Towards an Understanding of Protein-DNA Recognition

Daniela Rhodes, John W. R. Schwabe, Lynda Chapman and Louise Fairall

Phil. Trans. R. Soc. Lond. B 1996 **351**, 501-509
doi: 10.1098/rstb.1996.0048

Towards an understanding of protein–DNA recognition

DANIELA RHODES, JOHN W. R. SCHWABE*, LYNDA CHAPMAN AND LOUISE FAIRALL

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, U.K.

SUMMARY

Understanding how proteins recognize DNA in a sequence-specific manner is central to our understanding of the regulation of transcription and other cellular processes. In this article we review the principles of DNA recognition that have emerged from the large number of high-resolution crystal structures determined over the last 10 years. The DNA-binding domains of transcription factors exhibit surprisingly diverse protein architectures, yet all achieve a precise complementarity of shape facilitating specific chemical recognition of their particular DNA targets. Although general rules for recognition can be derived, the complex nature of the recognition mechanism precludes a simple recognition code. In particular, it has become evident that the structure and flexibility of DNA and contacts mediated by water molecules contribute to the recognition process. Nevertheless, based on known structures it has proven possible to design proteins with novel recognition specificities. Despite this considerable practical success, the thermodynamic and kinetic properties of protein/DNA recognition remain poorly understood.

1. INTRODUCTION

Protein–DNA recognition is central to many important cellular processes such as transcription and replication. In the last ten years, complexes between DNA-binding proteins and their specific DNA targets have been the focus of extensive structural analyses to gain an understanding of sequence specific recognition. What are the molecular mechanisms that these studies seek to explain? In biological terms it is important for DNA-binding proteins to bind to their DNA target site with an appropriate affinity and specificity, as well as binding to and releasing from their DNA targets with appropriate kinetics. This means that studies on these complexes need to address the thermodynamics and kinetics of DNA-binding, in addition to the more straight forward structural aspects of the components. In this brief review we will focus on the emerging principles of protein–DNA recognition and illustrate these primarily using the structure of two zinc-binding eukaryotic DNA binding domains recently solved in our laboratory (Fairall *et al.* 1993; Schwabe *et al.* 1993). We will also ask whether there are general rules governing protein–DNA interactions that can be formulated into some type of recognition code, and whether our understanding is sufficient to design new DNA-binding proteins with defined specificity.

2. RECONCILING THE PHYSICAL CHEMISTRY WITH THE STRUCTURAL BIOLOGY

Protein and DNA molecules will interact if there is a loss of Gibbs free energy on the formation of a complex. The change in free energy (ΔG) during complex formation depends upon the change in both entropy (ΔS) and enthalpy (ΔH) such that $\Delta G = \Delta H - (T \times \Delta S)$. What does this mean in terms of the structural details of protein–DNA complexes? The enthalpy term (ΔH) arises from the many very short-range non-covalent interactions between protein and DNA. The entropy term (ΔS) depends upon the nature of the solvent on the interacting surfaces of the protein and DNA before and after complex formation. If a significant number of ordered water molecules are displaced on complex formation, then the entropy term can favour the interaction. It is clear therefore that for a favourable contribution to ΔG, both the enthalpy (ΔH) and entropy (ΔS) terms require the protein to have a surface shape that is highly complementary to that of its DNA target. This so-called 'shape recognition' constitutes molecular recognition in its broadest sense. However, in addition to a complementary shape, the chemistry of the interacting surfaces must also be complementary i.e. the precise nature and three-dimensional arrangement of the functional groups on the protein must match those of the DNA target site. The arrangement of these interacting groups determines the specificity of DNA recognition at an atomic level. In conclusion, when

[*] Present address: The Salk Institute for Biological Studies, P.O. Box 85800, San Diego, California 92186–5800, U.S.A.

proteins interact with DNA both the enthalpic and entropic contributions must be considered. The recognition process itself can be conceptually divided into the recognition of complementary molecular shapes and chemical recognition at an atomic level. If one views specific protein–DNA interactions from the position of the protein, then it becomes clear that shape recognition concerns the global architecture of the protein, whereas chemical recognition is determined by the stereochemical arrangement of amino acids on the surface of the protein.

3. SHAPE RECOGNITION

Typically the protein component of protein–DNA complexes is illustrated by a schematic diagram highlighting the elements of secondary structure from which the protein is built (figure 1). Although these diagrams give an excellent impression of the general architecture of the protein, they may lead the casual observer to overlook one of the most striking features of these structures: the remarkable complementarity of shape achieved between protein and DNA. Figure 2 better illustrates the complementary shape of the protein and DNA for the zinc-fingers of the transcription factor Tramtrack (TTKDBD) and the DNA binding domain of the oestrogen receptor (ERDBD). Essentially all the structures of protein–DNA complexes determined to date have clearly illustrated this principle of shape recognition. The first structures solved, those of several prokaryotic proteins, showed how a helical element on the surface of the protein, buttressed against a second helix, is used to interact with bases in the major groove (see below). This arrangement of helices, the helix-turn-helix motif, is precisely docked on the DNA by numerous contacts to the sugar phosphate backbone made from amino acids of the scaffold (or backing structure). These structures led to the idea that specific interactions (with the DNA bases) made by the so-called helical 'reading head' were structurally separable from the non-specific interactions (with the sugar phosphate backbone). As will be seen below, specific and non-specific interactions are rarely distinctly separable. It is clear however, that to gain access to the bases, the protein must 'reach' into the major groove of the DNA. A great many DNA-binding proteins employ an α-helical element to interact with bases in the major groove. The precise orientation of this helix varies greatly from one DNA-binding motif to another (compare the orientations of the α-helices of the TTKDBD and ERDBD in figure 1) but the use of an α-helix is by no means universal. The prokaryotic met repressor employs an anti-parallel two-stranded β-sheet to serve the same function (Somers & Phillips 1992). Furthermore neither the NF-κB protein (Ghosh *et al.* 1995; Müller *et al.* 1995) nor the TATA-binding protein (TBP) (Kim *et al.* 1995*a*, *b*) employ an α-helical reading head. In the NF-κB P50 homodimer all the amino acid to base contacts are made by residues in loops protruding from the surface of the protein. For TBP, the nature of the complex with DNA is strikingly different: the DNA target is substantially deformed, with the protein interacting along the length of the minor groove, which is splayed open and curved away from the protein. The interacting surface on the protein is a β-barrel structure.

