Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 99–102, January 1988
Biophysics

Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins

JEREMY M. BERG

Department of Chemistry, Johns Hopkins University, 34th and Charles Streets, Baltimore, MD 21218

Communicated by Donald D. Brown, August 31, 1987

ABSTRACT **Several proteins, including the gene regulatory protein transcription factor IIIA, have been shown to contain tandem repeats of sequences of ≈30 amino acids that are believed to form structural domains around bound zinc ions ("zinc fingers"). The consensus sequence for these repeats is (Phe, Tyr)-Xaa-Cys-$(Xaa)_{2or4}$-Cys-$(Xaa)_3$-Phe-$(Xaa)_5$-Leu-$(Xaa)_2$-His-$(Xaa)_3$-His-$(Xaa)_5$, where Xaa is any amino acid. Comparisons with metalloproteins with known structures have allowed the development of a detailed three-dimensional model for these domains consisting of an antiparallel β-sheet followed by an α-helix. The proposed structure provides a basis for understanding the detailed roles of the conserved residues and allows construction of a model for the interaction of these proteins with nucleic acids in which the proteins wrap around the nucleic acids in the major groove.**

Recently a class of proteins has been discovered that contains from two to nine imperfect repeats of sequences of the form (Phe, Tyr)-Xaa-Cys-$(Xaa)_{2or4}$-Cys-$(Xaa)_3$-Phe-$(Xaa)_5$-Leu-$(Xaa)_2$-His-$(Xaa)_3$-His-$(Xaa)_5$ (1, 2). This consensus sequence is quite well defined since a total of 47 sequences of this form have been observed in nine proteins (3–11). The prototype for these proteins is transcription factor IIIA (TFIIIA), a protein that specifically binds to the 50-base-pair internal control region of 5S RNA genes and to the 5S RNA itself. The TFIIIA sequence has nine such repeats (4, 5). Based on the four conserved potential metal-binding residues (2 cysteines, 2 histidines) and on the observation that the TFIIIA:5S RNA complex contains 7–11 zinc ions per protein (4), it was proposed that each 30-amino acid repeat folds into an independent structural domain organized around a tetrahedrally coordinated Zn^{2+} ion (4, 5). This domain was termed a "zinc finger" (4) and the proteins were termed the "finger proteins."* This hypothesis is now well supported by several lines of evidence, including limited proteolysis studies (4), analysis of the TFIIIA gene structure (12), x-ray absorption studies of the zinc site (13), and preliminary studies of a peptide corresponding to a single domain (14). However, no three-dimensional structural information is yet available for any of these proteins.

The prediction of protein structures from amino acid sequences remains a highly desirable goal but a general solution seems out of reach at present. However, some progress has been made for proteins that contain features that can be identified as analogous to elements (e.g., secondary structures, ligand interactions) from proteins with known three-dimensional structures (15). The zinc finger presents an attractive problem for structural prediction since (*i*) the conserved residues that are presumably responsible for structural integrity are well established as noted above; (*ii*) the sequence is relatively small; and (*iii*) it seems to occur quite commonly in eukaryotic systems, playing roles in developmental and metabolic control through nucleic acid recognition. Several structurally characterized metalloproteins (16) have spacings between pairs of their metal-binding residues that match those from the "finger" sequences. An analysis of these proteins revealed that the metal-binding structures are quite conserved. These substructures have been combined to produce a relatively detailed model for the finger zinc-binding domains.

The consensus sequence may be divided into four parts:

[(Phe, Tyr)-Xaa-Cys-Xaa-Xaa-{Xaa-Xaa}-Cys-Xaa-Xaa-Xaa-Phe] - [Xaa_4] -
Cys-Cys loop Tip
[Xaa-Leu-Xaa-Xaa-His-Xaa-Xaa-Xaa-His] - [Xaa_5]
His-His loop Linker

