Reports

Potential Metal-Binding Domains in Nucleic Acid Binding Proteins

JEREMY M. BERG

A systematic search for sequences that potentially could form metal-binding domains in proteins has been performed. Five classes of proteins involved in nucleic acid binding or gene regulation were found to contain such sequences. These observations suggest numerous experiments aimed at determining whether metal-binding domains are present in these proteins and, if present, what roles such domains play in the processes of nucleic acid binding and gene regulation.

CHARACTERIZING THE MECHANISMS by which proteins bind to nucleic acids is of fundamental importance to understanding the processes of gene expression and replication. Recently Miller, McLachlan, and Klug proposed a novel mechanism for nucleic acid binding by transcription factor IIIA of *Xenopus laevis* (TFIIIA), a protein that binds to 5*S* RNA genes and to their RNA transcripts (*1*). They demonstrated that this protein binds 7 to 11 Zn^{2+} ions per ribonucleoprotein particle (*1*) and that it contains nine homologous units of about 30 amino acids (*1, 2*). Each of these contains a sequence of the form Cys-X_{2-5}-Cys-X_{12}-His-X_{2-3}-His, where X may be any amino acid. They proposed that the repeated units represent Zn^{2+} binding domains that interact with nucleic acids via hydrophilic residues. I report here that a systematic search for analogous sequences in other proteins has revealed that potential metal-binding domains occur in several other classes of proteins that have been implicated in nucleic acid binding. These results suggest that metal complex–based units may be present in a number of nucleic acid binding and gene regulatory proteins.

Initial searches (*3*) of the National Biomedical Research Foundation protein sequence library with sequences from TFIIIA were unsuccessful (*4*). However, a more general search for sequences of the form Cys-X_{2-4}-Cys-X_{2-15}-a-X_{2-4}-a or a-X_{2-4}-a-X_{2-15}-Cys-X_{2-4}-Cys, where a may be either cysteine or histidine, yielded interesting results. Such sequences were chosen by analogy with the sequences in TFIIIA and by analysis of metal-binding sites in structurally characterized proteins (*5*). Examples of short Cys- or His-containing sequences that provide two ligands to single metal ions include: Cys-X_2-Cys [liver alcohol dehydrogenase (E.C. 1.1.1.1) (Zn), aspartate carbamoyltransferase (E.C. 2.1.3.2) (Zn), metallothionein (Zn, Cd), rubredoxin (Fe)]; Cys-X_2-His [plastocyanin (Cu)]; His-X_2-His [Cu-Zn superoxide dismutase (E.C. 1.15.11) (Zn)]; Cys-X_3-Cys [metallothionein (Cd)]; His-X_3-His [hemerythrin (Fe)]; Cys-X_4-Cys [aspartate carbamoyltransferase (Zn), ferredoxin (Fe)]; and Cys-X_4-His [azurin (Cu)].

The search identified several classes of proteins. These included certain cytochromes c (where a Cys-X_2-Cys sequence is covalently linked to heme), iron-sulfur proteins, small disulfide-rich proteins (protease inhibitors, hormones, and toxins), high-sulfur keratins, and metallothioneins. Another major class of proteins that was well represented are those that have been implicated in nucleic acid binding. The sequences found in nucleic acid binding proteins are shown in Table 1. These proteins may be divided into several groups.

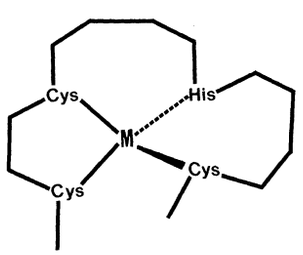

Fig. 1. A possible structure for a metal-binding domain from a retroviral low molecular weight nucleic acid binding protein. A tetrahedral coordination geometry is assumed. The other sequences found may adopt similar structures with different loop sizes. This structure is analogous to that proposed by Miller *et al.* (*1*).

