Reviews

Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins

Kevin Struhl

Four distinct structural motifs have been proposed for the DNA-binding domains of eukaryotic transcriptional regulatory proteins; the helix-turn-helix, two kinds of zinc finger, and the leucine zipper. Within each structural motif, there are often families of related proteins that recognize similar DNA sequences and are conserved throughout the eukaryotic kingdom. However, the processes of transcriptional activation and repression appear to be independent of the specific type of protein–DNA interaction.

Eukaryotic genes are regulated differentially in response to a complex set of environmental and developmental cues. In terms of transcriptional regulation, eukaryotic promoters are large, complex arrangements of short DNA sequences that are recognized by a wide variety of specific DNA-binding proteins that activate or repress transcription. The distinct transcriptional regulatory patterns of individual genes or sets of genes are determined primarily by the specific protein–DNA interactions that occur at the promoters.

Structural and functional analyses of eukaryotic DNA-binding proteins indicate that small autonomous domains containing less than 100 amino acid residues are sufficient for specific DNA-binding activity. Although detailed structural information is not yet available, the primary sequences and biochemical properties of various DNA-binding domains are suggestive of at least four basic structural motifs. These motifs are probably important for the overall structure of the DNA-binding domain rather than being directly involved in the specific contacts between protein and DNA, because different proteins containing a particular structural motif can recognize a variety of DNA sequences.

K. Struhl is at the Department of Biological Chemistry, Harvard Medical School, Boston, MA 02115, USA.

The helix-turn-helix

The first and by far the best characterized structural motif for a DNA-binding domain is the helix-turn-helix found in prokaryotic activator and repressor proteins (reviewed in Ref. 1). X-ray crystallographic analysis of several proteins and protein–DNA complexes have elucidated the structures in detail, and functionally important interactions have been uncovered by analysing mutant proteins and target DNA sites. As is obvious from the name, the crucial structure in this class of proteins contains two α-helices that are separated by a β-turn. Despite considerable sequence variability, the helix-turn-helix motifs have a highly conserved geometry. Amino acids within one of these helices, the so-called 'recognition helix', directly contact bases exposed in the major groove of the target DNA; the other α-helix lies across the major groove and makes some non-specific contacts to DNA (Fig. 1). The overall structural homology among helix-turn-helix motifs makes it possible in some cases to change the specificity of the protein–DNA interaction by altering or swapping recognition helices[2]. The prokaryotic helix-turn-helix proteins bind as dimers to DNA sequences that have dyad symmetric character.

In eukaryotic organisms, the helix-turn-helix motif was first invoked for a family of *Drosophila* proteins that control many of the key decisions in early development (reviewed in Refs 3, 4). These proteins contain a highly conserved 'homeodomain' about 60 amino acids in length within which lies a region with sequences common to many prokaryotic helix-turn-helix proteins. Homeodomains have now been found in a wide variety of eukaryotic organisms ranging from yeast to human, and more than 80 such proteins have already been identified. With the exception of the more diverged yeast proteins, different homeodomains show 40–90% amino acid sequence identity and can be classified into distinct subtypes based on the extent of homology. The subtypes often differ in their putative recognition helices, suggesting that they recognize different DNA sequences. Although there is no direct evidence that homeodomains contain helix-turn-helix motifs, it has been shown in several cases that the proteins specifically bind DNA *in vitro*. In cases where the functions of these proteins are known or inferred, homeodomains are frequently associated with proteins that control cell fate during development. (See *TIBS* 14, 52–56 [1989] for a review of vertebrate homeodomain proteins.)

