Zinc Finger Domains: From Predictions to Design

JEREMY M. BERG

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Received August 8, 1994

The prediction of the three-dimensional structures of proteins from amino acid sequence information represents one of the most challenging and important problems in biological chemistry. Because of developments in gene cloning and nucleic acid sequencing, protein sequence is now often one of the first pieces of information available about a biochemical system under characterization. However, since protein function is much more dependent on and related to tertiary structure than primary structure, this information is of limited utility. Nonetheless, the protein sequence does, in principle, contain the necessary information to dictate the three-dimensional structure. For most systems, knowledge of the three-dimensional structure must await experimental investigation via X-ray crystallographic or nuclear magnetic resonance methods. However, for one class of DNA-binding proteins, termed "zinc finger" proteins, predictions of structure based on analysis of amino acid sequence data have been made and subsequently proven to be essentially correct. It is the goal of this Account to illustrate the basis of these predictions and to describe the current state of knowledge concerning this large and important class of gene regulatory proteins. A large number of reviews describing other aspects of this rapidly growing field have appeared.[1−9]

Transcription Factor IIIA and the Zinc Finger Proteins

The first member of the zinc finger family to be characterized was transcription factor IIIA (TFIIIA) from the African clawed toad *Xenopus*. This protein is involved in control of expression of the 5S ribosomal RNA genes and binds both to specific DNA sequences within 5S RNA gene and to the 5S RNA molecule itself. In 1983, it was reported that TFIIIA contained bound zinc and that this zinc was required for the specific interaction between TFIIIA and DNA.[10] In 1984, the amino acid sequence was deduced from the sequence of a cDNA clone.[11] Subsequent analysis of this sequence of TFIIIA revealed that it is quasiperiodic.[12,13] It contains nine imperfectly repeated units of approximately 30 amino acids each. Miller et al. demonstrated that the TFIIIA−5S RNA complexes isolated from *Xenopus* oocytes contained 7−11 zinc ions per protein when purified under conditions that avoided potential chelators.[12] They went on to suggest a relationship between the nine repeated units in the amino acid sequence and the bound zinc. Each of the sequence repeats approximates the form (Tyr,Phe)-X-Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-$X_{3,4}$-His-X_{2-6} where X represents relatively non-conserved amino acids. They proposed that each of these sequences binds a zinc ion through the invariant cysteine and histidine residues to form a relatively independent structural domain that was referred to as a "zinc finger". The proposed "low-resolution" structure of TFIIIA is illustrated in Figure 1. This hypothesis explained the presence of the cysteine and histidine residues in the sequence; the three invariant hydrophobic residues were proposed to play some, at this point, undefined role in stabilizing the structure. In addition, the hypothesis provided appealing explanations for two additional observations.[12] First, TFIIIA protects approximately 45 base pairs of DNA from enzymatic digestion; the presence of tandem, small domains rather than one large globular domain potentially accounted for this elongated structure. Second, limited proteolysis studies of the TFIIIA−5S RNA complex revealed the presence of metastable fragments with molecular weights that were multiples of 3 kilodaltons (kDa) the size expected for sets of zinc finger domains.

Shortly after the analysis of the TFIIIA sequence, similar motifs were observed in the deduced sequences from a number of other proteins. These included the yeast transcription factor ADR1[14] (which contains two such sequences) and the *Drosophila* proteins serendipity β and δ,[15] and Krüppel[16] (with 6, 7, and 5 sequences, respectively). It was expected (and subsequently experimentally demonstrated) that each of these proteins is a sequence-specific DNA-binding protein. Since that time and continuing to the present, the number of known sequences (more than 1300 from

Jeremy M. Berg was born and raised in Stanford, CA. He received his B.S. and M.S. degrees in chemistry from Stanford University in 1980. He obtained his Ph.D. in chemistry from Harvard Chemistry in 1985 working with R. H. Holm. He then did a postdoctoral fellowship with Carl Pabo in the Department of Biophysics at The Johns Hopkins University School of Medicine. He moved to the Department of Chemistry at Johns Hopkins in 1986 as assistant professor. In 1980, he returned to the School of Medicine as professor and director of the Department of Biophysics and Biophysical Chemistry and director of the Markey Center for Macromolecular Structure and Function. His research interests are centered around the roles of metal ions in biological systems with particular emphasis on structural ions such as zinc and calcium. He has recently published a textbook entitled *Principles of Bioinorganic Chemistry* with Professor S. J. Lippard.

