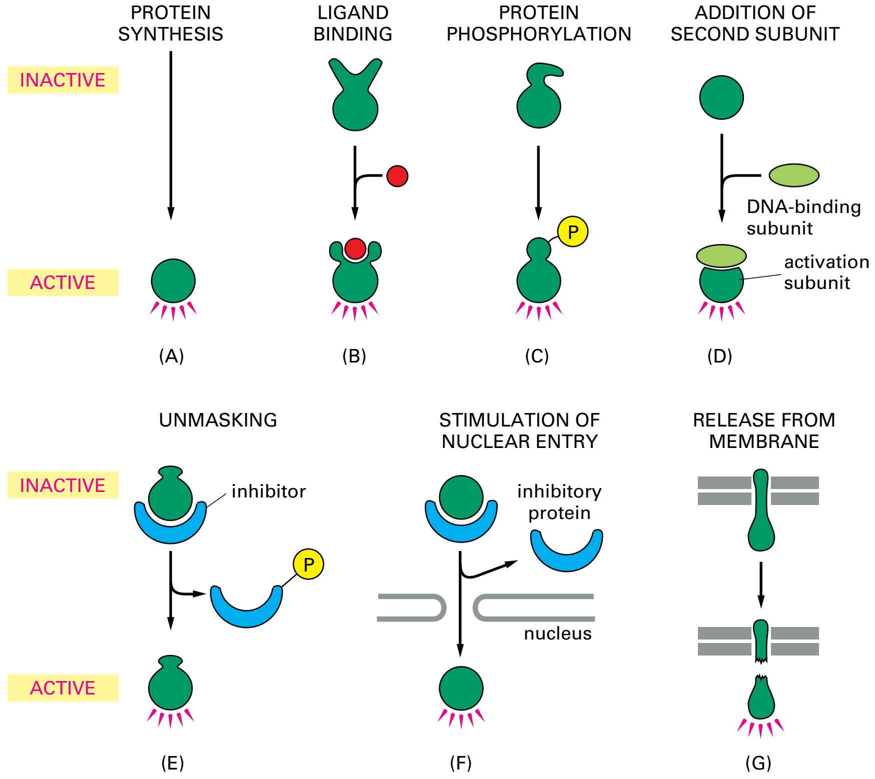
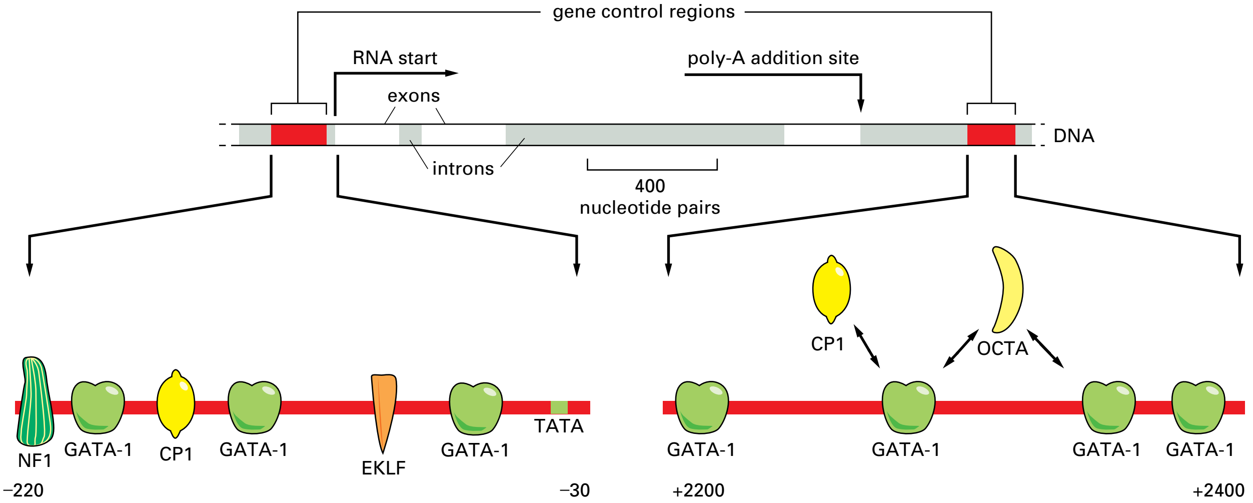


**Figure 7–58 Some ways in which the activity of gene regulatory proteins is regulated in eucaryotic cells.** (A) The protein is synthesized only when needed and is rapidly degraded by proteolysis so that it does not accumulate. (B) Activation by ligand binding. (C) Activation by phosphorylation. (D) Formation of a complex between a DNA-binding protein and a separate protein with a transcription-activating domain. (E) Unmasking of an activation domain by the phosphorylation of an inhibitor protein. (F) Stimulation of nuclear entry by removal of an inhibitory protein that otherwise keeps the regulatory protein from entering the nucleus. (G) Release of a gene regulatory protein from a membrane bilayer by regulated proteolysis.

Each of these mechanisms is typically controlled by extracellular signals which are communicated across the plasma membrane to the gene regulatory proteins in the cell. The ways in which this signaling occurs is discussed in Chapter 15. Mechanisms (A)–(F) are readily reversible and therefore also provide the means to selectively inactivate gene regulatory proteins.

transcription. This large number of genes reflects the exceedingly complex network of controls governing expression of mammalian genes. Each gene is regulated by a set of gene regulatory proteins; each of those proteins is the product of a gene that is in turn regulated by a whole set of other proteins, and so on. Moreover, the regulatory protein molecules are themselves influenced by signals from outside the cell, which can make them active or inactive in a whole variety of ways (Figure 7–58). Thus, pattern of gene expression in a cell can be viewed as the result of a complicated molecular computation that the intracellular gene control network performs in response to information from the cell's surroundings. We shall discuss this further in Chapter 21, dealing with multicellular development, but the complexity is remarkable even at the level of the individual genetic switch, regulating activity of a single gene. It is not unusual, for example, to find a mammalian gene with a control region that is 50,000 nucleotide pairs in length, in which many modules, each containing a number of regulatory sequences that bind gene regulatory proteins, are interspersed with long stretches of spacer DNA.

One of the best-understood examples of a complex mammalian regulatory region is found in the human β-globin gene, which is expressed exclusively in red blood cells and at a specific time in their development. A complex array of gene regulatory proteins controls the expression of the gene, some acting as activators and others as repressors (Figure 7–59). The concentrations (or activities) of many of these gene regulatory proteins are thought to change during development, and only a particular combination of all the proteins triggers transcription of the gene. The human β-globin gene is part of a cluster of globin genes (Figure 7–60A). The five genes of the cluster are transcribed exclusively in erythroid cells, that is, cells of the red blood cell lineage. Moreover, each gene is turned on at a different stage of development (see Figure 7–60B) and in different organs: the ε-globin gene is expressed in the embryonic yolk sac, γ in the yolk sac and the fetal liver, and δ and β primarily in the adult bone marrow. Each of the globin genes has its own set of regulatory proteins that are necessary to turn the gene on at the appropriate time and tissue. In addition to the individual regulation of each of the globin genes, the entire cluster appears to be subject to a shared control region called a *locus control region (LCR)*. The LCR lies far



upstream from the gene cluster (see Figure 7–60A), and we shall discuss its function next.

In cells where the globin genes are not expressed (such as brain or skin cells), the whole gene cluster appears tightly packaged into chromatin. In erythroid cells, by contrast, the entire gene cluster is still folded into nucleosomes, but the higher-order packing of the chromatin has become decondensed This change occurs even before the individual globin genes are transcribed, suggesting that there are two steps of regulation. In the first, the chromatin of the entire globin locus becomes decondensed, which is presumed to allow additional gene regulatory proteins access to the DNA. In the second step, the remaining gene regulatory proteins assemble on the DNA and direct the expression of individual genes.

The LCR appears to act by controlling chromatin condensation, and its importance can be seen in patients with a certain type of thalassemia, a severe inherited form of anemia. In these patients, the β-globin locus is found to have undergone deletions that remove all or part of the LCR, and although the β-globin gene and its nearby regulatory regions are intact, the gene remains transcriptionally silent even in erythroid cells. Moreover, the β-globin gene in the erythroid cells fails to undergo the normal chromatin decondensation step that occurs during erythroid cell development.

Many LCRs (that is, DNA regulatory sequences that regulate the accessibility and expression of distant genes or gene clusters) are present in the human genome, and they regulate a wide variety of cell-type specific genes. The way in which they function is not understood in detail, but several models have been proposed. The simplest is based on principles we have already discussed in this chapter: the gene regulatory proteins that bind to the LCR interact through DNA

**Figure 7–59 Model for the control of the human β-globin gene.** The diagram shows some of the gene regulatory proteins thought to control expression of the gene during red blood cell development (see Figure 7–60). Some of the gene regulatory proteins shown, such as CP1, are found in many types of cells, while others, such as GATA-1, are present in only a few types of cells—including red blood cells—and therefore are thought to contribute to the cell-type specificity of β-globin gene expression. As indicated by the *double-headed arrows*, several of the binding sites for GATA-1 overlap those of other gene regulatory proteins; it is thought that occupancy of these sites by GATA-1 excludes binding of other proteins. Once bound to DNA, the gene regulatory proteins recruit chromatin remodeling complexes, histone modifying enzymes, the general transcription factors and RNA polymerase to the promoter. (Adapted from B. Emerson, in Gene Expression: General and Cell-Type Specific [M. Karin, ed.], pp. 116–161. Boston: Birkhauser, 1993.)

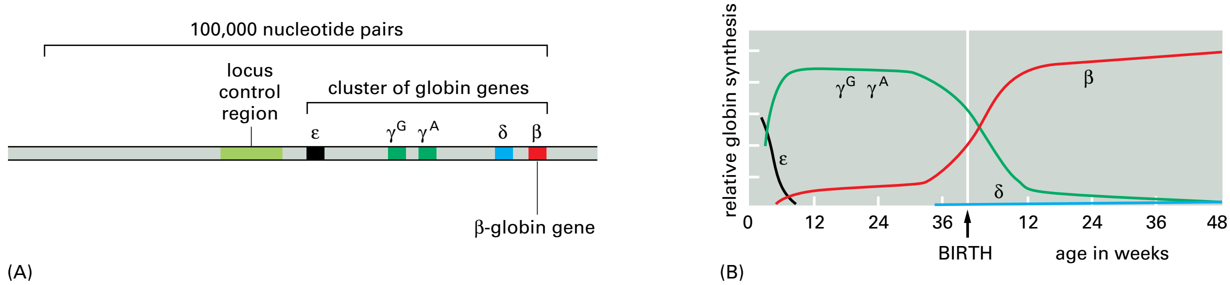


**Figure 7–60 The cluster of β-like globin genes in humans.** (A) The large chromosomal region shown spans 100,000 nucleotide pairs and contains the five globin genes and a locus control region (LCR). (B) Changes in the expression of the β-like globin genes at various stages of human development. Each of the globin chains encoded by these genes combines with an α-globin chain to form the hemoglobin in red blood cells (see Figure 7–115). (A, after F. Grosveld, G.B. van Assendelft, D.R. Greaves, and G. Kollias, *Cell* 51:975–985, 1987. © Elsevier.)

looping with proteins bound to the control regions of the genes they regulate. In this way, the proteins bound at the LCR could attract chromatin remodeling complexes and histone modifying enzymes that could alter the chromatin structure of the locus before transcription begins. Other models for LCRs propose a mechanism by which proteins initially bound at the LCR attract other proteins that assemble cooperatively and therefore spread along the DNA toward the genes they control, modifying the chromatin as they proceed.

## Insulators Are DNA Sequences That Prevent Eucaryotic Gene Regulatory Proteins from Influencing Distant Genes

All genes have control regions, which dictate at which times, under what conditions, and in what tissues the gene will be expressed. We also have seen that eucaryotic gene regulatory proteins can act across very long stretches of DNA. How then are control regions of different genes kept from interfering with one another? In other words, what keeps a gene regulatory protein bound on the control region of one gene from inappropriately influencing transcription of adjacent genes?

Several mechanisms have been proposed to account for this regulatory compartmentalization, but the best understood rely on **insulator elements**, also called *boundary elements*. Insulator elements (insulators, for short) are DNA sequences that bind specialized proteins and have two specific properties (Figure 7–61). First, they buffer genes from the repressing effects of heterochromatin. When a gene (from a fly or a mouse, for example) and its normal control region is inserted into different positions in the genome, it is often expressed at levels that vary depending on its site of insertion in the genome and are especially low when it is inserted amid heterochromatin. We saw an example of this *position effect* in Chapter 4, where genes inserted into heterochromatin are transcriptionally silenced (see Figure 4–45). When insulator elements that flank the gene and its control region are included, however, the gene is usually expressed normally, irrespective of its new position in the genome. The second property of insulators is in some sense the converse of this: they can block the action of enhancers (see Figure 7–61). For this to occur, the insulator must be located between the enhancer and the promoter of the target gene.

Thus insulators can define domains of gene expression, both buffering the gene from outside effects and preventing the control region of the gene (or cluster of genes) from acting outside the domain. For example, the globin LCR (discussed above) is associated with a neighboring insulator which allows the LCR to influence only the cluster of globin genes. Presumably, another insulator is located on the distal side of the globin cluster, serving to define the other end of the domain.

The distribution of insulators in a genome is therefore thought to divide it into independent domains of gene regulation and chromatin structure. Consistent with this idea, the distribution of insulators across a genome is roughly correlated with variations in chromatin structure. For example, an insulator-binding protein from flies is localized preferentially to interbands (and also to the edges of puffs) in polytene chromosomes (Figure 7–62).

The mechanisms by which insulators work are not currently understood, and different insulators may function in different ways. At least some pairs of insulators may define the basis of a looped chromosomal domain (see Figure 4–44). It has been proposed that chromosomes of all eucaryotes are divided by insulators into independent looped domains, each regulated separately from all the others.

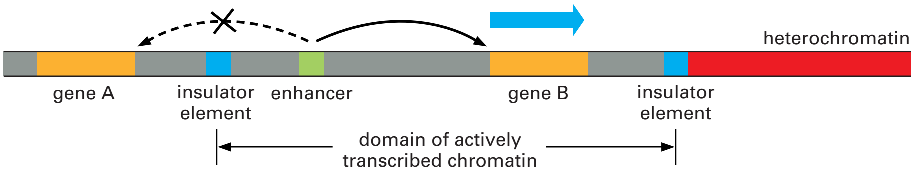


**Figure 7–61 Schematic diagram summarizing the properties of insulators.** Insulators both prevent the spread of heterochromatin *(right-hand side* of diagram) and directionally block the action of enhancers *(left-hand side)*. Thus gene B is properly regulated and gene B's enhancer is prevented from influencing the transcription of gene A.

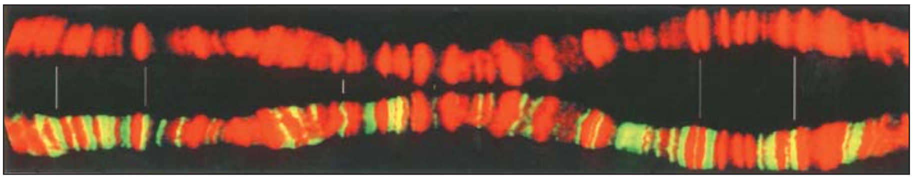


**Figure 7–62 Localization of a *Drosophila* insulator-binding protein on polytene chromosomes.** A polytene chromosome (see pp. 218–220) was stained with propidium iodide *(red)* to show its banding patterns—with bands appearing *bright red* and interbands as dark gaps in the pattern *(top)*. The positions on this polytene chromosome that are bound by a particular insulator protein (called BEAF) are stained *bright green* using antibodies directed against the protein *(bottom)*. BEAF is seen to be preferentially localized to interband regions suggesting a role for it in organizing domains of chromatin. For convenience, these two micrographs of the same polytene chromosome are arranged as mirror images. (Courtesy of Uli Laemmli, from K. Zhao et al., *Cell* 81:879–889, 1995. © Elsevier.)

## Bacteria Use Interchangeable RNA Polymerase Subunits to Help Regulate Gene Transcription

We have seen the importance of gene regulatory proteins that bind to regulatory sequences in DNA and signal to the transcription apparatus whether or not to start the synthesis of an RNA chain. Although this is the main way of controlling transcriptional initiation in both eucaryotes and procaryotes, some bacteria and their viruses use an additional strategy based on interchangeable subunits of RNA polymerase. As described in Chapter 6, a sigma (σ) subunit is required for the bacterial RNA polymerase to recognize a promoter. Many bacteria make several different sigma subunits, each of which can interact with the RNA polymerase core and direct it to a different set of promoters (Table 7–2). This scheme permits one large set of genes to be turned off and a new set to be turned on simply by replacing one sigma subunit with another; the strategy is efficient because it bypasses the need to deal with the genes one by one. It is often used subversively by bacterial viruses to take over the host polymerase and activate several sets of viral genes rapidly and sequentially (Figure 7–63).

In a sense, eucaryotes employ an analogous strategy through the use of three distinct RNA polymerases (I, II, and III) that share some of their subunits. Procaryotes, in contrast, use only one type of core RNA polymerase molecule, but they modify it with different sigma subunits.

## Gene Switches Have Gradually Evolved

We have seen that the control regions of eucaryotic genes are often spread out over long stretches of DNA, whereas those of procaryotic genes are typically closely packed around the start point of transcription. Several bacterial gene regulatory proteins, however, recognize DNA sequences that are located many nucleotide pairs away from the promoter, as we saw in Figure 7–40. This case provided one of the first examples of DNA looping in gene regulation and greatly influenced later studies of eucaryotic gene regulatory proteins.

It seems likely that the close-packed arrangement of bacterial genetic switches developed from more extended forms of switches in response to the evolutionary pressure on bacteria to maintain a small genome size. This compression comes at a price, however, as it restricts the complexity and adaptability of the control device. The extended form of eucaryotic control regions, in contrast, with discrete regulatory modules separated by long stretches of spacer DNA, would be expected to facilitate a reshuffling of the regulatory modules during evolution, both to create new regulatory circuits and to modify old ones. Unraveling the history of how gene control regions evolved presents a fascinating challenge, and many clues can be found in present-day DNA sequences. We shall take up this issue again at the end of this chapter.

**TABLE 7–2 Sigma Factors of *E. coli***

| SIGMA FACTOR | PROMOTERS RECOGNIZED |
|---|---|
| $\sigma^{70}$ | most genes |
| $\sigma^{32}$ | genes induced by heat shock |
| $\sigma^{28}$ | genes for stationary phase and stress response |
| $\sigma^{28}$ | genes involved in motility and chemotaxis |
| $\sigma^{54}$ | genes for nitrogen metabolism |

The sigma factor designations refer to their approximate molecular weights, in kilodaltons.

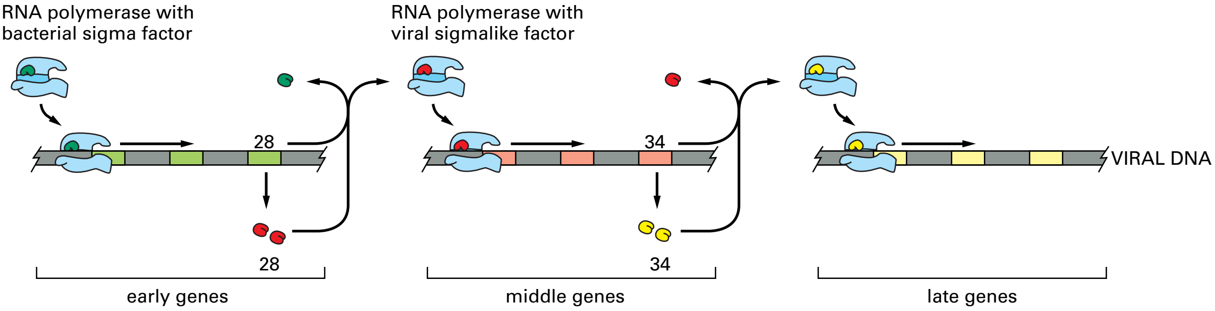

RNA polymerase with bacterial sigma factor
RNA polymerase with viral sigmalike factor
28
28
34
34
VIRAL DNA
early genes
middle genes
late genes

**Figure 7–63 Interchangeable RNA polymerase subunits as a strategy to control gene expression in a bacterial virus.** The bacterial virus SPO1, which infects the bacterium *B. subtilis,* uses the bacterial polymerase to transcribe its early genes immediately after the viral DNA enters the cell. One of the early genes, called 28, encodes a sigmalike factor that binds to RNA polymerase and displaces the bacterial sigma factor. This new form of polymerase specifically initiates transcription of the SPO1 "middle" genes. One of the middle genes encodes a second sigmalike factor, 34, that displaces the 28 product and directs RNA polymerase to transcribe the "late" genes. This last set of genes produces the proteins that package the virus chromosome into a virus coat and lyse the cell. By this strategy, sets of virus genes are expressed in the order in which they are needed; this ensures a rapid and efficient viral replication.

## Summary

*The transcription of individual genes is switched on and off in cells by gene regulatory proteins. In procaryotes these proteins usually bind to specific DNA sequences close to the RNA polymerase start site and, depending on the nature of the regulatory protein and the precise location of its binding site relative to the start site, either activate or repress transcription of the gene. The flexibility of the DNA helix, however, also allows proteins bound at distant sites to affect the RNA polymerase at the promoter by the looping out of the intervening DNA. Such action at a distance is extremely common in eucaryotic cells, where gene regulatory proteins bound to sequences thousands of nucleotide pairs from the promoter generally control gene expression. Eucaryotic activators and repressors act by a wide variety of mechanisms—generally causing the local modification of chromatin structure, the assembly of the general transcription factors at the promoter, and the recruitment of RNA polymerase.*

*Whereas the transcription of a typical procaryotic gene is controlled by only one or two gene regulatory proteins, the regulation of higher eucaryotic genes is much more complex, commensurate with the larger genome size and the large variety of cell types that are formed. The control region of the Drosophila eve gene, for example, encompasses 20,000 nucleotide pairs of DNA and has binding sites for over 20 gene regulatory proteins. Some of these proteins are transcriptional activators, whereas others are transcriptional repressors. These proteins bind to regulatory sequences organized in a series of regulatory modules strung together along the DNA, and together they cause the correct spatial and temporal pattern of gene expression. Eucaryotic genes and their control regions are often surrounded by insulators, DNA sequences recognized by proteins that prevent cross-talk between independently regulated genes.*

## THE MOLECULAR GENETIC MECHANISMS THAT CREATE SPECIALIZED CELL TYPES

Although all cells must be able to switch genes on and off in response to changes in their environments, the cells of multicellular organisms have evolved this capacity to an extreme degree and in highly specialized ways to form an organized array of differentiated cell types. In particular, once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the choice of fate is generally maintained through many subsequent cell generations, which means that the changes in gene expression involved in the choice must be remembered. This phenomenon of *cell memory* is a prerequisite for the creation of organized tissues and for the maintenance of stably differentiated cell types. In contrast, the simplest changes in gene expression in both eucaryotes and bacteria are only transient; the tryptophan repressor, for example, switches off the tryptophan genes in bacteria only in the presence of tryptophan; as soon as tryptophan is removed from the medium, the genes are switched back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan. Even in bacteria, however, a few types of changes in gene expression can be inherited stably.

In this section we examine how gene regulatory devices can be combined to create “logic” circuits through which cells differentiate, keep time, remember events in their past, and adjust the levels of gene expression over whole chromosomes. We begin by considering some of the best-understood genetic mechanisms of cell differentiation, which operate in bacterial and yeast cells.

## DNA Rearrangements Mediate Phase Variation in Bacteria

We have seen that cell differentiation in higher eucaryotes usually occurs without detectable changes in DNA sequence. In some procaryotes, in contrast, a stably inherited pattern of gene regulation is achieved by DNA rearrangements that activate or inactivate specific genes. Since changes in DNA sequence are copied faithfully during subsequent DNA replications, an altered state of gene activity will be inherited by all the progeny of the cell in which the rearrangement occurred. Some of these DNA rearrangements are, however, reversible so that occasional individuals switch back to original DNA configurations. The result is an alternating pattern of gene activity that can be detected by observations over long time periods and many generations.

A well-studied example of this differentiation mechanism occurs in *Salmonella* bacteria and is known as **phase variation**. Although this mode of differentiation has no known counterpart in higher eucaryotes, it can nevertheless have considerable impact on them because disease-causing bacteria use it to evade detection by the immune system. The switch in *Salmonella* gene expression is brought about by the occasional inversion of a specific 1000-nucleotide-pair piece of DNA. This change alters the expression of the cell-surface protein flagellin, for which the bacterium has two different genes (Figure 7–64). The inversion is catalyzed by a site-specific recombination enzyme and changes the orientation of a promoter that is within the 1000 nucleotide pairs. With the promoter in one orientation, the bacteria synthesize one type of flagellin; with the promoter in the other orientation, they synthesize the other type. Because inversions occur only rarely, whole clones of bacteria will grow up with one type of flagellin or the other.

Phase variation almost certainly evolved because it protects the bacterial population against the immune response of its vertebrate host. If the host makes antibodies against one type of flagellin, a few bacteria whose flagellin has been altered by gene inversion will still be able to survive and multiply.

Bacteria isolated from the wild very often exhibit phase variation for one or more phenotypic traits. These “instabilities” are usually lost with time from standard laboratory strains of bacteria, and underlying mechanisms have been studied in only a few cases. Not all involve DNA inversion. A bacterium that

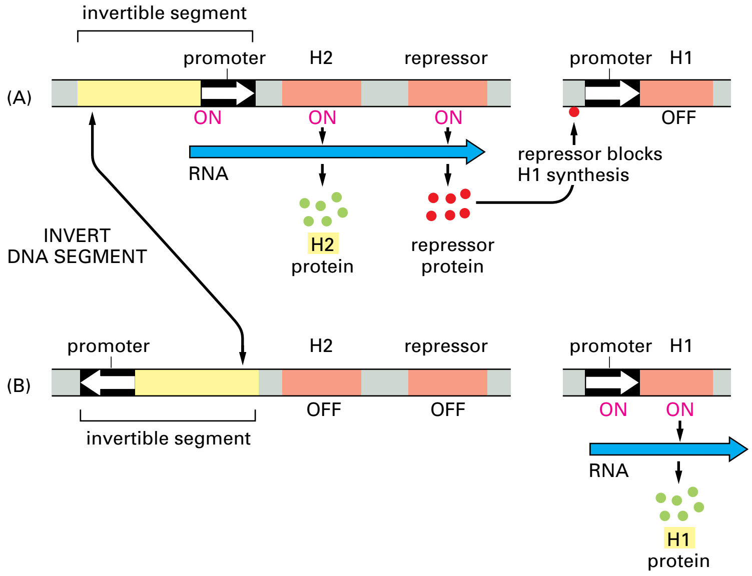


**Figure 7–64 Switching gene expression by DNA inversion in bacteria.** Alternating transcription of two flagellin genes in a *Salmonella* bacterium is caused by a simple site-specific recombination event that inverts a small DNA segment containing a promoter. (A) In one orientation, the promoter activates transcription of the H2 flagellin gene as well as that of a repressor protein that blocks the expression of the H1 flagellin gene. (B) When the promoter is inverted, it no longer turns on H2 or the repressor, and the H1 gene, which is thereby released from repression, is expressed instead. The recombination mechanism is activated only rarely (about once every $10^5$ cell divisions). Therefore, the production of one or other flagellin tends to be faithfully inherited in each clone of cells.

causes a common sexually transmitted human disease *(Neisseria gonorrhoeae),* for example, avoids immune attack by means of a heritable change in its surface properties that is generated by gene conversion (discussed in Chapter 5) rather than by inversion. This mechanism transfers DNA sequences from a library of silent "gene cassettes" to a site in the genome where the genes are expressed; it has the advantage of creating many variants of the major bacterial surface protein.

## A Set of Gene Regulatory Proteins Determines Cell Type in a Budding Yeast

Because they are so easy to grow and to manipulate genetically, yeasts have served as model organisms for studying the mechanisms of gene control in eucaryotic cells. The common baker's yeast, *Saccharomyces cerevisiae*, has attracted special interest because of its ability to differentiate into three distinct cell types. *S. cerevisiae* is a single-celled eucaryote that exists in either a haploid or a diploid state. Diploid cells form by a process known as **mating**, in which two haploid cells fuse. In order for two haploid cells to mate, they must differ in *mating type* (sex). In *S. cerevisiae* there are two mating types, α and **a**, which are specialized for mating with each other. Each produces a specific diffusible signaling molecule (mating factor) and a specific cell-surface receptor protein. These jointly enable a cell to recognize and be recognized by its opposite cell type, with which it then fuses. The resulting diploid cells, called **a**/α, are distinct from either parent: they are unable to mate but can form spores (sporulate) when they run out of food, giving rise to haploid cells by the process of meiosis (discussed in Chapter 20).

The mechanisms by which these three cell types are established and maintained illustrate several of the strategies we have discussed for changing the pattern of gene expression. The mating type of the haploid cell is determined by a single locus, the **mating-type (MAT) locus**, which in an **a**-type cell encodes a single gene regulatory protein, a1, and in an α cell encodes two gene regulatory proteins, Matα1 and Matα2. The Mata1 protein has no effect in the **a**-type haploid cell that produces it but becomes important later in the diploid cell that results from mating; meanwhile, the α-type haploid cell produces the proteins specific to its mating type by default. In contrast, the α2 protein acts in the α cell as a transcriptional repressor that turns off the **a**-specific genes, while the α1 protein acts as a transcriptional activator that turns on the α-specific genes. Once cells of the two mating types have fused, the combination of the a1 and α2 regulatory proteins generates a completely new pattern of gene expression, unlike that of either parent cell. Figure 7–65 illustrates the mechanism by which the mating-type-specific genes are expressed in different patterns in the three cell types. This was among the first examples of combinatorial gene control to be identified, and it remains one of the best understood at the molecular level.

Although in most laboratory strains of *S. cerevisiae,* the **a** and α cell types are stably maintained through many cell divisions, some strains isolated from the wild can switch repeatedly between the **a** and α cell types by a mechanism of gene rearrangement whose effects are reminiscent of the DNA rearrangements in *N. gonorrhoeae*, although the exact mechanism seems to be peculiar to yeast. On either side of the *MAT* locus in the yeast chromosome, there is a silent locus encoding the mating-type gene regulatory proteins: the silent locus on one side encodes α1 and α2; the silent locus on the other side encodes a1. Approximately every other cell division, the active gene in the *MAT* locus is excised and replaced by a newly synthesized copy of the silent locus determining the opposite mating type. Because the change involves the removal of one gene from the active "slot" and its replacement by another, this mechanism is called the *cassette mechanism*. The change is reversible because, although the original gene at the *MAT* locus is discarded, a silent copy remains in the genome. New DNA copies made from the silent genes function as disposable cassettes that will be inserted in alternation into the *MAT* locus, which serves as the "playing head" (Figure 7–66).

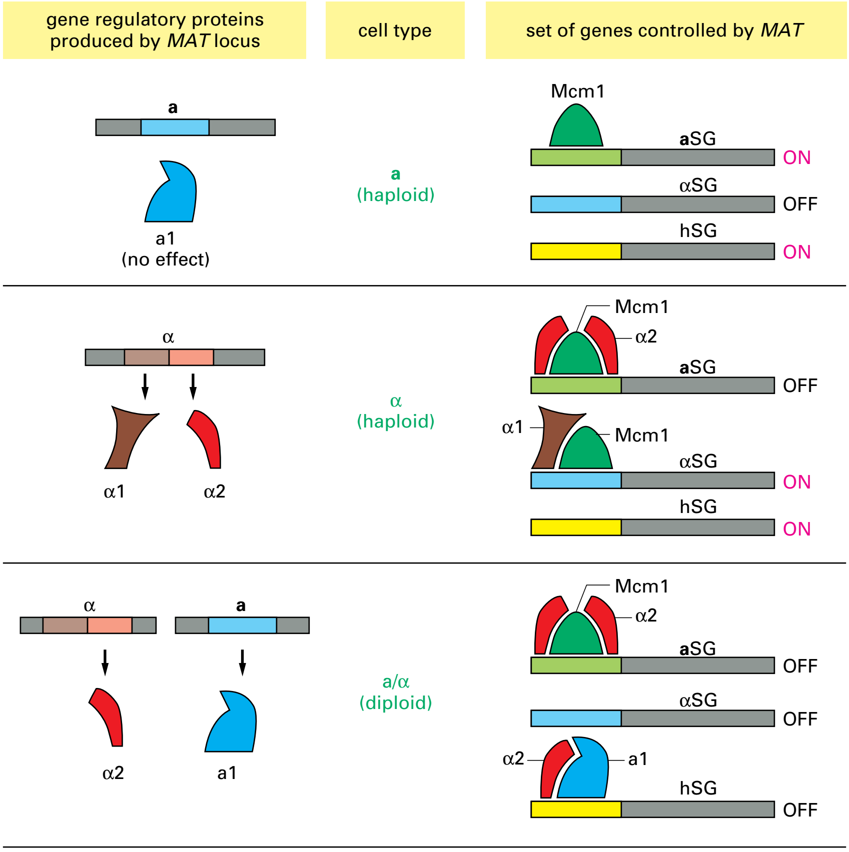


**Figure 7–65 Control of cell type in yeasts.** Yeast cell type is determined by three gene regulatory proteins (α1, α2, and a1) produced by the *MAT* locus. Different sets of genes are transcribed in haploid cells of type **a**, in haploid cells of type α, and in diploid cells (type a/α). The haploid cells express a set of haploid-specific genes (hSG) and either a set of α-specific genes (αSG) or a set of a-specific genes (aSG). The diploid cells express none of these genes. The α1, α2, and a1 proteins control many target genes in each type of cell by binding, in various combinations, to specific regulatory sequences upstream of these genes. Note that the α1 protein is a gene activator protein, whereas the α2 protein is a gene repressor protein. Both work in combination with a gene regulatory protein called Mcm1 that is present in all three cell types. In the diploid cell type, α2 and a1 form a heterodimer (shown in detail in Figure 7–23) that turns off a set of genes (including the gene encoding the α1 activator protein) different from that turned off by the α2 and Mcm1 proteins. This relatively simple system of gene regulatory proteins is an example of combinatorial control of gene expression (see Figure 7–50). The a1 and α2 proteins each recognize their DNA-binding sites using a homeodomain motif (see Figure 7–16).

