STRUCTURAL AND FUNCTIONAL ANALYSIS OF THE DNA-BINDING DOMAIN OF YEAST GCN4 PROTEIN

Kevin Struhl
Department of Biological Chemistry and Molecular Pharmacology
Harvard Medical School, Boston, MA 02115

GCN4, the yeast homologue of the AP-1 transcription factor family that includes the Jun and Fos oncoproteins, binds to the promoters of many amino acid biosynthetic genes and activates their transcription. The C-terminal 56 amino acids are sufficient for dimerization and specific binding to the dyad-symmetric sequence ATGA(C/G)TCAT. GCN4 interacts with non-equivalent and possibly overlapping half-sites (ATGAC and ATGAG) that have different affinities. The optimal half-site for binding is ATGAC, and the protein is surprisingly flexible because it can accommodate the insertion of a single base pair in the center of its compact binding site. The GCN4 DNA-binding domain contains a canonical leucine zipper, a coiled-coil dimerization element, and an adjacent basic that mediates specific DNA-binding. In accord with the scissors-grip hypothesis that the zipper symmetrically positions the two basic regions for specific DNA-binding to adjacent half-sites, a surprisingly wide variety of 7 amino acid sequences can be inserted between the leucine zipper and basic region without significantly impairing DNA binding. However, the invariant asparagine in the basic region, proposed to form an N-cap that breaks the α-helix, is not essential. An unusual feature of the GCN4 DNA-binding domain is that it undergoes a global folding transition upon interaction with DNA. The basic region, which is significantly unfolded in the absence of DNA, becomes almost completely α-helical, when bound to DNA, presumably reflecting a stabilization in protein conformation.

Gene Expression, The Second International Mochida Memorial Symposium,
Tokyo, September 25-26, 1990.

INTRODUCTION

At least 40 yeast genes from a wide variety of amino acid biosynthetic pathways are coordinately activated in response to conditions in which the synthesis or tRNA charging of any single amino acid is inhibited (reviewed in 1). This coordinate induction is mediated by GCN4, a protein that binds specifically to the promoters of the amino acid biosynthetic genes (2, 3). As a result of a novel translational control mechanism, GCN4 protein is synthesized only when cells are starved for amino acids, thus explaining why the amino acid biosynthetic genes are transcriptionally induced during starvation conditions (1, 4, 5). The general control of amino acid biosynthetic genes represents a mechanism to regulate protein synthesis by controlling the amount of amino acid precursors; hence, it is sensible that GCN4, the crucial regulator, is itself controlled by the translation process. As a consequence of its role in regulating protein synthesis, GCN4 is part of the global mechanism that controls cell growth and the decision to initiate new cell division cycles.

In many respects, GCN4 is a typical eukaryotic transcriptional activator protein. First, it contains functionally distinct and physically separate domains for specific DNA-binding to promoter sequences and for transcriptional activation (2). Second, the protein binds as a dimer to a dyad-symmetric sequence that is highly conserved among binding sites in native yeast promoters (6, 7). Third, GCN4 belongs to the leucine zipper class of eukaryotic transcription factors that is defined by novel structural motifs that mediate dimerization and specific DNA-binding (8). Fourth, GCN4 contains a short acidic activation region that is necessary and sufficient for stimulating transcription by RNA polymerase II (2). Fifth, GCN4 is structurally similar to the Jun oncoprotein (9), and both proteins recognize the same DNA sequences from which they activate transcription in yeast cells (10, 11); thus GCN4 is a member of the eukaryotic AP-1 transcription factor family (12). This paper will review work in this laboratory regarding the structural and functional analysis of the GCN4 DNA-binding domain, specifically in regard to dimerization and recognition of specific target sequences.

NATURE OF THE GCN4 RECOGNITION SITE

The GCN4 recognition sequence has been studied by saturation mutagenesis of the binding site in the wild-type *his3* promoter (6) and by selection of binding sites from random-sequence oligonucleotides (13). Both approaches indicate that a 9-bp dyad-symmetric sequence, ATGA(C/G)TCAT, is optimal for DNA-binding and that the central 7 base pairs are most important. The DNA sequence requirements for GCN4 binding *in vitro* and for transcriptional induction *in vivo* appear indistinguishable (6). The optimal GCN4 recognition sequence strongly resembles the consensus of binding sites from GCN4-regulated promoters, bit interestingly, none of the natural sites are identical to the consensus (6). Thus, yeast has evolved a coordinate regulatory system in which the individual promoters contain good, but not optimal binding sites. Presumably, this permits GCN4 to interact efficiently with a wid variety of sequences, thus allowing for regulatory and evolutionary flexibility.