In conclusion, when we look at protein–DNA complexes it is clear that the interacting surfaces have highly complementary shapes. Because at a coarse level the structure of DNA is essentially uniform, it is not surprising that diverse DNA-binding proteins have employed similar architectural strategies to achieve interfaces which are complementary in shape. It is the nature of the interactions between these complementary surfaces that is termed chemical recognition.

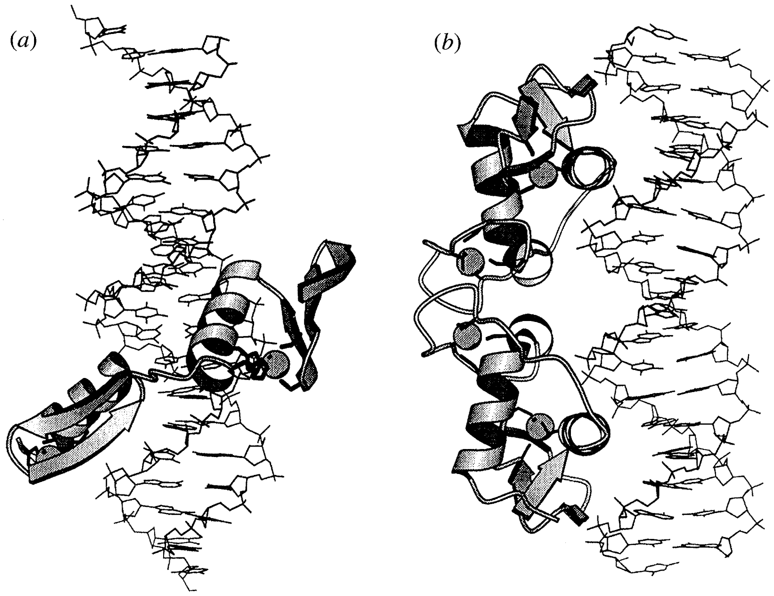


Figure 1. Schematic representations of the three-dimensional structures of two different protein–DNA complexes. (*a*) The TTKDBD-DNA complex. (*b*) The ERDBD-DNA complex. The protein is illustrated using the program MOLSCRIPT (Kraulis 1991). α-helices are shown as spirals and β-strands as arrows.