The silent cassettes are maintained in a transcriptionally inactive form by the same mechanism that is responsible for silencing genes located at the ends of the yeast chromosomes (see Figure 4–47); that is, the DNA at a silent locus is packaged into a highly organized form of chromatin that is resistant to transcription.

## Two Proteins That Repress Each Other's Synthesis Determine the Heritable State of Bacteriophage Lambda

The observation that a whole vertebrate or plant can be specified by the genetic information present in a single somatic cell nucleus (see Figure 7–2) eliminates the possibility that an irreversible change in DNA sequence is a major

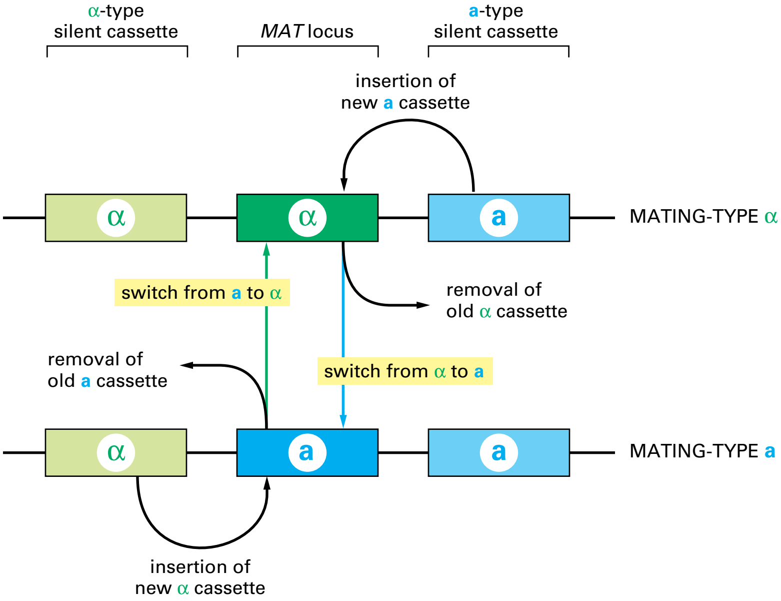


**Figure 7–66 Cassette model of yeast mating-type switching.** Cassette switching occurs by a gene-conversion process that involves a specialized enzyme (the HO endonuclease) that makes a double-stranded cut at a specific DNA sequence in the *MAT* locus. The DNA near the cut is then excised and replaced by a copy of the silent cassette of opposite mating type. The mechanism of this specialized form of gene conversion is similar to the general mechanism of homologous end joining discussed in Chapter 5 (see pp. 283–284).

mechanism in the differentiation of higher eucaryotic cells (although such changes are a crucial part of lymphocyte differentiation—discussed in Chapter 24). Reversible DNA sequence changes, resembling those just described for *Salmonella* and yeasts, in principle could still be responsible for some of the inherited changes in gene expression observed in higher organisms, but there is currently no evidence that such mechanisms are widely used.

Other mechanisms that we have touched upon in this chapter, however, are also capable of producing patterns of gene regulation that can be inherited by subsequent cell generations. Perhaps the simplest example is found in the bacterial virus (bacteriophage) lambda where a switch causes the virus to flip-flop between two stable self-maintaining states. This type of switch can be viewed as a prototype for similar, but more complex, switches that operate in the development of higher eucaryotes.

We mentioned earlier that bacteriophage lambda can in favorable conditions become integrated into the *E. coli* cell DNA, to be replicated automatically each time the bacterium divides. Alternatively, the virus can multiply in the cytoplasm, killing its host (see Figure 5–81). The switch between these two states is mediated by proteins encoded by the bacteriophage genome. The genome contains a total of about 50 genes, which are transcribed in very different patterns in the two states. A virus destined to integrate, for example, must produce the lambda *integrase* protein, which is needed to insert the lambda DNA into the bacterial chromosome, but must repress production of the viral proteins responsible for virus multiplication. Once one transcriptional pattern or the other has been established, it is stably maintained.

We cannot discuss the details of this complex gene regulatory system here and instead outline a few of its general features. At the heart of the switch are two gene regulatory proteins synthesized by the virus: the **lambda repressor protein** (cI protein), which we have already encountered, and the **Cro protein**. These proteins repress each other's synthesis, an arrangement giving rise to just two stable states (Figure 7–67). In state 1 (the *prophage state)* the lambda repressor occupies the operator, blocking the synthesis of Cro and also activating its own synthesis. In state 2 (the *lytic state)* the Cro protein occupies a different site in the operator, blocking the synthesis of repressor but allowing its own synthesis. In the prophage state most of the DNA of the stably integrated bacteriophage is not transcribed; in the lytic state, this DNA is extensively transcribed, replicated, packaged into new bacteriophage, and released by host cell lysis.

When the host bacteria are growing well, an infecting virus tends to adopt state 1, allowing the DNA of the virus to multiply along with the host chromosome. When the host cell is damaged, an integrated virus converts from state 1 to state 2 in order to multiply in the cell cytoplasm and make a quick exit. This conversion is triggered by the host response to DNA damage, which inactivates the repressor protein. In the absence of such interference, however, the lambda repressor both turns off production of the Cro protein and turns on its own synthesis, and this *positive feedback loop* helps to maintain the prophage state.

## Gene Regulatory Circuits Can Be Used to Make Memory Devices As Well As Oscillators

Positive feedback loops provide a simple general strategy for cell memory—that is, for the establishment and maintenance of heritable patterns of gene

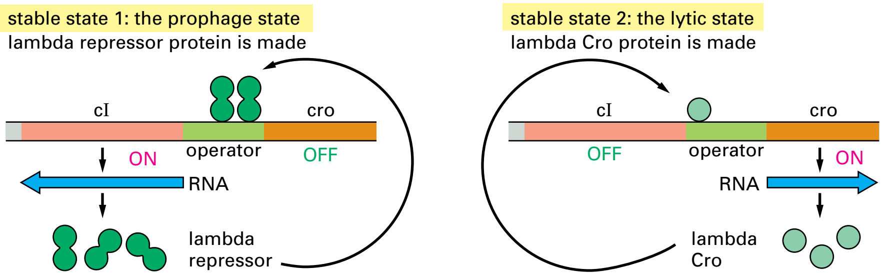

stable state 1: the prophage state
lambda repressor protein is made
cI
cro
operator
ON
OFF
RNA
lambda repressor
stable state 2: the lytic state
lambda Cro protein is made
cI
cro
OFF
operator
ON
RNA
lambda Cro

**Figure 7–67 A simplified version of the regulatory system that determines the mode of growth of bacteriophage lambda in the *E. coli* host cell.** In stable state 1 (the prophage state) the bacteriophage synthesizes a repressor protein, which activates its own synthesis and turns off the synthesis of several other bacteriophage proteins, including the Cro protein. In state 2 (the lytic state) the bacteriophage synthesizes the Cro protein, which turns off the synthesis of the repressor protein, so that many bacteriophage proteins are made and the viral DNA replicates freely in the *E. coli* cell, eventually producing many new bacteriophage particles and killing the cell. This example shows how two gene regulatory proteins can be combined in a circuit to produce two heritable states. Both the lambda repressor and the Cro protein recognize the operator through a helix–turn–helix motif (see Figure 7–14).

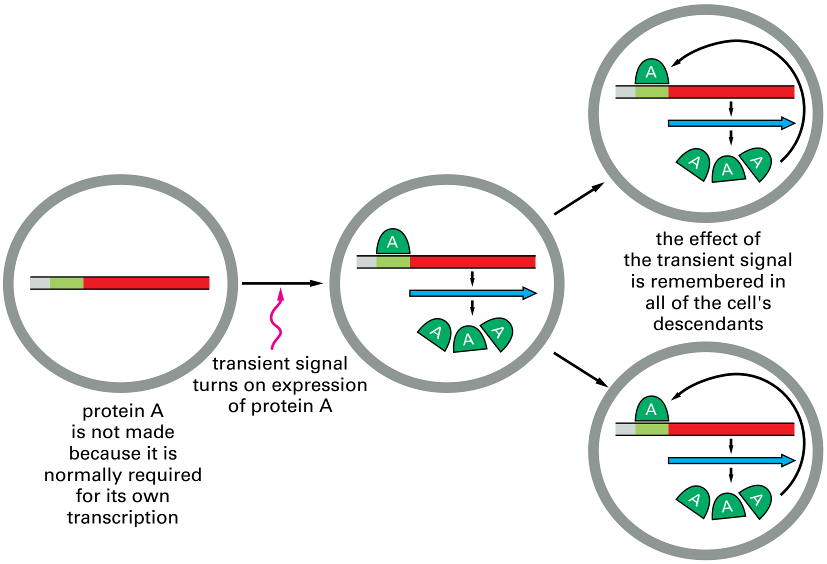

A
A
A
A
the effect of the transient signal is remembered in all of the cell's descendants
A
A
A
A
transient signal turns on expression of protein A
protein A is not made because it is normally required for its own transcription
A
A
A
A

**Figure 7–68 Schematic diagram showing how a positive feedback loop can create cell memory.** Protein A is a gene regulatory protein that activates its own transcription. All of the descendants of the original cell will therefore "remember" that the progenitor cell had experienced a transient signal that initiated the production of the protein.

transcription. Figure 7–68 shows the basic principle, stripped to its barest essentials. Variations of this simple strategy are widely used by eucaryotic cells. Several gene regulatory proteins that are involved in establishing the *Drosophila* body plan (discussed in Chapter 21), for example, stimulate their own transcription, thereby creating a positive feedback loop that promotes their continued synthesis; at the same time many of these proteins repress the transcription of genes encoding other important gene regulatory proteins. In this way, a sophisticated pattern of inherited behavior can be achieved with only a few gene regulatory proteins that reciprocally affect one another's synthesis and activities.

Simple gene regulatory circuits can be combined to create all sorts of control devices, just as simple electronic switching elements in a computer are combined to perform all sorts of complex logical operations. Bacteriophage lambda, as we have seen, provides an example of a circuit that can flip-flop between two stable states. More complex types of regulatory networks are not only found in nature, but can also be designed and constructed in the laboratory. Figure 7–69 shows, for example, how an engineered bacterial cell can switch between three states in a prescribed order, thus functioning as an oscillator or "clock."

## Circadian Clocks Are Based on Feedback Loops in Gene Regulation

Life on Earth evolved in the presence of a daily cycle of day and night, and many present-day organisms (ranging from archaea to plants to humans) have come to possess an internal rhythm that dictates different behaviors at different times of day. These behaviors range from the cyclical change in metabolic enzyme activities of a fungus to the elaborate sleep-wake cycles of humans. The internal oscillators that control such diurnal rhythms are called circadian clocks.

By carrying its own circadian clock, an organism can anticipate the regular daily changes in its environment and take appropriate action in advance. Of course, the internal clock cannot be perfectly accurate, and so it must be capable of being reset by external cues such as the light of day. Thus circadian clocks keep running even when the environmental cues (changes in light and dark) are removed, but the period of this free-running rhythm is generally a little less or a little more than 24 hours. External signals indicating the time of day cause small adjustments in the running of the clock, so as to keep the organism in synchrony with its environment. Following more drastic shifts, circadian cycles become gradually reset (entrained) by the new cycle of light and dark, as anyone who has experienced jet lag can attest.

One might expect that the circadian clock in a complex multicellular creature such as a human would itself be a complex multicellular device, with different groups of cells responsible for different parts of the oscillation mechanism. Remarkably, however, it turns out that in almost all organisms, including humans, the timekeepers are individual cells. Thus, our diurnal cycles of sleeping and waking, body temperature, and hormone release are controlled by a clock that operates in each member of a specialized group of cells (the SCN cells) in the hypothalamus (a part of the brain). Even if these cells are removed from the brain and dispersed in a culture dish, they will continue to oscillate individually, showing a cyclic pattern of gene expression with a period of approximately 24 hours. In the intact body, the SCN cells receive neural cues from the retina, entraining them to the daily cycle of light and dark, and they send information about the time of day to  other tissues such as the pineal gland, which relays the time signal to the rest of the body by releasing the hormone melatonin in time with the clock.

Although the SCN cells have a central role as timekeepers in mammals, it has been shown that they are not the only cells in the mammalian body that have an internal circadian rhythm or an ability to reset it in response to light. Similarly, in *Drosophila,* many different types of cells, including those of the thorax, abdomen, antenna, leg, wing, and testis all continue a circadian cycle when they have been dissected away from the rest of the fly. The clocks in these isolated tissues, like those in the SCN cells, can be reset by externally imposed light and dark cycles.

The working of circadian clocks, therefore, is a fundamental problem in cell biology. Although we do not yet understand all the details, studies in a wide variety of organisms have revealed many of the basic principles and molecular components. For animals, much of what we know has come from searches in *Drosophila* for mutations that make the fly's circadian clock run fast, or slow, or not all; and this work has led to the discovery that many of the same components  are involved in the circadian clock of mammals.  The mechanism of the clock in *Drosophila* is outlined in Figure 7–70. At the heart of the oscillator is a transcriptional feedback loop that has a time delay built into it: accumulation of certain key gene products switches off their transcription, but with a delay, so that—crudely speaking—the cell oscillates between a state where the products are present and transcription is switched off, and one where the products are absent and transcription is switched on.

Despite the relative simplicity of the basic principle behind circadian clocks, the details are complex. One reason for this complexity is that clocks must be buffered against changes in temperature, which typically speed up or slow down macromolecular association. They must also run accurately but be capable of being reset. Although it is not yet understood how biological clocks run at a constant speed despite changes in temperature, the mechanism for resetting the *Drosophila* clock is the light-induced destruction of one of the key gene regulatory proteins, as indicated in Figure 7–70.

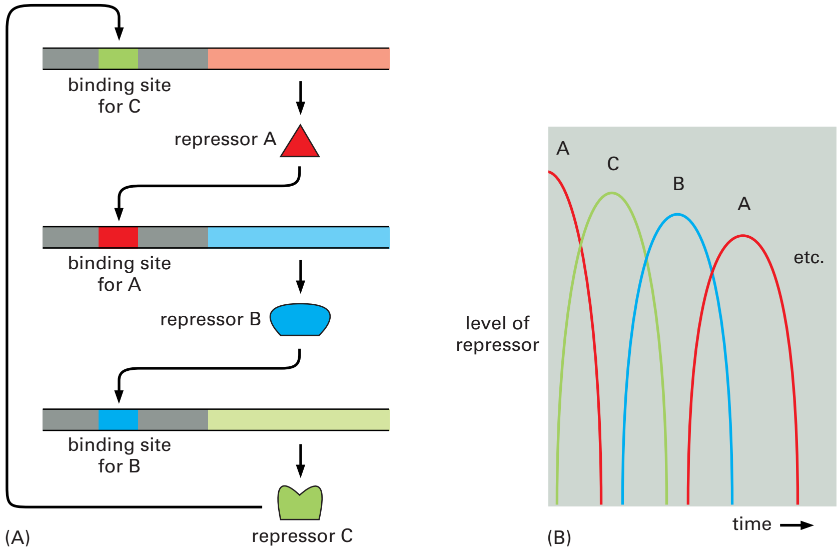


**Figure 7–69 A simple gene clock designed in the laboratory.** (A) Recombinant DNA techniques were used to place the genes for each of three different bacterial repressor proteins under the control of a different repressor. These repressors (denoted A, B, and C in the figure) are the *lac* repressor (see Figure 7–38), the lambda repressor (see Figure 7–67), and the *tet* repressor, which regulates genes in response to tetracycline. (B) A population of cells will cycle between the three different states shown in (A). For example, if the cells start in a state where only repressor A has accumulated to high levels, the gene for repressor B will be fully repressed.  As repressor C is gradually synthesized, it begins to shut off production of repressor A, and repressor B begins to accumulate and eventually shuts off production of repressor C. As this cycling continues a synchronized population of cells oscillates among three states in a specified order. (Adapted from M.B. Elowitz and S. Leibler, *Nature* 403:335–338, 2000.)

## The Expression of a Set of Genes Can Be Coordinated by a Single Protein

Cells need to be able to switch genes on and off individually but they also need to coordinate the expression of large groups of different genes. For example, when a quiescent eucaryotic cell receives a signal to divide, many hitherto unexpressed genes are turned on together to set in motion the events that lead eventually to cell division (discussed in Chapter 18). One way bacteria coordinate the expression of a set of genes is by having them clustered together in an *operon* under control of a single promoter (see Figure 7–33). In eucaryotes, however, each gene is transcribed from a separate promoter.

How do eucaryotes coordinate gene expression? This is an especially important question because, as we have seen, most eucaryotic gene regulatory proteins act as part of a "committee" of regulatory proteins, all of which are necessary to express the gene in the right cell, at the right time, in response to the proper signals, and to the proper level. How then can a eucaryotic cell rapidly and decisively switch whole groups of genes on or off? The answer is that even though control of gene expression is combinatorial, the effect of a single gene regulatory protein can still be decisive in switching any particular gene on or off, simply by completing the combination needed to maximally activate or repress that gene. This situation is analogous to dialing in the final number of a combination lock: the lock will spring open if the other numbers have been previously entered. Just as the same number can complete the combination for different locks, the same protein can complete the combination for several different genes. If a number of different genes contain the regulatory site for the same gene regulatory protein, it can be used to regulate the expression of all of them.

An example of this in humans is the control of gene expression by the *glucocorticoid receptor protein*. To bind to regulatory sites in DNA, this gene regulatory protein must first form a complex with a molecule of a glucocorticoid steroid hormone, such as cortisol (see Figures 15–12 and 15–13). This hormone is released in the body during times of starvation and intense physical activity, and among its other activities, it stimulates cells in the liver to increase the production of glucose from amino acids and other small molecules. To make this response, liver cells increase the expression of many different genes, coding for metabolic enzymes and other products. Although these genes all have different and complex control regions, their maximal expression depends on the binding of the hormone-glucocorticoid receptor complex to a regulatory site in the DNA of each gene. When the body has recovered and the hormone is no longer present, the expression of each of these genes drops to its normal level in the liver. In this way a single gene regulatory protein can control the expression of many different genes (Figure 7–71).

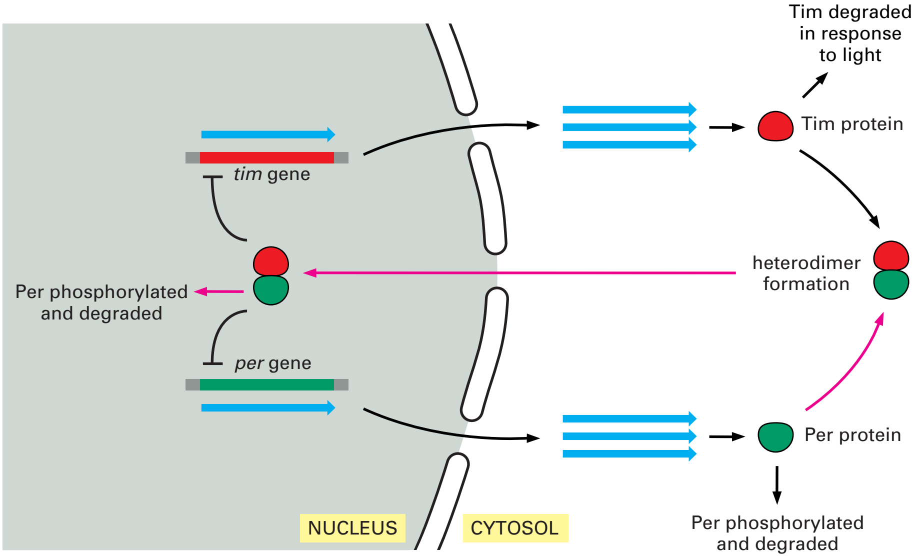

Tim degraded in response to light
Tim protein
tim gene
heterodimer formation
Per phosphorylated and degraded
per gene
Per protein
Per phosphorylated and degraded
NUCLEUS
CYTOSOL

**Figure 7–70 Outline of the mechanism of the circadian clock in *Drosophila* cells.** The central feature of the clock is the periodic accumulation and decay of two gene regulatory proteins, Tim (short for timeless, based on the phenotype of a gene mutation) and Per (short for period). These proteins are translated in the cytosol, and, when they have accumulated to critical levels, they form a heterodimer. This heterodimer is transported into the nucleus where it regulates a number of genes in concert with the clock. The Tim–Per heterodimer also represses the *tim* and *per* genes, creating a feedback system that causes the levels of Tim and Per to rise and fall periodically. In addition to this transcriptional feedback, the clock depends on the phosphorylation and subsequent degradation of the Per protein, which occurs in both the nucleus and the cytoplasm and is regulated by an additional clock protein, Dbt (short for double-time). This degradation imposes delays in the periodic accumulation of Tim and Per, which are crucial to the functioning of the clock. For example, the accumulation of Per in the cytoplasm is delayed by the phosphorylation and degradation of free Per monomers. Steps at which specific delays are imposed are shown in *red*.

Entrainment (or resetting) of the clock occurs in response to new light-dark cycles. Although most *Drosophila* cells do not have true photoreceptors, light is sensed by intracellular flavoproteins, and it rapidly causes the destruction of the Tim protein, thus resetting the clock.

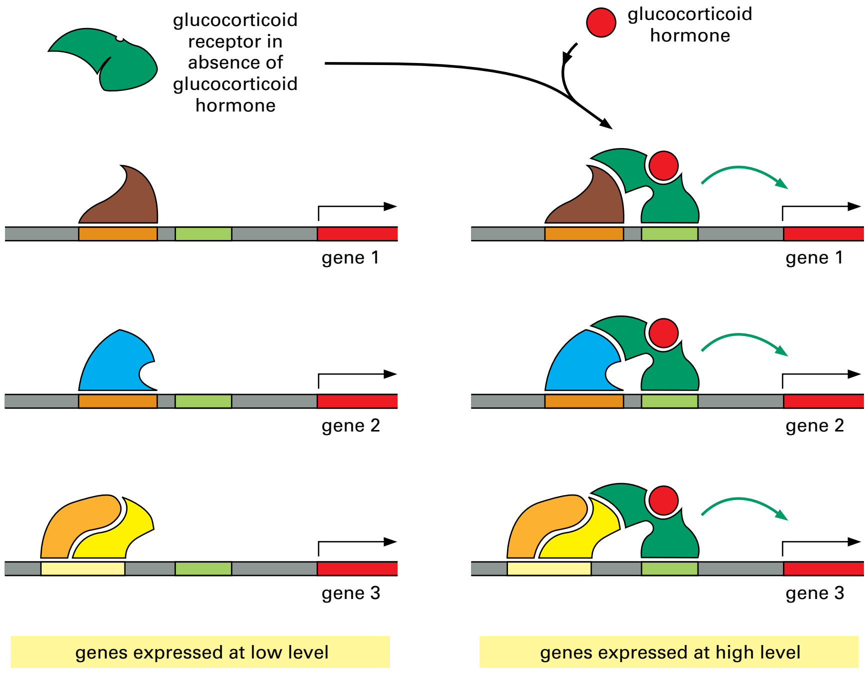

glucocorticoid receptor in absence of glucocorticoid hormone
glucocorticoid hormone
gene 1
gene 2
gene 3
gene 1
gene 2
gene 3
genes expressed at low level
genes expressed at high level

**Figure 7–71 A single gene regulatory protein can coordinate the expression of several different genes.** The action of the glucocorticoid receptor is illustrated schematically. On the *left* is a series of genes, each of which has various gene activator proteins bound to its regulatory region. However, these bound proteins are not sufficient on their own to fully activate transcription. On the *right* is shown the effect of adding an additional gene regulatory protein—the glucocorticoid receptor in a complex with glucocorticoid hormone—that can bind to the regulatory region of each gene. The glucocorticoid receptor completes the combination of gene regulatory proteins required for maximal initiation of transcription, and the genes are now switched on as a set. In the absence of the hormone, the glucocorticoid receptor is retained in the cytosol and is therefore unavailable to bind to DNA. In addition to activating gene expression, the hormone-bound form of the glucocorticoid receptor represses transcription of certain genes, depending on the gene regulatory proteins already present on their control regions. The effect of the glucocorticoid receptor on any given gene therefore depends upon the type of cell, the gene regulatory proteins contained within it, and the regulatory region of the gene. The structure of the DNA-binding portion of the glucocorticoid receptor is shown in Figure 7–19.

The effects of the glucocorticoid receptor are not confined to cells of the liver. In other cell types, activation of this gene regulatory protein by hormone also causes changes in the expression levels of many genes; the genes affected, however, are often different from those affected in liver cells. As we have seen, each cell type has an individualized set of gene regulatory proteins, and because of combinatorial control, these critically affect the action of the glucocorticoid receptor. Because the receptor is able to assemble with many different sets of cell-type specific gene regulatory proteins, it can produce a distinct spectrum of effects in different cell types.

## Expression of a Critical Gene Regulatory Protein Can Trigger Expression of a Whole Battery of Downstream Genes

The ability to switch many genes on or off coordinately is important not only in the day-to-day regulation of cell function. It is also the means by which eucaryotic cells differentiate into specialized cell types during embryonic development. The development of muscle cells provides a striking example.

A mammalian skeletal muscle cell is a highly distinctive giant cell, formed by the fusion of many muscle precursor cells called *myoblasts*, and therefore containing many nuclei. The mature muscle cell is distinguished from other cells by a large number of characteristic proteins, including specific types of actin, myosin, tropomyosin, and troponin (all part of the contractile apparatus), creatine phosphokinase (for the specialized metabolism of muscle cells), and acetylcholine receptors (to make the membrane sensitive to nerve stimulation). In proliferating myoblasts these muscle-specific proteins and their mRNAs are absent or are present in very low concentrations. As myoblasts begin to fuse with one another, the corresponding genes are all switched on coordinately as part of a general transformation of the pattern of gene expression.

This entire program of muscle differentiation can be triggered in cultured skin fibroblasts and certain other cell types by introducing any one of a family of helix–loop–helix proteins—the so-called myogenic proteins (MyoD, Myf5, myogenin, and Mrf4)—normally expressed only in muscle cells (Figure 7–72A). Binding sites for these regulatory proteins are present in the regulatory DNA sequences adjacent to many muscle-specific genes, and the myogenic proteins thereby directly activate transcription of many muscle-specific structural genes. In addition, the myogenic proteins stimulate their own transcription as well as that of various other gene regulatory proteins involved in muscle development, creating an elaborate series of positive feedback loops that amplify and maintain the muscle developmental program, even after the initiating signal has dissipated (Figure 7–72B; see also Chapter 22).

It is probable that the fibroblasts and other cell types that are converted to muscle cells by the addition of myogenic proteins have already accumulated a number of gene regulatory proteins that can cooperate with the myogenic proteins to switch on muscle-specific genes. In this view it is a specific combination of gene regulatory proteins, rather than a single protein, that determines muscle differentiation. This idea is consistent with the finding that some cell types fail to be converted to muscle by myogenin or its relatives; these cells presumably have not accumulated the other gene regulatory proteins required.

The conversion of one cell type (fibroblast) to another (skeletal muscle) by a single gene regulatory protein reemphasizes one of the most important principles discussed in this chapter: dramatic differences between cell types—in size, shape, chemistry, and function—can be produced by differences in gene expression.

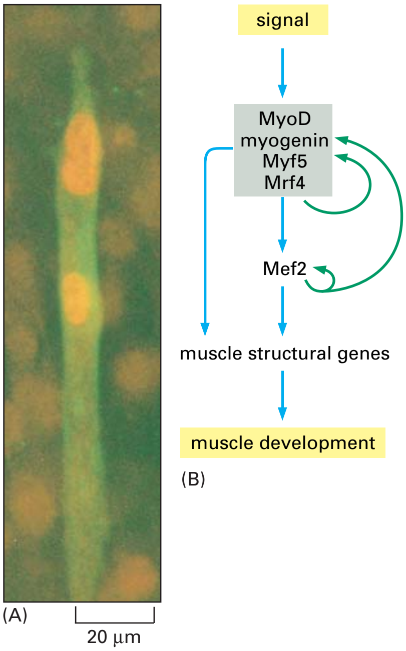


**Figure 7–72 Role of the myogenic regulatory proteins in muscle development.** (A) The effect of expressing the MyoD protein in fibroblasts. As shown in this immunofluorescence micrograph, fibroblasts from the skin of a chick embryo have been converted to muscle cells by the experimentally induced expression of the *myoD* gene. The fibroblasts that have been induced to express the *myoD* gene have fused to form elongated multinucleate muscle-like cells, which are stained *green* with an antibody that detects a muscle-specific protein. Fibroblasts that do not express the *myoD* gene are barely visible in the background. (B) Simplified scheme for some of the gene regulatory proteins involved in skeletal muscle development. The commitment of mesodermal progenitor cells to the muscle-specific pathway involves the synthesis of the four myogenic gene regulatory proteins, MyoD, Myf5, myogenin and Mrf4. These proteins directly activate transcription of muscle structural genes as well as the *MEF2* gene, which encodes an additional gene regulatory protein. Mef2 acts in combination with the myogenic proteins to further activate transcription of muscle structural genes and to create a positive feedback loop that acts to maintain transcription of the myogenic genes. (A, courtesy of Stephen Tapscott and Harold Weintraub; B, adapted from J.D. Molkentin and E.N. Olson, *Proc. Natl. Acad. Sci. USA* 93:9366–9373, 1996.)

## Combinatorial Gene Control Creates Many Different Cell Types in Eucaryotes

We have already discussed how multiple gene regulatory proteins can act in combination to regulate the expression of an individual gene. But, as the example of the myogenic proteins shows, combinatorial gene control means more than this: not only does each gene have many gene regulatory proteins to control it, but each regulatory protein contributes to the control of many genes. Moreover, although some gene regulatory proteins are specific to a single cell type, most are switched on in a variety of cell types, at several sites in the body, and at several times in development. This point is illustrated schematically in Figure 7–73, which shows how combinatorial gene control makes it possible to generate a great deal of biological complexity with relatively few gene regulatory proteins.

With combinatorial control, a given gene regulatory protein does not necessarily have a single, simply definable function as commander of a particular battery of genes or specifier of a particular cell type. Rather, gene regulatory proteins can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event. One requirement of combinatorial control is that many gene regulatory proteins must be able to work together to influence the final rate of transcription. To a remarkable extent, this principle is true: even unrelated gene regulatory proteins from widely different eucaryotic species can cooperate when experimentally introduced into the same cell. This situation reflects both the high degree of conservation of the transcription machinery and the nature of transcriptional activation itself. As we have seen, *transcriptional synergy*, in which multiple activator proteins can show more than additive effects on the final state of transcription, results from the ability of the transcription machinery to respond to multiple inputs (see Figure 7–47). It seems that the multifunctional, combinatorial mode of action of gene regulatory proteins has put a tight constraint on their evolution: they must interact with other gene regulatory proteins, the general transcription factors, the RNA polymerase holoenzyme, and the chromatin-modifying enzymes.

An important consequence of combinatorial gene control is that the effect of adding a new gene regulatory protein to a cell will depend on the cell's past history, since this history will determine which gene regulatory proteins are already present. Thus during development a cell can accumulate a series of gene regulatory proteins that need not initially alter gene expression. When the final member of the requisite combination of gene regulatory proteins is added, however, the regulatory message is completed, leading to large changes in gene expression. Such a scheme, as we have seen, helps to explain how the addition of a single regulatory protein to a fibroblast can produce the dramatic transformation of the fibroblast into a muscle cell. It also can account for the important difference, discussed in Chapter 21, between the process of cell determination—where a cell becomes committed to a particular developmental fate—and the process of cell differentiation, where a committed cell expresses its specialized character.

## The Formation of an Entire Organ Can Be Triggered by a Single Gene Regulatory Protein

We have seen that even though combinatorial control is the norm for eucaryotic genes, a single gene regulatory protein, if it completes the appropriate combination, can be decisive in switching a whole set of genes on or off, and we have seen how this can convert one cell type into another. A dramatic extension of the principle comes from studies of eye development in *Drosophila*, mice, and humans. Here, a gene regulatory protein (called Ey in flies and Pax-6 in vertebrates) is crucial. When expressed in the proper context, Ey can trigger the formation of not just a single cell type but a whole organ (an eye), composed of different types of cells, all properly organized in three-dimensional space.

