Protein Motifs for DNA Binding

Hang Xu, *University of Vermont, Burlington, Vermont, USA*

Scott W Morrical, *University of Vermont, Burlington, Vermont, USA*

Protein–DNA interactions play a central role in directing the flow of genetic information and in the control of life itself. Research has demonstrated that many DNA-binding proteins belong to distinct families containing common structural motifs.

Introductory article

Article Contents

Introduction

Polypeptides, the primary structure of proteins, can fold into different secondary structures, such as α helices, β strands, and loops/turns, according to the intrinsic properties of each polymer. The special combination and arrangement of these secondary structural elements is the supersecondary structure, which is the so-called protein motif. Here, we introduce some protein motifs that interact with another important group of macromolecules, deoxyribonucleic acid (DNA). In its common double-stranded form, DNA consists of two helices and the base pairs are exposed at the bottoms of two grooves on the DNA surface. The wider groove is called the major groove and the narrower groove is called the minor groove. Almost all DNA-binding protein motifs contain structural elements that fit into one or both of these grooves. In a DNA–protein complex, not only are the shapes of macromolecules complementary to each other, but also some chemical interactions between them are required to stabilize the complex. Three major types of chemical interactions stabilize protein–DNA complexes: electrostatic interactions, hydrophobic interactions and hydrogen bonds. Under physiological conditions, DNA carries negative charges due to its phosphate residues; therefore, a basic protein region carrying positive charges is able to attract the DNA and may form a stable complex with it. Hydrophobic interactions and hydrogen bonds between amino acid side-chains and the bases of DNA are also quite important. The combination of these interactions establishes the affinity between protein and DNA. Now it is necessary to clarify the concepts of affinity and specificity. In a complex of a certain protein and its specific DNA recognition sequence, it is safe to say that all of the interactions contribute to the affinity, but the specificity can only be observed in a particular set of interactions involving those specific bases. This limited set of interactions is called sequence-specific, while the remainder is called nonsequence-specific. Usually, interaction with the sugar–phosphate backbone of the DNA molecules is nonsequence-specific.

Helices in DNA Grooves

Helix–turn–helix motif and homeodomain

The helix–turn–helix (HTH) motif is the most common motif found in DNA-binding proteins in both prokaryotes and eukaryotes. This motif contains about 20 amino acids that form two α helices and a short four-residue turn in between (**Figure 1a**). When interacting with DNA, the second helix lies in the major groove and contacts DNA in a sequence-specific pattern (**Figure 1b**). Thus the second helix is called the 'recognition helix' for its DNA recognition function.

The prokaryotic representatives of the HTH family include the Cro and repressor proteins in lambdoid phages. The two proteins control the phage life cycle by binding to phage DNA and regulating the expression of a particular gene set. At the recognition site, hydrogen bonds as well as hydrophobic interactions between the edges of base pairs and the side-chains of amino acids define sequence-specific interactions. For example, in the complex formed between the phage 434 repressor and its DNA recognition sequence, OR1, the 5′-ACAA-3′ sequence in DNA is recognized by the Thr27, Gln28 and Gln29 amino acid residues in the repressor protein. The hydrogen bonds are formed between Gln28 and base pair 1 (adenine and thymine, AT), and between Gln29 and base pair 2 (cytosine and guanine, CG). In addition, the methyl group of T in base pair 3 (AT) is inserted into a hydrophobic pocket formed by Thr27 and Gln29. An interesting feature of the prokaryotic HTH proteins is that they usually function as dimers and their DNA recognition sequences are palindromic. Thus each monomer interacts with one half of the symmetric DNA sequence.

The HTH motif is also found in eukaryotes with slight modification to the prokaryotic HTH. The eukaryotic HTH motif, homeodomain, recognizes a DNA sequence called homeobox that has been identified in a large variety of genes. The homeodomain is folded into three α helices, with helices 2 and 3 resembling the HTH motif (**Figure 1c**). In contrast to the prokaryotic HTH, the homeodomain

