Molecular Cell, Vol. 8, 937–946, November, 2001, Copyright ©2001 by Cell Press

Recognition of Specific DNA Sequences

Review

Colin W. Garvie and Cynthia Wolberger[1]
Department of Biophysics and Biophysical Chemistry
and the Howard Hughes Medical Institute
Johns Hopkins University School of Medicine
Baltimore, Maryland 21205

Proteins that recognize specific DNA sequences play a central role in the regulation of transcription. The tremendous increase in structural information on protein-DNA complexes has uncovered a remarkable structural diversity in DNA binding folds, while at the same time revealing common themes in binding to target sites in the genome.

Introduction

The ability of a protein to bind selectively to a particular DNA site in the genome is the foundation upon which transcriptional regulatory pathways are built. The DNA binding proteins that regulate transcription are capable of selecting the correct binding site out of a vast number of potential sites in the genome. Our understanding of how these proteins do so has expanded tremendously in the 20 years since the determination of the first structures of sequence-specific DNA binding proteins and in the 15 years since the first structure determinations of protein-DNA complexes at atomic resolution. Since then, structures of proteins in complex with their DNA sites have been determined at an ever-increasing rate, reflecting the tremendous advances in structure determination methods as well as in cloning and expression. With more than 250 structures of protein-DNA complexes deposited in the Protein Data Bank as of August 2001, we now have a fuller picture of the architecture of DNA binding proteins and how they bind to DNA (Figure 1). Combined with biochemical and genetic studies, these structures have illuminated the various strategies by which these proteins bind selectively to particular DNA sequences.

Just 10 years ago, it was possible to write a short yet comprehensive review of the known DNA binding proteins (Harrison, 1991). The large number of protein-DNA structures determined since then makes it impossible to cover them all within the scope of this review, even when one considers only those of proteins bound in a specific manner to DNA. A comprehensive cataloging of DNA binding folds and their manner of interacting with DNA can be found in several recent reviews (Luscombe et al., 2000, 2001). Our intent here is to describe the architectural principles of DNA binding proteins, the structural variations in the DNA to which these proteins bind, and the manner in which specific DNA sequences are recognized, presenting selected examples to illustrate points. This review focuses on proteins that regulate transcription, but we note that there are many other examples of sequence-specific recognition among proteins that mediate recombination, DNA cleavage, methylation, and other processes. The broad principles of DNA binding outlined here pertain to these proteins as well, although the structural details differ.

Basic Requirements for DNA Binding

Double-stranded DNA is a polymer of relatively uniform structure, with a highly negatively charged sugar-phosphate backbone and a core of stacked base pairs whose edges are exposed in the major and minor grooves. Since each base pair has a characteristic set of functional groups, each DNA sequence has a chemical "signature" characterized by the pattern of these groups exposed in the DNA grooves. It is this chemical surface, along with sequence-dependent variations in DNA structure and flexibility, that is recognized by proteins. Proteins recognize a particular sequence by having a surface that is chemically complementary to that of the DNA, forming a series of favorable electrostatic and van der Waals interactions between the protein and the base pairs. In addition, all protein-DNA complex structures contain a large number of contacts with the negatively charged phosphates that include salt bridges with positively charged side chains and hydrogen bonds with uncharged main chain or side chain atoms in the protein (Luscombe et al., 2001).

The great majority of protein-DNA complex structures contain DNA that is essentially B-form, with only a moderate degree of bending and deformation. In these cases, it is the surface of the protein that conforms to the DNA structure, most commonly through the use of α helices or β sheets that protrude from the surface of the DNA binding protein and penetrate the DNA grooves. Initial proposals based on the structure of B-DNA suggested that either α helices or antiparallel β strands have the ideal proportions for insertion into the B-DNA major groove (Church et al., 1977), and indeed early structures of phage and bacterial repressor proteins bore this out (Anderson et al., 1982; McKay and Steitz, 1981; Pabo and Lewis, 1982; Somers and Phillips, 1992). As structures of a more diverse array of DNA binding domains were solved, however, it emerged that secondary structural units could be inserted into the DNA in a variety of ways and still form base contacts. While secondary structural elements most commonly mediate base contacts, the immunoglobulin-like proteins NF-κB (Ghosh et al., 1995; Muller et al., 1995), NFAT (Chen et al., 1998b), and STAT (Chen et al., 1998c) use a series of loops to form complementary contacts with undistorted DNA. In some notable examples, however, the DNA is significantly deformed to accommodate the protein fold. The most dramatic example of this remains that of the TATA binding protein (TBP) (Figure 1L), which binds to DNA that is highly underwound and bent. Less dramatic but still significant DNA kinking and bending can be seen in complexes with CAP (Figure 1D) and BmrR (Figure 1E). While deforming the DNA requires energy, the cost can clearly be compensated by a sufficient number of favorable contacts with the protein. In most cases, the

