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. We begin with an overview of some basic principles of gene control in multicellular organisms.

In This Chapter

AN OVERVIEW OF GENE CONTROL 411

DNA-BINDING MOTIFS IN GENE REGULATORY PROTEINS 416

HOW GENETIC SWITCHES WORK 432

THE MOLECULAR GENETIC MECHANISMS THAT CREATE SPECIALIZED CELL TYPES 454

POST-TRANSCRIPTIONAL CONTROLS 477

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. Evidence that they generally do this without altering the sequence of their DNA 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 7–2**A). 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 demonstrated in mammals, including sheep, cattle, pigs, goats, dogs, 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).

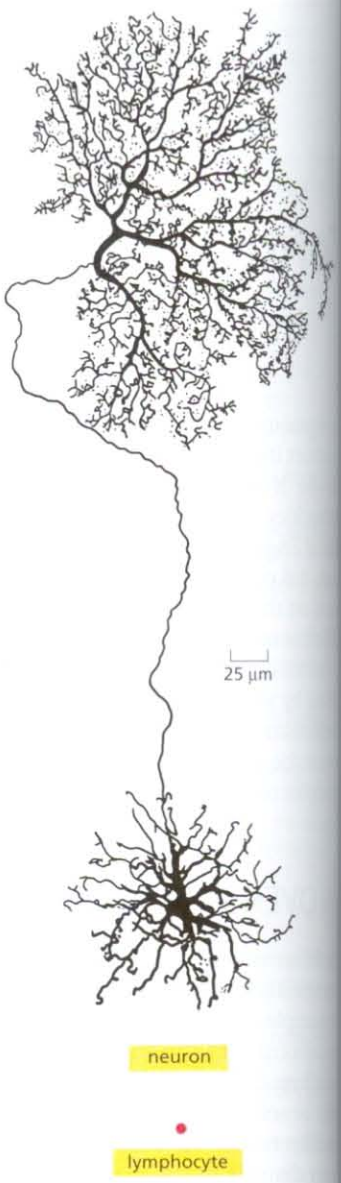


Figure 7–1 A mammalian neuron and a lymphocyte. The long branches of this neuron from the retina enable it to receive electrical signals from many cells and carry those signals to many neighboring cells. The lymphocyte is a white blood cell involved in the immune response to infection and moves freely 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.)

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 differentiated cells in the human body appear to be identical. Moreover, comparisons of the genomes of different cells based on recombinant DNA technology have confirmed, as a general rule, that the changes in gene expression that underlie the development of multicellular organisms do not rely on 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, which we discuss in Chapter 25.

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 have a detailed 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 30–60% of its approximately 25,000 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. Even genes that are expressed in all cell types vary in their level of expression from one cell type to the next. 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. For example, 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 methods that directly display the levels of proteins and their post-translational modifications (**Figure 7–4**).