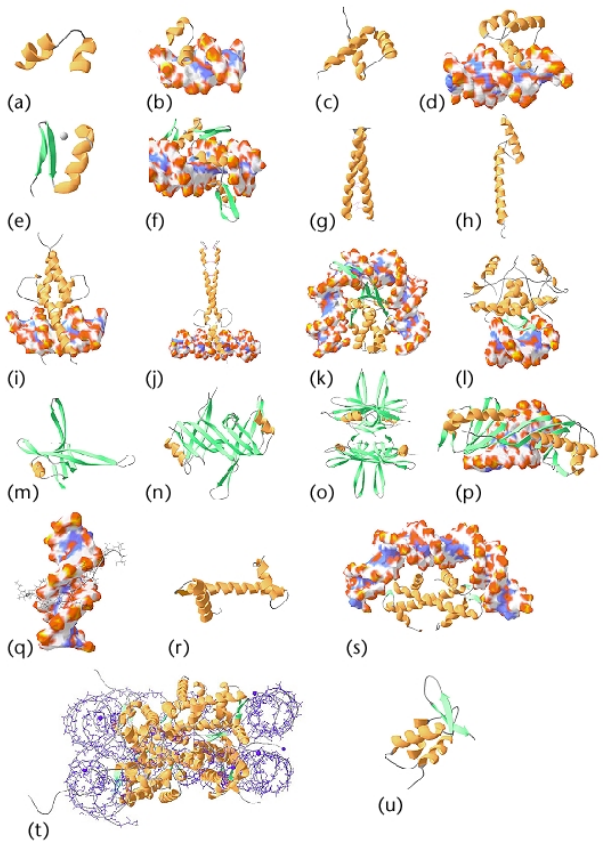

Figure 1 (a) Helix–turn–helix (HTH) motif; (b) HTH–DNA complex; (c) homeodomain; (d) homeodomain–DNA complex; (e) a single zinc finger (zn, grey); (f) three consecutive zinc fingers bound to DNA; (g) leucine zipper; (h) helix–loop–helix (HLH) monomer; (i) HLH dimer–DNA complex; (j) HLH–bZIP–DNA complex; (k) integration host factor (IHF)–DNA complex; (l) MetJ–DNA complex; (m) *Escherichia coli* single-strand DNA-binding protein (SSB) monomer; (n) *E. coli* SSB dimer, showing the extended β sheet; (o) *E. coli* SSB tetramer; (p) TATA-binding protein (TBP)–DNA complex; (q) AT hook motif–DNA complex; (r) histone fold; (s) histone fold pair–DNA complex; (t) nucleosome core particle (DNA backbones, blue); (u) histone H5 globular domain.

binds to DNA as a monomer, with a more extended recognition helix, helix 3 (**Figure 1d**). Despite the extensive nonspecific contacts on DNA from the loop and other helices, the specificity is determined by the extended part of helix 3. The side-chains of residues 47, 50, 51 and 54 interact with the consensus sequence in the major groove. Different DNA sequences require different combinations of these residues. For example, isoleucine or valine at position 47 form hydrophobic interactions with T in DNA, while an asparagine replaces isoleucine or valine at position 47 when C replaces T at the same position in the DNA. In conclusion, these residues are highly responsible for the specific binding of homeodomain to the homeobox.

Zinc fingers

The zinc-finger proteins comprise another class of eukaryotic DNA-binding proteins, many of which are identified as transcription factors. Zinc-finger proteins, as their name implies, contain Zn^{2+} ions in their structures. In the absence of zinc, the zinc-finger proteins lose the ability to bind DNA. What is the role of the zinc ion? Is it directly involved in protein–DNA contacts, or is it only a structure-stabilizing component? Both X-ray and nuclear magnetic resonance (NMR) techniques revealed that this DNA-binding motif (the 'finger') is composed of a two-stranded antiparallel β sheet and an α helix, with the zinc ion buried in the interior (**Figure 1e**). Therefore, the zinc is unlikely to interact directly with DNA. Instead, it is part of the hydrophobic core of the motif and helps to fold the polypeptide into the appropriate conformation. Other structural details of this motif include the following. This ββα fold unit possesses not only a hydrophobic core, but also many exposed polar side-chains that are capable of interacting either with DNA or with other parts of the protein. The zinc atom is coordinated by two cysteine and two histidine residues. The two cysteine ligands are located on the β sheet and the two histidines are on the α helix. The binding pattern of the zinc finger is quite similar to that of the HTH proteins. The α helix is projected into the major groove and makes specific contacts with the DNA. It is noteworthy that a single zinc finger can bind to DNA but in a nonspecific manner. So, in nature, the zinc-finger protein is usually composed of from two to as many as 30 repeating 'fingers', which bind continuously to the track of the major groove of a DNA molecule (**Figure 1f**, showing three 'fingers'). This pattern accounts for the sequence specificity of DNA binding.