[1] Correspondence: cwolberg@jhmi.edu

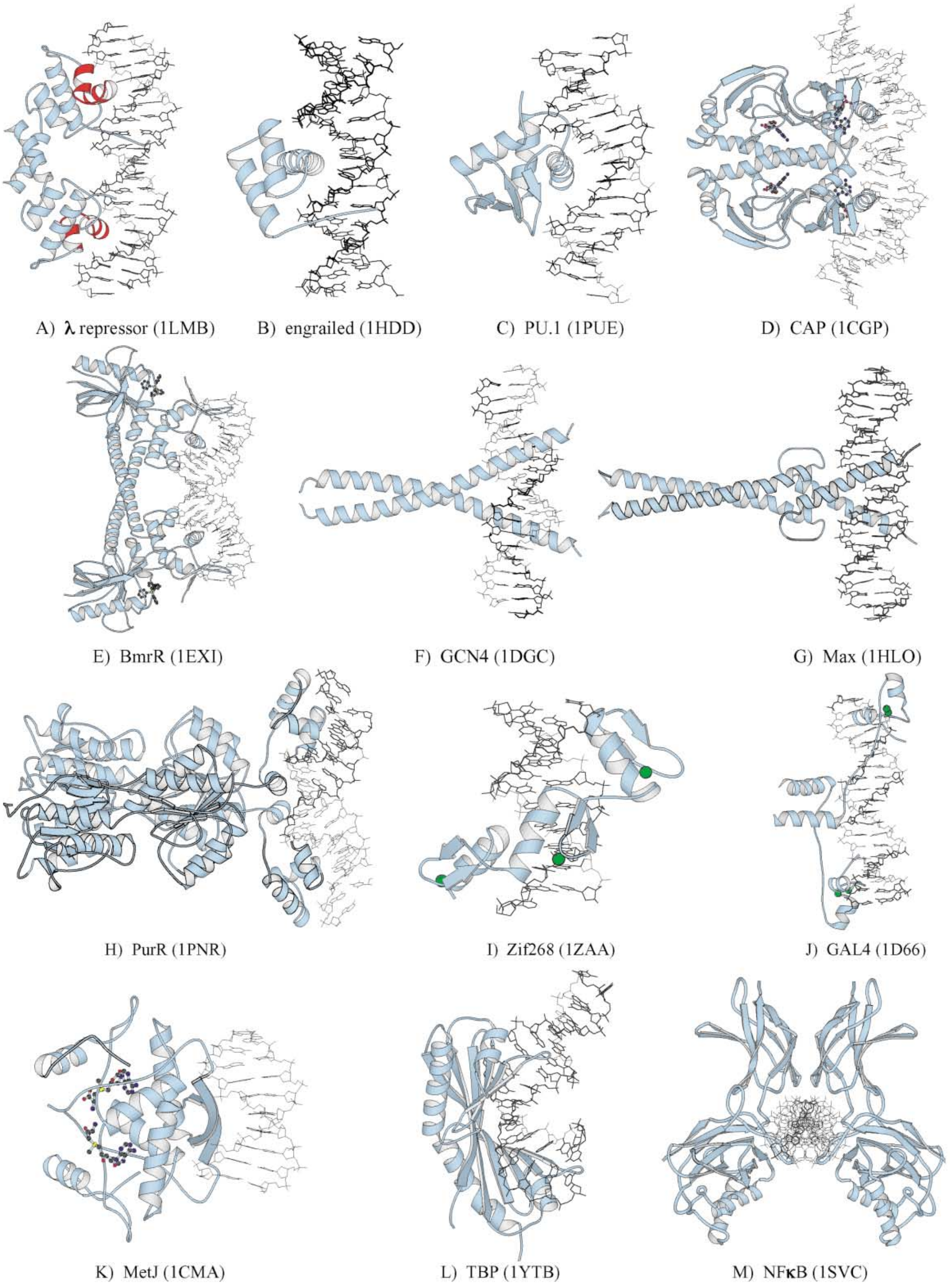


Figure 1. Representative Examples of Different DNA Binding Folds

PDB accession codes are shown in parentheses. (A) Bacterial helix-turn-helix protein: λ repressor. The helix-turn-helix is highlighted in red. (B) Homeodomain: engrailed. (C) Winged helix-turn-helix: PU.1 ETS domain. (D) Helix-turn-helix: CAP. (E) BmrR (unclassified). (F) bZIP: GCN4. (G) bHLH: Max. (H) LacI family member: PurR. (I) Zn finger: Zif268. (J) Zinc binding domain, GAL4 type: GAL4. (K) β sheet recognition: MetJ. (L) TATA binding protein (TBP). (M) Rel homology domain (immunoglobulin-type fold): NF-κB.

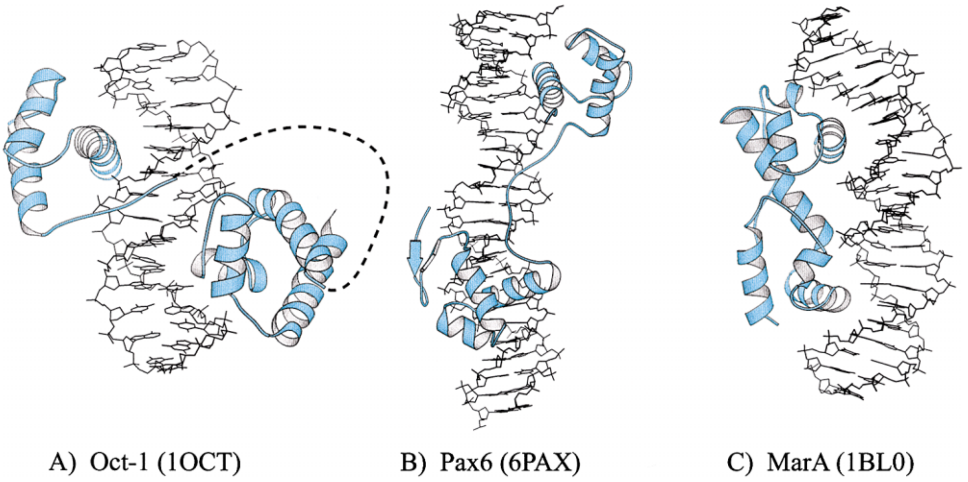


Figure 2. DNA Binding Motifs with More Than One Reading Head

PDB accession codes are shown in parentheses. (A) POU domain: Oct-1. (B) Paired domain: Pax6. (C) AraC family: MarA.

deformation appears to be incidental to complex formation, although in proteins such as BmrR the observed DNA distortions are thought to have functional significance for gene regulation (Heldwein and Brennan, 2001).

Overview of DNA Binding Protein Architecture

Proteins that recognize specific DNA sequences exhibit remarkably diverse architecture. The first three DNA binding proteins whose structures were determined, λ repressor (Pabo and Lewis, 1982), λ cro (Anderson et al., 1982), and CAP (McKay and Steitz, 1981), all contain a two-helix DNA binding motif termed the helix-turn-helix (Figure 1A), which made the design principles of DNA binding proteins appear deceptively simple. Since then, new DNA binding motifs have continued to be identified, revealing just how many different protein folds have evolved to bind DNA. A variety of classification schemes have been used to categorize the different folds. A recent analysis (Luscombe et al., 2001) identified more than 30 different families of DNA binding folds in transcriptional regulators that bind DNA and still others that are found in recombination proteins, endonucleases, methylases, and other classes of DNA binding proteins. Searches of sequence databases have shown that some of these folds, such as the Cys_2His_2 zinc finger and the homeodomain, are quite common, while others occur rarely or in a more restricted set of organisms. A representative sample of DNA binding domains is found in Figure 1.