The first part to be considered is the Cys-Cys loop, which will be labeled

$$\overset{1}{(\text{Phe, Tyr})}\text{-Xaa-}\overset{2}{\text{Cys}}\text{-Xaa-}\overset{3}{\text{Xaa}}\text{-}\overset{4}{\text{Xaa}}\text{-}\overset{5}{\text{Xaa}}\text{-}\{\overset{5a}{\text{Xaa}}\text{-}\overset{5b}{\text{Xaa}}\}\text{-}\overset{6}{\text{Cys}}\text{-}\overset{7}{\text{Xaa}}\text{-}\overset{8}{\text{Xaa}}\text{-}\overset{9}{\text{Xaa}}\text{-}\overset{10}{\text{Phe}}.$$

Two proteins have pairs of metal-binding cysteine residues that are separated by two or four residues and are not part of a larger cysteine-rich metal-binding sequence (such as that in, for example, metallothionein). The protein rubredoxin (17) contains two Cys-Xaa-Xaa-Cys sequences that are involved in binding a single metal ion (Fe), whereas the regulatory subunit of aspartate transcarbamoylase (18) has one Cys-Xaa-Xaa-Cys and one Cys-Xaa-Xaa-Xaa-Xaa-Cys sequence that together bind a zinc ion. Inspection of the structures of these proteins has revealed that the conformations of these regions are remarkably similar. Each of the Cys-Xaa-Xaa-Cys structures lies in a loop at the base of an antiparallel β-sheet with a hydrogen bond occurring between the NH group of the residue corresponding to $\overset{3}{\text{Cys}}$ and the carbonyl group of the residue corresponding to $\overset{8}{\text{Xaa}}$. Pairs of these structures may be superimposed with root-mean-square deviations for the metal, the sulfurs, and the backbone atoms of the residues corresponding to $\overset{2}{\text{Xaa}}$ and $\overset{9}{\text{Xaa}}$ of <1.0 Å. The Cys-Xaa-Xaa-Xaa-Xaa-Cys sequence occurs in a related conformation with an additional β-bend occurring involving residues corresponding to $\overset{5}{\text{Xaa}}$, $\overset{5a}{\text{Xaa}}$, $\overset{5b}{\text{Xaa}}$, and $\overset{6}{\text{Xaa}}$. The presence of NH to cysteine sulfur hydrogen bonds, first noted in rubredoxin and ferredoxin (19), provides a clear explanation for the similarity in these structures. For each structure a hydrogen bond or incipient hydrogen bond is observed between the sulfur atom of $\overset{n}{\text{Cys}}$ and the NH group of $\overset{n+2}{\text{Xaa}}$ with N—S distances ranging from 3.6 to 4.2 Å and N—H···S angles >150°. These interactions orient the peptide units in a manner that largely determines the overall conformation.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: TFIIIA, transcription factor IIIA.

*The term "zinc finger" is used here in a limited sense to refer only to proteins that are highly similar to the TFIIIA-like repeats and not to other proteins that have sequences suggestive of metal-binding domains.

The analogy between these structures and the zinc finger domains extends further to include the conserved hydrophobic residues. The sequences for the structures noted above are

Rubredoxin

^{4}
1 **Tyr**-Thr-Cys-Thr-Val-Cys-Gly-Tyr-Ile-**Tyr**

^{37}
2 **Trp**-Val-Cys-Pro-Leu-Cys-Gly-Val-Gly-Lys

Aspartate transcarbamoylase

^{136}
3 **Leu**-Lys-Cys-Lys-Tyr-Cys-Glu-Lys-Glu-**Phe**

^{107}
4 **Leu**-Val-Cys-Pro-Asn-Ser-Asn-Cys-Ile-Ser-His-Ala

In each case, a hydrophobic residue occurs at a position corresponding to position 1. Furthermore, for two of the sequences (**1** and **3**) an aromatic residue occurs at a position corresponding to position 10. The occurrence of these residues is particularly striking for sequence **3**, which is quite hydrophilic overall. Importantly, for these two structures (and not the others) the β-sheet extends through this region with hydrogen bonds at positions corresponding to $\overset{1}{N}$–$\overset{10}{O}$ and $\overset{1}{O}$–$\overset{10}{N}$. The two hydrophobic residues lie on the same side of the β-sheet as the metal ion and are packed against other hydrophobic residues of the proteins. The α-carbon atoms for residues 1–10 of these two structures may be superimposed with a root-mean-square deviation of 0.72 Å. These observations suggest that the conserved residues in the zinc finger sequences may play similar roles and that the structures from these proteins may be used to model that in the zinc finger domain. The structure corresponding to sequence **3** was used with a phenylalanine residue replacing leucine for model-building purposes.

