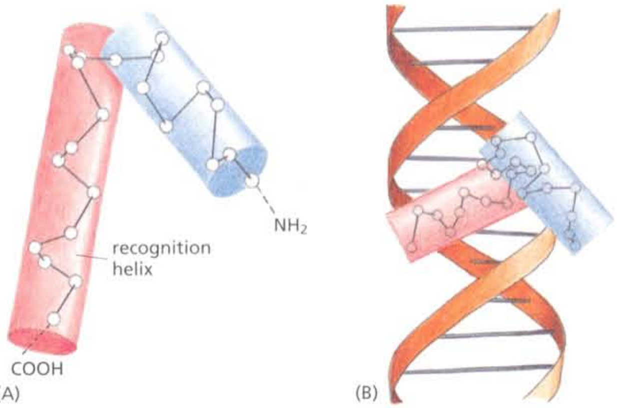


**Figure 7–10 The DNA-binding helix–turn–helix motif.** The motif is shown in (A), where each *white circle* denotes the central carbon of an amino acid. The C-terminal α helix *(red)* is called the recognition helix because it participates in sequence-specific recognition of DNA. As shown in (B), this 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.

The group of helix–turn–helix proteins shown in Figure 7–11 demonstrates a common feature of 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 7–12**). 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 22, they have since proved to have a fundamental role in the development of higher animals as well. Mutations in these genes can 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 code for 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

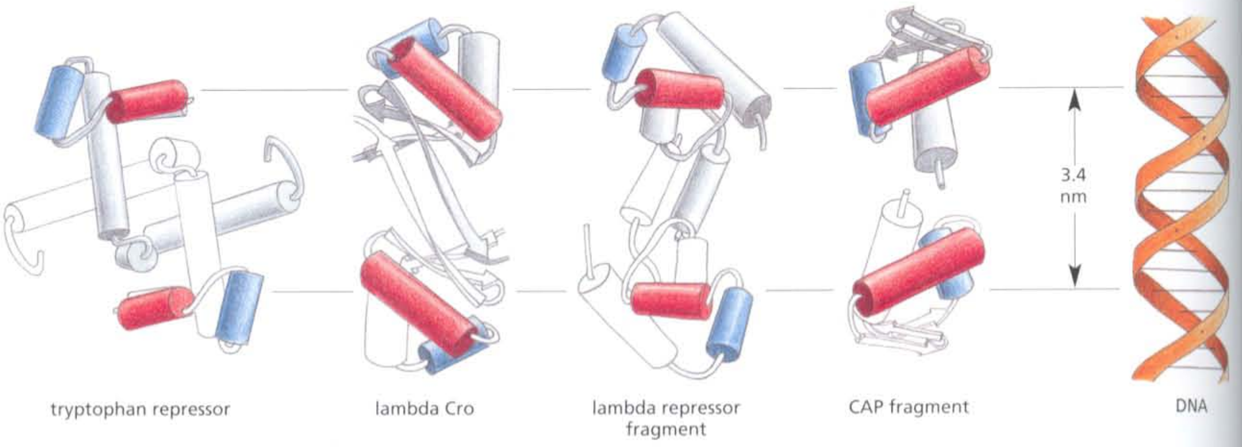


**Figure 7–11 Some helix–turn–helix DNA-binding proteins.** All of the proteins bind DNA as dimers in which the two copies of the recognition helix *(red cylinder)* are separated by exactly one turn of the DNA helix (3.4 nm). The other helix of the helix–turn–helix motif is colored *blue*, as in Figure 7–10. 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.

the bacter
the princi
organisms
ered in *Dr*
tually all
humans.
The s
shown in
ulatory p
helix–tur
structure
is always
shown th
tein have
same ma
(see Figu

There A

The heli
tant grou
compon
**zinc fing**
ings dat
studies
which w
activate
structur
7–14B).
helix of
uous st
DNA–p
(**Figure**
And
recepto

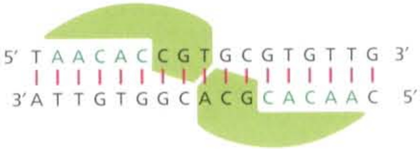

5′ T A A C A C C G T G C G T G T T G 3′
3′ A T T G T G G C A C G C A C A A C 5′

**Figure 7–12 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 7–11 for the actual structure of the protein.

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 7–13**. 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 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 3–13).

## 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 includes 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 7–14**A). Subsequent structural studies have shown that they fall into several distinct structural groups, two of which we consider here. The first type was initially discovered in the protein that activates the transcription of a eucaryotic ribosomal RNA gene. It has a simple structure, in which the zinc holds an α helix and a β sheet together (Figure 7–14B). This type of zinc finger is often found in tandem clusters 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 7–15).

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

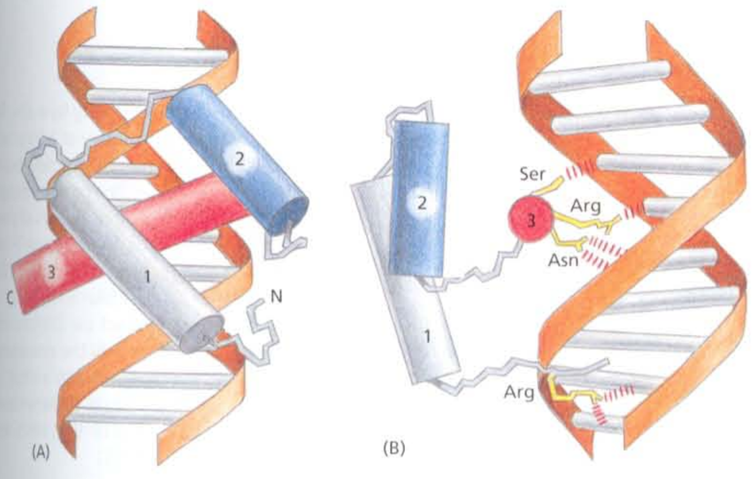


**Figure 7–13 A homeodomain bound to its specific DNA sequence.** Two different views of the same structure are shown. (A) The homeodomain is folded into three α helices, which are packed tightly together by hydrophobic interactions. The part containing helices 2 and 3 closely resembles the helix–turn–helix motif. (B) The recognition helix (helix 3, *red*) forms important contacts with the major groove of DNA. The asparagine (Asn) of helix 3, for example, contacts an adenine, as shown in Figure 7–9. A flexible arm attached to helix 1 forms contacts with nucleotide pairs in the minor groove. The homeodomain shown here is from a yeast gene regulatory protein, but it closely resembles homeodomains from many eucaryotic organisms. <ACGT> (Adapted from C. Wolberger et al., *Cell* 67:517–528, 1991. With permission from Elsevier.)