DNA binding folds typically contain globular domains whose surface side chains interact with DNA. However, many motifs contain flexible segments of polypeptide chain that become ordered upon binding to DNA and mediate important base and phosphate backbone contacts. A number of folds have flexible N- or C-terminal tails that are unstructured in the absence of the DNA but bind in one of the DNA grooves. The λ repressor has an N-terminal arm that contacts bases in the major groove (Jordan and Pabo, 1988), while homeodomain proteins have N-terminal arms that dock in the minor groove of the DNA (Gehring et al., 1994a). These flexible arms work together with the globular DNA binding domain to add additional base specificity. The basic region-leucine zipper (bZIP) fold is an example of a motif in which the entire DNA "reading head" folds upon binding to DNA (Ellenberger et al., 1992; Weiss et al., 1990) (described in detail below).

The effective length of a DNA binding site depends upon the conformation and size of the DNA binding domain, the number of "reading heads", and whether the protein forms homodimers or oligomeric interactions with other DNA binding partners. The portion of the DNA contacted by a single DNA binding domain typically spans 4–10 base pairs. Some DNA binding motifs contain multiple independently folding protein domains and thereby span a larger site. The POU domain (Figure 2A) was first identified on the basis of sequence comparisons, but in fact contains two protein domains, the POU-specific domain and the POU homeodomain, that are connected by a linker region (Klemm et al., 1994). The linker in this case serves as a flexible tether between the two domains and can permit different relative arrangements of the subdomains on the DNA (Scully et al., 2000). The paired domain (Figure 2B) contains N- and C-terminal globular domains connected by a linker region that mediates important minor groove contacts (Xu et al., 1999). MarA (Figure 2C), a member of the AraC family, consists of a single globular domain containing two reading heads that insert into successive major grooves of the DNA (Rhee et al., 1998).

In the sections that follow, we present several examples of DNA binding proteins that participate in transcriptional regulation and describe how they bind to specific DNA sequences. For the sake of brevity, we do not cover all DNA binding folds, but rather have chosen examples that illustrate important architectural features of DNA binding domains.

Different Uses of α Helices

The α helix is the most common protein structural element used for base recognition, typically through contacts in the major groove. Early model building (Church

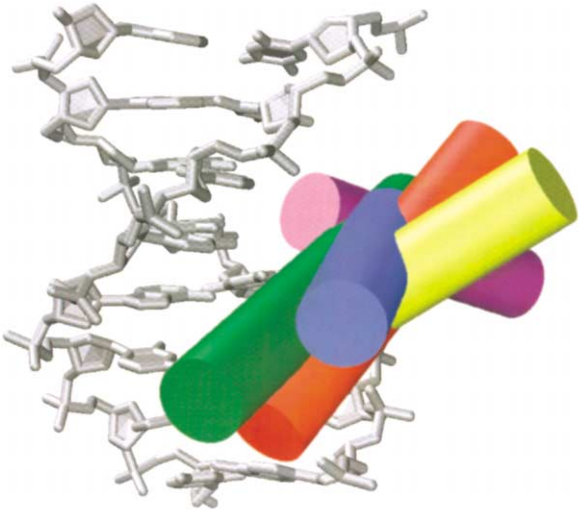


Figure 3. The Different Orientations of α Helices in the Major Groove of DNA

The position of the DNA recognition helix alone is shown for each of the structures indicated. Color key: red, IRF; blue, MATα2 homeodomain; purple, Trp repressor; yellow, Tox repressor; green, SAP-1.

et al., 1977) had shown that the proportions of the α helix were ideal for presenting side chains for interaction with bases in the major groove of undistorted B-DNA. The maximum number of DNA base contacts can be achieved when the α helix inserts into the major groove with its axis parallel to the flanking DNA backbone (Suzuki and Gerstein, 1995). This type of orientation is found in folds such as the homeodomain (Gehring et al., 1994b) (Figure 1B) and the basic region-leucine zipper proteins (Figure 1F). However, a survey of known DNA binding folds shows very wide variation in placement of the α helix in the major groove, as illustrated in Figure 3. These range from the tracking arrangement described above, to the end-on insertion of a helix into the major groove found in Trp repressor (Otwinowski et al., 1988) and in Cys_2 His_2 zinc finger proteins such as Zif268 (Pavletich and Pabo, 1991), to intermediate orientations used by members of the winged helix family (Gajiwala and Burley, 2000), CAP (Schultz et al., 1991), and other proteins. Helices can also insert into the minor groove, as occurs in the lac repressor family (Lewis et al., 1996; Schumacher et al., 1994). In these cases, kinking of the DNA is required to open the minor groove and thereby accommodate the helix. The examples that follow show how different DNA binding folds use α helices in conjunction with other structural elements to contact the DNA.

Helix-Turn-Helix Proteins

The proteins in this "family" in fact span a broad range of protein folds that contain a conserved bihelical motif termed the helix-turn-helix (HTH) (Figures 1A–1D), but are generally dissimilar in structure outside the HTH region. The two helices are related by a relatively fixed angle and are connected by a tight bend, although the length of each helix varies among different subclasses of this broadly defined family. The second of the two α helices, referred to as the recognition helix, inserts into the major groove and forms both base and sugar-phosphate backbone contacts. The first helix, while not embedded in the major groove, in some cases makes additional DNA contacts.

