DNA Binding by Proteins

ROBERT SCHLEIF

Study of proteins that recognize specific DNA sequences has yielded much information, but the field is still in its infancy. Already two major structural motifs have been discovered, the helix-turn-helix and zinc finger, and numerous examples of DNA-binding proteins containing either of them are known. The restriction enzyme Eco RI uses yet a different motif. Additional motifs are likely to be found as well. There is a growing understanding of some of the physical chemistry involved in protein-DNA binding, but much remains to be learned before it becomes possible to engineer a protein that binds to a specific DNA sequence.

DNA CONTAINS THE INFORMATION NECESSARY IN EACH cell for it to grow, divide, respond to its environment, and differentiate. Over the years, study on hundreds of different genes has demonstrated that most frequently the cellular regulation of transcription is at the level of initiation. This review focuses on a subproblem of the regulation question. For regulation to occur, the regulatory sequences associated with a specific gene must be identifiable in the presence of thousands or hundreds of thousands of other genes. As far as we know, this recognition process is performed almost entirely by proteins. How do proteins recognize specific DNA sequences?

Leaving aside special means for increasing information content in small areas such as auxiliary methylation of certain bases in special contexts, the sequence of bases in DNA contains all the information that can be stored there. Most of the DNA in cells is thought to exist in the B form, in which the base pairs that cross from one deoxyribose-phosphate backbone to the other are displaced from the helical axis. This asymmetric location creates a smaller minor groove and a larger major groove in the DNA (Fig. 1).

In principle, sequence can be directly "read" by hydrogen-bonding from the major groove of B form DNA without disruption of the double-stranded structure (*1*). This is shown in a more detailed drawing of the T-A and C-G base pairs, (Fig. 2). Both hydrogen-bond donors and hydrogen-bond acceptors are exposed in the major and minor grooves, as shown in the schematic representation (Fig. 2, bottom) (*2*). These drawings illustrate, as first pointed out by Seeman *et al.* (*1*), that with reasonable flexibility in protein structures and utilizing only hydrogen bonding to bases, formation of a minimum of two hydrogen bonds in the major groove is necessary to unambiguously read the identity of a base pair. For example, detecting a hydrogen-bond acceptor at the position of the acceptor on a T does not distinguish the T from a G. In the minor groove, A and T are indistinguishable and G and C are barely distinguishable; however, an A-T or T-A base pair can be distinguished from a G-C or C-G base pair. The methyl group on the T provides an additional sequence recognition element, that can be identified from the major groove. The major groove of B form DNA will nicely accommodate an α-helix of a protein, and therefore suitable hydrogen bond–forming residues on an exterior face of the helix should be capable of "reading" a DNA sequence.

In addition, sequence information can also be expressed, and presumably identified, by means of structural inhomogeneities along the DNA helix. Each of the repeating units of phosphate-sugar-base along one strand of DNA possesses six bonds about which rotation is possible, in addition to the various pucker conformations of the sugar. Therefore it is not surprising that DNA can exist in a variety of helical forms, and that the helical forms are not perfectly regular. X-ray crystallography of synthetic oligonucleotides has shown a surprisingly large nucleotide-to-nucleotide variation between the repeating units along the DNA helix (*3*). The helical twist from one nucleotide to the next, the tilt of a base pair around an axis within the plane of the hydrogen bonds and perpendicular to the hydrogen bonds between the bases, the roll of a base pair about an axis parallel to the hydrogen bonds between the bases, the propeller twist of one base with respect to the other in a base pair, and the buckle of the bases out of one plane after removal of the propeller pitch, all vary from one base pair to the next along the DNA. In principle then, a protein reading the DNA sequence

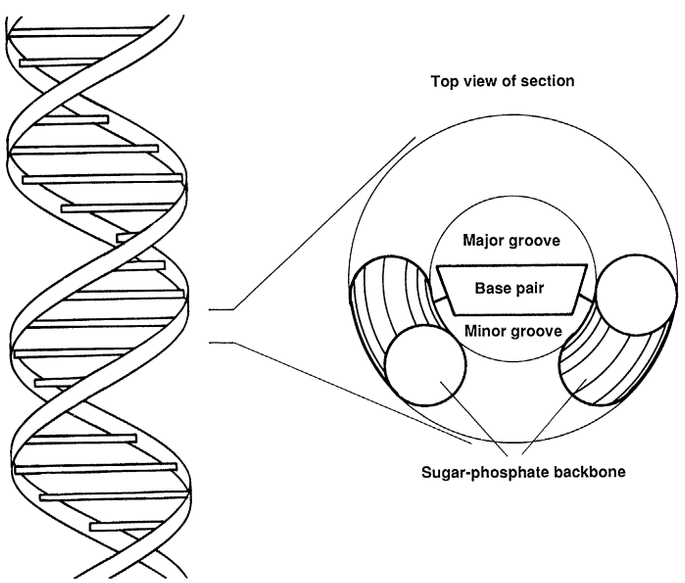

Fig. 1. The B form of DNA. (Left) Schematic drawing of the DNA double helix with the ribbon indicating the sugar-phosphate backbone, and the crosspieces indicating the base pairs. (Right) Schematic of a section viewing from the top showing one base pair. This indicates the locations of the major and minor grooves and shows the relation of the base pair to the sugar-phosphate backbone of the DNA.