The classical zinc-finger proteins are described above. The zinc-finger protein family also includes other members: in some proteins, an additional N-terminal β strand is involved and the motif becomes one of the βββα type. In addition to this structural variant, variations in the zinc ligand pattern have been identified. The classical zinc finger is of the Cys-Cys-His-His type, while others (e.g. glucocorticoid receptor) use a Cys-Cys-Cys-Cys scheme, and still others (e.g. gag protein of the human immunodeficiency virus) use a Cys-Cys-His-Cys pattern.

Leucine zippers

The leucine zipper motif was first discovered in certain transcription factors, such as the GCN4 protein in yeast. The amino acid sequences of these proteins exhibit a unique feature: in a region of about 30–40 residues, every seventh residue is a leucine. Since the helical repeat of the α helix is 3.6 residues per turn, the leucines are located at approximately two-turn intervals, and on the same side of the helix. Based on this, the structure of this motif was first hypothesized to be a dimer consisting of two antiparallel

α helices, with the leucine residues in each helix alternating in the dimerization region to construct a zipper-like pattern. The leucine residues were regarded as the teeth of that 'zipper', and hence the motif was named. But this model was later shown to be only partly correct. The proteins do dimerize, but the helices are in a parallel orientation, and the leucine residues are adjacent to their counterpart in the opposite helix, rather than alternated (**Figure 1g**). Although the 'zipper' model was rejected, the name stuck and is still in use. Because there is a basic region N-terminal to the leucine-rich region, another model called 'scissors-grip' was proposed. In this model, a continuous α helix is formed, containing both the basic region and the leucine-rich region. Two such α helices then dimerize through contacts at the leucine-rich regions, while the basic regions diverge to interact with the major groove of DNA. The dimerization region can be characterized as a coiled coil, in which long helices intertwine to form a superhelical structure. The formation of the coiled coil slightly interrupts the α helix and alters its helical repeat to 3.5 residues per turn. Thus in the coiled-coil region, the leucine residues are positioned precisely at two-turn intervals. This model has been verified by the structures of several 'leucine zipper' proteins bound to their corresponding DNA fragments. The leucine regions are not responsible for DNA binding, but they provide the correct arrangement of the two long helices, therefore enabling the correct positioning of the basic regions against the major groove and the subsequent base-specific DNA binding.

Helix–loop–helix motif and variants

Many helix–loop–helix (HLH) proteins are transcriptional activators. They contain a common structural element consisting of two α helices connected by a loop of variable length (**Figure 1h**). The number of residues in the loop ranges widely from five to 24. The helices are amphipathic (i.e. they contain hydrophobic residues on one side and hydrophilic residues on the other) and the hydrophobic residues are highly conserved. These residues are involved in the formation of the interface between two HLH monomers. The interface is a parallel, left-handed, four-helix bundle stabilized by a hydrophobic core. In addition to the hydrophobic region, there is another important region in the HLH motif, composed of basic amino acid residues which make contacts with DNA (**Figure 1i**).

The HLH proteins exist as dimers in solution and interact with DNA in dimeric form. In the complex with DNA, the two basic regions (one from each monomer) interact with the major groove-like forceps. This binding pattern is similar to that of the leucine zipper proteins. Interestingly, some proteins combine elements of both HLH and leucine zipper motifs (**Figure 1j**). In this structural hybrid, the second helix in an HLH motif is extended and intertwines with its counterpart from the other monomer to form a coiled-coil structure, as in the leucine zipper proteins. Such a protein is classified as an HLH-bZIP protein (bZIP is another name for leucine zipper).

β Ribbons in DNA Grooves

MetJ bound to the major groove

The *metJ* gene in *Escherichia coli* encodes the MetJ protein, which is the repressor of the *met* operon. MetJ is the prototype of one class of DNA-binding proteins, employing a β ribbon to interact with the DNA major groove.

The MetJ protein contains 104 amino acids, folded into three α helices and one β strand. The protein can form a stable dimer in solution and the monomer–monomer interface includes a two-stranded antiparallel β sheet (**Figure 1l**). This particular β sheet protrudes from the surface of the dimeric molecule and lies in the major groove to contact specifically its consensus DNA sequence 5′-AGACGTCT-3′. Although some loop structures outside of the β sheet also interact with the DNA molecule, their major function is to stabilize the protein–DNA complex by hydrogen bonding to the DNA backbones. In other words, these loops account for some of the affinity, but not the specificity, of binding. The specificity is actually determined by the β sheet consisting of β strands from different monomers. Upon binding, both the protein and DNA components essentially maintain their original structures, which is unusual in protein–DNA interactions. The only conformational change observed in the DNA is that the corresponding major groove is narrowed slightly to make a better fit with the MetJ β sheet.

