Origins of Specificity in Protein-DNA Recognition

Remo Rohs,[1,2,*] Xiangshu Jin,[1,2,*] Sean M. West,[1,2] Rohit Joshi,[2] Barry Honig,[1,2] and Richard S. Mann[2]

[1]Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics, [2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032; email: bh6@columbia.edu, rsm10@columbia.edu

Annu. Rev. Biochem. 2010. 79:233–69

First published online as a Review in Advance on March 24, 2010

The *Annual Review of Biochemistry* is online at biochem.annualreviews.org

This article's doi:
10.1146/annurev-biochem-060408-091030

0066-4154/10/0707-0233$20.00

*These authors contributed equally to this work.

Key Words

protein-DNA binding, direct readout, indirect readout, DNA base recognition, DNA shape recognition, narrow minor groove

Abstract

Specific interactions between proteins and DNA are fundamental to many biological processes. In this review, we provide a revised view of protein-DNA interactions that emphasizes the importance of the three-dimensional structures of both macromolecules. We divide protein-DNA interactions into two categories: those when the protein recognizes the unique chemical signatures of the DNA bases (base readout) and those when the protein recognizes a sequence-dependent DNA shape (shape readout). We further divide base readout into those interactions that occur in the major groove from those that occur in the minor groove. Analogously, the readout of the DNA shape is subdivided into global shape recognition (for example, when the DNA helix exhibits an overall bend) and local shape recognition (for example, when a base pair step is kinked or a region of the minor groove is narrow). Based on the >1500 structures of protein-DNA complexes now available in the Protein Data Bank, we argue that individual DNA-binding proteins combine multiple readout mechanisms to achieve DNA-binding specificity. Specificity that distinguishes between families frequently involves base readout in the major groove, whereas shape readout is often exploited for higher resolution specificity, to distinguish between members within the same DNA-binding protein family.

Contents

1. INTRODUCTION

Genomes are composed of both protein-coding and nonprotein-coding DNA sequences. Cells have the remarkable ability to decipher the information that is incorporated in both types of sequences. Biologists, on the other hand, are currently unable to do what the cell does—to interpret nonprotein-coding DNA sequences. An important step toward achieving this goal is to have a better understanding of protein-DNA recognition mechanisms. Traditionally, the analysis of noncoding DNA sequences has treated DNA as a linear string of nucleotides, which does not take into account the three-dimensional structure of DNA. In this review, we provide a new perspective on the problem of protein-DNA recognition, one that emphasizes the three-dimensional structures of both the DNA and the protein.

1.1. General Comments

More than 50 years after the structure of DNA was first proposed by Watson & Crick (1), biologists are still working to achieve a complete understanding of how proteins interact with genomes. One of the most important questions that remain is one of specificity—how do the large and diverse number of DNA-binding proteins encoded by eukaryotic genomes recognize their specific binding sites? Moreover, most DNA-binding proteins are part of large families that share DNA-binding domains with very similar biochemical properties. How do proteins with closely related DNA-binding domains carry out their unique functions in vivo? Providing answers to these questions is especially timely given the need to accurately annotate the many complete genome sequences that are now available, an endeavor that is still a major unsolved challenge.

The size and complexity of this problem has recently been underscored by several publications that use high-throughput approaches, such as protein-binding microarrays or the bacterial one-hybrid system, to generate an unprecedented database of the DNA sequence preferences for a large number of DNA-binding proteins (2–5). In one such recent report (6), the binding-site preferences for 104 mouse transcription factors, often including multiple members from the same transcription factor family, were described. To highlight just one example, the DNA-binding site preferences for 21 members of the Sox (SRY-related high-mobility group box)/TCF (T cell factor) family of transcriptional regulators were compared. Remarkably, although each factor executes unique functions, 14 of the 21 prefer to bind the sequence ACAAT. Moreover, although small differences in sequence preference were identified, these did not always correlate with the extent of sequence identity of the DNA-binding domains. For example, Sox1 preferred the sequence ATTTAAAT, whereas its two most closely related relatives (Sox14 and Sox21), as well as a much more distantly related family member, sex-determining region Y (SRY), preferred the sequence ACAAT. This study also revealed that many transcription factors have the capacity to recognize two distinct binding sites (so-called primary and secondary binding sites) and that there is a previously underappreciated interdependence between neighboring base pairs within a binding site.

Observations such as these raise a number of fundamental questions regarding protein-DNA recognition whose answers require a better understanding of the rules that govern how proteins bind to DNA sequences. We suggest that the linear sequence of base pairs in a binding site is only a small part of the story and that the three-dimensional structures of both macromolecules must be taken into account to fully understand protein-DNA recognition. In particular, local variations in DNA structure—DNA topography—may be as important as protein structure. A recent study that examined the evolutionary constraints on DNA topology strongly supports this point of view (7). Remarkably, the authors found that DNA topography of the human genome, as measured by hydroxyl radical cleavage patterns, is evolutionarily constrained. Moreover, these cleavage patterns, which are correlated with the solvent accessibility of the DNA helix (8), were found to be a much better predictor of functional DNA elements than the linear DNA sequence (7). Thus, to more fully understand the rules that govern protein-DNA recognition, we must consider both DNA structure and protein structure as equal partners.

1.2. Previous Definitions: Direct versus Indirect Readout Mechanisms

Understanding how proteins recognize their DNA-binding sites has a long history. Initially, on the basis of early low-resolution X-ray structures of nucleic acid duplexes (9), it was realized that the major groove of the DNA helix offered a set of base-specific hydrogen bond donors, acceptors, and nonpolar groups that could be recognized by a complementary set of donors and acceptors presented by amino acid side chains (10). Accordingly, the idea soon evolved that short DNA sequences could serve as binding sites that were specifically read by a complementary sequence of amino acids (11). This mechanism of protein-DNA recognition, now commonly referred to as direct readout, is evident in nearly all of the >1500 structures of protein-DNA complexes that have been solved and deposited in the Protein Data Bank (PDB). Nevertheless, as was realized many years ago (12), there is not a simple recognition code or one-to-one correspondence between DNA and protein sequences. Thus, direct readout, by itself, cannot be sufficient to account for the specificities of protein-DNA interactions.

Although elements of direct readout contribute to nearly all protein-DNA complexes, these structures also reveal that bound DNA frequently deviates from a standard B-form double helix. In some cases, deviations from a B-form helix are large and clearly contribute to DNA-binding specificity [e.g., the papillomavirus E2 protein and the TATA box-binding protein (TBP)] (13–15). In these cases, a bend or some other deformation of the DNA helix is required to establish a set of hydrogen bonds or nonpolar interactions between the protein and DNA that are much less likely to occur in the absence of the deformation. From such observations, the term indirect readout was coined (12). Indirect readout is defined as protein-DNA interactions that depend on base pairs that are not directly contacted by the protein (16). This broad definition includes situations where the DNA sequence creates or facilitates a DNA structure that is subsequently recognized by a protein, but also when the protein-DNA contact is mediated by a water molecule. In addition, over time, the term has been taken to mean any interaction between DNA and protein where the DNA is not a B-form helix. This even looser definition has limited value because it simply encompasses all interactions that are not direct.

1.3. Goals for this Review

In this review, we reevaluate the mechanisms that underlie protein-DNA recognition in light of new and previous structures of protein-DNA complexes. We suggest that the terms direct and indirect readout both describe idealized extremes that rarely exist in isolation in real protein-DNA complexes and therefore have limited value. For example, rarely are direct hydrogen bonds formed between protein side chains and DNA in the complete absence of any deviation from an ideal B-form helix. Conversely, rarely are protein-DNA interactions purely indirect. As detailed below, this reevaluation suggests that protein-DNA recognition utilizes a continuum of readout mechanisms that depend on the structural features and flexibility of both macromolecules, including the sequence-dependent propensity of DNA to assume conformations that deviate from ideal B-DNA. This more nuanced view suggests that protein-DNA and protein-protein recognition are in many ways analogous phenomena.

In order to reassess protein-DNA readout mechanisms, we divide this review into three main sections. In the first, we briefly discuss the range of protein structures that bind DNA. Because there are excellent recent reviews that already cover this topic (17–19), we simply summarize the major protein superfamiles that are observed in DNA-binding proteins. Second, because interactions between proteins and DNA depend on the interplay between both macromolecules, we review how DNA structures vary and the relationships between these structures and DNA sequence. Finally, with these structural considerations as a background, we review the range of interactions that are observed at protein-DNA interfaces, identifying common themes that are used both across and within individual families of DNA-binding proteins. We propose replacing the terms direct readout and indirect readout with the more informative terms, base readout and shape readout, which we further subdivide to reflect the way proteins recognize DNA sequences. Our goal is to present a richer and more subtle view of protein-DNA recognition that more accurately reflects the way in which evolution has fine-tuned these essential interactions.

Because the perspective offered here is structural in its origins, we do not review thermodynamic measurements of protein-DNA interactions nor do we summarize the many insights available from the application of simulation methodologies to the recognition problem (20). Rather, our goal is to review recent structural evidence regarding readout mechanisms of DNA sequences, recognizing that a deeper understanding of the underlying forces and their interactions requires the application of a variety of experimental and computational approaches to specific systems and on a genome-wide scale. It is our hope that the presentation and integration of structural data presented in this review serves to facilitate and to focus such studies.

2. STRUCTURE OF DNA-BINDING PROTEINS

The first protein-DNA complexes for which structural information was derived from X-ray crystallography were the catabolite gene activator protein (CAP) (21), Cro repressor (22), and λ repressor (23) bound to their binding sites. Since then, more than 1500 structures of protein-DNA complexes have been deposited in the Protein Data Bank.

Proteins utilize a wide range of DNA-binding structural motifs, such as the helix-turn-helix (HTH) motif of homeodomains, to recognize DNA. Many proteins also contain flexible segments outside a globular core that mediate important specific and nonspecific interactions. For example, λ repressor has an N-terminal arm that contacts bases in the major groove (24), the phage Φ29 transcriptional regulator p4 uses N-terminal β-turn substructures to make base-specific contacts in the major groove (25), and homeodomain proteins have N-terminal arms and linker regions that dock in the minor groove of the DNA (26–29). These flexible regions, which are sometimes not included in the strict definition of these DNA-binding domains, can have profound and essential roles in binding specificity.

According to the Structural Classification of Proteins (SCOP) database (30), DNA-binding proteins, whose structures are currently available in complexes with DNA, are grouped into more than 70 SCOP superfamilies (**Table 1**). Because of this large number, it is not possible to discuss each superfamily here, and thus, we focus only on a few representative examples. In **Table 1**, we group DNA-binding proteins into the following categories on the basis of the overall secondary structure content of the DNA-binding domains: mainly α, mainly β, mixed α/β, and multidomain proteins that have more than one of the aforementioned three domains. It is evident from the table that certain local motifs, such as the HTH motif, are used repeatedly and can be found within different global domain architectures. Moreover, depending on the protein and DNA-binding site, any one type of motif can be used in multiple ways to interact with DNA. These observations support one of the main points of this review: Protein-DNA interactions depend on the interplay between two equal partners, the DNA and the protein, and both macromolecules have their own characteristic three-dimensional structures that must accommodate the other to achieve specificity.

2.1. Mainly α

Proteins in 17 SCOP superfamilies have DNA-binding domains with mainly α-helical architecture, for example, homeodomains, leucine zipper proteins, and λ-repressor-like proteins. The α-helix is the most frequently used secondary structure element for specific DNA recognition in the major groove. The positioning of the helix in the major groove can vary between different protein families and also among different proteins within the same family, as reviewed previously (17). The Lac repressor (31, 32) and intron endonucleases (33–35) demonstrate that α-helices can also be used to interact with DNA in the minor groove. On the basis of the structural context in which the α-helices are found, the mainly α-class of proteins uses a number of local structural motifs for DNA binding.

2.1.1. Helix-turn-helix motif. The HTH motif is seen in many proteins in different SCOP superfamilies and is one of the most frequently represented structural motifs in DNA-binding proteins. The "recognition helix" of the HTH motif binds DNA through a series of hydrogen bonds and hydrophobic interactions with exposed bases, and the other helix stabilizes the interaction between the protein and DNA, but does not play a particularly strong role in recognition. Although the HTH motif is highly conserved, its structural context and precise orientation relative to the DNA-binding sites it recognizes can vary between different proteins, and the structures outside the HTH core region can differ greatly among various proteins. For example, in homeodomains, the second and third helices of the three-helix bundle comprise the HTH motif with the third helix (the recognition helix) contacting the major groove, in an orientation that is nearly parallel to the flanking DNA backbones. The motility gene repressor (MogR) DNA-binding domain contains seven α-helices connected by short loops: The first three helices form a three-helix bundle, the fourth helix forms a small dimerization interface, and helices 5–7 form a three-helix bundle DNA-binding domain that contains a HTH motif (α6 and α7), in which α7 is the recognition helix (36). Although the HTH motif is used most often in the major groove, some proteins use this motif to interact with the minor groove, for example, O6-alkylguanine-DNA alkyltransferase (AGT) (37).

A large class of HTH motif-containing proteins have an additional antiparallel β-sheet, hence its name "winged helix-turn-helix" (wHTH) motif (38). Proteins in many SCOP families contain the wHTH motif, including the hepatocyte nuclear factors-3 (HNF-3)/forkhead family of transcription factors (39),

Table 1 Architecture of DNA-binding proteins from the Structural Classification of Proteins (SCOP) database[a]

SCOP superfamily[b]	Number of PDB entries	Architecture of DNA-binding domains	DNA-binding motif
DNA/RNA polymerases	186	Multidomain, mixed α/β	
Nucleotidyltransferase	127	Multidomain, mixed α/β	
Ribonuclease H-like	104	Multidomain, mixed α/β	
Restriction endonuclease-like	89	Mixed α/β	
Homeodomain-like	75	Mainly α	Helix-turn-helix
Winged helix DNA-binding domain	75	Mainly α with a small β-ribbon (wing)	Winged helix-turn-helix
Lesion bypass DNA polymerase	60	Multidomain, mixed α/β	
Lambda repressor-like DNA-binding domains	57	Mainly α	Helix-turn-helix
Glucocorticoid receptor-like	53	Mixed α/β	Zinc finger
p53-like transcription factors	53	Mainly β	Immunoglobulin-like β-sandwich
DNA breaking-rejoining enzymes	45	Multidomain, mixed α/β	
DNA glycosylase	40	Mixed α/β	
S-adenosyl-L-methionine-dependent methyltransferases	40	Mixed α/β	
Histone fold	29	Mainly α	
Leucine zipper domain	27	Mainly α	Helix-loop-helix
TATA-box-binding protein-like	24	Mainly β	TBP β-sheet
Homing endonucleases	24	Mixed α/β	
C2H2 and C2HC zinc fingers	22	Mixed α/β	Zinc finger
E-set domains	21	Mainly β	Immunoglobulin-like β-sandwich
Chromo domain-like	19	Mainly β	β-barrel
DNA repair protein MutS	18	Multidomain, mixed α/β	
Ribbon-helix-helix	16	Mixed α/β	Ribbon-helix-helix
Uracil-DNA glycosylase-like	16	Mixed α/β	
His-Me finger endonucleases	14	Mixed α/β	
HMG box	13	Mainly α	Helix-turn-helix
Origin of replication-binding domain, RBD-like	13	Mixed α/β	
P-loop-containing nucleoside triphosphate hydrolases	12	Multidomain, mixed α/β	
Putative DNA-binding domain	12	Mainly α	
Zn2Cys6 DNA-binding domain	11	Mixed α/β	Zinc finger
IHF-like DNA-binding proteins	10	Mixed α/β	
RNase A-like	9	Mixed α/β	
Helix-loop-helix DNA-binding domain	8	Mainly α	Helix-loop-helix
SRF-like	8	Mixed α/β	
Zn2Cys4 DNA-binding domain	8	Mixed α/β	Zinc finger
C-terminal effector domain of the bipartite response regulators	7	Mainly α	Helix-turn-helix
DNase I-like	5	Mixed α/β	
Retrovirus zinc finger-like domains	5	Mixed α/β	Zinc finger
TrpR-like	5	Mainly α	Helix-turn-helix