The dyad-symmetric recognition site is recognized by a GCN4 dimer (7) indicating that the complex consists of two protein monomers interacting with adjacent DNA half-sites. However, the nine bp GCN4 binding site is unusually short with the crucial positions being contiguous and within a single turn of the DNA helix. In contrast, the critical residues required for binding by many bacterial repressor and activator proteins are located in two non-contiguous 4-5 bp regions in adjacent helical turns of the DNA. In these cases, the two monomer:half-site interactions are separated by a central region of highly variable sequence that is not in direct contact with the protein.

The highly compact nature of the target sequence as well as other observations suggest that GCN4 dimers bind to overlapping and nonequivalent half-sites. The optimal binding site is inherently asymmetric because it contains an odd number of base pairs and because mutation of the central C:G base pair reduces specific DNA-binding (6). Moreover, the GCN4 binding sites selected from random-sequence oligonucleotides show distinct sequence preferences at symmetrically equivalent positions (13). These observations are most consistent with the view that the central C:G base pair is specifically recognized by GCN4 and hence part of both half-sites. The overlap at the central C:G base pair indicates that the adjacent half-sites in the optimal recognition sequence have distinct DNA sequences, ATGAC and ATGAG, with different affinities.

The contributions of the individual half-sites were determined by analyzing symmetrical derivatives of the optimal binding sequence that delete or insert a single base pair at the center of the site (14). GCN4 binds efficiently to the sequence ATGACGTCAT, but it fails to bind ATGAGCTCAT or ATGATCAT. GCN4 therefore recognizes the central base pair, and the optimal half-site for GCN4 binding is ATGAC, not ATGAG. When GCN4 interacts with the optimal 9-bp target sequence, the left half-site (ATGAC) contributes more to the overall affinity than the right half-site (ATGAG), because the GCN4 monomer interacting with the left-half site presumably contacts the central base pair, whereas the monomer interacting with the right half-site does not. Since alterations in the right half-site are tolerated better than symmetrically equivalent alterations in the left half-site (13), GCN4 prefers to bind a sequence with one optimal and one weak half-site rather than a sequence with two moderate half-sites; this probably reflects cooperative binding to adjacent half-sites.

GCN4 is a surprisingly flexible protein because it can accommodate a major structural disruption, the insertion of a single base pair, in the center of its otherwise compact binding site. Although many DNA-binding proteins are highly sensitive to spacing changes in the target site, some proteins tolerate or even prefer different spacings between half-sites (15-17). However, in all these cases of flexibility, the sequence at the center of the binding site is relatively unimportant, and the dimerization region lies within a distinct structural domain from the region needed for DNA contacts. Thus it is very likely that the DNA interaction surfaces of the two monomers are structurally independent. In contrast, the dimerization and DNA-binding functions of GCN4 are localized to the 60 C-terminal residues (2, 7), a region that appears, by proteolytic mapping, to be a single structural domain (18).

The ATGACGTCAT sequence recognized by GCN4 strongly resembles sites bound by the yeast and mammalian ATF/CREB family of proteins (19, 20). Like GCN4, these proteins bind as dimers, and they contain leucine zipper motifs and adjacent basic regions (21-23); thus, it is extremely likely that the ATF/CREB family recognizes adjacent ATGAC half-sites. This suggests that GCN4 and the ATF/CREB protein family recognize similar half-sites, but have different spacing requirements. In this regard, the mammalian AP-1 protein family, which recognizes the same sequences as GCN4 (10, 24), is immunologically related to the ATF/CREB protein family (20). Thus, the GCN4/AP-1 and ATF/CREB classes of proteins likely belong to the same evolutionarily conserved superfamily of proteins that recognize essentially identical half-sites. A precedent for members of a protein superfamily that recognize similar half-sites with distinct spatial constraints has been suggested to explain the DNA-binding properties of the estrogen and thyroid hormone receptors (25, 26).

NATURE OF THE GCN4 DNA-BINDING DOMAIN

Extensive deletion analysis of the 281 amino acid GCN4 protein indicates that the 56 C-terminal amino acids are sufficient both for dimerization and for specific DNA-binding (2, 7, 27). The DNA-binding domain can be isolated from the full-length protein as a proteolytically stable fragment, indicating that it folds independently of the remainder of the protein (18). Moreover, GCN4 and the Jun oncoprotein bind the same DNA sequences (10), although amino acid sequence conservation between these proteins is restricted to the 65 C-terminal residues (9).