(1) Klug, A.; Rhodes, D. *Trends Biochem. Sci.* **1987**, *12*, 464.
(2) Berg, J. M. *J. Biol. Chem.* **1990**, *265*, 6513.
(3) Kaptein, R. *Curr. Opin. Struct. Biol.* **1991**, *1*, 63.
(4) Kaptein, R. *Curr. Opin. Struct. Biol.* **1992**, *2*, 109.
(5) Berg, J. M. *Curr. Opin. Struct. Biol.* **1993**, *3*, 11.
(6) Schmiedeskamp, M.; Klevit, R. E. *Curr. Opin. Struct. Biol.* **1994**, *4*, 28.
(7) Vallee, B. L.; Auld, D. S. *Acc. Chem. Res.* **1993**, *26*, 543.
(8) Rhodes, D.; Klug, A. *Sci. Am.* **1993**, 56.
(9) Schwabe, J. W. R.; Klug, A. *Nat. Struct. Biol.* **1994**, *1*, 345.
(10) Hanas, J. S.; Hazuda, D. J.; Bogenhagen, D. F.; Wu, F. H.-Y.; Wu, C.-W. *J. Biol. Chem.* **1983**, *258*, 14120.
(11) Ginsberg, A. M.; King, B. O.; Roeder, R. G. *Cell* **1984**, *39*, 479.
(12) Miller, J.; McLachlan, A. D.; Klug, A. *EMBO J.* **1985**, *4*, 1609.
(13) Brown, R. S.; Sander, C.; Argos, P. *FEBS Lett.* **1985**, *186*, 271.
(14) Hartshorne, T. A.; Blumberg, H.; Young, E. T. *Nature* **1986**, *320*, 283.
(15) Vincent, A.; Colot, H. V.; Rosbash, M. *J. Mol. Biol.* **1985**, *186*, 149.
(16) Rosenberg, U. B.; Schröder, C.; Preiss, A.; Kienlin, A.; Côte, S.; Jäckle, H. *Nature* **1986**, *319*, 336.

0001-4842/95/0128-0014$09.00/0

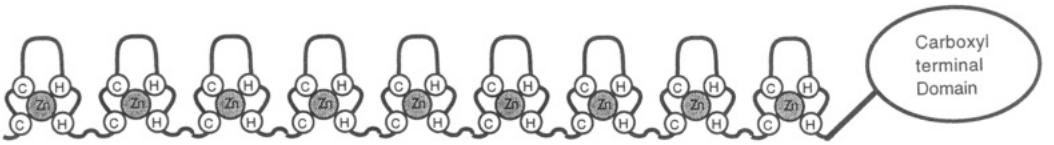

Figure 1. The "zinc finger" hypothesis. A schematic view of the quasiperiodic nature of transcription factor IIIA showing the nine proposed zinc-binding domains.[12] The carboxyl terminal domain does not appear to bind zinc.

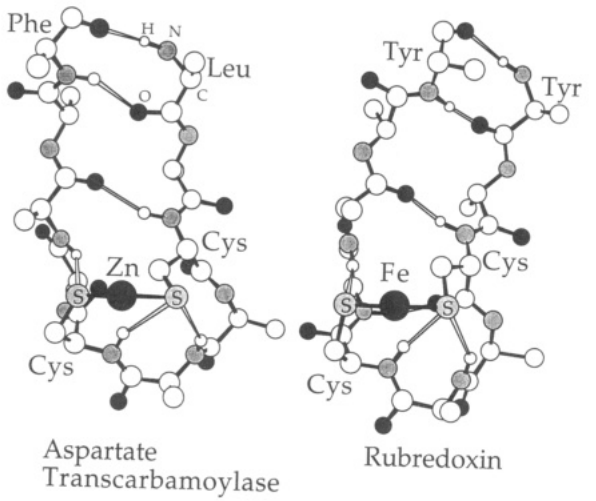

Figure 2. The structures of metal-binding X_2-Cys-X_2-Cys-X_4 sequences from the regulatory domain of aspartate transcarbamoylase and rubredoxin. The sequences adopt β hairpin structures with peptide amide to cysteinate sulfur hydrogen bonds.

