CHAPTER **EIGHT** **8**

Control of Gene Expression

An organism's DNA encodes all of the RNA and protein molecules that are needed to make 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 in each human cell—does not enable us to reconstruct that organism any more than a list of all the English words in a dictionary enables us to reconstruct a Shakespeare play. We need to know how the elements in the DNA sequence or the words on a list work together to produce the masterpiece.

For cells, the answer comes down to *gene expression*. Even the simplest single-celled bacterium can use its genes selectively—for example, switching genes on and off to make the enzymes needed to digest whatever food sources are available. In multicellular plants and animals, gene expression is even more elaborate. Over the course of embryonic development, a fertilized egg cell gives rise to many cell types that differ dramatically in both structure and function. The differences between an information-processing nerve cell and toxin-neutralizing liver cell, for example, are so extreme that it is difficult to imagine that the two cells contain the same DNA (**Figure 8–1**). For this reason, and because cells in an adult organism rarely lose their distinctive characteristics, biologists originally suspected that certain genes might be selectively eliminated from cells as they become specialized. We now know, however, that nearly all the cells of a multicellular organism contain the same genome. Cell *differentiation* is instead achieved by changes in gene expression.

In mammals, hundreds of different cell types carry out a range of specialized functions that depend upon genes that are switched on in that cell type but not in most others: for example, the β cells of the pancreas

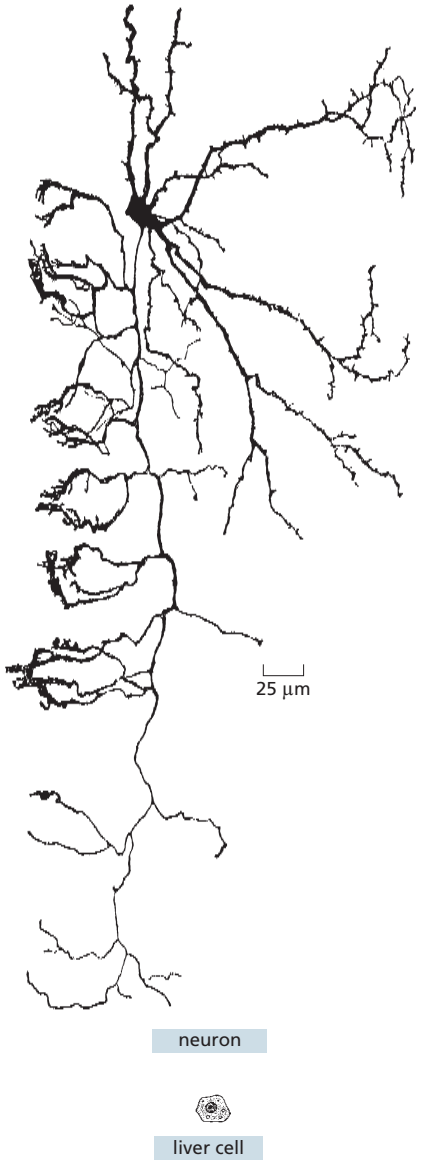

Figure 8–1 A neuron and a liver cell share the same genome. The long branches of this neuron from the retina enable it to receive electrical signals from numerous other neurons and pass these signals along to many neighboring neurons. The liver cell, which is drawn to the same scale, is involved in many metabolic processes, including digestion and the detoxification of alcohol and other drugs. Both of these mammalian cells contain the same genome, but they express different RNAs and proteins. (Neuron adapted from S. Ramón y Cajal, Histologie du Système Nerveux de l'Homme et de Vertébrés, 1909–1911. Paris: Maloine; reprinted, Madrid: C.S.I.C., 1972.)

make the protein hormone insulin, while the α cells of the pancreas make the hormone glucagon; the B lymphocytes of the immune system make antibodies, while developing red blood cells make the oxygen-transport protein hemoglobin. The differences between a neuron, a white blood cell, a pancreatic β cell, and a red blood cell depend on the precise control of gene expression. A typical differentiated cell expresses only about half the genes in its total repertoire. This selection, which differs from one cell type to the next, is the basis for the specialized properties of each cell type.

In this chapter, we discuss the main ways in which gene expression is regulated, with a focus on those genes that encode proteins as their final product. Although some of these control mechanisms apply to both eukaryotes and prokaryotes, eukaryotic cells—with their larger number of genes and more complex chromosomes—have some additional ways of controlling gene expression that are not found in bacteria.

AN OVERVIEW OF GENE EXPRESSION

Gene expression is a complex process by which cells selectively direct the synthesis of the many thousands of proteins and RNAs encoded in their genome. But how do cells coordinate and control such an intricate process—and how does an individual cell specify which of its genes to express? This decision is an especially important problem for animals because, as they develop, their cells become highly specialized, ultimately producing an array of muscle, nerve, and blood cells, along with the hundreds of other cell types seen in the adult. Such cell **differentiation** arises because cells make and accumulate different sets of RNA and protein molecules: that is, they express different genes.

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

The evidence that cells have the ability to change which genes they express without altering the nucleotide sequence of their DNA comes from experiments in which the genome from a differentiated cell is made to direct the development of a complete organism. If the chromosomes of the differentiated cell were altered irreversibly during development—for example, by jettisoning some of their genes—they would not be able to accomplish this feat.

Consider, for example, an experiment in which the nucleus is taken from a skin cell in an adult frog and injected into a frog egg from which the nucleus has been removed. In at least some cases, that doctored egg will develop into a normal tadpole (**Figure 8–2**). Thus, the nucleus from the transplanted skin cell cannot have lost any critical DNA sequences. Nuclear transplantation experiments carried out with differentiated cells taken from adult mammals—including sheep, cows, pigs, goats, and mice—have shown similar results. And in plants, individual cells removed from a carrot, for example, can regenerate an entire adult carrot plant.