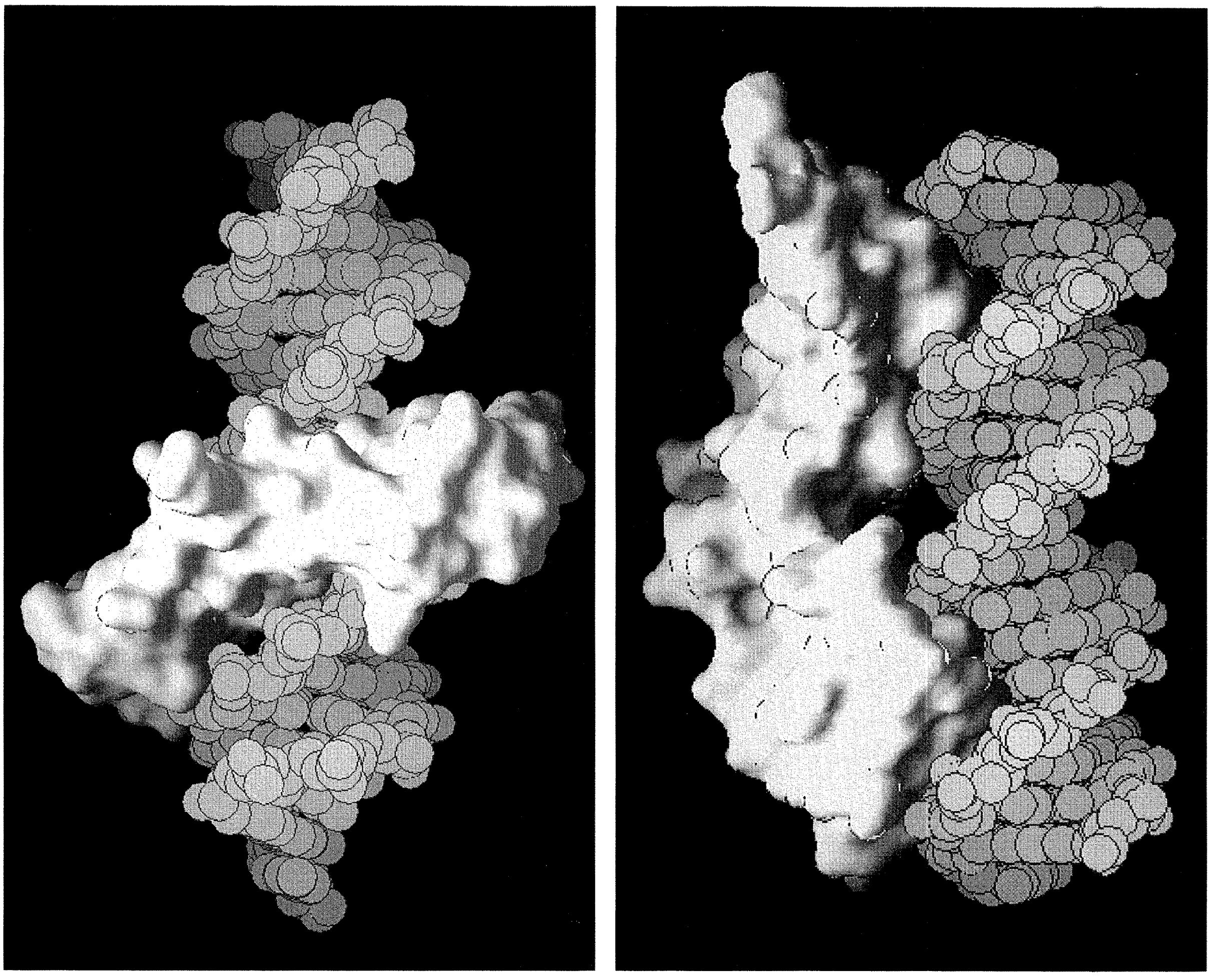


Figure 2. The shape of the protein and DNA are complementary. Space-filling models of: the TTKDBD-DNA complex (left-hand figure); and the ERDBD-DNA complex with the protein shown as a molecular surface (right-hand figure), drawn using the program GRASP (Nicholls *et al.* 1993).

4. CHEMICAL RECOGNITION

The same physical rules that determine protein and nucleic acid structure govern their specific and non-specific interactions. The forces involved include hydrogen bonds, van der Waals (or 'London dispersion') forces, hydrophobic interactions, global electrostatic interaction and local salt bridge interactions. High resolution crystal structures of protein–DNA complexes have revealed the three dimensional network of interactions that ties the two molecules together. As expected, contacts involving the phosphate backbone of the DNA appear to orient the protein in such a way that specific contacts can be made in the major groove. Certain common themes have emerged. In particular arginine and glutamine residues have long side chains that are able to make bidentate contacts to individual bases. Indeed, the most commonly occurring interaction is that of arginine with the N7 and O6 of guanine. Glutamine (and asparagine) can interact similarly with the N7 and N6 of adenine. However, both these amino acids are seen to make a variety of other contacts to bases. Finally, the bulky 5′-methyl group of thymine is suited to making van der Waals contacts with the methyls of several amino acids, although it may also play an important negative role in sterically preventing incorrect binding.

5. HOW TO BE MORE SPECIFIC

One problem faced by the generally small and compact DNA-binding domains, is that one such domain is not able to make a sufficient number of contacts with the DNA to specify a unique target site and bind with reasonable affinity. It seems that several strategies have been employed to overcome this problem. The first is simply to add on arms or tails that recognize additional features of the DNA, particularly in the minor groove (e.g. the homeodomains: see Kissinger *et al.* 1990; Wolberger *et al.* 1991). The second is to double up on the recognition by forming either homo- or hetero-dimers (Schwabe *et al.* 1993; Glover & Harrison 1995), thus specifying a longer DNA sequence. In the latter case this also vastly increases the recognition possibilities through a combinatorial approach. The third method of increasing specificity is to employ multiple DNA-binding domains, either by using tandem repeats of the same type of DNA-binding motif e.g. the zinc-finger motif (Schwabe & Klug 1994), or by linking together different types of motif (Klemm *et al.* 1994).

6. THE ROLE OF DNA STRUCTURE

Given the greater diversity of structures exhibited by proteins compared with DNA, the specific interaction between proteins and DNA has most frequently been considered from the viewpoint of the protein recognizing DNA. Although this approach has been most instructive, it has generally resulted in too little emphasis being placed upon the importance of the structural diversity of the DNA target sites.

Even the earliest fibre diffraction studies of DNA revealed that the double helix could adopt discrete conformations depending on the sequence and upon the degree of hydration (Arnott & Selsing 1974*a*, *b*). Two models for the extreme conformations (differing in their helical parameters) were termed A and B forms (Fuller *et al.* 1965; Langridge *et al.* 1960). It is now generally accepted that both the conformation and rigidity of DNA are determined by local base stacking and that the regular A and B-forms are not adequate to describe DNA in solution. For any one stretch of helix the width and depth of the major and minor grooves, the displacement and orientation of the base pairs relative to the helix axis, the helical periodicity and the global bend of the DNA are determined by the sequence of bases. Consequently, the structure and flexibility of the double helix is continuously variable and this must play a role in protein–DNA recognition.

If proteins are to recognize specific DNA sequences they need to 'read' the base sequence either through direct interactions, or through recognizing features of the overall structure of the DNA that are dependent upon the base sequence. In the latter case the protein needs to recognize the precise relative positions of the phosphate and sugar moieties that comprise the backbone of the DNA as well as the bases. For direct interactions proteins can access the functional groups of the base pairs in either the major or minor grooves. Because the structural variation and deformability of the double helix affects the accessibility and position of hydrogen bonding groups in the major and minor grooves, these aspects also play a role in direct recognition.

In general terms, in DNA close to the B-form structure, the major groove is wider and better suited to accommodate protein secondary structure than the minor groove. In A-like DNA the converse is true. Furthermore, in the major groove the pattern of hydrogen bond donors and acceptors is unique for each base-pair, whereas in the minor groove it is not possible to distinguish between AT and TA base-pairs nor between GC and CG base-pairs. Consequently, a major groove with B-like properties is best suited for allowing direct, sequence-specific interactions.

When we examine the known structures of protein–DNA complexes, we see that in all cases the protein interacts with both the base-pairs and the phosphate backbone of the DNA so that it has the potential to recognize the base sequence both directly and indirectly. So although direct recognition is more important, indirect recognition plays a role which varies between different complexes.