The most striking evidence for the role of Ey comes from experiments in fruit flies in which the *ey* gene is artificially expressed early in development in groups of cells that normally will go on to form leg parts. This abnormal gene expression causes eyes to develop in the middle of the legs (Figure 7–74). The *Drosophila* eye is composed of thousands of cells, and the question of how a regulatory protein coordinates the specification of a whole array in a tissue is a central topic in *developmental biology*. As discussed in Chapter 21, it involves cell–cell interactions as well as intracellular gene regulatory proteins. Here, we note that Ey directly controls the expression of many other genes by binding to

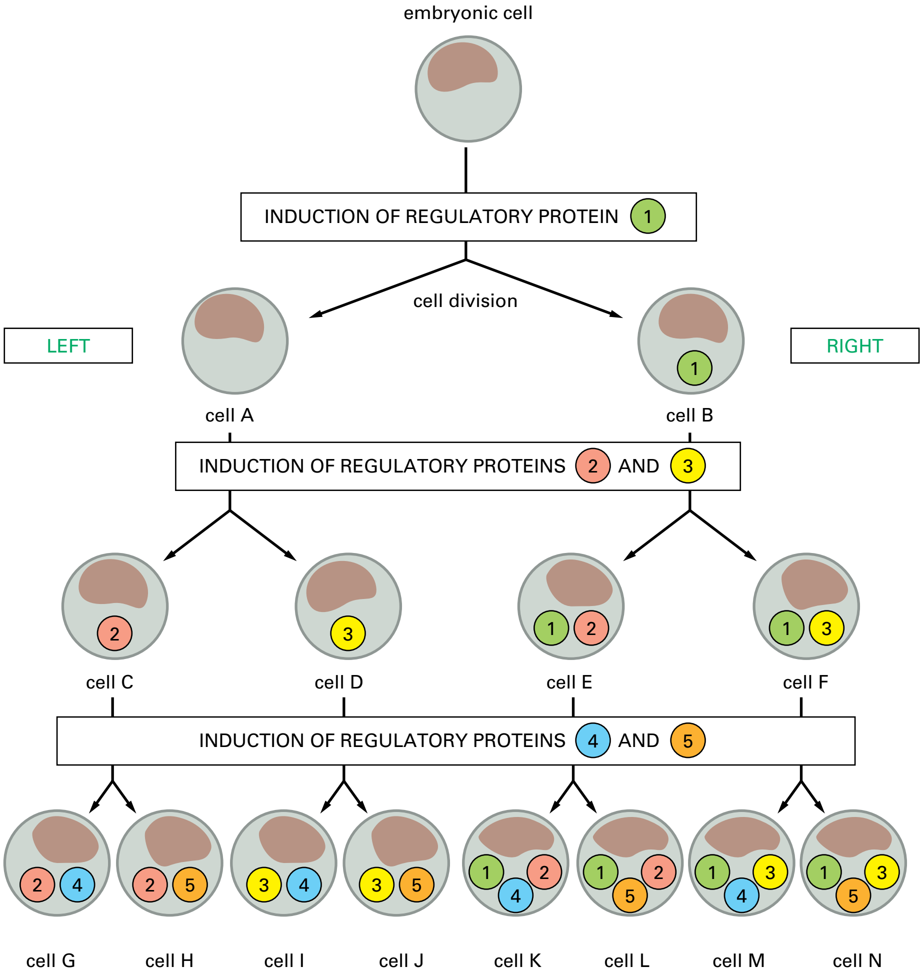


**Figure 7–73 The importance of combinatorial gene control for development.** Combinations of a few gene regulatory proteins can generate many cell types during development. In this simple, idealized scheme a "decision" to make one of a pair of different gene regulatory proteins (shown as numbered *circles)* is made after each cell division. Sensing its relative position in the embryo, the daughter cell toward the *left side* of the embryo is always induced to synthesize the even-numbered protein of each pair, while the daughter cell toward the *right side* of the embryo is induced to synthesize the odd-numbered protein. The production of each gene regulatory protein is assumed to be self-perpetuating once it has become initiated (see Figure 7–68). In this way, through cell memory, the final combinatorial specification is built up step by step. In this purely hypothetical example, eight final cell types (G–N) have been created using five different gene regulatory proteins.

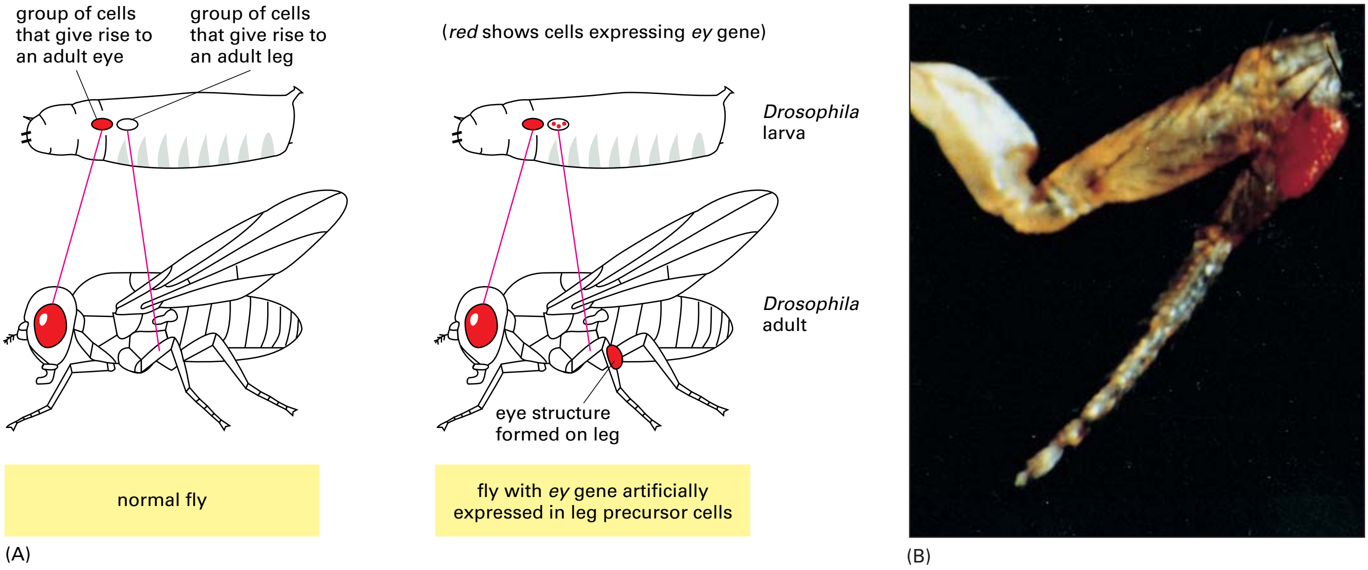

group of cells that give rise to an adult eye
group of cells that give rise to an adult leg
(*red* shows cells expressing *ey* gene)
*Drosophila* larva
*Drosophila* adult
eye structure formed on leg
normal fly
fly with *ey* gene artificially expressed in leg precursor cells
(A)
(B)

**Figure 7–74 Expression of the *Drosophila ey* gene in precursor cells of the leg triggers the development of an eye on the leg.** (A) Simplified diagrams showing the result when a fruit fly larva contains either the normally expressed *ey* gene (*left*) or an *ey* gene that is additionally expressed artificially in cells that normally give rise to leg tissue (*right*). (B) Photograph of an abnormal leg that contains a misplaced eye. (B, courtesy of Walter Gehring.)

their regulatory regions. Some of the genes controlled by Ey are themselves gene regulatory proteins that, in turn, control the expression of other genes. Moreover, some of these regulatory genes act back on *ey* itself to create a positive feedback loop that ensures the continued synthesis of the Ey protein (Figure 7–75). In this way, the action of just one regulatory protein can turn on a cascade of gene regulatory proteins and cell–cell interaction mechanims whose actions result in an organized group of many different types of cells. One can begin to imagine how, by repeated applications of this principle, a complex organism is built up piece by piece.

## Stable Patterns of Gene Expression Can Be Transmitted to Daughter Cells

Once a cell in an organism has become differentiated into a particular cell type, it generally remains specialized in that way, and if it divides, its daughters inherit the same specialized character. For example, liver cells, pigment cells, and endothelial cells (discussed in Chapter 22) divide many times in the life of an individual. This means that the pattern of gene expression specific to a differentiated cell must be remembered and passed on to its progeny through all subsequent cell divisions.

We have already described several ways of ensuring that daughter cells can “remember” what kind of cells they are supposed to be. One of the simplest is through a positive feedback loop (see Figures 7–68, 7–72B and 7–75) where a key gene regulatory protein activates transcription of its own gene (either directly or indirectly) in addition to that of other cell-type specific genes. The simple flip-flop switch shown in Figure 7–67 is a variation on this theme: by inhibiting expression of its own inhibitor, a gene product indirectly activates and maintains its own expression. Another very different way of maintaining cell type in

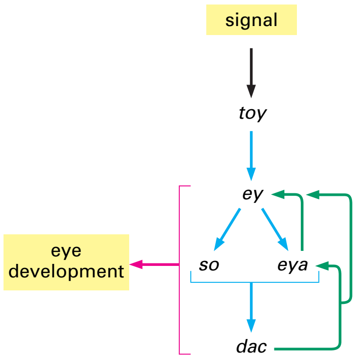


**Figure 7–75 Gene regulatory proteins that specify eye development in *Drosophila*.** *toy (twin of eyeless)* and *ey (eyeless)* encode similar gene regulatory proteins, Toy and Ey, either of which, when ectopically expressed, can trigger eye development. In normal eye development, expression of *ey* requires the *toy* gene. Once its transcription is activated by Toy, Ey activates transcription of *so (sine oculis)* and *eya (eyes absent)* which act together to express the *dac (dachshund)* gene. As indicated by the *green* arrows, some of the gene regulatory proteins form positive feedback loops which reinforce the initial commitment to eye development. The Ey protein is known to bind directly to numerous target genes for eye development, including those encoding lens crystallins (see Figure 7–119), rhodopsins, and other photoreceptor proteins. (Adapted from T. Czerny et al., *Mol. Cell* 3:297–307, 1999.)

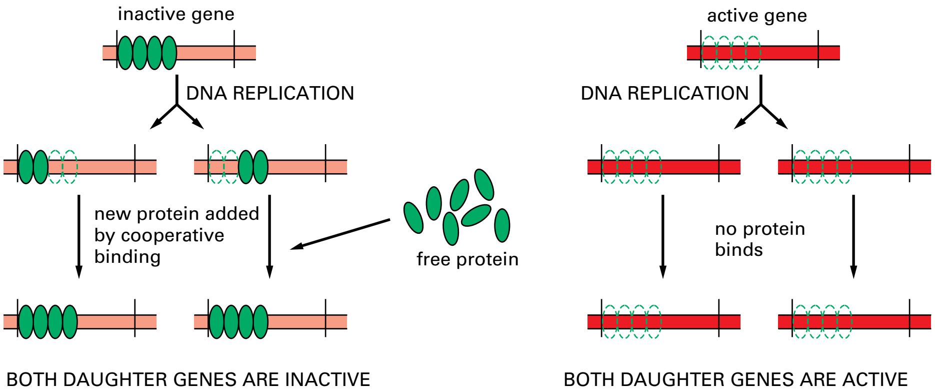

inactive gene
DNA REPLICATION
new protein added by cooperative binding
free protein
BOTH DAUGHTER GENES ARE INACTIVE
active gene
DNA REPLICATION
no protein binds
BOTH DAUGHTER GENES ARE ACTIVE

**Figure 7–76 A general scheme that permits the direct inheritance of states of gene expression during DNA replication.** In this hypothetical model, portions of a cooperatively bound cluster of chromosomal proteins are transferred directly from the parental DNA helix *(top left)* to both daughter helices. The inherited cluster then causes each of the daughter DNA helices to bind additional copies of the same proteins. Because the binding is cooperative, DNA synthesized from an identical parental DNA helix that lacks the bound proteins *(top right)* will remain free of them. If the bound proteins turn off gene transcription, then the inactive gene state will be directly inherited, as illustrated. If the bound proteins activate transcription, then the active gene state will be directly inherited (not shown).

When the cooperative protein binding requires specific DNA sequences, these events will be limited to specific gene control regions; if the binding can be propagated all along the chromosome, however, it could account for the spreading effect associated with the heritable chromatin states discussed in Chapter 4. Although the proteins are depicted as being identical, the same principle can explain how cooperatively assembling combinations of different proteins can be propagated stably.

eucaryotes is through the faithful propagation of chromatin structures from parent to daughter cells, as discussed in Chapter 4. Once a differentiated cell type has been specified by gene regulatory proteins, developmental decisions can be reinforced by packaging unexpressed genes into more compacted forms of chromatin and "marking" that chromatin as silent (see Figure 4–35). The chromatin of actively transcribed genes can also be marked and propagated by the same type of mechanism. The packing of selected regions of the genome into condensed chromatin is a genetic regulatory mechanism that is not available to bacteria, and it is thought to allow eucaryotes to maintain extraordinarily stable patterns of gene expression over many generations. This stability is particularly crucial in multicellular organisms, where abnormal gene expression in a single cell can have profound developmental consequences for the entire organism.

If maintenance of the pattern of gene expression depends on the pattern of chromatin packing, how is this chromatin configuration passed on faithfully from one cell to its daughters? Some possibilities have already been discussed in Chapter 4 (see Figure 4–48). One general mechanism depends on the cooperative binding of proteins to DNA (Figure 7–76). When the cell replicates its DNA, each DNA strand can inherit a share of the protein molecules bound to a given segment of the original double helix, and these inherited molecules can then recruit freshly made molecules to reconstruct a complete copy of the original chromatin complex in each daughter cell. This mechanism of cell memory can be based on cooperative binding of specific gene regulatory proteins, or of general chromatin structural components, or of both classes of molecules acting together. Thus, an initial pattern of binding of specific gene regulatory proteins can initiate a pattern of chromatin condensation that is subsequently maintained.

Yet another strategy of cell memory is based on self-propagating patterns of enzymatic modification of the chromatin proteins (as we saw in Chapter 4) or even of the DNA itself, as we explain later. But first we look more closely at a specific example in which cell memory clearly involves changes of chromatin structure.

## Chromosome Wide Alterations in Chromatin Structure Can Be Inherited

We saw in Chapter 4 that chromatin states can be heritable, and that they can be used to establish and preserve patterns of gene expression over great distances along DNA and for many cell generations. A striking example of such long-range effects of chromatin organization occurs in mammals, where an alteration in the chromatin structure of an entire chromosome is used to modulate levels of expression of all genes on that chromosome.

Males and females differ in their *sex chromosomes*. Females have two X chromosomes, whereas males have one X and one Y chromosome. As a result, female cells contain twice as many copies of X-chromosome genes as do male cells. In mammals, the X and Y sex chromosomes differ radically in gene content: the X chromosome is large and contains more than a thousand genes, whereas the Y chromosome is smaller and contains less than 100 genes. Mammals have evolved a *dosage compensation* mechanism to equalize the dosage of X chromosome gene products between males and females. Mutations that interfere with dosage compensation are lethal, demonstrating the necessity of maintaining the correct ratio of X chromosome to *autosome* (non-sex chromosome) gene products.

In mammals dosage compensation is achieved by the transcriptional inactivation of one of the two X chromosomes in female somatic cells, a process known as **X-inactivation**. Early in the development of a female embryo, when it consists of a few thousand cells, one of the two X chromosomes in each cell becomes highly condensed into a type of heterochromatin. The condensed X chromosome can be easily seen under the light microscope in interphase cells; it was originally called a *Barr body* and is located near the nuclear membrane. As a result of X-inactivation, two X chromosomes can coexist within the same nucleus exposed to the same transcriptional regulatory proteins, yet differ entirely in their expression.

The initial choice of which X chromosome to inactivate, the maternally inherited one ($X_m$) or the paternally inherited one ($X_p$), is random. Once either $X_p$ or $X_m$ has been inactivated, it remains silent throughout all subsequent cell divisions of that cell and its progeny, indicating that the inactive state is faithfully maintained through many cycles of DNA replication and mitosis. Because X-inactivation is random and takes place after several thousand cells have already formed in the embryo, every female is a mosaic of clonal groups of cells in which either $X_p$ or $X_m$ is silenced (Figure 7–77). These clonal groups are distributed in small clusters in the adult animal because sister cells tend to remain close together during later stages of development. For example, X-chromosome inactivation causes the red and black "tortoise-shell" coat coloration of some female cats. In these cats, one X chromosome carries a gene that produces red hair

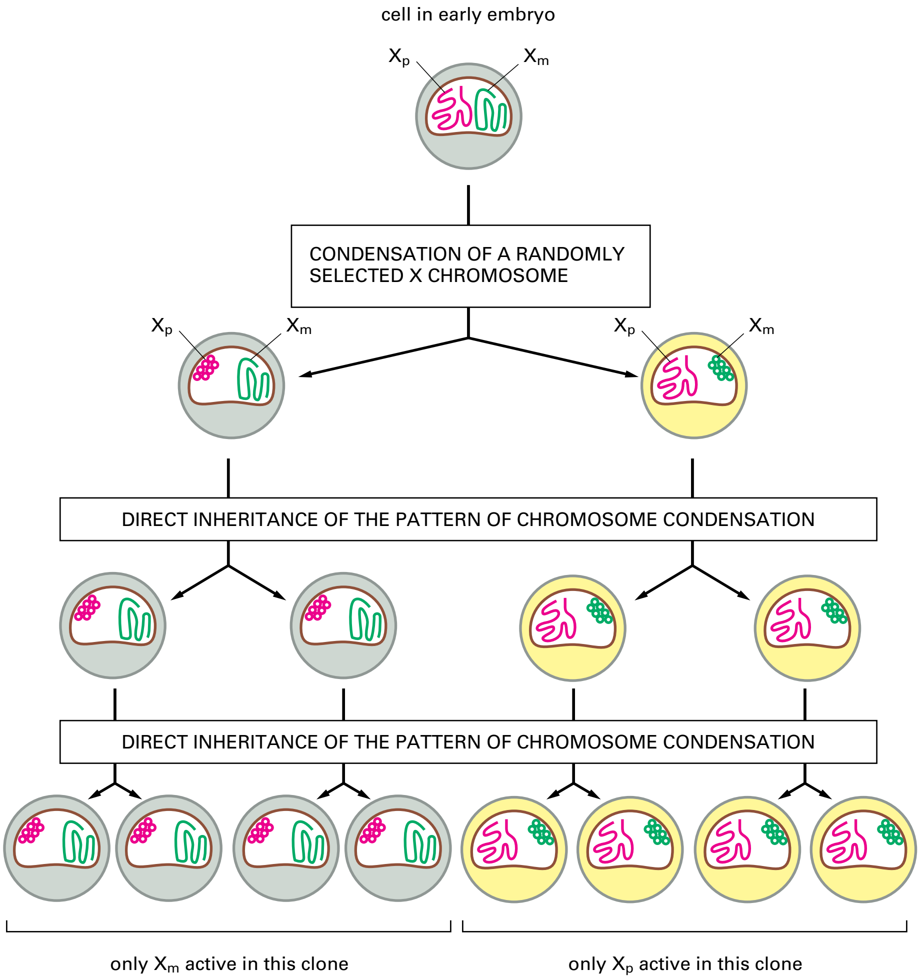

cell in early embryo
$X_p$
$X_m$
CONDENSATION OF A RANDOMLY SELECTED X CHROMOSOME
$X_p$
$X_m$
$X_p$
$X_m$
DIRECT INHERITANCE OF THE PATTERN OF CHROMOSOME CONDENSATION
DIRECT INHERITANCE OF THE PATTERN OF CHROMOSOME CONDENSATION
only $X_m$ active in this clone
only $X_p$ active in this clone

**Figure 7–77 X-inactivation.** The clonal inheritance of a condensed inactive X chromosome that occurs in female mammals.

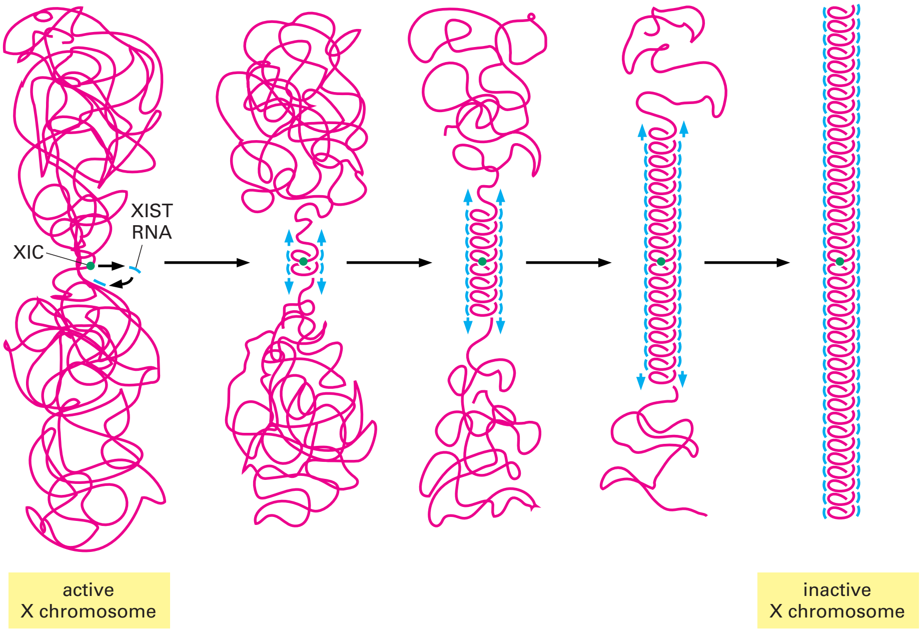


**Figure 7–78 Mammalian X-chromosome inactivation.** X-chromosome inactivation begins with the synthesis of XIST (X-inactivation specific transcript) RNA from the XIC (X-inactivation center) locus. The association of XIST RNA with the X chromosome is correlated with the condensation of the chromosome. Both XIST association and chromosome condensation gradually move from the XIC locus outward to the chromosome ends. The details of how this occurs remain to be deciphered.

color, and the other X chromosome carries an allele of the same gene that results in black hair color; it is the random X-inactivation that produces patches of cells of two distinctive colors. In contrast to the females, male cats of this genetic stock are either solid red or solid black, depending on which X chromosome they inherit from their mothers. Although X-chromosome inactivation is maintained over thousands of cell divisions, it is not always permanent. In particular, it is reversed during germ cell formation, so that all haploid oocytes contain an active X chromosome and can express X-linked gene products.

How is an entire chromosome transcriptionally inactivated? X-chromosome inactivation is initiated and spreads from a single site in the middle of the X chromosome, the **X-inactivation center** (**XIC**). Portions of the X chromosome that are removed from the XIC and fused to an autosome escape inactivation. In contrast, autosomes that are fused to the XIC of an inactive X chromosome are transcriptionally silenced. The XIC (a DNA sequence of approximately $10^6$ nucleotide pairs) can therefore be considered as a large regulatory element that seeds the formation of heterochromatin and facilitates its bi-directional spread along the entire chromosome. Encoded within the XIC is an unusual RNA molecule, *XIST RNA*, which is expressed solely from the inactive X chromosome and whose expression is necessary for X-inactivation. It does not get translated into protein; rather the XIST RNA remains in the nucleus, where it eventually coats the inactive X chromosome. The spread of XIST RNA from the XIC over the entire chromosome correlates with the spread of gene silencing, indicating that XIST RNA participates in the formation and spread of heterochromatin (Figure 7–78). In addition to containing XIST RNA, the X-chromosome heterochromatin is characterized by a specific variant of histone 2A, by hypoacetylation of histones H3 and H4, by methylation of a specific position on histone H3 and by methylation of the underlying DNA, a topic we will discuss below. Presumably all these features make the inactive X chromosome unusually resistant to transcription.

Many features of mammalian X-chromosome inactivation remain to be discovered. How is the initial decision made as to which X chromosome to inactivate? What mechanism prevents the other X chromosome from also being inactivated? How does XIST RNA coordinate the formation of heterochromatin? How is the inactive chromosome maintained through many cell divisions? We are just beginning to understand this mechanism of gene regulation that is crucial for the survival of our own species.

X-chromosome inactivation in females is only one way that sexually reproducing organisms solve the problem of dosage compensation. In *Drosophila,* all the genes on the single X chromosome present in male cells are transcribed at two-fold higher levels than their counterparts in female cells. This male-specific "up-regulation" of transcription results from an alteration in chromatin structure over the entire male X chromosome. As in mammals, this alteration involves the association of a specific RNA molecule with the X chromosome; however, in *Drosophila,* the X-chromosome-associated RNA increases gene activity rather than blocking it. The male X chromosome also contains a specific pattern of histone acetylation which may help to attract the transcription machinery to this chromosome (see Figures 4–35 and 7–46).

Dosage compensation in the nematode worm occurs by a third strategy. Here, the two sexes are male (with one X chromosome) and hermaphrodite (with two X chromosomes), and dosage compensation occurs by a two-fold "down-regulation" of transcription from each of the two X chromosomes in cells of the hermaphrodite. This is brought about through chromosome-wide structural changes in the X chromosomes of hermaphrodites (Figure 7–79). These changes involve the X-specific assembly of proteins, some of which are shared with the *condensins* that helps condense chromosomes during mitosis (see Figures 4–56 and 18–3).

Although the strategies for dosage compensation differ between mammals, flies, and worms, they all involve structural alterations over the entire X chromosome. It is likely that features of chromosome structure that are quite general were adapted and harnessed during evolution to overcome a highly specific problem in gene regulation encountered by sexually reproducing animals.

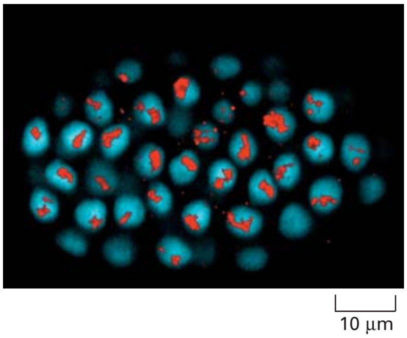


**Figure 7–79 Localization of dosage compensation proteins to the X chromosomes of *C. elegans* hermaphrodite (XX) nuclei.** Many nuclei from a developing embryo are visible in this image. Total DNA is stained *blue* with the DNA-intercalating dye DAPI, and the Sdc-2 protein is stained *red* using anti-Sdc-2 antibodies coupled to a fluorescent dye. This experiment shows that the Sdc-2 protein associates with only a limited set of chromosomes, identified by other experiments to be the two X chromosomes. Sdc-2 is bound along the entire length of the X chromosome and attracts other proteins, including a condensin-like complex, that complete the specialized structure of these chromosomes. (From H.E. Dawes et al., *Science* 284:1800–1804, 1999. © AAAS.)

## The Pattern of DNA Methylation Can Be Inherited When Vertebrate Cells Divide

Thus far, we have emphasized the regulation of gene transcription by proteins that associate with specific DNA sequences. However, DNA itself can be covalently modified, and in the following sections we shall see that this, too, provides opportunities for the regulation of gene expression. In vertebrate cells the methylation of cytosine seems to provide an important mechanism for distinguishing genes that are active from those that are not. The methylated form of cysteine, 5-methylcytosine (5-methyl C), has the same relation to cytosine that thymine has to uracil and the modification likewise has no effect on base-pairing (Figure 7–80). The methylation in vertebrate DNA is restricted to cytosine (C) nucleotides in the sequence CG, which is base-paired to exactly the same sequence (in opposite orientation) on the other strand of the DNA helix. Consequently, a simple mechanism permits the existing pattern of **DNA methylation** to be inherited directly by the daughter DNA strands. An enzyme called *maintenance methyltransferase* acts preferentially on those CG sequences that are base-paired with a CG sequence that is already methylated. As a result, the pattern of DNA methylation on the parental DNA strand serves as a template for the methylation of the daughter DNA strand, causing this pattern to be inherited directly following DNA replication (Figure 7–81).

The stable inheritance of DNA methylation patterns can be explained by maintenance DNA methyltransferases. DNA methylation patterns, however, are dynamic during vertebrate development. Shortly after fertilization there is a genome-wide wave of demethylation, when the vast majority of methyl groups are lost from the DNA. This demethylation may occur either by suppression of maintenance DNA methyltransferase activity, resulting in the passive loss of methyl groups during each round of DNA replication, or by a specific demethylating enzyme. Later in development, at the time that the embryo implants in the wall of the uterus, new methylation patterns are established by several *de novo DNA methyltransferases* that modify specific unmethylated CG dinucleotides. Once the new patterns of methylation are established, they can be propagated through rounds of DNA replication by the maintenance methyl transferases. Mutations in either the maintenance or the *de novo* methyltransferases

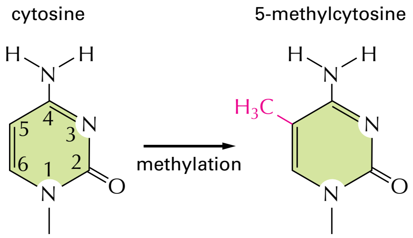
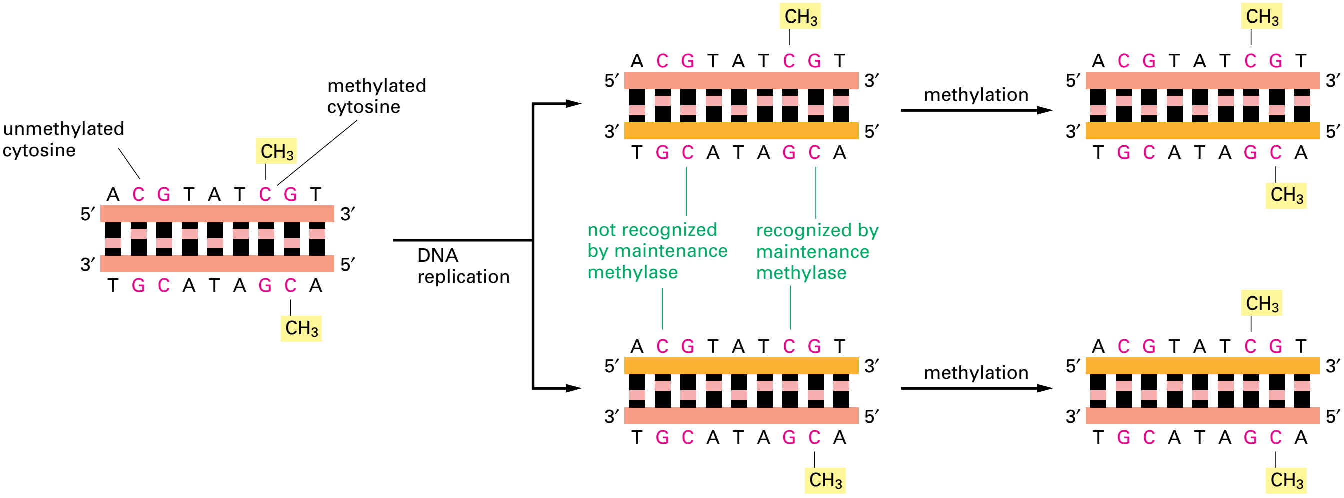


**Figure 7–80 Formation of 5-methylcytosine occurs by methylation of a cytosine base in the DNA double helix.** In vertebrates this event is confined to selected cytosine (C) nucleotides located in the sequence CG.



result in early embryonic death in mice, indicating that establishing and maintaining correct methylation patterns is crucial for normal development.

**Figure 7–81 How DNA methylation patterns are faithfully inherited.** In vertebrate DNAs a large fraction of the cytosine nucleotides in the sequence CG are methylated (see Figure 7–80). Because of the existence of a methyl-directed methylating enzyme (the maintenance methyltransferase), once a pattern of DNA methylation is established, each site of methylation is inherited in the progeny DNA, as shown.

## Vertebrates Use DNA Methylation to Lock Genes in a Silent State

In vertebrates DNA methylation is found primarily on transcriptionally silent regions of the genome, such as the inactive X chromosome or genes that are inactivated in certain tissues, suggesting that it plays a role in gene silencing. Vertebrate cells contain a family of proteins that bind methylated DNA. These DNA-binding proteins, in turn, interact with chromatin remodeling complexes and histone deacetylases that condense chromatin so it becomes transcriptionally inactive. In spite of this, DNA methylation is not sufficient to signal the inactivation of a gene, as the following examples demonstrate. Plasmid DNA encoding a muscle-specific actin gene can be prepared *in vitro* in both fully methylated and fully unmethylated forms, using bacterial proteins that methylate or demethylate DNA. When these two versions of the plasmid are introduced into cultured muscle cells, the methylated plasmid is transcribed at the same high rate as the unmethylated copy. Moreover, when a silent, methylated gene is turned on during the normal course of development, methylation is lost only after the gene has been transcribed for some time. Finally, during X chromosome inactivation, condensation and silencing occur before an increase in levels of DNA methylation can be detected. These results all suggest that methylation reinforces transcriptional repression that is initially established by other mechanisms. DNA methylation seems to be used in vertebrates mainly to ensure that once a gene is turned off, it stays off completely (Figure 7–82).

Experiments designed to test whether a DNA sequence that is transcribed at high levels in one vertebrate cell type is transcribed at all in another have demonstrated that rates of gene transcription can differ between two cell types by a factor of more than $10^6$. Thus unexpressed vertebrate genes are much less “leaky” in terms of transcription than are unexpressed genes in bacteria, in which the largest known differences in transcription rates between expressed and unexpressed gene states are about 1000-fold. DNA methylation of unexpressed vertebrate genes, with the consequent changes in their chromatin structures, accounts for at least part of this difference. Leaky transcription of the many thousands of genes that are normally turned off completely in each vertebrate cell may be the cause of early embryonic death in mice that lack the maintenance DNA methyltransferase.