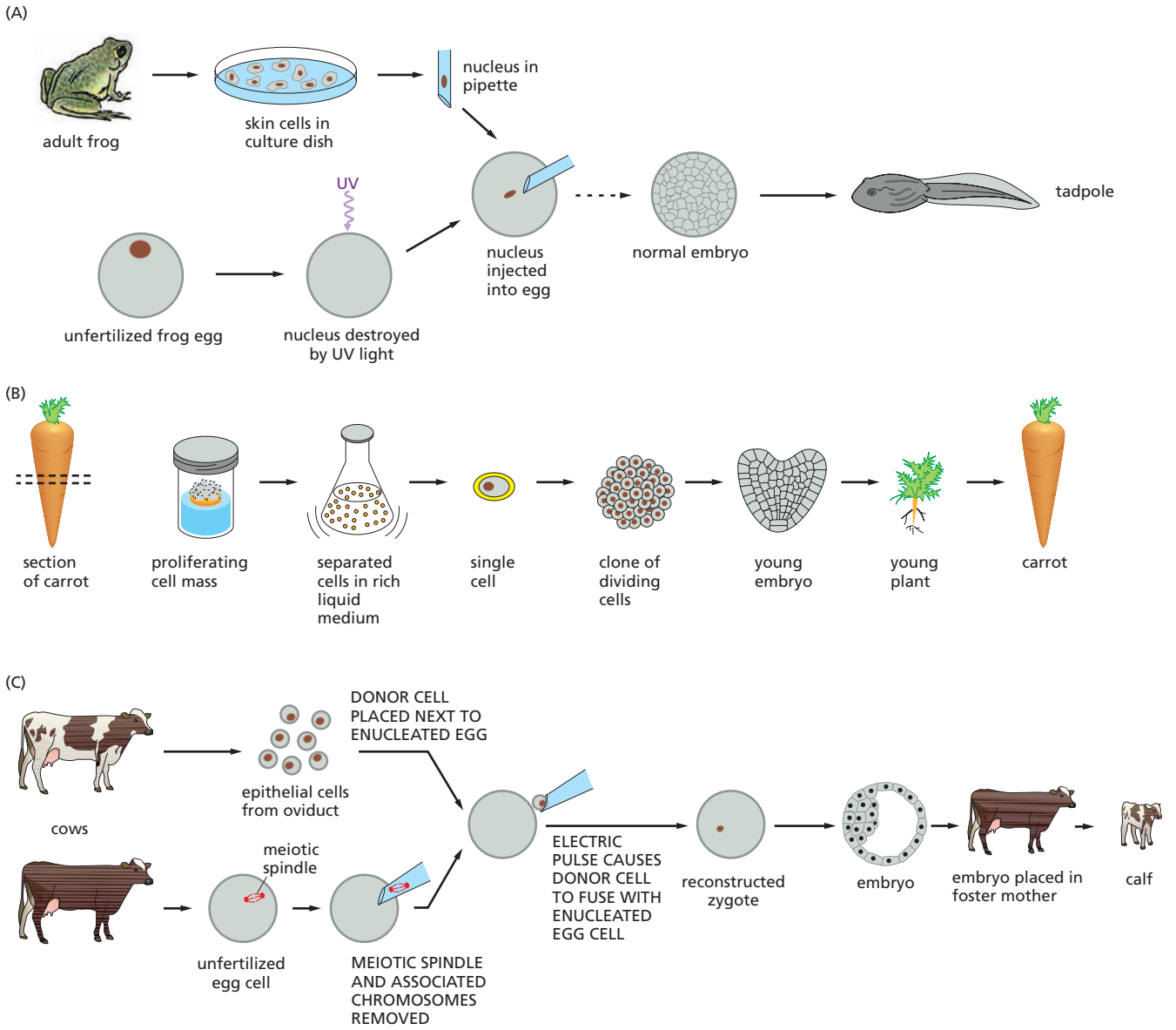

Figure 8–2 Differentiated cells contain all the genetic instructions needed to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an "enucleated" egg—one whose nucleus has been destroyed—can give rise to an entire tadpole. The broken arrow 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 "de-differentiate," so that a single cell can proliferate to form a clone of progeny cells that later give rise to an entire plant. (C) A nucleus removed from a differentiated cell of an adult cow can be introduced into an enucleated egg from a different cow to give rise to a calf. Different calves produced from the same differentiated cell donor are all clones of the donor and are therefore genetically identical. The cloned sheep Dolly was produced by this type of nuclear transplantation. (A, modified from J.B. Gurdon, *Sci. Am.* 219:24–35, 1968.)

These experiments all demonstrate that the DNA in specialized cell types of multicellular organisms still contains the entire set of instructions needed to form a whole organism. The various cell types of an organism therefore differ not because they contain different genes, but because they express them differently.

Different Cell Types Produce Different Sets of Proteins

The extent of the differences in gene expression between different cell types may be roughly gauged by comparing the protein composition of cells in liver, heart, brain, and so on. In the past, such analysis

was performed by two-dimensional gel electrophoresis (see Panel 4–5, p. 167). Nowadays, the total protein content of a cell can be rapidly analyzed by a method called mass spectrometry (see Figure 4–56). This technique is much more sensitive than electrophoresis and it enables the detection of proteins that are produced even in minor quantities.

Both techniques reveal that many proteins are common to all the cells of a multicellular organism. These *housekeeping* proteins include, for example, RNA polymerases, DNA repair enzymes, ribosomal proteins, enzymes involved in glycolysis and other basic metabolic processes, and many of the proteins that form the cytoskeleton. In addition, each different cell type also produces specialized proteins that are responsible for the cell's distinctive properties. In mammals, for example, hemoglobin is made almost exclusively in developing red blood cells.

Gene expression can also be studied by cataloging a cell's RNA molecules, including the mRNAs that encode protein. The most comprehensive methods for such analyses involve determining the nucleotide sequence of all RNAs made by the cell, an approach that can also reveal the relative abundance of each. Estimates of the number of different mRNA sequences in human cells suggest that, at any one time, a typical differentiated human cell expresses perhaps 5000–15,000 protein-coding genes from a total of about 19,000. And studies of a variety of tissue types confirm that the collection of expressed mRNAs differs from one cell type to the next.

