Protein–DNA Complexes: Specific

Mark A Strauch, *University of Maryland, Baltimore, Maryland, USA*

Sequence-specific DNA-binding proteins recognize and interact with discrete base sequences in the genome to regulate fundamental metabolic processes.

Secondary article

Article Contents

Introduction

Sequence-specific protein–DNA binding is a process fundamental to life on earth. From the identification of replication sites on the chromosome to the differential expression of genes during development, a multitude of sequence-specific DNA-binding proteins regulate the fundamental biochemical processes of life. Understanding how these different proteins are able to find and bind selectively to only one, or just a few, specific sequences out of the millions present in a genome is a major goal of molecular biology.

DNA Structure From the Standpoint of Sequence Recognition

Protein binding to a specific DNA sequence entails more than a mere recognition and interaction with the linear arrangement of the constituent base pairs (direct readout). The precise three-dimensional structure of the region of the DNA molecule containing the base pair-specific elements that can potentially interact with the protein must be taken into account. This so-called 'structural presentation' of potential recognition surfaces on DNA can play a critical role in protein–DNA binding interactions (Kim *et al*., 1997; Koudelka, 1998). The shape of a DNA helix is never a smoothly uniform structure due to the effect of primary sequence on such parameters as tilt, roll and twist angles of base pair steps and propeller twisting between bases of a pair. Sequence-dependent variability of these parameters along the DNA molecule leads to localized variations in the width and depth of the major and minor grooves and the propensity of some regions to assume a noncanonical B form, to be more easily distorted or even to adopt an intrinsic bend or kink. Because different base sequences dictate different spatial positionings not only of interactive sites on the bases themselves (e.g. hydrogen bond donors and acceptors), but also of atoms in the sugar–phosphate backbone, theoretically it is possible that specific sequence recognition could be achieved without any contact between a protein and the base moieties (indirect readout). However, there has yet to be found a clear-cut example of a sequence-specific DNA-binding protein that achieves high-affinity recognition based solely upon an indirect readout mechanism. Nevertheless, understanding sequence-specific recognition by any DNA-binding protein must take into account the contributions of both direct and indirect readout mechanisms (Harrington, 1992).

Factors beyond primary base sequence can also influence the three-dimensional characteristics of DNA regions and thus affect the structural presentation of a specific recognition sequence. Changes in the degrees of localized supercoiling, chromatin compaction, osmotic forces and ionic strength are among the factors that can alter the presentation or availability of a sequence in the cell. Additionally, other bound proteins may affect DNA regions spatially removed from their own binding sites and so affect the presentation of other recognition sequences. Of course, chemical modifications (such as covalent methylations at certain base positions) of moieties in recognition sequences change DNA structure by definition and obviously can have profound effects on binding interactions.

Structural presentation may be thought of as having an influence primarily on initial selection and discrimination of sequences. However, most DNA-binding proteins introduce significant changes in DNA structure (relative to the unbound state), some of which can be quite spectacular (Kim *et al*., 1993). It is becoming clear that the ability of a given DNA sequence to be structurally deformed into an alternate conformation plays a crucial role in many sequence-specific binding interactions (Lesser *et al*., 1993), presumably due to favourable energetic effects caused by increasing the interfacial complementarity between DNA and protein, the latter usually undergoing conformational changes as well (Hegde *et al*., 1998). The ability of a specific DNA sequence to be deformed appears to be an especially important recognition mechanism utilized by some DNA-binding proteins showing strict

sequence selectivity, such as restriction endonucleases and methyltransferases.

Points of Recognition in the Major and Minor Grooves

Sequence specificity via a direct readout mechanism entails noncovalent interactions (hydrogen bonding, electrostatic forces, salt bridges, van der Waals contacts) between the functional groups of the amino acids or backbone atoms of the protein and the functional groups available on the bases in the major and minor grooves of the DNA. Assuming that the normal Watson–Crick hydrogen bonding has not been disrupted, each of the four types of base pair (AT, TA, GC, CG) presents a distinct pattern of hydrogen donor, hydrogen acceptor and hydrophobic groups available for bonding (**Figure 1**). A protein, by using two hydrogen bonds and interacting in the major groove, could easily discriminate between the four possibilities. In the minor groove, the presence of a hydrogen donor from the 2 amino group of guanine can be utilized for differentiation of GC or CG from AT or TA. However, any discrimination of GC from CG or AT from TA in the minor groove would have to involve a highly sensitive means of distinguishing either the slightly different steric positioning and spacing between atoms or differentiation between the slight difference in hydrogen bond energies to the O2 of pyrimidines versus the N3 of purines. Interestingly, there is evidence that such selectivity is possible (Wong and Bateman, 1994; Kielkopf *et al*., 1998).

