Cell, Vol. 49, 295–297, May 8, 1987, Copyright © 1987 by Cell Press

Promoters, Activator Proteins, and the Mechanism of Transcriptional Initiation in Yeast

Minireview

Kevin Struhl
Department of Biological Chemistry
Harvard Medical School
Boston, Massachusetts 02115

Initiation and regulation of mRNA transcription in eukaryotes depend on several proteins in addition to RNA polymerase II. The prevailing view is that these proteins, called transcription factors, interact with the multiple sequence elements that form the eukaryotic promoter. Our knowledge of the molecular mechanisms involved in the initiation of transcription has come in part from studies on the yeast Saccharomyces cerevisiae, in which proteins that activate transcription of specific genes in vivo have been characterized both genetically and biochemically.

Yeast Promoter Elements

Yeast promoters are composed of upstream (UAS), TATA, and initiator (I) elements that are necessary for the regulation, amount, and accuracy of transcriptional initiation (Figure 1). Although a simple promoter could contain just one of each kind of element, most native yeast promoters are more complex. In addition, some promoters contain operator (OP) elements that are involved in the repression of transcription (Johnson and Herskowitz, Cell *42*, 237–247, 1985; reviewed by Brent, Cell *42*, 3–4, 1985).

Upstream elements are required for transcription, and they usually determine the particular regulatory properties of a given promoter (reviewed by Guarente, Cell *36*, 799–800, 1984). As expected, genes subject to a common control mechanism contain upstream elements that are similar in DNA sequence. Yeast upstream elements resemble mammalian enhancer sequences; they function in both orientations and at long and variable distances (up to at least 600 bp) with respect to other promoter elements and the mRNA initiation site. Unlike enhancers, they do not appear to activate transcription when located downstream from the initiation site (Guarente and Hoar, PNAS *81*, 7860–7864, 1984; Struhl, PNAS *81*, 7865–7869, 1984). In general, it is believed that upstream elements are DNA-binding sites for transcriptional regulatory proteins, and as will be discussed below, this has been demonstrated in several cases. However, upstream elements for several constitutively expressed genes are poly(dA-dT) homopolymer sequences, and it has been proposed that these act by excluding nucleosomes, not by binding specific proteins (Struhl, PNAS *82*, 8419–8423, 1985).

TATA elements (consensus sequence TATAAA) are necessary but not sufficient for transcriptional initiation of most yeast genes. The distance between yeast TATA elements and mRNA initiation sites ranges between 40–120 bp depending on the promoter; in contrast, higher eukaryotic TATA sequences are almost always located 25–30 bp away from the initiation site (Breathnach and Chambon, Ann. Rev. Biochem. *50*, 349–383, 1981). Because they are present in many different kinds of promoters, TATA elements are presumed to have a general role in the transcription process such as binding of a general transcription factor. However, two functionally distinct classes of TATA elements have been defined (Struhl, MCB *6*, 3847–3853, 1986), and some genes such as *PGK* (phosphoglycerate kinase) are highly expressed in the absence of any TATA-like sequences (Ogden et al., MCB *6*, 4335–4343, 1986). Thus, yeast may contain several different "general" transcription factors that define distinct classes of promoters. Such factors may be analogous to the different σ factors that interact with E. coli RNA polymerase and determine its promoter specificity.

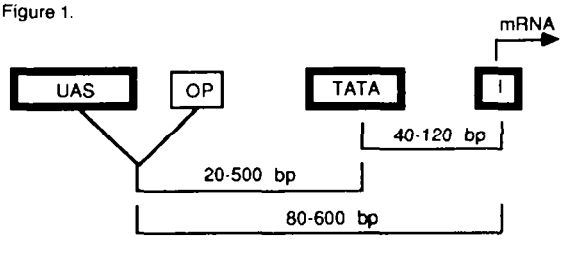
The initiator element, located near the actual mRNA start site, has little effect on the overall RNA level, but it is the primary determinant of where transcription begins. This is in apparent contrast to the situation in higher eukaryotic promoters where initiation sites are determined primarily by the distance from the TATA element. In yeast, accurate initiation is still observed when the distance to the TATA element is varied or when transcription depends on "foreign" TATA elements located at different positions from the natural element (Chen and Struhl, EMBO J. *4*, 3273–3280, 1985; Hahn et al., PNAS *82*, 8562–8566, 1985; Nagawa and Fink, PNAS *82*, 8557–8561, 1985; McNeil and Smith, JMB *187*, 363–378, 1986). Since it is located so close to the initiation site, the initiator element may represent the particular DNA sequences preferred by RNA polymerase II itself or by a closely associated initiation factor such as that described by Burton et al. (EMBO J. *5*, 2923–2930, 1986).