Dr. Schleif is a professor in the Graduate Department of Biochemistry, Brandeis University, Waltham, MA 02254.

need pay no attention to the chemical or hydrogen bonding differences between the four bases along the DNA. Instead, it could indirectly read information or bind to a specific site by recognizing the overall structure generated by a sequence, for example, by locating the precise positions of phosphates along the DNA backbone.

The large T antigen protein of simian virus SV40, which regulates gene expression and DNA replication of the virus, appears to read part of its recognition sequence indirectly. Between two regions of the DNA that are contacted by the protein lies a stretch of nucleotides that appear not to be contacted, but which generate a bend necessary for tight binding by the protein (*4*). A similar situation exists for the phage 434 repressor protein as discussed below. Nucleosomes also bind to DNA at locations defined by natural bends in the DNA (*5*). In reality a protein is likely to read sequence both by utilizing the pattern of base-specific hydrogen bonding and methyl groups and by recognizing overall structure.

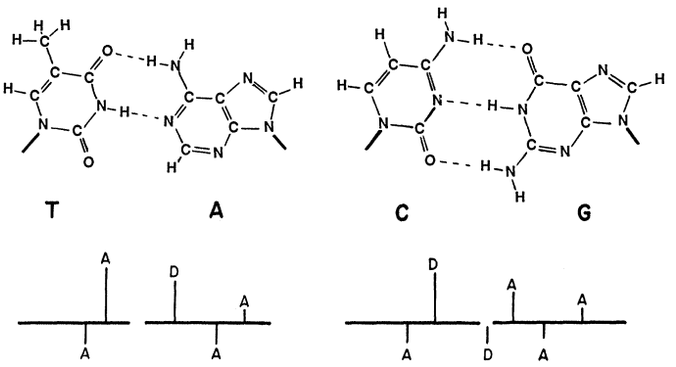

Fig. 2. T-A and C-G base pairs. (Top) Covalent structure. The orientation is that of Fig. 1. The top edge represents the portion exposed in the major groove, and the bottom edge represents that exposed in the minor groove. (Bottom) Schematic of the base pairs above indicating the approximate locations of hydrogen-bond donors (D) and acceptors (A). Above the line is the region in the major groove, and below the line is the minor groove region.

Why Many Protein-Binding Sites Are Symmetric or Repeated

Regulatory proteins must possess great selectivity for the correct binding site by binding with high affinity to the correct site and with relatively low affinity to most other sites. Other DNA-binding proteins do not need to bind with such great selectivity.

How can the affinity or binding selectivity of a monomeric DNA-binding protein be increased? Consider binding by a protein that makes a fixed number of hydrogen bonds and van der Waals interactions. These provide a fixed amount of binding energy, or more properly, enthalpy. In the process of binding, the center of mass of the protein must be properly located with respect to the DNA site and the protein must assume the correct angular orientation. Finally, the protein and DNA likely undergo conformational fluctuations. We expect that only a subset of the protein's conformational states is capable of binding to the DNA site. We can view this conformational flexibility as reducing the effective concentration of the protein, or equivalently, as a situation requiring binding energy to drive the protein into the correct conformational state. With a fixed set of interactions, binding can be increased by any means that hold the protein in the correct place or orientation, or that freeze the protein in the correct conformational state.

Many protein-binding sites are symmetric, and the proteins that bind to these sites possess either two or four identical subunits. There is a good reason for this. Rather than simply increasing the size of a monomeric protein so that it could make more interactions with DNA, becoming dimeric and contacting a repeat of the binding site avoids the need for synthesizing excessively long proteins, a process that can become inefficient because of the increasing probability of error.

A dimeric protein binds even more tightly to DNA than we might at first expect from the affinity of the individual monomers. For a typical monomeric protein the entropy change in being correctly positioned and oriented as it binds is substantial. However, similar entropy is required to position and orient the dimer. Therefore, for a dimer, the binding of one monomer correctly positions the second monomer so that its binding interactions can be used primarily to increase tightness of binding rather than properly locating or orienting the protein.

Cooperativity is another way to think of the "extra" affinity provided by two connected binding domains. The interaction energy between monomers generates cooperativity in binding so that the concentration of monomer required to achieve binding to both sites is lower than if no interaction energy were present. The binding of a dimeric protein to two repeated sites is merely an extreme in which the interaction energy between the dimers is sufficient to hold them together even before they have bound to the DNA.

Ion Displacement by Binding Proteins

The phosphate backbone of DNA creates a problem for DNA-binding proteins. Electrostatic attraction between the negatively charged phosphate groups and positively charged ions or molecules in solution or in the cell increases the local concentration of the charge-neutralizing positive ions along the DNA backbone.