By definition, sequence-specific DNA-binding proteins show a degree of selectivity or preference for binding certain base sequences over others. In many (but not all) cases, examination of different sites bound by a protein reveals a readily identifiable consensus sequence responsible for recognition. For proteins recognizing sequences via direct readout, these consensus elements must reflect a preferred spatial arrangement of different functional groups of the bases. In the case of indirect readout, the consensus sequence must be specifying a particular and unique three-dimensional DNA structure not randomly found. However, it should be kept in mind that most, if not all, sequence-specific DNA-binding proteins utilize a combination of direct and indirect readout mechanisms.

Some proteins, such as certain restriction/modification enzymes, exhibit a very strict consensus recognition element and are unable to tolerate even a single mismatch.

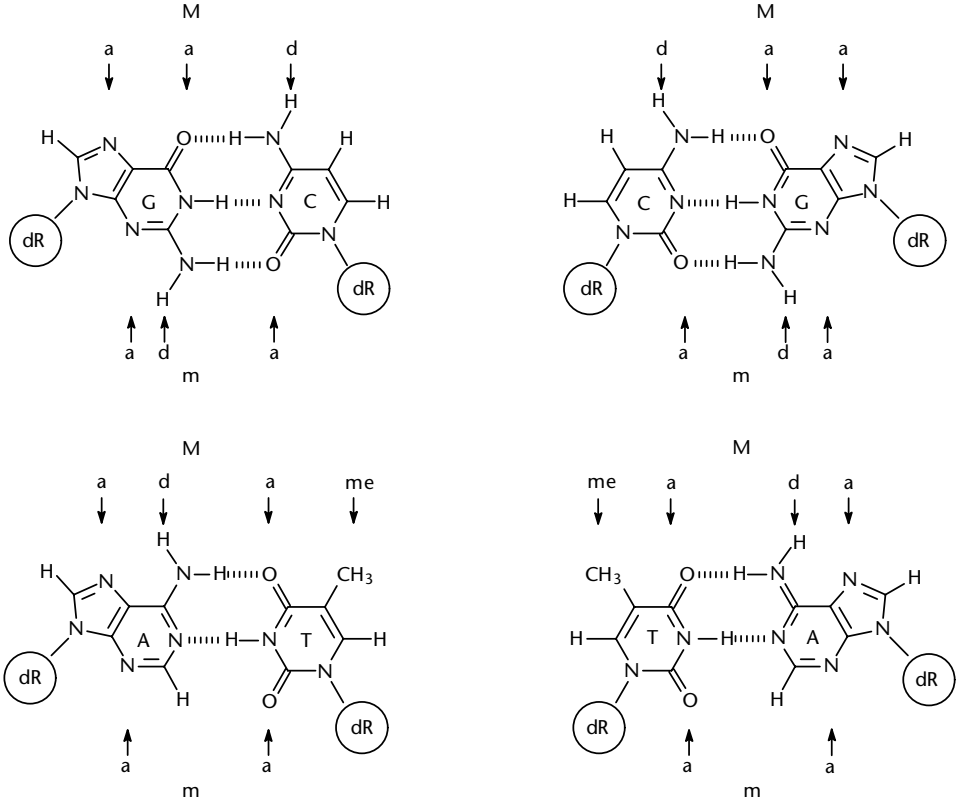

Figure 1 Points of recognition in the major (M) and minor (m) grooves of DNA for each of the four base pairs. a, electron acceptor; d, electron donor; me, methyl group. Hydrogen bonding in base pairs is indicated by dashed lines. dR in circles denotes the deoxyribose-phosphate backbone of DNA.