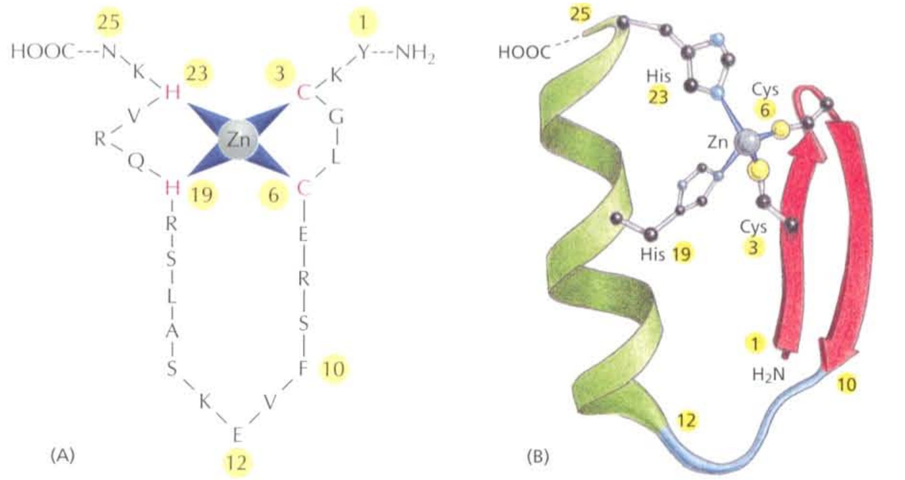


**Figure 7–14 One type of zinc finger protein.** This protein belongs to the Cys–Cys–His–His family of zinc finger proteins, named after the amino acids that grasp the zinc. (A) Schematic drawing of the amino acid sequence of a zinc finger from a frog protein of this class. (B) The three-dimensional structure of this same 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. With permission from AAAS.)

structure (similar in some respects to the helix–turn–helix motif) in which two α helices are packed together with zinc atoms (**Figure 7–16**). 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. 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.

14th

## 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 large group of gene regulatory proteins, however, has evolved an entirely different recognition strategy. In this case, a two-stranded β sheet, with amino acid side chains extending from the sheet toward the DNA, reads the information on the surface of the major groove (**Figure 7–17**). As in the case of a 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.

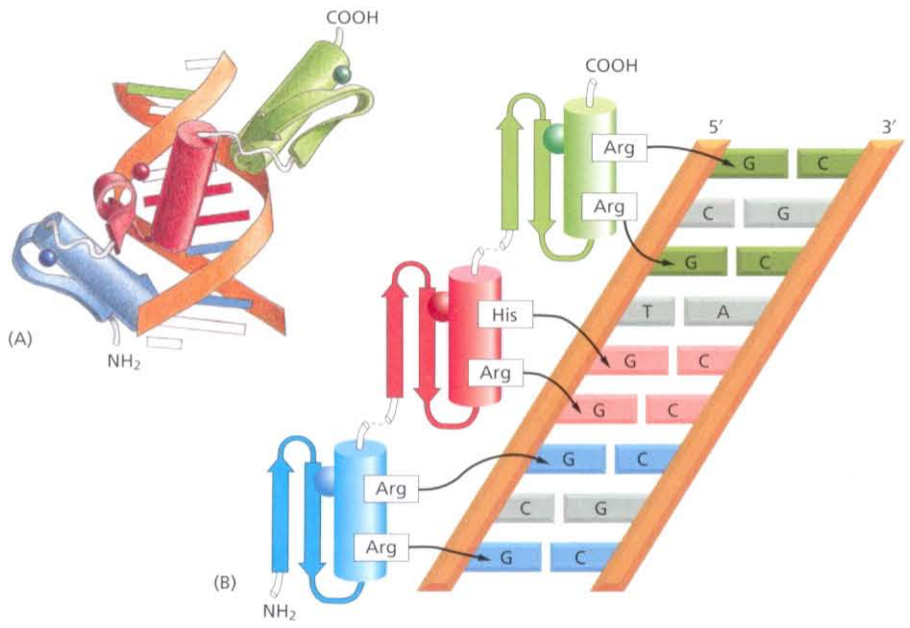


**Figure 7–15 DNA binding by a zinc finger protein.** (A) The structure of a fragment of a mouse gene regulatory protein bound to a specific DNA site. This protein recognizes DNA by using three zinc fingers of the Cys–Cys–His–His type (see Figure 7-14) arranged as direct repeats. <ATCT> (B) The three 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. With permission from AAAS.)

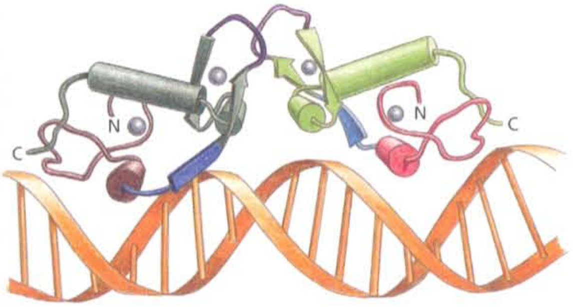


Figure 7–16 **A dimer of the zinc finger domain of the intracellular receptor family bound to its specific DNA sequence.** Each zinc finger domain contains two atoms of Zn (indicated by the small *gray spheres*); one stabilizes the DNA recognition helix (shown in *brown* in one subunit and *red* in the 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 7–11, 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. With permission from Macmillan Publishers Ltd.)

## Some Proteins Use Loops That Enter the Major and Minor Grooves to Recognize DNA

A few DNA-binding proteins use protruding peptide loops to read nucleotide sequences, rather than α helices and β sheets. For example, p53, a critical *tumor suppressor* in humans, recognizes nucleotide pairs from both the major and minor grooves using such loops (**Figure 7–18**). The normal role of the p53 protein is to tightly regulate cell growth and proliferation. Its importance can be appreciated by the fact that nearly half of all human cancers have acquired somatic mutations in the gene for p53; this step is key to the progression of many tumors, as we shall see in Chapter 20. Many of the p53 mutations observed in cancer cells destroy or alter its DNA-binding properties; indeed, Arg 248, which contacts the minor groove of DNA (see Figure 7–18) is the most frequently mutated p53 residue in human cancers.

## 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 7–12). Usually, the portion of the protein responsible for dimerization is distinct from the portion that is responsible for DNA binding. One motif, however, combines these two functions elegantly and economically. 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 3–9). 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 7–19**).

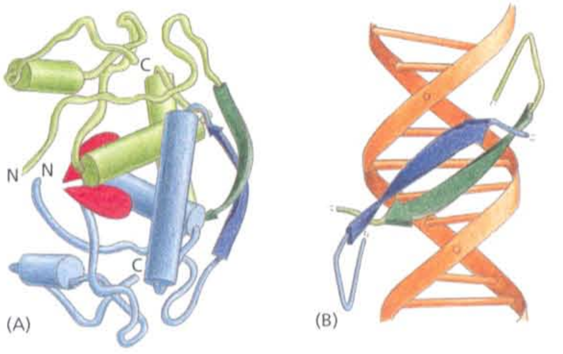


Figure 7–17 **The bacterial Met repressor protein.** The bacterial Met repressor regulates the genes encoding the enzymes that catalyze methionine synthesis. When this amino acid is abundant, it binds to the repressor, causing a change in the structure of the protein that enables it to bind to DNA tightly, shutting off the synthesis of the enzyme. (A) In order to bind to DNA tightly, the Met repressor must be complexed with *S*-adenosyl methionine, outlined 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, with permission from Elsevier; B, adapted from W. Somers and S. Phillips, *Nature* 359:387–393, 1992, with permission from Macmillan Publishers Ltd.)

