Cell, Vol. 65, 717–719, May 31, 1991, Copyright © 1991 by Cell Press

Modular Structure of Transcription Factors: Implications for Gene Regulation

Minireview

Alan D. Frankel* and Peter S. Kim*†
*Whitehead Institute for Biomedical Research
Cambridge, Massachusetts 02142
†Howard Hughes Medical Institute
Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142

Ever since the first three-dimensional structure of a protein, myoglobin, was determined, proteins have generally been thought of as rigid, globular structures composed largely of well-defined units of secondary structure. The structures of enzyme active sites and antibody-binding pockets illustrate how proteins can contain well-ordered ligand-binding sites. Recent structural studies of eukaryotic transcription factors indicate that the control of transcription also involves proteins that are well ordered. There are, however, other studies that reveal some surprising features of transcription factor structure and organization, causing us to rethink some simple concepts of protein structure–function relationships. In all cases, the use of structural modules in transcription proteins is becoming clear. Here we describe studies that indicate the following: first, modules can adopt rigid structures; second, modules can undergo disorder-to-order transitions; and finally, they can interact with other modules to enhance specificity. We suggest reasons why these characteristics of modules are important for the function of eukaryotic transcription complexes.

Transcription Factor Modules

The modular organization of eukaryotic transcription factors was discovered by Brent and Ptashne (1985) in "domain-swap" experiments with LexA and GAL4. In these experiments, the DNA-binding domain of LexA was fused to the activation domain of GAL4, resulting in a transcriptional activator that operated through a LexA binding site. The remarkably modular nature of transcription factors has been confirmed in many other systems. In the case of the estrogen and glucocorticoid nuclear receptors, modules can be interchanged to switch DNA-binding, ligand-binding, and activation functions (see Hollenberg and Evans, 1988, and references therein). The precise positioning of these modules within the hybrid proteins is highly flexible, suggesting that each module represents an independent structural domain. While it is not surprising that proteins as large as GAL4 or the receptors (which have more than 600 amino acids) contain multiple structural domains, it is remarkable that the domains can be mixed and matched with such flexibility. The flexibility is further illustrated by the fact that modules can either be covalently attached to each other or interact with each other through intermediary proteins (see Liu and Green, 1990).

Structured Modules

Small modular domains from eukaryotic transcription factors can adopt highly ordered conformations, consistent with previous ideas about protein structure. The structure of the DNA-binding domain of the glucocorticoid and estrogen receptors (66 amino acids) has been solved by two-dimensional NMR (Hard et al., 1990; Schwabe et al., 1990). This domain contains two α helices that are stabilized by the binding of two zinc ions to cysteine side chains. Model building suggests that one α helix from each monomer of the receptor dimer interacts specifically with the major groove of DNA.

The structure of the TFIIIA-like zinc finger (30 amino acids) has been solved by NMR (Lee et al., 1989; Klevit et al., 1990), and a complex with DNA has been solved by X-ray crystallography (Pavletich and Pabo, 1991). The domain is a two-stranded β sheet and an α helix with one zinc ion coordinated to two cysteine and two histidine side chains. Side chains from the α helix provide specific DNA contacts.

The structure of a third DNA-binding domain, the homeodomain (61 amino acids), has been solved by two-dimensional NMR (Otting et al., 1990) and X-ray crystallography (Kissinger et al., 1990) and reveals a unit of three α helices that interacts with DNA in a manner somewhat similar to bacterial helix-turn-helix proteins.

In addition to nucleic acid-binding domains, the structure of the leucine zipper dimerization domain of GCN4 (33 amino acids) has been shown by X-ray scattering (Rasmussen et al., 1991) to be a two-stranded, parallel α-helical coiled coil. This domain contains all the information required to mediate specific homodimer and heterodimer formation (O'Shea et al., 1989).

Induced Structure

In contrast to highly ordered modules, other transcription factor domains are not so highly ordered on their own but appear to become structured only upon interaction with other molecules. This may have been suggested initially by the finding that the function of transcriptional activation domains does not require a well-defined amino acid sequence. By selecting from random sequences, a high density of negative charges, rather than the specific protein sequence, was shown to be the major determinant of an activating region (Ma and Ptashne, 1987). Based on their sequence flexibility, it was suggested that acidic regions may be unstructured "negative noodles" that become structured only upon interaction with some part of the transcription apparatus (Sigler, 1988). Although the interacting partner of activation domains has not yet been conclusively identified, recent studies suggest that the acidic region of VP16 may interact with TFIIB (Lin and Green, 1991) or with TFIID (Stringer et al., 1990) and may allow the "negative noodle" hypothesis to be tested.