(Two caveats should be mentioned with respect to these DNA-binding enzymes. First, binding affinity should not be equated with enzymatic activity: *Eco*RV appears to bind equally well to most sequences but only catalyses cleavage at GATATC sequences due to unique chemical properties and structural deformations that this stretch of DNA can undergo during the transition state of catalysis. Second, changing environmental conditions, such as alteration of salt concentration or pH, can relax specificity.) At the other end of the sequence-specific binding spectrum are proteins whose derived consensus recognition sequences show degeneracies at a number of positions and can bind with high affinity to a broad, but not random, subset of base sequences. Degeneracy in the consensus could imply a number of different characteristics about the interaction of the protein with base pair-specific recognition points at that position. (1) An amino acid side-chain might interact with different base pairs with little or no steric occlusion presented by other functional groups in the vicinity; for example, a hydrogen donating group bonding with the N7 position of a purine regardless of whether a 6-amino group (A) or 6-keto group (G) were present. (2) The protein might have the flexibility to assume different conformations (or induce different DNA conformations) in order to adjust the spatial positions of interacting groups; for example, a hydrogen-accepting side-chain interacting with the 4-amino group of cytosine whether from a GC or CG pair at a particular position. (3) Limited degeneracies might indicate positions where the bases do not provide functional groups directly interacting with the protein but where certain bases affect the structural presentation of recognition points in neighbouring positions. Conversely, the invariant occurrence of a specific base at a given position does not necessarily indicate that it contributes groups that interact directly with the protein: it could be strictly required for structural presentation of other interacting elements. (4) Degeneracies might also be seen at positions where a certain base can contribute direct points of contact, but the absence of such contacts does not affect the overall ability of the protein to bind, given that enough other contacts are available at other positions. However, such situations would probably result in the binding interaction having different thermodynamic or kinetic properties (i.e. alterations in ΔG, dissociation constant, on or off rates, etc.) depending upon the actual DNA sequence.

Although it may be possible to define one particular sequence variation as having optimal recognition points and binding stability via *in vitro* biochemical or biophysical measurements, this should not be confused with the concept of optimal *in vivo* functioning, which is ultimately the result of protein and DNA-binding site coevolving to accomplish a physiological role within the constraints of the cellular milieu.

Overview of DNA-binding Motifs for Sequence-specific Binding

Many sequence-specific DNA-binding proteins have been grouped into families based upon the type of structural motifs used for recognition and interaction with their DNA targets. A variety of classification schemes have been devised, each involving somewhat different groupings, but an initial broad classification can be made on the basis of which type of protein secondary structure is used for recognition: α helix, β sheet or some type of loop. Further subdivision into specific families is usually based upon the manner in which the recognition element is spatially related to the surrounding protein structure or the method of multimerization between individual subunits in a multimeric protein. Not all known DNA-binding proteins can be unambiguously placed into a family grouping and there is no reason to assume that all families have been discovered. At present, generally agreed upon major families of motifs include the helix–turn–helix (HTH), homeodomain, HNF-3/fkh winged helix, zinc fingers, zinc-containing steroid receptors, leucine zipper, helix–loop–helix (HLH), Rel and β ribbon. Since detailed descriptions and discussion of these major families can be found elsewhere in these volumes and in sources cited at the end of this entry, only a few general principals will be briefly touched upon here.

The majority of sequence-specific DNA-binding proteins examined, including members of the HTH, homeodomain, steroid receptor, winged helix, and zinc finger families, employ α helices for recognition. (Leucine zipper and HLH proteins contain a basic domain rich in arginine and lysine residues that appears to be only partially helical in solution but which undergoes a conformational transition into a typical α helix upon binding to DNA.) Usually, the so-called recognition helix interacts with specificity determinants in the major groove. However, a variety of ways for inserting a recognition helix into the major groove have been observed, even among members of the same family. For example, the bacteriophage λCI and *Escherichia coli* TrpR proteins are both HTH family members, yet the recognition helix of CI lies lengthwise in the groove while TrpR orients its recognition helix perpendicularly in the groove. It should be emphasized that no single recognition helix can in and of itself bind to DNA in a sequence-specific manner. A three-dimensional protein architecture is required not only for proper positioning of the recognition helix, but also to provide additional DNA contacts, either to the bases or the phosphate backbone, which are necessary for binding stability and accurate sequence discrimination. In fact, the notion of a protein having its sole binding specificity determinants present within a single domain or motif is rather misleading: multiple motifs or contact domains are usually required for high-fidelity sequence specificity. The occurrence of more

than one motif, either of the same or different type, is often the result of multimer formation, but multiple motifs may also be present within a single polypeptide chain.