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

Although each cell type in a multicellular organism expresses its own group of genes, these collections are not static. Specialized cells are capable of altering their patterns of gene expression in response to extracellular cues. For example, if a liver cell is exposed to the steroid hormone cortisol, the production of several proteins is dramatically increased. Released by the adrenal gland during periods of starvation, intense exercise, or prolonged stress, cortisol signals liver cells to boost the production of glucose from amino acids and other small molecules. The set of proteins whose production is induced by cortisol includes enzymes such as tyrosine aminotransferase, which helps convert tyrosine to glucose. When the hormone is no longer present, the production of these proteins returns to its resting level.

Other cell types respond to cortisol differently. In fat cells, for example, the production of tyrosine aminotransferase is reduced; some other cell types do not respond to cortisol at all. The fact that different cell types often respond in different ways to the same extracellular signal contributes to the specialization that gives each cell type its distinctive character.

Gene Expression Can Be Regulated at Various Steps from DNA to RNA to Protein

If differences among the various cell types of an organism depend on the particular genes that each cell expresses, at what level is this control of gene expression exercised? As we discussed in the previous chapter, there are many steps in the pathway leading from DNA to protein, and each of them can in principle be regulated. Thus a cell can control the proteins it contains by (1) controlling when and how often a given gene is transcribed, (2) controlling how an RNA transcript is spliced or otherwise processed, (3) selecting which mRNAs are exported from the nucleus to the cytosol, (4) regulating how quickly certain mRNA molecules are

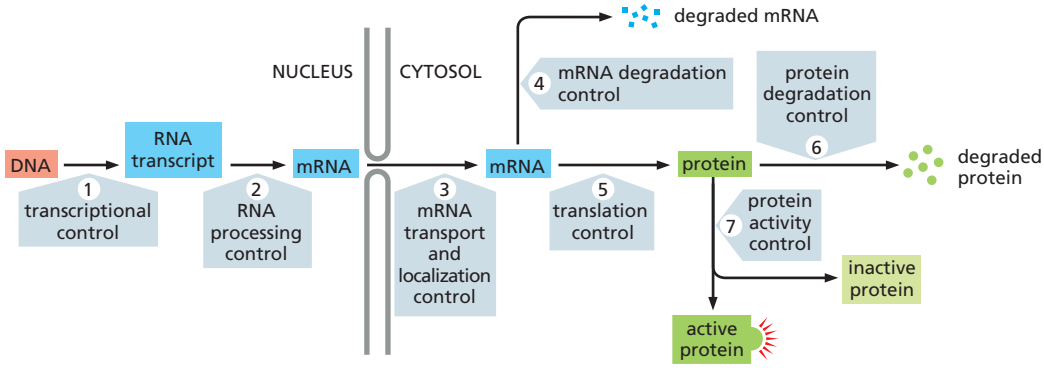

Figure 8–3 Gene expression in eukaryotic cells can be controlled at various steps. Examples of regulation at each of these steps are known, although for most genes the main site of control is step 1: transcription of a DNA sequence into RNA.

degraded, (5) selecting which mRNAs are translated into protein by ribosomes, or (6) regulating how rapidly specific proteins are destroyed after they have been made; in addition, the activity of individual proteins, once they have been synthesized, can be further regulated in a variety of ways.

In eukaryotic cells, gene expression can be regulated at each of these steps (**Figure 8–3**). For most genes, however, the control of transcription (shown in step 1) is paramount. This makes sense because only transcriptional control can ensure that no unnecessary intermediates are synthesized. Thus it is the regulation of transcription—and the DNA and protein components that determine which genes a cell transcribes into RNA—that we address first.

HOW TRANSCRIPTION IS REGULATED

Until 50 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 studies of how *E. coli* bacteria adapt to changes in the composition of their growth medium. Many of the same principles apply to eukaryotic cells. However, the enormous complexity of gene regulation in organisms that possess a nucleus, combined with the packaging of their DNA into chromatin, creates special challenges and some novel opportunities for control—as we will see. We begin with a discussion of the *transcription regulators* (often loosely referred to as transcription factors), proteins that bind to specific DNA sequences and control gene transcription.

Transcription Regulators Bind to Regulatory DNA Sequences

Nearly all genes, whether bacterial or eukaryotic, contain sequences that direct and control their transcription. In Chapter 7, we saw that the **promoter** region of a gene binds the enzyme *RNA polymerase* and correctly orients the enzyme to begin its task of making an RNA copy of the gene. The promoters of both bacterial and eukaryotic genes include a *transcription initiation site*, where RNA synthesis begins, plus nearby sequences that contain recognition sites for proteins that associate with RNA polymerase: sigma factor in bacteria (see Figure 7–9) or the general transcription factors in eukaryotes (see Figure 7–12).

In addition to the promoter, the vast majority of genes include **regulatory DNA sequences** that are used to switch the gene on or off. Some regulatory DNA sequences are as short as 10 nucleotide pairs and act as simple switches that respond to a single signal; such simple regulatory switches predominate in bacteria. Other regulatory DNA sequences, especially those in eukaryotes, are very long (sometimes spanning more than 100,000 nucleotide pairs) and act as molecular microprocessors,