There is significant variation in how the recognition helices of different types of HTH proteins insert into the DNA. The λ repressor recognition helix is inserted into the major groove with its N-terminal portion embedded more deeply than the C-terminal end, which tilts away from the DNA (Figure 1A). Other bacterial and phage HTH proteins also insert the N-terminal halves of their recognition helices into the major groove. In the case of TrpR (Otwinowski et al., 1988), the recognition helix is introduced end-on into the DNA helix, sharply angling away from the DNA axis (Figure 3). By contrast, the homeodomain recognition helix "tracks" the major groove and is oriented nearly parallel to the neighboring sugar-phosphate backbone (Figures 1B and 3). The homeodomain HTH, both of whose helices are longer than their phage and bacterial counterparts, is also shifted somewhat relative to the DNA so that the central portion of the recognition helix contacts the DNA rather than its N terminus. Thus, while the helix-turn-helix in all cases is used to interact with the major groove, the docking arrangement on the DNA is governed by the remainder of the domain in which the HTH is embedded.

While the helix-turn-helix motif is responsible for many of the key DNA contacts made by proteins in this family, there are generally additional contacts made by other DNA binding domain residues that lie outside the helix-turn-helix region. As mentioned above, λ repressor makes additional major groove contacts with a flexible peptide "arm" located at the N terminus of the DNA binding domain (Figure 1A). A subclass of the HTH family known as the winged helix-turn-helix proteins are so named due to the presence of an additional wing immediately C-terminal to the HTH unit that mediates additional contacts with the DNA. An example can be seen in the ETS domain protein, PU.1 (Figure 1C) (Kodandapani et al., 1996).

Basic Region-Leucine Zipper and Helix-Loop-Helix Proteins

These two classes of dimeric eukaryotic DNA binding proteins have a common mechanism of binding DNA and distinct but related modes of dimerization. Basic region-leucine zipper (bZIP) proteins consist of long, uninterrupted α helices of about 60 residues. These proteins associate via their C-terminal halves, forming a parallel coiled-coil with leucine residues at the hydrophobic dimer interface (O'Shea et al., 1991), while the N-terminal portions of the dimerized helices splay out and insert into the major groove on either side of the DNA (Ellenberger et al., 1992) (Figure 1F). A striking feature of these proteins is that the helical structure of the entire DNA sequence reading head is coupled to DNA binding, as these residues are unstructured in the absence of DNA (Weiss et al., 1990). The basic region-helix-loop-helix (bHLH) proteins share with the bZIP proteins a very similar mode of DNA binding (Figure 1G). The salient difference lies in the dimerization region, which is composed of two helices separated by a loop. The HLH portions of the protein associate to form a four-helix bundle. Some bHLH proteins, such as Max (Ferre-D'Amare et al., 1993) (Figure 1G), are followed by a leucine zipper dimerization region, while others such as E47 (Ellenberger et al., 1994) and MyoD (Ma et al.,

1994) lack the leucine zipper region. Both bZIP and bHLH proteins have many members that form both homodimers and heterodimers, a feature that expands the repertory of DNA sequences that the proteins can recognize.

Lac and Purine Repressor: α Helices in the Minor Groove

The lacI family of proteins provides an example of how the minor groove can, in fact, accommodate an α helix, provided the DNA is sufficiently distorted. As first seen in the structure of the purine repressor dimer (PurR) bound to DNA (Schumacher et al., 1994), each monomer contains two separate modules that contact DNA: a helix-turn-helix "headpiece" that contacts bases in the major groove and a two-turn "hinge" helix that contacts bases in the minor groove (Figure 1H). The hinge helices from each monomer associate via hydrophobic interactions and insert into the minor groove at the center of the dyad-symmetric binding site. Since the minor groove dimensions of undistorted B-DNA are too small to accommodate one, let alone two α helices, the DNA becomes kinked and underwound at the point of insertion. A 45° kink that bends the DNA away from the protein, together with unwinding and base unstacking at the central base step, opens the minor groove sufficiently to enable residues in the hinge helix pair to form direct contacts with bases in the minor groove. The kink is facilitated by intercalation of leucine side chains that help pry apart the central base step. A further example of intercalating hydrophobic side chains that promote DNA distortion will be seen below in the case of the TATA binding protein (TBP), and they are also found among the non-sequence-specific, "architectural" DNA binding proteins belonging to the HMG box family (Murphy et al., 1999).

Zinc-Coordinating Proteins

Protein domains with one or more coordinated zinc ions at their core form a superfamily of eukaryotic DNA binding proteins. In all cases, the zinc serves a structural role in maintaining the protein fold and does not interact with the DNA. It is important to note that the various families of zinc-coordinating DNA binding domains differ significantly in overall protein fold and DNA binding, and it is therefore more useful to consider the individual types of zinc binding domains. Prominent among these is the zinc finger family, the most abundant class of DNA binding proteins in the human genome (Lander et al., 2001), whose members contain multiple copies of a compact, ~30-amino acid DNA binding domain (Figure 1I). This domain is also referred to as the TFIIIA/zif268, or Cys_2His_2-type zinc finger. It is the most minimal of DNA binding domains: a relatively short α helix, two antiparallel strands of β sheet, and a core Zn^{2+} ion coordinated by two cysteine and two histidine residues (Pavletich and Pabo, 1991). Proteins that bind DNA with this motif typically have multiple copies of the zinc finger domain connected by short linker regions, with two fingers being the minimal length necessary for DNA binding. Canonical fingers bind DNA by inserting the α helix end-on into the major groove (Figure 1I), recognizing a 3–4 base pair site. In the case of proteins such as zif268 (Pavletich and Pabo, 1991) and Gli (Pavletich and Pabo, 1993), successive fingers track the DNA major groove, with the center-to-center spacing between each finger's subsite determined by the length of the linker region (Wolfe et al., 2000). However, the structure of a six-finger fragment of TFIIIA shows that some of the fingers serve a spacer-like function, straddling the minor groove of the DNA and connecting flanking fingers that bind in the major groove of DNA in the canonical manner (Nolte et al., 1998).