A particularly clear example of induced protein structure, reminiscent of Koshland's induced-fit hypothesis of enzyme–substrate interactions, is seen in the DNA-binding basic region of leucine zipper proteins. For GCN4, this region can be reduced to 31 residues without loss of specific binding activity, provided that a disulfide bond is used in place of the leucine zipper (Talanian et al., 1990). Although the basic regions from leucine zipper proteins

are only partially structured in solution, they can be induced to form an α-helical structure upon specific interaction with DNA (Talanian et al., 1990; O'Neil et al., 1990; Weiss et al., 1990; Patel et al., 1990). Thus, protein folding is coupled to specific DNA recognition.

An analogous example can be found in the HIV Tat protein, which contains an arginine-rich region (9 amino acids). This region is an independent RNA-binding domain that seems to become structured upon RNA binding (Calnan et al., 1991). Like the acidic activating regions, this module of Tat shows substantial sequence flexibility, with a high density of basic residues, rather than the precise amino sequence, being important for specific RNA binding (Calnan et al., 1991). Other domains of transcription complexes are likely to require interactions with other molecules in order to adopt a specific conformation. These candidates include the heptapeptide repeat of the large subunit of RNA polymerase II, which has been proposed to form a β-turn-like structure that intercalates into DNA (Suzuki, 1990), and glutamine-rich and proline-rich transcriptional activation domains (see Mitchell and Tjian, 1989), which may require protein–protein interactions to function.

A critical question is whether the induced fit observed with peptides occurs in the context of the intact protein. Preliminary difference circular dichroism experiments with intact GCN4 suggest that α-helical structure is induced upon DNA binding by the intact protein (O'Neil et al., 1990). For Tat, preliminary measurements indicate that the RNA-binding affinity of the short peptide is similar to that measured with the intact protein (Calnan et al., 1991; Dingwall et al., 1990). These results suggest that, at least in some cases, part of the intact protein remains relatively unstructured until it is bound to DNA or RNA.

If proteins of the transcription apparatus do indeed contain disordered regions for which structure can be induced, then how are these regions protected from proteolytic degradation in the cell? Perhaps localization within the nucleus, which contains few proteases, is sufficient to prevent degradation; alternatively, these unstructured regions may interact with other parts of the same protein, with other proteins, or with nonspecific nucleic acids to escape degradation. If these domains are sensitive to proteases in vivo, then the possibility exists for additional levels of regulation analogous to the cleavage of λ repressor (Sauer et al., 1990).

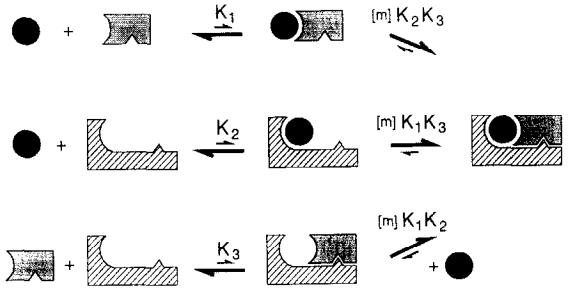
Why do some transcription factor modules undergo induced fit? It seems reasonable that, on their own, some domains will contain exposed amino acids that are destined to interact with adjacent modules or with specific nucleic acid–binding sites. In the absence of such a partner, the structure of a domain may be unstable, perhaps because it has solvent-exposed hydrophobic residues or electrostatically unfavorable neighboring charged residues. Induced structure may also be important for proteins that wrap around DNA or RNA, as seen with the N-terminal arm of λ repressor (Sauer et al., 1990). Although the conformational changes that accompany induced-fit binding will cost energy and lower the affinity, the specificity of the interaction can still be very high.