Transcriptional silencing in vertebrate genomes is also particularly important to repress the proliferation of transposable elements (see Figure 4–17). While coding sequences make up only a few percent of a typical vertebrate genome, transposable elements can comprise nearly half of these genomes. As we saw in Chapter 5, transposable elements can make copies of themselves and

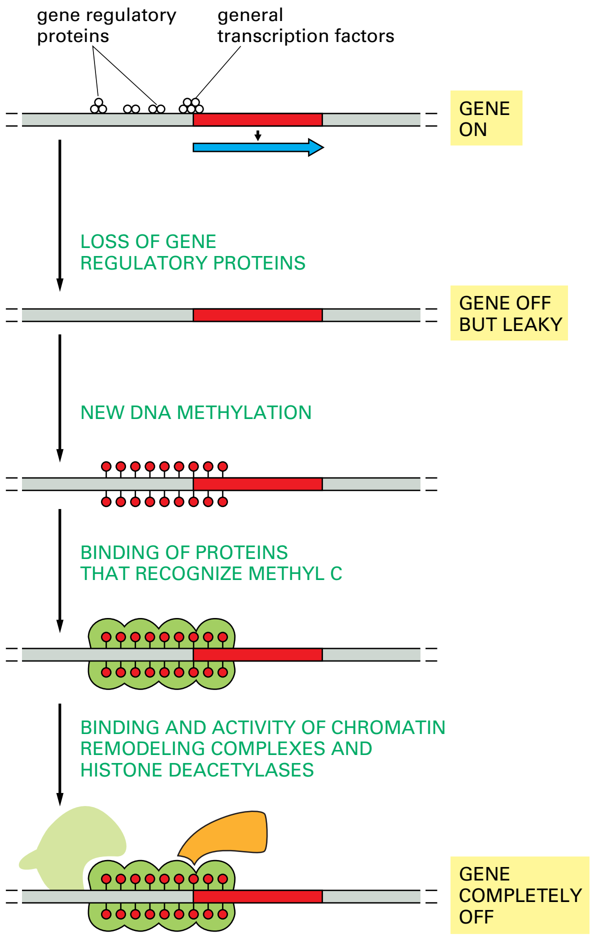

gene regulatory proteins
general transcription factors
GENE ON
LOSS OF GENE REGULATORY PROTEINS
GENE OFF BUT LEAKY
NEW DNA METHYLATION
BINDING OF PROTEINS THAT RECOGNIZE METHYL C
BINDING AND ACTIVITY OF CHROMATIN REMODELING COMPLEXES AND HISTONE DEACETYLASES
GENE COMPLETELY OFF

**Figure 7–82 How DNA methylation may help turn off genes.** The binding of gene regulatory proteins and the general transcription machinery near an active promoter may prevent DNA methylation by excluding *de novo* methylases. If most of these proteins dissociate from the DNA, however, as generally occurs when a cell no longer produces the required activator proteins, the DNA becomes methylated, which enables other proteins to bind, and these shut down the gene completely by further altering chromatin structure (see Figure 7–49).

insert these copies elsewhere in the genome, potentially disrupting genes or important regulatory sequences. By suppressing the transcription of transposable elements, DNA methylation limits their spread and thereby maintains the integrity of the genome. In addition to these varied uses, DNA methylation is also required for at least one special type of cellular memory, as we discuss next.

## Genomic Imprinting Requires DNA Methylation

Mammalian cells are diploid, containing one set of genes inherited from the father and one set from the mother. In a few cases the expression of a gene has been found to depend on whether it is inherited from the mother or the father, a phenomenon called **genomic imprinting**. The gene for *insulin-like growth factor-2* (*Igf2*) is one example of an imprinted gene. *Igf2* is required for prenatal growth, and mice that do not express *Igf2* are born half the size of normal mice. Only the paternal copy of *Igf2* is transcribed. As a result, mice with a mutated paternally derived *Igf2* gene are stunted, while mice with a defective maternally derived *Igf2* gene are normal.

During the formation of germ cells, genes subject to imprinting are marked by methylation according to whether they are present in a sperm or an egg. In this way, the parental origin of the gene can be subsequently detected in the embryo; DNA methylation is thus used as a mark to distinguish two copies of a gene that may be otherwise identical (Figure 7–83). Because imprinted genes are not affected by the wave of demethylation that takes place shortly after fertilization (see p. 430), this mark enables somatic cells to “remember” the parental origin of each of the two copies of the gene and to regulate their expression accordingly. In most cases, the methyl imprint silences nearby gene expression using the mechanisms shown in Figure 7–82. In some cases, however, the methyl imprint can activate expression of a gene. In the case of *Igf2*, for example, methylation of an insulator element (see Figure 7–61) on the paternally derived

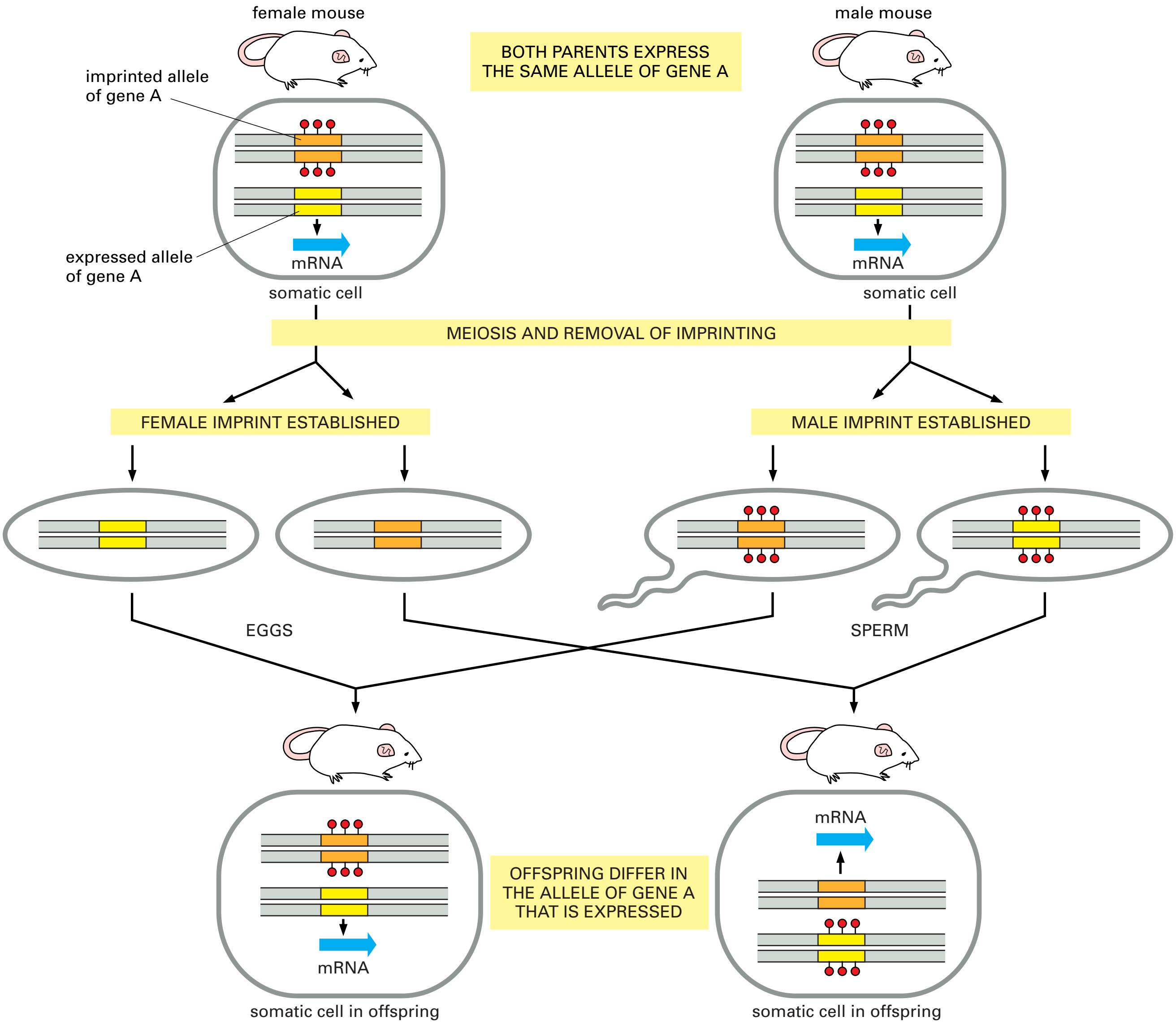

female mouse
male mouse
BOTH PARENTS EXPRESS THE SAME ALLELE OF GENE A
imprinted allele of gene A
expressed allele of gene A
mRNA
mRNA
somatic cell
somatic cell
MEIOSIS AND REMOVAL OF IMPRINTING
FEMALE IMPRINT ESTABLISHED
MALE IMPRINT ESTABLISHED
EGGS
SPERM
mRNA
OFFSPRING DIFFER IN THE ALLELE OF GENE A THAT IS EXPRESSED
mRNA
somatic cell in offspring
somatic cell in offspring

**Figure 7–83 Imprinting in the mouse.** The *top* portion of the figure shows a pair of homologous chromosomes in the somatic cells of two adult mice, one male and one female. In this example, both mice have inherited the top homolog from their father and the bottom homolog from their mother, and the paternal copy of a gene subject to imprinting (indicated in *orange)* is methylated, which prevents its expression. The maternally-derived copy of the same gene *(yellow)* is expressed. The remainder of the figure shows the outcome of a cross between these two mice. During meiosis and germ cell formation, the imprints are first erased and then reimposed *(middle* portion of figure). In eggs produced from the female, neither allele of the A gene is methylated. In sperm from the male, both alleles of gene A are methylated. Shown at the *bottom* of the figure are two of the possible imprinting patterns inherited by the progeny mice; the mouse on the *left* has the same imprinting pattern as each of the parents, whereas the mouse on the *right* has the opposite pattern. If the two alleles of A gene are distinct, these different imprinting patterns can cause phenotypic differences in the progeny mice, even though they carry exactly the same DNA sequences of the two A gene alleles. Imprinting provides an important exception to classical genetic behavior, and more than 100 mouse genes are thought to be affected in this way. However, the great majority of mouse genes are not imprinted, and therefore the rules of Mendelian inheritance apply to most of the mouse genome.

chromosome blocks its function and allows a distant enhancer to activate transcription of the *Igf2* gene. On the maternally derived chromosome, the insulator is not methylated and the *Igf2* gene is therefore not transcribed (Figure 7–84).

Imprinting is an example of an *epigenetic change*, that is, a heritable change in phenotype that does not result from a change in DNA nucleotide sequence. Why imprinting should exist at all is a mystery. In vertebrates, it is restricted to placental mammals, and all the imprinted genes are involved in fetal development. One idea is that imprinting reflects a middle ground in the evolutionary struggle between males to produce larger offspring and females to limit offspring

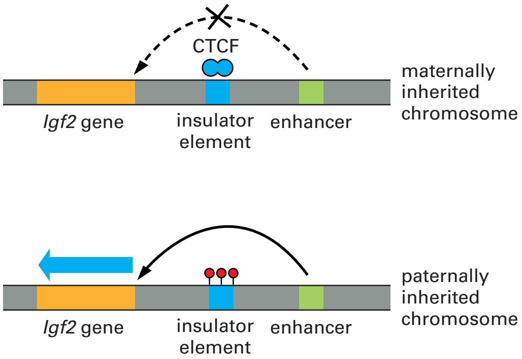


**Figure 7–84 Mechanism of imprinting of the mouse *Igf2* gene.** On chromosomes inherited from the female, a protein called CTCF binds to an insulator, (see Figure 7–61) blocking communication between the enhancer *(green)* and the *Igf2* gene *(orange)*. *Igf2* is therefore not expressed from the maternally inherited chromosome. Because of imprinting, the insulator on the male-derived chromosome is methylated; this inactivates the insulator, by blocking the binding of the CTCF protein, and allows the enhancer to activate transcription of the *Igf2* gene. In other examples of imprinting, methylation blocks gene expression by interfering with the binding of proteins required for the gene's transcription. The methylation patterns (imprints) on the chromosome, inherited by the zygote after fertilization, are maintained in subsequent generations by maintenance methyl transferases (see Figure 7–81).

size. Whatever its purpose might be, imprinting provides startling evidence that features of DNA other than its sequence of nucleotides can be inherited.

## CG-rich Islands Are Associated with About 20,000 Genes in Mammals

Because of the way DNA repair enzymes work, methylated C nucleotides in the genome tend to be eliminated in the course of evolution. Accidental deamination of an unmethylated C gives rise to U, which is not normally present in DNA and thus is recognized easily by the DNA repair enzyme uracil DNA glycosylase, excised, and then replaced with a C (as discussed in Chapter 5). But accidental deamination of a 5-methyl C cannot be repaired in this way, for the deamination product is a T and so indistinguishable from the other, nonmutant T nucleotides in the DNA. Although a special repair system exists to remove these mutant T nucleotides, many of the deaminations escape detection, so that those C nucleotides in the genome that are methylated tend to mutate to T over evolutionary time.

During the course of evolution, more than three out of every four CGs have been lost in this way, leaving vertebrates with a remarkable deficiency of this dinucleotide. The CG sequences that remain are very unevenly distributed in the genome; they are present at 10 to 20 times their average density in selected regions, called **CG islands**, that are 1000 to 2000 nucleotide pairs long. These islands, with some important exceptions, seem to remain unmethylated in all cell types. They often surround the promoters of the so-called *housekeeping genes*—those genes that code for the many proteins that are essential for cell viability and are therefore expressed in most cells (Figure 7–85). In addition, some *tissue–specific genes*, which code for proteins needed only in selected types of cells, are also associated with CG islands.

The distribution of CG islands (also called CpG islands to distinguish the CG dinucleotides from CG nucleotide pairs) can be explained if we assume that CG methylation was adopted in vertebrates primarily as a way of maintaining DNA in a transcriptionally inactive state (Figures 7–82 and 7–86). In vertebrates, new methyl-C to T mutations can be transmitted to the next generation only if they

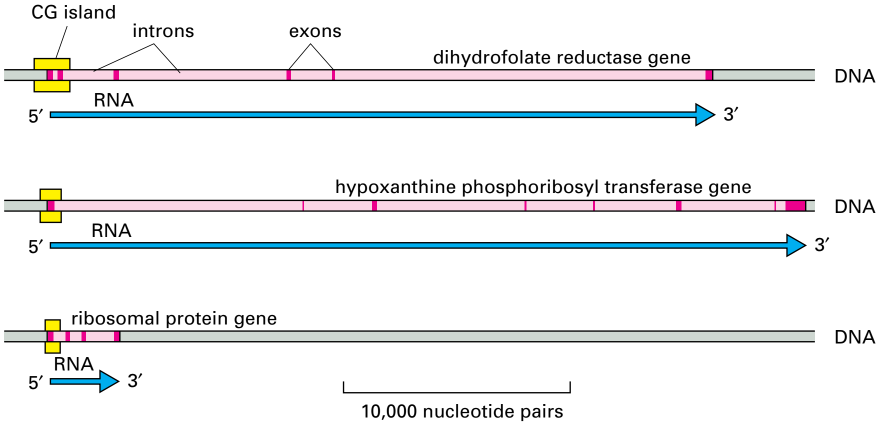


**Figure 7–85 The CG islands surrounding the promoter in three mammalian housekeeping genes.** The *yellow* boxes show the extent of each island. As for most genes in mammals (see Figure 6–25), the exons *(dark red)* are very short relative to the introns *(light red)*. (Adapted from A.P. Bird, *Trends Genet.* 3:342–347, 1987.)

occur in the germ line, the cell lineage that gives rise to sperm or eggs. Most of the DNA in vertebrate germ cells is inactive and highly methylated. Over long periods of evolutionary time, the methylated CG sequences in these inactive regions have presumably been lost through spontaneous deamination events that were not properly repaired. However promoters of genes that remain active in the germ cell lineages (including most housekeeping genes) are kept unmethylated, and therefore spontaneous deaminations of Cs that occur within them can be accurately repaired. Such regions are preserved in modern day vertebrate cells as CG islands. In addition, any mutation of a CG sequence in the genome that destroyed the function or regulation of a gene in the adult would be selected against, and some CG islands are simply the result of a higher than normal density of critical CG sequences.

The mammalian genome contains an estimated 20,000 CG islands. Most of the islands mark the 5′ ends of transcription units and thus, presumably, of genes. The presence of CG islands often provides a convenient way of identifying genes in the DNA sequences of vertebrate genomes.

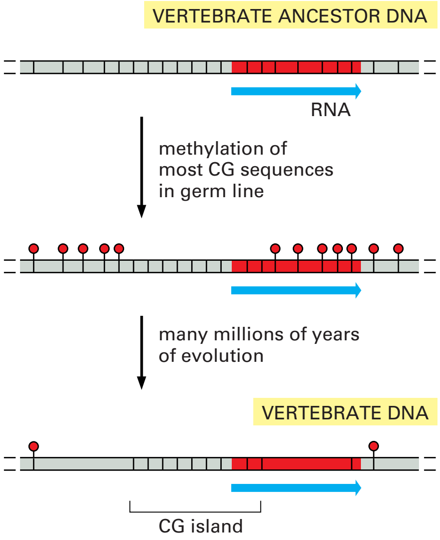

VERTEBRATE ANCESTOR DNA
RNA
methylation of most CG sequences in germ line
many millions of years of evolution
VERTEBRATE DNA
CG island

**Figure 7–86 A mechanism to explain both the marked overall deficiency of CG sequences and their clustering into CG islands in vertebrate genomes.** A *black line* marks the location of a CG dinucleotide in the DNA sequence, while a *red* "lollipop" indicates the presence of a methyl group on the CG dinucleotide. CG sequences that lie in regulatory sequences of genes that are transcribed in germ cells are unmethylated and therefore tend to be retained in evolution. Methylated CG sequences, on the other hand, tend to be lost through deamination of 5-methyl C to T, unless the CG sequence is critical for survival.

## Summary

*The many types of cells in animals and plants are created largely through mechanisms that cause different genes to be transcribed in different cells. Since many specialized animal cells can maintain their unique character through many cell division cycles and even when grown in culture, the gene regulatory mechanisms involved in creating them must be stable once established and heritable when the cell divides. These features endow the cell with a memory of its developmental history. Bacteria and yeasts provide unusually accessible model systems in which to study gene regulatory mechanisms. One such mechanism involves a competitive interaction between two gene regulatory proteins, each of which inhibits the synthesis of the other; this can create a flip-flop switch that switches a cell between two alternative patterns of gene expression. Direct or indirect positive feedback loops, which enable gene regulatory proteins to perpetuate their own synthesis, provide a general mechanism for cell memory. Negative feedback loops with programmed delays form the basis for cellular clocks.*

*In eucaryotes the transcription of a gene is generally controlled by combinations of gene regulatory proteins. It is thought that each type of cell in a higher eucaryotic organism contains a specific combination of gene regulatory proteins that ensures the expression of only those genes appropriate to that type of cell. A given gene regulatory protein may be active in a variety of circumstances and typically is involved in the regulation of many genes.*

*In addition to diffusible gene regulatory proteins, inherited states of chromatin condensation are also used by eucaryotic cells to regulate gene expression. An especially dramatic case is the inactivation of an entire X chromosome in female mammals. In vertebrates DNA methylation also functions in gene regulation, being used mainly as a device to reinforce decisions about gene expression that are made initially by other mechanisms. DNA methylation also underlies the phenomenon of genomic imprinting in mammals, in which the expression of a gene depends on whether it was inherited from the mother or the father.*

# POSTTRANSCRIPTIONAL CONTROLS

In principle, every step required for the process of gene expression could be controlled. Indeed, one can find examples of each type of regulation, although any one gene is likely to use only a few of them. Controls on the initiation of gene transcription are the predominant form of regulation for most genes. But other controls can act later in the pathway from DNA to protein to modulate the amount of gene product that is made. Although these **posttranscriptional controls**, which operate after RNA polymerase has bound to the gene's promoter and begun RNA synthesis, are less common than *transcriptional control*, for many genes they are crucial.

In the following sections, we consider the varieties of posttranscriptional regulation in temporal order, according to the sequence of events that might be experienced by an RNA molecule after its transcription has begun (Figure 7–87).

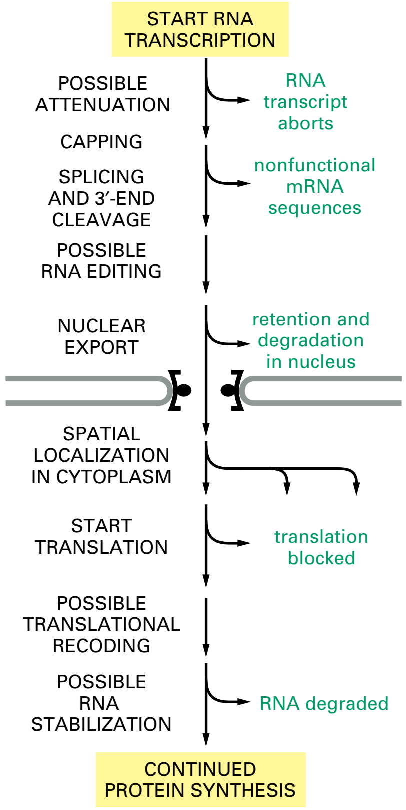

START RNA TRANSCRIPTION
POSSIBLE ATTENUATION
RNA transcript aborts
CAPPING
SPLICING AND 3′-END CLEAVAGE
nonfunctional mRNA sequences
POSSIBLE RNA EDITING
NUCLEAR EXPORT
retention and degradation in nucleus
SPATIAL LOCALIZATION IN CYTOPLASM
START TRANSLATION
translation blocked
POSSIBLE TRANSLATIONAL RECODING
POSSIBLE RNA STABILIZATION
RNA degraded
CONTINUED PROTEIN SYNTHESIS

**Figure 7–87 Possible post-transcriptional controls on gene expression.** Only a few of these controls are likely to be important for any one gene.

## Transcription Attenuation Causes the Premature Termination of Some RNA Molecules

In bacteria the expression of certain genes is inhibited by premature termination of transcription, a phenomenon called **transcription attenuation**. In some of these cases the nascent RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription. When the gene product is required, regulatory proteins bind to the nascent RNA chain and interfere with attenuation, allowing the transcription of a complete RNA molecule.

In eucaryotes transcription attenuation can occur by a number of distinct mechanisms. A well-studied example is found in HIV (the human AIDS virus). Once it has been integrated into the host genome, the viral DNA is transcribed by the cellular RNA polymerase II (see Figure 5–73). However, the host polymerase usually terminates transcription (for reasons that are not well-understood) after synthesizing transcripts of several hundred nucleotides and therefore does not efficiently transcribe the entire viral genome. When conditions for viral growth are optimal, this premature termination is prevented by a virus-encoded protein called Tat, which binds to a specific stem-loop structure in the nascent RNA that contains a “bulged base.” Once bound to this specific RNA structure (called Tar), Tat assembles several cellular proteins which allow the RNA polymerase to continue transcribing. The normal role of at least some of these cellular proteins is to prevent pausing and premature termination by RNA polymerase when it transcribes normal cellular genes. Eucaryotic genes often contain long introns; to transcribe a gene efficiently, RNA polymerase II cannot afford to linger at nucleotide sequences that happen to promote pausing. Thus a normal cellular mechanism has apparently been adapted by HIV to permit transcription of its genome to be controlled by a single viral protein.

## Alternative RNA Splicing Can Produce Different Forms of a Protein from the Same Gene

As discussed in Chapter 6, the transcripts of many eucaryotic genes are shortened by RNA splicing, in which the intron sequences are removed from the mRNA precursor. We saw that a cell can splice the “primary transcript” in different ways and thereby make different polypeptide chains from the same gene—a process called **alternative RNA splicing** (see Figures 6–27 and 7–88). A substantial proportion of higher eucaryotic genes (at least a third of human genes, it is estimated) produce multiple proteins in this way.

When different splicing possibilities exist at several positions in the transcript, a single gene can produce dozens of different proteins. In one extreme case, a *Drosophila* gene may produce as many as 38,000 different proteins from a single gene through alternative splicing (Figure 7–89), although only a small fraction of these forms have thus far been experimentally observed. Considering that the *Drosophila* genome has approximately 14,000 identified genes, it is clear that the protein complexity of an organism can greatly exceed the number of its genes. This example also illustrates the perils in equating gene number with organism complexity. For example, alternative splicing is relatively rare in single-celled budding yeasts but very common in flies. Budding yeast has ~6200

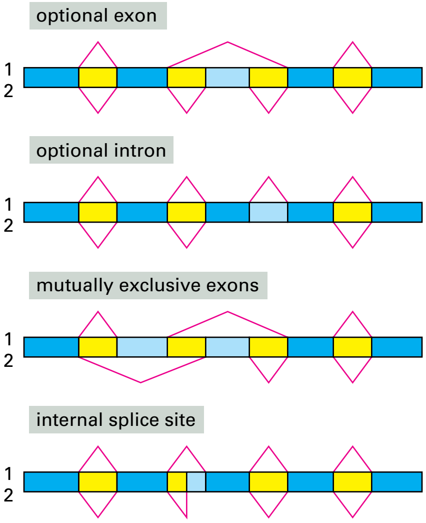


**Figure 7–88 Four patterns of alternative RNA splicing.** In each case a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark exon sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences that are included in only one of the mRNAs. The boxes are joined by *red lines* to indicate where intron sequences *(yellow)* are removed. (Adapted with permission from A. Andreadis, M.E. Gallego, and B. Nadal-Ginard, *Annu. Rev. Cell Biol.* 3:207–242, 1987.)

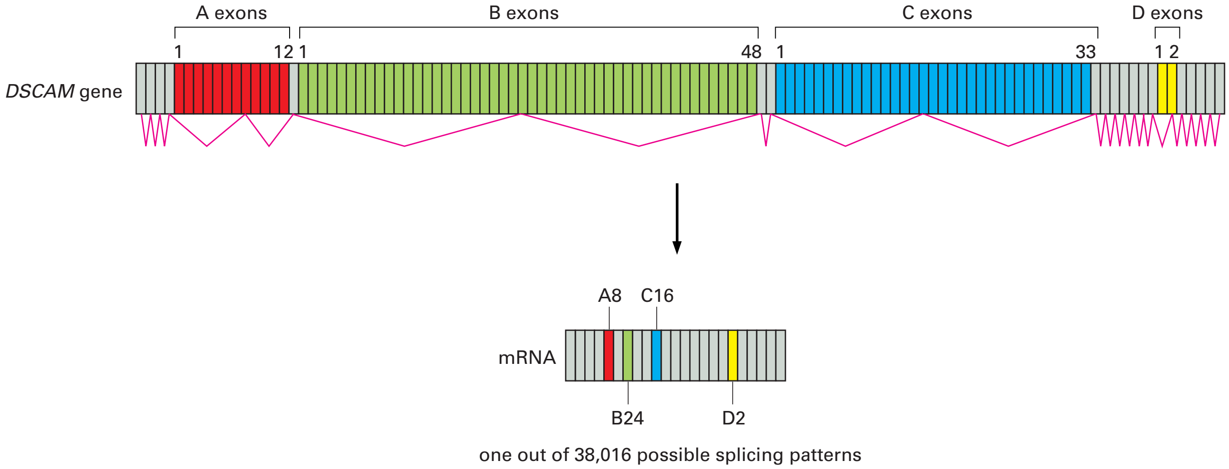


**Figure 7–89 Alternative splicing of RNA transcripts of the *Drosophila DSCAM* gene.** DSCAM proteins are axon guidance receptors that help to direct growth cones to their appropriate targets in the developing nervous system. The final mRNA contains 24 exons, four of which (denoted A, B, C, and D) are present in the *DSCAM* gene as arrays of alternative exons. Each RNA contains 1 of 12 alternatives for exon A *(red)*, 1 of 48 alternatives for exon B *(green)*, 1 of 33 alternatives for exon C *(blue)*, and 1 of 2 alternatives for exon D *(yellow)*. If all possible splicing combinations are used, 38,016 different proteins could in principle be produced from the *DSCAM* gene. Only one of the many possible splicing patterns (indicated by the *red line* and by the mature mRNA below it) is shown. Each variant DSCAM protein would fold into roughly the same structure [predominantly a series of extracellular immunoglobulin-like domains linked to a membrane-spanning region (see Figure 24–71)], but the amino acid sequence of the domains would vary according to the splicing pattern. It is possible that this receptor diversity contributes to the formation of complex neural circuits, but the precise properties and functions of the many DSCAM variants are not yet understood. (Adapted from D.L. Black, *Cell* 103:367–370, 2000.)

genes, only 327 of which are subject to splicing, and nearly all of these have only a single intron. To say that flies have only 2–3 times as many genes as yeasts is to greatly underestimate the difference in complexity of these two genomes.

In some cases alternative RNA splicing occurs because there is an *intron sequence ambiguity:* the standard spliceosome mechanism for removing intron sequences (discussed in Chapter 6) is unable to distinguish cleanly between two or more alternative pairings of 5′ and 3′ splice sites, so that different choices are made by chance on different transcripts. Where such constitutive alternative splicing occurs, several versions of the protein encoded by the gene are made in all cells in which the gene is expressed.

In many cases, however, alternative RNA splicing is regulated rather than constitutive. In the simplest examples, regulated splicing is used to switch from the production of a nonfunctional protein to the production of a functional one. The transposase that catalyzes the transposition of the *Drosophila* P element, for example, is produced in a functional form in germ cells and a nonfunctional form in somatic cells of the fly, allowing the P element to spread throughout the genome of the fly without causing damage in somatic cells (see Figure 5–70). The difference in transposon activity has been traced to the presence of an intron sequence in the transposase RNA that is removed only in germ cells.

In addition to switching from the production of a functional protein to the production of a nonfunctional one, the regulation of RNA splicing can generate different versions of a protein in different cell types, according to the needs of the cell. Tropomyosin, for example, is produced in specialized forms in different types of cells (see Figure 6–27). Cell-type-specific forms of many other proteins are produced in the same way.

RNA splicing can be regulated either negatively, by a regulatory molecule that prevents the splicing machinery from gaining access to a particular splice site on the RNA, or positively, by a regulatory molecule that helps direct the splicing machinery to an otherwise overlooked splice site (Figure 7–90). In the case of the *Drosophila* transposase, the key splicing event is blocked in somatic cells by negative regulation.

Because of the plasticity of RNA splicing (see pp. 324–325), the blocking of a "strong" splicing site will often expose a "weak" site and result in a different pattern of splicing. Likewise, activating a suboptimal splice site can result in alternative splicing by suppressing a competing splice site. Thus the splicing of a pre-mRNA molecule can be thought of as a delicate balance between competing splice sites—a balance that can easily be tipped by regulatory proteins.

## The Definition of a Gene Has Had to Be Modified Since the Discovery of Alternative RNA Splicing

The discovery that eucaryotic genes usually contain introns and that their coding sequences can be assembled in more than one way raised new questions about the definition of a gene. A gene was first clearly defined in molecular

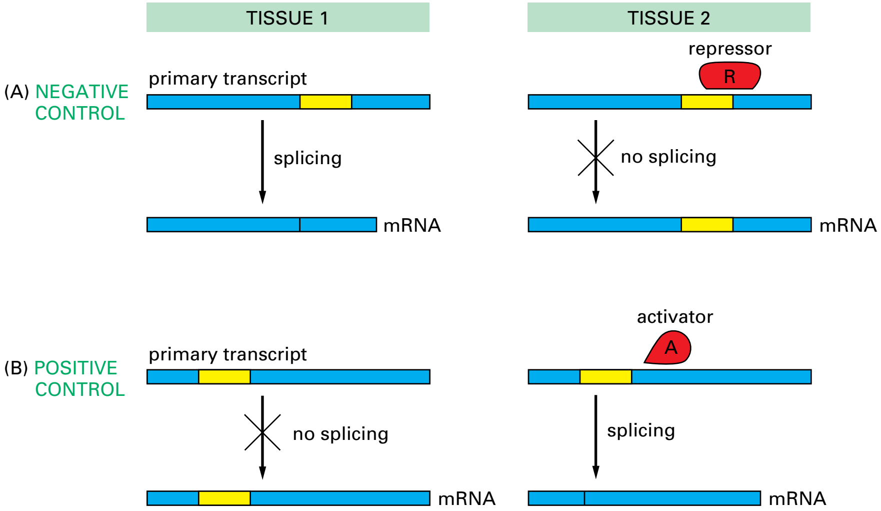


**Figure 7–90 Negative and positive control of alternative RNA splicing.** (A) Negative control, in which a repressor protein binds to the primary RNA transcript in tissue 2, thereby preventing the splicing machinery from removing an intron sequence. (B) Positive control, in which the splicing machinery is unable to efficiently remove a particular intron sequence without assistance from an activator protein.

terms in the early 1940s from work on the biochemical genetics of the fungus *Neurospora*. Until then, a gene had been defined operationally as a region of the genome that segregates as a single unit during meiosis and gives rise to a definable phenotypic trait, such as a red or a white eye in *Drosophila* or a round or wrinkled seed in peas. The work on *Neurospora* showed that most genes correspond to a region of the genome that directs the synthesis of a single enzyme. This led to the hypothesis that one gene encodes one polypeptide chain. The hypothesis proved fruitful for subsequent research; as more was learned about the mechanism of gene expression in the 1960s, a gene became identified as that stretch of DNA that was transcribed into the RNA coding for a single polypeptide chain (or a single structural RNA such as a tRNA or an rRNA molecule). The discovery of split genes and introns in the late 1970s could be readily accommodated by the original definition of a gene, provided that a single polypeptide chain was specified by the RNA transcribed from any one DNA sequence. But it is now clear that many DNA sequences in higher eucaryotic cells can produce a set of distinct (but related) proteins by means of alternative RNA splicing. How then is a gene to be defined?

In those relatively rare cases in which two very different eucaryotic proteins are produced from a single transcription unit, the two proteins are considered to be produced by distinct genes that overlap on the chromosome. It seems unnecessarily complex, however, to consider most of the protein variants produced by alternative RNA splicing as being derived from overlapping genes. A more sensible alternative is to modify the original definition to count as a gene any DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (protein isoforms). This definition of a gene also accommodates those DNA sequences that encode protein variants produced by posttranscriptional processes other than RNA splicing, such as translational frameshifting (see Figure 6–78), regulated poly-A addition, and RNA editing (to be discussed below).