The first and most prominent group contains the low molecular weight nucleic acid binding proteins encoded within the *gag* genes of retroviruses (*6–8*). These proteins, which contain fewer than 100 amino acids, bind to single-stranded DNA and to RNA. Specific interactions with homologous viral RNA have also been demonstrated (*9*). The proteins from some viruses contain one copy of a sequence of the form Cys-X_2-Cys-X_4-His-X_4-Cys while others contain two such sequences separated by 5 to 11 residues. A possible structure for these units is given in Fig. 1. The hypothesis that these sequences form metal-binding units rather than some other structure is supported by the following observations. (i) Some proteins contain one sequence of this form while others contain two, suggesting that each unit forms an independent structure. (ii) The four potential metal-binding residues are the only completely conserved residues in such sequences from these proteins. Searches (*3*) for homologous sequences revealed only a nucleic acid binding protein from simian sarcoma virus that has four additional residues inserted before the final Cys. (iii) Experimental attempts to identify unique disulfide bonds present in the native proteins have been unsuccessful (*8*). (iv) The Gly residues that frequently occur in positions 5 and 8 within these sequences are analogous to the Gly residues in the metal-binding sequences of rubredoxins (Cys-X_2-Cys-Gly) (*10*) and azurins (Cys-X_3-Gly-His) (*11*), respectively. While the sequence homology among the retroviral proteins has been described previously (*12*), the potential of the sequence to form a metal-binding domain had not been noted.

The second group of proteins consists of the adenovirus E1A gene products. These proteins contain sequences of the form Cys-X_2-Cys-X_{13}-Cys-X_2-Cys. The sequences contain a number of basic and other potentially hydrogen-bonding residues, particularly in their amino terminal halves. Two overlapping messenger RNA's are transcribed from the E1A genes. These differ by a small (93 to 138 nucleotide) internal sequence that is removed by splicing such that the two messages remain in the same reading frame (*13*). Thus, two protein products are produced that differ by an internal sequence of 31 to 46 amino acids. Results from recent studies with mutant E1A genes have revealed that the two proteins have different functions, with the larger protein more effectively stimulating transcription from early promoters (*14–16*). The potential metal-binding sequences are encoded in the sections of the E1A genes that are removed by splicing for all three known adenovirus E1A sequences (*13, 17*). This finding implies that these sequences play a crucial role in positive regulation.

The third group consists of several transfer RNA (tRNA) synthetases. Two methionyl-tRNA synthetases (E.C. 6.1.1.10)

Department of Biophysics, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205.

(MetRS) have sequences of the form Cys-X_2-Cys-X_9-Cys-X_2-Cys. *Escherichia coli* isoleucyl-tRNA synthetase (E.C. 6.1.1.5) (IleRS) has a similar but somewhat longer sequence. The *E. coli* MetRS and IleRS have been shown to contain one Zn^{2+} ion per polypeptide chain and to be inhibited by 1,10-phenanthroline (*18*). While the crystal structure of a fully active and Zn-containing proteolytic fragment of the *E. coli* MetRS enzyme has been reported, the Zn site had not been unambiguously located, and some sections of the polypeptide chain had not been traced (*19*). A sequence of the form Cys-X_2-Cys-X_6-His-X_2-His is found in *E. coli* alanyl-tRNA synthetase (E.C. 6.1.1.7). This sequence is weakly homologous to sequences in *Bacillus stearothermophilus* tyrosyl-tRNA synthetase (E.C. 6.1.1.1) and *E. coli* MetRS that lack the second Cys residue and are not involved in metal binding. However, the *E. coli* alanyl-tRNA synthetase sequence is unlikely to form a structure similar to those of the regions of weak homology in *B. stearothermophilus* tyrosyl-tRNA synthetase and *E. coli* MetRS (*20*).