β Sheets arranged in antiparallel fashion (β ribbons) have also been observed to function as sequence-specific recognition elements via interactions in the major groove. Additionally, there are examples where β ribbons make sequence-specific contacts within the minor groove. The component β sheets forming the ribbon can either be present on the same polypeptide (as in the case of the eukaryotic TATA-binding proteins) or come together via dimerization of identical polypeptide chains (see below).

Recently, a new motif which relies upon neither an α helix nor a β sheet for sequence recognition has been defined by structural studies of a variety of eukaryotic transcription factors involved in cellular stress responses and developmental events. Members of this family, called Rel, utilize a projecting hydrophilic loop of polypeptide to make sequence-specific contacts with five contiguous base pairs in a major groove (Chytil and Verdine, 1996). Many DNA-binding proteins use various other types of loops and strands to make DNA contacts, in addition to the contacts made by the major recognition elements. The contribution to overall binding energy provided by these additional contacts is usually absolutely necessary for the stability of binding. For the most part, the heterogeneity in structure of these loops and strands prevents them from being easily grouped into defined classes of motifs.

A number of variations on a few themes have evolved to achieve sequence-specific DNA recognition by proteins. Although many motifs can be grouped into families, each DNA-binding protein is unique. It will be especially interesting if future research leads to the discovery that different proteins, either naturally occurring or constructed using protein engineering techniques, can recognize the same DNA sequence with comparable affinities using entirely different types of recognition motifs (e.g. α helix versus β ribbon).

Examples of Amino acid–Nucleotide Interaction

Noncovalent bonds made by amino acid side-chains or the main chain of the protein (usually the amides) to the bases and phosphate backbone of the DNA target underlie all sequence-specific interactions. The majority of contacts are usually made via hydrogen bonds but van der Waals contacts and electrostatic interactions can occur and are often critically involved. Of particular interest for most sequence-specific interactions are the contacts made between the amino acid side-chains and the bases. Given a proper context, it appears that nearly every type of free side-chain can make some type of contact with at least one of the four bases. Practically every polar amino acid has been observed to be capable of making a base pair contact and some examples are shown in **Figure 2**. Additionally, nonpolar hydrophobic side-chains of alanine and isoleucine have been seen to make van der Waals contact with the 5-methyl group of thymidine, and even the aromatic ring of phenylalanine is intercalated between base pairs in some complexes.

No general recognition code of one-to-one correspondence relating a specific amino acid to a single specific base with which it can interact exists. Many amino acid side-chains can interact with more than one type of base and any given type of base can be contacted by different side-chains. Often, more than one side-chain contacts a given base, and in other instances a single side-chain can contact more than one base pair simultaneously. Although no simple rules of recognition having applicability to all sequence-specific interactions seem to exist, there may exist codes governing the interactions seen for members of some families, in particular the zinc-finger proteins (Choo and Klug, 1997).

Hydrogen-bonding interactions between protein and DNA need not be direct: frequently they entail one or more water-mediated contacts. The extensive and important role that water plays in sequence-specific protein–DNA binding has been one of the most surprising and exciting results of recent structural determinations. For example, in the crystal structure of the bacterial Trp repressor protein bound to its cognate DNA target, the important contacts between residues of the recognition helix and base pairs required for sequence specificity were all observed to be mediated via ordered water molecules, a finding confirmed by many subsequent studies. A fluctuating variability in positions and bonding interactions of water molecules mediating contact between the antennapaedia homeodomain and its DNA target dramatically illustrates the dynamic nature of protein–DNA interactions, even in stable complexes (see Schwabe, 1997).