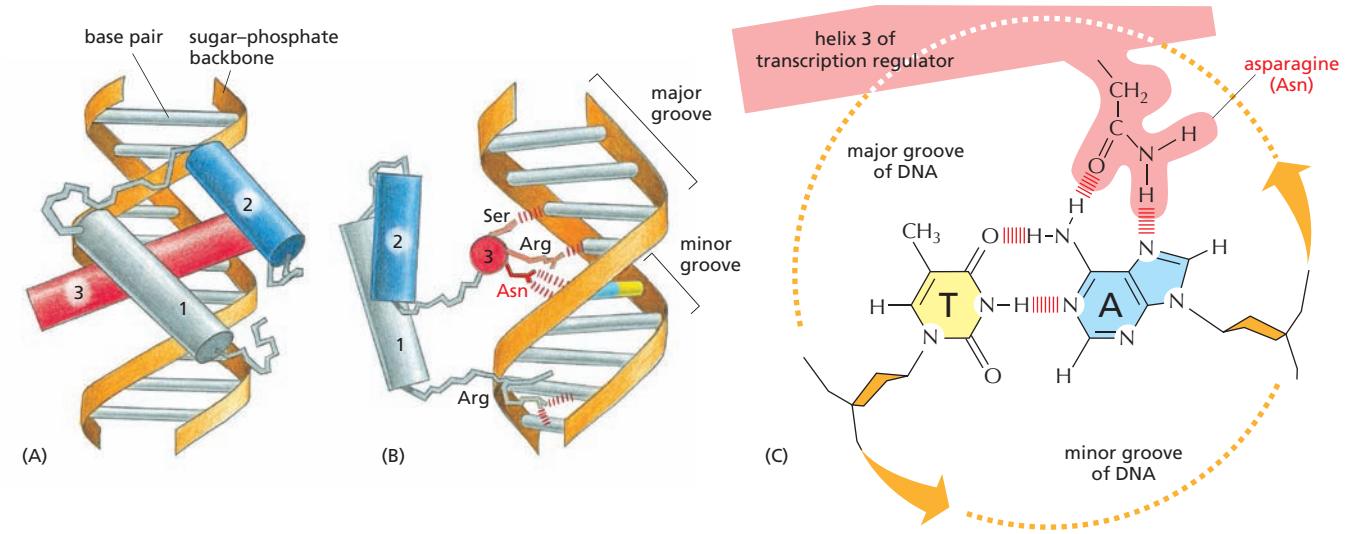

Figure 8–4 A transcription regulator interacts with the DNA double helix. (A) The regulator shown recognizes DNA via three α helices, drawn as numbered cylinders, which allow the protein to fit into the major groove and form tight associations with the base pairs in a short stretch of DNA. This particular structural motif, called a *homeodomain*, is found in many eukaryotic DNA-binding proteins (**Movie 8.1**). (B) Most of the contacts with the DNA bases are made by helix 3 (*red*), which is shown here end-on. (C) An asparagine side chain from helix 3 forms two hydrogen bonds with the adenine in an A-T base pair. The view is end-on, looking down the center of the DNA double helix, and the protein contacts the base pair from the major-groove side. Note that the interactions between the protein and DNA take place along the edges of the nucleotide base and do not disrupt the hydrogen bonds that hold the base pairs together. For simplicity, only one amino acid–base contact is shown; in reality, transcription regulators form hydrogen bonds (as shown here), ionic bonds, and hydrophobic interactions with multiple bases. Most of these contacts occur in the major groove, but some proteins also interact with bases in the minor groove, as shown in (B). Typically, the protein–DNA interface would consist of 10–20 such contacts, each involving a different amino acid and each contributing to the overall strength of the protein–DNA interaction.

integrating information from a variety of signals into a command that determines how often transcription of the gene is initiated.

Regulatory DNA sequences do not work by themselves. To have any effect, these sequences must be recognized by proteins called **transcription regulators**. It is the binding of a transcription regulator to a regulatory DNA sequence that acts as the switch to control transcription. The simplest bacterium produces several hundred different transcription regulators, each of which recognizes a different DNA sequence and thereby regulates a distinct set of genes. Humans make many more—2000 or so—indicating the importance and complexity of this form of gene regulation in the development and function of a complex organism.

Proteins that recognize a specific nucleotide sequence do so because the surface of the protein fits tightly against the surface features of the DNA double helix in that region. Because these surface features will vary depending on the nucleotide sequence, different DNA-binding proteins will recognize different nucleotide sequences. In most cases, the protein inserts into the major groove of the DNA double helix and makes a series of intimate, noncovalent molecular contacts with the nucleotide pairs within the groove (**Figure 8–4, Movie 8.2**). Although each individual contact is weak, the 10 to 20 contacts that typically form at the protein–DNA interface combine to ensure that the interaction is both highly specific and very strong; indeed, protein–DNA interactions are among the tightest and most specific molecular interactions known in biology.

Many transcription regulators bind to the DNA helix as dimers. Such dimerization roughly doubles the area of contact with the DNA, thereby greatly increasing the potential strength and specificity of the protein–DNA interaction (**Figure 8–5, Movie 8.3**).

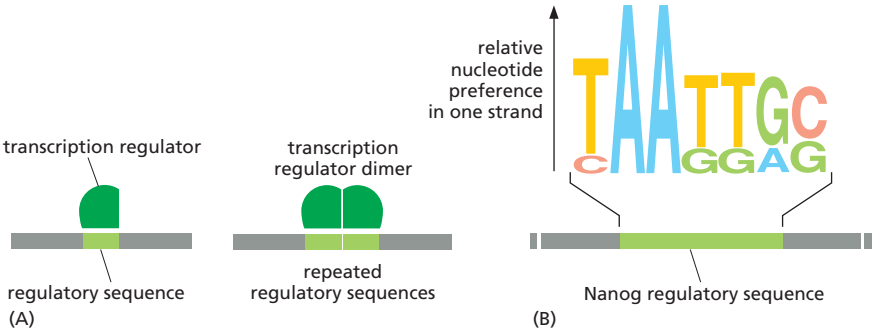

Figure 8–5 Many transcription regulators bind to DNA as dimers. (A) As shown, such dimerization doubles the number of protein–DNA contacts. Here, and throughout the book, regulatory sequences are represented by colored bars; each bar represents a double-helical segment of DNA, as in Figure 8–4. (B) Shown here is a regulatory sequence recognized by Nanog, a homeodomain family member that is a key regulator in embryonic stem cells. This diagram, called a "logo," represents the preferred nucleotide at each position of the sequence; the height of each letter is proportional to the frequency with which this base is found at that position in the regulatory sequence. In the first position, for example, T is found more often than C, while A is the only nucleotide found in the second and third position of the sequence. Although regulatory sequences in the cell are double-stranded, a logo typically shows the sequence of only one DNA strand; the other strand is simply the complementary sequence. Logos are useful because they reveal at a glance the range of DNA sequences to which a given transcription regulator will bind.

Transcription Switches Allow Cells to Respond to Changes in Their Environment

The simplest and best-understood examples of gene regulation occur in bacteria. The genome of the bacterium *E. coli* 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. Bacteria regulate the expression of many of their genes according to the food sources that are available in the environment. In *E. coli*, for example, five genes code for enzymes that manufacture tryptophan when this amino acid is scarce. These genes are arranged in a cluster on the chromosome and are transcribed from a single promoter as one long mRNA molecule; such coordinately transcribed clusters are called *operons* (**Figure 8–6**). Although operons are common in bacteria (see Figure 7–40), they are rare in eukaryotes, where genes are transcribed and regulated individually.