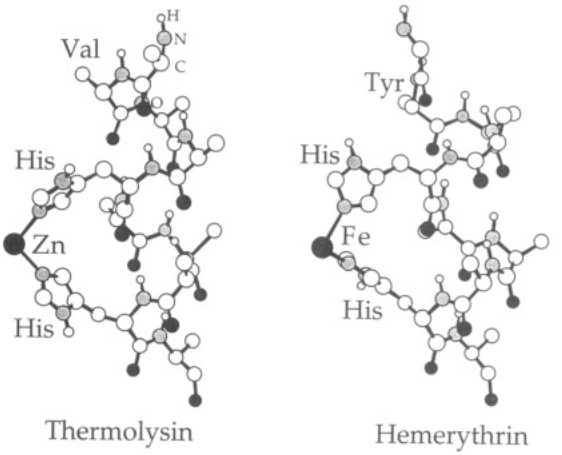

Figure 3. The structures of metal-binding X_3-His-X_3-His-X sequences from thermolysin and hemerythrin. The sequences adopt α helical structures with the His residues coordinated through their ϵ-nitrogen atoms.

over 200 different proteins[17]) has continued to explode. These sequences closely match the consensus that was deduced from TFIIIA.

Structure Predictions and Determination

Using a data base of 37 sequences, Brown and Argos proposed that part of the carboxyl terminal portion of each of these sequences would form an α helix, based on secondary structure prediction algorithms.[18] At this point, I arrived at a much more detailed structure prediction for these domains by examining the structures of short metal-chelating sequences in the Brookhaven Data Base of known protein structures.[19] Proteins that contained metal-chelating sequences of the Cys-X_{2-4}-Cys and His-X_3-His were examined. Two recurring substructures were found. The structures adopted by the sequences **Leu**Lys**Cys**LysTry**Cys**-GluLysGlu**Phe** from *Escherichia coli* aspartate transcarbamoylase and **Tyr**Thr**Cys**ThrVal**Cys**GlyTyrIle-**Tyr** from *Clostridium pasteurianum* rubredoxin are shown in Figure 2. Each consists of an antiparallel β hairpin with the two metal-binding cysteinate residues at the base. Remarkably, these sequences also have hydrophobic residues that match the positions of two of the hydrophobic residues from the zinc finger sequences. These hydrophobic residues lie on the same face of the β sheet as the metal−cysteinate unit. Similarly, structures for the sequences **Val**ValAla-**His**GluLeuThr**His** from thermolysin and **Tyr**AlaGlu-**His**LysLysAla**His** from *Thermiste dyscritum* hemerythrin are shown in Figure 3. Here, each substructure consists of an α helix with the two histidines coordinated to the metal ion via their ϵ-nitrogens. Again, each sequence has a hydrophobic residue in the position corresponding to the recurring leucine residue in the zinc finger sequences, and each of these hydrophobic residues lies on the same face of the helix as the metal coordination unit.

These two substructures may be combined to yield potential models for a complete zinc finger domain. Because the metal-binding site was believed (and has subsequently been shown) to be tetrahedral, there are only two fundamentally different ways of joining the two substructures. One of these is incompatible with the five amino acids in the "finger tip" available to connect the two sequences. The other places the carboxyl terminus of the X_2-Cys-X_2-Cys-X_4 unit within 7.5 Å of the amino terminus of the X_3-His-X_3-His substructure. In addition, this arrangement packs the hydrophobic residues against one another and against the hydrophobic portions of the metal coordination unit. The structure can be completed with the five amino acid connecting loop. The construction of the predicted structure is summarized in Figure 4. A similar structure was independently proposed which differed in pairing across the β sheet and in the length of the helix.[20]

The first experimental structural information about these domains came from NMR studies of single-domain peptides. It had been previously demonstrated by other methods that a 30 amino peptide corresponding to one of the domains from TFIIIA was unfolded in the absence of metal but folded into a stable structure upon addition of stoichiometric amounts of zinc(II) or cobalt(II).[21] The first NMR data reported for a zinc finger peptide involved a similar

(17) Jacobs, G. H. *EMBO J.* **1992**, *11*, 4507.
(18) Brown, R. S.; Argos, P. *Nature* **1986**, *324*, 215.
(19) Berg, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 99.
(20) Gibson, T. J.; Postma, J. P. M.; Brown, R. S.; Argos, P. *Protein Eng.* **1988**, *2*, 209.
(21) Frankel, A. D.; Berg, J. M.; Pabo, C. O. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 4841.