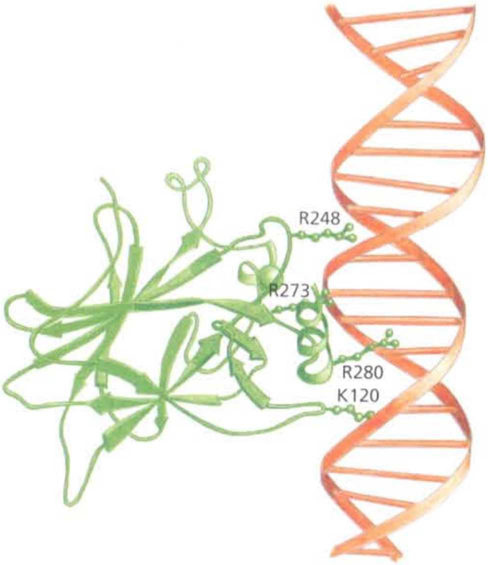


Figure 7–18 **DNA recognition by the p53 protein.** The most important DNA contacts are made by arginine 248 and lysine 120, which extend from the protruding loops entering the minor and major grooves. The folding of the p53 protein requires a zinc atom (shown as a sphere), but the way in which the zinc is grasped by the protein is completely different from that of the zinc finger proteins, described previously.

## Heterodimerization Expands the Repertoire of DNA Sequences That Gene Regulatory Proteins Can Recognize

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 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 in this way greatly expands the repertoire of DNA-binding specificities that these proteins can display. As illustrated in **Figure 7–20**, three distinct DNA-binding specificities could, in principle, be generated from two types of leucine zipper monomers, while six could be created from three types of monomers, and so on.

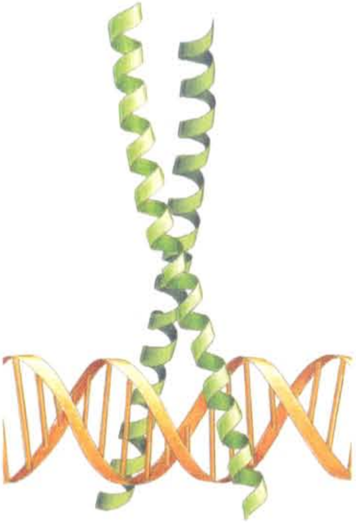


Figure 7–19 **A leucine zipper dimer bound to DNA.** Two α-helical DNA-binding domains *(bottom)* dimerize through their α-helical leucine zipper region *(top)* to form an inverted Y-shaped structure. Each arm of the Y is formed by a single α helix, one from each monomer, that mediates binding to a specific DNA sequence in the major groove of DNA. <TGTT> 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. With permission from Elsevier.)

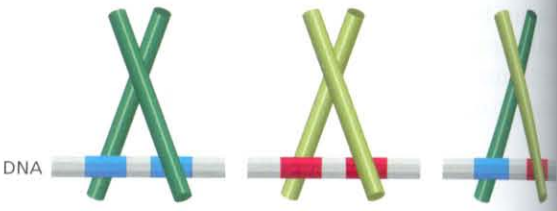


Figure 7–20 **Heterodimerization of leucine zipper proteins can alter their DNA-binding specificity.** Leucine zipper homodimers bind to symmetric DNA sequences, as shown in the left-hand and center drawings. These two proteins recognize different DNA sequences, as indicated by the *red* and *blue* regions 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.

There
types of leu
the amour
presumabl
erodimer
leucine zi
exact amir
protein in
proteins.
Heter
binations
process. H
expressio
**7–21**). Co
edly in th
plexes is c
trol gene
Certa
in the cel
rearrange
single po
**7–22**).

## The Hel
## Binding

Another
**helix–lo**
discusse
to a seco

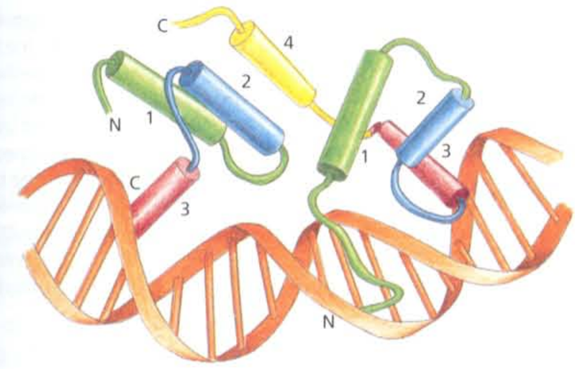


**Figure 7–21 A heterodimer composed of two homeodomain proteins bound to its DNA recognition site.** The *yellow* helix 4 of the protein on the right (Matα2) is unstructured in the absence of the protein on the left (Mata1), forming a helix only upon heterodimerization. The DNA sequence is recognized jointly by both proteins; some of the protein–DNA contacts made by Matα2 were shown in Figure 7–13. 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 7–13. (Adapted from T. Li et al., *Science* 270:262–269, 1995. With permission from AAAS.)

There are, however, limits to this promiscuity: for example, 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 presumably 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 cell process. Heterodimerization as a mechanism for combinatorial control of gene expression occurs in many different types of gene regulatory proteins (**Figure 7–21**). Combinatorial control is a major theme that we shall encounter repeatedly in this chapter, and the formation of heterodimeric gene regulatory complexes is only one of many ways in which proteins work in combinations to control gene expression.

Certain combinations of gene regulatory proteins have become "hardwired" in the cell; for example, two distinct DNA-binding domains can, through gene rearrangements occurring over evolutionary time scales, become joined into a single polypeptide chain that displays a novel DNA-binding specificity (**Figure 7–22**).

## 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 differs from the 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

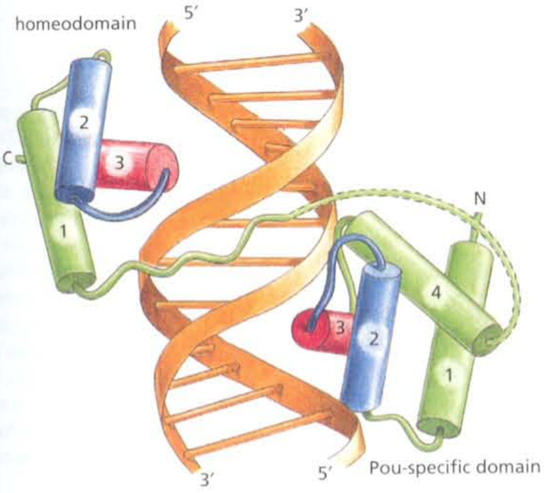


**Figure 7–22 Two DNA-binding domains covalently joined by a flexible polypeptide.** The structure shown (called a Pou-domain) consists of both a homeodomain and a helix–turn–helix structure joined by a flexible polypeptide "leash," indicated by the broken lines. A single gene encodes the entire protein, which 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 Oct1 protein. (Adapted from J.D. Klemm et al., *Cell* 77:21–32, 1994. With permission from Elsevier.)