When tryptophan concentrations are low, the operon is transcribed; the resulting mRNA is translated to produce a full set of biosynthetic enzymes, which work in tandem to synthesize the amino acid. When tryptophan is abundant, however—for example, when the bacterium is in the gut of a mammal that has just eaten a protein-rich meal—the amino acid is imported into the cell and shuts down production of the enzymes, which are no longer needed.

We understand in considerable detail how this repression of the tryptophan operon comes about. Within the operon's promoter is a short DNA sequence, called the operator (see Figure 8–6), that is recognized by a transcription regulator. When this regulator binds to the *operator*, it blocks access of RNA polymerase to the promoter, thus preventing transcription of the operon and, ultimately, the production of the tryptophan-synthesizing enzymes. The transcription regulator is known as the *tryptophan repressor*, and it is controlled in an ingenious way: the repressor can bind to DNA only if it is also bound to tryptophan (**Figure 8–7**).

The tryptophan repressor is an allosteric protein (see Figure 4–44): the binding of tryptophan causes a subtle change in its three-dimensional

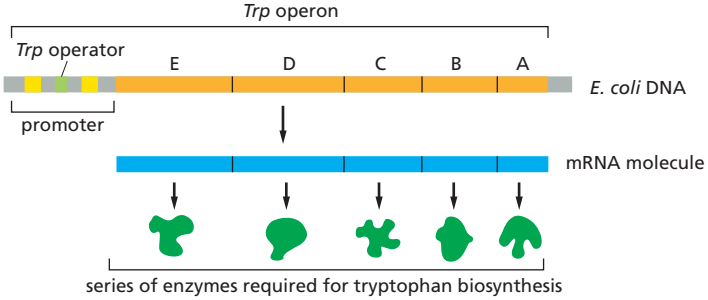

Figure 8–6 A cluster of bacterial genes can be transcribed from a single promoter. Each of these five genes encodes a different enzyme; all of the enzymes are needed to synthesize the amino acid tryptophan from simpler molecular building blocks. The genes are transcribed as a single mRNA molecule, a feature that allows their expression to be coordinated. Such clusters of genes, called operons, are common in bacteria. In this case, the entire operon is controlled by a single regulatory DNA sequence, called the *Trp operator* (*green*), situated within the promoter. The *yellow* blocks in the promoter represent DNA sequences that bind RNA polymerase.

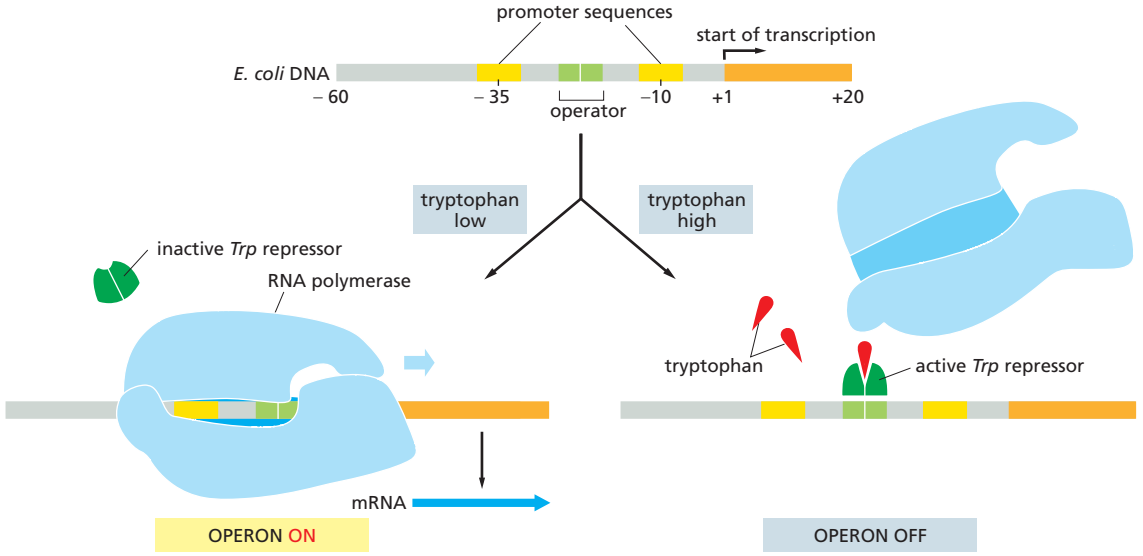

Figure 8–7 Genes can be switched off by repressor proteins. If the concentration of tryptophan inside a bacterium is low (*left*), RNA polymerase (*blue*) binds to the promoter and transcribes the five genes of the tryptophan operon. However, if the concentration of tryptophan is high (*right*), the repressor protein (*dark green*) becomes active and binds to the operator (*light green*), where it blocks the binding of RNA polymerase to the promoter. Whenever the concentration of intracellular tryptophan drops, the repressor falls off the DNA, allowing the polymerase to again transcribe the operon. The promoter contains two key blocks of DNA sequence information, the –35 and –10 regions, highlighted in *yellow*, which are recognized by RNA polymerase (see Figure 7–10). The complete operon is shown in Figure 8–6.

structure so that the protein can bind to the operator sequence. When the concentration of free tryptophan in the bacterium drops, the repressor no longer binds to DNA, and the tryptophan operon is transcribed. The repressor is thus a simple device that switches production of a set of biosynthetic enzymes on and off according to the availability of tryptophan—a form of feedback inhibition (see Figure 4–42).

The tryptophan repressor protein itself is always present in the cell. The gene that encodes it is continuously transcribed at a low level, so that a small amount of the repressor protein is always being made. Thus the bacterium can respond very rapidly to increases and decreases in tryptophan concentration.

Repressors Turn Genes Off and Activators Turn Them On

The tryptophan repressor, as its name suggests, is a **transcriptional repressor** protein: in its active form, it switches genes off, or *represses* them. Some bacterial transcription regulators do the opposite: they switch genes on, or *activate* them. These **transcriptional activator** proteins work on promoters that—in contrast to the promoter for the tryptophan operon—are only marginally able to bind and position RNA polymerase on their own. These inefficient promoters can be made fully functional by activator proteins that bind to a nearby regulatory sequence and make contact with the RNA polymerase, helping it to initiate transcription (**Figure 8–8**).