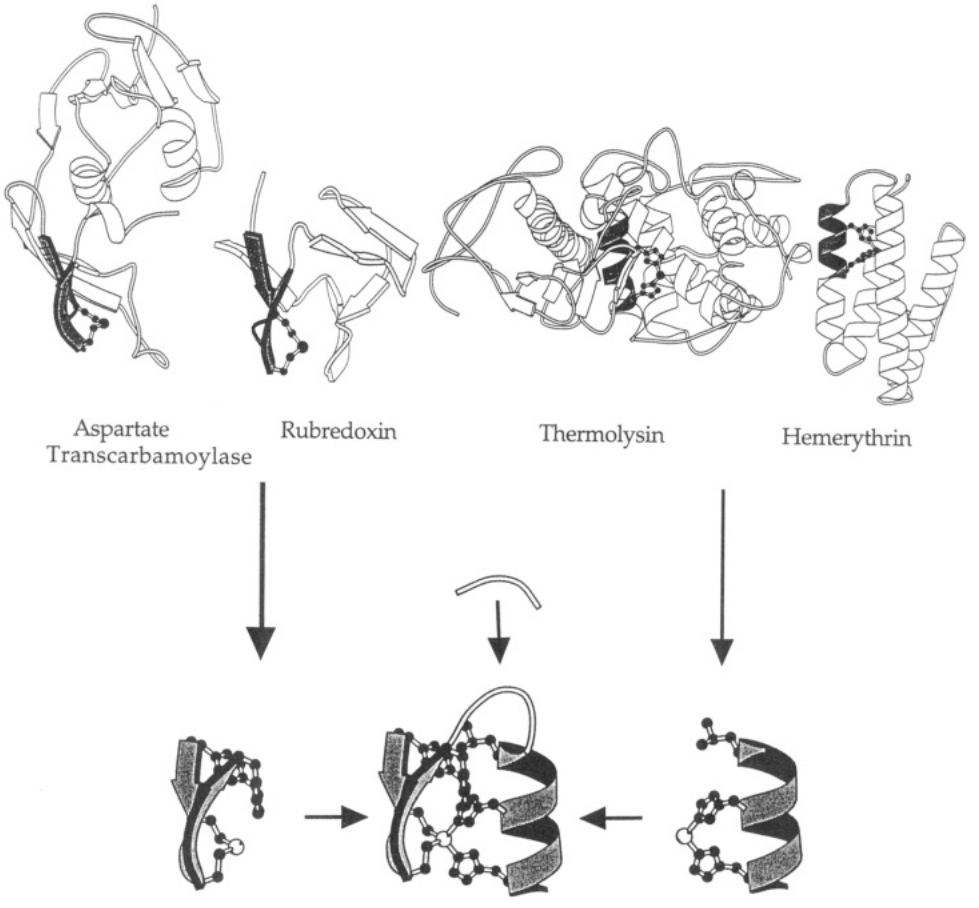

Figure 4. A summary of the generation of the predicted structure for the zinc finger domains.[19] The recurring metal-chelating substructures from the four proteins are shown and combined to yield the predicted zinc finger domain structure. This structure generated through the use of MOLSCRIPT.[44]

peptide from the yeast transcription factor ADR1.[22] A preliminary structure was reported that demonstrated the presence of a helix that overlapped the region predicted as well as the hydrophobic core involving at least the conserved Phe and Leu residues. A more complete structure was reported for a peptide based on the 31st of 37 domains from the *Xenopus* protein Xfin.[23] This structure determination indicated the presence of the β sheet in the predicted position as well as confirming the chirality at the metal center and the presence of the helix. The major difference between the predicted structure and the experimentally derived one involved the region joining the β sheet and the helix. This was modeled as a distorted β turn in the model whereas, in the NMR structure, the helix extended further to connect to the top of the sheet. The predicted and experimental structures are overlaid in Figure 5.