The fourth group of proteins are the Large T antigens from simian virus 40 (SV40) and polyoma viruses. These sequences have the form Cys-X_2-Cys-X_{11-13}-His-X_2-His, with available hydrogen-bonding residues distributed throughout. The sequence in SV40 T antigen occurs in or adjacent to sites that have been deleted in mutant proteins that are deficient in binding to the origin of replication (*21*). However, other sites in the T antigen sequence also have been implicated in DNA binding.

The last group of proteins are from bacteriophages. A sequence of the form Cys-X_3-His-X_5-Cys-X_2-Cys is found beginning at residue 77 of the helix-destabilizing protein from phage T4. This sequence lies within the region (residues 72 to 116) implicated in nucleic acid binding on the basis of spectroscopic and chemical modification studies (*22*).

A systematic search has revealed that a variety of nucleic acid binding proteins contain sequences with four Cys or His residues arranged in a manner suggestive of a metal-binding domain. Moreover, in a number of cases the sequences are found in regions of the proteins that have been implicated by other studies as being functionally significant. These observations suggest that metal-binding domains may be present in, and functionally important to, these proteins. Metal-binding units might function so that relatively short sequences can form independently structured domains, as they apparently do in TFIIIA. Another example of such a domain is found in the Zn-domain of the regulatory chain of aspartate carbamoyltransferase (*23*), which contains a sequence of the form Cys-X_4-Cys-X_{25}-Cys-X_2-Cys.

The hypothesis that the sequences found represent metal-binding domains suggests several types of experiments, including the determination of the metal content of these proteins and their sensitivity to chelating agents. Furthermore, the hypothesis suggests target sites for directed mutagenesis studies. The results of such experiments will allow an evaluation of the general significance of metal-binding domains in protein–nucleic acid interactions and in gene regulation.

Note added in proof: Cysteine-rich sequences in the nucleic acid binding regions of several other proteins have recently been noted: human glucocorticoid receptor (*26*), human estrogen receptor (*27*), and yeast GAL4, PPRI, and ARGII proteins (*28*).

Table 1. Potential metal-binding sequences from nucleic acid binding proteins.

Protein*	Position†	Sequence‡
Retroviral low molecular weight nucleic acid binding proteins		
HTLV-I§	11 (357)	C FR C GKAG H WSRD C
	36 (380)	C PL C QDPT H WKRD C
HTLV-II	13 (363)	C FR C GKVG H WSRD C
	36 (386)	C PL C QDPS H WKRD C
HTLV-III	29 (392)	C FN C GKEG H TARN C
	50 (413)	C WK C GKEG H QMKD C
VLV (*24*)	19 (387)	C YN C GKPG H LARQ C
	38 (406)	C HH C GKRG H MQKD C
EIAV (*25*)	24 (383)	C YN C GKPG H LSSQ C
	43 (402)	C FK C KQPG H FSKQ C
BLV	24 (347)	C YR C LKEG H WARD C
	49 (372)	C PI C KDPS H WKRD C
RSV	21 (509)	C YT C GSPG H YQAQ C
	47 (535)	C QL C NGMG H NAKQ C
AKV MuLV	(503)	C AY C KEKG H WAKD C
RMuLV	26	C AY C KEKG H WAKD C
MMuLV	(504)	C AY C KEKG H WAKD C
MMuSV	(504)	C TY C EEQG H WAKD C
BaEV	26 (498)	C AY C KERG H WTKD C
FeLV	30	C AY C KEKG H WVRD C
SSV	(492)	C AY C KEKG H WDEEIAPA C
Adenovirus E1A gene products		
Adenovirus 2,5	154	C RS C HYHRRNTGDPDIM C SL C
Adenovirus 7	163	C KS C EFHRNNTGMKELL C SL C
Adenovirus 12	159	C KS C EHHRNSTGNTDLM C SL C
Aminoacyl tRNA synthetases		
E. coli Met-RS	145	C PK C LSPDQYGDN C EV C
Yeast Met-RS	337	C PK C HYDDARGDQ C DK C
E. coli Ile-RS	109	C PR C WHYTQDVGKVAFHAEI C GR C
E. coli Ala-RS	179	C DP C EIFYD H GD H
Large T antigens		
SV40	302	C LK C IKKEQPSHYKY H EK H
HuBKPV	304	C KK C QKKDQPYHFKY H EK H
HuJCPV	303	C KK C EKKDQPNHFNH H EK H
MuPV	452	C RS C SKEETRLQIHWKN H RK H
Bacteriophage proteins		
T4 helix-destabilizing protein	77	C SST H GDYDA C PV C
Phi-X174 gene A protein	246	C YQYF C VPEYGTANGRL H AV H
G4 gene A protein	287	C YQYF C VPEYGTQHGRL H AV H
T4 DNA primase	17	C DN C GSSDGNSLFSDGHTF C YV C
Lambda DNA packaging protein A	142	H FMRF H VA C PH C