(Continued)

Table 1 ***(Continued)***

SCOP superfamily[b]	Number of PDB entries	Architecture of DNA-binding domains	DNA-binding motif
Viral DNA-binding domain	5	Mixed α/β	
PIN domain-like	5	Mixed α/β	Ribbon-helix-helix
Zinc finger design	4	Mixed α/β	Zinc finger
DNA-binding domains of HMG-I(Y)	4	Peptide	AT hook
Transcription factor IIA (TFIIA)	4	Mainly β	β-barrel
Replication terminator protein (Tus)	4	Multidomain, mixed α/β	
UDP/glycosyltransferase, glycogen phosphorylase	4	Mixed α/β	
Replication modulator SeqA, C-terminal DNA-binding domain	4	Mainly α	
DNA-binding domain	4	Mixed α/β	β-sheet
FMT C-terminal domain-like	4	Mixed α/β	
Sigma3 and sigma4 domains of RNA polymerase sigma factors	3	Mainly α	Helix-turn-helix
Methylated DNA-protein cysteine methyltransferase domain	3	Mixed α/β	
DNA-binding domain of intron-encoded endonucleases	3	Mixed α/β	
Cryptochrome/photolyase FAD-binding domain	3	Mixed α/β	
T4 endonuclease V	2	Mainly α	Helix-turn-helix
SMAD MH1 domain	2	Mixed α/β	
KorB DNA-binding domain-like	2	Mainly α	Helix-turn-helix
DNA topoisomerase IV, alpha subunit	2	Multidomain, mixed α/β	
SMAD MH1 domain	2	Mixed α/β	
5′ to 3′ exonuclease catalytic domain	2	Mixed α/β	
Metallo-dependent phosphatases	2	Multidomain, mixed α/β	
WD40 repeat-like	2	Mainly β	
Xylose isomerase-like	1	Mixed α/β	
RNA polymerase	1	Multidomain, mixed α/β	
GCM domain	1	Mixed α/β	β-sheet
ATP-dependent DNA ligase DNA-binding domain	1	Multidomain, mixed α/β	
Transposase IS200-like	1	Mixed α/β	
Thioredoxin-like	1	Multidomain, mixed α/β	
Holliday junction resolvase RusA	1	Mixed α/β	
Skn-1	1	Mainly α	
ARID-like	1	Mainly α	Helix-turn-helix
GCM domain	1	Mixed α/β	β-sheet
Phage replication organizer domain	1	Mainly α	
Bet v1-like	1	Mixed α/β	
AbrB/MazE/MraZ-like	1	Mainly β	

[a] This table lists DNA-binding protein domains in different SCOP superfamilies, whose structures in complexes with DNAs are available in the Protein Data Bank as of August 2009. When they are well defined, the DNA-binding motifs used by these SCOP superfamilies are listed in the fourth column.
[b] Abbreviations: Please see **http://supfam.mrc-lmb.cam.ac.uk/** for the nomenclature used in names of SCOP superfamilies.

Ets domain (40), and multiple antibiotic resistance (MarR)-like transcription factors (41). The "wing" typically sits over the minor groove to make additional DNA contacts. However, in some cases, the wings rather than the HTH motif contact the DNA in the major groove, as seen in regulatory factor X1 (RFX1) (42). Many proteins also contain a second wing, which makes additional DNA contacts.

2.1.2. Helix-loop-helix and leucine zipper motifs. The helix-loop-helix motif consists of a short α-helix connected by a loop to a longer α-helix. Part of this motif is a dimerization domain that interacts with other helix-loop-helix proteins to form homo- or heterodimers; the dimerization partner often determines DNA-binding affinity and specificity because two α-helices, one from each monomer, bind to the major groove of the target DNA (43–46).

2.2. Mainly β

Although less common than α-helices, β-strands and intervening loops embedded in the mainly β-domain structures are used by proteins in seven SCOP superfamilies to recognize specific DNA sequences.

2.2.1. TATA box-binding protein. TBPs use a large β-sheet surface to recognize DNA by binding in the minor groove (14, 15). Insertion of the concave, 10-stranded β-sheet of TBP into the groove requires profound DNA distortion. As discussed in the following sections, the TATA box DNA undergoes dramatic unwinding and bending that allows for contacts between the protein's concave surface and the edges of the base pairs in the otherwise recessed minor groove.

2.2.2. Immunoglobulin-like β-sandwich. Immunoglobulin-like structural domains are used for DNA binding in diverse families of proteins, such as p53-like transcription factors (47), E-set domains (48, 49), and Runt domains (50). The sequence conservation of the immunoglobulin-like domains in different families is low, and the structures outside the domain diverge significantly. Although the overall fold is a β-sandwich, DNA recognition is achieved mainly by intervening loops. Like the mainly α-helical DNA-binding domains, the orientation of the β-sandwich domains relative to the DNA varies among different proteins and different families of proteins.

2.2.3. β-trefoil. The β-trefoil is a capped β-barrel with an approximate threefold symmetry, i.e., four strands are repeated in a threefold arrangement, where strands 1 and 4 form the walls of the β-barrel and strands 2 and 3 contribute to the cap structure to give a 12-stranded structure. The β-trefoil domain of CSL [CBF-1, Su(H), Lag-1], the nuclear effector of Notch signaling, contacts DNA via the loop between strands βA1 and βA2 (51).

2.2.4. β-β-β-sandwich. The structure of AgrA_C (52) reveals a novel topology of 10 β-strands arranged into three antiparallel β-sheets, which are arranged roughly parallel to each other in an elongated β-β-β-sandwich, and a small two-turn α-helix that is not involved in DNA binding. Base-specific contacts are made with residues from intervening loops at both the major and minor grooves.

2.3. Mixed α/β

A large number of proteins, which belong to 48 SCOP superfamilies, use mixed α/β domains to bind DNA, although the major secondary structure elements used for recognition can be any one or any combination of α-helix, β-strand, or loop.

2.3.1. Zinc finger proteins. The zinc finger is a compact ~30-amino acid DNA-binding domain. Zinc fingers are the most minimal of DNA-binding domains, with a relatively short α-helix, a two-stranded antiparallel β-sheet, and a Zn^{2+} ion coordinated by cysteine and histidine residues (53). Zinc fingers are classified by the type and order of the zinc coordinating residues, e.g., Cys_2His_2, Cys_4, and Cys_6. Zinc fingers often occur as tandem repeats with two, three, or more fingers that can bind in the major groove, typically spaced at 3-bp intervals. The α-helix of each domain (the recognition helix) makes sequence-specific contacts to DNA bases in the major groove; residues from a single recognition helix can contact four or more bases to yield an overlapping pattern of contacts with adjacent zinc fingers.

2.3.2. Ribbon-helix-helix motif. A family of transcription factors from bacteria contains the ribbon-helix-helix (RHH) motif (54) that consists of a two-stranded antiparallel β-ribbon followed by two α-helices. DNA recognition is achieved by insertion of the β-ribbon into the major groove, whereas the two helices comprise most of the hydrophobic core and are involved in dimerization. The prototypical examples are Met repressor MetJ (55) and Arc repressor (56).

2.3.3. Other mixed α/β domains. Structural studies of seemingly dissimilar restriction endonucleases with remarkable DNA sequence specificity demonstrated that they all share a common structural core with a mixed α/β architecture (57). A large amount of structural data also reveal that DNA polymerases, DNA lesion repair enzymes, and DNA-modifying enzymes all have mixed α/β domain structures (**Table 1**).

2.4. Multidomain Proteins

Many DNA-binding proteins contain multiple DNA-binding domains, which can work together to recognize different regions of a target sequence, achieving high affinity and recognition specificity. For example, POU domain proteins, such as Oct-1 (58) and Brn-5 (59), contain a homeodomain (POU_{HD}) and POU-specific domain (POU_{S}) that are connected by a flexible linker, and MarA (multiple antibiotic resistance A) consists of two HTH motifs that contact two successive major grooves (60). Other examples are the Rel-homology domain proteins, such as NF-κB p50, that have two immunoglobulin-like domains in each monomer: The N-terminal domain mediates DNA contacts primarily in the major groove, and the C-terminal domain mediates homo- and heterodimer interactions in addition to contacting DNA (48, 61). The side chains involved in dimer interactions lie along one face of the β-sandwich, leaving the loops free to contact the DNA. The *Escherichia coli* transcription factor Rob, which belongs to the AraC/XylS family, has two HTH domains: One binds specifically to DNA, whereas the other only forms a single salt bridge with the DNA backbone (62, 63). TCF (T cell factor) binds to specific DNA sequences through a high-mobility group (HMG) domain. Recent data suggest that DNA recognition by *Drosophila* TCF occurs through a bipartite mechanism, involving both the HMG domain and the C-clamp, which enables TCF to locate and activate wingless-regulated enhancers in the nucleus (64).

3. SEQUENCE-DEPENDENT VARIATIONS OF DNA STRUCTURE

Most current analyses of the information content in a nucleotide sequence view DNA as a one-dimensional string of letters based on an alphabet consisting of only four characters: A, C, G, and T. Yet these bases are chemical entities that, along with the inclusion of the backbone sugar and phosphate groups, create a three-dimensional double-stranded structure in which each base pair has a specific chemical and conformational signature (10). Although this textbook view of the double helix is well-known, what is much less appreciated is that DNA structures vary in a sequence-dependent manner (20, 65) and that structural variations are used by proteins to recognize DNA sequences (66).

In this section, we review the main ways in which DNA structures are known to deviate from idealized B-DNA. We distinguish between effects that vary the geometry of the helix in a localized manner (local shape, e.g., minor groove width and DNA kinks) from those that deform the overall cylindrical shape

Table 2 Tendency of DNA sequence elements to have specific structural characteristics[a]

Sequence element[b]	Structural characteristics	References
AT rich	B-DNA	72, 114
GC rich	A-DNA at low humidity	76, 77, 188
A-tract	B′-DNA, narrow minor groove, bending, rigid for $\geq$4 bp	81–83, 86, 217
TATA box	High deformability, A-DNA, TA-DNA upon TBP binding	78, 189
RY alternating (especially GC alternating)	Z-DNA at high salt concentration, upon cytosine methylation or supercoiling	79, 80, 197
YpR step (especially TpA step)	Compresses major groove, high deformability, hinge step, kinking	84, 88–90
RpY step	Compresses minor groove, low deformability	84, 88, 89

[a] The table reflects general tendencies for some sequences to have particular structural characteristics. It is important to stress, however, that DNA conformation depends on environmental conditions (e.g., humidity and salt concentration) and the larger sequence context (65, 76). For example, although AT-rich DNA is usually observed in B form, TATA box-containing oligonucleotides were crystallized in A form (189), which is the basis for TATA-binding protein specificity. In addition, owing to their high deformability, the structure of TATA boxes is affected by long-range sequence effects (218) and by supercoiling. A TATA box flanked by GC alternating regions can also assume a Z-DNA conformation (219).

[b] Abbreviations: A, adenine; C, cytosine; G, guanine; T, thymine; R, purine; Y, pyrimidine. The lower case "p" between nucleotides stands for phosphate to distinguish a base pair step from a base pair.

of the double helix (global shape, e.g., DNA bending, A-DNA, and Z-DNA). In addition, although some DNA sequences do not produce a well-defined structure per se, they may be highly flexible and therefore have a strong propensity to assume a non-B-like structure when bound to a protein. This property, commonly referred to as deformability, is another sequence-dependent feature that is used by proteins to recognize specific DNA sequences. To help make the connection between DNA sequence and DNA structure, **Table 2** lists DNA sequences that have a tendency to assume a particular DNA structure.

Differences in DNA shape can produce electrostatic potentials of varying magnitudes, a characteristic that can be read by proteins. For example, narrow minor grooves locally enhance the negative electrostatic potential of DNA through electrostatic focusing (66), which describes the deformation of field lines owing to the shape of the dielectric boundary between solute and solvent (67). This phenomenon was first described for a cavity of the superoxide dismutase protein (68) but has also been shown to play a role in codon-anticodon recognition in transfer RNAs (69), in shaping electrostatic potentials around diverse RNA structures (70), and in shifting pK_as in RNA catalytic sites (71). As discussed below, the effect appears to play an important role in protein-DNA recognition. In the following sections, we therefore refer to the electrostatic potential surfaces shown in **Figure 1**, which illustrate the close connection between shape and electrostatic potential in different DNA structures.

3.1. Global Shape Variations

In this section, we discuss the major ways in which DNA shape can vary in a global manner. These include different helical topologies and overall deformations of the DNA helix.

3.1.1. Polymorphisms of the double helix.

Global shape variations include previously recognized polymorphisms of the double helix, B-DNA, A-DNA, TA-DNA, and Z-DNA, which we briefly discuss here.

3.1.1.1. B-DNA. The most common form of double-stranded DNA is B-DNA, which is generally favored in aqueous solution similar to the environment in cells (72). Most DNA-binding proteins recognize B-DNA and its structural variants. B-DNA is a right-handed double helix with base pairs oriented approximately perpendicular to the helix axis. Ideal