In a number of structures of protein–DNA complexes it is clear that the structure of DNA plays an important part in the recognition process. Evidence for this is that sometimes bases are highly conserved in different DNA targets, yet are not in direct contact with the protein (see Schwabe *et al.* 1993). A striking example of the role of non-contacted bases comes from mutations of the central base pairs of the 434 repressor binding site (Koudelka *et al.* 1987). Mutation from TA to AT had no effect, whereas mutation to GC or CG decreased the affinity of binding 50-fold. In a number of complexes it is apparent that the flexibility of DNA is important for protein–DNA recognition. Indeed, there are now many examples of protein–DNA complexes in which the DNA is significantly distorted. The binding of CAP and E2 proteins (Schultz *et al.* 1991; Hegde *et al.* 1992) to DNA results in significant local and global bending of the DNA double helix so that the DNA takes up the shape of the protein surface. Even subtle changes in the local structure of DNA can have important consequences for recognition. For example when TTK binds to its target, an ATA step at the 3′ side of the binding site is associated with a bend of 20° towards the protein in the binding site of finger 1 (Fairall *et al.* 1993). The A-T and T-A steps associated with this bend have a helical twist of 24° and 40° respectively. Although alternate low and high twist are characteristic of this sequence (Yoon *et al.* 1988), the twist of the A-T step is particularly low and results in the A and T bases of the binding site stacking directly over each other. This results in the T being displaced towards the protein by 2.5 Å, compared with the equivalent base in the binding site for finger 2. This permits a serine, which has a short side chain, to interact with the O4 of the T.

The most extreme example of DNA distortion is seen in the structure of the TATA binding protein bound to its DNA target in which the DNA has two 90° bends (Kim *et al.* 1993*a*, *b*). In this case (as mentioned above) the protein binds lengthways in the minor groove which is splayed apart. Clearly the structural properties of the TATAAAA sequence are an important factor in facilitating this distortion.

7. RECOGNIZING MORE THAN ONE DNA-SEQUENCE

Most specific DNA-binding proteins do not recognize a unique DNA target. Rather they recognize a family of related DNA sequences. Although in most structural analyses proteins are crystallized initially with their consensus targets, in a few cases there is structural information for the protein bound to more than one DNA sequence. These studies show that there are three ways in which proteins can interact with a non-consensus DNA target. In the simplest case, it appears that the protein can rearrange the conformation of surface side chains so as to create a slightly different network of hydrogen bonds with the alternative sequence. This is exemplified by the ERDBD bound to a non-consensus target in which a G-C base pair is replaced by an A-T (Schwabe *et al.* 1995). In this case a lysine sidechain moves so as to make alternative hydrogen bonds. However this mutation results in a reduced binding affinity. A similar rearrangement is likely to take place when TTK binds to different DNA-targets. In several of the TTK binding sites the thymine is substituted by a cytosine (Harrison & Travers 1990; Brown *et al.* 1991; Read & Manley 1992). Presumably a rotation about the Cβ-Oγ bond would allow the serine to accept a hydrogen bond from N4 of the cytosine with no overall change in the geometry of the interaction. Although serine can act as both a hydrogen bond acceptor and donor and hence hydrogen bond to all of the bases, only substitution of the thymine by a cytosine, but not by a purine, should permit the DNA deformation described above. So in this case the specificity is being sensed in terms of the DNA structure. For other proteins (for example the phage 434 repressor; see Rodgers & Harrison 1993) a comparison of complexes containing different DNA targets reveals that in addition to sidechain rearrangements, the DNA structure and relative position of the protein can also change.

Interestingly, for certain dimeric proteins the spacing between half sites is invariant (e.g. the ERDBD and other hormone receptors; see Schwabe *et al.* 1993), whereas for others the spacing can vary by one nucleotide without compromising specific DNA-binding, e.g. GCN4 and NF-κB. In these latter cases the protein adapts to the different spacing by distorting the DNA and protein structure on the different sequences, such that the interactions of the individual monomers with DNA are essentially identical (Ellenberger *et al.* 1992; König & Richmond 1993; Ghosh *et al.* 1995; Müller *et al.* 1995).

The third and final way in which proteins seem to overcome the problem of non-consensus DNA-targets is through part of the protein (e.g. one monomer in a dimer) binding non-specifically to DNA. This type of binding appears to be important for the nuclear hormone receptors for which one half site of the palindromic binding site frequently bears little resemblance to the consensus. The structure of the GRDBD-DNA complex shows very clearly how one half of the dimer adjusts to non-specific interaction with DNA (Luisi *et al.* 1991).

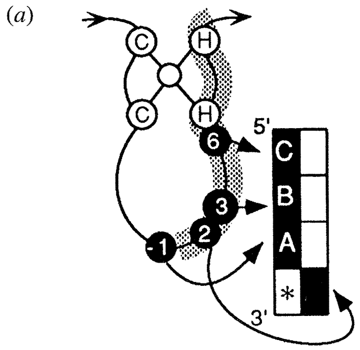
8. IS THERE A DISCERNIBLE RECOGNITION CODE?

A long-standing question regarding protein–DNA recognition is whether or not there is a recognition code, in some way analogous to the genetic code. In 1976 Seeman *et al.* (Seeman *et al.* 1976) recognized that sequence specific DNA-binding proteins were likely to interact with bases in the major groove of the double helix, where the pattern of hydrogen bond donors and acceptors is unique for each base-pair. Their classic paper also predicted that certain amino acids were ideally suited to recognizing certain base pairs e.g. Asn or Gln contacting adenine and Arg contacting guanine. Although these contacts are seen in several of the protein–DNA complexes, there seems to be great variation in the way that each amino acid is employed to interact with DNA base-pairs. This led to the view that protein–DNA interfaces were rather too complex to allow detailed predictions of the contacts involved, and that there was no simple recognition code like that of the genetic code.