Low-Specificity Interactions

Another unexpected feature of many transcription factor domains, especially eukaryotic ones, is that they sometimes show only a modest degree of specificity and affinity in their interactions with ligands. For example, λ repressor binds to specific operator sequences with affinities as high as 10^{-13} M and binds to specific DNA 500,000-fold better than to nonspecific DNA (Sauer et al., 1990). In contrast, the binding to DNA by some steroid receptors occurs with nanomolar affinity and with a specificity of less than 100-fold (e.g., Schauer et al., 1989). As noted above, it is possible for interactions to be very specific even if the affinity is low, and it is probably specificity, rather than affinity, that is the key ingredient for assembling functional transcription complexes.

How are transcription factor interactions of only modest specificity tolerated in a eukaryotic cell? A likely answer is that transcription complexes are held together by many interactions. Cooperative, multiple interactions assure that the overall specificity of the transcription complex is high, even if some individual interactions are of low specificity. It would be difficult to dissociate or regulate the activity of a transcription complex if every component had an exceedingly high affinity or specificity for every other component. Furthermore, extremely tight or specific interactions might interfere with the combinatorial use of factors by many promoters. The design of larger, cooperatively interacting complexes may be the price that the eukaryotic cell has to pay for increased flexibility: individual interactions with only modest specificity seem to be inherent to these designs. As in enzyme–substrate reactions and protein folding, cooperativity results largely from entropic factors (Creighton, 1983; see figure).

Ribosomes, which are composed of many types of interactions with varying specificities and affinities, may be a particularly apt analogy to the assembly of transcription complexes. In general, complex macromolecular assemblies seem to be built up of cooperative protein–DNA, protein–RNA, and protein–protein interactions; only when the entire complex is assembled is the true specificity of the system revealed.



Multiple Low-Specificity Interactions Can Lead to Cooperativity in Binding

The equilibrium constant for adding the ball to a binary complex, $[m]K_1K_2$, is larger than the product of the equilibrium constants for adding the ball to each individual component (K_1K_2). This results in large part from the loss in entropy that has already occurred in forming the binary complex. m, cooperativity factor.

Why Modules?

Modular protein domains or structural units are ideally suited for complexes in gene regulation, because they facilitate the design of interacting protein structures that are so important for cooperative and/or allosteric interactions. For example, interactions mediated by the C-terminal domain of λ repressor lead to cooperative changes in DNA binding and gene expression (Sauer et al., 1990), while inducer binding to *lac* repressor provides the classic example of allosteric effects on DNA binding. As an example in a eukaryotic system, transfer of the steroid-binding domain of nuclear receptors to other DNA-binding proteins can confer ligand responsiveness (e.g., see Eilers et al., 1989); ligand binding may affect protein–protein interactions or may induce an allosteric transition in the receptor.

Quaternary structure is also important for cooperative binding reactions, even without allostery. Studies of λ repressor (Sauer et al., 1990) and CAP (Brown and Crothers, 1989) show that protein dimerization can lead to cooperative DNA binding. The potential for regulation of DNA binding by protein homo- and heterodimerization has been emphasized in the many studies of leucine zipper and helix-loop-helix proteins (see Jones, 1990). Concentration gradients of interacting transcription factors, such as those established in embryos, could lead to cooperative patterns of DNA binding and the formation of sharp boundaries of gene expression. The sharpness of the boundary would then depend on the number of interactions and the degree of cooperativity.

Modular structural units permit combinatorial use of factors. The many advantages of combinatorial use of factors have been reviewed recently (Jones, 1990). A particularly elegant example is provided by composite glucocorticoid responsive elements that allow interaction between the glucocorticoid receptor and the transcription factor AP-1. Either repression or activation can result, depending on the levels of receptor, Fos, and Jun present in the cell (Diamond et al., 1990; Schüle et al., 1990; Yang-Yen et al., 1990; Jonat et al., 1990). There is also the well-recognized evolutionary advantage of structural modules, which is borne out by the remarkable number of transcription factors that belong to the leucine zipper and helix-loop-helix classes.

The modular view of transcription complexes is discouraging—full understanding of a regulatory complex may require the simultaneous presence of an entire set of interactions, some of which may be difficult to discern individually because the affinity or specificity is low. The encouraging aspect of the modular view, however, is that many structural questions can be addressed by studying small, discrete domains. Genetic approaches will continue to be instrumental in identifying these domains and their interacting partners. Biophysical approaches, including X-ray crystallography and two-dimensional NMR, can then be combined with bacterial expression or peptide synthesis to undertake structural studies of transcription factor domains. Placing the domains together with their interacting partners into complete views of transcription complexes will be a major and exciting challenge.