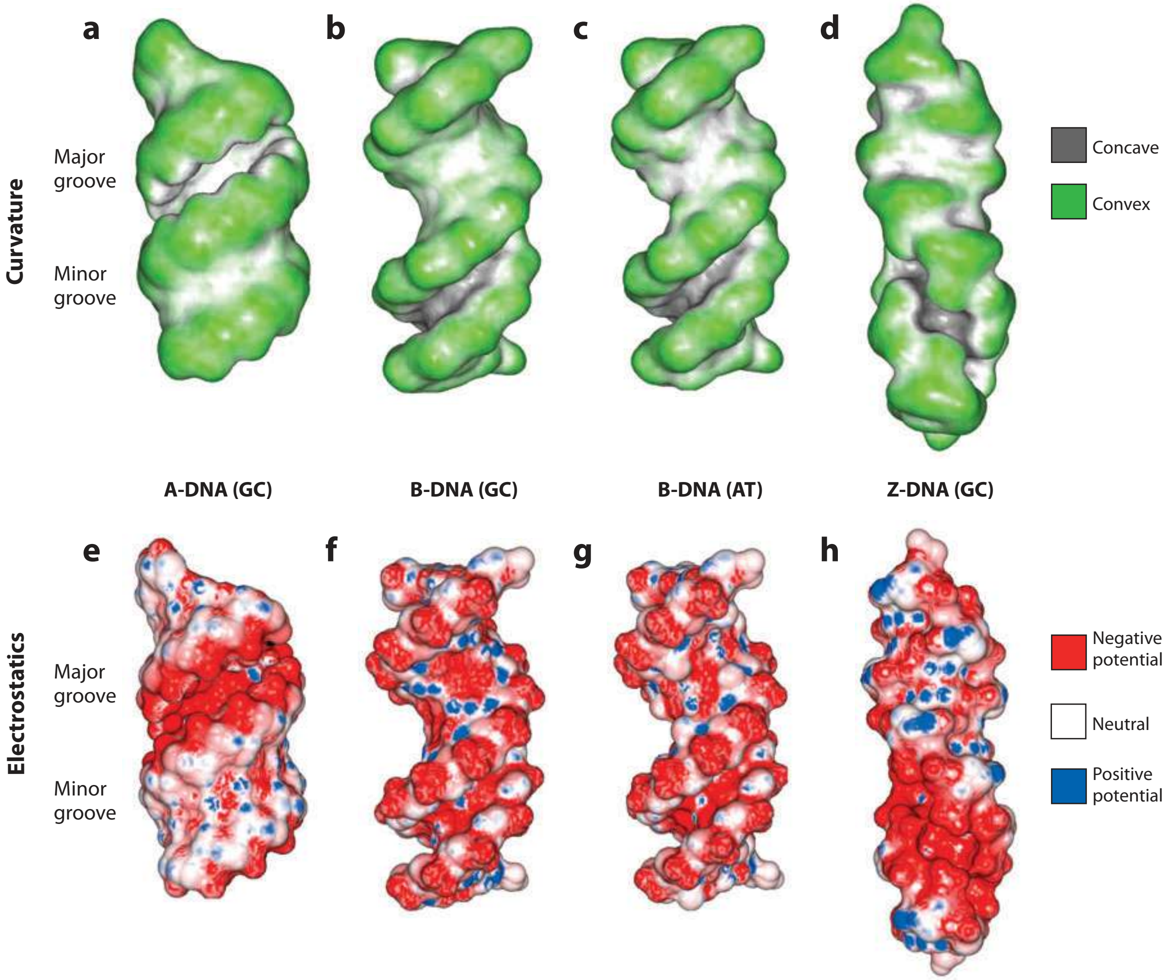

Figure 1

Molecular shape and electrostatic potential of A-DNA, B-DNA, and Z-DNA. The upper panels show the molecular shape in GRASP2 images (convex surfaces in *green* and concave surfaces in *dark gray*) (220) of the three helical forms of DNA constructed with the software tool, 3DNA (92) from fiber diffraction data (72, 80). Each DNA helix comprises 14 mers. The width and depth stated below were calculated with the software tool, Curves (221). The lower panels show how the electrostatic potential at the molecular surface varies owing to shape and atomic charges. The electrostatic potentials were calculated as described in (66) by solving the Poisson-Boltzmann equation with DelPhi (67, 222) at a salt concentration of 0.145 M (negative electrostatic potentials are shown in *red* and positive electrostatic potentials in *blue*). (*a*) A-DNA with a narrow, deep major groove (2.2-Å wide and 9.5-Å deep) and a wide, shallow minor groove (10.9-Å wide and no defined depth). The model is of the alternating sequence d(GC)$_7$. (*b*) B-DNA [alternating sequence d(GC)$_7$] with a wide, shallow major groove (11.4-Å wide and 4.0-Å deep) and a narrow, deep minor groove (5.9-Å wide and 5.5-Å deep). (*c*) B-DNA [alternating sequence d(AT)$_7$]. Because the models are built on the basis of fiber diffraction data, the shape of GC and AT alternating B-DNA does not reflect a sequence dependency. (*d*) Z-DNA lacks a major groove (13.2-Å wide and no defined depth), and the minor groove is narrow and deep (2.4-Å wide and 5.0-Å deep). The model is of the alternating sequence d(GC)$_7$. (*e*) A-DNA exhibits a strongly negative major groove but a hydrophobic minor groove surface, which is partially owing to its exposed C3′ endo sugar moieties. (*f*) B-DNA [alternating sequence d(GC)$_7$] exhibits a negative minor groove and less negative major groove. (*g*) B-DNA [alternating sequence d(AT)$_7$]. Variations in electrostatic potential between GC and AT alternating B-DNA reflect the different functional groups of the base pairs (e.g., positive guanine amino group in the GC minor groove and neutral thymine methyl group in the AT major groove). (*h*) Z-DNA exhibits a negative minor groove and a positive surface on opposing edges of the bases.

B-DNA exhibits a wide, shallow major groove and a narrow, deep minor groove (**Figure 1*b*,*c***) (65). As is evident from (**Figure 1*f*,*g***), the minor groove of B-DNA generally exhibits a more electronegative potential than the major groove. The differences in the potential in either groove between AT- and GC-rich sequences are due to the disposition of polar groups at the base edges; specifically AT-rich sequences display more negative electrostatic potentials in the minor groove than GC-rich sequences (**Figure 1*f*,*g***) (73, 74). These effects are further enhanced by sequence-dependent effects on groove width, as discussed below.

3.1.1.2. A-DNA. A-DNA is observed under dehydrated conditions and in some protein-DNA complexes (75). GC-rich sequences have an increased tendency to assume A-DNA or A/B intermediate conformations (76). This property is, in part, because GC base pairs have three hydrogen bonds, whereas AT base pairs have only two. This property makes GC base pairs more planar, allowing consecutive GC base pairs to slide relative to each other, which promotes the A/B transition (77). Although less pronounced, such a tendency is also observed for TATA boxes partly because the TpA step counters propeller twisting. A-DNA is also a right-handed double helix with the base pairs shifted toward the minor groove and, compared to B-DNA, tilted with respect to the helix axis by about 20°. This results in a narrow, very deep major groove and a wide, very shallow minor groove (**Figure 1*a***) (65). On the basis of this geometry, the A-DNA major groove resembles the shape of the B-DNA minor groove, which explains why, in contrast to B-DNA, the A-form major groove has a more negative electrostatic potential than its shallow minor groove (**Figure 1*e***) (70).

3.1.1.3. TA-DNA. TA-DNA is a variant of A-DNA observed in TATA boxes, which are specifically recognized by TBPs. It differs from A-DNA mainly by a larger base pair inclination of around 50° relative to the helix axis. This feature led to the description of TA-DNA as tilted A-DNA (78). The TA-DNA geometry exhibits a positive roll (rotation between adjacent base pairs with respect to the base pairing axis), which explains the opening of the TATA box minor groove observed in TATA-TBP complexes (14, 15).

3.1.1.4. Z-DNA. Alternating purine-pyrimidine sequences were observed to form a left-handed double helix under high salt concentrations (79, 80). Because of the zigzag conformation of its backbone, this topology was coined Z-DNA. Thought to form when B-DNA is deformed by supercoiling, Z-DNA does not have a pronounced major groove; instead, the base edges form a convex surface. The minor groove, however, resembles the dimensions of the B-DNA minor groove, but with a zigzag trajectory of the backbone (**Figure 1*d***) and a uniform negative electrostatic potential (**Figure 1*h***).

3.1.2. DNA bending. We define DNA bending as a curvature distributed over a stretch of several base pairs, leading to a different orientation of the regions on both sides of the curvature (**Figure 2*a***). Bending has frequently been observed for sequences that contain A-tracts, which are stretches of A:T base pairs that include ApA (TpT) and ApT, but not TpA, steps (81–83). Various models have been established to explain the molecular origin of bending (84, 85). These models either associate bending with wedge angles between adjacent base pairs, which can involve both roll and tilt, or with junctions between regions with negative base pair inclination (A-tracts) and regions with positive inclinations (83, 86).

It is likely that the phasing of wedge angles is the critical factor for overall curvature. If short A-tracts (regions with negative roll) are phased by half a helical turn, the overall curvature cancels owing to bending toward opposite sides of the helix. In a sequence where regions with negative roll are phased by a helical turn, the overall curvature is enhanced. The effect is further enhanced if regions with negative roll are in phase of half a helical turn with regions of positive roll

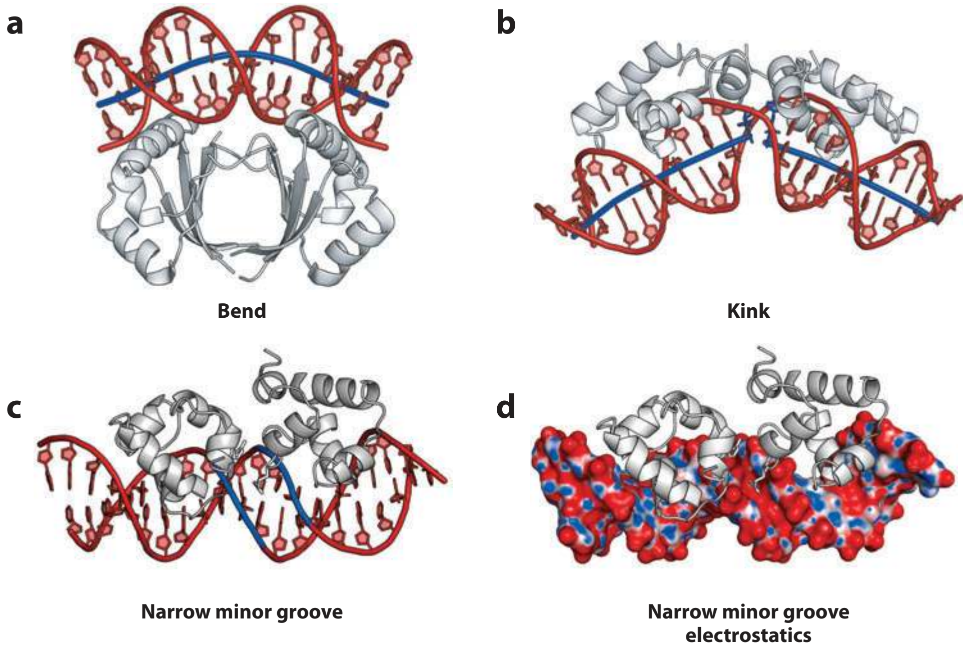

Figure 2

Illustration of DNA bending, kinking, and minor groove narrowing in protein-DNA complexes. (*a*) HPV-18 E2 bound to DNA (PDB ID 1jj4) shows bending over a large stretch of the helix. The smooth curvature is visualized by the helix axis (*blue*), calculated with Curves (221). (*b*) The Lac repressor kinks the DNA at a central CpG base pair step, stabilized by the partial intercalation of leucines (PDB ID 2kei). The helix axes calculated for both sides of the kink (*blue*) show an abrupt change in the helix trajectory caused by the kink. (*c*) Phage 434 repressor recognizes local shape deformations of its operator with arginine residues (PDB ID 2or1) (66). The narrow region of the minor groove that is contacted by arginines is highlighted in blue. (*d*) For the same structure shown in panel *c*, the electrostatic potential of the operator, calculated in the absence of the repressor, is plotted on the molecular surface. In comparison with **Figure 1*f*,*g***, the bottom of the minor groove is uniformly red, indicating enhanced negative electrostatic potential (66).

as both regions would bend the double helix in the same direction. Such a pattern has been reported for the nucleosome (84) and the papillomavirus E2-binding site (87). Ultimately, the source of sequence-dependent bending can be traced to the conformational properties of individual dinucleotide steps (88, 89), their tendency to form wedge angles, and the composition of these dinucleotide steps in a DNA sequence.

3.2. Local Shape Variations

Unlike global shape variations, we use the term local shape variations to refer to deviations from ideal B-DNA that originate from an individual base pair (e.g., a kink) or are localized in a small region of the double helix (e.g., minor groove narrowing).

3.2.1. DNA kinks. We distinguish a kink from a DNA bend by defining a kink as a local disruption of an otherwise linear helix (**Figure 2*b***). DNA kinks result from the complete or partial loss in stacking at a single base pair step. The pyrimidine-purine (YpR) steps TpA, CpA (TpG), and CpG are least stabilized through base stacking interactions, and of these, the TpA step has the weakest stacking interactions (**Table 2**) (65, 90). Therefore, it is the most flexible of the 10 unique dinucleotides and is referred to as a "hinge" step (86, 89). Because kinks occur at individual base pair steps, regions adjacent to a kink can remain in a straight B-form conformation or can be curved. Bending and kinking can enhance each other as is the case for CpA steps adjacent to an A-tract (91). Kinks are often stabilized by protein binding in cases where the loss of stacking interactions is compensated by the intercalation of hydrophobic side chains, which usually further deforms the kinked dinucleotide.

3.2.2. Minor groove narrowing. Minor groove width is another feature that varies locally in DNA structures (**Figure 2*c***) (66). Differences in minor groove width arise from differences in the hydrogen bonding pattern of each base pair and from differing stacking interactions for each dinucleotide step. To optimize both types of interactions, DNA structures vary with respect to three rotational parameters: *roll (*the relative rotation between adjacent base pairs with respect to the base pairing axis), *helix twist* (the relative rotation between adjacent base pairs with respect to the helix axis), and *propeller twist* (the relative rotation between bases within a base pair with respect to the base pairing axis) (92). ApT base pair steps usually have negative roll angles, which lead to a compression of the minor groove (**Table 2**) (84). In an A-tract sequence, ApT and ApA (TpT) steps exhibit a negative roll, and the bifurcated hydrogen bonds of A:T base pairs lead to propeller twisting, both enhancing minor groove narrowing (83, 87). In addition, several A:T base pairs in a row enhance propeller twisting by allowing the formation of interbase pair hydrogen bonds in the major groove (81). In contrast to ApA (TpT) and ApT, propeller-twisted TpA steps lead to a steric clash of the cross-strand adenines (86). Therefore, TpA steps tend to locally widen the minor groove and break rigid A-tract structures, and are thus referred to as hinge steps (**Table 2**) (89).

4. MECHANISMS OF PROTEIN-DNA RECOGNITION

Macromolecular interactions, whether they be protein-protein or protein-DNA in nature, depend on the three-dimensional structures of both interacting partners. In this section, we classify the types of readout mechanisms used by proteins to recognize DNA sequences in light of the types of DNA structures defined above.

4.1. General Comments

Protein-DNA interfaces involve on average 24 protein residues and 12 nucleotides (93), making it likely that each interface is composed of many different types of interactions. Although all interactions contribute to binding affinity, specificity can be viewed as resulting from a subset of interactions that are sequence specific. It is these specificity-determining contacts that we are most concerned with here.

Given our focus on specificity, it is important to define what we mean by this term and to point out that DNA-binding proteins generally exhibit multiple tiers of specificity. All homeodomains, for example, have an asparagine at position 51 (Asn51), which is important for the specific binding of these proteins to AT-rich sequences, such as TAAT (e.g., Engrailed and Antennapedia) (26, 94, 95). Thus, Asn51 can be considered to be a critical determinant of homeodomain DNA-binding specificity. However, as all homeodomains have Asn51, this residue cannot contribute to specificity within this superfamily. On a finer level, position 50 of the homeodomain partially fulfills this role: When it is a glutamine (Gln), the preferred binding sites are TAAT$\underline{\text{TG}}$ or TAAT$\underline{\text{TA}}$ (where the Gln contacts are underlined), but when it is a lysine, the preferred binding site is TAAT$\underline{\text{CC}}$ (96–99). However, the subset of homeodomain proteins that have a glutamine at position 50 is still very large and includes all of the Hox homeodomains, of which there are 39 in humans alone. Therefore, Gln50 cannot contribute to specificity within this subset of homeodomain proteins. In addition to Asn51 and Gln50, which are presented from a HTH recognition helix in the major groove, Hox proteins also bind to the minor groove, where DNA shape, in particular minor groove width, is read (29). As discussed below, this mode of protein-DNA recognition contributes to specificity within the Hox family. From this one example, we see that DNA-binding proteins use multiple readout mechanisms and that specificity is ultimately achieved by combinations of these mechanisms that successively fine-tune the selection of binding sites.

Although contacts between proteins and the DNA backbone are typically considered to have little impact on specificity (100), backbone contacts may play a role in specificity through the positioning of protein recognition elements in orientations that allow them to make other, more specific contacts, such as hydrogen bonds to the bases (101, 102). Indeed, protein families often contain conserved backbone-contacting residues that preserve the interface orientation for an entire family (102). In addition, specificity may depend on contacts to the DNA backbone if these contacts can only be made when the DNA assumes a sequence-dependent structure that deviates from ideal B-DNA (referred to below as nonideal B-DNA). An example is the readout of narrow minor groove regions, where the phosphates are located in positions that differ from ideal B-DNA. The Arg repressor from *Mycobacterium tuberculosis*, for instance, specifically recognizes a narrow minor groove region via extensive phosphate contacts from a four-stranded β-sheet that lies above the groove without inserting any side chain into the groove (103).