Recently, the sequence of the hunchback gene from *Drosophila* was reported (20). The predicted amino acid sequence contains six finger motifs that generally fit the consensus sequence noted above except for the fact that five of the six sequences lack the aromatic residue in position 10. However, three of these five sequences have a phenylalanine or tyrosine residue in position 8. The aromatic ring from this position with the t90 side-chain conformation (21) can occupy a similar position in space as that from $\overset{10}{Phe}$ in the −90 side-chain conformation. Indeed, this is observed in rubredoxin, where the aromatic ring of Tyr occupies a similar position (relative to the metal coordination unit) to that of Phe in aspartate transcarbamoylase. Thus, the hunchback sequences are predicted to adopt a similar structure as the other finger sequences with the hydrophobic residue in position 8 playing the role of that in position 10 in the more typical sequences.

The His-His loop has the sequence $\overset{1}{Xaa}$-$\overset{2}{Leu}$-$\overset{3}{Xaa}$-$\overset{4}{Xaa}$-$\overset{5}{His}$-$\overset{6}{Xaa}$-$\overset{7}{Xaa}$-$\overset{8}{Xaa}$-$\overset{9}{His}$. Three structurally characterized proteins have a metal ion bound to two histidines separated by three residues: thermolysin (Zn) (22), hemerythrin (Fe) (23), and hemocyanin (Cu) (24), which has two such structures. In each case the region adopts an α-helical conformation with each histidine coordinated to the metal ion through the ε-nitrogen. The hypothesis that this region of the zinc finger domain is α-helical is supported by two additional observations. First, the conserved leucine precedes the first histidine by two residues such that it would lie on the same face of the helix as the histidines. Second, secondary structure predictions averaged over the known finger sequences (11) suggest the presence of a helix in the region corresponding to residues 1–7. The structure from thermolysin with leucine placed in position 2 was used for model-building purposes. It should be noted that a few of the finger sequences have four rather than three residues between the two histidine residues. No structurally characterized metalloproteins show this spacing between bound histidine residues. However, the region including residues $\overset{101}{}$–$\overset{106}{}$ in hemerythrin is part of a distorted helix with $\overset{101}{His}$ and $\overset{106}{Asp}$ coordinated to an iron atom. Thus, the presence of four residues between coordinated histidines may not preclude the extension of a helix through this region of a finger structure.

The Cys-Cys loop and the His-His loop structures found in this manner may be combined around a zinc ion in two ways (corresponding to the two chiralities of the tetrahedral zinc). The first places the carboxyl terminus of the Cys-Cys loop and the amino terminus of the His-His loop >15 Å apart, a distance too great to be spanned by the four residues of the tip region. The second (corresponding to the S absolute configuration with priorities assigned $\overset{3}{Cys} > \overset{6}{Cys} > \overset{5}{His} > \overset{9}{His}$) places these termini within 8 Å of each other and neatly packs the two aromatic residues against $\overset{2}{Leu}$ and $\overset{5}{His}$. This packing buries much of the hydrophobic surface of the two aromatic residues from the Cys-Cys loop and of $\overset{2}{Leu}$ and $\overset{5}{His}$. It remains to join the two structures by way of the tip region. This region was modeled as a type I β-bend. This is analogous to the region that precedes the zinc-binding α-helix from thermolysin (25). The overall structure was then adjusted by way of the side-chain torsional angles of the histidine residues (which vary significantly from one protein to another and control the angle between the helix and zinc coordination sphere) and the backbone torsional angles in the tip region to connect the three substructures. Several alternative depictions of the structure are shown in Fig. 1.† The structure has the form β-β-α with the α-helix packed against one face of the antiparallel sheet at an angle only slightly different than zero. Analysis of known protein structures has shown that helices tend to pack against sheets with their axes nearly parallel to the β-strands (26). Furthermore, the β-β-α unit does occur in other (non-metal-containing) proteins although it is relatively rare (26).