## Sex Determination in *Drosophila* Depends on a Regulated Series of RNA Splicing Events

We now turn to one of the best understood examples of regulated RNA splicing. In *Drosophila* the primary signal for determining whether the fly develops as a male or female is the X chromosome/autosome ratio. Individuals with an X chromosome/autosome ratio of 1 (normally two X chromosomes and two sets of autosomes) develop as females, whereas those with a ratio of 0.5 (normally one X chromosome and two sets of autosomes) develop as males. This ratio is assessed early in development and is remembered thereafter by each cell. Three crucial gene products transmit information about this ratio to the many other genes that specify male and female characteristics (Figure 7–91). As explained in

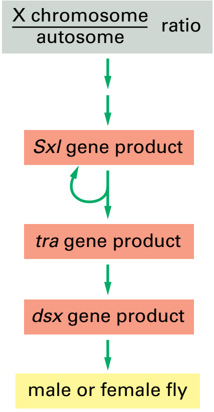


**Figure 7–91 Sex determination in *Drosophila*.** The gene products shown act in a sequential cascade to determine the sex of the fly according to the X chromosome/autosome ratio. The genes are called *sex-lethal (Sxl), transformer (tra),* and *doublesex (dsx)* because of the phenotypes that result when the gene is inactivated by mutation. The function of these gene products is to transmit the information about the X chromosome/autosome ratio to the many other genes that create the sex-related phenotypes. These other genes function as two alternative sets: those that specify female features and those that specify male features (see Figure 7–92).

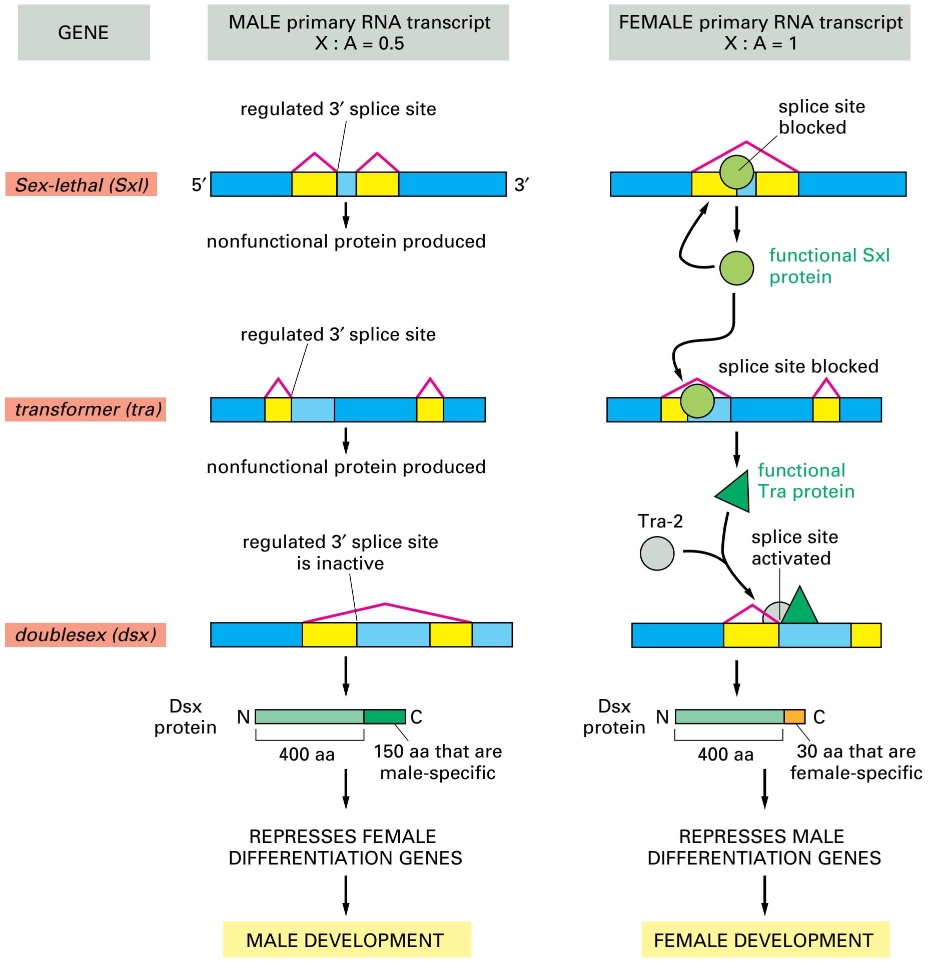


**Figure 7–92 The cascade of changes in gene expression that determines the sex of a fly through alternative RNA splicing.** An X chromosome/autosome ratio of 0.5 results in male development. Male is the "default" pathway in which the *Sxl* and *tra* genes are both transcribed, but the RNAs are spliced constitutively to produce only nonfunctional RNA molecules, and the *dsx* transcript is spliced to produce a protein that turns off the genes that specify female characteristics. An X chromosome/autosome ratio of 1 triggers the female differentiation pathway in the embryo by transiently activating a promoter within the *Sxl* gene that causes synthesis of a special class of *Sxl* transcripts that are constitutively spliced to give functional Sx1 protein. Sxl is a splicing regulatory protein with two sites of action: (1) it binds to the constitutively produced *Sxl* RNA transcript, causing a female-specific splice that continues the production of a functional Sxl protein, and (2) it binds to the constitutively produced *tra* RNA and causes an alternative splice of this transcript, which now produces an active Tra regulatory protein. The Tra protein acts with the constitutively produced Tra-2 protein to produce the female-specific spliced form of the *dsx* transcript; this encodes the female form of the Dsx protein, which turns off the genes that specify male features. The components in this pathway were all initially identified through the study of *Drosophila* mutants that are altered in their sexual development. The *dsx* gene, for example, derives its name *(doublesex)* from the observation that a fly lacking this gene product expresses both male- and female-specific features. Note that, although both the Sxl and the Tra proteins bind to specific RNA sites, Sxl is a repressor that acts negatively to block a splice site, whereas the Tra proteins are activators that act positively to induce a splice (see Figure 7–90). Sx1 binds to the pyrimidine-rich stretch of nucleotides that is part of the standard splicing consensus sequence (see Figure 6–28) and blocks access by the normal splicing factor, U2AF (see Figure 6–29). Tra binds to specific RNA sequences in an exon and activates a normally suboptimal splicing signal.

Figure 7–92, sex determination in *Drosophila* depends on a cascade of regulated RNA splicing events that involves these three gene products.

Although *Drosophila* sex determination provides one of the best-understood examples of a regulatory cascade based on RNA splicing, it is not clear why the fly should use this strategy. Other organisms (the nematode, for example) use an entirely different scheme for sex determination—one based on transcriptional and translational controls. Moreover, the *Drosophila* male-determination pathway requires that a number of nonfunctional RNA molecules be continually produced, which seems unnecessarily wasteful. One speculation is that this RNA-splicing cascade exploits an ancient control device, left over from the early stage of evolution where RNA was the predominant biological molecule, and controls of gene expression would have had to be based almost entirely on RNA–RNA interactions (discussed in Chapter 6).

## A Change in the Site of RNA Transcript Cleavage and Poly-A Addition Can Change the C-terminus of a Protein

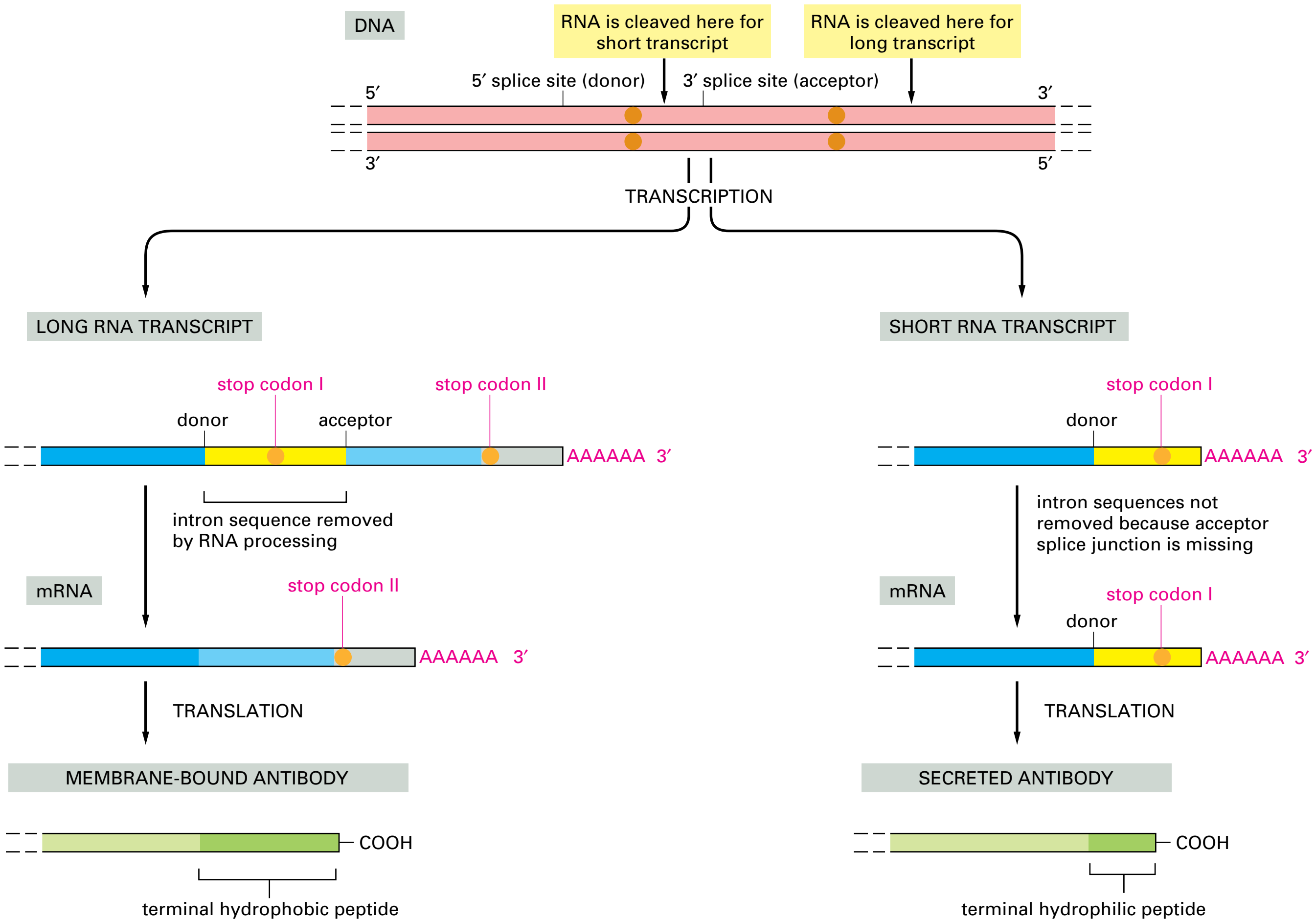
We saw in Chapter 6 that the 3′ end of a eucaryotic mRNA molecule is not formed by the termination of RNA synthesis by the RNA polymerase. Instead, it results from an RNA cleavage reaction that is catalyzed by additional factors while the transcript is elongating (see Figure 6–37). A cell can control the site of this cleavage so as to change the C-terminus of the resultant protein.

A well-studied example is the switch from the synthesis of membrane-bound to secreted antibody molecules that occurs during the development of B lymphocytes. Early in the life history of a B lymphocyte, the antibody it produces is anchored in the plasma membrane, where it serves as a receptor for antigen. Antigen stimulation causes B lymphocytes to multiply and to begin secreting their antibody. The secreted form of the antibody is identical to the membrane-bound form except at the extreme C-terminus. In this part of the protein, the membrane-bound form has a long string of hydrophobic amino acids that traverses the lipid bilayer of the membrane, whereas the secreted form has a much shorter string of hydrophilic amino acids. The switch from membrane-bound to secreted antibody therefore requires a different nucleotide sequence at the 3′ end of the mRNA; this difference is generated through a change in the length of the primary RNA transcript, caused by a change in the site of RNA cleavage, as shown in Figure 7–93. This change is caused by an increase of the concentration of a subunit of CStF, the protein that binds to the G/U rich sequences of RNA cleavage and poly-A addition sites and promotes RNA cleavage (see Figures 6–37 and 6–38). The first cleavage-poly-A addition site encountered by an RNA polymerase transcribing the antibody gene is suboptimal and is usually skipped in unstimulated B lymphocytes, leading to production of the longer RNA transcript. When antibody stimulation causes an increase in CSTF concentration, cleavage now occurs at the suboptimal site, and the shorter transcript is produced. In this way a change in concentration of a general RNA processing factor can produce specific effects on a relatively small number of genes.

## RNA Editing Can Change the Meaning of the RNA Message

The molecular mechanisms used by cells are a continual source of surprises. An example is the process of **RNA editing**, which alters the nucleotide sequences of mRNA transcripts once they are transcribed. In Chapter 6, we saw that rRNAs and tRNAs are modified posttranscriptionally. In this section we see that some mRNAs are modified in ways that change the coded message they carry. The most dramatic form of RNA editing was discovered in RNA transcripts that code for proteins in the mitochondria of trypanosomes. Here, one or more U nucleotides are inserted (or, less frequently, removed) from selected regions of a

**Figure 7–93 Regulation of the site of RNA cleavage and poly-A addition determines whether an antibody molecule is secreted or remains membrane-bound.** In unstimulated B lymphocytes *(left)*, a long RNA transcript is produced, and the intron sequence near its 3′ end is removed by RNA splicing to give rise to an mRNA molecule that codes for a membrane-bound antibody molecule. In contrast, after antigen stimulation *(right)* the primary RNA transcript is cleaved upstream from the splice site in front of the last exon sequence. As a result, some of the intron sequence that is removed from the long transcript remains as coding sequence in the short transcript. These are the nucleotide sequences that encode the hydrophilic C-terminal portion of the secreted antibody molecule.



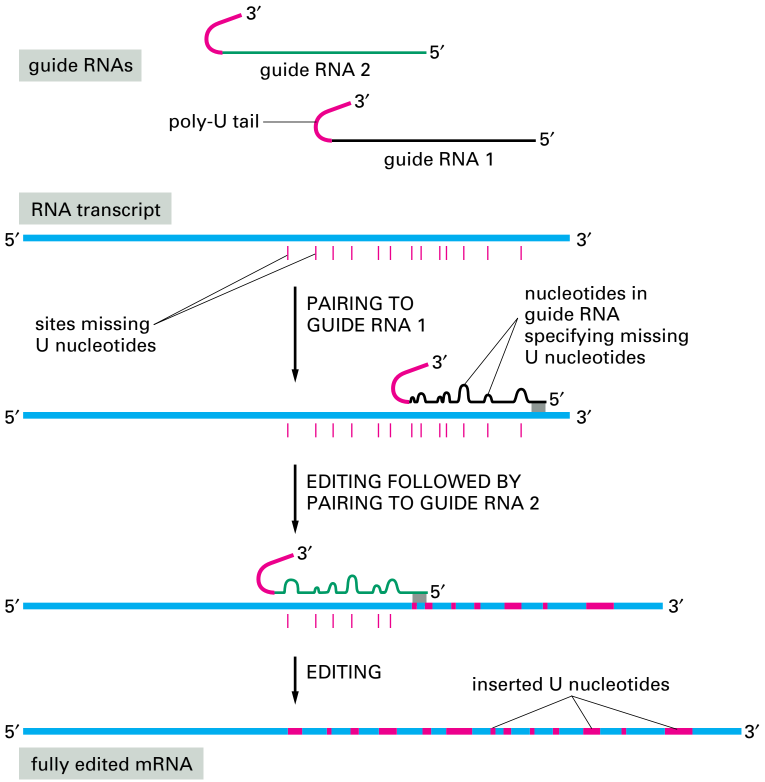


**Figure 7–94 RNA editing in the mitochondria of trypanosomes.** Guide RNAs contain at their 3′ end a stretch of poly U, which donates U nucleotides to sites on the RNA transcript that mispair with the guide RNA; thus the poly-U tail gets shorter as editing proceeds (not shown). Editing generally starts near the 3′ end and progresses toward the 5′ end of the RNA transcript, as shown, because the "anchor sequence" at the 5′ end of most guide RNAs can pair only with edited sequences.

transcript, causing major modifications in both the original reading frame and the sequence, thereby changing the meaning of the message. For some genes the editing is so extensive that over half of the nucleotides in the mature mRNA are U nucleotides that were inserted during the editing process. The information that specifies exactly how the initial RNA transcript is to be altered is contained in a set of 40– to 80–nucleotide-long RNA molecules that are transcribed separately. These so-called *guide RNAs* have a 5′ end that is complementary in sequence to one end of the region of the transcript to be edited; this is followed by a sequence that specifies the set of nucleotides to be inserted into the transcript, which is followed in turn by a continuous run of U nucleotides. The editing mechanism is remarkably complex, with the U nucleotides at the 3′ end of the guide RNA being transferred directly into the transcript, as illustrated in Figure 7–94.

Extensive editing of mRNA sequences has also been found in the mitochondria of many plants, with nearly every mRNA being edited to some extent. In this case, however, RNA bases are changed from C to U, without nucleotide insertions or deletions. Often many of the Cs in an mRNA are affected by editing, changing 10% or more of the amino acids that the mRNA encodes.

We can only speculate as to why the mitochondria of trypanosomes and plants make use of such extensive RNA editing. The suggestions that seem most reasonable are based on the premise that mitochondria contain a primitive genetic system that offers scanty opportunities for other forms of control. There is evidence that editing is regulated to produce different mRNAs under different conditions, so that RNA editing can be viewed as a primitive way to change the expression of genes, a relic, perhaps, of mechanisms that operated in very ancient cells, where most catalyses were probably carried out by RNA molecules rather than by proteins.

RNA editing of a much more limited kind occurs in mammals. One of the most important types is the enzymatic deamination of adenine to produce inosine (see Figure 6–55), which occurs at selected positions in some pre-mRNAs. In some cases, this modification changes the splicing pattern of the RNA; in others, it changes the meanings of codons. Because inosine base pairs with cytosine, A-to-I editing can result in a protein with an altered amino acid sequence.

This editing is carried out by protein enzymes called *ADARs* (adenosine deaminases acting on RNA); these enzymes recognize a double-stranded RNA structure that is formed by base pairing between the site to be edited and a complementary sequence located elsewhere on the same RNA molecule, typically in a 3′ intron (Figure 7–95). An especially important example of A-to-I editing takes place in the pre-mRNA that codes for a transmitter-gated ion channel in the brain. A single edit changes a glutamine to an arginine; the affected amino acid lies on the inner wall of the channel, and the editing change alters the $Ca^{2+}$ permeability of the channel. The importance of this edit in mice has been demonstrated by deleting the relevant ADAR gene. The mutant mice are prone to epileptic seizures and die during or shortly after weaning. If the gene for the gated ion channel is mutated to produce the edited form of the protein directly, mice lacking the ADAR develop normally, showing that editing of the ion channel RNA is normally crucial for proper brain development. Mice and humans have several additional ADAR genes, and deletion of one of these in mice causes death of the mouse embryo before birth. Because these mice have severe defects in the production of red blood cell precursors, it is likely that RNA editing is also essential for the proper development of the hemopoietic system.

C-to-U editing has also been observed in mammals. In one example, that of the *apolipoprotein-B* mRNA, a C to U change creates a stop codon that causes a truncated version of this large protein to be made in a tissue-specific manner. Why editing in mammalian cells exists at all is a mystery. One idea is that it arose in evolution to correct “mistakes” in the genome. Another is that is provides yet another way for the cell to produce a variety of related proteins from a single gene. A third view is that editing is merely one of a large number of haphazard, makeshift devices that have originated through random mutation and have been perpetuated because they happen to contribute to a useful effect.

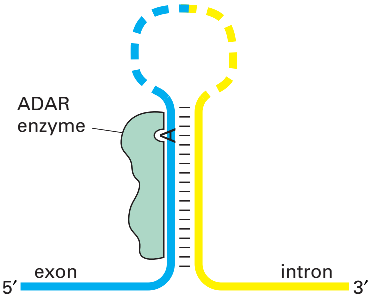


**Figure 7–95 Mechanism of A-to-I RNA editing in mammals.** The position of an edit is signaled by RNA sequences carried on the same RNA molecule. Typically, a sequence complementary to the position of the edit is present in an intron, and the resulting double-stranded RNA attracts the A-to-I editing enzyme ADAR. Mice and humans have three ADAR enzymes: ADR1 is required in the liver for proper red blood cell development, ADR2 is required for proper brain development (as described in the text), and the role of ADR3 is not yet known.

## RNA Transport from the Nucleus Can Be Regulated

It has been estimated that in mammals only about one-twentieth of the total mass of RNA synthesized ever leaves the nucleus. We saw in Chapter 6 that most mammalian RNA molecules undergo extensive processing and the “left-over” RNA fragments (excised introns and RNA sequences 3′ to the cleavage/poly-A site) are degraded in the nucleus. Incompletely processed and otherwise damaged RNAs are also eventually degraded in the nucleus as part of the quality control system of RNA production. This degradation is carried out by the **exosome**, a large protein complex that contains, as subunits, several different RNA exonucleases.

As described in Chapter 6, the export of RNA molecules from the nucleus is delayed until processing has been completed. Therefore any mechanism that prevents the completion of RNA splicing on a particular RNA molecule could in principle block the exit of that RNA from the nucleus. This feature forms the basis for one of the best understood examples of **regulated nuclear transport** of mRNA, which occurs in HIV, the virus that causes AIDS.

HIV is a retrovirus—an RNA virus that, once inside a cell, directs the formation of a double-stranded DNA copy of its genome which is then inserted into the host genome (see Figure 5–73). Once inserted, the viral DNA is transcribed as one long RNA molecule by the host cell RNA polymerase II. This transcript is then spliced in many different ways to produce over 30 different species of mRNA, which in turn, are translated into a variety of different proteins (Figure 7–96). In order to make progeny virus, entire, unspliced viral transcripts must be exported from the nucleus to the cytosol where they are packaged into viral capsids (see Figure 5–73). Moreover, several of the HIV mRNAs are alternatively spliced in such a way that they still carry complete introns. The host cell’s block to the nuclear export of unspliced RNA (and its subsequent degradation) therefore presents a special problem for HIV, and it is overcome in an ingenious way.

The virus encodes a protein (called Rev) that binds to a specific RNA sequence (called the Rev responsive element, RRE) located within a viral intron. The Rev protein interacts with a nuclear export receptor (exportin 1), which

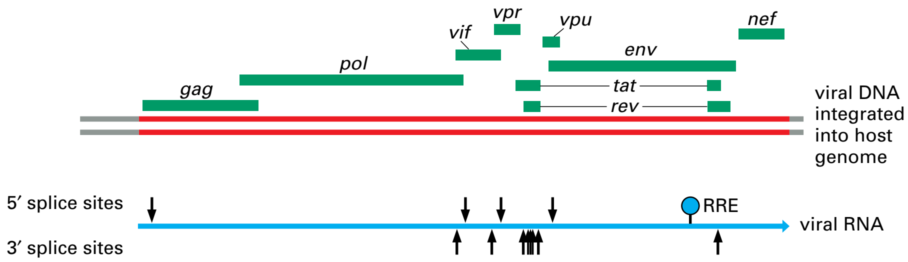


**Figure 7–96 The compact genome of HIV, the human AIDS virus.** The positions of the nine HIV genes are shown in *green*. The *red double line* indicates a DNA copy of the viral genome which has become integrated into the host DNA *(gray)*. Note that the coding regions of many genes overlap, and those of *tat* and *rev* are split by introns. The *blue line* at the *bottom* of the figure represents the pre-mRNA transcript of the viral DNA showing the locations of all the possible splice sites *(arrows)*. There are many alternative ways of splicing the viral transcript; for example the *env* mRNAs retain the intron that has been spliced out of the *tat* and *rev* mRNAs. The Rev response element (RRE) is indicated by a blue ball and stick. It is a 234-nucleotide long stretch of RNA that folds into a defined structure; Rev recognizes a particular hairpin (see Figure 6–94) within this larger structure.

The *gag* gene codes for a protein that is cleaved into several smaller proteins that form the viral capsid. The *pol* gene codes for a protein that is cleaved to produce reverse transcriptase (which transcribes RNA into DNA) as well as the integrase involved in integrating the viral genome (as double-stranded DNA) into the host genome. Pol is produced by ribosomal frameshifting of translation that begins at *gag* (see Figure 6–78). The *env* gene codes for the envelope proteins (see Figure 5–73). Tat, Rev, Vif, Vpr, Vpu, and Nef are small proteins with a variety of functions. For example, Rev regulates nuclear export (see Figure 7–97) and Tat regulates the elongation of transcription across the integrated viral genome (see p. 436).

directs the movement of viral RNAs through nuclear pores into the cytosol despite the presence of intron sequences. We discuss in detail the way that export receptors function in Chapter 12.

The regulation of nuclear export by Rev has several important consequences for HIV growth and pathogenesis. In addition to ensuring the nuclear export of specific unspliced RNAs, it divides the viral infection into an early phase (where Rev is translated from a fully spliced RNA and RNAs containing an intron are retained in the nucleus and degraded) and a late phase (where unspliced RNAs are exported due to Rev function). This timing helps the virus replicate by providing the gene products roughly in the order in which they are needed (Figure 7–97). It is also possible that regulation by Rev helps the HIV virus to achieve latency, a condition where the HIV genome has become integrated into the host cell genome but the production of viral proteins has temporarily ceased. If, after its initial entry into a host cell, conditions became unfavorable for viral transcription and replication, Rev is made at levels too low to promote export of unspliced RNA. This situation stalls the viral growth cycle. When conditions for viral replication improve, Rev levels increase, and the virus can enter the replication cycle.

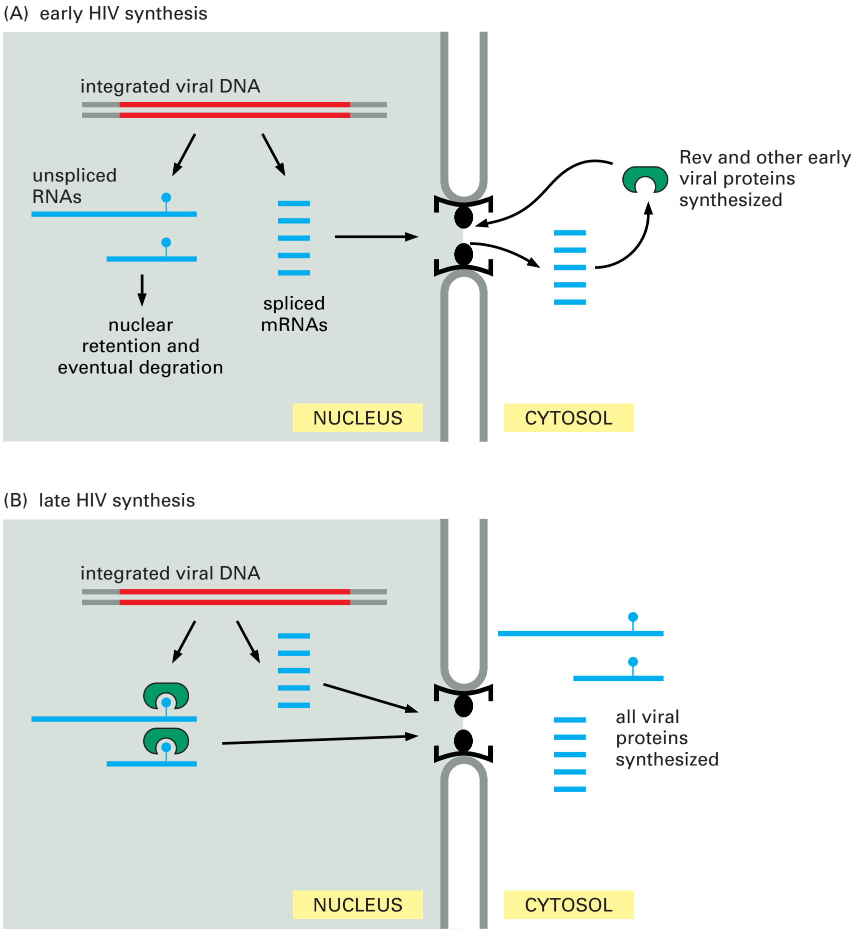


**Figure 7–97 Regulation of nuclear export by the HIV Rev protein.** Early in HIV infection (A), only the fully spliced RNAs (which contain the coding sequences for Rev, Tat, and Nef) are exported from the nucleus and translated. Once sufficient Rev protein has accumulated and been transported into the nucleus (B), unspliced viral RNAs can be exported from the nucleus. Many of these RNAs are translated into protein, and the full length transcripts are packaged into new viral particles.

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

Once a newly made eucaryotic mRNA molecule has passed through a nuclear pore and entered the cytosol, it is typically met by ribosomes, which translate it into a polypeptide chain (see Figure 6–40). If the mRNA encodes a protein that is destined to be secreted or expressed on the cell surface, it will be directed to the endoplasmic reticulum (ER) by a signal sequence at the protein's amino terminus; components of the cell's protein-sorting apparatus recognize the signal sequence as soon as it emerges from the ribosome and direct the entire complex of ribosome, mRNA, and nascent protein to the membrane of the ER, where the remainder of the polypeptide chain is synthesized, as discussed in Chapter 12. In other cases the entire protein is synthesized by free ribosomes in the cytosol, and signals in the completed polypeptide chain may then direct the protein to other sites in the cell.

Some mRNAs are themselves directed to specific intracellular locations before translation begins. Presumably it is advantageous for the cell to position its mRNAs close to the sites where the protein produced from the mRNA is required. The signals that direct mRNA localization are typically located in the 3′ *untranslated region* (*UTR*) of the mRNA molecule—a region that extends from the stop codon, which terminates protein synthesis, to the start of the poly-A tail (see Figure 6–22A). A striking example of mRNA localization is seen in the *Drosophila* egg, where the mRNA encoding the bicoid gene regulatory protein is localized by attachment to the cytoskeleton at the anterior tip of the developing egg. When the translation of this mRNA is triggered by fertilization, a gradient of the bicoid protein is generated that plays a crucial part in directing the development of the anterior part of the embryo (shown in Figure 7–52 and discussed in more detail in Chapter 21). Many mRNAs in somatic cells are localized in a similar way. The mRNA that encodes actin, for example, is localized to the actin-filament-rich cell cortex in mammalian fibroblasts by means of a 3′ UTR signal.

RNA localization has been observed in many organisms, including unicellular fungi, plants, and animals, and it is likely to be a common mechanism used by cells to concentrate high-level production of proteins at specific sites. Several distinct mechanisms for mRNA localization have been discovered (Figure 7–98), but all of them require specific signals in the mRNA itself, usually concentrated in the 3′ UTR (Figure 7–99).

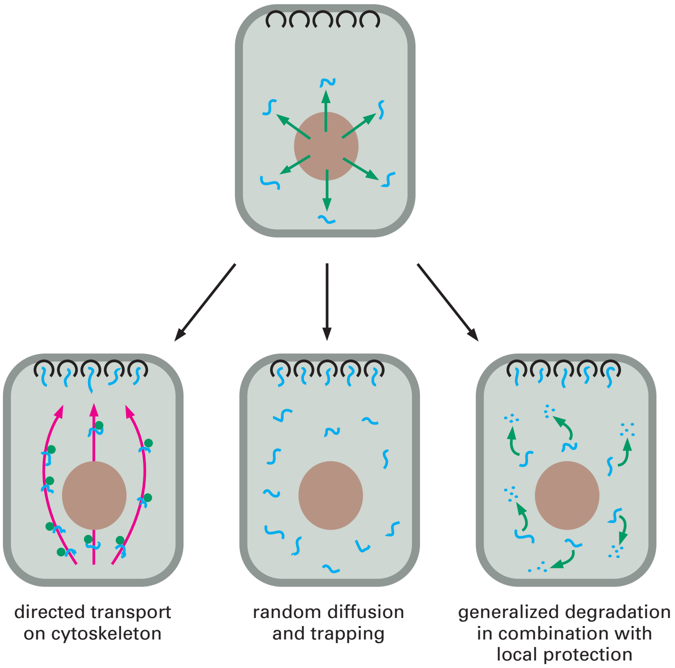

directed transport on cytoskeleton
random diffusion and trapping
generalized degradation in combination with local protection

**Figure 7–98 Three mechanisms for the localization of mRNAs.** The mRNA to be localized leaves the nucleus through nuclear pores *(top)*. Some localized mRNAs *(left diagram)* travel to their destination by associating with cytoskeleton motors *(green)*. As described in Chapter 16, these motors use the energy of ATP hydrolysis to move unidirectionally along components of the cytoskeleton. At their destination, the mRNAs are held in place by anchor proteins *(black)*. Other mRNAs randomly diffuse through the cytosol and are simply trapped and therefore concentrated at their sites of localization *(center diagram)*. Still other mRNAs *(right diagram)* are degraded in the cytosol unless they have bound, through random diffusion, a localized protein complex that anchors and protects the mRNA from degradation *(black)*. Each of these mechanisms requires specific signals on the mRNA, which are typically located in the 3′ UTR (see Figure 7–99). In many cases of mRNA localization, additional mechanisms block the translation of the mRNA until it is properly localized. (Adapted from H.D. Lipshitz and C.A. Smibert, *Curr. Opin. Gen. Dev.* 10:476–488, 2000.)