Biological Ranges of Dissociation Constants for Sequence-specific Protein Binding to DNA

A variety of *in vitro* methodologies are available for measuring the thermodynamic and kinetic properties of protein–DNA binding. The most common, and frequently easiest, parameter to derive is the apparent equilibrium dissociation constant (K_d) of the overall interaction. K_d measurements made *in vitro* are extremely useful for addressing numerous questions when examining a particular sequence-specific binding reaction. However, caution must be exercised in interpreting these results as being truly reflective of *in vivo* thermodynamic properties since the experimental conditions used to derive K_d values bear little resemblance to intracellular situations (*in vitro*, small

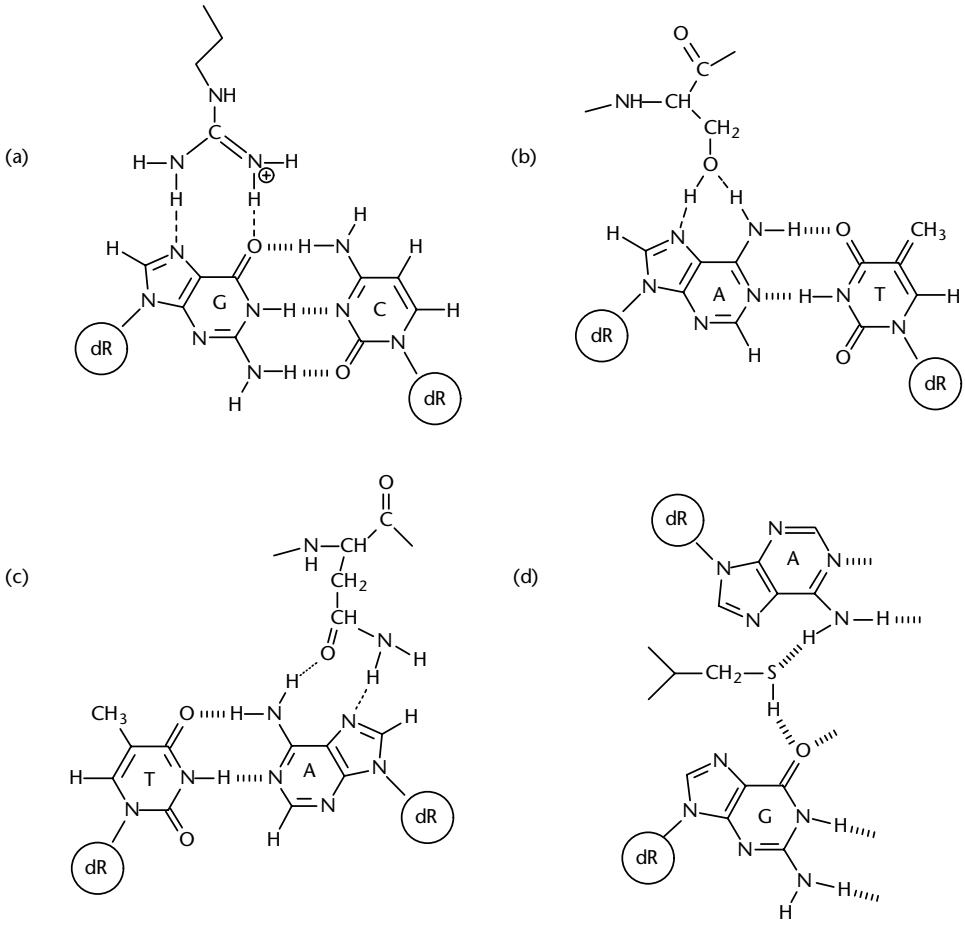

Figure 2 Examples of amino acid–base pair interactions. (a) Arginine bonding to guanine. (b) Serine bonding to adenine. (c) Asparagine bonding to adenine. (d) Cysteine double contact to adenine and guanine of adjacent base pairs. In (d), the adenine is attached to the opposite strand relative to the guanine but the bases have been drawn on the same plane. Hydrogen bonds are denoted by dashed lines.

defined pieces of DNA containing the target sequence are used and thus the target is not within its natural surroundings subject to chromatin condensation and packaging forces and variations; rarely are other DNA-binding proteins or enzymes present, let alone the multitude present in a nucleus or cell; ionic and osmolarity conditions are usually substantially different from those *in vitro*, and so on). Unfortunately, accurate thermodynamic assessments of specific protein–DNA binding reactions *in vivo* are not possible using current technologies.

The above caveats being given, most sequence-specific protein–DNA complexes have K_d values in the range of 10^{-11}–10^{-8} mol L^{-1}. In general, proteins with more strict sequence specificity, such as bacterial repressors acting at one or just a few sites in the genome, exhibit higher affinities *in vitro* (i.e. have lower K_d) than do proteins with a more 'relaxed' sequence specificity. In cases where reasonably accurate assessments are possible, the difference in K_d values for specific targets versus random DNA (the sequence discrimination factor) is in the range of 10^2–10^4.