References

Brent, R., and Ptashne, M. (1985). Cell *43*, 729–736.

Brown, A. M., and Crothers, D. M. (1989). Proc. Natl. Acad. Sci. USA *86*, 7387–7391.

Calnan, B. J., Biancalana, S., Hudson, D., and Frankel, A. D. (1991). Genes Dev. *5*, 201–210.

Diamond, M. I., Miner, J. N., Yoshinaga, S. K., and Yamamoto, K. R. (1990). Science *249*, 1266–1272.

Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Heaphy, S., Karn, J., Lowe, A. D., Singh, M., and Skinner, M. A. (1990). EMBO J. *9*, 4145–4153.

Eilers, M., Picard, D., Yamamoto, K. R., and Bishop, J. M. (1989). Nature *340*, 66–68.

Härd, T., Kellenbach, E., Boelens, R., Maler, B. A., Dahlman, K., Freedman, L. P., Carlstedt-Duke, J., Yamamoto, K. R., Gustafsson, J.-Å., and Kaptein, R. (1990). Science *249*, 157–160.

Hollenberg, S. M., and Evans, R. M. (1988). Cell *55*, 899–906.

Jonat, G., Rahmsdorf, H. J., Park, K.-K., Cato, A. C. B., Gebel, S., Ponta, H., and Herrlich, P. (1990). Cell *62*, 1189–1204.

Jones, N. (1990). Cell *61*, 9–11.

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990). Cell *63*, 579–590.

Klevit, R. E., Herriott, J. R., and Horvath, S.J. (1990). Proteins *7*, 215–226.

Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A., Wright, P. E. (1989). Science *245*, 635–637.

Lin, Y.-S., and Green, M. R. (1991). Cell *64*, 971–981.

Liu, F., and Green, M. R. (1990). Cell *61*, 1217–1224.

Ma, J., and Ptashne, M. (1987). Cell *51*, 113–119.

Mitchell, P. J., and Tjian, R. (1989). Science *245*, 371–378.

O'Neil, K. T., Hoess, R. H., and DeGrado, W. F. (1990). Science *249*, 774–778.

O'Shea, E. K., Rutkowski, R., Stafford, W. F., III, and Kim, P. S. (1989). Science *245*, 646–648.

Otting, G., Qian, Y. Q., Billeter, M., Müller, M., Affolter, M., Gehring, W. J., and Wüthrich, K. (1990). EMBO J. *9*, 3085–3092.

Patel, L., Abate, C., and Curran, T. (1990). Nature *347*, 572–574.

Pavletich, N. P., and Pabo, C. O. (1991). Science, in press.

Rasmussen, R., Benvegnu, D., O'Shea, E. K., Kim, P. S., and Alber, T. (1991). Proc. Natl. Acad. Sci. USA *88*, 561–564.

Sauer, R. T., Jordan, S. R., and Pabo, C. O. (1990). Adv. Prot. Chem. *40*, 1–61.

Schauer, M., Chalepakis, G., Willman, T., and Beato, M. (1989). Proc. Natl. Acad. Sci. USA *86*, 1123–1127.

Schüle, R., Rangarajan, P., Kliewer, S., Ransone, L. J., Bolado, J., Yang, N., Verma, I. M., and Evans, R. M. (1990). Cell *62*, 1217–1226.

Schwabe, J. W. R., Neuhaus, D., and Rhodes, D. (1990). Nature *348* 458–461.

Sigler, P. B. (1988). Nature *333*, 210–212.

Stringer, K. F., Ingles, C. J., and Greenblatt, J. (1990). Nature *345* 783–786.

Suzuki, M. (1990). Nature *344*, 562–565.

Talanian, R. V., McKnight, C. J., and Kim, P. S. (1990). Science *249*, 769–771.

Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., and Struhl, K. (1990). Nature *347*, 575–578.

Yang-Yen, H.-F., Chambard, J.-C., Sun, Y.-L., Smeal, T., Schmidt, T. J., Drouin, J., and Karin, M. (1990). Cell *62*, 1205–1215.