The last part of the structure to be considered is the linker that connects adjacent domains. This sequence is quite conserved in some of the finger proteins, such as the *Drosophila* Krüppel gene product (6, 9) with the form Thr-Gly-Glu-Lys-Pro, whereas significant variations in sequence and length occur in other proteins, such as TFIIIA. A search was performed of the amino acid sequences of proteins in the Brookhaven Data Bank (16) for sequences of the form (Ser, Thr)-(Gly, Pro)-Glu-(Lys, Arg)-(Gly, Pro)-(Phe, Tyr), where the last residue represents the first residue from the Cys-Cys loop from the next domain. Among the best matches found were $\overset{202}{Ser}$-Pro-Glu-Arg-Pro-Phe from horse muscle phosphoglycerate kinase (27) and $\overset{56}{Ser}$-Gly-Ala-Lys-Gly-Phe from *Escherichia coli* L-arabinose-binding protein (28). In these cases the serine residue is the last residue of an α-helix with the next four residues in a relatively extended loop and the phenylalanine residue lying at the amino-terminal end of a β-sheet. These observations are consistent with the secondary structures at the termini of the zinc finger domain and suggest that the linker is in a similar extended conformation with the conserved glycine and proline residues acting to terminate the α-helix and β-sheet.

Based on nuclease digestion and chemical protection studies of TFIIIA bound to a 5S RNA gene, Fairall, Rhodes, and Klug (29) proposed two models for the interactions of TFIIIA with DNA. In model I, TFIIIA wraps around the

†The coordinates for the proposed structure are available from the author upon request.