The yeast MATα2 protein, which regulates cell type by binding to operator sequences and repressing transcription, is the best characterized eukaryotic protein with a putative helix-turn-helix motif[5]. The unusually large (30 bp) α2 operators have highly conserved sequences at both ends with an approximate twofold symmetry, but lack sequence similarities in the middle. In accord with this unusual arrangement, α2 dimers directly interact with half-sites at each end of the operator but do not contact the center of the operator. The protein (α2) is remarkably flexible as it can bind to an operator with the central 13 bp deleted such that the half-sites are immediately adjacent and located on the opposite side of the DNA helix. Like the prokaryotic proteins, α2 contains two domains; a C-terminal homeodomain that interacts specifically with operator DNA although with a reduced affinity compared to the full length protein, and an N-terminal region that does not bind DNA but facilitates dimerization. Amino acid substitutions within the α2 homeo-like region abolish function,

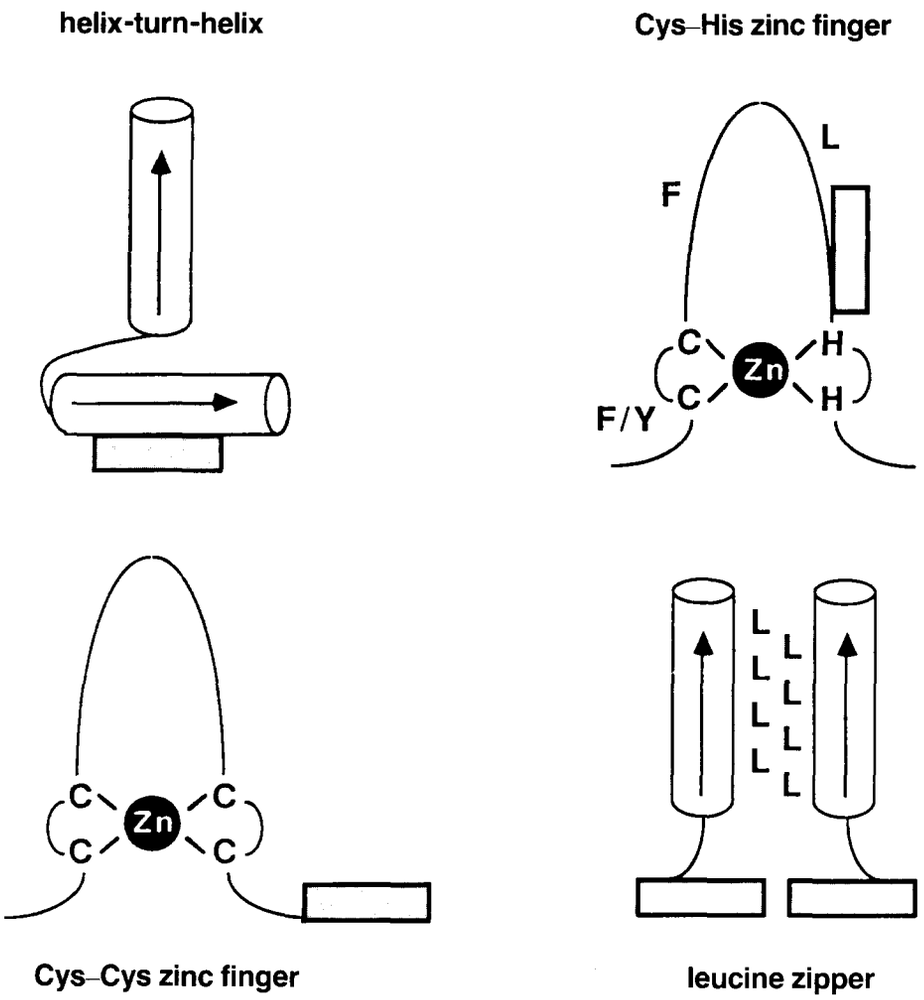

Fig. 1. Schematic representation of the four structural motifs described in the text. α-Helices are represented by cylinders with arrows indicating directionality; conserved amino acid residues are shown in one-letter code, and zinc ions shown as black circles. The shaded boxes indicate the regions of the proteins proposed to be involved in specific contacts to DNA.

presumably by virtue of a loss of DNA-binding activity[6].