As a protein binds to a specific site on the DNA, what happens to the neutralizing charges? If they are displaced, then they ought to be considered part of the binding reaction, and as a result, the equilibrium binding constant will depend not only on the concentration of the protein and DNA, but also on the concentration of the charge-neutralizing ion. Since as many as 10 to 15 such ions may be displaced as a protein binds to DNA, the affinity of a protein for DNA can vary markedly with salt concentration (*6*).

For example, the affinity of *lac* repressor to *lac* operator DNA in vitro changes by a factor of 20 as the concentration of NaCl changes from 0.1*M* to 0.2*M* (*7, 8*). If the salt concentration in bacterial growth media varies, the intracellular concentration of ions also varies. Ion displacement should generate huge changes in the affinity of some proteins for DNA. Despite this, cells grow under these different conditions, and preliminary measurements suggest that changes in the growth medium generate little change in the binding of RNA polymerase or even *lac* repressor (*9*). How the cells manage to compensate for the changes in salt concentrations is not yet known. Perhaps the main ions near the phosphates of DNA are polycations whose concentration does not change with changes in the growth medium. Alternatively, many proteins may use a mechanism of binding ions to themselves in the binding process to compensate for the ions displaced from the DNA (*10*).

Crystallography of Helix-Turn-Helix Proteins

The first DNA-binding proteins whose structures were determined by x-ray crystallography were the cyclic AMP receptor protein (CRP) of *Escherichia coli*, the bacteriophage λ regulatory

protein Cro, and the NH_2-terminal domain of λ repressor (*11*). More recently the structure of the repressor of the tryptophan operon, TrpR has also been solved (*12*). All these proteins were crystallized in the absence of DNA. Therefore, one problem after determining their structures was recognizing the DNA-contacting portion of the proteins. Among other features, their surfaces possess protrusions separated by about 35 Å and, in some of the proteins, oriented at roughly the same angle as the major groove in DNA. This protrusion structure consists of a short α-helix that would be oriented across the major groove, a turn, and a second helix that is predicted to lie partly within the major groove where it could make specific residue-base interactions (Fig. 3). This second helix is termed the recognition helix. It is not a small or simple task to identify the DNA-contacting portion of the proteins on the basis of their shapes, and initially CRP was proposed to bind to left helical DNA.

Recently the 3.2 to 4.5 Å electron density map of co-crystals of the DNA-binding domain of phage 434 repressor and its DNA-binding site have been published (*13*). These as well as lower resolution structures (*14*) show that the helix-turn-helix portion of the protein does contact the DNA. The NH_2-terminal end of the recognition helix points somewhat into the groove, much as λ repressor protein was predicted to contact its DNA-binding site. Each of the subunits contacts the edges of bases in a major groove region from one side of the DNA (Fig. 4). Of necessity then, a minor groove of the DNA lies between these two major groove regions. The protein makes contact with the phosphates on the backbone in this region, but it does not appear to have any base-specific contacts within the minor groove. Binding, however, is sensitive to the identity of the four noncontacted bases between the two contacted regions. These bases can be A or T, but not G or C. Apparently A and T permit the DNA to be overwound in this area. As a result the minor groove is narrowed, and perhaps the contacted regions are brought into more favorable relative orientation (*15*).

All proteins need not contact two major groove regions in the same manner as the 434 repressor. For example, the recognition helices can conceivably approach the DNA from either side (Fig. 4). In this case ten contiguous base pairs of DNA can be contacted by the protein. Because of the helix geometry, the two recognition helices need not encounter steric hinderance with each other, even though both may be contacting the same base pair.

At the resolution presently available, it appears that the 434 repressor makes three or four hydrogen bonds to phosphates of the backbone, primarily with peptide amino groups. Three Gln's in the recognition helix seem to make the primary contacts to the bases. One of these forms two hydrogen bonds to a base, one appears to make one hydrogen bond and have a van der Waals interaction with the methyl group on a T, and one Gln makes a single hydrogen bond to each of two consecutive bases.

The structures of the CRP, Cro, λ repressor, and 434 repressor in solution display significant differences in the orientations of the recognition helices relative to the major groove (*16*). The structure of Cro is most compatible with the recognition helix being parallel and tangential to the major groove of the DNA. The orientation of λ repressor DNA-binding domains suggest, instead, that the NH_2-terminal ends of the recognition helices would point inward toward the DNA, and therefore their more important DNA contacts would derive from amino acids near the NH_2-terminus of this helix. Uncertainties remained, however, since the λ repressor NH_2-terminal fragment used in the structure determination contains only 92 of the 236 amino acids of the complete repressor, and the subunit-subunit interface in Cro appears to be flexible. Biochemical experiments show that Cro and λ repressor contact DNA so similarly that their recognition helices most likely are oriented similarly. Observations of the NH_2-terminal fragment of the 434 phage repressor in association with DNA suggest that the inward-pointing recognition helix is more common.