The GCN4 DNA-binding domain contains a leucine zipper, a conserved structural motif found in a class of eukaryotic transcription factors that includes C/EBP and the Jun and Fos oncoproteins (8). The leucine zipper consists of four or five leucine residues (GCN4 has four) spaced exactly seven amino acids apart embedded within a region whose sequence is consistent with the formation of an amphipathic α-helix. Adjacent to the leucine zipper is a conserved region that is rich in basic residues and also includes an invariant asparagine. The spacing between the leucine zipper and adjacent basic region is absolutely maintained in this family of DNA-binding proteins.

Distinct sub-domains for dimerization and DNA-binding. In the initial structural model, it was proposed that the leucine zipper provides the dimerization function by virtue of interdigitated α-helices from each monomer, and that it properly positions the adjacent basic regions for specific contacts to the adjacent DNA half-sites (8). A more recent elaboration, the scissors grip model, imagines that the entire DNA-binding domain is largely α-helical when associated with its target site, and that the invariant asparagine forms an N-cap that breaks the α-helix within the basic region and permits its reorientation with respect to the major groove of the DNA (28).

Chimeric proteins have been used to prove that the GCN4 leucine zipper confers the specific dimerization properties of the intact protein and that the adjacent basic region is sufficient for specific DNA-binding. The basis of such experiments is that the various leucine zipper proteins have distinct dimerization and DNA-binding properties despite having common sequence motifs. For dimerization, GCN4, Jun, and Fos contain the conserved leucines in the zipper motif and interact with the same DNA sites, yet the only functional species are GCN4 homodimers, Jun homodimers, and Fos-Jun heterodimers; Fos homodimers, Fos-GCN4 heterodimers and GCN4-Jun heterodimers can not be formed (12, 29, 30). However, precise replacement of the Fos zipper by the GCN4 zipper generates a Fos-GCN4 chimera with GCN4 dimerization specificity; it binds DNA as a homodimer or as a heterodimer with GCN4, but not as a heterodimer with Jun (29, 30). Conversely, GCN4 and C/EBP recognize different DNA sequences, and analysis of similar zipper-basic region chimeric proteins indicate that DNA-binding specificity tracks with the basic region (31). The fact that leucine zipper and basic regions are can be interchanged between different family members to generate chimeric proteins with predicted dimerization and DNA-binding specificities indicates that these conserved motifs encode distinct structural sub-domains.

The leucine zipper. The original structural concept of the leucine zipper invoked an α-helical dimer formed primarily by interdigitation of leucine residues within the hydrophobic interface (8). In support of this idea, a GCN4 leucine zipper peptide (the 33 C-terminal residues) forms stable α-helical dimers in solution (32), and the same region exists as a dimeric α-helical structure in the functional DNA-binding domain (27). However, in contrast to the prediction of the initial interdigitation model, the α-helices associate in the parallel rather than anti-parallel arrangement (32). In addition, since the canonical leucine residues are common to all zipper proteins, non-conserved residues in the various zipper regions must have critical roles in generating distinct dimerization specificities and hence zipper association properties. From these observations, it was proposed that the leucine zipper is similar to the coiled coil structure found in muscle filament proteins (32). In the coiled coil, the dimerization interface is not formed by leucine interdigitation, but rather by pairwise interaction of the leucines with hydrophobic residues predicted to lie on the same side of the α-helix.

Another observation more consistent with the coiled coil model is that the GCN4 leucine zipper is surprisingly tolerant of mutations in the leucine residues (Pu et al, unpublished). 19 of 20 single substitution proteins are functionally indistinguishable from wild-type GCN4 when assayed for DNA-binding *in vitro* and transcriptional activation *in vivo*. Each of the 4 leucine residues can be changed, and a wide variety of substitutions are permitted including basic (arg267 and arg274) and acidic (glu260) amino acids. The sole exception, gly267, displays a reduced but clearly detectable level of function, probably a consequence of the α-helix destabilizing nature of glycine residues; interestingly, gly267 forms fully functional DNA-binding heterodimers with wild-type GCN4. GCN4 derivatives containing two leucine substitutions display low or no detectable function *in vivo*, but most of these tested bind DNA weakly as homodimers and strongly as heterodimers with wild-type GCN4. The observations do not imply that the leucines are functionally unimportant, but rather indicate that numerous other interactions within the coiled coil are crucial for efficient dimerization.