Now, 20 years later, we have many more structures of protein–DNA complexes. It is therefore pertinent to ask again whether we can devise a recognition code, or indeed predict that such a code would exist. It has become clear that for each of the different DNA-binding motifs, such as the helix-turn-helix motif, homeodomain, zinc-finger and hormone receptor, there is a pattern of contacts that is reasonably conserved for members of the same family. This type of observation has led some researchers to attempt to devise recognition codes for many of the different DNA-binding motifs (Suzuki & Yagi 1994). Unfortunately, a major limitation of these studies is that the effect of DNA structure and the role of water molecules are very difficult to predict and furthermore, unless different families have structural features in common, such predictions are limited to members of the same type of DNA-binding domain.

Amongst DNA-binding motifs, the classical C2-H2 zinc-finger motif seems to provide the best candidate for understanding the rules for recognition. Whether this can be called a code is questionable. The framework of the zinc-finger is very simple and its orientation with respect to the DNA is probably dominated by the amino acid to base contacts. These occur from four main positions in the zinc-finger: one immediately preceding the α-helix and the other 3 from within the α-helix (figure 3*a*) (Pavletich & Pabo 1991; Fairall *et al.* 1993). The binding site for a zinc-finger generally spans 3 bases on one strand with a single base contact to the opposite strand, but each zinc-finger does not necessarily contact all of these positions. Also, generally, there is one to one recognition in that a single amino-acid makes contact to a single base.

The apparent simplicity of the zinc-finger led to it being the target of mutagenesis experiments aimed at deriving a recognition code (Desjarlais & Berg 1992; Nardelli *et al.* 1992). This has not proved easy, because concerted changes are in some instances required to obtain a new specificity. This approach showed that a simple recognition code did not exist, but that it should be possible to build a catalogue of zinc-fingers to recognize different DNA sequences. The phage display system appears to be the most efficient way of obtaining a large library of zinc-finger motifs. In this method zinc-fingers are cloned as fusions to the coat proteins of filamentous bacteriophage and as such are displayed on the capsid which encloses the viral genome. The zinc-fingers are then randomized in the positions important for sequence specific DNA-binding and the DNA sequence of interest is used to select the corresponding zinc-finger phage coat fusion protein. Multiple rounds of selection and amplification result in an enrichment of the relevant zinc-finger fusion proteins. In this way Choo & Klug have been able to devise some rules for DNA recognition by zinc-fingers (figure 3*b*) (Choo & Klug 1994). However, the phage display system has not been successful in selecting zinc-fingers for all DNA sequences. Clearly this approach is



(b)

position

base	5' C	B	A 3'

G	Arg(6)
	Ser(6)/Asp(2*)
	Thr(6)/Asp(2*)
		His(3)	Arg(-1)/Asp(2)

A		Asn(3)	Gln(-1)/Ala(2)

T	Ser(6)/Asp(2*)
	Thr(6)/Asp(2*)
		Ala(3)
		Ser(3)
		Val(3)
			Asn(-1)
			Gln(-1)/Ser(2)

C		Asp(3)
		Leu(3)
		Thr(3)
		Val(3)
			Asp(-1)

Figure 3. DNA-recognition by C2-H2 zinc-fingers: (*a*) Schematic representation of a zinc-finger showing the positions of the amino-acid to base contacts. The α-helix is shown shaded; (*b*) Consensus zinc-finger recognition code derived from screening a zinc-finger library (Choo & Klug 1994). The helical position of residues are given in brackets. Note that an asterisk indicates a base within the binding site of an adjacent zinc-finger.

much better for designing new DNA-binding proteins (see later section), than for predicting the binding site for a protein whose binding site is unknown. This is because, for example, in the case of multi zinc-finger proteins, not all zinc-fingers contact the DNA and also not all zinc-fingers bind to DNA in the simple way described above (Pavletich & Pabo 1993).

9. THE ROLE OF SOLVENT IN SPECIFICITY

In most of the protein–DNA crystal structures determined to date there are ordered water molecules present at the protein–DNA interface. The number of water molecules and their role seems to differ in the different complexes. Whereas this is partly a consequence of the resolution at which the structures have been solved, it also seems that the nature of different protein–DNA interfaces can differ significantly in this regard. In the structure of the *trp* repressor bound to its DNA target there are few direct amino acid to base contacts (Otwinowski *et al.* 1988). However there are many water molecules that participate in a network of hydrogen bonds with protein and DNA. Significantly, many of these water molecules were found to reside in essentially identical positions on the DNA in the absence of protein (Shakked *et al.* 1994). Thus it appears that the protein is specifically recognizing not just the DNA, but the associated water structure as well. The structure of the ERDBD bound to its DNA target site (Schwabe *et al.* 1993) shows a number of well ordered water molecules (low crystallographic temperature factors) that appear to have no direct access to bulk solvent. Water molecules are present in the same position in two different crystal forms (and in the related TR/RXR heterodimer bound to DNA) and appear to be necessary for specific recognition of the DNA target. In contrast, in the structure of the related GRDBD bound to its DNA target there are far fewer ordered water molecules, resulting in a much less polar protein–DNA interface. For C2-H2 zinc-finger proteins water molecules present at the protein–DNA interface seem to play a somewhat less important role. Consistent with the role of water in protein–DNA recognition derived from X-ray analysis, NMR spectroscopic techniques also indicate that water molecules are present at the protein–DNA interface, but these are often in quite fast exchange with bulk solvent (Qian *et al.* 1993).