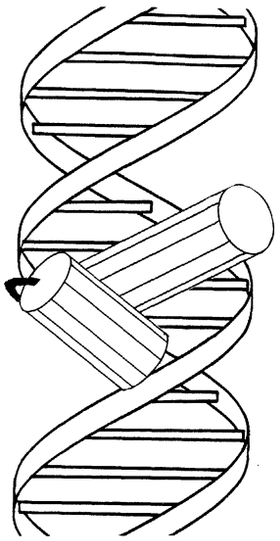

Fig. 3. The helix-turn-helix with the recognition helix in the major groove of DNA.

Not only are the amino acid sequences in the helix-turn-helix regions of Cro, λ repressor, and CRP similar to each other, but they are also found in other DNA-binding proteins and rarely in proteins that do not bind DNA (*17, 18*). Thus, the presence of such a sequence in a protein suggests that the protein binds DNA and that the helix-turn-helix structure is the major contacting region (*18*). Recently a conserved sequence of about 60 amino acids called the homeo domain has been observed in proteins that determine developmental fates of cells (*19*). A part of the homeo domain possesses significant homology to the helix-turn-helix sequence, and the homeo domain has been shown to bind to specific DNA sequences (*20*).

Genetic Studies of the Helix-Turn-Helix Structure

In the initial absence of direct structure determinations of protein-DNA complexes, it was genetics that provided strong evidence that the helix-turn-helix structure contacted DNA and recognized specific sequences. The experiments were greatly facilitated by the fact that the proteins were dimers with each subunit contacting DNA. An alteration in the DNA-contacting surface of one subunit should not interfere with its ability to fold and to dimerize with a wild-type subunit. The heterodimer thus formed should bind DNA less well however, since one of its DNA-contacting domains is defective. The gene encoding the defective subunit acts as a dominant negative since it can sequester the wild-type subunits in defective dimers and thus lower their effective concentration.

Dominant negative mutations in λ repressor are concentrated in the helix-turn-helix region (*21*). Similar mutations have been found in the *lac* repressor and *tet* repressor (*22*) in regions identified as helix-turn-helix regions on the basis of their amino acid sequences. Although the majority of dominant negatives lie in the DNA helix-turn-helix portion of proteins, some may result from changes to residues outside the helix-turn-helix regions or from alterations that misorient the DNA-contacting domains of the two subunits.

What about the reverse type of mutation? If simple modifications to the basic structure of a protein can increase the number of interactions, then we would expect several types of mutations that

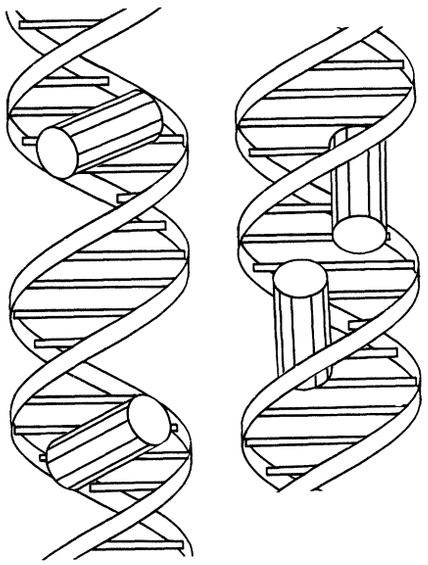

Fig. 4. Phage 434 repressor bound to operator DNA. (Left) The locations of the recognition helices; contact is to the major grooves from above. (Right) An hypothetical structure in which two recognition helices have entered from either side of the DNA so as to contact ten contiguous base pairs.

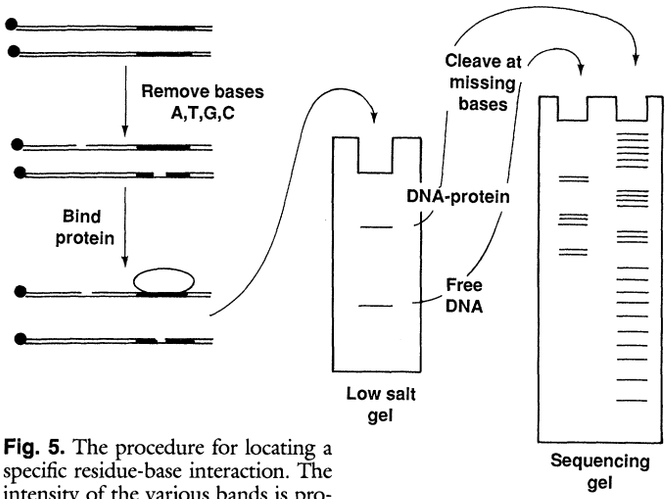

Fig. 5. The procedure for locating a specific residue-base interaction. The intensity of the various bands is proportional to the numbers of DNA molecules that were missing a base at that position. By this technique, those bands that are enriched in the free DNA lane correspond to positions where a missing base interferes with protein binding, and the bands enriched in the protein-DNA lane correspond to positions where a missing base does not affect binding.

bind more tightly to DNA. Such mutations have been isolated and found to lie in three locations in λ repressor (*23*). One of the tighter binding mutations introduced a Lys in a position where it could interact with a phosphate group on the DNA backbone. Most likely such a new interaction would displace a charge-neutralizing cation from the vicinity of the phosphate as the mutant repressor binds to DNA. Indeed, the mutant repressor has increased affinity for its normal binding site as well as increased affinity for nonoperator DNA. Furthermore, its binding affinity shows a higher salt dependence than wild type, indicating that it displaces more cations than the wild-type repressor.