Protein-DNA recognition is also more complex than a simple docking process of two structurally preformed macromolecules. Some proteins fold only in the presence of DNA. For example, the leucine zippers of Fos and Jun are helical only when they form a heterodimer, and the basic regions are helical only when the dimer binds DNA (104, 105). Moreover, other domains in both proteins appear to be unstructured until bound by cofactors such as CREB-binding protein (CBP/p300) (106). Lymphoid enhancer factor-1 (LEF-1) also transitions from a relatively unstructured state to a well-folded domain upon DNA binding (107). The sequence-specific binding of Cys_2His_2 zinc finger proteins to DNA causes their linker regions to fold, cap, and thereby stabilize the preceding helix, which helps to orient the next zinc finger correctly for binding in the major groove (108). Finally, binding of the zinc finger domain of retinoid X receptor (RXR) to DNA leads to folding of the dimerization region, which is disordered in the unbound protein (109). DNA binding can also induce conformational changes in the bound protein that can change its properties. For example, the binding of the glucocorticoid receptor (GR) to its response elements induces conformational changes that expose transcriptional activation surfaces (110). Moreover, different GR-binding sites result in distinct GR activities, which, on the basis of X-ray data, could be explained by changes in the orientation of a GR loop induced by a modification of DNA backbone contacts (111).

DNA can also change conformation, and preexisting sequence-dependent conformations can be stabilized or enhanced upon protein binding (**Figure 3**). For example, in specifically designed noncognate GR complexes, the DNA is able to distort so as to maximize the number of cognate interface interactions, even if these are only maintained by a single strand (102, 112). Such effects make it difficult to unambiguously determine if nonideal B-DNA structures observed in protein-DNA complexes are intrinsic to the DNA sequence, induced by the protein, or some combination of the two. The relative impact of intrinsic versus induced effects on DNA structure can only be assessed with certainty by comparing the structure of the free DNA-binding site with its protein-bound form. Such structural information is currently restricted to the binding sites of only a handful of proteins, including the EcoRI endounclease (113, 114), Trp repressor (12, 115), Met repressor (55, 116), purine repressor (31, 116), NF-κB (48, 49, 117), Zif268 zinc fingers (**Figure 3*a***) (116, 118, 119), papillomavirus E2 protein (**Figure 3*b***) (13, 82, 120), and the Runt domain (50, 121, 122). The limited size of this group is largely because of the lack of free DNA structures (20). In their place, theoretical

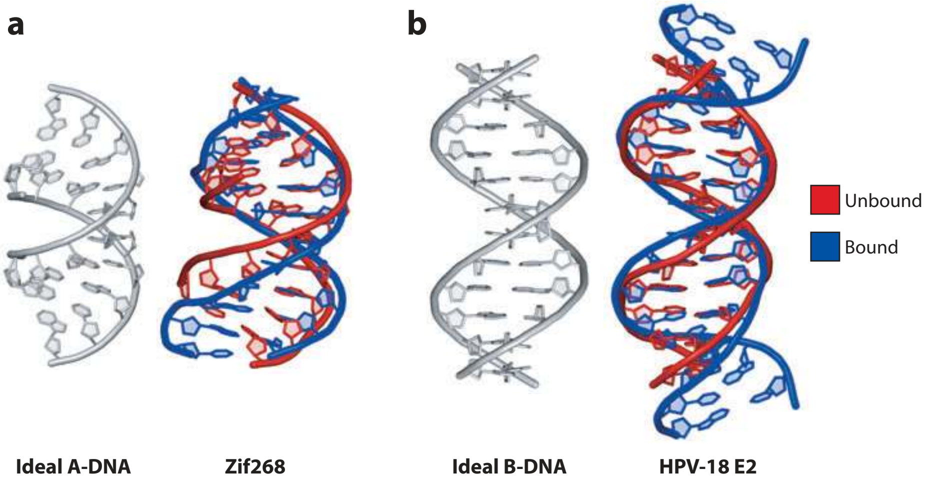

Figure 3

DNAs bound to proteins have features already present in unbound DNAs. (*a*) The structure of the unbound FIN-B sequence (PDB ID 2b1c) is similar to ideal A-DNA (*gray*), whereas the bound structure of the Zif268-DNA complex (PDB ID 1a1f) has some A-DNA characteristics, notably a wider minor groove than normally found in B-DNA. (*b*) The specific HPV-18 E2 site (PDB ID 1ilc) contains an A-tract AATT in the central region of the helix, which, although not contacted by the protein, bends the free-DNA structure (*red*) in a manner similar to that seen in the bound structure (*blue*) of the HPV-18 E2-DNA complex (PDB ID 1jj4). In comparison to ideal B-DNA (*gray*), the bending is reflected by a minor groove narrowing in the center of the free and bound DNA.

approaches have been developed to estimate the impact of intrinsic versus induced effects when only the bound form is available (123) or to predict the structure of the unbound DNA-binding site (20, 29, 87).

With this background in mind, below we discuss the various mechanisms proteins use to recognize their binding sites, attempting to organize them from a structure-based perspective (**Figure 4**). Note, we only have space in this review to support each readout mechanism with a small number of examples. Furthermore, because any one DNA-binding protein typically uses a variety of readout mechanisms, the same example may be used multiple times.

4.2. Base Readout

One well-established way for proteins to achieve DNA-binding specificity is through contacts with the bases in either the major or minor groove that recognize the chemical signature of the base or base pair. This type of recognition is generally mediated by the formation of hydrogen bonds between amino acids and bases, which convey the highest degree of specificity and, in some cases, by water-mediated hydrogen bonds or hydrophobic contacts (**Figure 4**).

4.2.1. Base-specific interactions in the major groove. In this section, we discuss the two main types of base readout mechanisms that occur in the major groove of the DNA, hydrogen bonds and hydrophobic interactions (**Figure 4**).

4.2.1.1. Hydrogen bonds with bases. Hydrogen bonds with bases can confer greater specificity in the major groove than in the minor groove because the four possible base pairs have a unique pattern of hydrogen bond donors and acceptors in the major but not in the minor groove (**Figure 5**) (10, 124). Proteins that form hydrogen bonds with bases in the major groove use HTH domains (e.g., homeodomains,

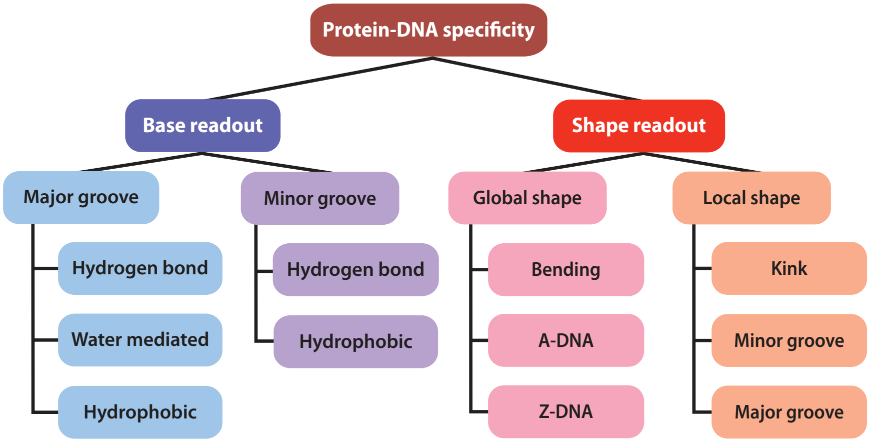

Figure 4

Types of protein-DNA recognition mechanisms used for specificity. We distinguish between two main classes of recognition: base readout and shape readout, which are further subdivided as illustrated.

434 repressor, λ repressor, Trp repressor, Myb), zinc finger domains (e.g., TFIIIA), immunoglobin fold domains (e.g., p53, NF-κB, STAT, and NFAT), and the N-terminal end of basic leucine zipper (bZip) domains or the Max transcription factor (17–19).

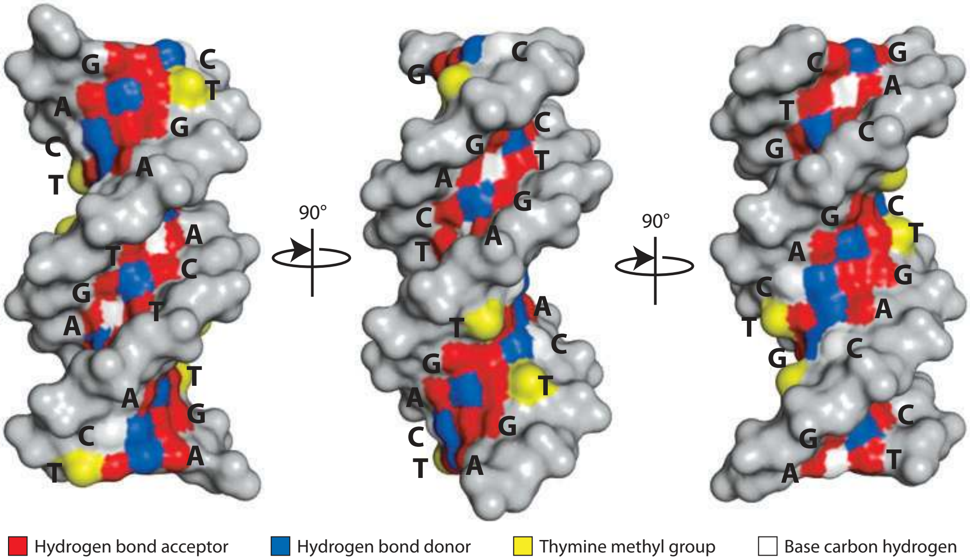

Figure 5

Base recognition in the major and minor groove. Sequence-specific patterns on the edges of the bases in the major groove underlie the ability of proteins to readout base pairs through hydrogen bonds and hydrophobic contacts (hydrogen bond acceptors in red, donors in blue, thymine methyl group in yellow, and base carbon hydrogens in white). In contrast, A:T versus T:A and C:G versus G:C are indistinguishable in the minor groove. The three panels show successive rotations of 90° around the helix axis. The dodecamer $d(GACT)_3$ was built on the basis of fiber diffraction data with 3DNA (92).

As noted above, the orientation of the recognition helix in the major groove is similar for homeodomain-DNA interfaces (125) but can vary among different families (17) and even within a given family, as between the Trp and λ repressors (100). In some cases, as observed for the KorA repressor, the recognition helix induces a widening of the major groove (126). In addition to α-helices, hydrogen bonds between β-sheets and bases can be used as well in specific recognition. Hydrogen bonds between bases in the major groove with the convex side of a β-sheet are observed in the binding of the MetJ and Arc repressors to DNA (127). The width of the major groove adjusts to the size of the β-sheet (widened in Arc repressor and narrowed in MetJ repressor), and the side of the β-sheet interacting with DNA generally exhibits more positive electrostatic potentials (127).

Specificity conveyed through hydrogen bonds in either groove depends on the number of contacts formed between protein residues and DNA bases but also on the uniqueness of the hydrogen bonding geometry. Bidentate hydrogen bonds (two hydrogen bonds with different donor and acceptor atoms) have the highest degree of specificity, followed by bifurcated hydrogen bonds (two hydrogen bonds that share the donor) and single hydrogen bonds. Whereas single hydrogen bonds usually do not contribute to specificity, bidentate hydrogen bonds are a source of remarkable selectivity (128). Bidentate hydrogen bonds can be formed with one base, two bases in a base pair, two adjacent bases in one strand, or two bases diagonally in different base pairs and opposite strands.

As discussed above, the specificity achieved through hydrogen bonds with bases depends on the pattern of donors and acceptors at the base edges in both grooves (**Figure 5**). Because DNA usually occurs in Watson-Crick geometry (1), this pattern is specific for each of the four base pairs in the major groove. However, base pair geometry can vary. For instance, Hoogsteen base pairs (129) have been observed in structures with deformed DNA sequences, such as the TBP/TATA box complex (130) and at the ends of oligonucleotides where the helical structure is preserved through stacking interactions [e.g., in a p53 tetramer complex (122)]. To date, a Hoogsteen base pair not present at the end of an oligonucleotide has only been observed in one complex with undistorted B-DNA, i.e., the MATα2 homeodomain bound to a specific binding site (131). Interestingly, the Hoogsteen base pair occurs in the center of the binding site CATGT$\underline{\text{A}}$ATT (underlined A) and was seen in crystals generated under various conditions (131). A transition from a Watson-Crick to Hoogsteen geometry alters the pattern of hydrogen bond donors and acceptors in both grooves and the conformation of the double helix. Although this single example should be interpreted with caution, it raises the possibility that non-Watson-Crick base pairs may contribute in important ways to binding specificity. As high-resolution structures are required to visualize such geometries, non-Watson-Crick base pairs may be present at a greater frequency than is evident in existing structures (see Note Added in Proof).

In many structures, the hydrogen bonds between protein and DNA are mediated by intervening water molecules. The bridging of hydrogen bonds by water molecules has frequently been observed for enzymes (132), and most hydrogen bonds in the Trp repressor-DNA interface are water mediated (12, 124). Mutagenesis experiments have shown that the CTAG tracts in both half sites of the Trp repressor's binding site are most critical for its sequence specificity (133). Highly ordered water molecules also mediate the specific readout of bases in the RXR-retinoid acid receptor (RAR)-DNA complex involving several arginine and lysine residues (134). Interestingly for the Lac repressor, the protein-DNA interface retains a significant portion of its hydration when it binds nonspecifically but not in the specific complex (135).

These data suggest that water-mediated hydrogen bonds in the major groove can be used for specific readout because they often reflect the positions of hydrogen bond donors and acceptors at the base edges. This is not the case for water molecules in the minor groove where the donor-acceptor patterns become unrecognizable.

4.2.1.2. Hydrophobic contacts with bases. Whereas hydrogen bonds with bases are highly specific in recognizing purines, hydrophobic contacts with bases are mainly used to read pyrimidines. Protein side chains employ hydrophobic interactions to differentiate thymine from cytosine (124) as in the bacteriophage 434 repressor and 434 Cro binding to their operator sites (136, 137). Four thymine methyl groups form a cleft that is specifically recognized by a valine in the lambdoid bacteriophage P22 c2 repressor-operator complex (138).

Hydrophobic contacts with bases also play a key role in the sequence-specific recognition of single-stranded DNA by bacterial cold-shock proteins, which recognize polythymine strands through stacking interactions with phenylalanines and histidines and distinguish thymine from cytosine through hydrogen bonding (139, 140).

4.2.2. Base-specific interactions in the minor groove. This section discusses the two forms of base readout observed in the minor groove: hydrogen bonds and hydrophobic contacts (**Figure 4**).

4.2.2.1. Hydrogen bonds with bases. Although, as discussed above, the pattern of donors and acceptors in the minor groove does not distinguish A:T from T:A or G:C from C:G base pairs (10) (**Figure 5**). Some proteins, such as zinc finger proteins with Cys_2Cys_2 GATA-like domains, that form hydrogen bonds in the major groove also bind in the minor groove (19). HMG proteins form hydrogen bonds in the minor groove (19) but rely on the recognition of DNA shape and flexibility, discussed below, to achieve specificity. This is also apparent for the binding of TBP to the minor groove as the six observed hydrogen bonds with the TATA box are not sufficient for the protein to attain specificity (14, 15, 141).

In some cases, a spine of hydration, a continuous string of water molecules, in narrow minor groove regions is contacted by proteins, as observed in the DNA complexes formed by the IFN-β enhanceosome (142, 143) and the integration host factor (IHF) (141). In other cases, only individual water molecules are displaced from narrow minor groove regions when amino acids intrude into the groove, e.g., α2-Arg7 in the MATa1-MATα2-DNA complex (144). The displacement of water molecules from the narrow minor groove has been shown to provide a strong thermodynamic driving force for DNA binding (145–147).

4.2.2.2. Hydrophobic contacts with bases. Architectural proteins only contact the minor groove, which is often associated with a dramatic widening and extensive hydrophobic contacts (141). This mechanism is employed by the TBP, SRY, and LEF-1. The TBP/TATA box interface is completely dehydrated, and the abundance of hydrophobic contacts in the interface (148) suggests that they contribute to specificity. Although 12 of the 16 hydrogen bond acceptors in the minor groove remain unsatisfied upon TBP binding, these base atoms mainly engage in hydrophobic contacts with nonpolar side chains (14, 15, 141).