We saw in Chapter 6 that the 5′ cap and the 3′ poly-A tail are necessary for efficient translation, and their presence on the same mRNA molecule thereby signals to the translation machinery that the mRNA molecule is intact. As just described, the 3′ UTR often contains a "zip code," which directs mRNAs to different places in the cell. In this chapter, we will also see that mRNAs also carry information specifying the average length of time each mRNA persists in the cytosol and the efficiency with which each mRNA is translated into protein. In a broad sense, the untranslated regions of eucaryotic mRNAs resemble the transcriptional control regions of genes: their nucleotide sequences contain information specifying the way the RNA is to be used, and proteins that interpret this information bind specifically to these sequences. Thus, over and above the specification of the amino acid sequences of proteins, mRNA molecules are rich with many additional types of information.

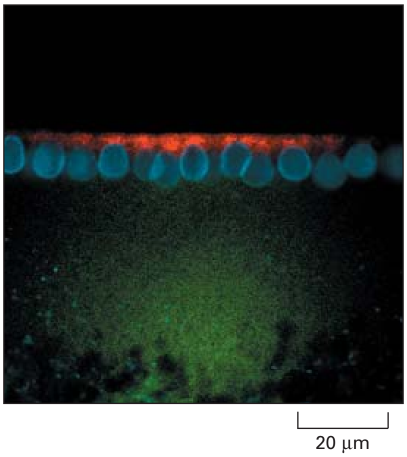


**Figure 7–99 The importance of the 3′ UTR in localizing mRNAs to specific regions of the cytoplasm.** For this experiment, two different fluorescently-labeled RNAs were prepared by transcribing DNA *in vitro* in the presence of fluorescently-labeled derivatives of UTP. One RNA (labeled with a *red* fluorochrome) contains the coding region for the *Drosophila* hairy protein and includes the adjacent 3′ UTR. The other RNA (labeled *green)* contains the hairy coding region but the 3′ UTR has been deleted. The two RNAs were mixed and injected into a *Drosophila* embryo at a stage of development when multiple nuclei reside in a common cytoplasm (see Figure 7–52). When the fluorescent RNAs were visualized 10 minutes later, the full-length hairy RNA *(red)* was localized to the apical side of nuclei *(blue)* but the transcript missing the 3′ UTR *(green)* failed to localize. Hairy is one of many gene regulatory proteins that specifies positional information in the developing *Drosophila* embryo discussed in Chapter 21). The localization of its mRNA (shown in this experiment to depend on its 3′ UTR) is thought to be critical for proper fly development. (Courtesy of Simon Bullock and David Ish-Horowicz.)

## Proteins That Bind to the 5′ and 3′ Untranslated Regions of mRNAs Mediate Negative Translational Control

Once an mRNA has been synthesized, one of the most common ways of regulating the levels of its protein product is by controlling the step in which translation is initiated. Even though the mechanistic details of translation initiation differ between eucaryotes and bacteria (as we saw in Chapter 6), some of the same basic regulatory strategies are used.

In bacterial mRNAs a conserved stretch of six nucleotides, the *Shine-Dalgarno sequence*, is always found a few nucleotides upstream of the initiating AUG codon. This sequence forms base pairs with the 16S RNA in the small ribosomal subunit, correctly positioning the initiating AUG codon in the ribosome. Because this interaction makes a major contribution to the efficiency of initiation, it provides the bacterial cell with a simple way to regulate protein synthesis through **negative translational control** mechanisms. These mechanisms generally involve blocking the Shine-Dalgarno sequence, either by covering it with a bound protein or by incorporating it into a base-paired region in the mRNA molecule. Many bacterial mRNAs have specific *translational repressor proteins* that can bind in the vicinity of the Shine-Dalgarno sequence and thereby inhibit translation of only that species of mRNA. For example, some ribosomal proteins can repress translation of their own mRNAs by binding to the 5′ untranslated region. This mechanism comes into play only when the ribosomal proteins are produced in excess over ribosomal RNA and are therefore not incorporated into ribosomes. In this way, it allows the cell to maintain correctly balanced quantities of the various components needed to form ribosomes. It is not hard to guess how this mechanism might have evolved. Ribosomal proteins assemble into ribosomes by binding to specific sites in rRNA; ingeniously, some of them exploit this RNA-binding ability to regulate their own production by binding to similar sites present in their own mRNAs.

Eucaryotic mRNAs do not contain a Shine-Dalgarno sequence. Instead, as discussed in Chapter 6, the selection of an AUG codon as a translation start site is largely determined by its proximity to the cap at the 5′ end of the mRNA molecule, which is the site at which the small ribosomal subunit binds to the mRNA and begins scanning for an initiating AUG codon. Despite the differences in translation initiation, eucaryotes also utilize translational repressors. Some bind to the 5′ end of the mRNA and thereby inhibit translation initiation. Others recognize nucleotide sequences in the 3′ UTR of specific mRNAs and decrease translation initiation by interfering with the communication between the 5′ cap and 3′ poly-A tail, which is required for efficient translation (see Figure 6–71).

A well-studied form of negative translational control in eucaryotes allows the synthesis of the intracellular iron storage protein ferritin to be increased rapidly if the level of soluble iron atoms in the cytosol rises. The iron regulation depends on a sequence of about 30 nucleotides in the 5′ leader of the ferritin mRNA molecule. This iron-response element folds into a stem-loop structure that binds a translation repressor protein called aconitase, which blocks the translation of any RNA sequence downstream (Figure 7–100). Aconitase is an iron-binding protein, and exposure of the cell to iron causes it to dissociate from the ferritin mRNA, releasing the block to translation and increasing the production of ferritin by as much as hundredfold.

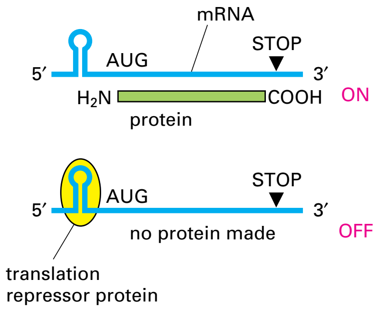


**Figure 7–100 Negative translational control.** This form of control is mediated by a sequence-specific RNA-binding protein that acts as a translation repressor. Binding of the protein to an mRNA molecule decreases the translation of the mRNA. Several cases of this type of translational control are known. The illustration is modeled on the mechanism that causes more ferritin (an iron storage protein) to be synthesized when the free iron concentration in the cytosol rises; the iron-sensitive translation repressor protein is called aconitase (see also Figure 7–105). In other examples, a complementary RNA molecule, rather than a protein, regulates translation initiation by blocking a critical region of the mRNA through the formation of a short region of double-helical RNA.

## The Phosphorylation of an Initiation Factor Globally Regulates Protein Synthesis

Eucaryotic cells decrease their overall rate of protein synthesis in response to a variety of situations, including deprivation of growth factors or nutrients, infection by viruses, and sudden increases in temperature. Much of this decrease is caused by the phosphorylation of the translation initiation factor eIF-2 by specific protein kinases that respond to the changes in conditions.

The normal function of eIF-2 was outlined in Chapter 6. It forms a complex with GTP and mediates the binding of the methionyl initiator tRNA to the small ribosomal subunit, which then binds to the 5′ end of the mRNA and begins scanning along the mRNA. When an AUG codon is recognized, the bound GTP is hydrolyzed to GDP by the eIF-2 protein, causing a conformational change in the protein and releasing it from the small ribosomal subunit. The large ribosomal subunit then joins the small one to form a complete ribosome that begins protein synthesis (see Figure 6–71).

Because eIF-2 binds very tightly to GDP, a guanine nucleotide exchange factor (see Figure 15–54), designated eIF-2B, is required to cause GDP release so that a new GTP molecule can bind and eIF-2 can be reused (Figure 7–101A). The reuse of eIF-2 is inhibited when it is phosphorylated—the phosphorylated eIF-2 binds to eIF-2B unusually tightly, inactivating eIF-2B. There is more eIF-2 than eIF-2B in cells, and even a fraction of phosphorylated eIF-2 can trap nearly all of the eIF-2B. This prevents the reuse of the nonphosphorylated eIF-2 and greatly slows protein synthesis (Figure 7–101B).

Regulation of the level of active eIF-2 is especially important in mammalian cells, being part of the mechanism that allows them to enter a nonproliferating, resting state (called $G_0$)—in which the rate of total protein synthesis is reduced to about one-fifth the rate in proliferating cells (discussed in Chapter 17).

## Initiation at AUG Codons Upstream of the Translation Start Can Regulate Eucaryotic Translation Initiation

We saw in Chapter 6 that eucaryotic translation typically begins at the first AUG downstream of the 5′ end of the mRNA, as it is the first AUG encountered by a

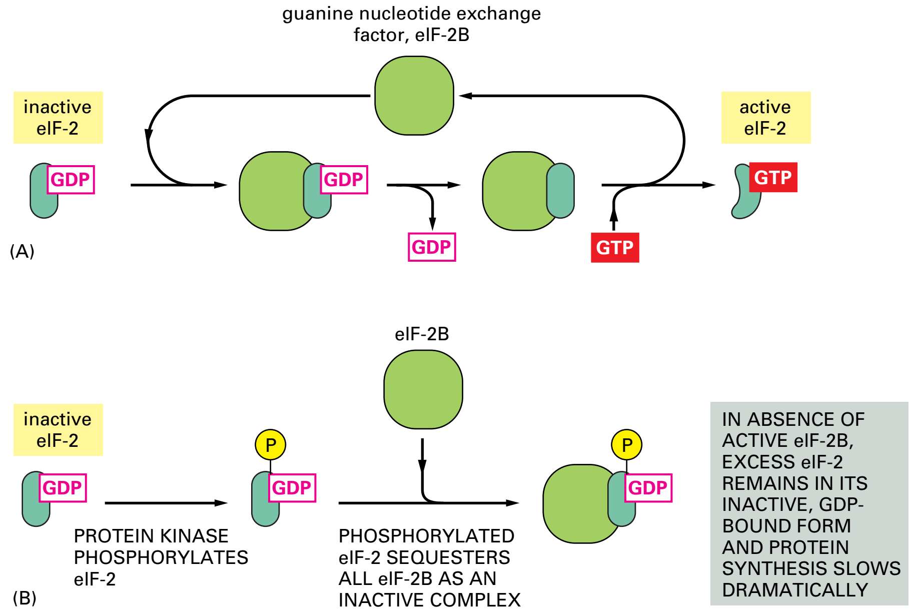


**Figure 7–101 The eIF-2 cycle.** (A) The recycling of used eIF-2 by a guanine nucleotide exchange factor (eIF-2B). (B) eIF-2 phosphorylation controls protein synthesis rates by tying up eIF-2B.

scanning small ribosomal subunit. But the nucleotides immediately surrounding the AUG also influence the efficiency of translation initiation. If the recognition site is poor enough, scanning ribosomal subunits will ignore the first AUG codon in the mRNA and skip to the second or third AUG codon instead. This phenomenon, known as "leaky scanning," is a strategy frequently used to produce two or more closely related proteins, differing only in their amino termini, from the same mRNA. For example, it allows some genes to produce the same protein with and without a signal sequence attached at its amino terminus so that the protein is directed to two different locations in the cell. In some cases, the cell can regulate the relative abundance of the protein isoforms produced by leaky scanning; for example, a cell-type specific increase in the abundance of the initiation factor eIF-4F favors usage of the AUG closest to the 5′ end of the mRNA.

Another type of control found in eucaryotes uses one or more short open reading frames that lie between the 5′ end of the mRNA and the beginning of the gene. Often, the amino acid sequences coded by these upstream open reading frames (uORFs) are not critical; rather the uORFs serve a purely regulatory function. An uORF present on an mRNA molecule will generally decrease translation of the downstream gene by trapping a scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the protein coding sequences.

When the activity of a general translation factor, such as eIF-2 (discussed above), is reduced, one might expect that the translation of all mRNAs would be reduced equally. Contrary to this expectation, however, the phosphorylation of eIF-2 can have selective effects, even enhancing the translation of specific mRNAs that contain uORFs. This can enable yeast cells, for example, to adapt to starvation for specific nutrients by shutting down the synthesis of all proteins except those that are required for synthesis of the nutrients that are missing. The details of this mechanism have been worked out for a specific yeast mRNA that encodes a protein called Gcn4, a gene regulatory protein that is required for the activation of many genes encoding proteins that are important for amino acid synthesis.

The *GCN4* mRNA contains four short uORFs, and these are responsible for selectively increasing the translation of *GCN4* in response to eIF-2 phosphorylation provoked by amino acid starvation. The mechanism by which *GCN4* translation is increased is complex. In outline, ribosomal subunits move along the mRNA, encountering each of the uORFs but translating only a subset of them; if the fourth uORF is translated, as is the case in unstarved cells, the ribosomes dissociate at the end of the uORF, and translation of *GCN4* is inefficient. The decrease in eIF-2 activity makes it more likely that a scanning ribosome will move through the fourth uORF before it acquires the ability to initiate translation. Such a ribosome can then efficiently initiate translation on the *GCN4* sequences, leading to the production of proteins that promote amino acid synthesis inside the cell.

## Internal Ribosome Entry Sites Provide Opportunities for Translation Control

Although approximately 90% of eucaryotic mRNAs are translated beginning with the first AUG downstream from the 5′ cap, certain AUGs, as we saw in the last section, can be skipped over during the scanning process. In this section, we discuss yet another way that cells can initiate translation at positions distant from the 5′ end of the mRNA. In these cases, translation is initiated directly at specialized RNA sequences, each of which is called an **internal ribosome entry site** (**IRES**). An IRES can occur in many different places in an mRNA. In some unusual cases, two distinct protein coding sequences are carried in tandem on the same eucaryotic mRNA; translation of the first occurs by the usual scanning mechanism and translation of the second through an IRES. IRESs are typically several hundred nucleotides in length and fold into specific structures that bind many, but not all, of the same proteins that are used to initiate normal cap-dependent translation (Figure 7–102). In fact, different IRESs require different subsets of initiation factors. However, all of them bypass the need for a 5′ cap structure and the translation initiation factor that recognizes it, eIF-4E.

IRESs were first discovered in certain mammalian viruses where they provide a clever way for the virus to take over its host cell's translation machinery. On infection, these viruses produce a protease (encoded in the viral genome) that cleaves the cellular translation factor eIF-4G and thereby renders it unable to bind to eIF-4E, the cap-binding complex. This shuts down the great majority of host cell translation and effectively diverts the translation machinery to the IRES sequences, which are present on many viral mRNAs. The truncated eIF-4G remains competent to initiate translation at these internal sites and may even stimulate the translation of certain IRES-containing viral mRNAs.

The selective activation of IRES-mediated translation also occurs on cellular mRNAs. For example, when eucaryotic cells enter M phase of the cell cycle, the overall rate of translation drops to approximately 25% that in interphase cells. This drop is largely caused by a cell-cycle dependent dephosphorylation of the cap binding complex , eIF-4E, which lowers its affinity for the 5′ cap. IRES-containing mRNAs, however, are immune to this effect, and their relative translation rates therefore increase as the cell enters M phase.

Finally, when mammalian cells enter the programmed cell death pathway (discussed in Chapter 17), eIF-4G is cleaved, and a general decrease in translation ensues. Some proteins critical for the control of cell death are translated from IRES-containing mRNAs, and they continue to be synthesized. It seems that one of the main advantages of the IRES mechanism for the cell is that it allows selected mRNAs to be translated at a high rate despite a general decrease in the cell's capacity to initiate protein synthesis.

## Gene Expression Can Be Controlled By a Change In mRNA Stability

The vast majority of mRNAs in a bacterial cell are very unstable, having a half-life of about 3 minutes. Exonucleases, which degrade in the 3′ to 5′ direction, are usually responsible for the rapid destruction of these mRNAs. Because its mRNAs are both rapidly synthesized and rapidly degraded, a bacterium can adapt quickly to environmental changes.

The mRNAs in eucaryotic cells are more stable. Some, such as that encoding β-globin, have half-lives of more than 10 hours. Many, however, have half-lives of 30 minutes or less. These unstable mRNAs often code for regulatory proteins, such as growth factors and gene regulatory proteins, whose production rates need to change rapidly in cells.

Two major degradation pathways exist for eucaryotic mRNAs, and sequences in each mRNA molecule determine the pathway and kinetics of degradation. The most common pathway involves the gradual shortening of the poly-A tail. We saw in Chapter 6 that capping and polyadenylation of mRNA

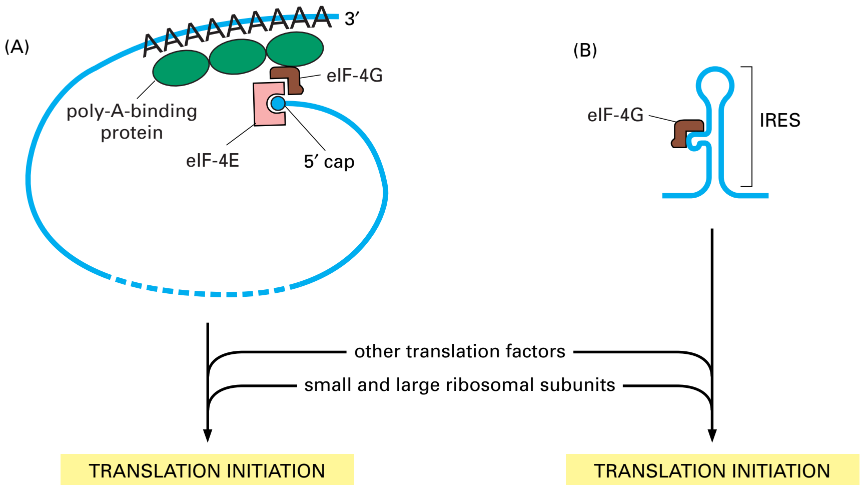


**Figure 7–102 Two mechanisms of translation initiation.** (A) The cap-dependent mechanism requires a set of initiation factors whose assembly on the mRNA is stimulated by the presence of a 5′ cap and a poly-A tail (see also Figure 6–71). (B) The IRES-dependent mechanism requires only a subset of the normal translation initiating factors, and these assemble directly on the folded IRES. (Adapted from A. Sachs, *Cell* 101:243–245, 2000.)

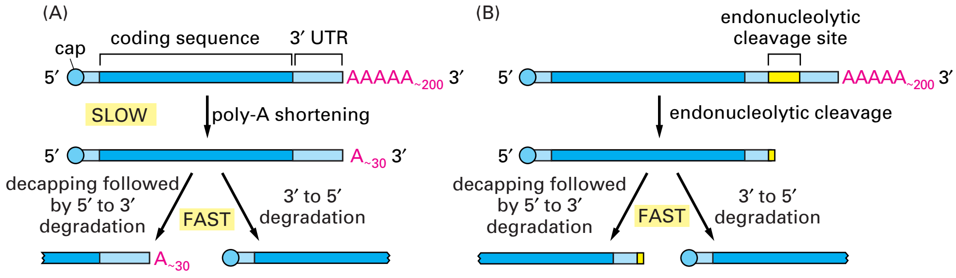


**Figure 7–103 Two mechanisms of eucaryotic mRNA decay.** (A) Deadenylation-dependent decay. Most eucaryotic mRNAs are degraded by this pathway. The critical threshold of poly-A tail length that induces decay may correspond to the loss of the poly-A binding proteins (see Figure 6–40). As shown in Figure 7–104, the deadenylation enzyme associates with both the 3′ poly-A tail and the 5′ cap, and this arrangement may coordinate decapping with poly-A shortening. Although 5′ to 3′ and 3′ to 5′ degradation are shown on separate RNA molecules, these two processes can occur together on the same molecule.
(B) Deadenylation-independent decay. It is not yet known with certainty whether decapping follows endonucleolytic cleavage of the mRNA. (Adapted from C.A. Beelman and R. Parker, *Cell* 81:179–183, 1995.)

molecules occurs in the nucleus. Once in the cytosol, the poly-A tails (which average about 200 As in length) are gradually shortened by an exonuclease that chews away the tail in the 3′ to 5′ direction. Once a critical threshold of tail shortening has been reached (approximately 30 A's remaining), the 5′ cap is removed (a process called "decapping"), and the RNA is rapidly degraded (Figure 7–103A).

Nearly all mRNAs are subject to poly-A tail shortening, decapping, and eventual degradation, but the rate at which this occurs differs from one species of mRNA to the next. The proteins that carry out tail-shortening compete directly with the machinery that catalyzes translation; therefore, any factors that affect the translation efficiency of an mRNA will tend to have the opposite effect on its degradation (Figure 7–104). In addition, many mRNAs carry in their 3′ UTR sequences binding sites for specific proteins that increase or decrease the rate of poly-A tail shortening. For example, many unstable mRNAs contain stretches of AU sequences, which greatly enhance the shortening rate.

A second pathway by which mRNA is degraded begins with the action of specific endonucleases, which simply cleave the poly-A tail from the rest of the mRNA in one step (see Figure 7–103). The mRNAs that are degraded in this way carry specific nucleotide sequences, typically in their 3′ UTR, that serve as recognition sequences for the endonucleases.

The stability of an mRNA can be changed in response to extracellular signals. For example, the addition of iron to cells decreases the stability of the mRNA that encodes the receptor protein that binds the iron-transporting protein transferrin, causing less of this receptor to be made. Interestingly, this effect is mediated by the iron-sensitive RNA-binding protein aconitase, which, as we discussed above, also controls ferritin mRNA translation. Aconitase can bind to the 3′ UTR of the transferrin receptor mRNA and cause an increase in receptor production by blocking endonucleolytic cleavage of the mRNA. On the addition of iron, aconitase is released from the mRNA, decreasing mRNA stability (Figure 7–105).

## Cytoplasmic Poly-A Addition Can Regulate Translation

The initial polyadenylation of an RNA molecule (discussed in Chapter 6) occurs in the nucleus, apparently automatically for nearly all eucaryotic mRNA precursors. As we have just seen, the poly-A tails on most mRNAs gradually shorten in

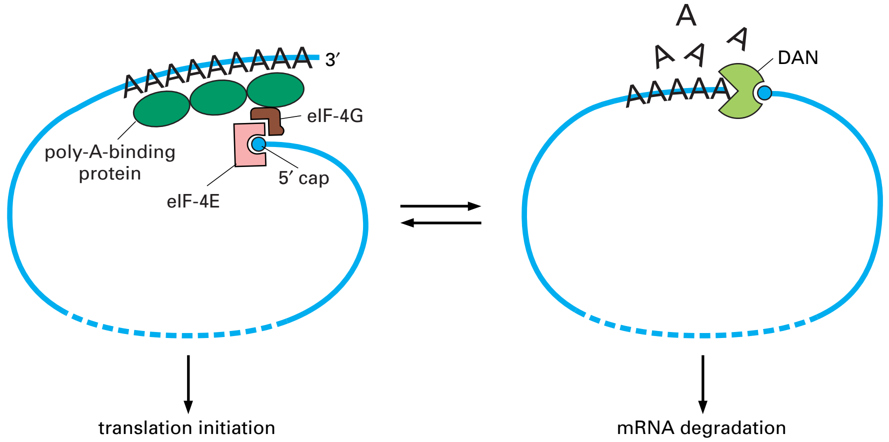


**Figure 7–104 The competition between mRNA translation and mRNA decay.** The same two features of mRNA—the 5′ cap and the 3′ poly-A site—are used in both translation initiation and deadenylation-dependent mRNA decay (see Figure 7–103). The enzyme (called DAN) that shortens the poly-A tail in the 3′ to 5′ direction associates with the 5′ cap. As described in Chapter 6 (see Figure 6–71), the translation initiation machinery also associates with both the 5′ cap and the poly-A tail. (Adapted from M. Gao et al., *Mol. Cell* 5:479–488, 2000.)

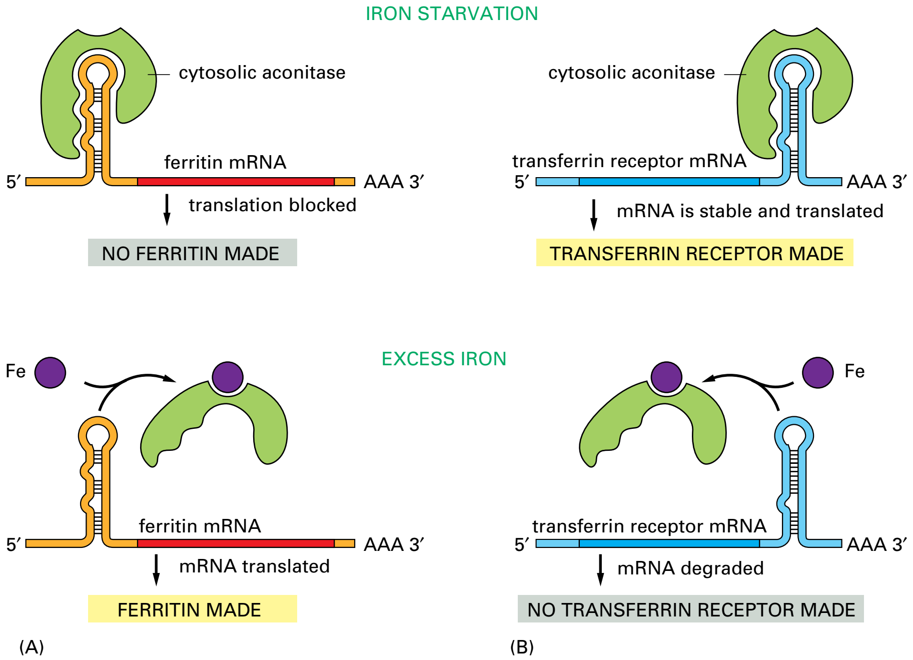


**Figure 7–105 Two posttranslational controls mediated by iron.** In response to an increase in iron concentration in the cytosol, a cell increases its synthesis of ferritin in order to bind the extra iron (A) and decreases its synthesis of transferrin receptors in order to import less iron across the plasma membrane (B). Both responses are mediated by the same iron-responsive regulatory protein, aconitase, which recognizes common features in a stem-and-loop structure in the mRNAs encoding ferritin and transferrin receptor. Aconitase dissociates from the mRNA when it binds iron. But because the transferrin receptor and ferritin are regulated by different types of mechanisms, their levels respond oppositely to iron concentrations even though they are regulated by the same iron-responsive regulatory protein. The binding of aconitase to the 5′ UTR of the ferritin receptor mRNA blocks translation initiation; its binding to the 3′ UTR of the ferritin receptor mRNA blocks an endonuclease cleavage site and thereby stabilizes the mRNA (see Figure 7–103). (Adapted from M.W. Hentze et al., *Science* 238:1570–1573, 1987 and J.L. Casey et al., *Science* 240:924–928, 1988.)

the cytosol, and the RNAs are eventually degraded. In some cases, however, the poly-A tails of specific mRNAs are lengthened in the cytosol, and this mechanism provides an additional form of translational regulation.

Maturing oocytes and eggs provide the most striking example. Many of the normal mRNA degradation pathways seem to be disabled in these giant cells, so that the cells can build up large stores of mRNAs in preparation for fertilization. Many mRNAs are stored in the cytoplasm with only 10 to 30 As at their 3′ end, and in this form they are not translated. At specific times during oocyte maturation and postfertilization, when the proteins encoded by these mRNAs are required, poly A is added to selected mRNAs, greatly stimulating the initiation of their translation.

## Nonsense-mediated mRNA Decay Is Used as an mRNA Surveillance System in Eucaryotes

We saw in Chapter 6 that mRNA production in eucaryotes occurs by an elaborately choreographed series of synthesis and processing steps. Only when all of the steps of mRNA production have been completed are the mRNAs exported from the nucleus to the cytosol for translation into protein. If any of those steps go awry, the RNA is eventually degraded in the nucleus (along with excised introns) by the *exosome,* a large protein complex that contains at least ten 3′-to-5′ RNA exonucleases. The eucaryotic cell has an additional mechanism, called **nonsense-mediated mRNA decay,** that eliminates certain types of aberrant mRNAs before they can be efficiently translated into protein. This mechanism was discovered when mRNAs that contain misplaced in-frame translation stop codons (UAA, UAG, or UGA) were found to be rapidly degraded. These stop codons can arise either from mutation or from incomplete splicing: in both cases, the phenomenon was observed. This mRNA surveillance system therefore prevents the synthesis of abnormally truncated proteins which, as we have seen, can be especially dangerous to the cell. But how are these potentially harmful mRNAs recognized by the cell?

In vertebrates, the critical feature of mRNA that is sensed by the nonsense-mediated decay system is the spatial relationship between the first in-frame termination codon and the exon–exon boundaries formed by RNA splicing. If the stop codon lies downstream (3′) of all the exon–exon boundaries, the mRNA is spared from nonsense-mediated decay; if, on the other hand, a stop codon is located upstream (5′) to an exon–exon boundary, the mRNA is degraded. Translating ribosomes, in conjunction with other surveillance proteins, assess this relationship for each individual mRNA. Exactly how this is accomplished is not understood in detail, but it is easy to understand why ribosomes must play a part: only in-frame termination codons trigger nonsense-mediated decay, and it is the relationship between the ribosome and the mRNA that defines the reading frame. According to one model (Figure 7–106), proteins in the nucleus bind to and thereby mark the exon–exon junctions following RNA splicing. As the mRNA leaves the nucleus, it remains at the nuclear periphery and is joined by a set of additional surveillance proteins as translation begins. The first round of translation of an individual mRNA molecule would, in this view, be used simply to assess the fitness of the mRNA for further rounds of translation. If the mRNA passes this test, translation begins in earnest as the mRNA is released to diffuse through the cytosol.

Nonsense-mediated decay may have been especially important in evolution, allowing eucaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing—by selecting only those mRNAs for translation that produce a full-length protein. Nonsense-mediated decay is also important in cells of the developing immune system, where the extensive DNA rearrangements that occur (see Figure 24–37) often generate premature termination codons. The mRNAs produced from such rearranged genes are degraded by this surveillance system, thereby avoiding the toxic effects of truncated proteins.

## RNA Interference Is Used by Cells to Silence Gene Expression

Eucaryotic cells use a specialized type of RNA degradation as a defense mechanism to destroy foreign RNA molecules, specifically those that can be identified by virtue of their occurrence within the cell in double-stranded form. Termed **RNA interference** (**RNAi**), this mechanism is found in a wide variety of organisms, including single-celled fungi, plants, worms, mice, and probably humans—suggesting that it is an evolutionarily ancient defense mechanism. In plants, RNA interference protects cells against RNA viruses. In other types of organisms, it is thought to protect against the proliferation of transposable elements that replicate via RNA intermediates (see Figure 5–76). Many transposable elements and plant viruses produce double-stranded RNA, at least transiently, in their life cycles. RNAi not only helps to keep such infestations in check, but also provides scientists with a powerful experimental technique to turn off the expression of individual cellular genes (see Figure 8–65).