Mechanisms for Sequence Location; Effects of Nonspecific Sequences

Before stably binding at a specific DNA sequence, the protein must first locate the sequence rapidly and selectively from among the millions of possible sequence permutations present in the genome. It is unreasonable to believe that this could be effectively accomplished by random collisions of the reactants in proper orientation. As alluded to above, all sequence-specific DNA-binding proteins possess a degree of nonspecific affinity for DNA, primarily through electrostatic interactions with the sugar–phosphate backbone. The free energy component allowing these nonspecific binding interactions is believed to be mostly entropic in nature and derived from counter-ion displacement from the backbone as well as the disordering of water molecules that normally hydrate the DNA. Nonspecific affinity allows the protein to rapidly sample sequence permutations, presumably by some type

of 'sliding' mechanism along the DNA and 'jumping' between proximal segments. When the protein encounters its specific target and interacts with sequence-specific recognition determinants, the electrostatic interactions with the sugar–phosphate backbone responsible for the nonspecific binding mode still play a critical role: they contribute a significant amount of the favourable binding free energy without which stable complex formation would be impossible.

In any large genome, the chances are that a number of sequences resembling a specific binding protein's recognition site exist and that they have some degree of intermediary affinity for the protein. A number of scenarios can be envisioned whereby evolution could have exploited these 'semispecific' sites for regulatory purposes. Congregation of semispecific sites near a particular physiological target site could have evolved as a means to increase the localized concentration of the protein near that target and so ensure a differential probability of protein binding there versus at a target not surrounded by semispecific sites. Even more intriguing would be if one or more optimal binding sites for a particular protein existed but were inaccessible under some conditions (due to some factor such as chromatin condensation) and the protein was exerting a regulatory effect by binding at suboptimal sites under those conditions. If the optimal sites became accessible in response to some cell cycle or environmental signal, then they might serve as a 'sink' to titrate the protein away from the suboptimal sites and thus change the regulatory effect.

Role of Multiple Subunits in DNA-binding Proteins

The majority of sequence-specific DNA-binding proteins examined to date are multimeric in solution (often either homodimeric or homotetrameric) and in these cases multimerization is usually necessary for high-affinity binding. However, there are many examples of sequence-specific proteins that bind as monomers so multimerization is not an absolute requirement for high affinity. (Some, but not all, of these monomeric binders do possess two or more separable recognition motifs in the same polypeptide.)

Relative to a monomeric–single site interaction, using multimers to recognize multiple adjacent sites would obviously increase the overall number of protein–DNA contacts in the complex and so would be expected to increase binding strength. It is not difficult to imagine that many homomultimeric DNA-binding proteins were once monomeric–single site binders that coevolved with their target sites under selective pressures favouring tighter binding. There is also a source of recognition flexibility and variability attendant upon the use of multimeric DNA-binding proteins. For example, members of the leucine zipper and helix–loop–helix families of DNA-binding transcription factors can form heterodimers with other members of their own family. Since each individual monomer brings with it different DNA-binding properties, heterodimer formation adds a degree of regulatory versatility that can be altered by changing the relative expression levels of the subunits in response to different stimuli.

For some proteins, prior multimerization is an absolute requirement for sequence-specific DNA binding: the individual monomers have no DNA-binding ability and multimerization is necessary to actually form the DNA recognition/binding motif. The β-ribbon family members MetJ, Arc and Mnt are excellent examples of this category.

Multimerization is the basis for cooperative binding behaviour exhibited by many sequence-specific DNA-binding proteins. Binding cooperativity can result from multimer formation at two different stages: assembly of the multimer prior to DNA interaction or assembly of the multimeric complex on the DNA. For the former, dramatic cooperative effects would be observed if the dissociation constant for protein multimerization was significantly greater than the dissociation constant for the multimer–DNA binding reaction. In the latter case, cooperativity could be due to either a conformational change in the protein entity initially binding DNA which makes it a better target for further protein–protein interactions necessary for stable complex formation, or due to a conformational change in DNA structure adjacent to the initially bound protein which causes that particular stretch of DNA to be a better substrate for additional protein binding (Vashee *et al.*, 1998). In these latter scenarios, the initially bound protein entity could be either monomeric or already multimeric, and either homologous or heterologous proteins could be the subsequent binders.