The zinc finger

The unusual sequence of TFIIIA, a protein required for transcription of the 5S RNA genes by RNA polymerase III, led to the proposal of a distinct DNA-binding motif, the zinc finger[7]. TFIIIA contains 7–11 zinc atoms per molecule, and it is composed of nine repeating units of approximately 30 residues. Each unit contains two invariant pairs of cysteines and histidines as well as other conserved amino acids. In the original zinc-finger model, the cysteine and histidine pairs serve as a tetrahedral coordination site for a single zinc ion, and the amino acids between these coordination sites project out as fingers[7] (Fig. 1). (See also *TIBS* 12, 464–469 [1987] for review.) A different model, which invokes an antiparallel β-sheet and α-helix between the zinc coordination sites, has been proposed by analogy with structures of other metalloproteins[8]. Deletion analysis of both TFIIIA protein and its large target site within the 5S RNA structural gene indicate that the zinc fingers contribute to specific DNA binding in a modular fashion, and can be viewed as being arranged in a linear array along the target sequence[9,10]. However, individual fingers do not interact precisely with adjacent nucleotides, but rather clusters of fingers bind to the three functionally important regions of the binding site.

DNA sequence analysis of cloned genes indicates that more than 15 proteins from a variety of eukaryotic organisms from yeast to man contain regions that strongly resemble the putative zinc-finger motif of TFIIIA (reviewed in Ref. 11). Some of these proteins are known to be transcriptional activators, while others have presumptive roles in development or in sex determination. Depending on the protein, the number of potential zinc fingers ranges from two to more than ten. In several cases, it has been demonstrated that these proteins bind specifically to DNA, and that both the zinc-finger region and zinc are necessary for binding. In both models, specific DNA contacts are proposed to occur with residues in the finger regions, possibly the putative α-helix[7,8]. However, a synthetic peptide containing the two zinc fingers of yeast ADR1 is unable to recognize its target site, even though it forms a discrete structure in the presence of zinc that can interact non-specifically with DNA[12]. This suggests that the zinc finger has an essential structural role for a functional DNA-binding domain, but may not always be directly involved in protein–DNA interactions.

There is a related, but distinct, class of specific DNA-binding proteins, exemplified by the yeast GAL4 transcriptional activator and mammalian steroid receptors, in which putative zinc fingers are formed by coordination with four cysteine residues; replacement of one cysteine pair by a histidine pair abolishes DNA-binding activity (reviewed in Ref. 11) (Fig. 1). In the case of GAL4, an important role for zinc is inferred from mutants that function only at high zinc concentrations *in vitro* and *in vivo*[13]. Finger swapping experiments suggest that the all cysteine zinc fingers, though essential for DNA-binding, are not important for specificity. For example, the yeast PPR1 zinc finger can functionally replace the GAL4 zinc finger, but the resulting chimeric protein still binds to GAL4 recognition sites (S. Johnston, pers. commun.). Similarly, binding specificities of the glucocorticoid and thyroid receptors can be interchanged by replacing short regions between the two zinc fingers (R. Evans, pers. commun.). Thus, as may be the case for the TFIIIA-like proteins, it is likely that the zinc finger is important for maintaining overall structure and perhaps non-specific interactions to DNA, whereas distinct sequences are involved in direct contact to DNA.

The leucine zipper

Recently, a new class of proteins has been identified, and it has been proposed that they utilize a novel motif for DNA binding, the leucine zipper[14]. These proteins, which include the yeast GCN4 transcriptional activator, the jun, fos and myc oncoproteins, and the C/EBP enhancer binding protein, all contain four or five leucine residues that are spaced exactly seven residues

apart and hence could be viewed as being repeated every two turns of an α-helix. In the initial structural model, it was proposed that the leucine residues are important for interdigitating two α-helices, one from each monomer unit, that provide the structural basis for the dimer formation[14] (Fig. 1).

Consistent with this model, GCN4, C/EBP, and jun bind as dimers to their target sites, and fos and jun can form DNA-binding heterodimers[14-18]; in the case of GCN4, a 60-residue region containing the putative leucine zipper is fully competent for dimer formation and specific DNA-binding. Moreover, a synthetic peptide corresponding to the GCN4 leucine zipper forms stable dimers that are essentially 100% α-helical[19]. Some problems with the original leucine zipper model are that the α-helices are parallel[19], and that a variety of mutations of the conserved leucines in GCN4 and fos have little if any functional effect[18] (Sellers and Struhl, unpublished). It has been suggested that the leucine zipper may indeed be a more conventional coiled coil structure[19].