10. UNDERSTANDING THE THERMODYNAMICS AND KINETICS

In addition to participating in the ordered interface between protein and DNA (in some cases playing a role in the specificity of protein–DNA interactions), water molecules may play a very important role in the energetics of protein–DNA interactions. In dilute solution the role of water seems clear. Upon complex formation many waters bound at the interface of the protein and DNA will be displaced (Garner & Rau 1995). This results in an entropic contribution to DNA binding that may favour complex formation. Those water molecules that remain at the interface participate in a hydrogen bonding network that makes an enthalpic contribution that may also favour complex formation. It is difficult to assess the thermodynamic importance of these two factors, because we can rarely assess the hydration state of the molecules before complex formation. Furthermore, the available crystal structures only reveal those water molecules that are extremely well ordered. However, it is likely that the *in vitro* experimental conditions probably bear little resemblance to the environment *in vivo*. Inside the cell (in both prokaryotes and eukaryotes), there is an extremely high concentration of dissolved macromolecules (300–400 mg ml^{-1}, perhaps somewhat lower in eukaryotes) (reviewed in Garner & Burg 1994). This means that the environment inside the cell is much more like that in a macromolecular crystal than in solution. This macromolecular crowding has a number of effects. Firstly, it enhances the affinity of intermolecular interactions and secondly, it reduces the macromolecular diffusion rate. Recent experimental data suggest that the change in hydration when proteins bind to DNA is a key thermodynamic variable in protein–DNA interactions, and that in general as the two surfaces come together, the interaction energies associated with hydration forces increase exponentially with the number of waters displaced (Garner & Rau 1995).

So is the displacement of water molecules sufficient to explain protein–DNA binding affinities? Experimental studies show that on the formation of protein–DNA complexes there is a large increase in heat capacity (Spolar & Record 1994). This is similar to the increase in heat capacity when protein molecules undergo a transition from a denatured to a folded state, and there is evidence that a similar transition may occur when proteins interact with DNA (Percipalle *et al.* 1995). In other words the formation of a complex is associated with local folding events. Many DNA-binding proteins appear to be somewhat disordered in the absence of DNA - although this is not always apparent from individual crystal structures. However for the *trp* repressor, two crystal forms of the unbound protein are significantly different from each other and from the protein in the complex with the DNA (Schevitz *et al.* 1985; Lawson *et al.* 1988; Otwinowski *et al.* 1988). These differences are mainly confined to the DNA-binding surface of the repressor. Another example is the ERDBD which is a monomer in solution, but binds to DNA as a dimer. The dimer interface appears disordered by NMR analysis before binding to DNA, but becomes ordered upon DNA-binding (Schwabe *et al.* 1990, 1993). Adjacent zinc-finger domains are clearly flexibly oriented with respect to each other, before binding to DNA (Nakaseko *et al.* 1992), but they have a fixed relative orientation when bound to DNA. All of these examples support the idea that the formation of a protein–DNA complex involves local folding events and these are coupled to the thermodynamics of binding.

In addition to the need to account for the thermodynamics (binding affinity) of protein–DNA interactions, we need to understand the kinetic events of binding and release of proteins from their DNA targets. These are particularly important if we are to understand the processes by which genes are turned on and off. As mentioned above, one consequence of macromolecular crowding is that the rate of macromolecular diffusion is reduced. It has long been proposed that DNA-binding proteins find their target so rapidly that they must employ some form of scanning along the DNA, rather that simple three dimensional diffusion. This reduced rate of diffusion in the cellular environment further highlights our poor understanding of the kinetic processes that allow rapid complex formation.

11. CUSTOM BUILT DNA-BINDING PROTEINS

It is clear that we have some way to go before we fully understand how proteins recognize DNA. However the available structural information has been of immense help in allowing us to design custom-built DNA-binding proteins that will recognize either designed, or specific naturally occurring DNA targets. This is clearly very important for both biotechnological and medical applications. Two different strategies have proven successful. The first is to take existing DNA-binding domains and link them together. The best example of this is the artificial protein ZFHD1, that was constructed by linking two zinc-fingers with a homeodomain (Pomerantz *et al.* 1995). This protein recognizes the combined binding sites of the zinc-fingers and the homeodomain. When this protein is attached to an activation domain it has been shown to regulate transcription *in vivo* in a sequence specific manner. A disadvantage of this methodology is that it is based upon existing DNA-binding domains and therefore the number of sequences that can be recognized is limited.

A more general approach has been to use the architecture of the classical zinc finger to design DNA-binding proteins with genuinely novel DNA-binding specificity. The most comprehensive and efficient approach is to employ the phage display technique described above, to select for zinc fingers that recognize desired base triplets. These may then be linked together to recognize a binding site of the desired sequence and length. The success of this strategy has been strikingly demonstrated through the design of a novel three-zinc-finger peptide that recognizes the oncogene BCR-ABL and remarkably, inhibits its transcription (Choo *et al.* 1994).

In conclusion, structural studies of protein–DNA complexes have revealed stunningly beautiful images of nature at work. These structures have greatly enhanced our understanding of protein–DNA recognition and have culminated in our ability to design novel proteins to recognize specific DNA sequences. Despite these successes however, an in depth understanding of the thermodynamics and kinetics of protein–DNA recognition lies some distance in the future.

REFERENCES

Arnott, S. & Selsing, E. 1974*a* The structure of polydeoxyguanylic acid.polydeoxycytidylic acid. *J. Molec. Biol.* **88**, 551–552.

Arnott, S. & Selsing, E. 1974*b* Structures for the polynucleotide complexes poly(dA).poly(dT) and poly(dT).poly(dA).poly(dT). *J. Molec. Biol.* **88**, 509–521.