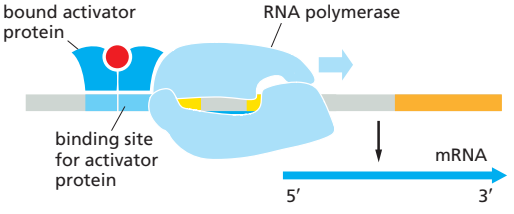

Figure 8–8 Genes can be switched on by activator proteins. An activator protein binds to a regulatory sequence on the DNA and then interacts with the RNA polymerase to help it initiate transcription. Without the activator, the promoter fails to initiate transcription efficiently. In bacteria, the binding of the activator to DNA is often controlled by the interaction of a metabolite or other small molecule (*red* circle) with the activator protein.

Like the tryptophan repressor, activator proteins often have to interact with a second molecule to be able to bind DNA. For example, the bacterial activator protein *CAP* has to bind cyclic AMP (cAMP) before it can bind to DNA (see Figure 4–20). Genes activated by CAP are switched on in response to an increase in intracellular cAMP concentration, which occurs when glucose, the bacterium's preferred carbon source, is no longer available; as a result, CAP drives the production of enzymes that allow the bacterium to digest other sugars.

The Lac Operon Is Controlled by an Activator and a Repressor

In many instances, the activity of a single promoter is controlled by two different transcription regulators. The *Lac operon* in *E. coli*, for example, is controlled by both the *Lac repressor* and the CAP activator that we just discussed. The *Lac* operon encodes proteins required to import and digest the disaccharide lactose. In the absence of glucose, the bacterium makes cAMP, which activates CAP to switch on genes that allow the cell to utilize alternative sources of carbon—including lactose. It would be wasteful, however, for CAP to induce expression of the *Lac* operon if lactose itself were not present. Thus the *Lac* repressor shuts off the operon in the absence of lactose. This arrangement enables the control region of the *Lac* operon to integrate two different signals, so that the operon is highly expressed only when two conditions are met: glucose must be absent and lactose must be present (**Figure 8–9**). This circuit thus behaves much like a switch that carries out a logic operation in a computer. When lactose is present AND glucose is absent, the cell executes the appropriate program—in this case, transcription of the genes that permit the uptake and utilization of lactose. None of the other combinations of conditions produce this result.

The elegant logic of the *Lac* operon first attracted the attention of biologists more than 50 years ago. The molecular basis of the switch in *E. coli* was uncovered by a combination of genetics and biochemistry, providing the first insight into how transcription is controlled. In a eukaryotic

QUESTION 8–1

Bacterial cells can take up the amino acid tryptophan (Trp) from their surroundings or, if there is an insufficient external supply, they can synthesize tryptophan from other small molecules. The *Trp* repressor is a transcription regulator that shuts off the transcription of genes that code for the enzymes required for the synthesis of tryptophan (see Figure 8–7).

A. What would happen to the regulation of the tryptophan operon in cells that express a mutant form of the tryptophan repressor that (1) cannot bind to DNA, (2) cannot bind tryptophan, or (3) binds to DNA even in the absence of tryptophan?

B. What would happen in scenarios (1), (2), and (3) if the cells, in addition, produced normal tryptophan repressor protein from a second, normal gene?

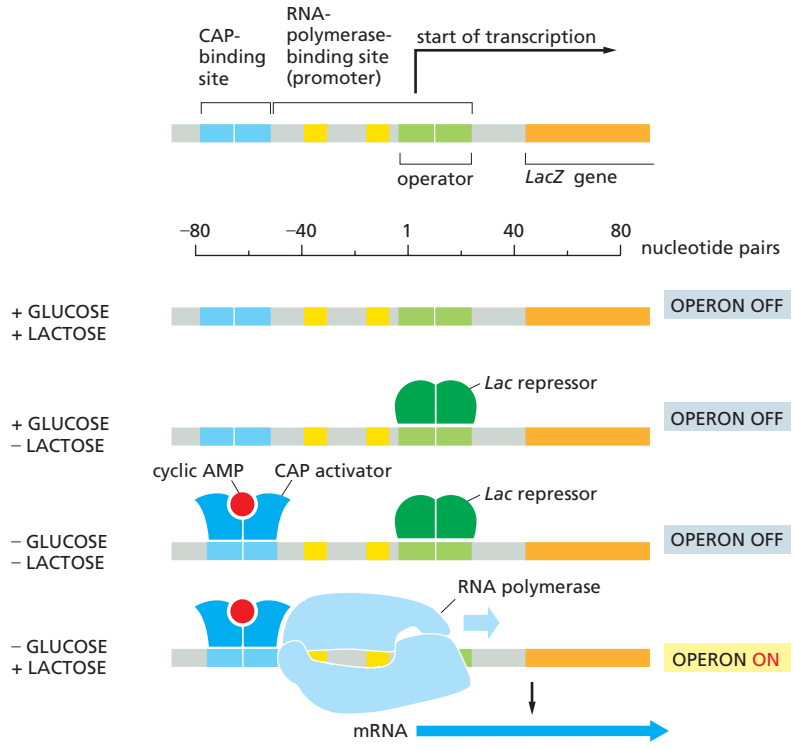

Figure 8–9 The *Lac* operon is controlled by two transcription regulators, the *Lac* repressor and CAP. When lactose is absent, the *Lac* repressor binds to the *Lac* operator and shuts off expression of the operon. Addition of lactose increases the intracellular concentration of a related compound, allolactose; allolactose binds to the *Lac* repressor, causing it to undergo a conformational change that releases its grip on the operator DNA (not shown). When glucose is absent, cyclic AMP (*red* circle) is produced by the cell, and CAP binds to DNA. For the operon to be transcribed, glucose must be absent (allowing the CAP activator to bind) and lactose must be present (releasing the *Lac* repressor). *LacZ*, the first gene of the operon, encodes the enzyme β-galactosidase, which breaks down lactose to galactose and glucose (**Movie 8.4**).

cell, similar transcription regulatory devices are combined to generate increasingly complex circuits, including those that enable a fertilized egg to form the tissues and organs of a multicellular organism.