*Abbreviations used: HTLV-I, human T-cell leukemia virus I; HTLV-II, human T-cell leukemia virus II; HTLV-III, human T-cell leukemia virus III; VLV, visna lentivirus; EIAV, equine infectious anemia virus; BLV, bovine leukemia virus; RSV, Rous sarcoma virus; AKV MuLV, AKV murine leukemia virus; RMuLV, Rauscher murine leukemia virus; MMuLV, Moloney murine leukemia virus; MMuSV, Moloney murine sarcoma virus; BaEV, baboon endogenous virus; FeLV, feline leukemia virus; SSV, simian sarcoma virus; SV40, simian virus 40; HuBKPV, human BK polyomavirus; HuJCPV, human JC polyomavirus; MuPV, murine polyomavirus. †The number given is the residue number of the first amino acid in the sequence shown. The number given in parentheses is the residue number of the first amino acid in the sequence within the *gag* polyprotein (for the retroviral nucleic acid binding proteins). The starting residues for the proteins from HTLV-III, VLV, and EIAV are from (*25*). ‡Standard single-letter amino acid abbreviations: A, Ala; R, Arg; N, Asn; D, Asp; C, Cys; Q, Gln; E, Glu; G, Gly; H, His; I, Ile; L, Leu; K, Lys; M, Met; F, Phe; P, Pro; S, Ser; T, Thr; W, Trp; Y, Tyr; and V, Val. §Similar sequences with single amino acid changes have been observed in lymphadenopathy-associated virus-1a and in AIDS-associated retrovirus-2.

REFERENCES AND NOTES

1. J. Miller, A. D. McLachlan, A. Klug, *EMBO J.* **4**, 1609 (1985).
2. The repeating nature of the TFIIIA sequence has been independently noted [R. S. Brown, C. Sander, P. Argos, *FEBS Lett.* **186**, 271 (1985)].
3. D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985).
4. Recently sequences homologous to the repeats in TFIIIA have been discovered [A. Vincent *et al.*, *J. Mol. Biol.* **186**, 149 (1985); U. B. Rosenberg *et al.*, *Nature (London)* **319**, 336 (1986)].