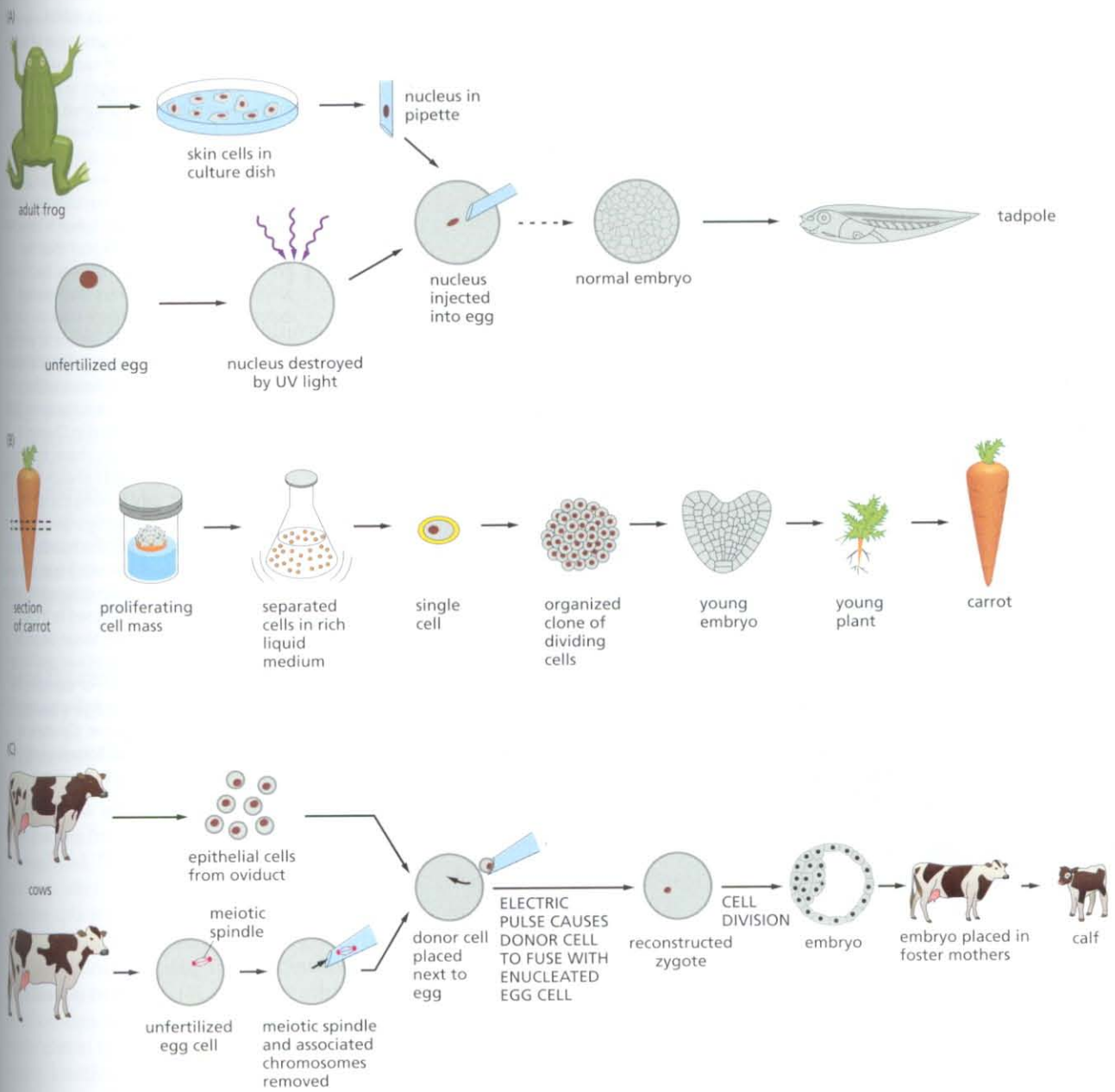


Figure 7–2 Evidence that a differentiated cell contains all the genetic instructions necessary to direct the formation of a complete organism. (A) The nucleus of a skin cell from an adult frog transplanted into an enucleated egg can give rise to an entire tadpole. The *broken arrow* indicates that, to 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 nucleus 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.* 219:24–35, 1968. With permission from Scientific American.)