Eukaryotic Transcription Regulators Control Gene Expression from a Distance

Eukaryotes, too, use transcription regulators—both activators and repressors—to regulate the expression of their genes. The DNA sites to which eukaryotic gene activators bind are termed *enhancers*, because their presence dramatically enhances the rate of transcription. However, biologists discovered that eukaryotic activator proteins could enhance transcription even when they are bound thousands of nucleotide pairs upstream—or downstream—of the gene's promoter. These observations raised several questions. How do enhancer sequences and the proteins bound to them function over such long distances? How do they communicate with a gene's promoter?

Many models for this "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, bringing the activator protein into close proximity with the promoter (**Figure 8–10**). The DNA thus acts as a tether, allowing a protein that is bound to an enhancer—even one that is thousands of nucleotide pairs away—to interact with the proteins in the vicinity of the promoter (see Figure 7–12). Often, additional proteins serve as adaptors to close the loop; the most important of these is a large complex of proteins known as *Mediator*. Together, all of these proteins ultimately attract and position the general transcription factors and RNA polymerase at the promoter, forming a *transcription initiation complex* (see Figure 8–10). Eukaryotic repressor proteins do the opposite: they decrease transcription by preventing the assembly of this complex.

Eukaryotic Transcription Regulators Help Initiate Transcription by Recruiting Chromatin-Modifying Proteins

In a eukaryotic cell, the proteins that guide the formation of the transcription initiation complex must also deal with the problem of DNA packaging. As discussed in Chapter 5, eukaryotic DNA is wound around clusters of histone proteins to form nucleosomes, which, in turn, are

QUESTION 8–2

Explain how DNA-binding proteins can make sequence-specific contacts to a double-stranded DNA molecule without breaking the hydrogen bonds that hold the bases together. Indicate how, through such contacts, a protein can distinguish a T-A from a C-G pair. Indicate the parts of the nucleotide base pairs that could form noncovalent interactions—hydrogen bonds, electrostatic attractions, or hydrophobic interactions (see Panel 2–3, pp. 70–71)—with a DNA-binding protein. The structures of all the base pairs in DNA are given in Figure 5–4.

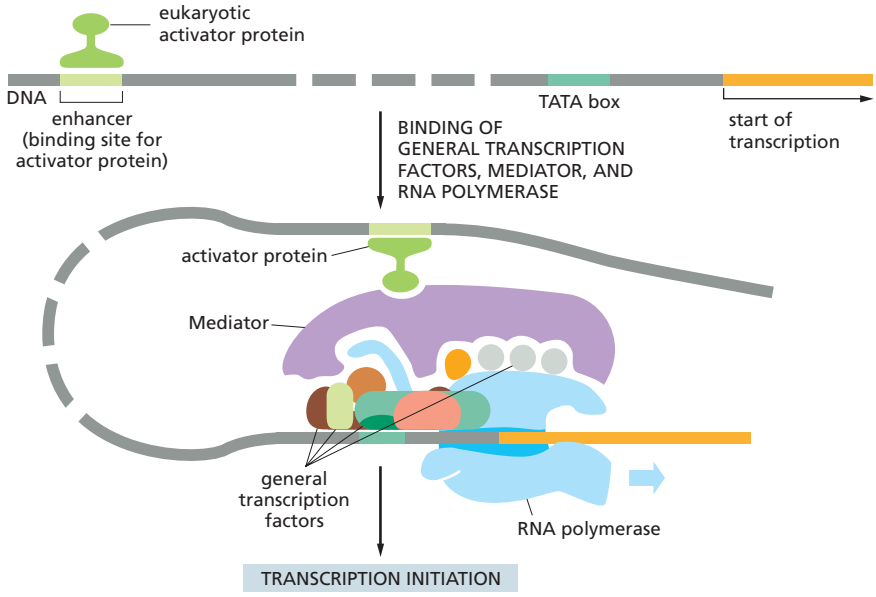

Figure 8–10 In eukaryotes, gene activation can occur at a distance. An activator protein bound to a distant enhancer attracts RNA polymerase and the general transcription factors to the promoter. Looping of the intervening DNA permits contact between the activator and the transcription initiation complex bound to the promoter. In the case shown here, a large protein complex called Mediator serves as a go-between. The broken stretch of DNA signifies that the segment of DNA between the enhancer and the start of transcription varies in length, sometimes reaching tens of thousands of nucleotide pairs. The TATA box is a DNA recognition sequence for the first general transcription factor that binds to the promoter (see Figure 7–12). Some eukaryotic activator proteins bind to DNA as dimers, but others bind DNA as monomers, as shown.

folded into higher-order structures. How do transcription regulators, general transcription factors, and RNA polymerase gain access to the underlying DNA? Although some of these proteins can bind efficiently to DNA that is wrapped up in nucleosomes, others are thwarted by these compact structures. More critically, nucleosomes that are positioned over a promoter can inhibit the initiation of transcription by physically blocking the assembly of the general transcription factors and RNA polymerase on the promoter. Such packaging may have evolved in part to prevent leaky gene expression by blocking the initiation of transcription in the absence of the proper activator proteins.

In eukaryotic cells, activator and repressor proteins can exploit the mechanisms used to package DNA to help turn genes on and off. As we saw in Chapter 5, chromatin structure can be altered by *chromatin-remodeling complexes* and by enzymes that covalently modify the histone proteins that form the core of the nucleosome (see Figures 5–24 and 5–25). Many gene activators take advantage of these mechanisms by attracting such chromatin-modifying proteins to promoters. For example, the recruitment of *histone acetyltransferases* promotes the attachment of acetyl groups to selected lysines in the tail of histone proteins; these acetyl groups themselves attract proteins that promote transcription, including some of the general transcription factors (**Figure 8–11**). And the recruitment of chromatin-remodeling complexes makes nearby DNA more accessible. These actions enhance the efficiency of transcription initiation.