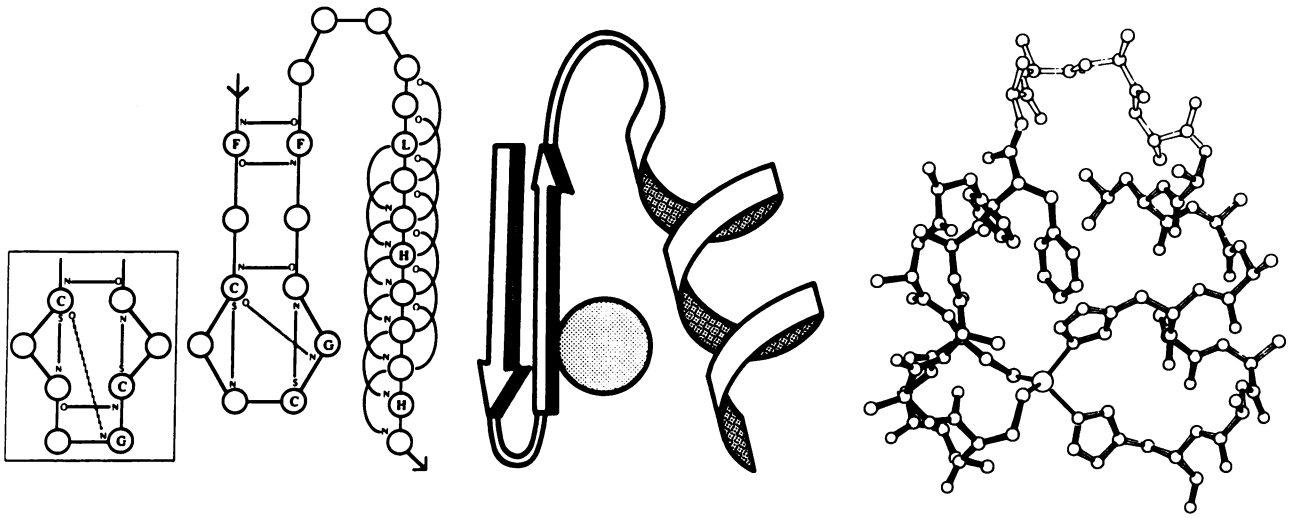

FIG. 1. (*Left*) Hydrogen-bonding pattern for the predicted zinc-binding domain structure. The main figure shows the hydrogen bonding for a Cys-Xaa-Xaa-Cys loop; the insert corresponds to a Cys-Xaa-Xaa-Xaa-Xaa-Cys loop. Glycine (G) residues, labeled in outline, occur frequently in the positions shown and allow the additional hydrogen bonds (dotted) to form. (*Center*) Schematic drawing of the proposed zinc finger structure. (*Right*) An ORTEP drawing of the proposed structure showing the side chains of the conserved metal-binding and hydrophobic residues. The Cys-Cys loop (dark), His-His loop (medium), and Tip (light) regions are indicated.

DNA with the fingers following the helical path of the major groove. In model II, the protein lies on one face of the DNA with alternate fingers lying in two different planes at an angle to one another so that each finger lies in the major groove. Based on a detailed analysis of their data, Fairall *et al.* favored model II. These two models require substantially different structures for the individual fingers. In model II, the true repeat for the structure is two fingers, so that alternate fingers contact the DNA through opposite faces, suggesting a symmetrical structure for the finger. Furthermore, model II suggests that the amino terminus and carboxyl terminus of the finger structure should exit from the same end of the finger. In contrast, in the model I, each finger interacts with the DNA in the same manner. In addition, the amino and carboxyl termini should exit from opposite ends of the structure. The proposed structure developed herein clearly favors model I with the protein following the major groove of the DNA, as shown in Fig. 2. The proposed structure suggests that the α-helix lies in the major groove of DNA with the Cys-Cys loop β-structure lying farther away from the DNA helical axis. This suggests that specific contacts are made by the carboxyl-terminal residues of the tip region and by the helical His-His loop, particularly by residues $X^{3}aa$ and $X^{4}aa$. This arrangement is quite appealing in that interactions between the amino-terminal ends of α-helices and the major groove of DNA are a feature common to several classes of structurally characterized DNA-binding proteins (30, 31).

FIG. 2. Model for the interaction between a protein consisting of tandemly repeated zinc finger domains and DNA based on the predicted structure for the individual domains. The α-helix from each domain lies in the major groove of the DNA and makes sequence-specific contacts with the edges of the base pairs, whereas the β-sheet lies further away from the DNA helical axis and contacts the DNA backbone. The linker also lies in the major groove and may make important contacts.

Nonspecific backbone contacts would be made by the edges of the β-sheet, including residue $X^{8}aa$ of the Cys-Cys loop, which is the most conserved basic amino acid in the finger sequences (3–11). This residue is constrained by the conformation of the Cys-Cys loop to point toward the noncoding strand of the 5S gene, the strand with which TFIIIA is known to make more extensive contacts (32). Additional contacts could be made by the two hydrophilic amino acids of the linker region that would lie in the major groove as well.

Until an experimentally determined structure of a finger protein or a single zinc-binding domain peptide (14) becomes available, the structure described here provides a useful model for considering the mechanisms of action of these proteins. It should be noted that this model depends on the spacing between the metal-binding residues and between these and the conserved hydrophobic residues in the consensus sequence. Though most of the sequences do fit the consensus in this regard, differences do occur. Clearly, variations in the finger structure itself or in the length or characteristics of the linkers between fingers may allow variations in the types of interaction between finger proteins and nucleic acids. These variant structures may be crucial for the nucleic acid recognition process.

I thank Drs. Tom Tullius and Mark Snow for useful discussions and Eric Suchanek for assistance in preparation of Fig. 1. This work was supported by the Camille and Henry Dreyfus Foundation (Distinguished New Faculty Award) and the National Institutes of Health (GM 38230). The Interactive Graphics Facility of the Department of Biophysics at the Johns Hopkins School of Medicine was established and is maintained by National Institutes of Health and National Science Foundation grants and by a gift from the Richard-King Mellon Foundation.