Although the leucine zipper is probably important for dimer formation, it is likely that sequences outside this region are involved in DNA interactions[14]. (1) By analogy with zinc-finger proteins, the various leucine zippers show modest conservation beyond the leucine residues, and the proteins bind different DNA sequences. (2) GCN4 and jun are structurally related[20] and bind essentially identical DNA sequences but are unable to form heterodimers[21], suggesting that the residues involved in direct contacts to DNA are located within an adjacent 30-residue stretch that is most highly conserved. This adjacent region is also found in fos, which binds to the identical DNA sequence when complexed with jun as a heterodimer. (3) Leucine-zipper motifs have recently been observed in proteins that also contain zinc fingers or homeodomains; perhaps the leucine zipper acts as an independent dimerization region for different kinds of DNA-binding domains.

Related families of proteins recognizing similar DNA sequences

Although different eukaryotic DNA-binding proteins containing a particular structural motif often recognize unrelated DNA sequences, one fundamental aspect of eukaryotic organisms is that they contain families of proteins that interact with similar DNA sequences. In contrast, prokaryotic helix-turn-helix proteins generally have unique DNA sequence recognition properties; i.e. each protein recognizes a distinct set of target sites (the binding of the bacteriophage λ repressor and cro proteins to common operators is a notable exception). Related protein families have been identified for all the structural classes. Six different *Drosophila* homeodomains bind to a common AT-rich consensus sequence even though their protein sequences vary, and the mammalian OCT-1, OCT-2 and Pit-1 proteins (which have similar homeodomains and an additional POU domain[22]) recognize related sites[3]. For the zinc-finger proteins, the steroid receptors show considerable amino acid sequence homology and they bind related DNA sequences[11]. The retinoic acid and thyroid receptors have nearly identical specificities, and the consensus sequences for binding by the glucocorticoid receptor are similar at six out of eight positions. In the case of the putative leucine-zipper motif, mammalian cells have a variety of proteins (jun, junB, junD, fos, fra and others) that interact with a common sequence usually known as an AP-1 site[16-18]. Moreover, multiple leucine-zipper proteins recognizing a common target sequence are found even in a unicellular eukaryote; e.g. the yeast GCN4, AP-1 and probably other proteins[23,24].

The existence of multiple proteins that recognize related sequences increases the precision and flexibility for coordinately and independently regulating genes, particularly those involved in processes of fundamental importance such as cell growth. For example, the bacteriophage λ repressor and cro proteins recognize similar but not identical sequences that control the developmental decision between lysis and lysogeny[25]. In multicellular organisms, the related protein families might provide a mechanism for controlling a core group of genes in a variety of cell types and developmental stages (e.g. the *Drosophila* homeodomain proteins) or in response to extracellular signals (e.g. the steroid receptors). Precision and flexibility could be achieved by subtle differences in sequence recognition properties or DNA-binding affinities such that individual promoters would be strongly affected by particular proteins (or combination of proteins). In this regard, the dimeric nature of the leucine-zipper proteins (and probably the zinc-finger and helix-turn-helix proteins) suggests an additional mode of flexibility involving heterodimers between different proteins. Such heterodimers might yield new proteins with distinct recognition properties, or they might make it possible to influence gene expression only when two specific physiological conditions occur.

Evolutionary conservation of eukaryotic DNA-binding proteins

It is now clear that the mechanism of transcription is remarkably conserved throughout the eukaryotic kingdom. Yeast upstream activator proteins function in a variety of eukaryotic organisms, vertebrate transcription factors function in yeast cells, and the yeast and mammalian TATA factors are functionally interchangeable for transcription *in vitro*[26–28]. Such functional conservation undoubtedly indicates that the basic mechanism of transcriptional initiation has existed since the first eukaryotic organisms.