This structure appears to occur with some minor variations for all zinc finger domains of this type. NMR structures have been reported for a number of single domains and several two-domain peptides from several different proteins. In addition, two designed peptides have been characterized. The first of these is a consensus sequence, termed CP-1, which was designed by choosing the amino acid that occurred most frequently at each position in a data base of over 100 zinc finger sequences.[24] This peptide was found to form an unusually stable structure based on metal-binding properties and pH stability. The structure of the zinc complex of CP-1 has been determined by NMR methods revealing a structure very similar to those of the natural zinc finger peptide complexes.[25] The second is a "minimalist" zinc finger peptide with the sequences LysTyrAla**Cys**AlaAla**Cys**AlaAlaAlaPhe-AlaAlaLysAlaAlaLeuAlaAla**His**-AlaAlaAla**His**Ala-Lys, containing only the metal-binding and conserved hydrophobic residues as well as three lysines for aqueous solubility with all other residues replaced with alanine.[26] On the basis of the observation of the expected cross peaks in NOESY spectra as well as strong similarities in chemical shifts with other zinc finger peptides, it is clear that this peptide also adopts the canonical structure.

Zinc Finger Protein–DNA Complex Structure

Even before any experimental information on the structure of a zinc finger protein–DNA complex had been obtained, it was clear that the structure of a single domain had implications with regard to how a tandem array of such units would interact with DNA.[19] Most importantly, the amino and carboxyl termini of the domain are quite far apart, essentially as far apart as they can be, given the size of the

(22) Parrága, G.; Horvath, S. J.; Eisen, A.; Taylor, W. E.; Hood, L.; Young, E. T.; Klevit, R. *Science* **1988**, *241*, 1489.
(23) Lee, M.; Gippert, G. P.; Soman, K.; Case, D. A.; Wright, P. E. *Science* **1989**, *245*, 635.
(24) Krizek, B. A.; Amann, B. T.; Kilfoil, V. J.; Merkle, D. L.; Berg, J. M. *J. Am. Chem. Soc.* **1991**, *113*, 4518.
(25) Kilfoil, V. J. Ph.D. Thesis, The Johns Hopkins University, 1992.
(26) Michael, S. F.; Kilfoil, V. J.; Schmidt, M. H.; Amann, B. T.; Berg, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4796.

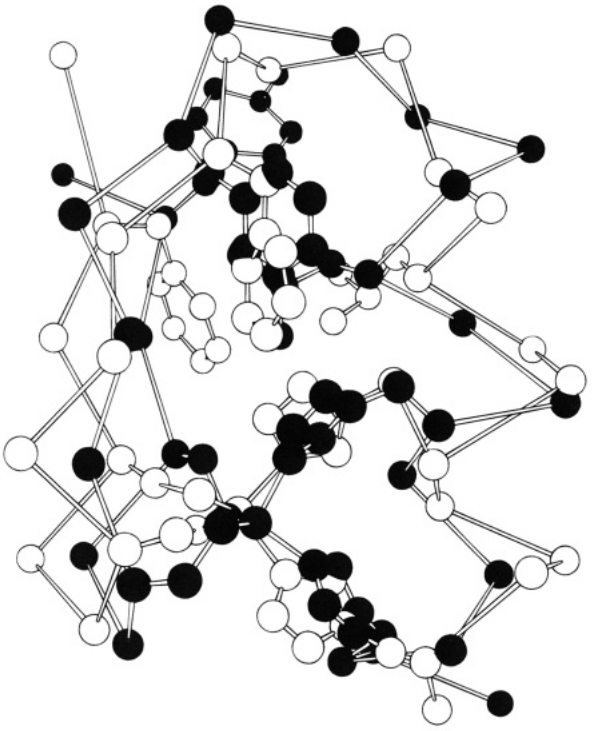

Figure 5. A comparison of the predicted and experimentally determined zinc finger domain structures. The α carbon and conserved metal-binding and hydrophobic residues for the predicted structure are shown in white, and those for the structure of a zinc finger domain determined by NMR methods[25] are shown in black. The root mean square deviation for the atoms shown is less than 2 Å.

structure. This property allows the straightforward construction of models for complexes where a multiple domain protein wraps around the DNA and makes it very difficult to construct models in which the protein lies along one face of the DNA with the "fingers" sticking up. A variety of chemical data indicated that these proteins interact primarily with the major groove of the DNA.