Other families of Zn^{2+}-coordinating DNA binding domains bind DNA as dimers and do not have the modular design of the zinc finger family. One example is the nuclear hormone receptor family (Mangelsdorf and Evans, 1995), whose DNA binding domains contain a zinc ion coordinated by four cysteines. These proteins, which bind DNA by inserting an α helix into the major groove (Figure 5A), bind DNA as either homodimers or heterodimers. The GAL4-type protein (Figure 1J) contains two structural Zn^{2+} ions per DNA binding domain and also inserts a helix into the major groove (Marmorstein et al., 1992). The DNA binding domain is joined via a linker to a leucine zipper dimerization motif, illustrating how a conserved dimerization motif can be used to mediate multimerization of different types of DNA binding domains.

DNA Recognition with β Sheets

Although the use of β sheets to mediate DNA contacts is not nearly as prevalent as the use of α helices, several examples exist. The ribbon-helix-helix proteins, exemplified by the MetJ (Somers and Phillips, 1992) and arc (Raumann et al., 1994) repressors, form dimers in which each monomer donates a single strand to a two-stranded antiparallel β sheet. The dimer β sheet inserts into the major groove with the side chains on the face of the β sheet contacting the base pairs and the strands parallel to the flanking sugar-phosphate backbone (Figure 1K). Proteins from the ribbon-helix-helix family bind cooperatively to two or more adjacent DNA binding sites, with the types of cooperative interactions and the relative arrangement of dimers on the DNA varying among different family members (Gomis-Ruth et al., 1998; Somers and Phillips, 1992). The insertion of three-stranded β sheets into the major groove of the DNA has been observed in the plant GCC box binding domain (Allen et al., 1998). Because of the staggered arrangement of the strands, though, there are at any location only two strands inserting into the major groove, thereby allowing the DNA to accommodate the protein with little distortion.

TATA binding proteins (TBP) use a large β sheet surface to recognize DNA sequence by binding in the minor groove (Kim et al., 1993a, 1993b). In contrast to β sheet recognition in the major groove, which is accompanied by a moderate degree of DNA bending, insertion of the concave, ten-stranded β sheet of TBP into the minor groove requires profound DNA distortion (Figure 1L). The DNA undergoes dramatic unwinding and bending that makes possible contacts between the protein's concave surface and the edges of the base pairs in the otherwise recessed minor groove. Phenylalanine side chains that intercalate into the DNA promote the DNA distortion, reminiscent of PurR-DNA binding.

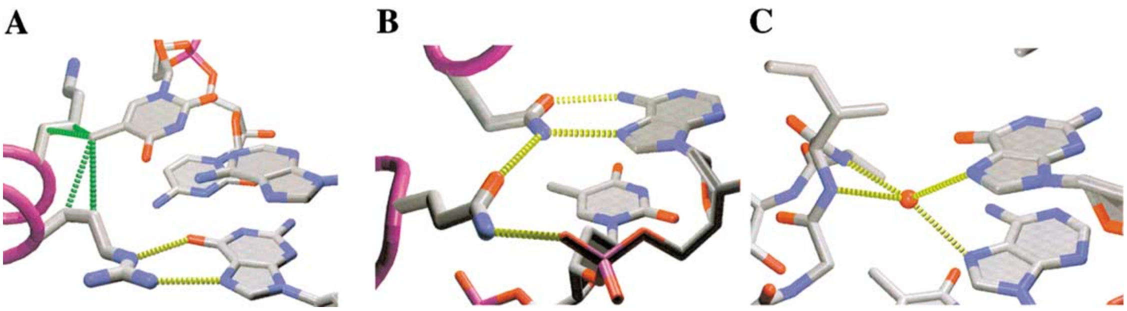


Figure 4. Details of Protein-DNA Contacts

(A) Bidentate contacts between arginine side chain and guanine base (yellow dashed lines) and hydrophobic contacts to a thymine methyl (green dashed lines).

(B) Bidentate contact between glutamine and adenine in the λ repressor-DNA complex. In addition to contacting the adenine, this side chain hydrogen bonds to a second glutamine side chain which in turn contacts a phosphate group.

(C) Water-mediated hydrogen bonds at the protein-DNA interface of the Trp repressor-DNA complex.