In addition to the conservation of the basic transcriptional activation mechanism, eukaryotic cells from yeast to human contain structurally similar and functionally analogous transcription factors that recognize essentially identical sequences. Some early examples include the yeast GCN4 protein and the vertebrate jun oncoprotein[20,21], the yeast PHO2 and *Drosophila* engrailed protein[29], the human CCAAT-binding protein CP-1 and the yeast HAP2,3 proteins[30]. Moreover, both CP-1 and HAP2,3 bind as heteromeric complexes involving two distinct proteins, and the yeast and human subunits can be functionally exchanged[30]. It seems likely that these examples represent the tip of the iceberg, and that striking similarities throughout the eukaryotic kingdom will soon become the rule rather than the exception.

Although eukaryotic cells have related transcription factors that recognize similar DNA sequences, the homologues often perform different functions in their respective organisms. For example, GCN4 and HAP2,3 activate the amino acid biosynthetic and oxygen regulated genes respectively in yeast, whereas their evolutionary counterparts, jun and CP-1, activate a variety of genes whose functions appear unrelated. This phenomenon may reflect the fact that for transcription factors that bind to multiple promoters, it would be difficult to alter the sequence recognition properties of the regulatory protein at any point in evolution without

affecting the transcription of many genes. However, the functions of such transcription factors could diverge in individual organisms. By analogy, the genetic code is essentially universal and eukaryotic TATA elements and prokaryotic −10 sequences are similar even though transcriptional and translational mechanisms are quite different.

Role of different structural motifs in transcriptional regulation

Eukaryotic transcription factors contain at least two essential functions, a DNA-binding domain and a transcriptional activation region, that are generally located in separate parts of the protein. Yeast transcriptional activation functions are defined by short acidic regions with minimal primary sequence-homology regions (reviewed in Refs 31, 32). Different acidic regions of GCN4 and GAL4 are equally capable of activating transcription even though their primary sequences are dissimilar. Moreover, acidic regions are observed in other yeast activator proteins, acidic character is the common feature of functional transcriptional activation regions selected from short *E. coli* DNA segments, and amino acid substitutions that alter the level of activation are usually associated with a change in net negative charge. Despite the importance of acidic character, functional transcriptional activation regions have additional structural features beyond negative charge. High resolution deletion analysis and proteolytic mapping suggest that activation regions are repeating structures composed of small units that might be α-helices with amphipathic character. It has been proposed that acidic activation regions are surfaces used for interactions with other proteins of the transcription machinery.

A crucial feature of acidic activation regions is that they function autonomously when fused to different DNA-binding domains (reviewed in Refs 31, 32). For example, functional transcriptional activator proteins can be produced by fusing the DNA-binding domain of the *E. coli* LexA repressor (a helix-turn-helix protein) to the activation regions of proteins containing zinc fingers (GAL4, HAP1), leucine zippers (GCN4, jun, fos, myc), and homeodomains (bicoid). Moreover, the distance and orientation of the GCN4 and GAL4 activation regions with respect to their DNA-binding domains is functionally unimportant.

Acidic regions of DNA-binding proteins are likely to be commonly employed for transcriptional activation in all eukaryotic organisms. GAL4 activates transcription from appropriate target promoters in mammalian cells, and the acidic activation region is required[33,34]. Conversely, transcriptional activation in yeast cells by the jun oncoprotein and the glucocorticoid receptor requires acidic sequences in addition to the DNA-binding domains[35,36]. Thus, acidic activation regions probably contact some part of the transcription machinery that is conserved functionally throughout the eukaryotic kingdom such as a TATA-binding protein or RNA polymerase II.

In this view, the DNA-binding domain has two important roles in transcriptional regulation. (1) It brings the protein to the DNA such that it can easily interact (through the activation region) with other components to form a functional transcriptional initiation complex. (2) The high specificity of the protein–DNA interactions provides the major mechanism by which genes are differentially expressed. For either of these roles, the distinct structural motifs are equivalent. Thus, the helix-turn-helix, zinc-finger, and leucine-zipper motifs utilized for recognizing specific DNA sequences represent different structural solutions to a common function. Did these structural motifs arise independently during evolution, or does one of them represent the primordial DNA-binding domain?