External Signals Can Cause a Cell to Change the Expression of Its Genes

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

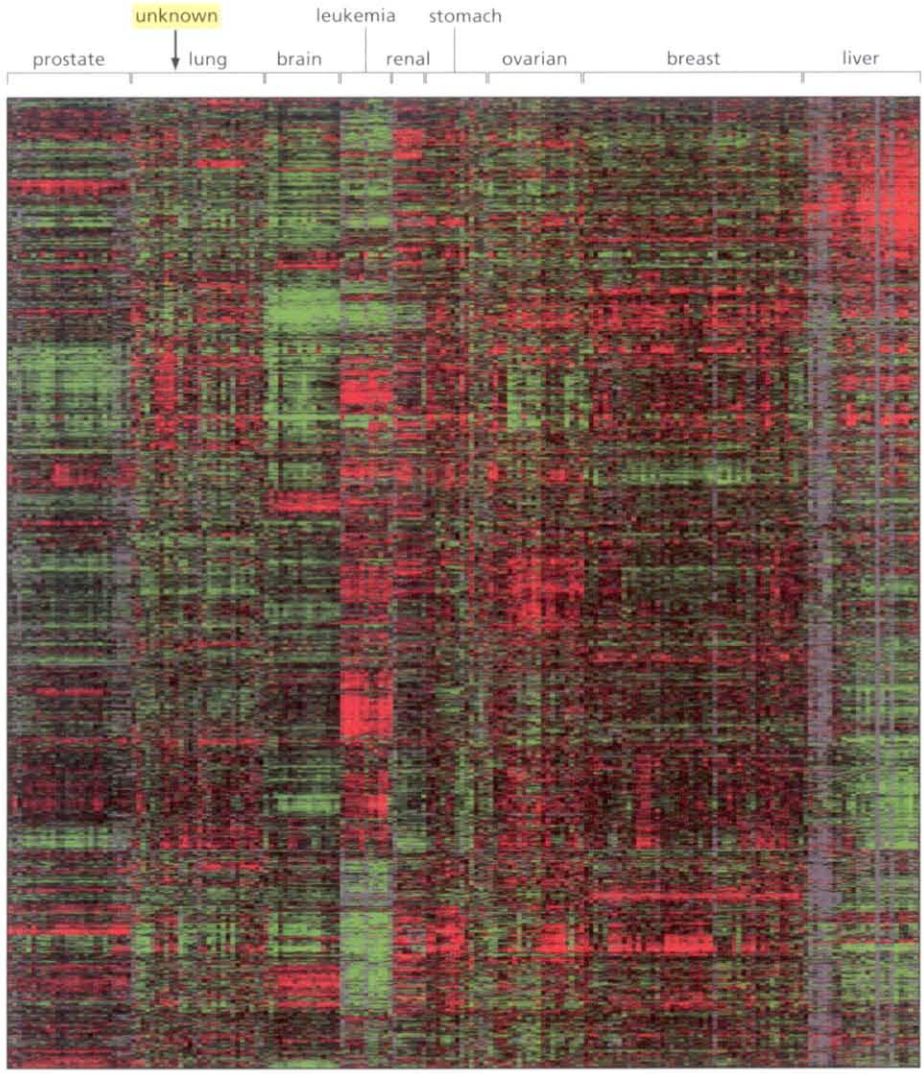


Figure 7–3 Differences in mRNA expression patterns among different types of human cancer cells. This figure summarizes a very large set of measurements in which the mRNA levels of 1800 selected genes (arranged *top* to *bottom*) were determined for 142 different human tumor cell lines (arranged *left* to *right*), each from a different patient. Each small *red* bar indicates that a 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 (pp. 574–575). 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 from *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.)

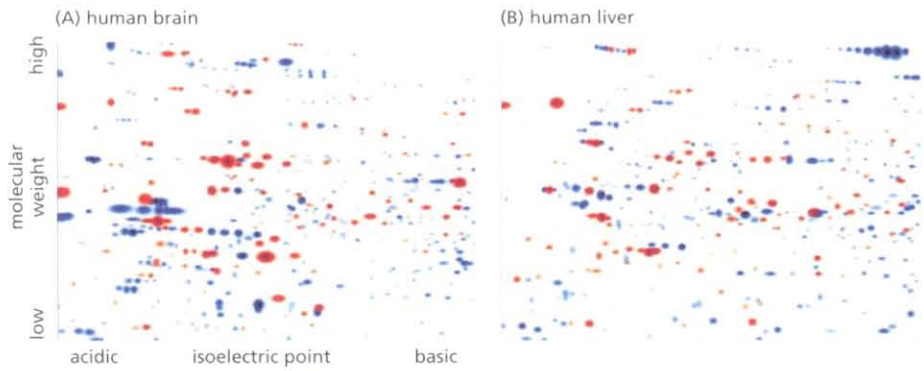


Figure 7–4 Differences in the proteins expressed by two human tissues. In each panel, the proteins are displayed using two-dimensional polyacrylamide-gel electrophoresis (see pp. 521–522). The proteins have been separated by molecular weight (*top* to *bottom*) and isoelectric point, the pH at which 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 abundances are usually different. Note that this technique separates proteins by both 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. Although two-dimensional gel electrophoresis provides a simple way to visualize the differences between two protein samples, methods based on mass spectrometry (see pp. 519–521) provide much more detailed information and are therefore more commonly used. (Courtesy of Tim Myers and Leigh Anderson, Large Scale Biology Corporation.)