Another tighter binding mutant was located in the dimerization portion of the protein. Its binding properties are consistent with the mutation altering the relative positions of the two DNA-contacting domains to correspond more closely to the DNA-binding site. The third mutation may increase the number of specific interactions between repressor and either a base or the phosphodiester backbone of the DNA.

How much does the recognition helix contribute to sequence recognition? The rest of the protein could provide a precisely shaped platform against which the recognition helix nestles and, in conjunction with this platform, the actual DNA-contacting amino acids would be precisely held in the correct positions. Alternatively, the recognition helix, or even the DNA-contacting amino acids themselves, could entirely determine the binding selectivity of the helix-turn-helix proteins. Experiments have revealed that the contacting residues can provide a surprisingly large amount of selectivity. First, the phage 434 repressor protein can be given part of the binding specificity of the 434 Cro protein and the Cro DNA-contacting pattern by replacing the entire recognition α-helix of the repressor with the corresponding Cro α-helix (*24*). Going a bit further, the 434 repressor could be given the DNA-binding specificity of the related *Salmonella* phage P22 repressor by substituting the solvent-exposed amino acid residues of 434 repressor with the corresponding residues of P22 repressor (*25*). Finally, the λ repressor could be given a significant amount of the Cro protein specificity by changing just one amino acid of the repressor that contacts a base (*26*). These phage proteins are closely related and we have not seen all combinations of these types of exchange, but it is impressive that, in at least some cases, significant selectivity is provided by just the DNA-contacting amino acids.

High-Resolution Probing of Amino Acid–Base Interactions

Remarkably, at this time, the resolution of biochemical and genetic approaches for determining specific amino acid–base interactions equals or exceeds that of x-ray crystallography. One use of this knowledge was the study on the specificity of λ repressor and Cro, described above. These techniques are not limited to proteins that can be crystallized with their DNA-binding sites.

If a smaller amino acid is substituted for a larger residue that previously contacted the DNA and if the protein and the DNA are not too flexible, the contacts provided by this one residue are lost. Such a change, of course, weakens the binding. If the contact were to the base and not to the phosphate or the deoxyribose, then a subsequent change in this base will not affect the binding affinity of the altered protein. In other words, the mutant protein with the smaller amino acid residue has become indifferent to the identity of the base that was previously contacted by the unaltered residue (*27*).

The simple concept of making a dent in the protein so it no longer sees a specific nucleotide is somewhat arduous to carry out. The gene for the protein must be altered, the wild-type and the mutant proteins purified (or partially purified), the binding site altered, and then the binding affinity of the wild-type and the mutant proteins to the wild-type and the mutant DNA's must be measured. An additional complication is that since many DNA-binding proteins contact DNA in a symmetric manner with two identical subunits, two nucleotides in the DNA must be simultaneously altered in these experiments.

The missing contact approach has been used to determine precise amino acid residue–base interactions made by λ repressor, Cro protein, and CRP (*27, 28*). First, the missing contact approach revealed that homologous amino acids in λ repressor and Cro contact the same bases, which indicates that the two proteins bind to DNA similarly and, hence, that their recognition helices are oriented similarly. Second, the two proteins' abilities to distinguish DNA sequences (as measured by relative binding affinities to different sequences) result from the ability of each protein to read bases unique to each class of sequence. In the case of CRP, substituting Glu^{181} by Val or Leu made the protein become independent of the identity of the base at positions 7 and 16 of the protein's binding site, thereby proving the Glu^{181}–position 7 interaction.

Recently a modification of the missing contact approach was described that has the potential for increasing the efficiency of identifying contacts (*29*). In this modification, the variant proteins are generated just as before. Then, the contacts made by the wild-type and the variant proteins are biochemically probed to locate the nucleotide or nucleotides that are contacted only by the wild-type protein (Fig. 5). This modified scheme can eliminate the necessity for correctly guessing the base or bases that are contacted.

One concern of the streamlined missing contact method is that the DNA missing a base would be significantly altered in structure or that the binding protein would contact the DNA differently because of the missing contact. At this point we do not know how frequently this will be a problem. However, the method correctly revealed the Ser^{45}-G^4 contact of λ repressor that was predicted from the known structure and biochemical probing experiments and confirmed by the more arduous missing contact method.