4.3. Shape Readout

For most DNA-binding proteins, the readout of base pairs through hydrogen bonds or hydrophobic contacts is not sufficient to explain specificity. Other factors that have been proposed to contribute to specificity are sequence-dependent DNA structure and deformability (20, 149). These readouts, which all depend on deviations from ideal B-DNA, comprise a diverse set of mechanisms that all fall under the general heading of binding a nonideal B-DNA shape. As such, we collectively refer to them as shape readout (**Figure 4**). Furthermore, we distinguish between *local shape readout* mechanisms, in which the DNA helix deviates from ideal B-DNA in a localized manner, and *global shape readout* mechanisms, in which most of the DNA-binding site is either deformed or in a nonideal B-form conformation (**Figure 4**).

Both local and global shape readouts can contribute to DNA-binding specificity. For local shape readout, such as minor groove narrowing, recent results suggest that the shape of the minor groove within a binding site can be "read" by a complementary set of basic side chains, most typically arginines, when presented in the correct conformation (66). In contrast, global shape readout, such as a gradual bend in the DNA helix, may position elements of the DNA backbone such that these otherwise nonspecific contacts can become highly specific. Below, we discuss each of these types of readouts, providing specific examples to illustrate them.

4.3.1. Local shape readout.

As described in the DNA structure section, the two predominant local shape deviations from ideal B-DNA are (*a*) small regions of 3–8 bp where the minor groove is narrow and (*b*) DNA kinks, which are caused by the unstacking of a single base pair step (**Figure 4**).

4.3.1.1. Minor groove shape.

The N-terminal arms of homeodomain proteins have been observed in the minor grooves of several structures, but only recently have they been shown to play a role in DNA-binding specificity. In particular, the binding of the Hox protein Sex combs reduced (Scr) and its cofactor Extradenticle (Exd) to Scr-specific (*fkh250*) and Hox-Exd consensus (*fkh250^{con}*) binding sites shows how N-terminal arm arginines recognize a minor groove shape to achieve specificity (29). Whereas both Arg3 and Arg5 of Scr are ordered in the minor groove of the specific binding site (**Figure 6*c***), Arg3 is disordered when presented with the Hox-Exd consensus site (**Figure 6*d***). Arg3 does not form direct base contacts but instead forms a hydrogen bond with His -12, which, in turn, contacts the bases through a water-mediated hydrogen bond. Mutagenesis studies have shown that Arg3 plays a critical role in Scr in vivo specificity (29).

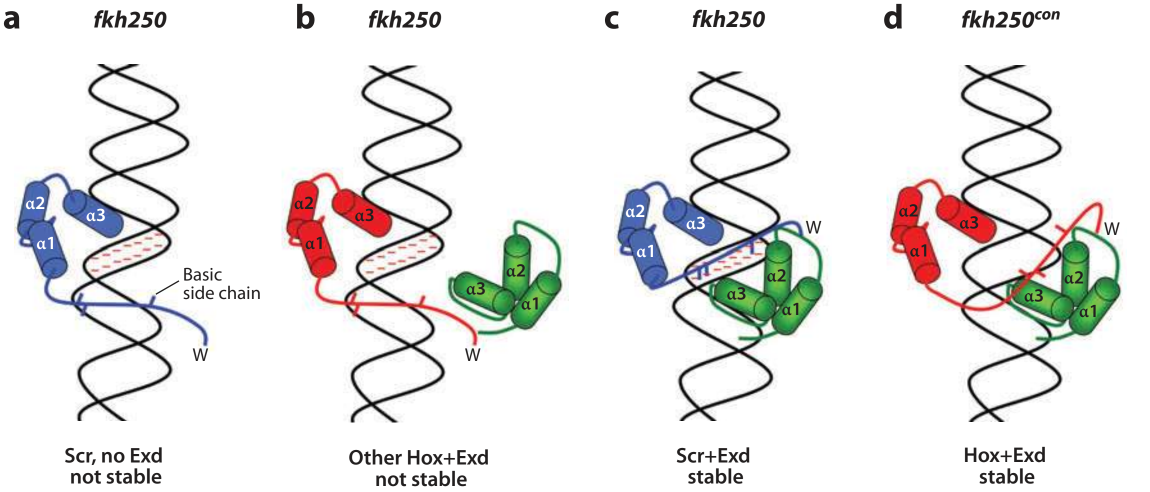

Figure 6

Hox DNA-binding specificity mediated by local shape recognition. All panels show either the *fkh250*-binding site or the *fkh250^{con}*-binding site. *fkh250*, but not *fkh250^{con}*, has two minor groove minima, which creates a more negative electrostatic potential (*minus signs*). The capital letter W refers to the Hox YPWM motif, which makes a direct contact with the cofactor Exd. See Reference 29 for details. (*a*) In the absence of Exd, Scr does not bind with high affinity to *fkh250* because basic side chains (*small bars*), in particular, arginines, on the N-terminal arm and linker of Scr are not positioned correctly. (*b*) Other Hox proteins do not bind well to *fkh250* even in the presence of Exd because their N-terminal arms and linker regions have different sequences. (*c*) The Scr-Exd heterodimer binds well to *fkh250* because the Scr N-terminal arm and linker region have the correct residues, and Exd positions them correctly by binding the YPWM motif (W). (*d*) Other Hox-Exd heterodimers bind well to *fkh250^{con}*. This binding site is not as selective because it has a less negative electrostatic potential. Thus, the sequences of the Hox N-terminal arms and linker regions are not as important for binding.

The Scr-specific and Hox-Exd consensus sites differ in minor groove shape, a structural feature that appears to be intrinsic to these sequences. These local variations in shape result in the enhancement of negative electrostatic potential at distinct positions that attract arginines into the minor groove (**Figure 2*c*,*d***) (20, 29). The Scr N-terminal arm uses these sequence-dependent variations in shape and electrostatic potential to achieve DNA-binding specificity (**Figures 6*c*,*d***) (150). Because narrow minor grooves are often associated with AT-rich sequences (**Table 2**), enhancement of negative electrostatic potential in the minor groove, which, in turn, is recognized by arginines, offers a general mechanism for sequence-specific recognition of DNA shape (66).

In addition to the results for Scr, mutagenesis studies on the Hox protein Ultrabithorax (Ubx) also suggest a role for linker and N-terminal arm residues in DNA-binding specificity, even when Ubx binds as a monomer (151, 152). Although no crystal structures are yet available to visualize these interactions, an intriguing possibility is that these residues may be reading differences in minor groove shape.

The use of arginines to bind to narrow regions of the minor groove is widespread among DNA-binding proteins (66). However, the manner in which the arginines are presented to the minor groove can differ (**Figure 7**). In the case of Scr-Exd, heterodimer formation between these two homeodomain proteins is necessary to position Arg3 and His-12, which are normally on an unstructured part of the Hox protein, so that these side chains can insert into the minor groove (**Figure 7*a***). In the case of the POU domain protein Brn-5 binding to its element *CRH-II*, the arginines that insert into a narrow region of the minor groove come from the linker region that separates the $\mathrm{POU_{HD}}$ from the $\mathrm{POU_S}$ domain (59). Thus, as with Scr-Exd, two DNA-binding domains are required to position the Brn-5 arginines, but in this case, both domains are in the same protein (**Figure 7*b***). Not all POU proteins use this method to position the relevant arginines (153). For example, the Oct-1/PORE complex uses the Arg2 and Arg5 side chains of two Oct-1 monomers to bind to two short A-tracts, ATTT and AAAT (154), and a Pit-1 dimer binds to DNA in a fashion similar to the Oct-1 dimer but uses Arg49 of the POU-specific domain to distinguish its ATAC site from the ATGC site of the Oct-1 dimer (153).

Proteins from families other than homeodomains also use the mechanism of local minor groove shape readout (66). The LEAFY gene regulator, for example, binds as a homodimer in which arginines present on the N-terminal arm of both monomers bind two distant narrow minor groove regions (155). In comparison, MogR uses arginines on a C-terminal extension from both monomers to contact a narrow minor groove composed of two antiparallel A-tracts that are separated by a TpA hinge step (**Figure 7*c***) (36). The γδ resolvase forms an arginine contact to a narrow minor groove with its N-terminal extension and uses another N-terminal arginine to contact the major groove (**Figure 7*d***) (156).

In all of the above examples, the arginines that insert into the minor groove come from otherwise unstructured strands that must be positioned owing to heterodimerization (Scr-Exd) or homodimerization (Oct-1, MogR), or via two adjacent DNA-binding domains in the same protein (Brn-5). Arginines that insert into minor grooves can also be integral to DNA-binding domains. For example, MEF2A, from the myocyte enhancer factor-2 family, uses its α1 helix, which is positioned on top of the minor groove, to contact the *MEF2A* minor groove (**Figure 7*e***) (157).

Minor groove-interacting arginines are often presented as part of short sequence motifs that include more than one arginine, such as RQR in Scr (29), RPR in Engrailed (26), RKKR in POU homeodomains (158), and RGHR in MATα2 (144). The observation that arginine-rich motifs bind to the minor groove was also made for the phage 434 repressor (KRPR) (**Figures 2*c*,*d***) and the Hin recombinase (GRPR), for which arginine mutations were shown to have a dramatic effect

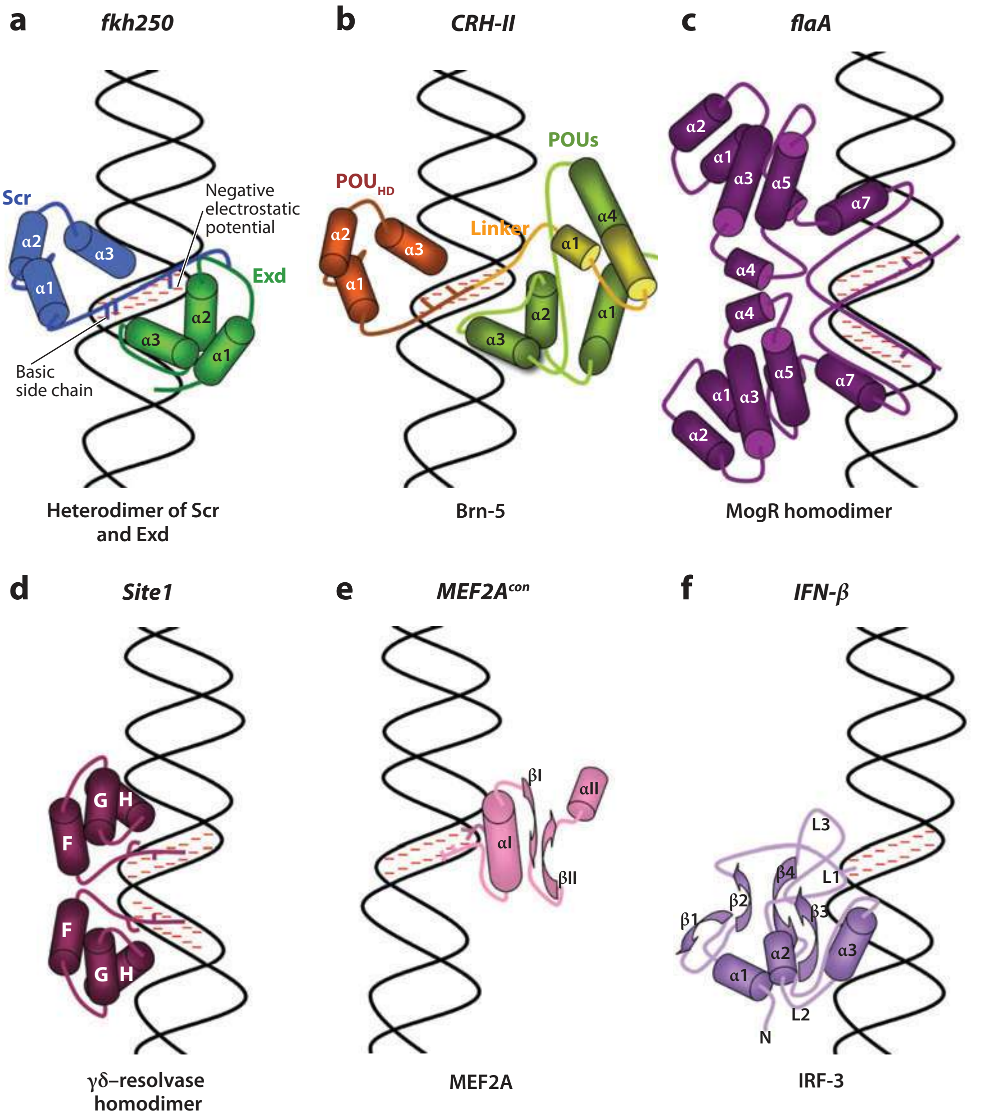

Figure 7

Examples of minor groove shape recognition. Each panel shows a different example in which basic side chains (*colored bars*) bind to minor grooves. (*a*) Arginine residues present on Scr's N-terminal arm and linker region require heterodimerization with Exd to be positioned correctly to insert into a narrow minor groove region of *fkh250* (PDB ID 2r5z). (*b*) Arginine residues present on the linker region that separates POU_{HD} from POU_{S} of Brn-5 insert into a narrow minor groove of the *CRH-II*-binding site (PDB ID 3d1n). (*c*) Arginine residues present on a C-terminal extension of a MogR homodimer insert into narrow regions of the *flaA*-binding site (PDB ID 3fdq). (*d*) An N-terminal extension from the γδ resolvase has an arginine that inserts into a narrow minor groove and a second arginine that inserts into the major groove of its binding site (PDB ID 1gdt) (*e*) MEF2A recognizes a narrow minor groove of the *MEF2A*-binding site via an arginine and glycine present on an N-terminal strand and via a lysine present on α-helix αI (PDB ID 1egw). (*f*) A histidine residue of IRF-3 inserts into a narrow minor groove region of the *IFN-β* enhancer (PDB ID 1t2k).

on binding affinity (159). The RQR motif of Scr introduces its arginines like a fork into the minor grove, with the glutamine pointing away from the DNA like the fork's handle (29). Other arginine-rich motifs orient the arginine side chains differently, allowing them to recognize distinct minor groove shapes.

Unlike homeodomain proteins, which rely on both major and minor groove interactions to achieve specificity, the architectural proteins TBP, SRY, LEF-1, IHF, and HMG-I(Y) only contact the minor groove. For example, the N-terminal arm of IHF inserts two arginines deep into a narrow region of the minor groove complemented by a third arginine that contacts a different narrow region (141). HMG-I(Y) proteins bind to AT-rich minor grooves but, in contrast to IHF, stabilize essentially straight instead of deformed DNA (141).

Although arginine is the most abundant residue that inserts into minor grooves, lysines can also be observed in such regions, although at a much lower frequency (66). The difference between these two basic amino acids is due, at least in part, to the higher free energy associated with removing lysines, which have a less delocalized positive charge distribution, from the aqueous phase (66). The importance of solvation effects is illustrated by the IFN-β enhanceosome structure, which exhibits a number of lysines in the proximity of the minor groove, clearly solvated rather than intruding into the groove (142, 143, 160). However, the enhanceosome uses histidines (from IRF-3 and IRF-7) to penetrate narrow minor groove regions formed by A-tracts (142, 143, 160). His40 of IRF-1, which is conserved across the IRF family, also inserts into narrow minor groove regions (**Figure 7*f***) (161, 162). A histidine is also observed to insert into the minor groove in the Scr-Exd-*fkh250* structure (29).

4.3.1.2. Major groove shape.

There are indications that sequence-dependent major groove shape is, like minor groove shape, also used as a readout mechanism. Indeed, minor and major groove geometries are correlated with each other (163). The human regulatory factor hRFX1 is a wHTH protein, which recognizes the DNA major groove with its β-hairpin wing in place of the recognition helix used by other wHTH proteins (42). In turn, hRFX1 protein places its H3 helix over the minor groove, from which a single lysine contacts the groove (42). The minor groove widens, resulting in a narrowing of the major groove that, in turn, improves major groove shape complementarity (38). In another example, domain 4 of the *E. coli* extracytoplasmic function σ factor, σ^E, specifically recognizes the GG$\underline{\text{AA}}$CTT element on the basis of major groove shape complementarity, which is achieved by narrowing the minor groove (164). The AT base pairs in the σ^E-binding site (underlined), which are highly conserved despite a lack of strong base contacts, are located in the center of a narrow minor groove (164) and were shown in genetic screening experiments to inhibit transcription when mutated (165).