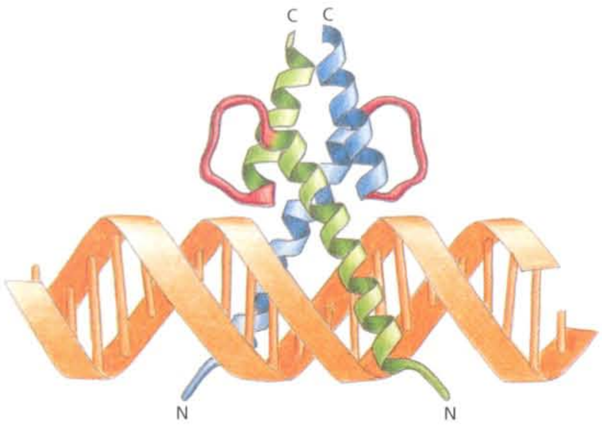


Figure 7–23 **A helix–loop–helix (HLH) dimer bound to DNA.** The two monomers are held together in a four-helix bundle: each monomer contributes two α helices connected by a flexible loop of protein *(red)*. A specific DNA sequence is bound by the two α helices that project from the four-helix bundle. (Adapted from A.R. Ferre-D'Amare et al., *Nature* 363:38–45, 1993. With permission from Macmillan Publishers Ltd.)

and pack against the other. As shown in **Figure 7–23**, this two-helix structure binds both to DNA and to the HLH motif of a second HLH protein. 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 cells with a widely used way to hold specific gene regulatory proteins in check (**Figure 7–24**).

## It Is Not Yet Possible to 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 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 is no, although certain types of amino acid-base interactions appear much more frequently than others (**Figure 7–25**). 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 patterns 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 7–26**). Nevertheless, molecular biologists are beginning to understand the principles of protein–DNA recognition well enough to design new proteins that will recognize a given DNA sequence.

Having outlined the general features of gene regulatory proteins, we turn to some of the methods that are now used to study them.

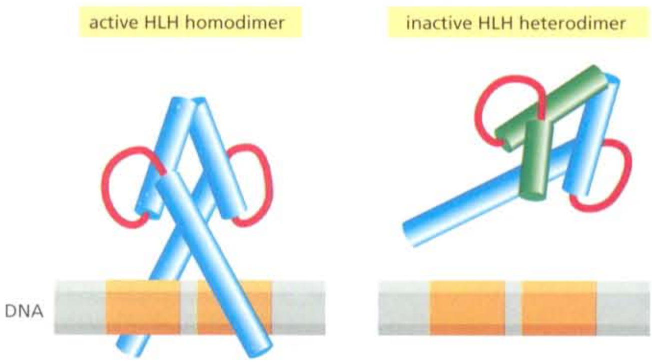


Figure 7–24 **Inhibitory regulation by truncated HLH proteins.** The HLH motif is responsible for both dimerization and DNA binding. On the *left*, an HLH homodimer recognizes a symmetric DNA sequence. On the *right*, the binding of a full-length HLH protein *(blue)* to a truncated HLH protein *(green)* 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.

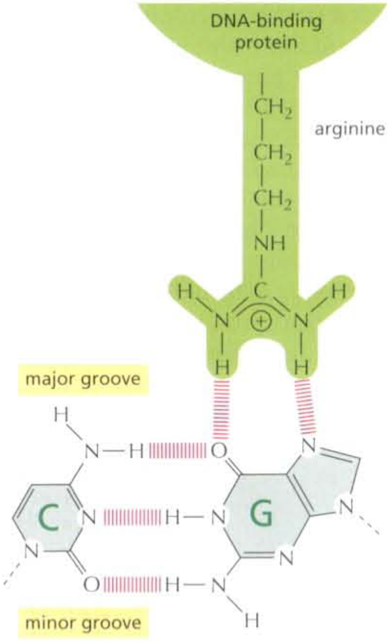
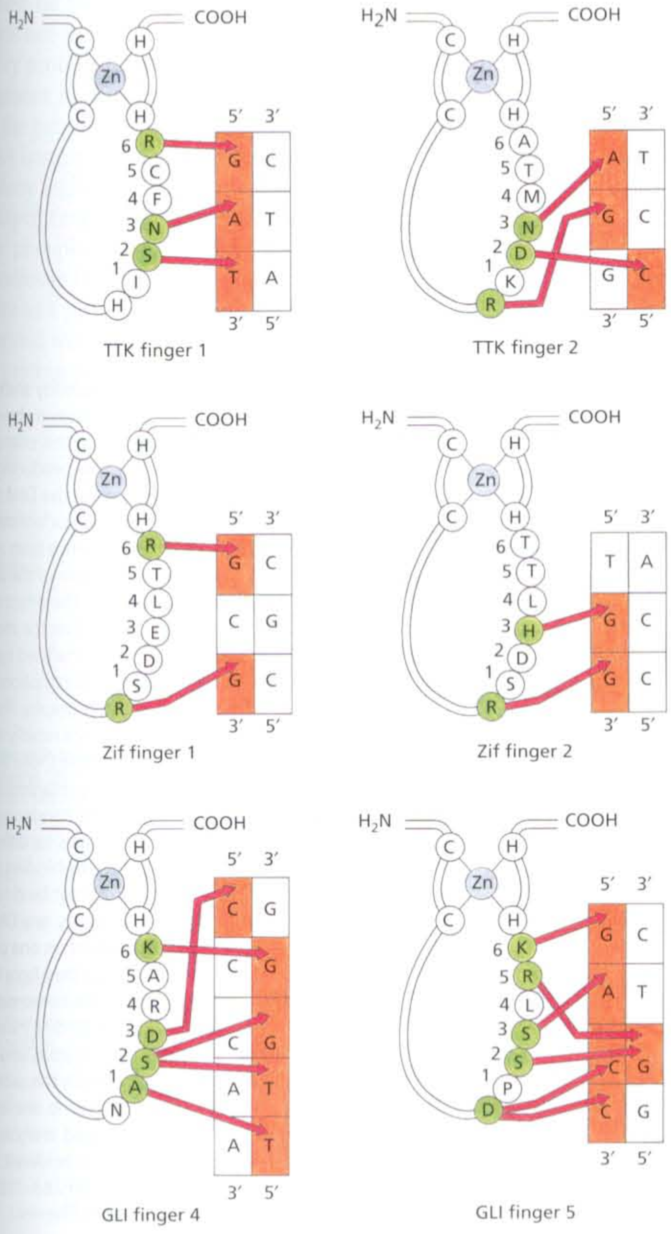


**Figure 7–25 One of the most common protein–DNA interactions.** Because of its specific geometry of hydrogen-bond acceptors (see Figure 7–7), the side chain of arginine unambiguously recognizes guanine. Figure 7–9 shows another common protein–DNA interaction.