1. Berg, J. M. (1986) *Nature (London)* **319**, 264–265.
2. Vincent, A. (1986) *Nucleic Acids Res.* **14**, 4385–4391.
3. Ginsberg, A. M., King, B. O. & Roeder, R. G. (1984) *Cell* **39**, 479–489.
4. Miller, J., McLachlan, A. D. & Klug, A. (1985) *EMBO J.* **4**, 1609–1614.
5. Brown, R. S., Sander, C. & Argos, P. (1985) *FEBS Lett.* **186**, 271–274.
6. Rosenberg, U. B., Schroder, C., Preiss, A., Kienlin, A., Côté, S., Riede, I. & Jäckle, H. (1986) *Nature (London)* **319**, 336–339.

7. Vincent, A., Colot, H. V. & Rosbash, M. (1985) *J. Mol. Biol.* **186,** 149–166.
8. Hartshorne, T. A., Blumberg, H. & Young, E. T. (1986) *Nature (London)* **320,** 283–287.
9. Schuh, R., Aicher, W., Gaul, U., Côté, S., Preiss, A., Maier, D., Siefert, E., Nauber, U., Schröder, C., Kemler, R. & Jäckle, H. (1986) *Cell* **47,** 1025–1032.
10. Chowdhury, K., Deutsch, U. & Gruss, P. (1987) *Cell* **48,** 771–778.
11. Brown, R. S. & Argos, P. (1986) *Nature (London)* **324,** 215.
12. Tso, J. Y., Van den Berg, D. J. & Korn, L. J. (1986) *Nucleic Acids Res.* **14,** 2187–2200.
13. Diakun, G., Fairall, L. & Klug, A. (1986) *Nature (London)* **324,** 698–699.
14. Frankel, A. D., Berg, J. M. & Pabo, C. O. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4841–4845.
15. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987) *Nature (London)* **326,** 347–352.
16. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
17. Watenpaugh, K. D., Sieker, L. V. & Jensen, L. H. (1980) *J. Mol. Biol.* **138,** 615–633.
18. Honzatko, R. B., Crawford, J. L., Monaco, H. L., Ladner, J. E., Ewards, B. F. P., Evans, D. R., Warren, S. G., Wiley, D. C., Ladner, R. C. & Lipscomb, W. N. (1982) *J. Mol. Biol.* **160,** 219–263.
19. Adman, E., Watenpaugh, K. D. & Jensen, L. H. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 4854–4858.
20. Tautz, D., Lehmann, R., Schnurch, H., Schuh, R., Seifert, E., Kienlin, A., Jones, J. & Jäckle, H. (1987) *Nature (London)* **327,** 383–389.
21. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
22. Holmes, M. A. & Matthews, B. W. (1980) *J. Mol. Biol.* **160,** 623–639.
23. Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1983) *Acta Crystallogr. B* **39,** 697–703.
24. Gaykema, W. P. J., Volbeda, A. & Hol, W. G. J. (1985) *J. Mol. Biol.* **187,** 255–275.
25. Matthews, B. W., Weaver, L. H. & Kester, W. R. (1974) *J. Biol. Chem.* **249,** 8030–8044.
26. Chothia, C. (1984) *Annu. Rev. Biochem.* **53,** 537–572.
27. Banks, R. D., Blake, C. C. F., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M. & Phillips, A. W. (1979) *Nature (London)* **279,** 773–777.
28. Gilliland, G. L. & Quiocho, F. A. (1981) *J. Mol. Biol.* **146,** 341–362.
29. Fairall, L., Rhodes, D. & Klug, A. (1986) *J. Mol. Biol.* **192,** 577–591.
30. Pabo, C. O. & Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53,** 293–321.
31. McClarin, J. A., Frederick, C. A., Wang, B.-C., Greene, P., Boyer, H. W., Grable, J. & Rosenberg, J. M. (1986) *Science* **234,** 1526–1541.
32. Smith, D. R., Jackson, I. J. & Brown, D. D. (1984) *Cell* **37,** 645–652.