The models proposed for the protein wrapping around the DNA were confirmed and extended by the determination of the structure of the complex between the three zinc finger DNA-binding domain from the mouse protein Zif268 and an oligonucleotide containing its binding site.[27] The structure of this complex is shown in Figure 6. In addition to demonstrating the gross features of a zinc finger protein–DNA complex, the Zif268 complex structure provided a great deal of information about the interactions between the protein and the DNA that led to specific sequence recognition. Each zinc finger domain was seen to contact three base pairs of DNA. Furthermore, the specific contacts were made essentially to one strand of the DNA. The more heavily contacted strand of the DNA lies "antiparallel" to the protein so that, if the protein is aligned from its amino terminus to its carboxyl terminus, the DNA sequence of this strand runs 3′ to 5′. The sequence specific contacts involved residues in positions 13, 16, and 19 within the sequence X-(Thr,Phe)-X-Cys-X_2-Cys-X_3-Phe-X-X^{13}-X-X-X^{16}-Leu-X-X^{19}-His-X_3-His. The first and third domains have X^{13} = Arg, X^{16} = Glu, X^{19} = Arg, and contact 5′-GCG-3′ whereas the second domain has X^{13} = Arg, X^{16} = His, X^{19} = Thr, and contacts 5′-TGG-3′. Five of the

(27) Pavletich, N. P.; Pabo, C. O. *Science* **1991**, *252*, 809.

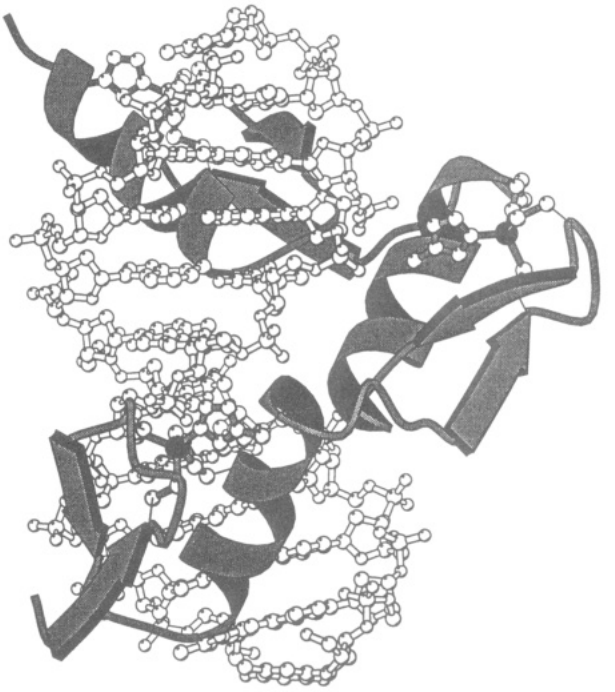

Figure 6. The structure of the complex between the three tandem zinc finger domains from Zif268 and its DNA-binding site determined by X-ray crystallography.[27]

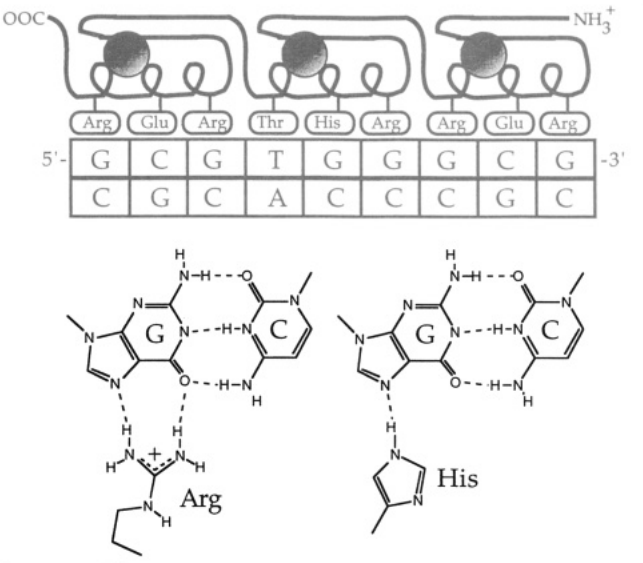

Figure 7. The interactions between Zif268 and its binding site. The positions of the three DNA-contacting residues are shown schematically only. The hydrogen-bonding interactions between Arg and guanine and between His and guanine responsible for sequence recognition are shown at the bottom of the figure.[27]

six direct interactions involved double hydrogen bond interactions between the Arg residues and guanine bases. The other involves a His to guanine hydrogen bond. These interactions are summarized in Figure 7.