## A Gel-Mobility Shift Assay Readily Detects Sequence-Specific DNA-Binding Proteins

Genetic analyses, which provided a route to the gene regulatory proteins of bacteria, yeast, and *Drosophila*, are 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. One of the most common ways to detect and study sequence-specific DNA-binding proteins is based on the effect of a bound protein on the migration of DNA molecules in an electric field.



TTK finger 1

TTK finger 2

Zif finger 1

Zif finger 2

GLI finger 4

GLI finger 5

**Figure 7–26 Summary of sequence-specific interactions between six different zinc fingers and their DNA recognition sequences.** Even though all six Zn fingers have the same overall structure (see Figure 7–14), each binds to a different DNA sequence. The numbered amino acids form the α helix that recognizes DNA (Figures 7–14 and 7–15), and those that make sequence-specific DNA contacts are *green*. Bases contacted by protein are *orange*. Although arginine–guanine contacts are common (see Figure 7–25), 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 7–15; and two are from a human protein (GL1) whose aberrant forms can cause certain types of cancers. (Adapted from C. Branden and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Publishing, 1999.)

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 (see p. 534), 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, as discussed in Chapter 8) 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 proteins bind, autoradiography (see pp. 602–603) 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 7–27**). Once a sequence-specific DNA protein has been purified, the gel-mobility shift assay can be used to study the strength and specificity of its interactions with different DNA sequences, the lifetime of DNA–protein complexes, and other properties critical to the functioning of the protein in the cell.

## DNA Affinity Chromatography Facilitates the Purification of Sequence-Specific DNA-Binding Proteins

A particularly powerful protein-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

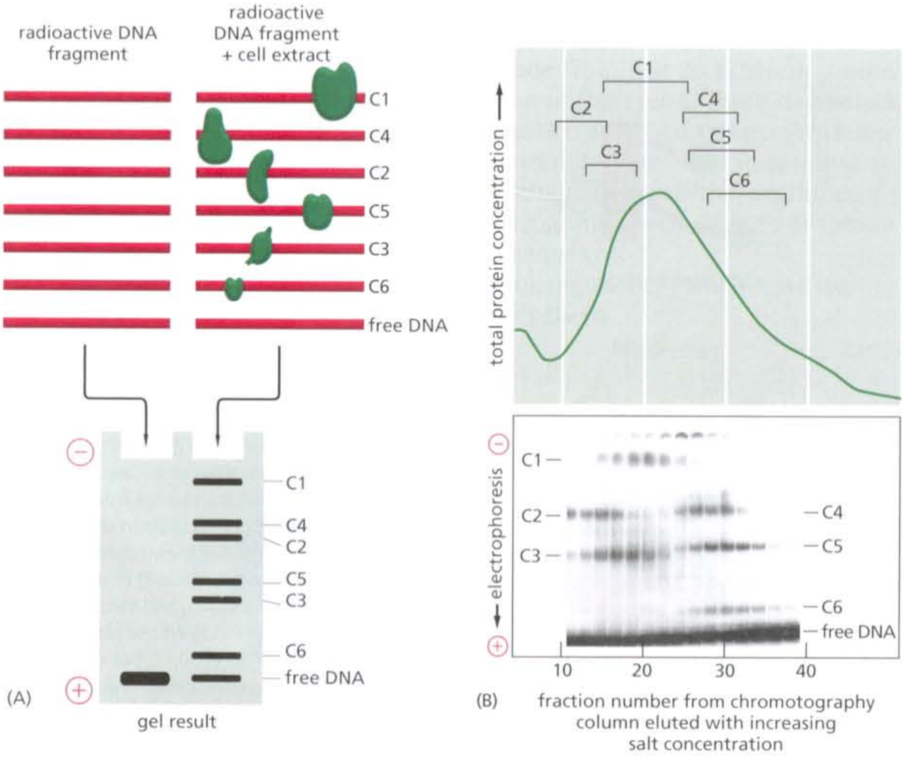

radioactive DNA fragment
radioactive DNA fragment + cell extract
C1
C4
C2
C5
C3
C6
free DNA
−
+
C1
C4
C2
C5
C3
C6
free DNA
(A)
gel result
C1
C2
C4
C3
C5
C6
total protein concentration
C1—
C2—
C3—
—C4
—C5
—C6
—free DNA
−
+
electrophoresis
10
20
30
40
(B)
fraction number from chromotography column eluted with increasing salt concentration

**Figure 7–27 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) a standard chromatography technique (see pp. 512–513) was used to fractionate the extract *(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. With permission from Elsevier.)

then us
the part
can be a
Alth
cell, enc
obtain a
cussed
known,
gene no
provides
genetic

The DN
Be Det

Gene re
ognize i
were dis
This allo
could th
*printing*
regulato
fied frag
site for
long DN
a regula
DNA foc
at every
phospho
nition si
A se
ulatory

colu
conta
diffe

assay.
n
le an
cell line
ragment
of a
gene
tibody
f the
bility of

is
free
the
gments
e
gests
ent
roteins
this DNA

otein
e
graphic
used to
ach
active
e of a
as in (A).
. Heguy
1987.

then used to construct a column that selectively binds proteins that recognize the particular DNA sequence (**Figure 7–28**). Purifications as great as 10,000-fold can be achieved by this means with relatively little effort.

Although most gene regulatory proteins are present at very low levels in the cell, 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 not only provides the complete amino acid sequence of the protein; it also provides the means to produce the protein in unlimited amounts through genetic engineering techniques, also discussed in Chapter 8.

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

Gene regulatory proteins can be discovered before the DNA sequence they recognize is 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 overexpressed in cultured cells and easily purified. *DNA footprinting* is one method of determining the DNA sequences recognized by a gene regulatory protein once it has been purified. This strategy also requires a purified fragment of duplex DNA that contains somewhere within it a recognition site for the protein. Short recognition sequences can occur by chance on any long DNA fragment, although it is often necessary to use DNA corresponding to a regulatory region for a gene known to be controlled by the protein of interest. DNA footprinting is based on nucleases or chemicals that randomly cleave DNA at every phosphodiester bond. A bound gene regulatory protein blocks the phosphodiester bonds from attack, thereby revealing the protein's precise recognition site as a protected zone, or footprint (**Figure 7–29**).