Brown, J. L., Sonoda, S., Ueda, H., Scott, M. P. & Wu, C. 1991 Repression of the *Drosophila fushi tarazu* (*ftz*) segmentation gene. *EMBO J.* **10**, 665–674.

Choo, Y. & Klug, A. 1994 Selection of DNA binding sites for zinc fingers using rationally randomised DNA reveals coded interactions. *Proc. natn. Acad. Sci. U.S.A.* **91**, 11168–11172.

Choo, Y., Sánchez-Garciá, I. & Klug, A. 1994 In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature, Lond.* **372**, 642–645.

Desjarlais, J. R. & Berg, J. M. 1992 Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. natn. Acad. Sci. U.S.A.* **89**, 7345–7349.

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

Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. 1993 The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature, Lond.* **366**, 483–487.

Fuller, W., Wilkins, M. H. F., Wilson, H. R. & Hamilton, L. D. 1965 The molecular configuration of deoxyribonucleic acid. IV. X-ray diffraction study of the A-form. *J. Molec. Biol.* **12**, 60–80.

Garner, M. M. & Burg, M. B. 1994 Macromolecular crowding and confinement in cells exposed to hypertonicity. *Am. J. Physiol. Cell Physiol.* **266**, 877–892.

Garner, M. M. & Rau, D. C. 1995 Water release associated with specific binding of *gal* repressor. *EMBO J.* **14**, 1257–1263.

Ghosh, G., Van Duyne, G., Ghosh, S. & Sigler, P. B. 1995 Structure of NF-κB p50 homodimer bound to a *κ*B site. *Nature, Lond.* **373**, 303–310.

Glover, J. N. M. & Harrison, S. C. 1995 Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature, Lond.* **373**, 257–261.

Harrison, S. D. & Travers, A. A. 1990 The *tramtrack* gene encodes a *Drosophila* finger protein that interacts with the *ftz* transcriptional regulatory region and shows a novel embryonic expression pattern. *EMBO J.* **9**, 207–216.

Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. 1992 Crystal-structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature, Lond.* **359**, 505–512.

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

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

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. 1990 Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell* **63**, 579–590.

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

König, P. & Richmond, T. J. 1993 The X-ray structure of the GCN4-b-Zip bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *J. Molec. Biol.* **233**, 139–154.

Koudelka, G. B., Harrison, S. C. & Ptashne, M. 1987 Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature, Lond.* **326**, 886–888.

Kraulis, P. J. 1991 MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. 1960 The molecular configuration of deoxyribonucleic acid. II. Molecular models and their fourier transforms. *J. Molec. Biol.* **2**, 38–64.

Lawson, C. L., Zhang, R. G., Schevitz, R. W., Otwinowski, Z., Joachimiak, A. & Sigler, P. B. 1988 Flexibility of the DNA-binding domains of *trp* repressor. *Proteins* **3**, 18–31.

Luisi, B. F., Xu, W. X., Owtinowski, Z., Freedman, L. P., Yamamoto, K. R. & Sigler, P. B. 1991 Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature, Lond.* **352**, 497–505.

Müller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. 1995 Structure of the NF-κB p50 homodimer bound to DNA. *Nature, Lond.* **373**, 311–317.

Nakaseko, Y., Neuhaus, D., Klug, A. & Rhodes, D. 1992 Adjacent zinc-finger motifs in multiple zinc-finger peptides from SWI5 form structurally independent, flexibly linked domains. *J. Molec. Biol.* **228**, 619–636.

Nardelli, J., Gibson, T. J. & Charnay, P. 1992 Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucl. Acids Res.* **20**, 4137–4144.

Nicholls, A., Bharadwaj, R. & Honig, R. 1993 GRASP – graphical representation and analysis of surface-properties. *Biophys. J* **64**, 166.

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

Pavletich, N. P. & Pabo, C. O. 1993 Crystal structure of a five-finger Gli-DNA complex: New perspectives on zinc fingers. *Science, Wash.* **261**, 1701–1707.

Pavletich, N. P. & Pabo, C. O. 1991 Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science, Wash.* **252**, 809–817.

Percipalle, P., Simoncsits, A., Zakhariev, S., Guarnaccia, C., Sanchez, R. & Pongor, S. 1995 Rationally designed helix-turn-helix proteins and their conformational changes upon DNA-binding. *EMBO J.* **14**, 3200–3205.

Pomerantz, J. L., Sharp, P. A. & Pabo, C. O. 1995 Structure-based design of transcription factors. *Science, Wash.* **267**, 93–96.

Qian, Y. Q., Otting, G. & Wüthrich, K. 1993 NMR detection of hydration water in the intermolecular interface of a protein–DNA complex. *J. Am. Chem. Soc.* **115**, 1189–1190.

Read, D. & Manley, J. L. 1992 Alternatively spliced transcripts of the *Drosophila tramtrack* gene encode zinc finger proteins with distinct DNA binding specificities. *EMBO J.* **11**, 1035–1044.

Rodgers, D. W. & Harrison, S. C. 1993 The complex between the phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half sites. *Structure* **1**, 227–240.

Schevitz, R. W., Otwinowski, Z., Joachimiak, A. J., Sigler, P. B. & Lawson, C. L. 1985 The 3-dimensional structure of *trp* repressor. *Nature, Lond.* **317**, 782–786.

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

Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. 1993 The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: How receptors discriminate between their response elements. *Cell* **75**, 567–578.

Schwabe, J. W. R., Chapman, L. & Rhodes, D. 1995 The oestrogen receptor recognizes an imperfectly palindromic response element through an alternative side-chain conformation. *Structure* **3**, 201–213.