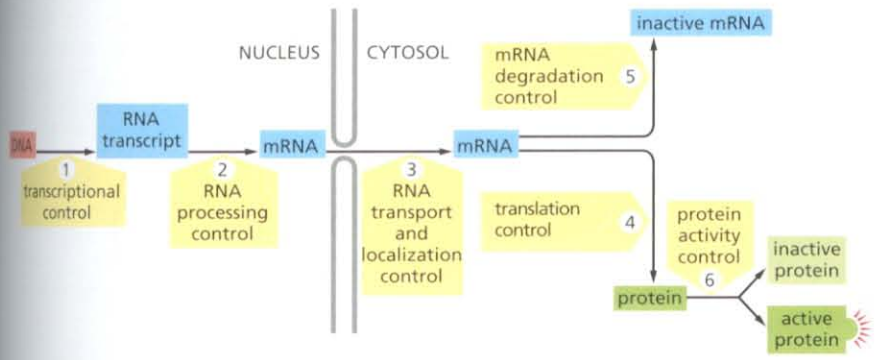


Figure 7–5 Six steps at which eucaryotic gene expression can be controlled. Controls that operate at steps 1 through 5 are discussed in this chapter. Step 6, the regulation of protein activity, occurs largely through covalent post-translational modifications including phosphorylation, acetylation, and ubiquitylation (see Table 3–3, p. 186) and is discussed in many chapters throughout the book.

production of glucose from amino acids and other small molecules; 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. Fat cells, for example, reduce the production of tyrosine aminotransferase, 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 differently 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.

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

What 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 previous chapter, there are many steps in the pathway leading from DNA to protein. We now know that 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 the splicing and processing of RNA transcripts (**RNA processing control**), (3) selecting which completed mRNAs are exported from the nucleus 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 locating 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 many 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 outlined in Chapter 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 thrown by a single signal. Many others are complex and resemble tiny microprocessors, responding to a variety of signals that they interpret and integrate in order to switch their neighboring gene on or off. Whether complex or simple, these switching devices are found in all cells and are composed of two types of fundamental components: (1) short stretches of DNA of defined sequence and (2) *gene regulatory proteins* that recognize and bind to this DNA.

We begin our discussion of gene regulatory proteins by describing how they 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** (often loosely called "transcription factors") 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 5–78). 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 and methods for studying protein–DNA interactions discussed later in this chapter.

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

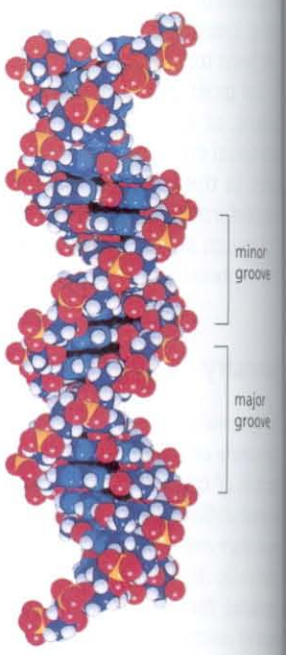


Figure 7–6 Double-helical structure of DNA. A space-filling model of DNA showing the major and minor grooves on the outside of the double helix. The atoms are colored as follows: carbon, *dark blue;* nitrogen, *light blue;* hydrogen, *white;* oxygen, *red;* phosphorus, *yellow.*