4.3.1.3. Kinks.

As discussed above, DNA kinks occur when the linearity of the helix is abruptly broken, often owing to the unstacking of a flexible base pair step, such as TpA (**Table 2**). Kinks can contribute to binding specificity by promoting conformations that optimize protein-DNA and protein-protein contacts. As an example, the conformational flexibility of the ATA region allows the Tramtrack-binding site to adjust to the contacting zinc finger (166). DNA recognition by endonuclease EcoRV also depends upon the deformability of a TpA step (167). The binding site of the γδ resolvase comprises a central TATA element and exhibits kinks at both TpA steps (156). The flexibility intrinsic to TpA steps also plays a role in the specific binding of the RevErb nuclear hormone receptor as it binds to a site that contains two TpA steps (168). Although neither of these steps engage in base-specific contacts with RevErb, they show different degrees of deformation, indicating the importance of their flexibility.

The DNA-binding site of the catabolite activator protein (CAP) shows dramatic kinks at two CpA (TpG) steps (16, 169), which cause, along with two additional smaller kinks, an overall bending of the DNA of about 90° around the protein (170, 171). The kink at the CpA (TpG) step makes it possible for an arginine residue to engage in partial stacking interactions with a thymine (124). The HincII endonuclease recognizes its cognate site GTYRAC on the basis of the deformability of its central YpR step and shows the highest affinity when this step is CpG (172). Similarly, the binding of the EcoRI endonuclease to the Dickerson dodecamer involves a kink at the center of its binding site (173).

4.3.1.4. Intercalation. Owing to weaker stacking interactions, kinks are often stabilized through the intercalation of protein side chains, which, in turn, causes further deformation of the DNA helix. The specific DNA-binding site of the Lac repressor adjusts to the protein by forming a kink of about 36° at its central CpG step, which widens the minor groove where two leucine residues interact with the kinked base pair step through partial intercalation (**Figure 2*b***) (135). By contrast, a nonspecific DNA sequence, which has been designed to be different in all positions from the Lac operator, does not form a kink upon binding to the Lac repressor, but the protein rearranges its backbone and side chain conformations to engage in phosphate contacts (174). When the purine repressor is bound to its cognate site GCAAACGTTTGC, a similar kink is observed at its central CpG step (underlined) and is stabilized by the partial intercalation of two leucine residues from the minor groove side (31). Although the conformations of the flanking A-tract regions are very similar in the structures of free and PurR-bound DNA, a kink is not observed in the unbound site (116). This observation argues that, in this case, it is not DNA structure per se but its deformability that is recognized by PurR.

The yeast TBP structure shows phenylalanine intercalations in the first and last base pair step (underlined) of its TATATAAA-binding site (14). Whereas the first intercalation site is a flexible TpA step, the second site is likely determined by spacing (141, 148). Architectural proteins that intercalate hydrophobic amino acids between base pairs from the minor groove are the HMG box proteins SRY and LEF-1 (141). These intercalating hydrophobic residues are conserved in HMG domains and are usually flanked by basic amino acids (175). SRY and LEF-1 both use Asn10 to convey specificity through tripartite polar contacts with base pairs preceding the intercalation pocket. Closely related to SRY, DNA-bending SOX domains represent another subgroup of HMG boxes (176). The SOX2-Oct-4-DNA ternary complex is characterized by the intercalation of methionine and phenylalanine residues into an ApA (TpT) step inducing a kink (154). The SOX17 protein also uses its HMG domain to cause a drastic kink of an ApA (TpT) step through the intercalation of a phenylalanine-methionine dipeptide (177).

4.3.2. Global shape readout. We include in this category the recognition of DNA sequences where the entire binding site is not in a classic B-form helix. Examples are the recognition of bent DNA, where the curvature is distributed along the entire helix, A-DNA, sequences that have elements of both A- and B-DNA, and Z-DNA (**Figure 4**).

4.3.2.1. Bent DNA. The papillomavirus E2 protein provides a clear example of DNA bending playing a role in protein-DNA recognition. The E2 protein binds as a dimer to two half sites separated by a linker of four base pairs (87, 178). Although only the underlined half sites of the ACCGN_4CGGT consensus-binding site are contacted by the protein, the variable linker optimizes these contacts through bending, which, in turn, enhances interactions between the protomers of the E2 dimer (13, 82). The DNA is similarly bent in complex with the E2 proteins of the bovine papillomavirus BPV-1 (13) and the human papillomavirus HPV-18 (**Figure 2*b***) (179). However, whereas the BPV-1 E2 protein binds with similar binding affinity to consensus sites with various linker sequences, the HPV-18 E2 protein shows a strong preference for AATT linkers (180), and the HPV-16

E2 protein shows a preference for AATT and AAAA linkers (178). X-ray crystallographic studies and Monte Carlo predictions stressed that the E2-binding site with AATT linker is also bent when not bound to the protein, whereas the site with ACGT linker is essentially straight (**Figure 3*b***) (82, 87, 120). A correlation of the structural data with binding studies suggests that high-affinity sites are prebent as seen in the E2-DNA complex, but low-affinity sites require the protein to induce the site to bend (178, 179).

Bending was also suggested to play a role in the specificity of homeodomains by facilitating contacts with the recognition helix (181). Specific DNA recognition by the phage 434 repressor is associated with the bending of its operator (149), which decreases with the number of G:C base pairs in its operator sequence (182). Long A-tracts are associated with bending and are present, for instance, in the binding sites of the MATa1-MATα2 heterodimer (144) and the NF-κB protein (48). The conformation of the NF-κB-binding site in its bound state is similar to the bending already present in its free state (117, 183). The RXR-RAR heterodimer recognizes the same half sites as the RXR homodimer. However, the smooth bending of the AAA region between both half sites in place of the kink induced by the RXR homodimer contributes to RXR-RAR specificity (134). The restriction endonucleases BglII and BamHI recognize DNA sites, AGATCT and GGATCC, with an identical core region (underlined), but bending differentiates both binding sites (184–186). In contrast, the similar binding sites of the endonucleases MunI and EcoRI, CAATTG and GAATTC, respectively, cannot be distinguished through bending and require an arginine contact to read the outer C:G base pair (186).

4.3.2.2. A-DNA. Whereas sugars are usually buried in the minor groove of B-DNA, they are exposed in A-DNA and provide about 50% of the protein-DNA interface in the TBP-DNA complex, where the DNA is in an A-form conformation (14). Although arginine and lysine frequently interact with nucleotides in B-DNA conformations, nonpolar amino acids, such as alanine, leucine, phenylalanine, and valine, contact nucleotides in A-DNA conformations (187). These types of contacts are thus associated with GC-rich sequences (76, 77, 188) and with TATA boxes (**Table 2**) (189). The higher accessibility of C3′-endo sugars of A-DNA in comparison to buried C2′-endo sugars of B-DNA (187) also contributes to the specificity of zinc finger proteins for GC-rich sequences (116) and of the TBPs for TATA boxes (78).

The transition from B-DNA to A-DNA that transforms the sugar conformations and widens the minor groove is often associated with the intrusion of hydrophobic residues into the minor groove (190). B- to A-transitions are often observed in complexes with endonucleases because A-DNA makes the phosphate oxygen of the bond that is cleaved more accessible (75). Other proteins that recognize A/B-intermediate conformations are the Trp repressor and the *Caenorhabditis elegans* Tc3 transposase (75). The transcription factor for polymerase IIIA (TFIIIA) also binds to an A-DNA-like binding site (191). In general, zinc finger proteins tend to bind A/B-intermediates in major grooves that are deep like A-DNA and wide like B-DNA (119) and that have the increased helix diameter typical for A-DNA (192). Zinc fingers from the human glioblastoma protein (GLI) show the base pair inclination that is distinct for A-DNA (193). In other complexes, only a limited number of base pairs exhibit A-DNA conformations, whereas the remaining site resembles B-DNA, as seen in two regions of the I-PolI-binding site (75).

Interestingly, binding sites of the mouse Cys_2His_2 zinc finger protein Zif268 crystallize in A-like conformations when both unbound and bound by the protein (**Figure 3*a***) (116, 118, 119). These observations suggest that this DNA sequence has an intrinsic tendency to assume an A-like conformation and that exposed hydrophobic surfaces of A-like sugars may be generally recognized by zinc fingers (191). Another example of the recognition of a DNA that has an A/B intermediate structure is the

Runt domain and its binding site (50). In this case, the unbound binding site was observed both in A-DNA (194) and B-DNA (121) conformations. Perhaps related to such observations is that some transcription factors, such as TFIIIA, Bicoid, and p53, bind to both DNA and RNA; the latter almost exclusively exhibits A-form topology (195).

4.3.2.3. Z-DNA. The zigzag positioning of phosphates along a left-handed Z-DNA helix is specifically recognized by the double-stranded RNA adenosine deaminase (ADAR1), which is an RNA-editing enzyme with a wHTH motif (196). Z-DNA structures have only been observed to form with purine-pyrimidine alternating sequences that can adopt a left-handed helix (79, 80, 197). The Zα-domain of ADAR1 has a conformation tailored to recognize a row of five phosphates in one zigzag-shaped backbone of Z-DNA. Because the tumor-associated DLM-1 protein also recognizes Z-DNA via five phosphates along a zigzag-shaped left-handed strand, phosphate positions seem to be the signature code recognized by Z-DNA-binding proteins (**Figures 1*d*,*h***) (198).

5. EXAMPLES OF HIGHER-ORDER PROTEIN-DNA INTERACTIONS

The above discussion highlights examples that illustrate specific readout mechanisms and thus provides a reductionist perspective on DNA recognition. However, individual DNA-binding proteins combine many, if not most, of these readout mechanisms to achieve the correct affinity and specificity required for function. To illustrate this, below we discuss a few examples of protein-DNA recognition in which combinations of readout mechanisms are clearly used.

5.1. The Nucleosome

The presence of nucleosomes in eukaryotic genomes profoundly affects the activity of transcription factors and other DNA-binding proteins (199–201). Although some factors can bind to nucleosomal DNA, others can only bind nucleosome-free DNA. For instance, the packaging of DNA in nucleosomes is expected to narrow the minor groove of TATA boxes, thus precluding TBP binding (148). In contrast, the bending of nucleosomal DNA was suggested to assist p53 binding at the DNA surface facing away from the histone core (91). Owing to the intimate relationship between protein-DNA recognition and nucleosome binding, attempts to predict nucleosome positions in genomic DNA have received a great deal of attention (202–205). Because DNA deformability (kinks), DNA bending, and local shape recognition all contribute to nucleosome positioning, these mechanisms need to be considered in any prediction algorithm.

The bendability of short sequences accommodates the wrapping of DNA around the histone core in the nucleosome (148, 149). The presence of short A-tracts of only three A:T base pairs stabilizes the deformation required for regions of the nucleosomal DNA facing the histones, where the minor groove is compressed (66, 206). Consequently, the distribution of short A-tracts in yeast in vivo sequences reflects the periodicity of a helical turn in congruence with the structural periodicity caused by the wrapping of nucleosomal DNA around the histone core (66). In addition, kinks caused by CpA steps adjacent to short A-tracts can enhance the overall curvature in regions where the minor groove faces the histones (91). And, because of their flexibility, the kinks resulting from TpA steps are also used to help wrap the DNA around the histone core. Taking both observations together, the deformability of short A-tracts and YpR steps provides more information about sequence periodicities than was originally observed for dinucleotides (202, 207–209).

The periodicity of short A-tracts in nucleosomal DNA also results in a periodic narrowing of the minor groove, which is, in turn, read by arginines present at the histone-DNA interface (66). Nucleosome-bound DNA contains, on average, 10 of these intrinsically narrow minor groove regions, most of which are likely to be contacted by arginines. Thus, in addition to DNA kinks and bends, nucleosome-DNA interactions also rely on the recognition of local variations in DNA shape (66).

5.2. *Escherichia coli* IHF

A combination of kinking, bending, and intercalation is used to achieve DNA-binding specificity for the *E. coli* nucleoid protein IHF, which also functions as a transcriptional activator (210). The IHF α/β heterodimer sharply bends DNA by about 160° to bring distant binding sites of the λ repressor into close proximity (211). IHF recognizes three DNA sites: TATCAA in the central region of its binding site, a 6-bp A-tract, and a TTG region at its flanks (210). The large bending is partially induced by the A-tract with its intrinsically narrow minor groove at one side of the IHF-DNA complex (212). On the other side of the complex, the TpG (CpA) step in the TTG element narrows the minor groove through kinking, which is recognized through the insertion of βArg46 (213). The TTG to TAG mutation, which shifts the YpR step 5′ by 1 bp, indicates that the IHF protein discriminates between A:T and T:A base pairs in this region due to the flexibility of the YpR step (213). The α-arm of the protein contacts the minor groove of the central consensus element with three arginine residues. Two large kinks at the ApA (TpT) steps caused by proline intercalations are the main contributors to the U-formed shape of the IHF-bound DNA (211).

5.3. Cooperativity

DNA-binding proteins often bind DNA cooperatively to create higher-order nucleoprotein complexes that reflect the combinatorial control of gene expression. DNA-binding cooperativity is most typically attributed to direct protein-protein interactions between adjacent DNA-binding factors that promote the assembly of higher-order complexes. Notable examples are Hox-Exd/Pbx heterodimers (28, 29, 214), the MATa1-MATα2 heterodimer (144), and the NFAT-Fos-Jun heterotrimer (215). Whereas cofactors in all of the previous examples directly bind to DNA, the cofactor CBFβ enhances the binding of the *Drosophila* Runt domain to DNA without forming any DNA contact (50).

In addition to this classical form of cooperativity, a sequence-dependent DNA structure may also promote the cooperative binding of multiple factors. One particularly striking example is the assembly of the IFN-β enhanceosome, which is composed of at least eight DNA-binding proteins: a heterodimer of ATF-2/c-Jun, a heterodimer of p50/Rel, and four IRF monomers, all bound to a highly conserved ∼55-bp element (142, 143). In addition, the architectural protein HMGA1 binds, perhaps transiently, in the minor groove to at least two positions, inducing DNA bends that facilitate the assembly of the enhanceosome (216). Remarkably, despite the binding of eight transcription factors, a paucity of protein-protein interactions is observed, arguing that cooperativity is likely to be achieved in some other manner (142). One appealing suggestion is that the final DNA structure, which is optimized for enhanceosome assembly, depends on the intrinsic deformability of the DNA (160). According to this view, the binding of each factor improves the binding of the other factors through an effect on DNA structure. This idea follows logically from many of the other examples described above where DNA shape and deformability contribute to specificity on a smaller scale. Thus, if correct, a sequence-dependent DNA structure may be a critical component in the binding not only of individual factors to their binding sites, but also in the assembly of higher-order, multiprotein complexes. This idea fits well with another recent observation that was also pointed out at the beginning of this review, namely, that DNA shape is under evolutionary selection and provides a better indicator of functional elements than conservation of the linear DNA sequence (7).

SUMMARY POINTS

1. DNA-binding proteins use a wide range of mechanisms to bind specifically to binding sites.
2. The three-dimensional structure of the binding site must be taken into consideration when understanding binding specificity.
3. The main readout mechanisms are (*a*) the recognition of bases and (*b*) the recognition of DNA shape.
4. The recognition of bases can be further subdivided into those interactions that occur in the major groove, which provides the greatest potential for specificity, and those that occur in the minor groove.
5. The recognition of DNA shape can be further subdivided into the recognition of local shape variation (e.g., minor groove width) and the recognition of global shape variation (e.g., bent DNA).
6. Any one DNA-binding protein is likely to use a combination of readout mechanisms.
7. Readout mechanisms are often interrelated (e.g., bending toward the minor groove also narrows it).
8. The formation of higher-order protein-DNA complexes may depend on sequence-dependent DNA structures that are optimized to promote assembly.