A variation of the missing contact method is to find variant proteins or DNA-binding sites that compensate for changes in the binding partner. By this means an interaction between Gln and A in position 1 of phage 434 operator was demonstrated. Upon changing the Gln to Ala, only a T in position 1 could be recognized by the repressor, presumably by means of a van der Waals interaction (*30*). Similarly, substituting Glu^{181} in CRP by Val or Leu made this protein become independent of the identity of the base at positions 7 and 16 of the protein's binding site, thereby proving the Glu^7 interaction (*28*).

NMR Studies on DNA-Binding Proteins

Nuclear magnetic resonance measurements (NMR) have provided two types of data in support of the helix-turn-helix mode of DNA binding. First, specific amino acids of the helix-turn-helix structure ought to contact the binding site, and both nuclear polarization and perturbations to the ^{19}F-NMR spectra of 3-fluorotyrosine–substituted *lac* repressor headpiece show the expected interactions (*31*). Secondly, a large collection of inter-proton distance measurements can be made by means of two-dimensional NMR. Triangulation with these distances can be used to derive both the secondary and tertiary structures of the protein. For the *lac* repressor headpiece, the NMR-derived secondary and tertiary structures of the helix-turn-helix region yielded the same structure that the homologous amino acids adopt in crystals of the above-mentioned proteins (*32*). As the power of NMR increases, we can expect more and larger protein structures to be solved without the need for crystallization. We should also see details of the interaction between proteins and DNA begin to be elucidated by NMR.

Structure of the Eco RI–DNA Complex

The structure of the restriction enzyme Eco RI complexed with a DNA oligonucleotide containing its DNA-binding and cleavage sequence (GAATTC) has recently been determined to 3 Å resolution (*33*). The basic structure of the sequence-recognition elements in Eco RI is different from that in the helix-turn-helix proteins and the DNA contacted is not the typical B form structure. As described above, complete identification of a base by hydrogen bonding requires formation of at least two hydrogen bonds per base pair. Since Eco RI uniquely recognizes its hexanucleotide binding site, it must make at least two such bonds with each of these base pairs or substitute van der Waals interactions for hydrogen bonds.

The electron density map indicates that the expected 12 hydrogen bonds are made. Instead of being made by α-helices lying within and nearly parallel to the major groove of the DNA, the Eco RI contacts are from the NH_2-terminal ends of four α-helices stuck end-on into the major groove of the DNA. These four helices cannot be accommodated by the normal width of the major groove. Binding of the protein separates the grooves by rolling the strands away from one another, thereby generating kinks in the phosphodiester backbone. This rolling appears to be generated by pockets on the enzyme that are lined with positively charged amino acids. These amino acids interact with the phosphate backbone of the DNA and pull the backbones of the two strands away from one another.

The contacts from each subunit of the dimeric Eco RI enzyme are made by one Gln residue and two Arg residues. In addition to these and the clefts for the phosphates of the DNA, each subunit also possesses an arm that reaches around the DNA and interacts with the backbone in a sequence-independent manner that strengthens binding without changing binding specificity (*34*). These arms possess a well-defined secondary structure and a well-defined tertiary structure. The λ repressor (but not Cro from λ or 434 phage, CRP, Trp repressor, or the 434 phage repressor) also possesses arms that reach around the DNA. The arms on λ repressor possess no elements of specific secondary structure, but they make a specific contact to the DNA from the back side (*35*). These encircling arms must get out of the way as the proteins bind to DNA.

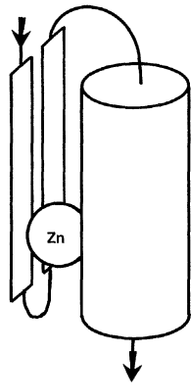

Fig. 6. Schematic structure of the structure hypothesized for a zinc finger (*44*).

Zinc Finger Proteins

The zinc finger is another major motif for a DNA-contacting domain of proteins. This structure was first identified in the *Xenopus* 5*S* RNA transcription factor TFIIIA (*36*). The presence of Zn^{2+} in TFIIIA was first suggested by the requirement for the presence of Zn^{2+} during the protein's purification. Proteolytic digestion showed that the protein contained about nine similar-sized substructures, and atomic absorption spectroscopy confirmed the presence of about ten Zn^{2+} in the protein. Analysis of the amino acid sequence of TFIIIA derived from the mRNA sequence showed that the protein contains nine near repeats of the unusual sequence Cys-x_{2-4}-Cys-x_{12}-His-x_3-His. Thus, it seemed likely that the Cys and His residues were coordinating the Zn^{2+}, and indeed, extended x-ray absorption fine structure (EXAFS) analysis showed that the zinc ions in TFIIIA were each coordinated with two Cys and two His residues (*37*).

In addition to TFIIIA, a number of other proteins known to bind DNA, or suspected of binding to DNA, also contain zinc finger sequences of doublets of Cys separated by about 12 amino acids from doublets of His (*38*). In some of these, a second Cys doublet substitutes for the His doublet. These include proteins identified as transcription factors, proteins known to be involved in development, and virus proteins. Additionally, indications have been found

of a "finger domain" that may be analogous to the homeo domain, for there appears to be a conserved family of nuclear proteins that contain sequences related to the zinc finger sequence of a *Drosophila* developmental gene (*39*).