Activator Proteins Bind to Upstream Promoter Elements

Attempts to identify yeast activator proteins have usually begun with mutations that define *trans*-acting factors which induce a specific gene or set of genes. However, biochemical evidence is critical to prove that a mutation directly affects transcription by altering a specific protein–DNA interaction. Specific binding to upstream promoter elements has been demonstrated for the GAL4, GCN4, and HAP1 proteins.

Transcription of the *GAL* (galactose metabolizing) genes, which occurs in galactose but not glucose medium, requires the GAL4 protein. As assayed by DNAaseI footprinting in vitro or methylation protection in vivo, GAL4 binds four sites within the upstream regulatory region required for bidirectional activation of the *GAL1* and *GAL10* genes (Bram and Kornberg, PNAS *82*, 43–47, 1985; Giniger et al., Cell *40*, 767–774, 1985). Sequence comparison of the binding sites suggests that a 17 bp sequence of imperfect dyad symmetry is important for recognition by GAL4.

Figure 1.



Although GAL4 protein is made at a constant level under all growth conditions, it binds to the *GAL* upstream elements when cells are grown in galactose but not glucose medium (Giniger et al., op. cit.). Presumably, the binding and nonbinding states of GAL4 account for the transcriptional regulatory patterns of the *GAL* structural genes.

The product of the *GCN4* gene is necessary for the transcriptional induction of many amino acid biosynthetic enzymes in response to amino acid starvation. GCN4 protein, synthesized from the cloned gene by transcription and translation in vitro or produced in E. coli, binds specifically to the promoter regions of all coregulated genes (Hope and Struhl, Cell *43*, 177–188, 1985; Arndt and Fink, PNAS *83*, 8516–8520, 1986). Saturation point mutagenesis of the *HIS3* regulatory site indicates that the DNA sequence requirements for GCN4-binding in vitro and transcriptional induction in vivo are indistinguishable (Hill et al., Science *234*, 451–457, 1986). Interestingly, although the optimal GCN4 binding site is a 9 bp palindrome, presumptive binding sites in fifteen coregulated promoters all have 1–2 deviations from the consensus. GCN4 protein is made only during conditions of amino acid deprivation even though *GCN4* mRNA is synthesized at all times (Thireos et al., PNAS *81*, 5096–5100, 1984; Hinnebusch, PNAS *81*, 6442–6446, 1984). This novel translational control of GCN4 synthesis accounts for the fact that amino acid biosynthetic enzymes are induced only when cells are starved for amino acids.

The HAP1 protein binds to the upstream regulatory elements of two cytochrome *c* genes, *CYC1* and *CYC7*, and induces their transcription (Pfeifer et al., Cell *49*, 9–18 and 19–27, 1987). Specific DNA-binding in vitro and transcriptional induction in vivo are stimulated by heme. Surprisingly, the *CYC1* and *CYC7* binding sites have no obvious sequence similarity, even though they compete for HAP1 binding and have comparable affinities for the protein. Moreover, a HAP1 mutant protein that binds normally to the *CYC7* site fails to bind the *CYC1* site. Thus, the HAP1 activator protein recognizes two different DNA sequences, possibly by using a single DNA-binding domain.

Analyses of truncated versions of GAL4 and GCN4 indicate that the DNA-binding domains are contained within short regions of the proteins. For GAL4 (881 amino acids), the N-terminal 73 amino acids are sufficient for binding to the cognate site (Keegan et al., Science *231*, 699–704, 1986). Since this N-terminal region contains two sequences that may be binding sites for zinc, it is possible that the protein interacts with DNA via "zinc-fingers," the model initially proposed for TFIIIA (Miller et al., EMBO J. *4*, 1609–1614, 1985). For GCN4 (281 amino acids), the 60 C-terminal amino acids are sufficient for specific DNA-binding (Hope and Struhl, Cell *46*, 885–894, 1986). The sequence of the GCN4 DNA-binding domain makes it unlikely that either of the two major structural motifs, zinc-finger or helix-turn-helix, are involved in DNA recognition.