A second way of determining the DNA sequences recognized by a gene regulatory protein requires no prior knowledge of what genes the protein might

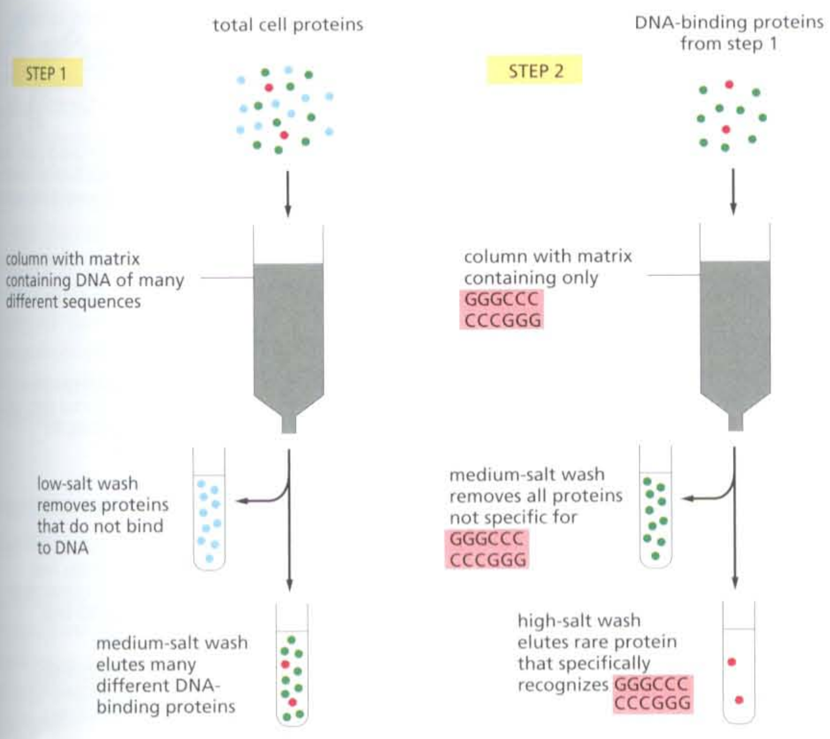

STEP 1
total cell proteins
column with matrix containing DNA of many different sequences
low-salt wash removes proteins that do not bind to DNA
medium-salt wash elutes many different DNA-binding proteins
STEP 2
DNA-binding proteins from step 1
column with matrix containing only
GGGCCC
CCCGGG
medium-salt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

**Figure 7–28 DNA affinity chromatography.** In the first step, all the proteins that can bind DNA are separated from the remainder of the cell 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.

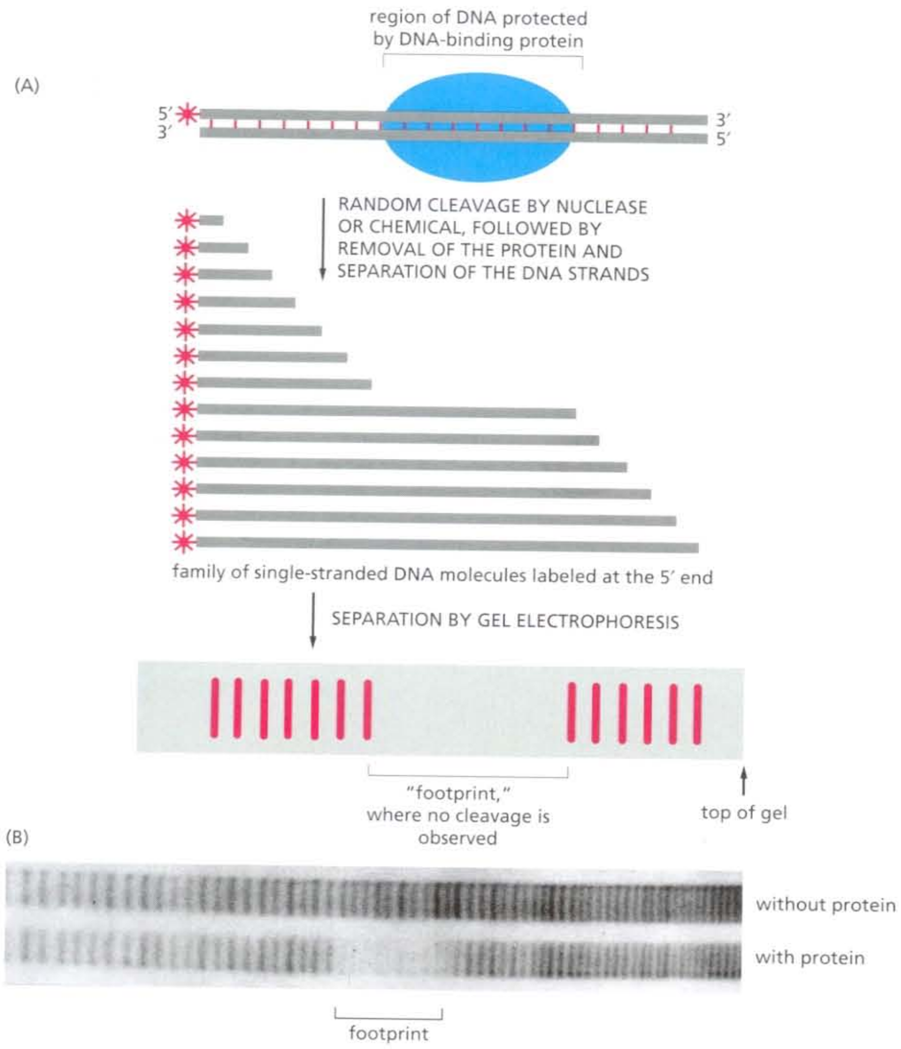

region of DNA protected by DNA-binding protein
(A)
5′
3′
3′
5′
RANDOM CLEAVAGE BY NUCLEASE OR CHEMICAL, FOLLOWED BY REMOVAL OF THE PROTEIN AND SEPARATION OF THE DNA STRANDS
family of single-stranded DNA molecules labeled at the 5′ end
SEPARATION BY GEL ELECTROPHORESIS
"footprint," where no cleavage is observed
top of gel
(B)
without protein
with protein
footprint

**Figure 7–29 DNA footprinting.** (A) Schematic of the method. A DN fragment is labeled at one end with procedure described in Figure 8-34 the DNA is cleaved with a nuclease chemical that makes random, single-stranded cuts. After the DNA molec denatured to separate its two strand resultant fragments from the labeled strand are separated on a gel and detected by autoradiography (see Fig 8–33). The pattern of bands from DN in the presence of a DNA-binding pro is compared with that from DNA cut absence. When protein is present, it covers the nucleotides at its binding and protects their phosphodiester bo from cleavage. As a result, those lab fragments that would otherwise terminate in the binding site are miss leaving a gap in the gel pattern calle "footprint." In the example shown, th DNA-binding protein protects seven phosphodiester bonds from the DN cleaving agent. (B) An actual footpr used to determine the binding site fo gene regulatory protein from human The cleaving agent was a small, iron containing organic molecule that normally cuts at every phosphodieste bond with nearly equal frequency. (B, courtesy of Michele Sawadogo and Robert Roeder.)