We now have excellent data that the zinc finger domains do contact DNA. The most dramatic demonstration was the replacement of a 66–amino acid portion of estrogen receptor protein containing at least two zinc finger domains by the corresponding portion of human glucocorticoid receptor protein (*40*). The resulting hybrid protein showed the DNA-binding specificity of the glucocorticoid receptor protein. Additional evidence for DNA contacts by the zinc finger domain is the fact that the majority of the defective mutations in the *GAL4* transcription-activating protein in yeast were found in the region of its two zinc fingers (*41*). Also, disrupting the zinc finger region of the human glucocorticoid receptor protein left it nonfunctional (*42*), and mutations in the zinc finger region of a yeast regulatory protein left it inactive unless elevated amounts of Zn^{2+} were added to the growth medium (*43*).

Thus far no zinc finger proteins have been crystallized, and so we have no definitive structure analogous to the helix-turn-helix. However, a number of metalloenzymes have been found to possess the Cys or His doublets, and structure of the zinc-containing regions in these proteins provides a reasonable first guess as to the structure of the zinc fingers (*44*). Instead of a loop, the domain may consist of a helix and two antiparallel beta sheets stabilized by the zinc ion (Fig. 6).

Additional DNA-Binding Structures

A proteolytic fragment of the *E. coli* DNA polymerase I containing the polymerizing and 3′-5′ exonuclease activities has been crystallized and its structure determined (*45*). The structure suggests that the protein wraps around the DNA and is kept from sliding along the DNA by a finger of two α-helices that stick into the major groove of the DNA. The structure of these helices does not closely resemble the helix-turn-helix structures. HU protein, a small, basic histone-like protein found in bacteria, binds to DNA without appreciable sequence dependence and may be comparable in function to histones in higher cells. The structure of HU from *Stearothermophilus* determined by x-ray diffraction suggests that this protein contacts the phosphates and sugars of DNA with two antiparallel beta sheets, but the structure does not favor the major or minor groove for these interactions (*46*).

Two small regulatory proteins of the *Salmonella* phage P22, the Arc and Mnt proteins, bind to DNA but seem not to possess either the zinc finger or the helix-turn-helix motifs (*47*). Because these are small proteins and the biological system is amenable to genetic studies, these ought to be fruitful objects for further study.