Schwabe, J. W. R. & Klug, A. 1994 Zinc mining for protein domains. *Nature Struct. Biol.* **1**, 345–349.

Schwabe, J. W. R., Neuhaus, D. & Rhodes, D. 1990 Solution structure of the DNA-binding domain of the oestrogen receptor. *Nature, Lond.* **348**, 458–461.

Seeman, N. D., Rosenberg, J. M. & Rich, A. 1976 Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. natn. Acad. Sci. U.S.A.* **73**, 804–808.

Shakked, Z., Guzikevichguerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. & Sigler, P. 1994 Determinants of repressor-operator recognition from the structure of the *trp* operator binding site. *Nature, Lond.* **368**, 469–473.

Somers, W. S. & Phillips, S. E. V. 1992 Crystal structure of the *met* repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β-strands. *Nature, Lond.* **359**, 387–393.

Spolar, R. S. & Record, M. T. 1994 Coupling of local folding to site-specific binding of proteins to DNA. *Science, Wash.* **263**, 777–784.

Suzuki, M. & Yagi, N. 1994 DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. natn. Acad. Sci. U.S.A.* **91**, 12357–12361.

Wolberger, C., Vershon, A. K., Lui, B., Johnson, A. D. & Pabo, C. O. 1991 Crystal structure of a MATa2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**, 517–528.

Yoon, C., Prive, G. G., Goodsell, D. S. & Dickerson, R. E. 1988 The structure of an alternating B-helix and its relationship to A-tract DNA. *Proc. natn. Acad. Sci. U.S.A.* **85**, 6332–6336.

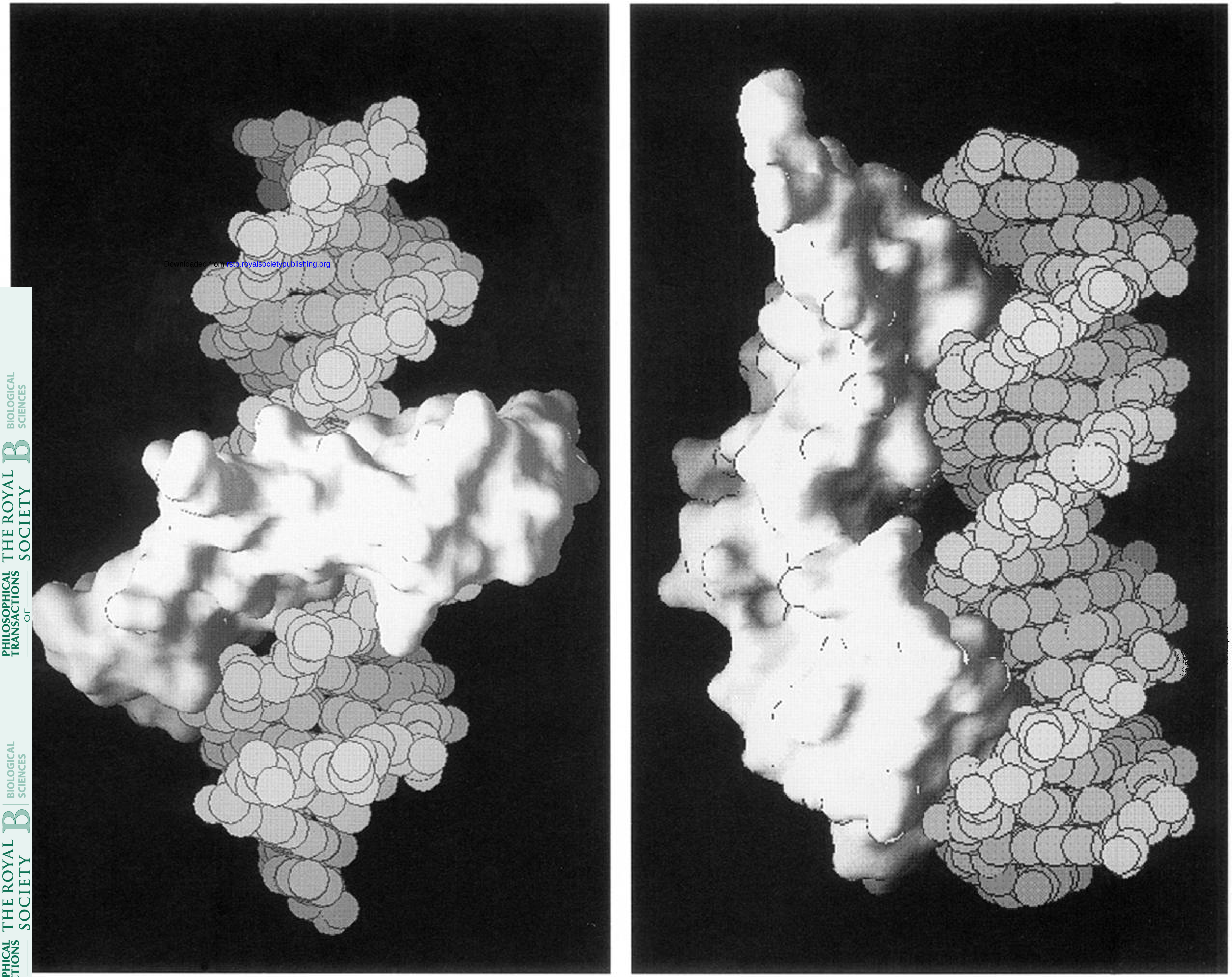


Figure 2. The shape of the protein and DNA are complementary. Space-filling models of: the TTKDBD-DNA complex (left-hand figure); and the ERDBD-DNA complex with the protein shown as a molecular surface (right-hand figure), drawn using the program GRASP (Nicholls *et al.* 1993).