The estimated lifetime of GCN4-p dimers at 25 ^{o}C is between 10 ms and 1 s, based on the exchange properties of NMR resonances assigned to the leucine zipper region. In conjunction with the modest dissociation constant of GCN4-p for DNA, these observations suggest that unfolding and reassembly of GCN4, and other transcription factors utilizing a coiled-coil dimerization interface, may occur without significant kinetic barriers, thereby facilitating subunit exchange.

DNA Binding. The suggestion that the leucine zipper correctly positions the basic region for specific DNA-binding was based on the precisely conserved spacing relationship between these two sub-domains (8, 28). In support of this idea, disruption of this spacing by insertion of two, four, or six amino acids between the GCN4 leucine zipper and basic region abolishes GCN4 function (31; Pu et al, submitted). More convincingly, a surprisingly wide variety of seven amino acid sequences results in proteins displaying weak to wild-type levels of GCN4 activity (Pu et al, submitted). Thus, the correct spatial relationship is retained upon the insertion of an integral number of α-helical turns (7 residues) between the zipper and basic region. Interestingly, heterodimers between GCN4 and heptapeptide insertion proteins fail to bind DNA; i.e. both proteins contain an acceptable spacing between the zipper and basic region, but the spacings are not mutually compatible. These results strongly suggest that the leucine zipper symmetrically orients the two basic regions along the adjacent half-sites and that the region between the two sub-domains in α-helical. Besides being consistent with predictions of the scissors grip model (28), these observations strongly suggests that GCN4 homodimers are the primary, and possibly the sole, mediators of GCN4 function in yeast cells.

Mutational analysis of the invariant asparagine in the basic region of GCN4 strongly argues against the model that this residue forms an N-cap structure that breaks the putative α-helix thereby allowing both halves of the basic region to interact with the major groove of the target site. Although most mutations of the asparagine 235 codon abolish GCN4 function, activity was observed with either the tryptophan or glutamine substitutions. In comparison to wild-type GCN4, the activity of the gln235 protein is reduced, but most unexpectedly, the trp235 protein appears to function more efficiently in the standard complementation assay. This increased function of the trp235 protein can also be seen by its ability to activate transcription from a *his3* promoter containing a single weak GCN4 target site (TTGACTCAA) that is unresponsive to wild-type GCN4 (Tzamarias, D., W.P., and K.S., unpublished data). Tryptophan and glutamine are strongly disfavored in the N-cap position of naturally occuring α-helices, and glutamine and asparagine are very non-conservative replacements in helices and at helix ends (33).

Highly conserved features of protein families are generally presumed to be fundamentally important for function and often serve as the principal basis for proposing structural models. However, many highly conserved features of leucine zipper proteins are not essential for GCN4 DNA-binding. The spacing between the zipper and basic region can be altered by inserting an integral number of helical turns, the invariant asparagine can be changed and indeed may not even be optimal, and the canonical leucine residues in the zipper can be varied considerably. Thus, it seems very likely that there are eukaryotic transcriptional regulatory factors that lack some or many of the defining characteristics of leucine zipper proteins, yet nevertheless are structurally and functionally homologous.

GCN4 undergoes a global folding transition upon specific DNA binding. DNA-binding domains such as the helix-turn-helix and the zinc finger are pre-folded structures that dock against the DNA double helix by virtue of complementary surfaces. In striking contrast, the GCN4 DNA-binding domain undergoes a global folding transition upon specific interaction with DNA (27). In the absence of DNA, the dimeric DNA-binding domain is approximately 70% α-helical at 25°C as determined by circular dichroism. This α-helicity is due to the leucine zipper, thereby implying that the adjoining basic region is largely unstructured in the absence of DNA. Strikingly, addition of a GCN4 binding site increases the α-helix content to at least 95%, indicating that the basic region acquires substantial α-helical structure when it specifically binds to DNA. The almost completely α-helical nature of GCN4 in the protein-DNA complex is consistent with, but not specific to, the scissors grip model (28).

Although the basic region is largely unstructured in the absence of DNA, the α-helical content of the GCN4 DNA-binding domain increases to about 80% when the temperature is lowered. This partial α-helical transition is observed with a 26 residue peptide corresponding to the basic region, suggesting that these conformations are locally determined and not dependent on the adjoining zipper. These observations suggest that in the absence of DNA, the basic region of GCN4 exists as an ensemble of structures, with the folded state being significantly populated only at low temperature. More importantly, specific target sequences stabilize the α-helical conformation of the basic region, thus inducing the fit between protein and DNA.