IHF bound to the minor groove

The binding of prokaryotic integration host factor (IHF) to DNA causes a surprising conformational change: a U-turn in the DNA. Although the biological function of IHF needs further clarification, some evidence suggests that IHF-induced bending of DNA could bring elements that are distant in the DNA sequence close together in space, thus enabling their cooperative function. For example, the ability to bend DNA appears to be important in IHF-dependent stimulation of bacterial DNA replication and of bacteriophage site-specific recombination. IHF may also function as transcriptional corepressor in some cases.

How does IHF bend the DNA so much? IHF is a heterodimer composed of two similar subunits, α and β. The overall shape of IHF can be described as a body with two extended arms (**Figure 1k**). The arms are two two-stranded β sheets, each from one of the subunits, and the body is formed by the remainder of the subunits. The top of the body, together with the arms, binds DNA molecules in

the minor groove and makes continuous contact with the top 'kink' of the bent DNA. The consensus DNA sequence for IHF binding is not palindromic, therefore the interactions between protein and DNA are asymmetric, despite the apparent overall symmetry of the complex. In order to bend the DNA, two prolines (Pro65α, Pro64β), located at the tips of opposite arms, intercalate the DNA at two sites that are nine base pairs apart and cause the kink in the middle of the DNA. These prolines are highly conserved among all members of the IHF family, and their intercalation widens the minor groove to ~ 10 Å, therefore enlarging the contact area between DNA and protein. Meanwhile, at opposite sides of the complex, by inserting side-chains into the minor groove and applying electrostatic interactions with the DNA backbones, the body of IHF clamps on to the rest of the DNA recognition site and accomplishes the overall U-turn in the DNA.

Other Types of DNA-binding Motifs

OB fold

The OB (oligonucleotide/oligosaccharide-binding) fold is found in some nonsequence-specific single-stranded DNA-binding proteins. It was first described in 1993 and its general structure was extracted from four members of this family. This fold is composed of a β barrel capped by an α helix at its edge. The five-stranded barrel contains two orthogonally packed β sheets, one of which is formed by strands 2, 3 and part of strand 1; the other is formed by strands 4, 5 and the rest of strand 1. A common bulge is observed in β strand 1, forcing it to change direction and participate in both β sheets. The capping α helix is always located between β strands 3 and 4. Members of the OB-fold family are not conserved in primary sequence but in three-dimensional structure. They all contain an oligonucleotide-or oligosaccharide-binding site on the same part of the β-barrel surface, involving several loop structures connecting strands 1 and 2, strands 4 and 5, plus strand 3 and the α helix.

The *E. coli* single-stranded DNA (ssDNA)-binding protein (SSB) contains an OB fold. SSB plays important roles in DNA replication, recombination and repair by removing unfavourable secondary structure from single-stranded DNA and by protecting the ssDNA from degradation. The OB fold in SSB fits the consensus structure with only one minor modification: strand 1 is divided into two strands, 1 and 1′ connected by a four-residue turn instead of the bulge (**Figure 1m**). It is noteworthy that the structure between strands 4 and 5, including a two-stranded β ribbon and a loop, participates in DNA binding and monomer–monomer interface formation.

SSB exists as a homotetramer organized as a dimer of dimers. Within each dimeric unit, the major monomer–monomer interface is formed by strand 1. Via the interaction between two strand 1s from both monomers, an extended six-stranded antiparallel β sheet (**Figure 1n**) is formed in the dimer and further acts as the dimer–dimer interface to build up the tetramer (**Figure 1o**). Since SSB functions to bind a very long stretch of ssDNA, the ssDNA is likely to wrap around the SSB tetramer so that not only the β-barrel surface but also other parts including the α helices and loops should participate in DNA binding. This has been verified by crystallographic studies of the SSB–ssDNA complex. These studies also elucidated the structural basis of two ssDNA-binding modes observed in SSB, the so-called $(SSB)_{65}$ mode in which 65 nucleotide residues are bound per tetramer, and the $(SSB)_{35}$ mode in which 35 nucleotide residues are bound per tetramer. In $(SSB)_{65}$, the ssDNA is wrapped around all of the monomers and occupies all four binding sites. In $(SSB)_{35}$, the wrapping mode is different and only one monomer's binding site is fully occupied by ssDNA. Binding sites on two other monomers are half occupied by ssDNA and the fourth monomer is completely unoccupied.