References

1 Pabo, C. O. and Sauer, R. T. (1984) *Annu. Rev. Biochem.* 53, 293–321
2 Wharton, R. P. and Ptashne, M. (1985) *Nature* 316, 601–605
3 Levine, M. and Hoey, T. (1988) *Cell* 55, 537–540
4 Scott, M. P., Tamkun, J. W. and Hartzell III, G. W. *Biochim. Biophys. Acta Reviews on Cancer* (in press)
5 Sauer, R. T., Smith, D. L. and Johnson, A. D. (1988) *Genes Dev.* 2, 807–816
6 Porter, S. D. and Smith, M. (1986) *Nature* 320, 766–768
7 Miller, J., McLachlan, A. D. and Klug, A. (1985) *EMBO J.* 4, 1609–1614
8 Berg, J. (1988) *Proc. Natl Acad. Sci. USA* 85, 99–102
9 Vrana, K. E., Churchill, M. E. A., Tullius, T. D. and Brown, D. D. (1988) *Mol. Cell. Biol.* 8, 1684–1696
10 Pieler, T., Hamm, J. and Roeder, R. G. (1987) *Cell* 48, 91–100
11 Evans, R. M. and Hollenberg, S. M. (1988) *Cell* 52, 1–3
12 Parraga, G., Horvath, S. J., Eisen, A., Taylor, W. E., Hood, L., Young, E. T. and Klevit, R. E. (1988) *Science* 241, 1489–1492
13 Johnston, M. (1987) *Nature* 328, 353–355
14 Landschulz, W. H., Johnson, P. F. and McKnight, S. L. (1988) *Science* 240, 1759–1764
15 Hope, I. A. and Struhl, K. (1987) *EMBO J.* 6, 2781–2784
16 Nakabeppu, Y., Ryder, K. and Nathans, D. (1988) *Cell* 55, 907–915
17 Halazonetis, T. D., Georgopoulos, K., Greenberg, M. E. and Leder, P. (1988) *Cell* 55, 917–924
18 Kouzarides, T. and Ziff, E. (1988) *Nature* 336, 646–651
19 O'Shea, E. K., Rutkowski, R. and Kim, P. (1989) *Science* 243, 538–542
20 Vogt, P. K., Bos, T. J. and Doolittle, R. F. (1987) *Proc. Natl Acad. Sci. USA* 84, 3316–3319
21 Struhl, K. (1987) *Cell* 50, 841–846
22 Herr, W., Storm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. and Horvitz, H. R. (1988) *Genes Dev.* 2, 1513–1516
23 Harshman, K. D., Moye-Rowley, W. S. and Parker, C. S. (1988) *Cell* 53, 321–330
24 Jones, R. H., Moreno, S., Nurse, P. and Jones, N. C. (1988) *Cell* 53, 659–667
25 Johnson, A. D., Poteete, A. R., Lauer, G., Sauer, R. T., Ackers, G. R. and Ptashne, M. (1981) *Nature* 294, 217–223
26 Guarente, L. (1988) *Cell* 52, 303–305
27 Buratowski, S., Hahn, S., Sharp, P. A. and Guarente, L. (1988) *Nature* 334, 37–42
28 Cavallini, B., Huet, J., Plassat, J-L., Sentenac, A., Egly, J-M. and Chambon, P. (1988) *Nature* 334, 77–80
29 Burglin, T. R. (1988) *Cell* 53, 339–340
30 Chodosh, L. A., Olesen, J., Hahn, S., Baldwin, A. S., Guarente, L. and Sharp, P. A. (1988) *Cell* 53, 25–35
31 Struhl, K. (1987) *Cell* 49, 295–297
32 Ptashne, M. (1988) *Nature* 335, 683–689
33 Kakidani, H. and Ptashne, M. (1988) *Cell* 52, 161–167
34 Webster, N., Jin, J. R., Green, S., Hollis, M. and Chambon, P. (1988) *Cell* 52, 169–178
35 Struhl, K. (1988) *Nature* 332, 649–650
36 Schena, M. and Yamamoto, K. R. (1988) *Science* 241, 965–967