5. F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977).
6. R. C. Nowinski, E. Fleissner, N. H. Sarkar, T. Aoki, *J. Virol.* **9**, 359 (1972).
7. C. W. Long, L. E. Henderson, S. Oroszlan, *Virology* **104**, 491 (1980).
8. L. E. Henderson, T. D. Copeland, R. C. Sowder, G. W. Smythers, S. Oroszlan, *J. Biol. Chem.* **256**, 8400 (1981).
9. A. Sen, C. J. Sherr, G. J. Todaro, *Cell* **10**, 489 (1977); J. L. Darlix and P.-F. Spahr, *J. Mol. Biol.* **160**, 147 (1982); C. Meric, J.-L. Darlix, P.-F. Spahr, *ibid.* **173**, 531 (1984).
10. E. Adman, K. D. Watenpaugh, L. H. Jensen, *Proc. Nat. Acad. Sci. U.S.A.* **72**, 4854 (1975).
11. L. Ryden and J.-O. Lundgren, *Nature (London)* **261**, 344 (1976).
12. T. D. Copeland, S. Oroszlan, V. S. Kalyanaraman, M. G. Sarngadharan, R. C. Gallo, *FEBS Lett.* **162**, 390 (1983).
13. M. Perricaudet, G. Akusjarvi, A. Virtanen, U. Pettersson, *Nature (London)* **281**, 694 (1979).
14. R. P. Ricciardi, R. L. Jones, C. L. Cepko, P. A. Sharp, B. E. Roberts, *Proc. Nat. Acad. Sci. U.S.A.* **78**, 6121 (1981).
15. C. Montell, E. F. Fisher, M. H. Caruthers, A. J. Beck, *Nature (London)* **295**, 380 (1982).
16. R. A. Guilfoyle, W. P. Osheroff, M. Rossini, *EMBO J.* **4**, 707 (1985).
17. R. Dijkema *et al.*, *Gene* **12**, 287 (1980); Y. Sawada and K. Fujinaga, *J. Virol.* **36**, 639 (1980).
18. L. H. Posorske, M. Cohn, N. Yanagisawa, D. S. Auld, *Biochim. Biophys. Acta* **576**, 128 (1979); J.-F. Mayaux and S. Blanquet, *Biochemistry* **20**, 4647 (1981).
19. C. Zelwer, J. L. Risler, S. Brunie, *J. Mol. Biol.* **155**, 63 (1982).
20. D. M. Blow *et al.*, *ibid.* **171**, 571 (1983).
21. R. Clark, K. Peden, J. M. Pipas, D. Nathans, R. Tjian, *Mol. Cell. Biol.* **3**, 220 (1983).
22. K. R. Williams, M. B. LoPresti, M. Setoguchi, *J. Biol. Chem.* **256**, 1754 (1981); R. V. Prigodich, J. Casas-Finet, K. R. Williams, W. Konigsberg, J. E. Coleman, *Biochemistry* **23**, 522 (1984).
23. R. B. Honzatko *et al.*, *J. Mol. Biol.* **160**, 219 (1982).
24. P. Sonigo *et al.*, *Cell* **42**, 369 (1985).
25. R. M. Stephens, J. W. Casey, N. R. Rice, *Science* **231**, 589 (1986).
26. C. Weinberger *et al.*, *Nature (London)* **318**, 670 (1985).
27. G. L. Greene *et al.*, *Science* **231**, 1150 (1986).
28. L. Keegan *et al.*, *ibid.*, p. 669.
29. I thank Carl Pabo for support and useful discussions and the Jane Coffin Childs Memorial Fund for a Fellowship.

1 November 1985; accepted 18 February 1986

Occurrence of Peptide and Clavine Ergot Alkaloids in Tall Fescue Grass

PHILIP C. LYONS, RONALD D. PLATTNER, CHARLES W. BACON

Evidence is presented that ergot alkaloids are ubiquitous in tall fescue pastures infected with the clavicipitaceous fungal endophyte *Sphacelia typhina* (or *Acremonium coenophialum*). Ergopeptide alkaloids, predominantly ergovaline, constituted 10 to 50 percent of the total ergot alkaloid concentration, which was as high as 14 milligrams per kilogram in sheaths and 1.5 milligrams per kilogram in blades. Ergot alkaloid concentrations were substantially increased by application of large amounts (10 millimoles per liter) of potassium nitrate or ammonium chloride to infected plants in the greenhouse. The results indicate that ergot alkaloids are probably responsible for the toxicity to cattle of this common pasture and lawn grass and that ergotism-like toxicoses may be caused by clavicipitaceous fungi other than *Claviceps*.