TATA-binding protein

In eukaryotes, transcription events can only occur after the assembly of a ribonucleic acid (RNA) polymerase complex around the initiation site. Upstream of the initiation site, there is a TA-rich sequence called TATA box, and the protein that specifically binds to it is the so-called TATA-binding protein (TBP). The recognition of TATA box by TBP plays a very important role in the initiation of transcription because it is considered the first step of the RNA polymerase complex assembly.

The C-terminal domain of TBP is highly conserved and consists of two subdomains, with 89 and 90 amino acid residues, respectively. It is surprising that these two subdomains are almost identical in overall structure with only 30% homology in sequence. Each subdomain of this pseudosymmetric structure contains two α helices and a five-stranded antiparallel β sheet (**Figure 1p**). The curved β sheets are responsible for the interaction with DNA, while the α helices provide the structural scaffold. The N-terminal domain of TBP is neither conserved nor necessary for TATA binding and will not be discussed here.

The TBP–TATA box interaction is different from other transcription factor interactions with their recognition sequences. TBP interacts with the minor groove of DNA, which is narrower and contains fewer potential hydrogen-bonding sites to base edges compared to the major groove. But TBP adapts to this situation quite well. First, by intercalating two pairs of phenylalanine residues (Phe207/Phe190, Phe99/Phe116) into the two ends of the 8-bp TATA box, TBP distorts the DNA structure and makes the

minor groove much wider (9 Å compared to 4 Å in B-DNA) and shallower to facilitate the contact. Second, the role of hydrogen bonding is reduced in this interaction compared to other sequence specific DNA-binding proteins; in fact only six are observed in the protein–DNA interface. Instead, hydrophobic and van der Waals interactions account for the most of the specificity and stability of the TBP–TATA box complex.

AT hook motif

The 'AT hook' motif is a short peptide sequence found in many DNA-binding proteins. This peptide is believed to bind to AT-rich DNA sequences and is named for this property. It has been shown that the peptide, which contains a palindromic consensus sequence 'Pro-Arg-Gly-Arg-Pro', binds to the minor groove of B-DNA. The peptide sequences flanking the consensus sequence are also important because they can account for up to approximately a 100-fold difference in DNA-binding affinity. Interestingly, the AT hook motif usually does not occur in a single copy in a protein, but instead as several potential DNA-binding sites which are connected by linker regions, suggesting that crosstalk between different DNA-binding sites is possible.

The eukaryotic high mobility group (HMG)-I family is regarded as a model of the AT hook class of DNA-binding proteins and has been studied most thoroughly. This family has three members, HMG-I, HMG-Y and HMGI-C. The HMG-I and HMG-Y proteins are products of the same gene, but the latter has an internal deletion due to alternative messenger RNA (mRNA) splicing. So these proteins are sometimes referred to collectively as the HMG-I (Y) protein. HMG-I (Y) contains three AT hook motifs, which are called DNA-binding domain (DBD)-1,-2 and-3. Of the three, DBD-2 has the highest affinity for DNA. The NMR structures of DBD-2 and DBD-3 complexed with DNA fragments were determined in 1997, and revealed the structural basis for the difference in their DNA-binding affinities. The consensus peptide backbone forms a U-shaped structure and the tip (also called 'basic core'), consisting of 'Arg-Gly-Arg', deeply inserts into the minor groove (**Figure 1q**). At the C-terminal side of the basic core in DBD-2, a β turn, together with the extension from the turn, bridges the minor groove by interacting with DNA backbones and bases on both sides of the groove. Although the structure of the consensus peptide sequence in DBD-3 is almost identical to that in DBD-2, it lacks the bridge structure and thus exhibits much lower affinity for DNA.

Histones

The eukaryotic genomic DNA is much longer and more complicated than its prokaryotic counterpart, and chromatin is formed instead of naked DNA. The basic unit of chromatin is the nucleosome, which is a bead-like particle formed from a standard length of DNA wrapped around histone proteins. Since sequence specificity is not applicable in this case, most interactions between histone proteins and DNA involve protein contacts with the sugar–phosphate backbones of DNA strands. The nucleosome is composed of four histone proteins, H2A, H2B, H3 and H4, which assemble into an octameric protein core containing two copies of each protein. In the nucleosome, DNA wraps around the histone octamer for $1\frac{3}{4}$ turns. Linker histones H1 or H5 organize recurring nucleosome beads into higher-ordered chromatin structures.