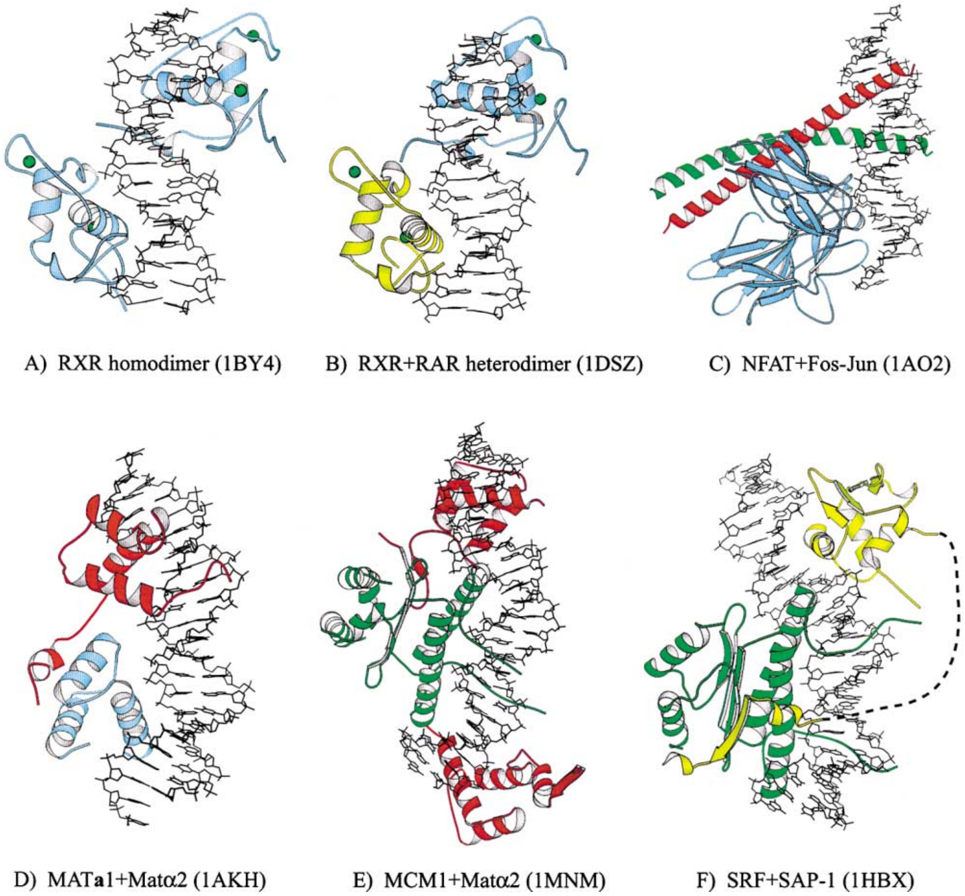
DNA Recognition with Loops

While α helices and β sheets provide a relatively rigid scaffold for side chain and main chain interaction with DNA, a superfamily of DNA binding proteins that contain an immunoglobulin-like fold use loops as the primary structural element for DNA contacts. Classified in the SCOP database (Lo Conte et al., 2000) as the p53-like transcription factors, the common element among the various subfamilies such as the Rel homology domain (Ghosh et al., 1995; Muller et al., 1995) (Figure 1M), runt domain (Tahirov et al., 2001), STAT (Chen et al., 1998c), and p53 (Cho et al., 1994) proteins is the β sheet immunoglobulin-like domain. This is one of the more diverse superfamilies of proteins, as the various subfamilies diverge significantly in structure outside of the immunoglobulin-like domain, and there is low sequence conservation in the Ig-like domain itself. Moreover, while a general feature of these proteins is DNA recognition with loops, there is much variation in the orientation of these β sandwich domains on the DNA and in the way they form base contacts. This domain can also serve multiple roles. In the case of the Rel-homology domain proteins such as NF-κB p50 (Figure 1M), there are two Ig-like domains in each monomer: the N-terminal domain mediates DNA contacts primarily in the major groove, while the C-terminal domain mediates homo- and heterodimer interactions in addition to contacting DNA (Chen et al., 1998a; Ghosh et al., 1995; Muller et al., 1995). The side chains involved in dimer interactions lie along one face of the β sandwich, leaving the loops free to contact the DNA. The DNA binding regions of other Ig-like proteins such as p53, Stat-1, and NFAT most closely resemble the N-terminal domain of the Rel proteins.

DNA Contacts and Specificity of Binding

The different types of DNA binding folds described above represent different evolutionary design solutions to the problem of how to present a set of side chains for contacts with the DNA. How do side chain contacts enable a protein to select a particular DNA sequence from among all possible binding sites in the genome? When a protein binds to its preferred sequence, it can form an optimal number of contacts with the base pairs and backbone. While the DNA backbone conformation is somewhat dependent upon sequence, it is predominantly the characteristic chemical signature of each base pair that is recognized by the protein. The great majority of proteins recognize functional groups in the major groove of the DNA, as it is here that each base pair can be uniquely distinguished. The pattern of hydrogen bond donors and acceptors is less varied in the minor groove, with A·T similar to T·A and G·C similar to C·G.

As many have pointed out, there is no simple protein code for base recognition (Mandel-Gutfreund et al., 1995; Matthews, 1988; Pabo and Nekludova, 2000), no particular set of contacts that universally specify a particular base sequence. Given the variety of protein folds, the flexibility of many side chains, and the modest energetic cost of small DNA distortions, there are simply many different ways to form a protein surface that is chemically complementary to DNA of a particular sequence. Yet despite the lack of a set of simple rules governing sequence recognition, some principles and common themes have emerged (Kono and Sarai, 1999; Luscombe et al., 2001; Mandel-Gutfreund et al., 1995; Pabo and Nekludova, 2000). It is thought that the bulk of the sequence specificity comes from hydrogen bonding interactions between the protein and the DNA, due to the requirement for near colinear apposition of donor and acceptor groups. Bidentate interactions, in which a single side chain forms two hydrogen bonds with the DNA, confer a even higher degree of specificity than single hydrogen bonds to a side chain. For example, arginine can recognize a guanine base through bidentate interactions of the Nϵ and Nζ of lysine with the N7 and O6 of guanine (Figure 4A). There is no other base that contains two hydrogen bond acceptors in the major groove, so no base substitutions can be made at this position without reducing DNA binding affinity. Indeed, arginine-guanine pairings are common in protein-DNA complexes, as are lysine-guanine (Luscombe et al., 2001). To a somewhat lesser extent, adenine is often contacted by either asparagine or glutamine, both of which can donate a hydrogen bond to the N6 and accept a hydrogen bond from the N6 (Figure 4B). These common pairings, however, have no predictive power: each of the side chains mentioned above has been observed to contact all four base pairs. Van der Waals interactions,



A) RXR homodimer (1BY4)   B) RXR+RAR heterodimer (1DSZ)   C) NFAT+Fos-Jun (1AO2)

D) MAT**a**1+Matα2 (1AKH)   E) MCM1+Matα2 (1MNM)   F) SRF+SAP-1 (1HBX)

Figure 5. Multiprotein Complexes

PDB accession codes are shown in parentheses. (A) Nuclear hormone receptor homodimer: RXR. (B) Nuclear hormone receptor heterodimer: RXR (blue)-RAR (yellow). (C) Heterotetrameric complex: FOS (red)-JUN (green) heterodimer and NFAT (blue). (D) Homeodomain heterodimer: MAT**a**1 (blue)-MATα2 (red). (E) Heterotetramer formed by the dimeric MADS box protein MCM1 (green) with two copies of the homeodomain protein MATα2 (red). (F) Interaction between the MADS box protein SRF (green) and the ETS domain protein SAP-1 (yellow).

because of their lack of directional requirements, are thought to play a lesser role in specificity. Nevertheless, the high proportion of van der Waals interactions found at most protein-DNA interfaces (Luscombe et al., 2001) imposes steric constraints on the types of side chains and bases that can be accommodated at particular positions, thereby also playing a role in sequence selectivity. In particular, van der Waals interactions between protein side chains and the methyl group of thymine (Figure 4A) have been observed to play an important role in sequence specificity in a number of cases.