A number of other interactions between amino acids in positions 13, 16, and 19 have been deduced from mutagenesis experiments.[28–33] Some of these include interactions of Gln in position 13 with thymine or

(28) Nardelli, J.; Gibson, T. J.; Vesque, C.; Charnay, P. *Nature* **1991**, *349*, 175.

(29) Nardelli, J.; Gibson, T.; Charnay, P. *Nucleic Acids Res.* **1992**, *20*, 4137.

(30) Desjarlais, J. R.; Berg, J. M. *Proteins: Struct., Funct., Gen.* **1992**, *12*, 101.

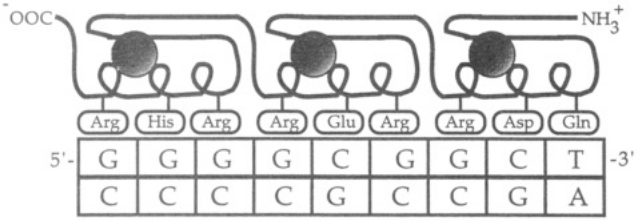

Figure 8. The alignment of the designed zinc finger protein on its binding site.[34]

adenine, of Asp and Asn in position 16 with cytosine or adenine, respectively, and of Lys in position 19 with guanine. The roles of the residues in these positions are not independent. For example, with Arg in position 13, His is quite effective in binding to guanine or adenine. With Gln in position 13, however, no strong binding to any sequence was observed.[32] Most of these observations can be rationalized in terms of side chain length. With a long Arg residue in position 13, relatively large residues such as Glu and His are effective, while with a shorter Gln in position 13, shorter residues such as Asp and Asn are more functional. Through these studies, a growing collection of sets of contact residues and their recognition sites is being developed.

Design of Sequence Specific DNA-Binding Zinc Finger Proteins

This collection is meaningful only if zinc finger domains and their contact residues can be transferred from one protein context to another with retention of their DNA recognition properties. This has been partially tested via a project in protein design.[34] A protein was designed that consisted of three zinc finger domains based on the consensus sequence of CP-1 but with the recognition residues modified to sets whose DNA recognition properties had been characterized in other systems. The first domain had X^{13} = Gln, X^{16} = Asp, and X^{19} = Arg and was expected to recognize the site 5′-GCT-3′ on the basis of studies of a mutant of the human transcription factor Spl.[32] The second domain had X^{13} = Arg, X^{16} = Glu, and X^{19} = Arg and was expected to bind 5′-GCG-3′ on the basis of Zif268[27] and Sp1.[32] The carboxyl terminal domain had X^{13} = Arg, X^{16} = His, X^{19} = Arg and was expected to prefer 5′-GGG-3′ on the basis of mutants of Spl[32] and Krox-20,[28] a relative of Zif268. Thus, the overall binding site for this protein was predicted to be 5′-GGG GCG GCT-3′. The overall design is summarized in Figure 8. The gene for the designed set of three tandem zinc finger domains was synthesized and expressed in *E. coli* to produce the protein.[34] The protein was found by quantitative DNase I footprinting studies to bind to a restriction fragment containing the predicted site with a dissociation constant of approximately 2−3 nM, comparable to those observed for many natural DNA-binding proteins. Furthermore, selection studies using a pool of randomized DNA sequences revealed that the predicted site was indeed the optimal site! This result indicated that the design of sequence specific DNA-binding proteins by "mixing and matching" zinc finger domains with known specificity properties is possible. As a further test of this observation, a permuted protein was produced in which the first and second zinc finger domains were interchanged. As expected, this protein was found to prefer the binding site 5′-GGG GCT GCG-3′ over the site above. More quantitative studies revealed that the permuted protein discriminated between the two sites by a factor of only 7 in dissociation constant compared with a factor of nearly 200 for the original protein. Given the context dependence observed within the zinc finger domain wherein the activity of one contact residue depends on the neighboring residues, it is not surprising that interdomain context is also important; only certain combinations and orders of zinc finger domains may be fully functional. Further work is required to elucidate these effects in more detail.