FUTURE ISSUES

1. The annotation of genomes must take into account DNA structure.
2. The rules governing the relationships between DNA sequence and DNA structure need to be better understood.
3. Understanding intrinsic versus induced effects on DNA structure is an important goal and would benefit from additional structural analyses of free DNAs.
4. Understanding the rules governing binding specificity within a protein family would benefit from comparisons of structures of multiple family members, each bound to specific and nonspecific binding sites.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) grants GM54510 (R.S.M.) and U54 CA121852 (B.H. and R.S.M.). The authors thank Z. Shakked, T. Tullius, T. Haran, and A. Aggarwal for helpful conversations.

LITERATURE CITED

1. Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171:737–38
2. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133:1266–76
3. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133:1277–89
4. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* 32:878–87
5. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 19:556–66
6. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* 324:1720–23
7. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324:389–92
8. Greenbaum JA, Pang B, Tullius TD. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.* 17:947–53
9. Rosenberg JM, Seeman NC, Kim JJ, Suddath FL, Nicholas HB, Rich A. 1973. Double helix at atomic resolution. *Nature* 243:150–54
10. Seeman NC, Rosenberg JM, Rich A. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA* 73:804–8
11. Viswamitra MA, Kennard O, Jones PG, Sheldrick GM, Salisbury S, et al. 1978. DNA double helical fragment at atomic resolution. *Nature* 273:687–88
12. Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, et al. 1988. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* 335:321–29
13. Hegde RS, Grossman SR, Laimins LA, Sigler PB. 1992. Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* 359:505–12
14. Kim Y, Geiger JH, Hahn S, Sigler PB. 1993. Crystal structure of a yeast TBP/TATA-box complex. *Nature* 365:512–20
15. Kim JL, Nikolov DB, Burley SK. 1993. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* 365:520–27
16. Lawson CL, Berman HM. 2008. Indirect readout of DNA sequence by proteins. In *Protein-Nucleic Acid Interactions: Structural Biology*, ed. PA Rice, CC Correll, pp. 66–90. Cambridge, UK: R. Soc. Chem.
17. Garvie CW, Wolberger C. 2001. Recognition of specific DNA sequences. *Mol. Cell* 8:937–46
18. Luscombe NM, Austin SE, Berman HM, Thornton JM. 2000. An overview of the structures of protein-DNA complexes. *Genome Biol.* 1:REVIEWS001
19. Hong M, Marmorstein R. 2008. Structural basis for sequence-specific DNA recognition by transcription factors and their complexes. See Ref. 16, pp. 47–65
20. Rohs R, West SM, Liu P, Honig B. 2009. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* 19:171–77
21. McKay DB, Steitz TA. 1981. Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. *Nature* 290:744–49
22. Anderson WF, Ohlendorf DH, Takeda Y, Matthews BW. 1981. Structure of the Cro repressor from bacteriophage lambda and its interaction with DNA. *Nature* 290:754–58
23. Pabo CO, Lewis M. 1982. The operator-binding domain of lambda repressor: structure and DNA recognition. *Nature* 298:443–47
24. Jordan SR, Pabo CO. 1988. Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions. *Science* 242:893–99
25. Badia D, Camacho A, Perez-Lago L, Escandon C, Salas M, Coll M. 2006. The structure of phage phi29 transcription regulator p4-DNA complex reveals an N-hook motif for DNA. *Mol. Cell* 22:73–81

26. Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. 1990. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. *Cell* 63:579–90

27. Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. 1991. Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* 67:517–28

28. Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK. 1999. Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* 397:714–19

29. Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, et al. 2007. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* 131:530–43

30. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–40

31. Schumacher MA, Choi KY, Zalkin H, Brennan RG. 1994. Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science* 266:763–70

32. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, et al. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 271:1247–54

33. Van Roey P, Waddling CA, Fox KM, Belfort M, Derbyshire V. 2001. Intertwined structure of the DNA-binding domain of intron endonuclease I-TevI with its substrate. *EMBO J.* 20:3631–37

34. Edgell DR, Derbyshire V, Van Roey P, LaBonne S, Stanger MJ, et al. 2004. Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. *Nat. Struct. Mol. Biol.* 11:936–44

35. Shen BW, Landthaler M, Shub DA, Stoddard BL. 2004. DNA binding and cleavage by the HNH homing endonuclease I-HmuI. *J. Mol. Biol.* 342:43–56

36. Shen A, Higgins DE, Panne D. 2009. Recognition of AT-rich DNA binding sites by the MogR repressor. *Structure* 17:769–77

37. Daniels DS, Woo TT, Luu KX, Noll DM, Clarke ND, et al. 2004. DNA binding and nucleotide flipping by the human DNA repair protein AGT. *Nat. Struct. Mol. Biol.* 11:714–20

38. Gajiwala KS, Burley SK. 2000. Winged helix proteins. *Curr. Opin. Struct. Biol.* 10:110–16

39. Clark KL, Halay ED, Lai E, Burley SK. 1993. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* 364:412–20

40. Kodandapani R, Pio F, Ni CZ, Piccialli G, Klemsz M, et al. 1996. A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. *Nature* 380:456–60

41. Hong M, Fuangthong M, Helmann JD, Brennan RG. 2005. Structure of an OhrR-*ohrA* operator complex reveals the DNA binding mechanism of the MarR family. *Mol. Cell* 20:131–41

42. Gajiwala KS, Chen H, Cornille F, Roques BP, Reith W, et al. 2000. Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature* 403:916–21

43. Ferre-D'Amare AR, Pognonec P, Roeder RG, Burley SK. 1994. Structure and function of the b/HLH/Z domain of USF. *EMBO J.* 13:180–89

44. Ma PC, Rould MA, Weintraub H, Pabo CO. 1994. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell* 77:451–59

45. Nair SK, Burley SK. 2003. X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* 112:193–205

46. Parraga A, Bellsolell L, Ferre-D'Amare AR, Burley SK. 1998. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 A resolution. *Structure* 6:661–72

47. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. 1994. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 265:346–55

48. Ghosh G, van Duyne G, Ghosh S, Sigler PB. 1995. Structure of NF-$_\kappa$B p50 homodimer bound to a $_\kappa$B site. *Nature* 373:303–10

49. Muller CW, Rey FA, Sodeoka M, Verdine GL, Harrison SC. 1995. Structure of the NF-kappa B p50 homodiner bound to DNA. *Nature* 373:311–17

50. Tahirov TH, Inoue-Bungo T, Morii H, Fujikawa A, Sasaki M, et al. 2001. Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell* 104:755–67

51. Kovall RA, Hendrickson WA. 2004. Crystal structure of the nuclear effector of Notch signaling, CSL, bound to DNA. *EMBO J.* 23:3441–51

52. Sidote DJ, Barbieri CM, Wu T, Stock AM. 2008. Structure of the *Staphylococcus aureus* AgrA LytTR domain bound to DNA reveals a beta fold with an unusual mode of binding. *Structure* 16:727–35

53. Pavletich NP, Pabo CO. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* 252:809–17

54. Schreiter ER, Drennan CL. 2007. Ribbon-helix-helix transcription factors: variations on a theme. *Nat. Rev. Microbiol.* 5:710–20

55. Somers WS, Phillips SE. 1992. Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. *Nature* 359:387–93

56. Raumann BE, Rould MA, Pabo CO, Sauer RT. 1994. DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature* 367:754–57

57. Pingoud A, Fuxreiter M, Pingoud V, Wende W. 2005. Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.* 62:685–707

58. Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. 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

59. Pereira JH, Kim SH. 2009. Structure of human Brn-5 transcription factor in complex with CRH gene promoter. *J. Struct. Biol.* 167:159–65

60. Rhee S, Martin RG, Rosner JL, Davies DR. 1998. A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl. Acad. Sci. USA* 95:10413–18

61. Chen FE, Huang DB, Chen YQ, Ghosh G. 1998. Crystal structure of p50/p65 heterodimer of transcription factor NF-κB bound to DNA. *Nature* 391:410–13

62. Kwon HJ, Bennik MH, Demple B, Ellenberger T. 2000. Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. *Nat. Struct. Biol.* 7:424–30

63. Muller CW. 2001. Transcription factors: global, detailed views. *Curr. Opin. Struct. Biol.* 11:26–32

64. Chang MV, Chang JL, Gangopadhyay A, Shearer A, Cadigan KM. 2008. Activation of wingless targets requires bipartite recognition of DNA by TCF. *Curr. Biol.* 18:1877–81

65. Shakked Z, Rabinovich D. 1986. The effect of the base sequence on the fine structure of the DNA double helix. *Prog. Biophys. Mol. Biol.* 47:159–95

66. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. 2009. The role of DNA shape in protein-DNA recognition. *Nature* 461:1248–53

67. Honig B, Nicholls A. 1995. Classical electrostatics in biology and chemistry. *Science* 268:1144–49

68. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. 1986. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins* 1:47–59

69. Sharp KA, Honig B, Harvey SC. 1990. Electrical potential of transfer RNAs: codon-anticodon recognition. *Biochemistry* 29:340–46

70. Chin K, Sharp KA, Honig B, Pyle AM. 1999. Calculating the electrostatic properties of RNA provides new insights into molecular interactions and function. *Nat. Struct. Biol.* 6:1055–61

71. Tang CL, Alexov E, Pyle AM, Honig B. 2007. Calculation of pK_as in RNA: on the structural origins and functional roles of protonated nucleotides. *J. Mol. Biol.* 366:1475–96

72. Leslie AG, Arnott S, Chandrasekaran R, Ratliff RL. 1980. Polymorphism of DNA double helices. *J. Mol. Biol.* 143:49–72

73. Jayaram B, Sharp KA, Honig B. 1989. The electrostatic potential of B-DNA. *Biopolymers* 28:975–93

74. Lavery R, Pullman B. 1981. The molecular electrostatic potential and steric accessibility of poly (dI.dC). Comparison with poly (dG.dC). *Nucleic Acids Res.* 9:7041–51

75. Lu XJ, Shakked Z, Olson WK. 2000. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* 300:819–40

76. Shakked Z, Guerstein-Guzikevich G, Eisenstein M, Frolow F, Rabinovich D. 1989. The conformation of the DNA double helix in the crystal is dependent on its environment. *Nature* 342:456–60

77. Ng HL, Kopka ML, Dickerson RE. 2000. The structure of a stable intermediate in the A ↔ B DNA helix transition. *Proc. Natl. Acad. Sci. USA* 97:2035–39

78. Guzikevich-Guerstein G, Shakked Z. 1996. A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nat. Struct. Biol.* 3:32–37

79. Wang AH, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, et al. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 282:680–86

80. Arnott S, Chandrasekaran R, Birdsall DL, Leslie AG, Ratliff RL. 1980. Left-handed DNA helices. *Nature* 283:743–45

81. Nelson HC, Finch JT, Luisi BF, Klug A. 1987. The structure of an oligo(dA)·oligo(dT) tract and its biological implications. *Nature* 330:221–26

82. Hizver J, Rozenberg H, Frolow F, Rabinovich D, Shakked Z. 2001. DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. USA* 98:8490–95

83. Haran TE, Mohanty U. 2009. The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* 42:41–81

84. Zhurkin VB, Tolstorukov MY, Xu F, Colasanti AV, Olson WK. 2005. Sequence-dependent variality of B-DNA: an update on bending and curvature. In *DNA Conformation and Transcription*, ed. T Ohyama, pp. 18–34. Georgetown, TX/New York: Landes Biosci./Springer Sci. Bus. Media

85. Goodsell DS, Kaczor-Grzeskowiak M, Dickerson RE. 1994. The crystal structure of C-C-A-T-T-A-A-T-G-G. Implications for bending of B-DNA at T-A steps. *J. Mol. Biol.* 239:79–96

86. Crothers DM, Shakked Z. 1999. DNA bending by adenine-thymine tracts. In *Oxford Handbook of Nucleic Acid Structures*, ed. S Neidle, pp. 455–70. London: Oxford Univ. Press

87. Rohs R, Sklenar H, Shakked Z. 2005. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* 13:1499–509

88. Gorin AA, Zhurkin VB, Olson WK. 1995. B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.* 247:34–48

89. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA* 95:11163–68

90. Mack DR, Chiu TK, Dickerson RE. 2001. Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.* 312:1037–49

91. Tolstorukov MY, Colasanti AV, McCandlish DM, Olson WK, Zhurkin VB. 2007. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.* 371:725–38

92. Lu XJ, Olson WK. 2008. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* 3:1213–27

93. Janin J, Rodier F, Chakrabarti P, Bahadur RP. 2007. Macromolecular recognition in the Protein Data Bank. *Acta Crystallogr. D* 63:1–8

94. Billeter M, Qian YQ, Otting G, Muller M, Gehring W, Wüthrich K. 1993. Determination of the nuclear magnetic resonance solution structure of an *Antennapedia* homeodomain-DNA complex. *J. Mol. Biol.* 234:1084–97

95. Ades SE, Sauer RT. 1995. Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry* 34:14601–8

96. Tucker-Kellogg L, Rould MA, Chambers KA, Ades SE, Sauer RT, Pabo CO. 1997. Engrailed (Gln50 → Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. *Structure* 5:1047–54

97. Grant RA, Rould MA, Klemm JD, Pabo CO. 2000. Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 → ala) complex at 2.0 A. *Biochemistry* 39:8187–92

98. Hanes SD, Brent R. 1989. DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell* 57:1275–83

99. Treisman J, Gonczy P, Vashishtha M, Harris E, Desplan C. 1989. A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* 59:553–62

100. Pabo CO, Sauer RT. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* 61:1053–95

101. Luscombe NM, Thornton JM. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* 320:991–1009

102. Siggers TW, Silkov A, Honig B. 2005. Structural alignment of protein-DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.* 345:1027–45
103. Cherney LT, Cherney MM, Garen CR, James MN. 2009. The structure of the arginine repressor from *Mycobacterium tuberculosis* bound with its DNA operator and co-repressor, L-arginine. *J. Mol. Biol.* 388:85–97
104. Ellenberger TE, Brandl CJ, Struhl K, Harrison SC. 1992. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell* 71:1223–37
105. Travers A. 1998. Transcription: activation by cooperating conformations. *Curr. Biol.* 8:R616–18
106. Weiss MA, Ellenberger T, Wobbe CR, Lee JP, Harrison SC, Struhl K. 1990. Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. *Nature* 347:575–78
107. Love JJ, Li X, Case DA, Giese K, Grosschedl R, Wright PE. 1995. Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature* 376:791–95
108. Laity JH, Dyson HJ, Wright PE. 2000. DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys_2-His_2 zinc fingers. *J. Mol. Biol.* 295:719–27
109. Holmbeck SM, Dyson HJ, Wright PE. 1998. DNA-induced conformational changes are the basis for cooperative dimerization by the DNA binding domain of the retinoid X receptor. *J. Mol. Biol.* 284:533–39
110. Lefstin JA, Yamamoto KR. 1998. Allosteric effects of DNA on transcriptional regulators. *Nature* 392:885–88
111. Meijsing SH, Pufall MA, So AY, Bates DL, Chen L, Yamamoto KR. 2009. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* 324:407–10
112. Luisi BF, Xu WX, Otwinowski Z, Freedman LP, Yamamoto KR, Sigler PB. 1991. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* 352:497–505
113. McClarin JA, Frederick CA, Wang BC, Greene P, Boyer HW, et al. 1986. Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. *Science* 234:1526–41
114. Drew HR, Wing RM, Takano T, Broka C, Tanaka S, et al. 1981. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. USA* 78:2179–83
115. Shakked Z, Guzikevich-Guerstein G, Frolow F, Rabinovich D, Joachimiak A, Sigler PB. 1994. Determinants of repressor/operator recognition from the structure of the *trp* operator binding site. *Nature* 368:469–73
116. Locasale JW, Napoli AA, Chen S, Berman HM, Lawson CL. 2009. Signatures of protein-DNA recognition in free DNA binding sites. *J. Mol. Biol.* 386:1054–65
117. Huang DB, Phelps CB, Fusco AJ, Ghosh G. 2005. Crystal structure of a free kappaB DNA: insights into DNA recognition by transcription factor NF-kappaB. *J. Mol. Biol.* 346:147–60
118. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. 1996. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* 4:1171–80
119. Elrod-Erickson M, Benson TE, Pabo CO. 1998. High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure* 6:451–64
120. Rozenberg H, Rabinovich D, Frolow F, Hegde RS, Shakked Z. 1998. Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc. Natl. Acad. Sci. USA* 95:15194–99
121. Bartfeld D, Shimon L, Couture GC, Rabinovich D, Frolow F, et al. 2002. DNA recognition by the RUNX1 transcription factor is mediated by an allosteric transition in the RUNT domain and by DNA bending. *Structure* 10:1395–407
122. Kitayner M, Rozenberg H, Kessler N, Rabinovich D, Shaulov L, et al. 2006. Structural basis of DNA recognition by p53 tetramers. *Mol. Cell* 22:741–53
123. Paillard G, Lavery R. 2004. Analyzing protein-DNA recognition mechanisms. *Structure* 12:113–22
124. Harrison SC, Aggarwal AK. 1990. DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* 59:933–69
125. Billeter M. 1996. Homeodomain-type DNA recognition. *Prog. Biophys. Mol. Biol.* 66:211–25
126. Konig B, Muller JJ, Lanka E, Heinemann U. 2009. Crystal structure of KorA bound to operator DNA: insight into repressor cooperation in RP4 gene regulation. *Nucleic Acids Res.* 37:1915–24