TALL FESCUE (*FESTUCA ARUNDINACEA* Schreb) is the predominant cool-season perennial forage grass in the United States, particularly in the transition zone of the eastern states. Unfortunately, it is frequently toxic to cattle. The most severe form of toxicosis, fescue foot, is a gangrene of the animal's extremities that is strikingly similar to ergotism; but it occurs in the absence of the ergot fungus *Claviceps* (*1–4*). A less severe but economically more significant toxic manifestation of tall fescue in cattle is the so-called "summer syndrome," which is characterized by weight loss or reduced weight gain, rough hair coat, and increased temperature and respiration (*5*). The occurrence of these toxic syndromes has been associated with endophytic infection of the grass by another clavicipitaceous fungus, *Sphacelia typhina* (or *Acremonium coenophialum*); however, the role of this endophyte in tall fescue toxicity is not presently understood (*6–8*).

The similarity of fescue foot to ergotism is the basis for postulating that vasoconstrictive substances such as ergot alkaloids, synthesized by the grass or the endophytic fungus associated with it, are responsible for this disorder (*9–17*). We now report that ergot alkaloids, including several toxic ergopeptide species, are commonly present in all aboveground parts in infected tall fescue.

To establish whether ergot alkaloids are commonly associated with endophyte infection, we obtained samples for analyses from eight infected and two uninfected pastures in northern Georgia. All the pastures except one were sampled once between June and October 1984 (one infected pasture was sampled twice, first in December 1983 and again in June 1984 after flowering). Several of the infected pastures had recent histories of toxicity. We estimated the infection levels in the pastures by staining sheath sections from 40 randomly chosen tillers with aniline blue and examining them microscopically for the fungus. Total concentrations of ergot alkaloids were measured colorimetrically on the basis of the formation of a blue complex with *p*-dimethylaminobenzaldehyde (*18*); ergopeptide alkaloids were identified and measured by tandem mass spectrometry (MS) (Finnigan 4535/TSQ quadropole mass spectrometer) in the negative chemical-ionization mode. This procedure separates all of the known ergopeptide alkaloids and is sensitive to the picogram level (*19–21*).

Ergot alkaloids were detected colorimetrically in all infected samples but not in uninfected samples (Table 1). The total ergot alkaloid concentration (micrograms of ergonovine per gram dry weight) varied among the samples from 1.0 to 14 μg/g in sheaths, where the fungus grows extensively, and from 0.4 to 1.5 μg/g in the blades, which are free of infection (Table 1). Ergot alkaloids were also present in inflorescence stems and inflorescences of the sample collected in June when the seeds were at late dough maturity. The concentrations in these tissues were comparable, respectively, to those in blades and sheaths (Table 1, sample 1-B). Both stems and inflorescences are infected by the endophyte.

Tandem MS revealed that ergopeptide alkaloids were present in all infected samples and accounted for 10 to 50 percent of the total ergot alkaloid concentration (Table 1). Five ergopeptide alkaloids were detected, of which three, ergovaline, ergosine, and ergonine, occurred in all samples in both blades and sheaths. These three alkaloids also were present in inflorescences and stems of the sample in which these parts were assayed. Ergoptine and ergocornine were detected in only a few samples and in small amounts. Ergopeptide alkaloid concentrations, based on tandem MS of samples spiked with known concentrations of ergovaline, varied from 0.1 to 0.3 μg/g in blades and from 0.3 to 2.8 μg/g in sheaths (Table 1). Ergovaline was the predominant species in all the samples, accounting for 84 to 97 percent of the total ergopeptide alkaloid fraction. Ergonine and ergosine were present in about equal concentrations. All five ergopeptide alkaloids were produced (in about the same relative proportions as in the

P. C. Lyons, Department of Plant Pathology, University of Georgia, Athens, GA 30602.
R. D. Plattner, Instrument Analysis Research Unit, Northern Regional Research Center, U.S. Department of Agriculture–Agricultural Research Service, Peoria, IL 61604.
C. W. Bacon, Toxicology and Biological Constituents Research Unit, R. B. Russell Agricultural Research Center, U.S. Department of Agriculture–Agricultural Research Service, Athens, GA 30613.