The presence of free, double-stranded RNA triggers RNAi by attracting a protein complex containing an RNA nuclease and an RNA helicase. This protein complex cleaves the double-stranded RNA into small (approximately 23 nucleotide pair) fragments which remain associated with the enzyme. The bound RNA fragments then direct the enzyme complex to other RNA molecules

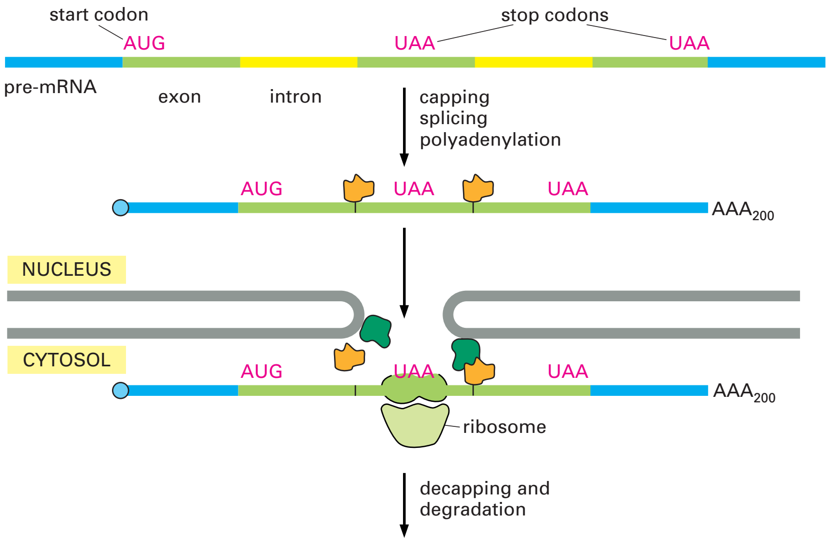
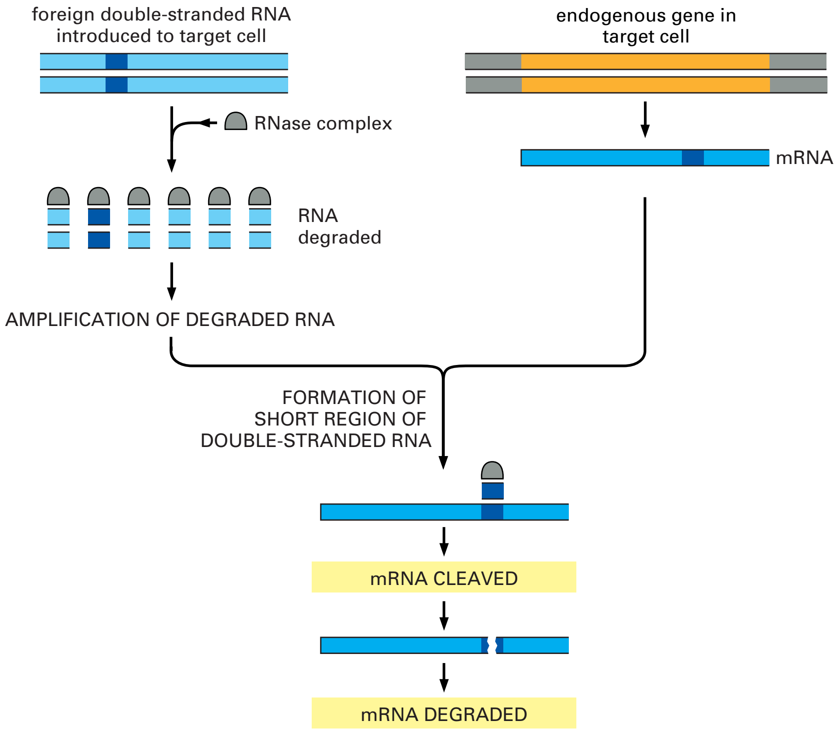


**Figure 7–106 A model for nonsense-mediated mRNA decay.** According to this model, nuclear proteins *(orange)* mark the exon–exon boundaries on a spliced mRNA molecule. These proteins are thought to assemble in concert with the splicing reaction and may also be involved in the transport of mature mRNAs from the nucleus (see Figure 6–40). A mature mRNA molecule is exported from the nucleus but remains in the vicinity of the nuclear envelope where a "test" round of translation is performed by the ribosome *(green)* aided by additional surveillance proteins *(dark green).* If an in-frame stop codon is encountered before the final exon–exon boundary is reached, the mRNA is subject to nonsense-mediated decay. If not, the mRNA is released from the nuclear envelope (perhaps because of a displacement of the exon–exon marking proteins by the ribosome) and is free to undergo multiple rounds of translation in the cytosol. According to the model shown, the test round of translation occurs just outside the nucleus; however, it is also possible that it takes place within the nucleus, just before the mRNA is exported. (Adapted from J. Lykke-Anderson et al., *Cell* 103:1121–1131, 2000.)



foreign double-stranded RNA introduced to target cell
RNase complex
RNA degraded
AMPLIFICATION OF DEGRADED RNA
endogenous gene in target cell
mRNA
FORMATION OF SHORT REGION OF DOUBLE-STRANDED RNA
mRNA CLEAVED
mRNA DEGRADED

**Figure 7–107 The mechanism of RNA interference.** On the left is shown the fate of foreign double-stranded RNA molecules. They are recognized by an RNase, present in a large protein complex, and degraded into short fragments that are approximately 23 nucleotide pairs in length. These fragments are sometimes amplified by an RNA-dependent RNA polymerase and, in this case, can be efficiently transmitted to progeny cells. If the foreign RNA has a nucleotide sequence similar to that of a cellular gene (*right* side of figure), mRNA produced by this gene will also be degraded, by the pathway shown. In this way, the expression of a cellular gene can be experimentally shut off by introducing double-stranded RNA into the cell that matches the nucleotide sequence of the gene. RNA interference also requires ATP hydrolysis and RNA helicases, probably to produce single-stranded RNA molecules that can form base pairs with additional RNA molecules.

that have complementary nucleotide sequences, and the enzyme degrades these as well. These other molecules can be either single- or double-stranded (as long as they have a complementary strand). In this way, the experimental introduction of a double-stranded RNA molecule can be used by scientists to inactivate specific cellular mRNAs (Figure 7–107).

Each time it cleaves a new RNA, the enzyme complex is regenerated with a short RNA molecule, so that an original double-stranded RNA molecule can act catalytically to destroy many complementary RNAs. In addition, the short double-stranded RNA cleavage products themselves can be replicated by additional cellular enzymes, providing an even greater amplification of RNA interference activity (see Figure 7–107). This amplification ensures that once initiated, RNA interference can continue even after all the initiating double-stranded RNA has been degraded or diluted out. For example, it permits progeny cells to continue carrying out RNA interference that was provoked in the parent cells. In addition, the RNA interference activity can be spread by the transfer of RNA fragments from cell to cell. This is particularly important in plants (whose cells are linked by fine connecting channels, as discussed in Chapter 19), because it allows an entire plant to become resistant to an RNA virus after only a few of its cells have been infected.

## Summary

*Many steps in the pathway from RNA to protein are regulated by cells to control gene expression. Most genes are thought to be regulated at multiple levels, although control of the initiation of transcription (transcriptional control) usually predominates. Some genes, however, are transcribed at a constant level and turned on and off solely by posttranscriptional regulatory processes. These processes include (1) attenuation of the RNA transcript by its premature termination, (2) alternative RNA splice-site selection, (3) control of 3′-end formation by cleavage and poly-A addition, (4) RNA editing, (5) control of transport from the nucleus to the cytosol, (6) localization of mRNAs to particular parts of the cell, (7) control of translation initiation, and (8) regulated mRNA degradation. Most of these control processes require the recognition of specific sequences or structures in the RNA molecule being regulated. This recognition is accomplished by either a regulatory protein or a regulatory RNA molecule.*

# HOW GENOMES EVOLVE

In this and the preceding three chapters, we discussed the structure of genes, the way they are arranged in chromosomes, the intricate cellular machinery that converts genetic information into functional protein and RNA molecules, and the many ways in which gene expression is regulated by the cell. In this section, we discuss some of the ways that genes and genomes have evolved over time to produce the vast diversity of modern-day life forms on our planet. Genome sequencing has revolutionized our view of this process of *molecular evolution*, uncovering an astonishing wealth of information about the family relationships among organisms and evolutionary mechanisms.

It is perhaps not surprising that genes with similar functions can be found in a diverse range of living things. But the great revelation of the past 20 years has been the discovery that the actual nucleotide sequences of many genes are sufficiently well conserved that **homologous** genes—that is, genes that are similar in their nucleotide sequence because of a common ancestry—can often be recognized across vast phylogenetic distances. For example, unmistakable homologs of many human genes are easy to detect in such organisms as nematode worms, fruit flies, yeasts, and even bacteria.

As discussed in Chapter 3 and again in Chapter 8, the recognition of sequence homology has become a major tool for inferring gene and protein function. Although finding such a homology does not guarantee similarity in function, it has proven to be an excellent clue. Thus, it is often possible to predict the function of a gene in humans for which no biochemical or genetic information is available simply by comparing its sequence to that of an intensively studied gene in another organism.

Gene sequences are often far more tightly conserved than is overall genome structure. As discussed in Chapter 4, features of genome organization such as genome size, number of chromosomes, order of genes along chromosomes, abundance and size of introns, and amount of repetitive DNA are found to differ greatly among organisms, as does the actual number of genes.

The number of genes is only very roughly correlated with the phenotypic complexity of an organism. Thus, for example, current estimates of gene number are 6,000 for the yeast *Saccharomyces cerevisiae*, 18,000 for the nematode *Caenorhabditis elegans*, 13,000 for *Drosophila melanogaster*, and 30,000 for humans (see Table 1–1). As we shall soon see, much of the increase in gene number with increasing biological complexity involves the expansion of families of closely related genes, an observation that establishes gene duplication and divergence as major evolutionary processes. Indeed, it is likely that all present-day genes are descendants—via the processes of duplication, divergence, and reassortment of gene segments—of a few ancestral genes that existed in early life forms.

## Genome Alterations are Caused by Failures of the Normal Mechanisms for Copying and Maintaining DNA

With a few exceptions, cells do not have specialized mechanisms for creating changes in the structures of their genomes: evolution depends instead on accidents and mistakes. Most of the genetic changes that occur result simply from failures in the normal mechanisms by which genomes are copied or repaired when damaged, although the movement of transposable DNA elements also plays an important role. As we discussed in Chapter 5, the mechanisms that maintain DNA sequences are remarkably precise—but they are not perfect. For example, because of the elaborate DNA-replication and DNA-repair mechanisms that enable DNA sequences to be inherited with extraordinary fidelity, only about one nucleotide pair in a thousand is randomly changed every 200,000 years. Even so, in a population of 10,000 individuals, every possible nucleotide substitution will have been “tried out” on about 50 occasions in the course of a million years—a short span of time in relation to the evolution of species.

Errors in DNA replication, DNA recombination, or DNA repair can lead either to simple changes in DNA sequence—such as the substitution of one base pair for another—or to large-scale genome rearrangements such as deletions, duplications, inversions, and translocations of DNA from one chromosome to another. It has been argued that the rates of occurrence of these mistakes have themselves been shaped by evolutionary processes to provide an acceptable balance between genome stability and change.

In addition to failures of the replication and repair machinery, the various mobile DNA elements described in Chapter 5 are an important source of genomic change. In particular, transposable DNA elements (transposons) play a major part as parasitic DNA sequences that colonize a genome and can spread within it. In the process, they often disrupt the function or alter the regulation of existing genes; and sometimes they even create altogether novel genes through fusions between transposon sequences and segments of existing genes. Examples of the three major classes of transposons were presented in Table 5–3, p. 287. Over long periods of evolutionary time, these transposons have profoundly affected the structure of genomes.

## The Genome Sequences of Two Species Differ in Proportion to the Length of Time That They Have Separately Evolved

The differences between the genomes of species alive today have accumulated over more than 3 billion years. Lacking a direct record of changes over time, we can nevertheless reconstruct the process of genome evolution from detailed comparisons of the genomes of contemporary organisms.

The basic tool of comparative genomics is the phylogenetic tree. A simple example is the tree describing the divergence of humans from the great apes (Figure 7–108). The primary support for this tree comes from comparisons of gene and protein sequences. For example, comparisons between the sequences of human genes or proteins and those of the great apes typically reveal the fewest differences between human and chimpanzee and the most between human and orangutan.

For closely related organisms such as humans and chimpanzees, it is possible to reconstruct the gene sequences of the extinct, last common ancestor of the two species (Figure 7–109). The close similarity between human and chimpanzee genes is mainly due to the short time that has been available for the accumulation of mutations in the two diverging lineages, rather than to functional constraints that have kept the sequences the same. Evidence for this view comes from the observation that even DNA sequences whose nucleotide order is functionally unconstrained—such as the sequences that code for the fibrinopeptides (see p. 236) or the third position of "synonymous" codons (codons specifying the same amino acid—see Figure 7–109)—are nearly identical.

For less closely related organisms such as humans and mice, the sequence conservation found in genes is largely due to **purifying selection** (that is, selection that eliminates individuals carrying mutations that interfere with important genetic functions), rather than to an inadequate time for mutations to occur. As

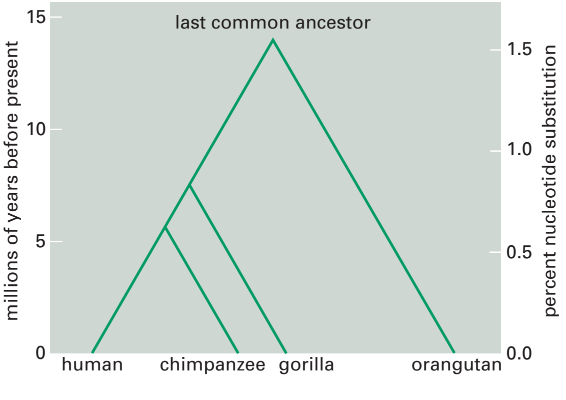
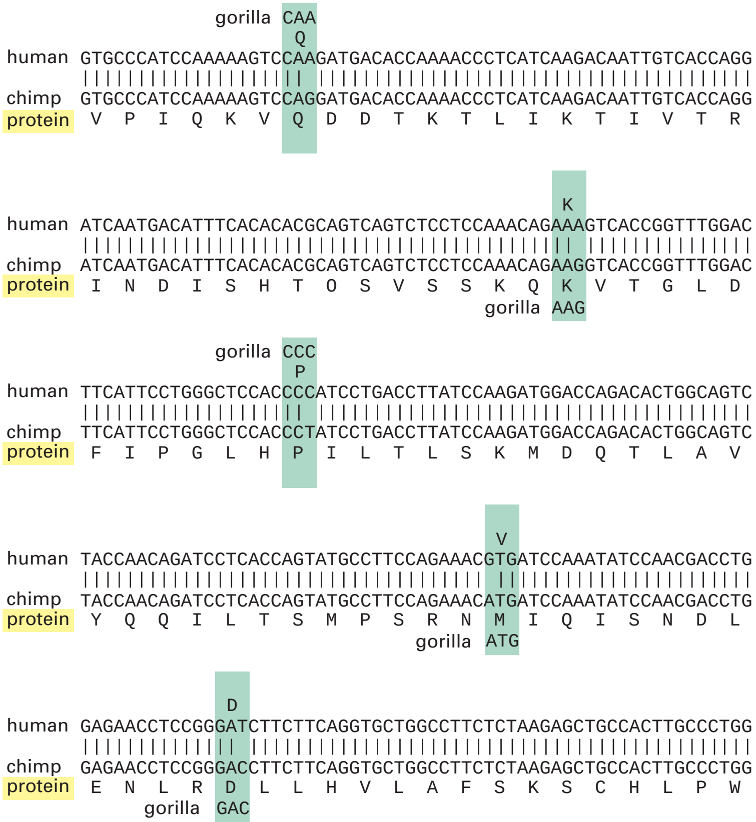


**Figure 7–108 A phylogenetic tree showing the relationship between the human and the great apes based on nucleotide sequence data.** As indicated, the sequences of the genomes of all four species are estimated to differ from the sequence of the genome of their last common ancestor by a little over 1.5%. Because changes occur independently on both diverging lineages, pairwise comparisons reveal twice the sequence divergence from the last common ancestor. For example, human–orangutan comparisons typically show sequence divergences of a little over 3%, while human–chimpanzee comparisons show divergences of approximately 1.2%. (Modified from F.-C. Chen and W.-H. Li, *Am. J. Hum. Genet.* 68:444–456, 2001.)



```
                  gorilla CAA
                           Q
human   GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
        |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||
chimp   GTGCCCATCCAAAAAGTCCAGGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
protein  V  P  I  Q  K  V  Q  D  D  T  K  T  L  I  K  T  I  V  T  R

                                                   K
human   ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAAGTCACCGGTTTGGAC
        |||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||
chimp   ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAGGTCACCGGTTTGGAC
protein  I  N  D  I  S  H  T  O  S  V  S  S  K  Q  K  V  T  G  L  D
                                          gorilla AAG

                  gorilla CCC
                           P
human   TTCATTCCTGGGCTCCACCCCATCCTGACCTTATCCAAGATGGACCAGACACTGGCAGTC
        |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||
chimp   TTCATTCCTGGGCTCCACCCTATCCTGACCTTATCCAAGATGGACCAGACACTGGCAGTC
protein  F  I  P  G  L  H  P  I  L  T  L  S  K  M  D  Q  T  L  A  V

                                             V
human   TACCAACAGATCCTCACCAGTATGCCTTCCAGAAACGTGATCCAAATATCCAACGACCTG
        |||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||
chimp   TACCAACAGATCCTCACCAGTATGCCTTCCAGAAACATGATCCAAATATCCAACGACCTG
protein  Y  Q  Q  I  L  T  S  M  P  S  R  N  M  I  Q  I  S  N  D  L
                                    gorilla ATG

                     D
human   GAGAACCTCCGGGATCTTCTTCAGGTGCTGGCCTTCTCTAAGAGCTGCCACTTGCCCTGG
        |||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||
chimp   GAGAACCTCCGGGACCTTCTTCAGGTGCTGGCCTTCTCTAAGAGCTGCCACTTGCCCTGG
protein  E  N  L  R  D  L  L  H  V  L  A  F  S  K  S  C  H  L  P  W
            gorilla GAC
```

**Figure 7–109 Tracing the ancestor sequence from a sequence comparison of the coding regions of human and chimpanzee leptin genes.** Leptin is a hormone that regulates food intake and energy utilization in response to the adequacy of fat reserves. As indicated by the codons boxed in *green*, only 5 (of 441 nucleotides total) differ between these two sequences. Moreover, when the amino acids encoded by both the human and chimpanzee sequences are examined, in only one of the 5 positions does the encoded amino acid differ. For each of the 5 variant nucleotide positions the corresponding sequence in the gorilla is also indicated. In two cases, the gorilla sequence agrees with the human sequence, while in three cases it agrees with the chimpanzee sequence. What was the sequence of the leptin gene in the last common ancestor? An evolutionary model that seeks to minimize the number of mutations postulated to have occurred during the evolution of the human and chimpanzee genes would assume that the leptin sequence of the last common ancestor was the same as the human and chimpanzee sequences when they agree; when they disagree, it would use the gorilla sequence as a tie-breaker. For convenience, only the first 300 nucleotides of the leptin coding sequences are given. The remaining 141 are identical between humans and chimpanzees.

a result, protein-coding sequences and regulatory sequences in the DNA that are constrained to engage in highly specific interactions with conserved proteins are often remarkably conserved. In contrast, most DNA sequences in the human and mouse genomes have diverged so far that it is often impossible to align them with one another.

Integration of phylogenetic trees based on molecular sequence comparisons with the fossil record has led to the best available view of the evolution of modern life forms. The fossil record remains important as a source of absolute dates based on the decay of radioisotopes in the rock formations in which fossils are found. However, precise divergence times between species are difficult to establish from the fossil record even for species that leave good fossils with distinctive morphology. Populations may be small and geographically localized for long periods before a newly arisen species expands in numbers sufficiently to leave a fossil record that is detectable. Furthermore, even when a fossil closely resembles a contemporary species, it is not certain that it is ancestral to it—the fossil may come from an extinct lineage, while the true ancestors of the contemporary species may remain unknown.

The integrated phylogenetic trees support the basic idea that changes in the sequences of particular genes or proteins occur at a constant rate, at least in the lineages of organisms whose generation times and overall biological characteristics are quite similar to one another. This apparent constancy in the rates at which sequences change is referred to as the molecular-clock hypothesis. As described in Chapter 5, the molecular clock runs most rapidly in sequences that are not subject to purifying selection—such as intergenic regions, portions of introns that lack splicing or regulatory signals, and genes that have been irreversibly inactivated by mutation (the so-called *pseudogenes)*. The clock runs most slowly for sequences that are subject to strong functional constraints—for example, the amino acid sequences of proteins such as actin that engage in specific interactions with large numbers of other proteins and whose structure, therefore, is highly constrained (see, for example, Figure 16–15).

Because molecular clocks run at rates that are determined both by mutation rates and by the amount of purifying selection on particular sequences, a different calibration is required for genes replicated and repaired by different systems within cells. Most notably, clocks based on functionally unconstrained mitochondrial DNA sequences run much faster than clocks based on functionally unconstrained nuclear sequences because of the high mutation rate in mitochondria.

Molecular clocks have a finer time resolution than the fossil record and are a more reliable guide to the detailed structure of phylogenetic trees than are classical methods of tree construction, which are based on comparisons of the morphology and development of different species. For example, the precise relationship among the great-ape and human lineages was not settled until sufficient molecular-sequence data accumulated in the 1980s to produce the tree that was shown in Figure 7–108.

## The Chromosomes of Humans and Chimpanzees Are Very Similar

We have just seen that the extent of sequence similarity between homologous genes in different species depends on the length of time that has elapsed since the two species last had a common ancestor. The same principle applies to the larger scale changes in genome structure.

The human and chimpanzee genomes—with their 5-million-year history of separate evolution—are still nearly identical in overall organization. Not only do humans and chimpanzees appear to have essentially the same set of 30,000 genes, but these genes are arranged in nearly the same way along the chromosomes of the two species (see Figure 4–57). The only substantial exception is that human chromosome 2 arose by a fusion of two chromosomes that are separate in the chimpanzee, the gorilla, and the orangutan.

Even the massive resculpting of genomes that can be produced by transposon activity has had only minor effects on the 5-million-year time scale of the human-chimpanzee divergence. For example, more than 99% of the one million copies of the Alu family of retrotransposons that are present in both genomes are in corresponding positions. This observation indicates that most of the Alu sequences in our genome underwent duplication and transposition before the divergence of the human and chimpanzee lineages. Nevertheless, the Alu family is still actively transposing. Thus, a small number of cases have been observed in which new Alu insertions have caused human genetic disease; these cases involve transposition of this DNA into sites unoccupied in the genomes of the patient’s parents. More generally, there exists a class of “human-specific” Alu sequences that occupy sites in the human genome that are unoccupied in the chimpanzee genome. Since perfect-excision mechanisms for Alu sequences appear to be lacking, these human-specific Alu sequences most likely reflect new insertions in the human lineage, rather than deletions in the chimpanzee lineage. The close sequence similarity among all of the human-specific Alu sequences suggests that they have a recent common ancestor; it may even be that only a single “master” Alu sequence remains capable of spawning new copies of itself in humans.

## A Comparison of Human and Mouse Chromosomes Shows How The Large-scale Structures of Genomes Diverge

The human and chimpanzee genomes are much more alike than are the human and mouse genomes. Although the size of the mouse genome is approximately the same and it contains nearly identical sets of genes, there has been a much longer time period over which changes have had a chance to accumulate—approximately 100 million years versus 5 million years. It may also be that rodents have significantly higher mutation rates than humans; in this case the great divergence of the human and mouse genomes would be dominated by a high rate of sequence change in the rodent lineage. Lineage-specific differences in mutation rates are, however, difficult to estimate reliably, and their contribution to the patterns of sequence divergence observed among contemporary organisms remains controversial.

As indicated by the DNA sequence comparison in Figure 7–110, mutation has led to extensive sequence divergence between humans and mice at all sites that are not under selection—such as the nucleotide sequences of introns. Indeed, human–mouse-sequence comparisons are much more informative of the functional constraints on genes than are human–chimpanzee comparisons. In the latter case, nearly all sequence positions are the same simply because not enough time has elapsed since the last common ancestor for large numbers of changes to have occurred. In contrast, because of functional constraints in human–mouse comparisons the exons in genes stand out as small islands of conservation in a sea of introns.

As the number of sequenced genomes increases, comparative genome analysis is becoming an increasingly important method for identifying their functionally important sites. For example, conservation of open-reading frames between distantly related organisms provides much stronger evidence that these sequences are actually the exons of expressed genes than does a computational analysis of any one genome. In the future, detailed biological annotation of the sequences of complex genomes—such as those of the human and the mouse—will depend heavily on the identification of sequence features that are conserved across multiple, distantly related mammalian genomes.

In contrast to the situation for humans and chimpanzees, local gene order and overall chromosome organization have diverged greatly between humans and mice. According to rough estimates, a total of about 180 break-and-rejoin events have occurred in the human and mouse lineages since these two species last shared a common ancestor. In the process, although the number of chromosomes is similar in the two species (23 per haploid genome in the human versus 20 in the mouse), their overall structures differ greatly. For example, while the centromeres occupy relatively central positions on most human chromosomes, they lie next to an end of each chromosome in the mouse. Nonetheless, even after the extensive genomic shuffling, there are many large blocks of DNA in which the gene order is the same in the human and the mouse. These regions of conserved gene order in chromosomes are referred to as **synteny** blocks (see Figure 4–18).

Analysis of the transposon families in the human and the mouse provide additional evidence of the long divergence time separating the two species. Although the major retrotransposon families in the human have counterparts in the mouse—for example, human Alu repeats are similar in sequence and transposition mechanism to the mouse B1 family—the two families have undergone separate expansions in the two lineages. Even in regions where human and mouse sequences are sufficiently conserved to allow reliable alignment, there is no correlation between the positions of Alu elements in the human genome and the B1 elements in corresponding segments of the mouse genome (Figure 7–111).

## It Is Difficult to Reconstruct the Structure of Ancient Genomes

The genomes of ancestral organisms can be inferred, but never directly observed: there are no ancient organisms alive today. Although a modern

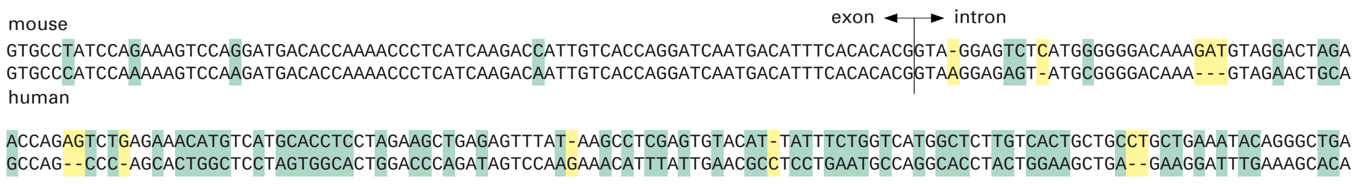

exon
intron
mouse
GTGCCTATCCAGAAAGTCCAGGATGACACCAAAACCCTCATCAAGACCATTGTCACCAGGATCAATGACATTTCACACACGGTA-GGAGTCTCATGGGGGGACAAAGATGTAGGACTAGA
GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTCACACACGGTAAGGAGAGT-ATGCGGGGACAAA---GTAGAACTGCA
human
ACCAGAGTCTGAGAAACATGTCATGCACCTCCTAGAAGCTGAGAGTTTAT-AAGCCTCGAGTGTACAT-TATTTCTGGTCATGGCTCTTGTCACTGCTGCCTGCTGAAATACAGGGCTGA
GCCAG--CCC-AGCACTGGCTCCTAGTGGCACTGGACCCAGATAGTCCAAGAAACATTTATTGAACGCCTCCTGAATGCCAGGCACCTACTGGAAGCTGA--GAAGGATTTGAAAGCACA

**Figure 7–110 Comparison of a portion of the mouse and human leptin genes.** Positions where the sequences differ by a single nucleotide substitution are boxed in *green*, and positions that differ by the addition or deletion of nucleotides are boxed in *yellow*. Note that the coding sequence of the exon is much more conserved than the adjacent intron sequence.

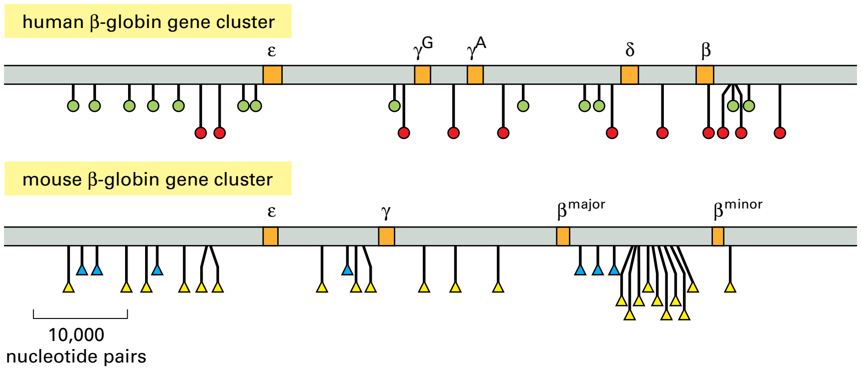


**Figure 7–111 A comparison of the β-globin gene cluster in the human and mouse genomes, showing the location of transposable elements.** This stretch of human genome contains five functional β-globin-like genes *(orange);* the comparable region from the mouse genome has only four. The positions of the human *Alu* sequence are indicated by *green circles*, and the human *L1* sequences by *red circles*. The mouse genome contains different but related transposable elements: the positions of *B1* elements (which are related to the human Alu sequences) are indicated by *blue triangles*, and the positions of the mouse *L1* elements (which are related to the human *L1* sequences) are indicated by *yellow triangles*. The absence of transposable elements from the globin structural genes can be attributed to purifying selection, which would have eliminated any insertion that compromised gene function. (Courtesy of Ross Hardison and Webb Miller.)

organism such as the horseshoe crab looks remarkably similar to fossil ancestors that lived 200 million years ago, there is every reason to believe that the horseshoe-crab genome has been changing during all that time at a rate similar to that occurring in other evolutionary lineages. Selective constraints must have maintained key functional properties of the horseshoe-crab genome to account for the morphological stability of the lineage. However, genome sequences reveal that the fraction of the genome subject to purifying selection is small; hence the genome of the modern horseshoe crab must differ greatly from that of its extinct ancestors, known to us only through the fossil record.

It is difficult to infer even gross features of the genomes of long-extinct organisms. An important example is the so-called introns-early versus introns-late controversy. Soon after the discovery in 1977 that the coding regions of most genes in metazoan organisms are interrupted by introns, a debate arose about whether introns reflect a late acquisition during the evolution of life on earth or whether they were instead present in the earliest genes. According to the introns-early model, fast-growing organisms such as bacteria lost the introns present in their ancestors because they were under selection for a compact genome adapted for rapid replication. This view is contested by an introns-late model, in which introns are viewed as having been inserted into intronless genes long after the evolution of single-cell organisms, perhaps through the agency of certain types of transposons.

There is presently no reliable way of resolving this controversy. Comparative studies of existing genomes provide estimates of rates of intron gain and loss in various evolutionary lineages. However, these estimates bear only indirectly on the question of how genomes were organized billions of years ago. Bacteria and humans are equally “modern” organisms, both of whose genomes differ so greatly from that of their last common ancestor that we can only speculate about the properties of this very ancient, ancestral genome.

When two modern organisms share nearly identical patterns of intron positions in their genes, we can be confident that the introns were present in the last common ancestor of the two species. An illuminating comparison involves humans and the puffer fish, *Fugu rubripes* (Figure 7–112). The *Fugu* genome is remarkable in having an unusually small size for a vertebrate (0.4 billion nucleotide pairs compared to 1 billion or more for many other fish and 3 billion for typical mammals). The small size of the *Fugu* genome is due almost entirely to the small size of its introns. Specifically, *Fugu* introns, as well as other noncoding segments of the *Fugu* genome, lack the repetitive DNA that makes up a large portion of the genomes of most well studied vertebrates. Nevertheless, the positions of Fugu introns are nearly perfectly conserved relative to their positions in mammalian genomes (Figure 7–113).

The question of why *Fugu* introns are so small is reminiscent of the introns-early versus introns-late debate. Obviously, either introns grew in many lineages while staying small in the *Fugu* lineage, or the *Fugu* lineage experienced massive loss of repetitive sequences from its introns. We have a clear understanding of how genomes can grow by active transposition since most transposition events are duplicative [*i.e.*, the original copy stays where it was while a copy inserts at

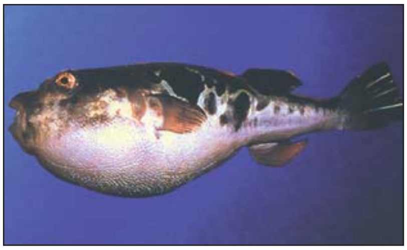


**Figure 7–112 The puffer fish, *Fugu rubripes*.** (Courtesy of Byrappa Venkatesh.)

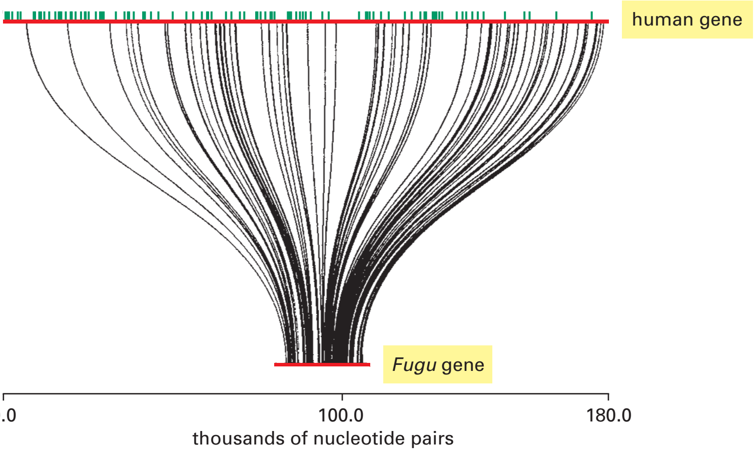


**Figure 7–113 Comparison of the genomic sequences of the human and *Fugu* genes encoding the protein huntingtin.** Both genes (indicated in *red)* contain 67 short exons that align in 1:1 correspondence to one another; these exons are connected by curved lines. The human gene is 7.5 times larger than the *Fugu* gene (180,000 versus 24,000 nucleotide pairs). The size difference is entirely due to larger introns in the human gene. The larger size of the human introns is due in part to the presence of retrotransposons, whose positons are represented by *green vertical lines;* the *Fugu* introns lack retrotransposons. In humans, mutation of the huntingtin gene causes Huntington's disease, an inherited neurodegenerative disorder (see p. 362). (Adapted from S. Baxendale et al., *Nat. Genet.* 10:67–76, 1995.)

the new site (see Figures 5–72 and 5–76)]. There is considerably less evidence in well-studied organisms for mutational processes that would efficiently delete transposons from immense numbers of sites without also deleting adjacent functionally critical sequences at rates that would threaten the survival of the lineage. Nonetheless, the origin of *Fugu*'s unusually small introns remains uncertain.

## Gene Duplication and Divergence Provide a Critical Source of Genetic Novelty During Evolution

Much of our discussion of genome evolution so far has emphasized neutral change processes or the effects of purifying selection. However, the most important feature of genome evolution is the capacity for genomic change to create biological novelty that can be positively selected for during evolution, giving rise to new types of organisms.

Comparisons between organisms that seem very different illuminate some of the sources of genetic novelty. A striking feature of these comparisons is the relative scarcity of lineage-specific genes (for example, genes found in primates but not in rodents, or those found in mammals but not in other vertebrates). Much more prominent are selective expansions of preexisting gene families. The genes encoding nuclear hormone receptors in humans, a nematode worm, and a fruit fly, all of which have fully sequenced genomes, illustrate this point (Figure 7–114). Many of the subtypes of these nuclear receptors (also called intracellular receptors) have close homologs in all three organisms that are more similar to each other than they are to other family subtypes present in the same species. Therefore, much of the functional divergence of this large gene family must have preceded the divergence of these three evolutionary lineages. Subsequently, one major branch of the gene family underwent an enormous expansion only in the worm lineage. Similar, but smaller lineage-specific expansions of particular subtypes are evident throughout the gene family tree, but they are particularly evident in the human—suggesting that such expansions offer a path toward increased biological complexity.