regulate. Here, the purified protein is used to select, from a large, randomly generated pool of different short DNA fragments, only those that bind tightly to it. After several rounds of such selection, the nucleotide sequences of the tightly bound DNAs are determined, and a consensus DNA recognition sequence for the gene regulatory protein can be formulated (**Figure 7–30**). Once the DNA sequence recognized by a gene regulatory protein is known, computerized genome searches can identify candidate genes whose transcription the gene

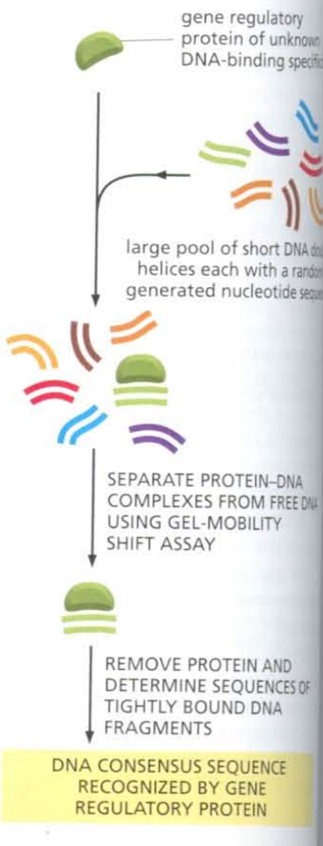
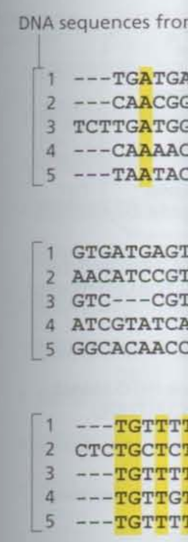

gene regulatory protein of unknown DNA-binding specific
large pool of short DNA do helices each with a random generated nucleotide seque
SEPARATE PROTEIN–DNA COMPLEXES FROM FREE DNA USING GEL-MOBILITY SHIFT ASSAY
REMOVE PROTEIN AND DETERMINE SEQUENCES OF TIGHTLY BOUND DNA FRAGMENTS
DNA CONSENSUS SEQUENCE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

**Figure 7–30 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 illustrated in Figure 7–27. After separation of the DNA–protein complexes from the free DNA, the DNA fragments are removed from the protein and typically used for several additional rounds of the same selection process (not shown). The nucleotide sequences of those DNA fragments that remain through multiple rounds of binding and release can be determined, and a consensus DNA recognition sequence can thus be generated.



regulatory prote
proof. For exam
latory proteins t
cannot resolve t
ulatory proteins
be tested experi

## Phylogenetic
## Through Com

The widespread
ingly simple me
when the gene r
genomes from s
chosen properly
lar, but the regio
will have diverg
vant and therefo
regulatory seque
served islands in
identity of the
sequences must
powerful metho
expression.

## Chromatin Im
## Gene Regulat

A gene regulator
in the genome a
be synthesized,
a heterodimer p
priate signal is r

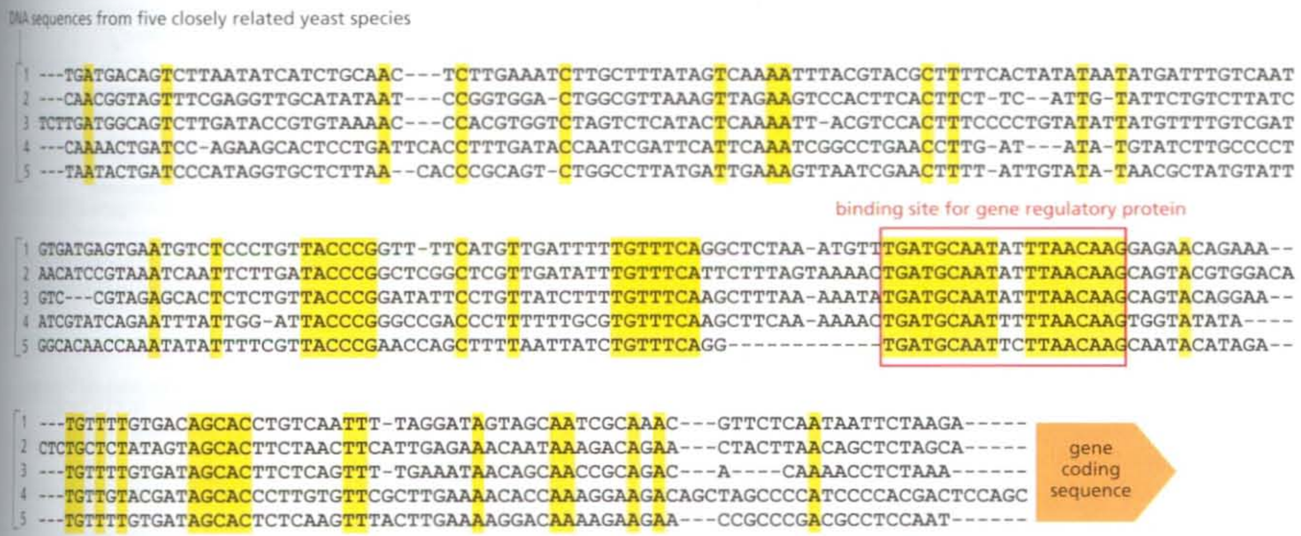

DNA sequences from five closely related yeast species
1 ---TGATGACAGTCTTAATATCATCTGCAAC---TCTTGAAATCTTGCTTTATAGTCAAAATTTACGTACGCTTTTCACTATATAATATGATTTGTCAAT
2 ---CAACGGTAGTTTCGAGGTTGCATATAAT---CCGGTGGA-CTGGCGTTAAAGTTAGAAGTCCACTTCACTTCT-TC--ATTG-TATTCTGTCTTATC
3 TCTTGATGGCAGTCTTGATACCGTGTAAAAC---CCACGTGGTCTAGTCTCATACTCAAAATT-ACGTCCACTTTCCCCTGTATATTATGTTTTGTCGAT
4 ---CAAAACTGATCC-AGAAGCACTCCTGATTCACCTTTGATACCAATCGATTCATTCAAATCGGCCTGAACCTTG-AT---ATA-TGTATCTTGCCCCT
5 ---TAATACTGATCCCATAGGTGCTCTTAA--CACCCGCAGT-CTGGCCTTATGATTGAAAGTTAATCGAACTTTT-ATTGTATA-TAACGCTATGTATT
binding site for gene regulatory protein
1 GTGATGAGTGAATGTCTCCCTGTTACCCGGTT-TTCATGTTGATTTTTGTTTCAGGCTCTAA-ATGTTTGATGCAATATTTAACAAGGAGAACAGAAA--
2 AACATCCGTAAATCAATTCTTGATACCCGGCTCGGCTCGTTGATATTTGTTTCATTCTTTAGTAAAACTGATGCAATATTTAACAAGCAGTACGTGGACA
3 GTC---CGTAGAGCACTCTCTGTTACCCGGATATTCCTGTTATCTTTTGTTTCAAGCTTTAA-AAATATGATGCAATATTTAACAAGCAGTACAGGAA--
4 ATCGTATCAGAATTTATTGG-ATTACCCGGGCCGACCCTTTTTTGCGTGTTTCAAGCTTCAA-AAAACTGATGCAATTTTTAACAAGTGGTATATA----
5 GGCACAACCAAATATATTTTCGTTACCCGAACCAGCTTTTAATTATCTGTTTCAGG-----------TGATGCAATTCTTAACAAGCAATACATAGA--
1 ---TGTTTTGTGACAGCACCTGTCAATTT-TAGGATAGTAGCAATCGCAAAC---GTTCTCAATAATTCTAAGA-----
2 CTCTGCTCTATAGTAGCACTTCTAACTTCATTGAGAAACAATAAAGACAGAA---CTACTTAACAGCTCTAGCA-----
3 ---TGTTTTGTGATAGCACTTCTCAGTTT-TGAAATAACAGCAACCGCAGAC---A----CAAAACCTCTAAA------
4 ---TGTTGTACGATAGCACCCTTGTGTTCGCTTGAAAACACCAAAGGAAGACAGCTAGCCCCATCCCCACGACTCCAGC
5 ---TGTTTTGTGATAGCACTCTCAAGTTTACTTGAAAAGGACAAAAGAAGAA---CCGCCCGACGCCTCCAAT------
gene coding sequence