Concluding Remarks

Sequence-specific DNA-binding proteins interact with spatial arrangements of atoms and reactive groups on DNA which are specified either by a particular sequence of base pairs or by closely related permutations of a sequence. In the past few years structural studies using X-ray crystallographic and nuclear magnetic resonance (NMR) techniques have elucidated the structure of many DNA-binding proteins in both the free and DNA-bound states. A wealth of information concerning how proteins and DNA interact to form stable, high-affinity bound complexes has been obtained. Further structural studies on different classes of DNA-binding proteins and mutant variants will provide new and deeper insights. It must be borne in mind, however, that protein–DNA interactions are dynamic processes that cannot be fully understood merely by solving a static or equilibrium structure. A major direction

of future research will be not only to dissect the biophysical and thermodynamic parameters responsible for stable complex formation, but also to integrate this information with knowledge about how changes affecting these parameters can bring about physiological changes in gene expression and DNA metabolism controlled by different sequence-specific DNA-binding proteins.

References

Choo Y and Klug A (1997) Physical basis of a protein–DNA recognition code. *Current Opinion in Structural Biology* **7**: 117–125.

Chytil M and Verdine GL (1996) The Rel family of eucaryotic transcription factors. *Current Opinion in Structural Biology* **6**: 91–100.

Harrington RE (1992) DNA curving and bending in protein–DNA recognition. *Molecular Microbiology* **6**: 2549–2555.

Hegde RS, Wang A-F, Kim SS and Schapira M (1998) Subunit rearrangement accompanies sequence-specific DNA binding by the bovine papillomavirus-1 E2 protein. *Journal of Molecular Biology* **276**: 797–808.

Kielkopf CL, White S, Szewczyk JW *et al*. (1998) A structural basis for recognition of A·T and T·A base pairs in the minor groove of B-DNA. *Science* **282**: 111–115.

Kim E, Albrechtsen N and Deppert W (1997) DNA-conformation is an important determinant of sequence-specific DNA binding by tumor suppressor p53. *Oncogene* **15**: 857–869.

Kim Y, Geiger JH, Hahn S and Sigler PB (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**: 512–519.

Koudelka GB (1998) Recognition of DNA structure by 434 repressor. *Nucleic Acids Research* **26**: 669–675.

Lesser DR, Kurpiewski MR, Waters T, Connolly BA and Jen-Jacobson L (1993) Facilitated distortion of the DNA site enhances *EcoRI* endonuclease-DNA recognition. *Proceedings of the National Academy of Sciences of the USA* **90**: 7548–7552.

Schwabe JWR (1997) The role of water in protein–DNA interactions. *Current Opinion in Structural Biology* **7**: 126–134.

Vashee S, Melcher K, Deng WV, Johnston SA and Kodadek T (1998) Evidence for two modes of cooperative DNA binding *in vivo* that do not involve protein–protein interactions. *Current Biology* **8**: 452–458.

Wong JM and Bateman E (1994) TBP–DNA interactions in the minor groove discriminate between A:T and T:A base pairs. *Nucleic Acids Research* **22**: 1890–1896.

Further Reading

Berg OG and von Hippel PH (1988) Selection of DNA binding sites by regulatory proteins. *Trends in Biochemical Sciences* **13**: 207–211.

Gehring WJ, Affolter M and Burglin T (1994) Homeodomain proteins. *Annual Review of Biochemistry* **63**: 487–526.

Harrison SC and Aggarwal AK (1990) DNA-recognition by proteins with the helix–turn–helix motif. *Annual Review of Biochemistry* **59**: 933–969.

Lilley DM (ed.) (1995) *DNA–Protein Structural Interactions*. Oxford: IRL Press.

Pabo CO and Sauer RT (1992) Transcription factors: structural families and principals of DNA recognition. *Annual Review of Biochemistry* **61**: 1053–1095.

Raumann BE, Brown BM and Sauer RT (1994) Major groove DNA recognition by β-sheets: the ribbon–helix–helix family of gene regulatory proteins. *Current Biology* **4**: 36–43.

Travers A (1993) *DNA–Protein Interactions*. London: Chapman and Hall.