127. Tateno M, Yamasaki K, Amano N, Kakinuma J, Koike H, et al. 1997. DNA recognition by beta-sheets. *Biopolymers* 44:335–59

128. Coulocheri SA, Pigis DG, Papavassiliou KA, Papavassiliou AG. 2007. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie* 89:1291–303

129. Hoogsteen K. 1963. Crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.* 16:907–916

130. Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK. 1999. TATA element recognition by the TATA box–binding protein has been conserved throughout evolution. *Genes Dev.* 13:3217–30

131. Aishima J, Gitti RK, Noah JE, Gan HH, Schlick T, Wolberger C. 2002. A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.* 30:5244–52

132. Tainer JA, Cunningham RP. 1993. Molecular recognition in DNA-binding proteins and enzymes. *Curr. Opin. Biotechnol.* 4:474–83

133. Joachimiak A, Haran TE, Sigler PB. 1994. Mutagenesis supports water mediated recognition in the trp repressor-operator system. *EMBO J.* 13:367–72

134. Rastinejad F, Wagner T, Zhao Q, Khorasanizadeh S. 2000. Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1. *EMBO J.* 19:1045–54

135. Kalodimos CG, Biris N, Bonvin AM, Levandoski MM, Guennuegues M, et al. 2004. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* 305:386–89

136. Aggarwal AK, Rodgers DW, Drottar M, Ptashne M, Harrison SC. 1988. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* 242:899–907

137. Wolberger C, Dong YC, Ptashne M, Harrison SC. 1988. Structure of a phage 434 Cro/DNA complex. *Nature* 335:789–95

138. Watkins D, Hsiao C, Woods KK, Koudelka GB, Williams LD. 2008. P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. *Biochemistry* 47:2325–38

139. Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U. 2007. Common mode of DNA binding to cold shock domains. Crystal structure of hexathymidine bound to the domain-swapped form of a major cold shock protein from *Bacillus caldolyticus*. *FEBS J.* 274:1265–79

140. Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U. 2006. T-rich DNA single strands bind to a preformed site on the bacterial cold shock protein Bs-CspB. *J. Mol. Biol.* 360:702–14

141. Bewley CA, Gronenborn AM, Clore GM. 1998. Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* 27:105–31

142. Panne D, Maniatis T, Harrison SC. 2007. An atomic model of the interferon-beta enhanceosome. *Cell* 129:1111–23

143. Escalante CR, Nistal-Villan E, Shen L, Garcia-Sastre A, Aggarwal AK. 2007. Structure of IRF-3 bound to the PRDIII-I regulatory element of the human interferon-beta enhancer. *Mol. Cell* 26:703–16

144. Li T, Jin Y, Vershon AK, Wolberger C. 1998. Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.* 26:5707–18

145. Crane-Robinson C, Dragan AI, Privalov PL. 2006. The extended arms of DNA-binding domains: a tale of tails. *Trends Biochem. Sci.* 31:547–52

146. Privalov PL, Dragan AI, Crane-Robinson C, Breslauer KJ, Remeta DP, Minetti CA. 2007. What drives proteins into the major or minor grooves of DNA? *J. Mol. Biol.* 365:1–9

147. Privalov PL, Dragan AI, Crane-Robinson C. 2009. The cost of DNA bending. *Trends Biochem. Sci.* 34:464–70

148. Patikoglou G, Burley SK. 1997. Eukaryotic transcription factor-DNA complexes. *Annu. Rev. Biophys. Biomol. Struct.* 26:289–325

149. Travers AA. 1989. DNA conformation and protein binding. *Annu. Rev. Biochem.* 58:427–52

150. Mann RS, Lelli KM, Joshi R. 2009. Hox specificity: unique roles for cofactors and collaborators. *Curr. Top. Dev. Biol.* 88:63–101

151. Liu Y, Matthews KS, Bondos SE. 2009. Internal regulatory interactions determine DNA binding specificity by a Hox transcription factor. *J. Mol. Biol.* 390:760–74

152. Liu Y, Matthews KS, Bondos SE. 2008. Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the *Drosophila* Hox protein Ultrabithorax. *J. Biol. Chem.* 283:20874–87

153. Phillips K, Luisi B. 2000. The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.* 302:1023–39

154. Remenyi A, Lins K, Nissen LJ, Reinbold R, Scholer HR, Wilmanns M. 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* 17:2048–59

155. Hames C, Ptchelkine D, Grimm C, Thevenon E, Moyroud E, et al. 2008. Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO J.* 27:2628–37

156. Yang W, Steitz TA. 1995. Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. *Cell* 82:193–207

157. Santelli E, Richmond TJ. 2000. Crystal structure of MEF2A core bound to DNA at 1.5 A resolution. *J. Mol. Biol.* 297:437–49

158. Remenyi A, Tomilin A, Pohl E, Lins K, Philippsen A, et al. 2001. Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. *Mol. Cell* 8:569–80

159. Churchill ME, Travers AA. 1991. Protein motifs that recognize structural features of DNA. *Trends Biochem. Sci.* 16:92–97

160. Panne D. 2008. The enhanceosome. *Curr. Opin. Struct. Biol.* 18:236–42

161. Escalante CR, Yie J, Thanos D, Aggarwal AK. 1998. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. *Nature* 391:103–6

162. Fujii Y, Shimizu T, Kusumoto M, Kyogoku Y, Taniguchi T, Hakoshima T. 1999. Crystal structure of an IRF-DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. *EMBO J.* 18:5028–41

163. Boutonnet N, Hui X, Zakrzewska K. 1993. Looking into the grooves of DNA. *Biopolymers* 33:479–90

164. Lane WJ, Darst SA. 2006. The structural basis for promoter -35 element recognition by the group IV σ factors. *PLoS Biol.* 4:e269

165. Miticka H, Rezuchova B, Homerova D, Roberts M, Kormanec J. 2004. Identification of nucleotides critical for activity of the sigmaE-dependent rpoEp3 promoter in *Salmonella enterica* serovar Typhimurium. *FEMS Microbiol. Lett.* 238:227–33

166. Fairall L, Schwabe JW, Chapman L, Finch JT, Rhodes D. 1993. The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* 366:483–87

167. Horton NC, Dorner LF, Perona JJ. 2002. Sequence selectivity and degeneracy of a restriction endonuclease mediated by DNA intercalation. *Nat. Struct. Biol.* 9:42–47

168. Sierk ML, Zhao Q, Rastinejad F. 2001. DNA deformability as a recognition feature in the reverb response element. *Biochemistry* 40:12833–43

169. Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM, Ebright RH. 2004. Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* 14:10–20

170. Schultz SC, Shields GC, Steitz TA. 1991. Crystal structure of a CAP-DNA complex: the DNA is bent by 90°. *Science* 253:1001–7

171. Parkinson G, Wilson C, Gunasekera A, Ebright YW, Ebright RE, Berman HM. 1996. Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.* 260:395–408

172. Little EJ, Babic AC, Horton NC. 2008. Early interrogation and recognition of DNA sequence by indirect readout. *Structure* 16:1828–37

173. Kim YC, Grable JC, Love R, Greene PJ, Rosenberg JM. 1990. Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing. *Science* 249:1307–9

174. Kalodimos CG, Boelens R, Kaptein R. 2004. Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. *Chem. Rev.* 104:3567–86

175. Travers A. 2000. Recognition of distorted DNA structures by HMG domains. *Curr. Opin. Struct. Biol.* 10:102–9

176. Weiss MA. 2001. Floppy SOX: mutual induced fit in HMG (high-mobility group) box-DNA recognition. *Mol. Endocrinol.* 15:353–62

177. Palasingam P, Jauch R, Ng CK, Kolatkar PR. 2009. The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *J. Mol. Biol.* 388:619–30

178. Hegde RS. 2002. The papillomavirus E2 proteins: structure, function, and biology. *Annu. Rev. Biophys. Biomol. Struct.* 31:343–60

179. Kim SS, Tam JK, Wang AF, Hegde RS. 2000. The structural basis of DNA target discrimination by papillomavirus E2 proteins. *J. Biol. Chem.* 275:31245–54

180. Hines CS, Meghoo C, Shetty S, Biburger M, Brenowitz M, Hegde RS. 1998. DNA structure and flexibility in the sequence-specific binding of papillomavirus E2 proteins. *J. Mol. Biol.* 276:809–18

181. Nelson HB, Laughon A. 1990. The DNA binding specificity of the *Drosophila* fushi tarazu protein: a possible role for DNA bending in homeodomain recognition. *New Biol.* 2:171–78

182. Koudelka GB, Carlson P. 1992. DNA twisting and the effects of noncontacted bases on affinity of 434 operator for 434 repressor. *Nature* 355:89–91

183. Edwards KJ, Brown DG, Spink N, Skelly JV, Neidle S. 1992. Molecular structure of the B-DNA dodecamer d(CGCAAATTTGCG)$_2$. An examination of propeller twist and minor-groove water structure at 2.2 Å resolution. *J. Mol. Biol.* 226:1161–73

184. Newman M, Strzelecka T, Dorner LF, Schildkraut I, Aggarwal AK. 1995. Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science* 269:656–63

185. Viadiu H, Aggarwal AK. 2000. Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Mol. Cell* 5:889–95

186. Lukacs CM, Aggarwal AK. 2001. BglII and MunI: what a difference a base makes. *Curr. Opin. Struct. Biol.* 11:14–18

187. Tolstorukov MY, Jernigan RL, Zhurkin VB. 2004. Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. *J. Mol. Biol.* 337:65–76

188. Eisenstein M, Shakked Z. 1995. Hydration patterns and intermolecular interactions in A-DNA crystal structures. Implications for DNA recognition. *J. Mol. Biol.* 248:662–78

189. Shakked Z, Rabinovich D, Kennard O, Cruse WB, Salisbury SA, Viswamitra MA. 1983. Sequence-dependent conformation of an A-DNA double helix. The crystal structure of the octamer d(G-G-T-A-T-A-C-C). *J. Mol. Biol.* 166:183–201

190. Travers AA. 1995. Reading the minor groove. *Nat. Struct. Biol.* 2:615–18

191. Choo Y, Klug A. 1997. Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.* 7:117–25

192. Nekludova L, Pabo CO. 1994. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc. Natl. Acad. Sci. USA* 91:6948–52

193. Pavletich NP, Pabo CO. 1993. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* 261:1701–7

194. Kitayner M, Rozenberg H, Rabinovich D, Shakked Z. 2005. Structures of the DNA-binding site of Runt-domain transcription regulators. *Acta Crystallogr. D* 61:236–46

195. Cassiday LA, Maher LJ 3rd. 2002. Having it both ways: transcription factors that bind DNA and RNA. *Nucleic Acids Res.* 30:4118–26

196. Schwartz T, Rould MA, Lowenhaupt K, Herbert A, Rich A. 1999. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science* 284:1841–45

197. Herbert A, Rich A. 1999. Left-handed Z-DNA: structure and function. *Genetica* 106:37–47

198. Schwartz T, Behlke J, Lowenhaupt K, Heinemann U, Rich A. 2001. Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat. Struct. Biol.* 8:761–65

199. Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* 128:707–19

200. Teytelman L, Ozaydin B, Zill O, Lefrancois P, Snyder M, et al. 2009. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* 4:e6700

201. Segal E, Widom J. 2009. From DNA sequence to transcriptional behavior: a quantitative approach. *Nat. Rev. Genet.* 10:443–56

202. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. 2006. A genomic code for nucleosome positioning. *Nature* 442:772–78

203. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, et al. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* 4:e1000216

204. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, et al. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res.* 17:1170–77

205. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362–66

206. Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191:659–75

207. Trifonov EN, Sussman JL. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA* 77:3816–20

208. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* 16:1505–16

209. Chung HR, Vingron M. 2009. Sequence-dependent nucleosome positioning. *J. Mol. Biol.* 386:1411–22

210. Swinger KK, Rice PA. 2004. IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.* 14:28–35

211. Ellenberger T, Landy A. 1997. A good turn for DNA: the structure of integration host factor bound to DNA. *Structure* 5:153–57

212. Rice PA, Yang S, Mizuuchi K, Nash HA. 1996. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* 87:1295–306

213. Lynch TW, Read EK, Mattis AN, Gardner JF, Rice PA. 2003. Integration host factor: putting a twist on protein-DNA recognition. *J. Mol. Biol.* 330:493–502

214. Piper DE, Batchelor AH, Chang CP, Cleary ML, Wolberger C. 1999. Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* 96:587–97

215. Chen L, Glover JN, Hogan PG, Rao A, Harrison SC. 1998. Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature* 392:42–48

216. Yie J, Liang S, Merika M, Thanos D. 1997. Intra- and intermolecular cooperative binding of high-mobility-group protein I(Y) to the beta-interferon promoter. *Mol. Cell. Biol.* 17:3649–62

217. Stefl R, Wu H, Ravindranathan S, Sklenar V, Feigon J. 2004. DNA A-tract bending in three dimensions: solving the dA4T4 vs dT4A4 conundrum. *Proc. Natl. Acad. Sci. USA* 101:1177–82

218. Faiger H, Ivanchenko M, Haran TE. 2007. Nearest-neighbor non-additivity versus long-range non-additivity in TATA-box structure and its implications for TBP-binding mechanism. *Nucleic Acids Res.* 35:4409–19

219. Ellison MJ, Feigon J, Kelleher RJ 3rd, Wang AH, Habener JF, Rich A. 1986. An assessment of the Z-DNA forming potential of alternating dA-dT stretches in supercoiled plasmids. *Biochemistry* 25:3648–55

220. Petrey D, Honig B. 2003. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* 374:492–509

221. Lavery R, Sklenar H. 1989. Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* 6:655–67

222. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* 23:128–37

223. Kitayner M, Rozenberg H, Rohs R, Suad O, Rabinovich D, et al. 2010. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* 17:423–29

NOTE ADDED IN PROOF

In section 4.2.1.1., we discuss the possibility of non-Watson-Crick base pairs playing a role in protein-DNA recognition. This hypothesis is supported by recent crystal structures of p53 tetramers bound to DNA-binding sites with contiguous half sites where the AT doublets of the CATG core regions exhibit Hoogsteen geometry (223). Although these Hoogsteen base pairs are embedded in essentially undistorted B-DNA, the alternate base pairing geometry affects local DNA shape. This observation expands the code of sequence readout.

Annual Review of Biochemistry

Volume 79, 2010

Contents

Preface

Prefatory Article

Recent Advances in Biochemistry

Indexes

Errata

An online log of corrections to *Annual Review of Biochemistry* articles may be found at http://biochem.annualreviews.org