The histone octamer with wrapped DNA is called the core particle, and all histone proteins in it contain a common motif called the 'histone fold' (**Figure 1r**). This special fold contains one long central α helix (α_2) and two short side helices (α_1 and α_3), connected by two loops, L1 and L2. While the overall organization of the fold is conserved between core histone proteins, some structural variation is observed between different types of histones, e.g. there is a slight bend in the central helix (α_2) found in histones H2A and H4; also, histones H2A and H3 contain an additional helix. In the assembly of the histone core, heterodimers of H2A/H2B and H3/H4 are formed first. Each pairing results in an antiparallel arrangement of two histone folds, stabilized by hydrophobic interactions between the two α_2 helices and by hydrogen bonding between the L1 and L2 loops of different monomers. Further assembly is brought about by helix bundle formation between heterodimers: the helix bundle between two H3 histones brings two H3-H4 dimers together to form a tetramer, and two H2B-H4 bundles add two H2A-H2B dimers on to the tetramer to form the final octamer.

In the nucleosome, 146-bp DNA is wrapped around the histone octamer. The histone fold pair found in each heterodimer contains three binding sites for DNA (**Figure 1s**). One site contains the two α_1 helices of the heterodimer, and is called the $\alpha_1\alpha_1$ site. The other two sites are found in the vicinity of the L1 and L2 loops. Histone fold interactions with DNA include the electrostatic positioning of DNA phosphate groups by the N-terminal ends of α_1 or α_2 helices, hydrogen bonds between DNA phosphate oxygens and histone main-chain amides, and the insertion of an arginine into the minor groove every time it faces the protein core. In addition, sequence nonspecific hydrophobic and hydrogen-bond networks contribute to the stability of the core particle (**Figure 1t**). Other interactions with DNA occur outside the histone fold.

All the linker histone proteins (H1, H5) contain a globular domain (**Figure 1u**) as well as unstructured N-and C-terminal tails. The structure of the globular domain consists of a three-helix bundle and a β 'wing'. The helix structure of the globular domain is a variant of HTH, in which a seven-residue loop replaces the standard

four-residue turn. Combined with the β wing, this motif constitutes the 'wing–helix' family.

Recurring Dimerization Motifs

In many cases discussed above, the protein functions as a dimer in the DNA–protein complex. Protein dimerization alternatively provides interactions with a palindromic DNA sequence (HTH motif), builds up the DNA-binding interface (MetJ), or promotes DNA bending (IHF, SSB, histones). What is the biochemical and evolutionary logic underlying dimerization in DNA-binding proteins? Structural biology provides some clues. First, DNA is symmetric. Some DNA sequences are palindromic, and thus suitable for binding to a protein dimer. In addition, disregarding the base sequence, double-stranded DNA contains a 2-fold rotational axis of symmetry perpendicular to the central axis of the duplex. This kind of axis can also be found in the protein dimers that are bound to DNA in a perpendicular mode, such as leucine zippers and HLH. The symmetric binding on two sites increases the affinity and may also increase the specificity by providing more bases for the interaction. Another possible benefit is that the affinity and specificity could be adjusted in a protein heterodimer by changing partners. Second, DNA bending is required for the crosstalk between distant elements in the DNA sequence. The simplest way to bend DNA is to apply a dimer, with each monomer contacting different sites on the same DNA molecule. IHF may be regarded as a model.

Summary

Structural characterization of many different DNA-binding proteins has revealed several conserved DNA-binding motifs. Different motifs employ either α-helix, β-ribbon or extensive β-sheet secondary structural elements in the protein–DNA interface. While some exceptions exist, the formation of a specific protein–DNA complex usually involves conformational changes in both the protein and the DNA components, resulting in an induced fit. Many DNA-binding proteins exist as dimers and bind to DNA in this form.

Further Reading

Bewley C, Gronenborn A and Clore GM (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annual Review of Biophysics and Biomolecular Structure* **27**: 105–131.

Billeter M (1996) Homeodomain-type DNA recognition. *Progress in Biophysics and Molecular Biology* **66**: 211–225.

Branden C and Tooze J (1991) *Introduction to Protein Structure*. New York: Garland.

Kohn W, Mant C and Hodges R (1997) α-Helical protein assembly motifs. *Journal of Biological Chemistry* **272**: 2583–2586.

Luger K, Mader A, Richmond R, Sargent D and Richmond T (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–260.

Raghunathan S, Kozlov A, Lohman T and Waksman G (2000) Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nature Structural Biology* **7**: 648–652.