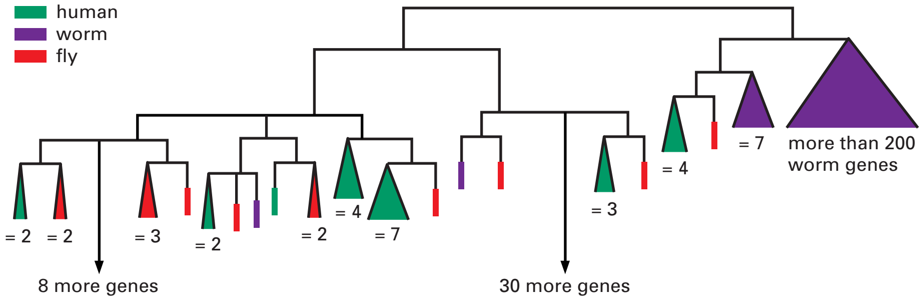


**Figure 7–114 A phylogenetic tree based on the inferred protein sequences for all nuclear hormone receptors encoded in the genomes of human *(H. sapiens)*, a nematode worm *(C. elegans)*, and a fruit fly *(D. melanogaster)*.** *Triangles* represent protein subfamilies that have expanded within individual evolutionary lineages; the width of these triangles indicates the number of genes encoding members of these subfamilies. *Colored vertical bars* represent a single gene. There is no simple pattern to the historical duplications and divergences that have created the gene families encoding nuclear receptors in the three contemporary organisms. The structure of a portion of a particular nuclear hormone receptor is shown in Figure 7–14, and a general description of their functions is discussed in Chapter 15. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001.)

Gene duplication appears to occur at high rates in all evolutionary lineages. An examination of the abundance and rate of divergence of duplicated genes in many different eucaryotic genomes suggests that the probability that any particular gene will undergo a successful duplication event (*i.e.*, one that spreads to most or all individuals in a species) is approximately 1% every million years. Little is known about the precise mechanism of gene duplication. However, because the two copies of the gene are often adjacent to one another immediately following duplication, it is thought that the duplication frequently results from inexact repair of double-strand chromosome breaks (see Figure 5–53).

## Duplicated Genes Diverge

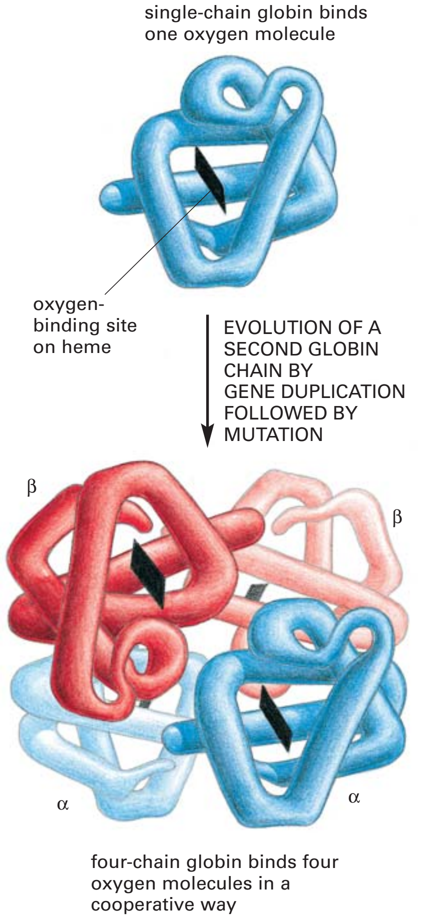
A major question in genome evolution concerns the fate of newly duplicated genes. In most cases, there is presumed to be little or no selection—at least initially—to maintain the duplicated state since either copy can provide an equivalent function. Hence, many duplication events are likely to be followed by loss-of-function mutations in one or the other gene. This cycle would functionally restore the one-gene state that preceded the duplication. Indeed, there are many examples in contemporary genomes where one copy of a duplicated gene can be seen to have become irreversibly inactivated by multiple mutations. Over time, the sequence similarity between such a **pseudogene** and the functional gene whose duplication produced it would be expected to be eroded by the accumulation of many mutational changes in the pseudogene—eventually becoming undetectable.

An alternative fate for gene duplications is for both copies to remain functional, while diverging in their sequence and pattern of expression and taking on different roles. This process of "duplication and divergence" almost certainly explains the presence of large families of genes with related functions in biologically complex organisms, and it is thought to play a critical role in the evolution of increased biological complexity.

Whole-genome duplications offer particularly dramatic examples of the duplication-divergence cycle. A whole-genome duplication can occur quite simply: all that is required is one round of genome replication in a germline cell lineage without a corresponding cell division. Initially, the chromosome number simply doubles. Such abrupt increases in the **ploidy** of an organism are common, particularly in fungi and plants. After a whole-genome duplication, all genes exist as duplicate copies. However, unless the duplication event occurred so recently that there has been little time for subsequent alterations in genome structure, the results of a series of segmental duplications—occurring at different times—are very hard to distinguish from the end product of a whole-genome duplication. In the case of mammals, for example, the role of whole genome duplications versus a series of piecemeal duplications of DNA segments is quite uncertain. Nevertheless, it is clear that a great deal of gene duplication has ocurred in the distant past.

Analysis of the genome of the zebrafish, in which either a whole-genome duplication or a series of more local duplications occurred hundreds of millions of years ago, has cast some light on the process of gene duplication and divergence. Although many duplicates of zebrafish genes appear to have been lost by mutation, a significant fraction—perhaps as many as 30–50%—have diverged functionally while both copies have remained active. In many cases, the most obvious functional difference between the duplicated genes is that they are expressed in different tissues or at different stages of development (see Figure 21–45). One attractive theory to explain such an end result imagines that different, mildly deleterious mutations quickly occur in both copies of a duplicated gene set. For example, one copy might lose expression in a particular tissue due to a regulatory mutation, while the other copy loses expression in a second tissue. Following such an occurrence, both gene copies would be required to provide the full range of functions that were once supplied by a single gene; hence, both copies would now be protected from loss through inactivating mutations. Over a longer period of time, each copy could then undergo further changes through which it could acquire new, specialized features.

**Figure 7–115 A comparison of the structure of one-chain and four-chain globins.** The four-chain globin shown is hemoglobin, which is a complex of two α- and β-globin chains. The one-chain globin in some primitive vertebrates forms a dimer that dissociates when it binds oxygen, representing an intermediate in the evolution of the four-chain globin.


single-chain globin binds one oxygen molecule
oxygen-binding site on heme
EVOLUTION OF A SECOND GLOBIN CHAIN BY GENE DUPLICATION FOLLOWED BY MUTATION
β
β
α
α
four-chain globin binds four oxygen molecules in a cooperative way

## The Evolution of the Globin Gene Family Shows How DNA Duplications Contribute to the Evolution of Organisms

The globin gene family provides a particularly good example of how DNA duplication generates new proteins, because its evolutionary history has been worked out particularly well. The unmistakable homologies in amino acid sequence and structure among the present-day globins indicate that they all must derive from a common ancestral gene, even though some are now encoded by widely separated genes in the mammalian genome.

We can reconstruct some of the past events that produced the various types of oxygen-carrying hemoglobin molecules by considering the different forms of the protein in organisms at different positions on the phylogenetic tree of life. A molecule like hemoglobin was necessary to allow multicellular animals to grow to a large size, since large animals could no longer rely on the simple diffusion of oxygen through the body surface to oxygenate their tissues adequately. Consequently, hemoglobin-like molecules are found in all vertebrates and in many invertebrates. The most primitive oxygen-carrying molecule in animals is a globin polypeptide chain of about 150 amino acids, which is found in many marine worms, insects, and primitive fish. The hemoglobin molecule in higher vertebrates, however, is composed of two kinds of globin chains. It appears that about 500 million years ago, during the evolution of higher fish, a series of gene mutations and duplications occurred. These events established two slightly different globin genes, coding for the α- and β-globin chains in the genome of each individual. In modern higher vertebrates each hemoglobin molecule is a complex of two α chains and two β chains (Figure 7–115). The four oxygen-binding sites in the $\alpha_2\beta_2$ molecule interact, allowing a cooperative allosteric change in the molecule as it binds and releases oxygen, which enables hemoglobin to take up and to release oxygen more efficiently than the single-chain version.

Still later, during the evolution of mammals, the β-chain gene apparently underwent duplication and mutation to give rise to a second β-like chain that is synthesized specifically in the fetus. The resulting hemoglobin molecule has a higher affinity for oxygen than adult hemoglobin and thus helps in the transfer of oxygen from the mother to the fetus. The gene for the new β-like chain subsequently mutated and duplicated again to produce two new genes, ε and γ, the ε chain being produced earlier in development (to form $\alpha_2\varepsilon_2$) than the fetal γ chain, which forms $\alpha_2\gamma_2$. A duplication of the adult β-chain gene occurred still later, during primate evolution, to give rise to a δ-globin gene and thus to a minor form of hemoglobin ($\alpha_2\delta_2$) found only in adult primates (Figure 7–116).

Each of these duplicated genes has been modified by point mutations that affect the properties of the final hemoglobin molecule, as well as by changes in regulatory regions that determine the timing and level of expression of the gene. As a result, each globin is made in different amounts at different times of human development (see Figure 7–60B).

The end result of the gene duplication processes that have given rise to the diversity of globin chains is seen clearly in the human genes that arose from the original β gene, which are arranged as a series of homologous DNA sequences located within 50,000 nucleotide pairs of one another. A similar cluster of α-globin genes is located on a separate human chromosome. Because the α- and β-globin gene clusters are on separate chromosomes in birds and mammals but are together in the frog *Xenopus,* it is believed that a chromosome translocation event separated the two gene clusters about 300 million years ago (see Figure 7–116).

There are several duplicated globin DNA sequences in the α- and β-globin gene clusters that are not functional genes, but pseudogenes. These have a close homology to the functional genes but have been disabled by mutations that prevent their expression. The existence of such pseudogenes make it clear

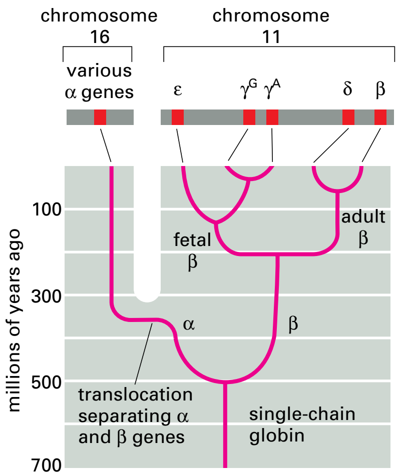


**Figure 7–116 An evolutionary scheme for the globin chains that carry oxygen in the blood of animals.** The scheme emphasizes the β-like globin gene family. A relatively recent gene duplication of the γ-chain gene produced $\gamma^G$ and $\gamma^A$, which are fetal β-like chains of identical function. The location of the globin genes in the human genome is shown at the top of the figure (see also Figure 7–60).

that, as expected, not every DNA duplication leads to a new functional gene. We also know that nonfunctional DNA sequences are not rapidly discarded, as indicated by the large excess of noncoding DNA that is found in mammalian genomes.

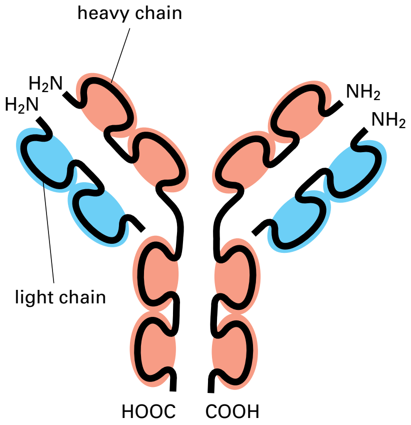


**Figure 7–117 Schematic view of an antibody (immunoglobulin) molecule.** This molecule is a complex of two identical heavy chains and two identical light chains. Each heavy chain contains four similar, covalently linked domains, while each light chain contains two such domains. Each domain is encoded by a separate exon, and all of the exons are thought to have evolved by the serial duplication of a single ancestral exon.

## Genes Encoding New Proteins Can Be Created by the Recombination of Exons

The role of DNA duplication in evolution is not confined to the expansion of gene families. It can also act on a smaller scale to create single genes by stringing together short, duplicated segments of DNA. The proteins encoded by genes generated in this way can be recognized by the presence of repeating, similar protein domains, which are covalently linked to one another in series. The immunoglobulins (Figure 7–117) and albumins, for example, as well as most fibrous proteins (such as collagens) are encoded by genes that have evolved by repeated duplications of a primordial DNA sequence.

In genes that have evolved in this way, as well as in many other genes, each separate exon often encodes an individual protein folding unit, or domain. It is believed that the organization of DNA coding sequences as a series of such exons separated by long introns has greatly facilitated the evolution of new proteins. The duplications necessary to form a single gene coding for a protein with repeating domains, for example, can occur by breaking and rejoining the DNA anywhere in the long introns on either side of an exon encoding a useful protein domain; without introns there would be only a few sites in the original gene at which a recombinational exchange between DNA molecules could duplicate the domain. By enabling the duplication to occur by recombination at many potential sites rather than just a few, introns increase the probability of a favorable duplication event.

More generally, we know from genome sequences that component parts of genes—both their individual exons and their regulatory elements—have served as modular elements that have been duplicated and moved about the genome to create the present great diversity of living things. As a result, many present-day proteins are formed as a patchwork of domains from different domain families, reflecting their long evolutionary history (Figure 7–118).

## Genome Sequences Have Left Scientists with Many Mysteries to Be Solved

Now that we know from genome sequences that a human and a mouse contain essentially the same genes, we are forced to confront one of the major problems that will challenge cell biologists throughout the next century. Given that a human and a mouse are formed from the same set of proteins, what has happened during the evolutionary process to make a mouse and a human so different? Although the answer is present somewhere among the three billion nucleotides in each sequenced genome, we do not yet know how to decipher this type of information—so that the answer to this critical, most fundamental question is not known.

Despite our ignorance, it is perhaps worth engaging in a bit of speculation, if only to help point the way forward to some of the hard problems ahead. In biology, timing is everything, as will become clear when we examine the elaborate mechanisms that allow a fertilized egg to develop into an embryo, and the embryo to develop into an adult (discussed in Chapter 21). The human body is formed as the result of many billions of decisions that are made during our development as to which RNA molecule and which protein are to be made where, as well as exactly when and in what amount each is to be produced.

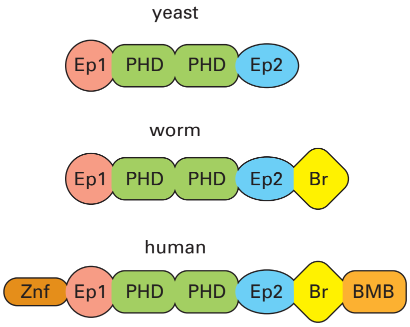


**Figure 7–118 Domain structure of a group of evolutionary related proteins that are thought to have a similar function.** In general, there is a tendency for the proteins in more complex organisms, such as ourselves, to contain additional domains—as is the case for the DNA-binding protein compared here.

These decisions are different for a human than for a chimpanzee or a mouse. The coding sequences of genomes represent a more or less standard set of the 30,000 or so basic parts from which all three organisms are made. It is therefore the many different types of controls on gene expression described in this Chapter that must largely create the difference between a human and other mammals.

Given these assumptions, it would be reasonable to expect genomes to have evolved in a way that allows organisms to experiment with altered gene timing and expression patterns in selected cells. We have already seen some evidence that this is so, when we discussed alternative RNA splicing and RNA editing mechanisms. There also appear to be mechanisms—some based on the movements of transposable DNA elements—that allow modules to be readily added to and subtracted from the regulatory regions of genes, so as to produce changes in the pattern of their transcription as organisms evolve. In fact, an analysis of these regulatory regions provides evidence to support the claim that most gene regulatory regions have been formed by the evolutionary mixing and matching of the DNA-binding sites that are recognized by gene regulatory proteins (Figure 7–119).

## Genetic Variation within a Species Provides a Fine-Scale View of Genome Evolution

In comparisons between two species that have diverged from one another by millions of years, it makes little difference which individuals from each species are compared. For example, typical human and chimpanzee DNA sequences differ from one another by 1%. In contrast, when the same region of the genome is sampled from two different humans, the differences are typically less than 0.1%. For more distantly related organisms, the inter-species differences overshadow intra-species variation even more dramatically. However, each "fixed difference" between the human and the chimpanzee (*i.e.*, each difference that is now characteristic of all or nearly all individuals of each species) started out as a new mutation in a single individual. If the size of the interbreeding population in which the mutation occurred is N, the initial **allele frequency** of a new mutation would be 1/2N for a diploid organism. How does such a rare mutation become fixed in the population, and hence become a characteristic of the species rather than of a particular individual genome?

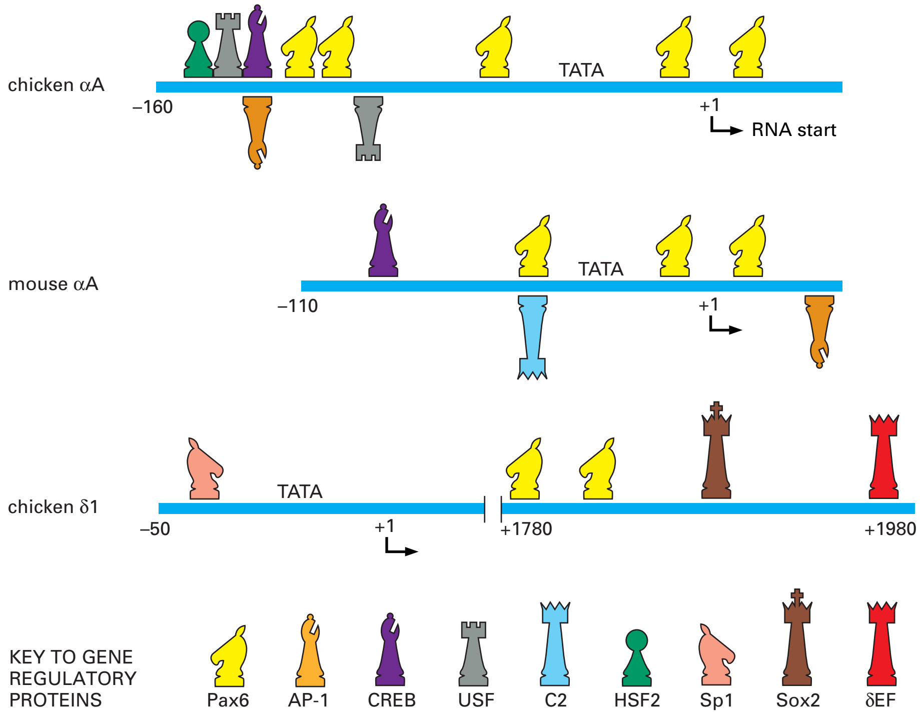


**Figure 7–119 Gene control regions for mouse and chicken eye lens crystallins.** Crystallins make up the bulk of the lens and are responsible for refracting and focusing light onto the retina. Many proteins in the cell have properties (high solubility, proper refractive index, etc.) suitable for lens function, and a wide variety of such proteins have been co-opted during evolution for use in the lens. For example, the α crystallins *(top two lines)* are closely related to heat shock proteins and are found in all vertebrate lenses. In contrast, δ crystallin *(third line)* is closely related to an enzyme involved in amino acid metabolism and is found only in birds and reptiles. The three crystallin gene control regions shown are a patchwork of different regulatory sequences that reflect the evolutionary history of each gene. The common feature of all three control regions is the presence of binding sites for the gene regulatory protein Pax6. Pax6 is the vertebrate homolog of the fly Toy and Eyeless proteins (see Figure 7–75) and is one of the key regulators that specifies eye development. Proteins above each gene control region are transcriptional activators and those below the line are repressors. (Adapted from E.H. Davidson, Genomic Regulatory Systems: Development and Evolution, pp. 191–201. San Diego: Academic Press, 2001 and A. Cvekl and J. Piatigarsky, *BioEssays* 18:621–630, 1996.)

The answer to this question depends on the functional consequences of the mutation. If the mutation has a significantly deleterious effect, it will simply be eliminated by purifying selection and will not become fixed. (In the most extreme case, the individual carrying the mutation will die without producing progeny.) Conversely, the rare mutations that confer a major reproductive advantage on individuals who inherit them will spread rapidly in the population. Because humans reproduce sexually and genetic recombination occurs each time a gamete is formed, the genome of each individual who has inherited the mutation will be a unique recombinational mosaic of segments inherited from a large number of ancestors. The selected mutation along with a modest amount of neighboring sequence—ultimately inherited from the individual in which the mutation occurred—will simply be one piece of this huge mosaic.

The great majority of mutations that are not harmful are not beneficial either. These *selectively neutral mutations* can also spread and become fixed in a population, and they make a large contribution to the evolutionary change in genomes. Their spread is not as rapid as the spread of the rare strongly advantageous mutations. The process by which such neutral genetic variation is passed down through an idealized interbreeding population can be described mathematically by equations that are surprisingly simple. The idealized model that has proven most useful for analyzing human genetic variation assumes a constant population size, and random mating, as well as selective neutrality for the mutations. While neither of these assumptions is a good description of human population history, they nonetheless provide a useful starting point for analyzing intra-species variation.

When a new neutral mutation occurs in a constant population of size $N$ that is undergoing random mating, the probability that it will ultimately become fixed is approximately ½$N$. For those mutations that do become fixed, the average time to fixation is approximately $4N$ generations. A detailed analysis of data on human genetic variation suggests an ancestral population size of approximately 10,000 during the period when the current pattern of genetic variation was largely established. Under these conditions, the probability that a new, selectively neutral mutation would become fixed was small ($5 \times 10^{-5}$), while the average time to fixation was on the order of 800,000 years. Thus, while we know that the human population has grown enormously since the development of agriculture approximately 15,000 years ago, most human genetic variation arose and became established in the human population much earlier than this, when the human population was still small.

Even though most of the variation among modern humans originates from variation present in a comparatively tiny group of ancestors, the number of variations encountered is very large. Most of the variations take the form of **single-nucleotide polymorphisms (SNPs)**. These are simply points in the genome sequence where one large fraction of the human population has one nucleotide, while another large fraction has another. Two human genomes sampled from the modern world population at random will differ at approximately $2.5 \times 10^6$ sites (1 per 1300 nucleotide pairs). Mapped sites in the human genome that are polymorphic—meaning that there is a reasonable probability that the genomes of two individuals will differ at that site—are extremely useful for genetic analyses, in which one attempts to associate specific traits (phenotypes) with specific DNA sequences for medical or scientific purposes (see p. 531).

Against the background of ordinary SNPs inherited from our prehistoric ancestors, certain sequences with exceptionally high mutation rates stand out. A dramatic example is provided by *CA repeats,* which are ubiquitous in the human genome and in the genomes of other eucaryotes. Sequences with the motif $(CA)_n$ are replicated with relatively low fidelity because of a slippage that occurs between the template and the newly synthesized strands during DNA replication; hence, the precise value of $n$ can vary over a considerable range from one genome to the next. These repeats make ideal DNA-based genetic markers, since most humans are heterozygous—carrying two values of $n$ at any particular CA repeat, having inherited one repeat length *(n)* from their mother and a different repeat length from their father. While the value of $n$ changes sufficiently rarely that most parent-child transmissions propagate CA repeats faithfully, the changes are sufficiently frequent to maintain high levels of heterozygosity in the human population. These and other simple repeats that display exceptionally high variability provide the basis for identifying individuals by DNA analysis in crime investigations, paternity suits, and other forensic applications (see Figure 8–41).

While most of the SNPs and other common variations in the human genome sequence are thought to have no effect on phenotype, a subset of them must be responsible for nearly all of the heritable aspects of human individuality. A major challenge in human genetics is to learn to recognize those relatively few variations that are functionally important—against the large background of neutral variation that distinguishes the genomes of any two human beings.

## Summary

***Comparisons of the nucleotide sequences of present-day genomes have revolutionized our understanding of gene and genome evolution. Due to the extremely high fidelity of DNA replication and DNA repair processes, random errors in maintaining the nucleotide sequences in genomes occur so rarely that only about 5 nucleotides in 1000 are altered every million years. Not surprisingly, therefore, a comparison of human and chimpanzee chromosomes—which are separated by about 5 million years of evolution—reveals very few changes. Not only are our genes essentially the same, but their order on each chromosome is almost identical. In addition, the positions of the transposable elements that make up a major portion of our noncoding DNA are mostly unchanged.***

***When one compares the genomes of two more distantly related organisms—such as a human and a mouse, separated by about 100 million years—one finds many more changes. Now the effects of natural selection can be clearly seen: through purifying selection, essential nucleotide sequences—both in regulatory regions and coding sequences (exon sequences)—have been highly conserved. In contrast, nonessential sequences (for example, intron sequences) have been altered to such an extent that an accurate alignment according to ancestry is often not possible.***

***Because of purifying selection, homologous genes can be recognized over large phylogenetic distances, and it is often possible to construct a detailed evolutionary history of a particular gene, tracing its history back to common ancestors of present-day species. We can thereby see that a great deal of the genetic complexity of present-day organisms is due to the expansion of ancient gene families. DNA duplication followed by sequence divergence has thus been a major source of genetic novelty during evolution.***

## References

### General

Carey M & Smale ST (2000) Transcriptional Regulation in Eukaryotes: Concepts, Strategies and Techniques. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Hartwell L, Hood L, Goldberg ML et al. (2000) Genetics: from Genes to Genomes. Boston: McGraw Hill.

Lewin B (2000) Genes VII. Oxford: Oxford University Press.

Lodish H, Berk A, Zipursky SL et al. (2000) Molecular Cell Biology, 4th edn. New York: WH Freeman.

McKnight SL & Yamamoto KR (eds) (1992) Transcriptional Regulation. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Mechanisms of Transcription. (1998) *Cold Spring Harb. Symp. Quant. Biol.* 63.

### An Overview of Gene Control

Campbell KH, McWhir J, Ritchie WA & Wilmut I (1996) Sheep cloned by nuclear transfer from a cultured cell line. *Nature* 380, 64–66.

Gurdon JB (1992) The generation of diversity and pattern in animal development. *Cell* 68, 185–199.

Ross DT, Scherf U, Eisen MB et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24, 227–235.

### DNA-binding Motifs in Gene Regulatory Proteins

Bulger M & Groudine M (199) Looping versus linking: toward a model for long-distance gene activation. *Genes Dev.* 13, 2465–2477.

Choo Y & Klug A (1997) Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* 7, 117–125.

Gehring WJ, Affolter M & Burglin T (1994) Homeodomain proteins. *Annu. Rev. Biochem.* 63, 487–526.

Harrison SC (1991) A structural taxonomy of DNA-binding domains. *Nature* 353, 715–719.

Jacob F & Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356.

Laity JH, Lee BM & Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* 11, 39–46.

Lamb P & McKnight SL (1991) Diversity and specificity in transcriptional regulation: the benefits of heterotypic dimerization. *Trends Biochem. Sci.* 16, 417–422.

McKnight SL (1991) Molecular zippers in gene regulation. *Sci. Am.* 264, 54–64.

Muller CW (2001) Transcription factors: global and detailed views. *Curr. Opin. Struct. Biol.* 11, 26–32.

Orlando V (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin. *Trends Biochem. Sci.* 25, 99–104.

Pabo CO & Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* 61, 1053–1095.

Ptashne M (1992) A Genetic Switch, 2nd edn. Cambridge, MA: Cell Press and Blackwell Press.

Rhodes D, Schwabe JW, Chapman L et al. (1996) Towards an understanding of protein–DNA recognition. *Proc. R. Soc. Lond.* B 351, 501–509.

Seeman NC, Rosenberg JM & Rich A (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA* 73, 804–808.

Wolberger C (1996) Homeodomain interactions. *Curr. Opin. Struct. Biol.* 6, 62–68.

### How Genetic Switches Work

Beckwith J (1987) The operon: an historical account. In *Escherichia coli* and *Salmonella typhimurium:* Cellular and Molecular Biology (Neidhart FC, Ingraham JL, Low KB et al. eds), vol 2, pp 1439–1443. Washington, DC: ASM Press.

Bell AC, West AG & Felsenfeld G (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic. *Science* 291, 447–450.

Buratowski S (2000) Snapshots of RNA polymerase II transcription initiation. *Curr. Opin. Cell Biol.* 12, 320–325.

Carey M (1998) The enhanceosome and transcriptional synergy. *Cell* 92, 5–8.

Fraser P & Grosveld F (1998) Locus control regions, chromatin activation and transcription. *Curr. Opin. Cell Biol.* 10, 361–365.

Kadonaga JT (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell* 92, 307–313.

Kercher MA, Lu P & Lewis M (1997) Lac repressor-operator complex. *Curr. Opin. Struct. Biol.* 7, 76–85.

Kornberg RD (1999) Eukaryotic transcriptional control. *Trends Cell Biol.* 9, M46–49.

Malik S & Roeder RG (2000) Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. *Trends Biochem. Sci.* 25, 277–283.

Merika M & Thanos D (2001) Enhanceosomes. *Curr. Opin. Genet. Dev.* 11, 205–208.

Myers LC & Kornberg RD (2000) Mediator of transcriptional regulation. *Annu. Rev. Biochem.* 69, 729–749.

Ptashne M & Gann A (1998) Imposing specificity by localization: mechanism and evolvability. *Curr. Biol.* 8, R812–R822.

Schleif R (1992) DNA looping. *Annu. Rev. Biochem.* 61, 199–223.

St Johnston D & Nusslein-Volhard C (1992) The origin of pattern and polarity in the *Drosophila* embryo. *Cell* 68, 201–219.

Struhl K (1998) Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* 12, 599–606.

### The Molecular Genetic Mechanisms that Create Specialized Cell Types

Bird AP & Wolffe AP (1999) Methylation-induced repression—belts, braces, and chromatin. *Cell* 99, 451–454.

Cross SH & Bird AP (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.* 5, 309–314.

Dasen J & Rosenfeld M (1999) Combinatorial codes in signaling and synergy: lessons from pituitary development. *Curr. Opin. Genet. Dev.* 9, 566–574.

Gehring WJ & Ikeo K (1999) Pax 6: mastering eye morphogenesis and eye evolution. *Trends Genet.* 15, 371–377.

Haber JE (1998) Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* 32, 561–599.

Marin I, Siegal ML & Baker BS (2000) The evolution of dosage-compensation mechanisms. *Bioessays* 22, 1106–1114.

Meyer BJ (2000) Sex in the worm: counting and compensating X-chromosome dose. *Trends Genet.* 16, 247–253.

Robertson BD & Meyer TF (1992) Genetic variation in pathogenic bacteria. *Trends Genet.* 8, 422–427.

Surani MA (1998) Imprinting and the initiation of gene silencing in the germ line. *Cell* 93, 309–312.

Weintraub H (1993) The MyoD family and myogenesis: redundancy, networks, and thresholds. *Cell* 75, 1241–1244.

Wolberger C (1999) Multiprotein-DNA complexes in transcriptional regulation. *Annu. Rev. Biophys. Biomol. Struct.* 28, 29–56.

Young MW (1998) The molecular control of circadian behavioral rhythms and their entrainment in Drosophila. *Annu. Rev. Biochem.* 67, 135–152.

### Posttranscriptional Controls

Baker BS (1989) Sex in flies: the splice of life. *Nature* 340, 521–524.

Benne R (1996) RNA editing: how a message is changed. *Curr. Opin. Genet. Dev.* 6, 221–231.

Cline TW & Meyer BJ (1996) Vive la difference: males vs females in flies vs. worms. *Annu. Rev. Genet.* 30, 637–702.

Dever TE (1999) Translation initiation: adept at adapting. *Trends Biochem. Sci.* 24, 398–403.

Frankel AD & Young JAT (1998) HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* 67, 1–25.

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107.

Gray NK & Wickens M (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.* 14, 399–458.

Hentze MW & Kulozik AE (1999) A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* 96, 307–310.

Hinnebusch AG (1997) Translational regulation of yeast GCN4. A window on factors that control initiator-tRNA binding to the ribosome. *J. Biol. Chem.* 272, 21661–21664.

Holcik M, Sonenberg N & Korneluk RG (2000) Internal ribosome initiation of translation and the control of cell death. *Trends Genet.* 16, 469–473.

Jansen RP (2001) mRNA localization: message on the move. *Nat Rev Mol Cell Biol* 2, 247–256.

Pollard VW & Malim MH (1998) The HIV-1 Rev protein. *Annu. Rev. Microbiol.* 52, 491–532.

Sharp PA (2001) RNA interference – 2001. *Genes Dev.* 15, 485–490.

Wilusz CJ, Wormington M & Peltz SW (2001) The cap-to-tail guide to mRNA turnover. *Nat. Rev. Mol. Cell Biol.* 2, 237–246.

### How Genomes Evolve

Dehal P, Predki P, Olsen AS et al. (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 293, 104–111.

Henikoff S, Greene EA, Pietrokovski S et al. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278, 609–614.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Kumar S & Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392, 917–920.

Li WH, Gu Z, Wang H & Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature* 409, 847–849.

Long M, de Souza SJ & Gilbert W (1995) Evolution of the intron-exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.* 5, 774–778.

Rowold DJ & Herrera RJ (2000) Alu elements and the human genome. *Genetica* 108, 57–72.

Stoneking M (2001) Single nucleotide polymorphisms. From the evolutionary past. *Nature* 409, 821–822.

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.