**Figure 7–31 Phylogenetic footprinting.** This example compares DNA sequences upstream of the same gene from five closely related yeasts; identical nucleotides are highlighted in *yellow*. Phylogenetic footprinting reveals DNA recognition sites for regulatory proteins, as they are typically more conserved than surrounding sequences. Only the region upstream of a particular gene is shown in this example, but the approach is typically used to analyze entire genomes. The gene regulatory proteins that bind to the site outlined in *red* are shown in Figure 7–21. Some of the shorter phylogenetic footprints in this example represent binding sites for additional gene regulatory proteins, not all of which have been identified. (From M. Kellis et al., *Nature* 423:241–254, 2003, with permission from Macmillan Publishers Ltd., and D.J. Galgoczy et al., *Proc. Natl Acad. Sci. U.S.A.* 101:18069–18074, 2004, with permission from National Academy of Sciences.)

15th

regulatory protein of interest might control. 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, in the end, be tested experimentally.

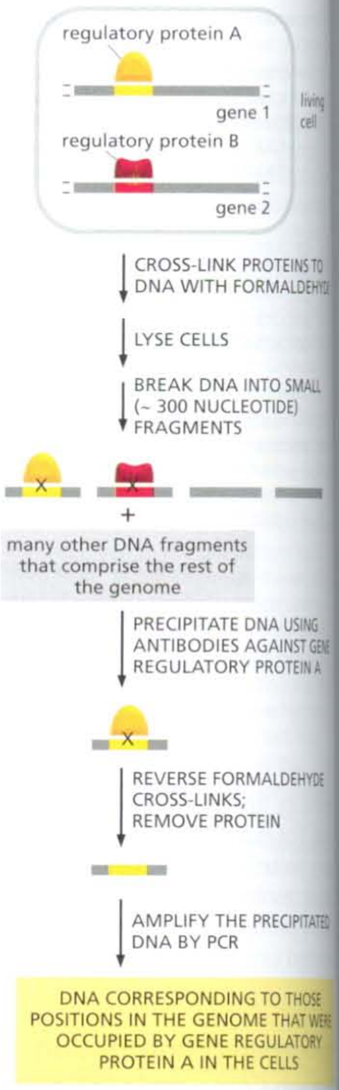
## Phylogenetic Footprinting Identifies DNA Regulatory Sequences Through Comparative Genomics

The widespread availability of complete genome sequences provides a surprisingly simple method for identifying important regulatory sites on DNA, even when the gene regulatory protein that binds them is unknown. In this approach, genomes from several closely related species are compared. If the species are chosen properly, the protein-coding portions of the genomes will be very similar, but the regions between sequences that encode protein or RNA molecules will have diverged considerably, as most of this sequence is functionally irrelevant and therefore not constrained in evolution. Among the exceptions are the regulatory sequences that control gene transcription. These stand out as conserved islands in a sea of nonconserved nucleotides (**Figure 7–31**). Although the identity of the gene regulatory proteins that recognize the conserved DNA sequences must be determined by other means, phylogenetic footprinting is a powerful method for identifying many of the DNA sequences that control gene expression.

## Chromatin Immunoprecipitation Identifies Many of the Sites That Gene Regulatory Proteins Occupy in Living Cells

A gene regulatory protein will not occupy all of its potential DNA-binding sites in the genome at a particular time. Under some conditions, the protein may not be synthesized, and so will be absent from the cell; it may be present but lacking a heterodimer partner; or it may be excluded from the nucleus until an appropriate signal is received from the cell's environment. Even if the gene regulatory

Figure 7–32 **Chromatin immunoprecipitation.** This method allows the identification of all the sites in a genome that a gene regulatory protein occupies *in vivo*. For the amplification of DNA by a polymerase chain reaction (PCR), see Figure 8–45. The identities of the precipitated, amplified DNA fragments can be determined by hybridizing the mixture of fragments to DNA microarrays, as described in Chapter 8.


regulatory protein A
gene 1
living cell
regulatory protein B
gene 2
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 CROSS-LINKS; REMOVE PROTEIN
AMPLIFY THE PRECIPITATED DNA BY PCR
DNA CORRESPONDING TO THOSE POSITIONS IN THE GENOME THAT WERE OCCUPIED BY GENE REGULATORY PROTEIN A IN THE CELLS

protein is present in the nucleus and is competent to bind DNA, components of chromatin or other gene regulatory proteins that can bind to the same or overlapping DNA sequences may occlude many of its potential binding sites on DNA.

**Chromatin immunoprecipitation** provides one way of empirically determining the sites on DNA that a given gene regulatory protein occupies under a particular set of conditions (**Figure 7–32**). In this approach, proteins are covalently cross-linked to DNA in living cells, the cells are broken open, and the DNA is mechanically sheared into small fragments. Antibodies directed against a given gene regulatory protein are then used to purify DNA that became covalently cross-linked to that protein in the cell. If this DNA is hybridized to microarrays that contain the entire genome displayed as a series of discrete DNA fragments (see Figure 8–73), the precise genomic location of each precipitated DNA fragment can be determined. In this way, all the sites occupied by the gene regulatory protein in the original cells can be mapped on the cell's genome (**Figure 7–33**).

Chromatin immunoprecipitation is also routinely used to identify the positions along a genome that are packaged by the various types of modified histones (discussed in Chapter 4). In this case, antibodies specific to the particular histone modification of interest are employed.

## 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 a gene regulatory protein recognizes. Heterodimerization increases the range of DNA sequences that can be recognized. Powerful techniques are now available for identifying and isolating these proteins, the genes that encode them, and the DNA sequences they recognize, and for mapping all of the genes that they regulate on a genome.*

# 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.

In the mid-twentieth century, 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 of 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