It is important to note that there is no fixed geometry by which a side chain can contact a base. Whether a side chain forms one or two hydrogen bonds with a base, hydrogen bonds to a water molecule, or participates in DNA backbone contacts, is strongly determined by the orientation in which the protein presents the side chain to the DNA and on the neighboring side chains (Pabo and Nekludova, 2000). Surrounding side chains can also have an impact, as there are many examples of DNA-contacting side chains that form additional hydrogen bonds with one another (Figure 4B). It is only when one compares closely related proteins that some common elements of recognition emerge. For example, glutamine is used to recognize adenine (Figure 4B) in an identical manner in both the λ repressor and 434 repressor complexes, since both dock on the DNA in a similar manner (Pabo et al., 1990). Similarly, the zif268-type zinc fingers bind DNA in a highly conserved manner, and some stronger trends are observed in the use of particular side chains to specify base identity at certain positions (Pabo and Nekludova, 2000). In general, members of a particular DNA binding fold family will have a characteristic manner of docking on the DNA, although the docking must be highly conserved in order for particular residues to play conserved roles in base recognition. There are, however, exceptions to the general rule: the hRFX1 protein, while a member of the winged helix family, binds DNA in a very different manner than all other known members of this fold family (Gajiwala et al., 2000).

The additional feature of bound water at protein-DNA interfaces expands the types of interactions possible between protein and DNA. Water-mediated hydrogen bonds are quite common at protein-DNA interfaces (Figure 4C), although their role in specificity has remained unclear. It has been proposed (Luscombe et al., 2001) that water frequently appears to serve as "filler" at the protein-DNA interface, occupying positions at the interface that would otherwise contain empty holes, which are energetically unfavorable. Another analysis of a limited set of protein-DNA complexes, however, suggested that some of the bound water molecules that hydrogen bond directly to bases are present even in the absence of protein and that the protein thereby appears to be recognizing a hydrated DNA structure (Woda et al., 1998). In select cases, notably that of Trp repressor, it appears that networks of water play a role in an indirect readout mechanism of DNA sequence recognition (Lawson and Carey, 1993).

Role of Multimerization and Cooperative Interactions

The ability to multimerize or bind DNA cooperatively with diverse partners expands the sequence recognition and regulatory possibilities of transcriptional regulators (Wolberger, 1999). For proteins that can bind DNA as either homo- or heterodimers, such as members of the nuclear hormone receptor family (Figure 5A and 5B) (Mangelsdorf and Evans, 1995), the different combinations of protein partners make possible recognition of a diverse set of DNA sequences. The added favorable energy of cooperative interactions also provides a way to assemble a tightly binding multimeric complex out of proteins that, on their own, bind DNA with low specificity or affinity, as is the case for the MATα2 and MAT**a**1 proteins (Goutte and Johnson, 1993). Similarly, cooperative interactions can make possible recruitment of proteins to suboptimal DNA sites (Fitzsimmons et al., 1996) in addition to uniquely positioning partners on the DNA (Chen et al., 1995).

The structural determinates for interactions between DNA-bound partners may lie within the DNA binding domain itself, as is the case in the NFAT/Fos-Jun complex (Figure 5C) (Chen et al., 1998b), or may be mediated by peptides or domains outside the conserved DNA binding domain. The MATα2 homeodomain protein, for example, can recruit either the MATa1 homeodomain protein (Li et al., 1995) or the MCM1 MADS box protein (Tan and Richmond, 1998), using a C-terminal tail to recruit MAT**a**1 (Figure 5D) and an N-terminal linker region to complex with MCM1 (Figure 5E). The latter interaction is strikingly similar to that between the SRF MADS box and the SAP-1 ETS domain protein (Figure 5F). In all of the aforementioned cases, the protein-protein interfaces serve simply to foster interactions between neighboring partners, without affecting the DNA-sequence selectivity of the individual partners. In the complex formed by Ets-1 and Pax5 on DNA, however, Pax5 alters the DNA contacts made by Ets-1 through direct interactions with the Ets-1 recognition helix (Garvie et al., 2001). The frequency with which this mechanism is used to regulate sequence recognition awaits further structural study of other classes of interacting proteins.

Conclusions

The solutions to the problem of how an organism evolves a protein to recognize a specific sequence of DNA turn out to be many and varied. This may be the result of other functional requirements that are related to regulation of DNA binding activity and the need to associate with other proteins that may or may not bind DNA. However, it also appears that the structural and "design" requirements for forming a complementary protein-DNA interface are simply not that great, making it easy to adapt a variety of protein folds to the task of binding to and recognizing particular DNA sequences. The examples we have outlined here illustrate many of the types of structural elements that can be used for DNA binding. While there is little doubt that there are still more DNA binding folds whose structures remain to be elucidated, we are likely to see recurring use of many of the same basic elements already seen in the sequence-specific DNA binding proteins that have been characterized to date.

References

Allen, M.D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., and Suzuki, M. (1998). A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. EMBO J. *17*, 5484–5496.

Anderson, W.F., Ohlendorf, D.H., Takeda, Y., and Matthews, B.W. (1982). Structure of the cro repressor from bacteriophage λ and its interaction with DNA. Nature *290*, 754–758.

Chen, L., Oakley, M.G., Glover, J.N., Jain, J., Dervan, P.B., Hogan, P.G., Rao, A., and Verdine, G.L. (1995). Only one of the two DNA-bound orientations of AP-1 found in solution cooperates with NFATp. Curr. Biol. *5*, 882–889.

Chen, F.E., Huang, D.B., Chen, Y.Q., and Ghosh, G. (1998a). Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA. Nature *391*, 410–413.

Chen, L., Glover, J.N., Hogan, P.G., Rao, A., and Harrison, S.C. (1998b). Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. Nature *392*, 42–48.

Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell, J.E., Jr., and Kuriyan, J. (1998c). Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. Cell *93*, 827–839.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346–355.

Church, G.M., Sussman, J.L., and Kim, S.H. (1977). Secondary structural complementarity between DNA and proteins. Proc. Natl. Acad. Sci. USA *74*, 1458–1462.