In the protein-DNA complex, GCN4 is structurally quite rigid due to its almost completely α-helical nature. However, the protein undergoes the same global folding transition when bound to the ATF/CREB site that contains an additional base pair between the adjacent half-sites, suggesting some degree of flexibility in the protein-DNA complex. Cirular dichroism spectra of the different target DNAs change only slightly and in a similar manner in the presence of GCN4, thus suggesting that the alternative half-site spacings are accomodated by flexibility in the protein rather than by major structural rearrangements in the DNA (27). The most likely structural basis for this flexibility is at or near the bifurcation where the helices of the two basic regions split off from the dimeric coiled-coil of the leucine zipper.

REFERENCES

1. Hinnebusch, A. G. (1988) *Microbiol. Rev.* **52,** 248-273.

2. Hope, I. A. & Struhl, K. (1986) *Cell* **46,** 885-894.

3. Arndt, K. & Fink, G. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83,** 8516-8520.

4. Thireos, G., Penn, M. D. & Greer, H. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81,** 5096-5100.

5. Hinnebusch, A. G. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81,** 6442-6446.

6. Hill, D. E., Hope, I. A., Macke, J. P. & Struhl, K. (1986) *Science* **234,** 451-457.

7. Hope, I. A. & Struhl, K. (1987) *EMBO J.* **6,** 2781-2784.

8. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1988) *Science* **240,** 1759-1764.

9. Vogt, P. K., Bos, T. J. & Doolittle, R. F. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84,** 3316-3319.

10. Struhl, K. (1987) *Cell* **50,** 841-846.

11. Struhl, K. (1988) *Nature* **332,** 649-650.

12. Curran, T. & Franza, B. J. (1988) *Cell* **55,** 395-397.

13. Oliphant, A. R., Brandl, C. J. & Struhl, K. (1989) *Mol. Cell. Biol.* **9,** 2944-2949.

14. Sellers, J. W., Vincent, A. C. & Struhl, K. (1990) *Mol. Cell. Biol.* **10,** 5077-5086.

15. Sadler, J. R., Sasmor, H. & Betz, J. L. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80,** 6785-6789.

16. Falvey, E. & Grindley, N. D. F. (1987) *EMBO J.* **6,** 815-821.

17. Sauer, R. T., Smith, D. L. & Johnson, A. D. (1988) *Genes and Devel.* **2,** 807-816.

18. Hope, I. A., Mahadevan, S. & Struhl, K. (1988) *Nature* **333,** 635-640.

19. Roesler, W. J., Vandenbark, G. R. & Hanson, R. W. (1988) *J. Biol. Chem.* **263,** 9063-9066.

20. Hai, T., Liu, F., Allegretto, E. A., Karin, M. & Green, M. R. (1988) *Genes Dev.* **2,** 1216-1226.

21. Hoeffler, J. P., Meyer, T. E., Yun, Y., Jameson, J. L. & Haebner, J. F. (1988) *Science* **242,** 1430-1433.

22. Gonzalez, G. A., Yamamoto, K. K., Fischer, W. H., Karr, D., Menzel, P., Biggs, W., Vale, W. W. & Montminy, M. R. (1989) *Nature* **337,** 749-752.

23. Hai, T., Liu, F., Coukos, W. J. & Green, M. R. (1989) *Genes Dev.* **3,** 2083-2090.

24. Bohmann, D., Bos, T. J., Admon, A., Nishimura, T., Vogt, P. K. & Tjian, R. (1987) *Science* **238,** 1386-1392.

25. Glass, C. K., Holloway, J. M., Devary, O. V. & Rosenfeld, M. G. (1988) *Cell* **54,** 313-323.

26. Umesono, K. & Evans, R. M. (1989) *Cell* **57,** 1139-1146.

27. Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C. & Struhl, K. (1990) *Nature* in press.

28. Vinson, C. R., Sigler, P. B. & McKnight, S. L. (1989) *Science* **246,** 911-916.

29. Kouzarides, T. & Ziff, E. (1989) *Nature (Lond).* **340,** 568-571.

30. Sellers, J. W. & Struhl, K. (1989) *Nature* **341,** 74-76.

31. Agre, P., Johnson, P. F. & McKnight, S. L. (1989) *Science* **246,** 922-926.

32. O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1989) *Science* **243,** 538-542.

33. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648-1652.