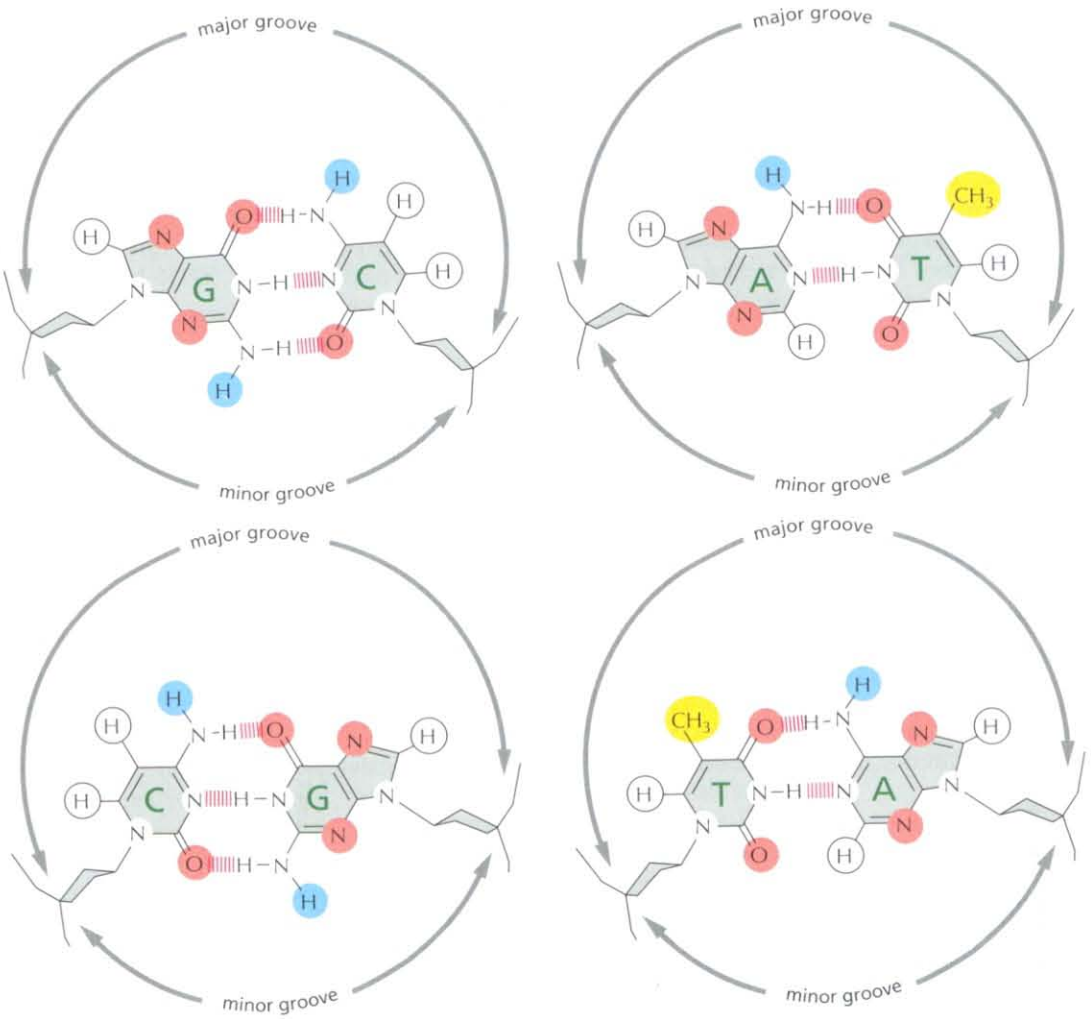
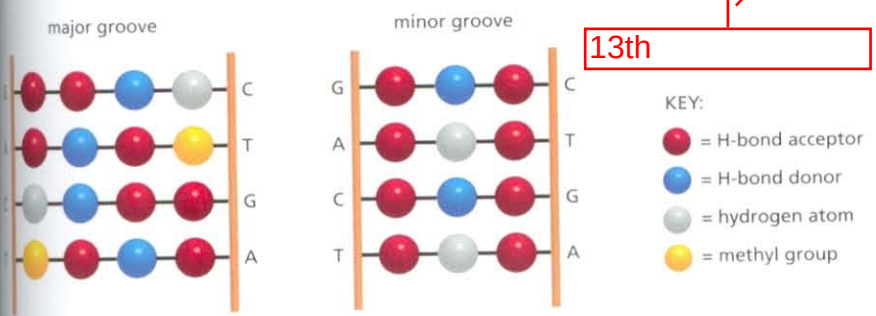