REFERENCES AND NOTES

1. N. Seeman, J. Rosenberg, A. Rich, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 804 (1976).
2. C. Woodbury, Jr., O. Hagenbüchle, P. von Hippel, *J. Biol. Chem.* **255**, 11534 (1980).
3. H. Wing *et al.*, *Nature* **287**, 755 (1980); A. Fratini, M. Kopka, H. Drew, R. Dickerson, *J. Biol. Chem.* **257**, 14686 (1982); H. Nelson, J. Finch, B. Luisi, A. Klug, *Nature* **330**, 221 (1987); W. Hunter *et al.*, *J. Biol. Chem.* **262**, 9962 (1987); M. McCall, T. Brown, W. N. Hunter, O. Kennard, *Nature* **322**, 661 (1986).
4. D. Ryder, S. Silver, A. DeLucia, E. Fanning, P. Tegtmeyer, *Cell* **44**, 719 (1986).
5. S. C. Satchwell, H. R. Drew, A. A. Travers, *J. Mol. Biol.* **191**, 659 (1986).
6. M. T. Record *et al.*, *Annu. Rev. Biochem.* **50**, 997 (1981).
7. R. B. Winter and P. H. von Hippel, *Biochemistry* **20**, 6948 (1981); M. C. Mossing and M. T. Record, Jr., *J. Mol. Biol.* **186**, 295 (1985).
8. R. B. Winter and P. H. von Hippel, *Biochemistry* **20**, 6948 (1981).
9. B. Richey *et al.*, *J. Biol. Chem.* **262**, 7157 (1987).
10. K. J. Martin and R. F. Schleif, *J. Mol. Biol.* **195**, 741 (1987).
11. D. B. McKay and T. A. Steitz, *Nature* **290**, 744 (1981); W. F. Anderson, D. H. Ohlendorf, Y. Takeda, B. W. Matthews, *ibid.*, p. 754; C. O. Pabo and M. Lewis, *ibid.* **298**, 443 (1982).
12. R. Schevitz, Z. Otwinowski, A. Joachimiak, C. L. Lawson, P. B. Sigler, *ibid.* **317**, 782 (1985); R. Zhang *et al.*, *ibid.* **327**, 591 (1987).
13. J. E. Anderson, M. Ptashne, S. C. Harrison, *ibid.* **326**, 849 (1987).
14. ———, *ibid.* **316**, 596 (1985).
15. G. Koudelka, S. Harrison, M. Ptashne, *ibid.* **326**, 886 (1987).
16. D. H. Ohlendorf, W. F. Anderson, M. Lewis, C. O. Pabo, B. W. Matthews, *J. Mol. Biol.* **169**, 757 (1983).
17. B. W. Matthews, D. H. Ohlendorf, W. F. Anderson, Y. Takeda, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1428 (1982); R. T. Sauer, R. R. Yocum, R. F. Doolittle, M. Lewis, C. O. Pabo, *Nature* **298**, 447 (1982); I. T. Weber, D. B. McKay, T. A. Steitz, *Nucleic Acids Res.* **10**, 5085 (1982).
18. C. O. Pabo and R. T. Sauer, *Annu. Rev. Biochem.* **53**, 293 (1984).
19. W. J. Gehring and Y. Hiromi, *Annu. Rev. Genet.* **20**, 147 (1986).
20. A. Laughon and M. Scott, *Nature* **310**, 25 (1984); M. N. Hall and A. D. Johnson, *Science* **237**, 1007 (1987); A. Fainsod *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9532 (1986).
21. M. H. Hecht, H. C. Nelson, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 2676 (1983).
22. P. Isackson and K. Bertrand, *ibid.* **82**, 6226 (1985); K. Adler *et al.*, *Nature* **237**, 322 (1972); M. Pfahl, C. Stockter, B. Gronenborn, *Genetics* **7**, 669 (1974).
23. H. Nelson and R. Sauer, *Cell* **42**, 529 (1985).
24. R. P. Wharton, E. L. Brown, M. Ptashne, *ibid.* **38**, 361 (1984).
25. R. P. Wharton and M. Ptashne, *Nature* **316**, 601 (1985).
26. A. Hochschild, J. Douhan III, M. Ptashne, *Cell* **47**, 807 (1986).
27. A. Hochschild and M. Ptashne, *ibid.* **44**, 925 (1986).
28. R. Ebright *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6083 (1987).
29. A. Brunelle and R. F. Schleif, *ibid.*, p. 6673.
30. R. Warton and M. Ptashne, *Nature* **326**, 888 (1987).
31. K. Arndt, H. Nick, F. Boschelli, P. Lu, J. Sadler, *J. Mol. Biol.* **161**, 439 (1982); F. Buck, H. Ruterjans, R. Kaptein, K. Beyreuther, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5145 (1980).
32. R. Kaptein, E. Zuiderweg, R. Scheck, R. Boelens, W. van Gunsteren, *J. Mol. Biol.* **182**, 179 (1985).
33. J. A. McClarin *et al.*, *Science* **234**, 1526 (1986); C. A. Frederick *et al.*, *Nature* **309**, 327 (1984).
34. L. Jen-Jacobson, D. Lesser, M. Kurpiewski, *Cell* **45**, 619 (1986).
35. J. Eliason, M. Weiss, M. Ptashne, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2339 (1985); C. O. Pabo, R. T. Sauer, J. M. Sturtevant, M. Ptashne, *ibid.* **76**, 1608 (1979).
36. J. Miller, A. McLachlan, A. Klug, *EMBO J.* **4**, 1609 (1985).
37. G. Diakun, L. Fairall, A. Klug, *Nature* **324**, 698 (1986).
38. J. M. Berg, *Science* **232**, 485 (1986); R. Evans and S. Hollenberg, *Cell* **52**, 1 (1988); P. Gauss, K. B. Krassa, D. S. McPheeters, M. A. Nelson, L. Gold, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8515 (1987); J. Milbrandt, *Science* **238**, 797 (1987); H. Blumberg, A. Sledziewski, D. Bader, E. Young, *Nature* **328**, 443 (1987).
39. R. Schuh *et al.*, *Cell* **47**, 1025 (1986).
40. S. Green and P. Chambon, *Nature* **325**, 75 (1987).
41. M. Johnston and J. Dover, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2401 (1987).
42. S. Hollenberg, V. Giguere, P. Segui, R. Evans, *Cell* **49**, 39 (1987).
43. H. Blumberg, A. Eisen, A. Sledziewski, D. Bader, E. Young, *Nature* **328**, 443 (1987).
44. J. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
45. D. L. Ollis, P. Brick, R. Hamlin, N. G. Xuong, T. A. Steitz, *Nature* **313**, 762 (1985).
46. I. Tanaka, K. Appelt, J. Dijk, S. W. White, K. S. Wilson, *ibid.* **310**, 376 (1984).
47. A. K. Vershon, S. M. Liao, W. R. McClure, R. T. Sauer, *J. Mol. Biol.* **195**, 323 (1987); *ibid.*, p. 311; R. Sauer, W. Krovatin, J. DeAnda, P. Youderian, M. Susskind, *ibid.* **168**, 699 (1983); A. K. Vershon, P. Youderian, M. M. Susskind, R. T. Sauer, *J. Biol. Chem.* **260**, 12124 (1985).

48. Supported by NIH and NSF grants. I thank K. Martin, P. Wensink, A. Brunelle, J. Tobin, R. Lobell, L. Huo, J. Carra, D. H. Lee, and G. Duffaud for comments on the manuscript.