Other Structures, Methods, and Classes of Zinc-Binding Domains

The crystal structures of two additional zinc finger protein−DNA complexes have been reported. The first of these is the human oncogene product GLI.[35] The DNA-binding domain from this protein contains five zinc finger domains. The cocrystal structure reveals significant differences from the Zif268 structure. In particular, only four of the five fingers contact the DNA and, while these do wrap around the major groove in a manner similar to that seen in the Zif268 structure, the same simple periodic DNA contacts are not observed. The second is a two zinc finger domain fragment from *Drosophila* Tramtrak bound to DNA.[36] Again, a structure similar to that of Zif268 was observed but with some novel features. In addition to these crystallographic studies, the structure of the TFIIIA−DNA complex has been elucidated by footprinting techniques.[37] It appears that domains one through three and seven through nine wrap around the DNA in a manner similar to that seen in the crystal structures whereas the middle three domains interact with DNA in a different manner. These structures illustrate that the general features of zinc finger protein−DNA complexes are likely to be common but that the details of the protein−DNA interactions may be more complex than can be described by simple rules.

Because of this complexity, it is desirable to have methods for producing DNA-binding zinc finger proteins with known specificity that do not rely entirely on design. Two classes of methods have been developed. The first involves determining the DNA-binding preferences of given zinc finger proteins. A common approach to this involves the selection of tightly bound DNA fragments from libraries containing a large number of randomized sites.[38,39] The sequences of individual sites are then determined and the results compiled to produce a consensus sequence. Such methods were used to confirm the optimal nature of

(31) Desjarlais, J. R.; Berg, J. M. *Proteins: Struct., Funct., Genet.* **1992**, *13*, 272.
(32) Desjarlais, J. R.; Berg, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7345.
(33) Thukrel, S. K.; Morrison, M. L.; Young, E. T. *Mol. Cell. Biol.* **1992**, *12*, 2784.
(34) Desjarlais, J. R.; Berg, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 2256.
(35) Pavletich, N. P.; Pabo, C. O. *Science* **1993**, *261*, 1701.
(36) Fairall, L.; Schwabe, J. W. R.; Chapman, L.; Finch, J. T.; Rhodes, D. *Nature* **1993**, *366*, 483.
(37) Hayes, J. J.; Tullius, T. D. *J. Mol. Biol.* **1992**, *227*, 407.
(38) Thiesen, H.-J.; Bach, C. *Nucleic Acids Res.* **1990**, *18*, 3203.
(39) Kinzler, K. W.; Vogelstein, B. *Nucleic Acids Res.* **1989**, *17*, 3645.

the binding site for the designed protein as discussed above[34] and to find the binding site for GLI used in the crystallographic studies.[39] A more recently developed variation on this relies on the use of a DNA library in which information concerning the sequences is encoded in the lengths of the DNA fragments used so that individual members do not have to be directly sequenced to obtain binding-site information.[40] The second class of methods is essentially the reverse of the selection process. A set of proteins with some of the positions thought to be involved in sequence specificity randomly mutated is screened by binding to a specific DNA binding site.[41,42] This is often accomplished by using a phage display system. In this way, a protein that binds to a site of interest can be directly obtained if it is contained within the library that has been prepared.

Finally, it is important to note that other classes of zinc-binding domains have been found to occur in other proteins including those involved in interactions with DNA and RNA. At least one member of a total of 10 classes of zinc-binding domains have now had their three-dimensional structures determined.[9] For some of these classes, the relationship between their structure and biological function is clear. For example, the DNA-binding domains of the steroid receptor superfamily of proteins are composed of sets of two adjacent zinc-binding domains that allow interactions with double-stranded DNA.[43] For others, the biochemical roles of the zinc-binding domains are less clear but the knowledge of the three-dimensional structures will certainly aid in their elucidation. Many more challenges await in the still large set of proteins which, based on patterns of cysteines and histidines in their deduced amino acid sequences, are likely to form zinc (or other metal ion)-based domains but for which no three-dimensional structural information is yet available.

(40) Desjarlais, J. R.; Berg, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11099.
(41) Rebar, E. J.; Pabo, C. O. *Science* **1994**, *263*, 671.
(42) Jamieson, A. C.; Kim, S.-H.; Wells, J. A. *Biochemistry* **1994**, *33*, 5689.
(43) Härd, T.; Gustafsson, J.-Å. *Acc. Chem. Res.* **1993**, *26*, 644.
(44) Kraulis, P. J. *J. Appl. Crystallog.* **1991**, *24*, 946.

I thank the members of the Berg laboratory, past and present, for their contributions to our work in this area.

AR940048F