Figure 7–7 How the different base pairs in DNA can be recognized from their edges without the need to open the double helix. The four possible configurations of base pairs are shown, with potential hydrogen bond donors indicated in *blue,* potential hydrogen bond acceptors in *red,* and hydrogen bonds 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 ed. New York: Garland Publishing, 1999.)

equence 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 that 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 make specific contacts with the major groove—as we shall see.



13th

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 ed. New York: Garland Publishing, 1999.)

Short DNA Sequences Are Fundamental Components of Genetic Switches

A specific nucleotide sequence can be "read" as a pattern of molecular 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.

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

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

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

	NAME	DNA SEQUENCE RECOGNIZED*
Bacteria	Lac repressor	5′ AATTGTGAGCGGATAACAATT 3′ TTAACACTCGCCTATTGTTAA
	CAP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
	Lambda repressor	TATCACCGCCAGAGGT ATAGTGGCGGTCTCCAT
Yeast	Gal4	CGGAGGACTGTCCTCCG GCCTCCTGACAGGAGGC
	Matα2	CATGTAATT GTACATTAA
	Gcn4	ATGACTCAT TACTGAGTA
Drosophila	Kruppel	AACGGGTTAA TTGCCCAATT
	Bicoid	GGGATTAGA CCCTAATCT
Mammals	Sp1	GGGCGG CCCGCC
	Oct1 Pou domain	ATGCAAAT TACGTTTA
	GATA1	TGATAG ACTATC
	MyoD	CAAATG GTTTAC
	p53	GGGCAAGTCT CCCGTTCAGA

*For convenience, only one recognition sequence, rather than a consensus sequence (see Figure 6–12), is given for each protein.

complementary to the special surface features of the double helix in that region. In most cases the protein makes a series of contacts with the DNA, involving hydrogen bonds, ionic bonds, and hydrophobic interactions. Although each individual contact is weak, the 20 or so that are typically formed at the protein–DNA interface add together to ensure that the interaction is both highly specific and very strong (**Figure 7–9**). 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 them contain one or another of a small set of DNA-binding structural motifs. These motifs generally use 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 many 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 7–10**). 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.

Outside the helix–turn–helix region, the structure of the various proteins that contain this motif can vary enormously (**Figure 7–11**). 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.

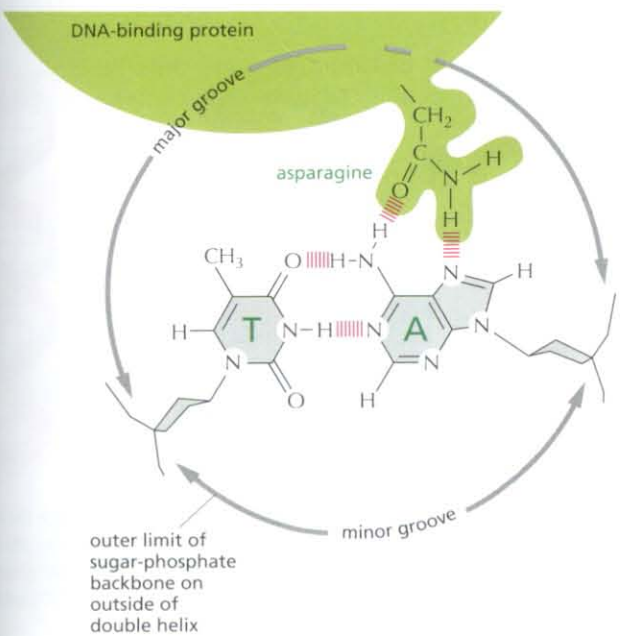


Figure 7–9 The binding of a gene regulatory protein to the major groove of DNA. Only a single contact is shown. Typically, the protein–DNA interface would consist of 10–20 such contacts, involving different amino acids, each contributing to the strength of the protein–DNA interaction.