In a similar way, gene repressor proteins can modify chromatin in ways that reduce the efficiency of transcription initiation. For example, many repressors attract *histone deacetylases*—enzymes that remove the acetyl groups from histone tails, thereby reversing the positive effects that acetylation has on transcription initiation. Although some eukaryotic repressor proteins work on a gene-by-gene basis, others can orchestrate the formation of large swathes of transcriptionally inactive chromatin. As discussed in Chapter 5, these transcription-resistant regions of DNA include the heterochromatin found in interphase chromosomes and the inactive X chromosome in the cells of female mammals.

QUESTION 8–3

Some transcription regulators bind to DNA and cause the double helix to bend at a sharp angle. Such "bending proteins" can stimulate the initiation of transcription without contacting either the RNA polymerase, any of the general transcription factors, or any other transcription regulators. Can you devise a plausible explanation for how these proteins might work to modulate transcription? Draw a diagram that illustrates your explanation.

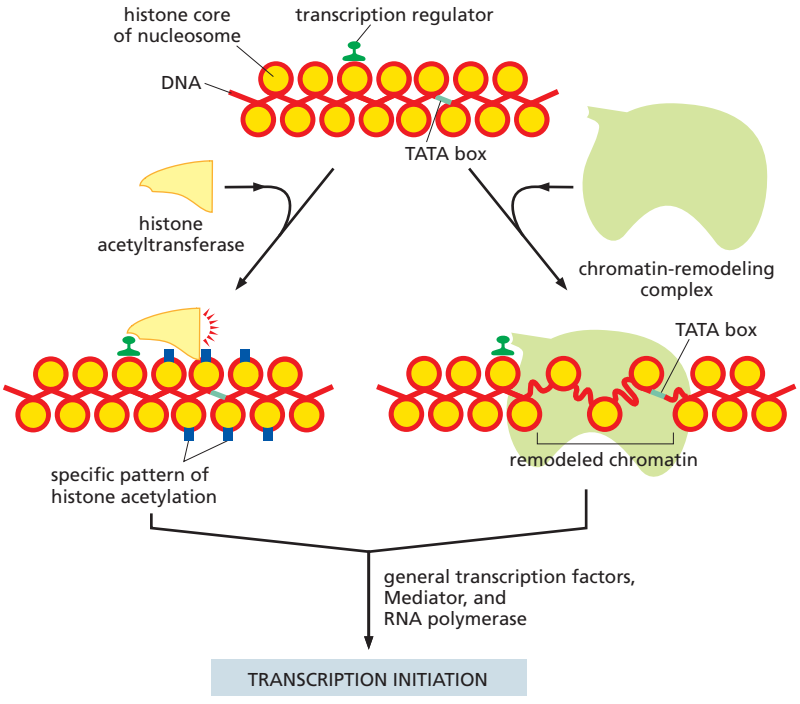

Figure 8–11 Eukaryotic transcriptional activators can recruit chromatin-modifying proteins to help initiate gene transcription. On the left, the recruitment of histone-modifying enzymes such as histone acetyltransferases adds acetyl groups to specific histones, which can then serve as binding sites for proteins that stimulate transcription initiation (not shown). On the right, chromatin-remodeling complexes render the DNA packaged in nucleosomes more accessible to other proteins in the cell, including those required for transcription initiation; notice, for example, the increased exposure of the TATA box.

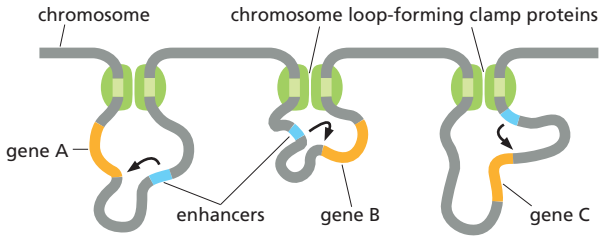

Figure 8–12 Animal and plant chromosomes are arranged in DNA loops. In this schematic diagram, specialized proteins (*green*) hold chromosomal DNA in loops, thereby favoring the association of each gene with its proper enhancer. The loops, sometimes called *topological associated domains* (TADs), range in size between thousands and millions of nucleotide pairs and are typically much larger than the loops that form between regulatory sequences and promoters (see Figure 8–10).

The Arrangement of Chromosomes into Looped Domains Keeps Enhancers in Check

We have seen that all genes have regulatory regions, which dictate at which times, under what conditions, and in what tissues the gene will be expressed. We have also seen that eukaryotic transcription regulators can act across very long stretches of DNA, with the intervening DNA looped out. What, then, prevents a transcripton regulator—bound to the control region of one gene—from looping in the wrong direction and inappropriately influencing the transcription of a neighboring gene?

To avoid such unwanted cross-talk, the chromosomal DNA of plants and animals is arranged in a series of loops that hold individual genes and their regulatory regions in rough proximity. This localization restricts the action of enhancers, preventing them from wandering across to adjacent genes. The chromosomal loops are formed by specialized proteins that bind to sequences that are then drawn together to form the base of the loop (**Figure 8–12**).

The importance of these loops is highlighted by the effects of mutations that prevent the loops from properly forming. Such mutations, which lead to genes being expressed at the wrong time and place, are found in numerous cancers and inherited diseases.

GENERATING SPECIALIZED CELL TYPES

All cells must be able to turn genes on and off in response to signals in their environment. But the cells of multicellular organisms have taken this type of transcriptional control to an extreme, using it in highly specialized ways to form organized arrays of differentiated cell types. Such decisions present a special challenge: once a cell in a multicellular organism becomes committed to differentiate into a specific cell type, the choice of fate is generally maintained through subsequent cell divisions. This means that the changes in gene expression, which are often triggered by a transient signal, must be remembered by the cell. Such *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 eukaryotes and bacteria are often only transient; the tryptophan repressor, for example, switches off the tryptophan operon in bacteria only in the presence of tryptophan; as soon as the amino acid is removed from the medium, the genes switch back on, and the descendants of the cell will have no memory that their ancestors had been exposed to tryptophan.

In this section, we discuss some of the special features of transcriptional regulation that allow multicellular organisms to create and maintain specialized cell types. These cell types ultimately produce the tissues and organs that give worms, flies, and even humans their distinctive characteristics.