Ellenberger, T.E., Brandl, C.J., Struhl, K., and Harrison, S.C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. Cell *71*, 1223–1237.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. *8*, 970–980.

Ferre-D'Amare, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363*, 38–45.

Fitzsimmons, D., Hodsdon, W., Wheat, W., Maira, S.M., Wasylyk, B., and Hagman, J. (1996). Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. Genes Dev. *10*, 2198–2211.

Gajiwala, K.S., and Burley, S.K. (2000). Winged helix proteins. Curr. Opin. Struct. Biol. *10*, 110–116.

Gajiwala, K.S., Chen, H., Cornille, F., Roques, B.P., Reith, W., Mach, B., and Burley, S.K. (2000). Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. Nature *403*, 916–921.

Garvie, C., Hagman, J., and Wolberger, C. (2001). Structural studies of Ets-1/Pax-5 complex formation on DNA. Mol. Cell, in press.

Gehring, W.J., Affolter, M., and Bürglin, T. (1994a). Homeodomain proteins. Annu. Rev. Biochem. *63*, 487–526.

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wutrich, K. (1994b). Homeodomain-DNA recognition. Cell *78*, 211–223.

Ghosh, G., van Duyne, G., Ghosh, S., and Sigler, P.B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. Nature *373*, 303–310.

Gomis-Ruth, F.X., Sola, M., Acebo, P., Parraga, A., Guasch, A., Eritja, R., Gonzalez, A., Espinosa, M., del Solar, G., and Coll, M. (1998). The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. EMBO J. *17*, 7404–7415.

Goutte, C., and Johnson, A.D. (1993). Yeast **a**1 and α2 homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. J. Mol. Biol. *233*, 359–371.

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

Heldwein, E.E., and Brennan, R.G. (2001). Crystal structure of the transcription activator BmrR bound to DNA and a drug. Nature *409*, 378–382.

Jordan, S.R., and Pabo, C.O. (1988). Structure of the lambda complex at 2.5 Å resolution: Details of the repressor-operator interactions. Science *242*, 893–899.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature *365*, 520–527.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512–520.

Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., and Pabo, C.O. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell *77*, 21–32.

Kodandapani, R., Pio, F., Ni, C.Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R.A., and Ely, K.R. (1996). A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. Nature *380*, 456–460.

Kono, H., and Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. Proteins *35*, 114–131.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lawson, C.L., and Carey, J. (1993). Tandem binding in crystals of a trp repressor/operator half-site complex. Nature *366*, 178–182.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science *271*, 1247–1254.

Li, T., Stark, M.R., Johnson, A.D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science *270*, 262–269.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. (2000). SCOP: a structural classification of proteins database. Nucleic Acids Res. *28*, 257–259.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol *1*, REVIEWS001.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.

Ma, P.C., Rould, M.A., Weintraub, H., and Pabo, C.O. (1994). Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. Cell *77*, 451–459.

Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J. Mol. Biol. *253*, 370–382.

Mangelsdorf, D.J., and Evans, R.M. (1995). The RXR heterodimers and orphan receptors. Cell *83*, 841–850.

Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. Nature *356*, 408–414.

Matthews, B.W. (1988). Protein-DNA interaction. No code for recognition. Nature *335*, 294–295.

McKay, D.B., and Steitz, T. (1981). Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. Nature *290*, 744–749.

Muller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., and Harrison, S.C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. Nature *373*, 311–317.

Murphy, F.V., IV, Sweet, R.M., and Churchill, M.E. (1999). The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition. EMBO J. *18*, 6610–6618.

Nolte, R.T., Conlin, R.M., Harrison, S.C., and Brown, R.S. (1998). Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. Proc. Natl. Acad. Sci. USA *95*, 2938–2943.

O'Shea, E.K., Klemm, J.D., Kim, P.S., and Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science *254*, 539–544.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321–329.

Pabo, C.O., and Lewis, M. (1982). The operator-binding domain of λ repressor: Structure and DNA recognition. Nature *298*, 443–447.

Pabo, C.O., and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J. Mol. Biol. *301*, 597–624.

Pabo, C.O., Aggarwal, A.K., Jordan, S.R., Beamer, L.J., Obeysekare, U.R., and Harrison, S.C. (1990). Conserved residues make similar contacts in two repressor-operator complexes. Science *247*, 1210–1213.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809–817.

Pavletich, N.P., and Pabo, C.O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science *261*, 1701–1707.

Raumann, B.E., Rould, M.A., Pabo, C.O., and Sauer, R.T. (1994). DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. Nature *367*, 754–757.

Rhee, S., Martin, R.G., Rosner, J.L., and Davies, D.R. (1998). A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. Proc. Natl. Acad. Sci. USA *95*, 10413–10418.

Schultz, S.C., Shields, G.C., and Steitz, T.A. (1991). Crystal structure of a CAP-DNA complex: The DNA is bent by 90°. Science *253*, 1001–1007.

Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science *266*, 763–770.

Scully, K.M., Jacobson, E.M., Jepsen, K., Lunyak, V., Viadiu, H., Carriere, C., Rose, D.W., Hooshmand, F., Aggarwal, A.K., and Rosenfeld, M.G. (2000). Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. Science *290*, 1127–1131.

Somers, W.S., and Phillips, S.E. (1992). Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. Nature *359*, 387–393.

Suzuki, M., and Gerstein, M. (1995). Binding geometry of alpha-helices that recognize DNA. Proteins *23*, 525–535.

Tahirov, T.H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., et al. (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell *104*, 755–767.

Tan, S., and Richmond, T.J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature *391*, 660–666.

Weiss, M.A., Ellenberger, T., Wobbe, C.R., Lee, J.P., Harrison, S.C., and Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature *347*, 575–578.

Woda, J., Schneider, B., Patel, K., Mistry, K., and Berman, H.M. (1998). An analysis of the relationship between hydration and protein-DNA interactions. Biophys. J. *75*, 2170–2177.

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

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct *29*, 183–212.

Xu, H.E., Rould, M.A., Xu, W., Epstein, J.A., Maas, R.L., and Pabo, C.O. (1999). Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. Genes Dev. *13*, 1263–1275.