DNA-Binding and Transcriptional Activation Are Separable Functions of Activator Proteins

Although the 73 N-terminal amino acids of GAL4 and the 60 C-terminal amino acids of GCN4 are sufficient for DNA binding, such truncated proteins do not activate transcription in vivo. Instead, DNA-binding by these shortened proteins actually represses transcription in situations where the recognition site is placed between a heterologous upstream element and TATA sequence (Keegan et al., op cit.; Hope and Struhl, 1986, op cit.). These derivatives, which are analogous to positive control mutants of the λcI or P22c2 proteins, strongly suggest that GAL4 and GCN4 contain a transcriptional activation function that is distinct from DNA binding.

A clever experiment by Brent and Ptashne (Cell *43*, 729–736, 1985) clearly shows that DNA binding and transcriptional activation are separate functions. Specifically, a hybrid protein was generated in which the GAL4 DNA-binding domain (the 73 N-terminal amino acids) was replaced by the DNA-binding domain of the E. coli LexA repressor. Although this LexA-GAL4 fusion protein was unable to activate the native *GAL* genes, it was able to activate transcription from a promoter in which the normal GAL4 binding site was replaced by a LexA operator. In other words, the hybrid protein bindings via the LexA repressor–operator interaction and stimulates transcription through the GAL4 activation function. Similarly, LexA-GCN4, a hybrid protein in which the LexA binding domain is fused near the N-terminus of GCN4, is a bifunctional activator protein that can stimulate transcription upon binding to sites recognized either by the GCN4 or LexA binding domains (Hope and Struhl, 1986, op cit.). This last result also suggests that the GCN4 activation function is effective whether a DNA-binding domain is located at its natural C-terminal position or at the inverted N-terminal position.

Transcriptional Activation Functions Are Defined by Short Acidic Regions with Little Sequence Homology

The transcriptional region of GCN4 has been localized by sequential deletion analysis of LexA-GCN4; in this way, the ability to stimulate transcription could be tested even when one of the two DNA-binding domains was removed (Hope and Struhl, 1986, op cit.). Surprisingly, almost all of GCN4 can be deleted without significantly affecting the transcriptional activation function. However, deletions from either direction that remove part or all of a 19 amino acid region reduce or eliminate transcriptional activation.

The most obvious structural feature of the 19 amino acid segment is its central location within an acidic region of GCN4 (30% acidic residues over 60 amino acids). Clearly, this acidic region is more than sufficient for transcriptional activtion by GCN4, because approximately 20 amino acids can be deleted from either end with minimal loss of function. These and additional experiments suggest that acidic regions of 35–40 amino acids are sufficient for full activation, and those of 20–30 amino acids activate to some extent. Interestingly, different portions of the acidic region are equally capable of transcriptional activation, suggesting that transcriptional activation regions do not have rigid sequence requirements.

Recent deletion analysis of the GAL4 activator has yielded similar results (Ma and Ptashne, Cell *48*, 847–853, 1987). Specifically, two short (50–100 amino acids) and separate regions are each capable of transcriptional acti-

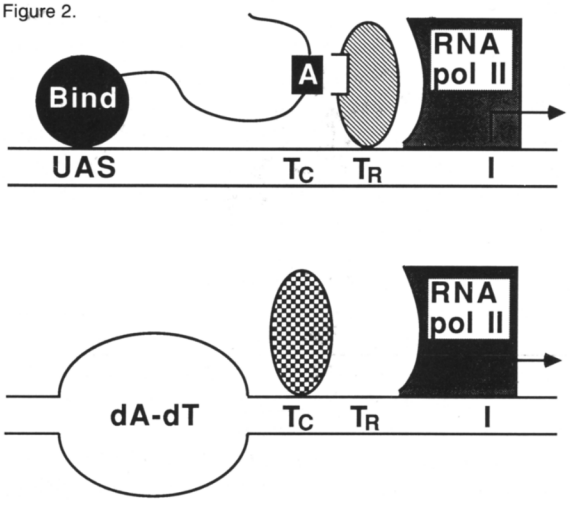
vation when fused to the DNA-binding domain, and more than 80% of the protein can be removed without major phenotypic effects. Moreover, these two transcriptional activation regions correspond to the only two acidic regions of GAL4, yet neither is homologous to the GCN4 transcriptional activation region. Similarly, several other presumptive yeast activators contain regions equally acidic as the GCN4 or GAL4 activation regions, yet there is no sequence homology. Taken together, these results strongly suggest that yeast transcriptional activation regions are short and are not defined by a specific primary sequence, but rather by a more general structural feature.

Molecular Mechanisms for Transcriptional Activation

The short acidic regions that are sufficient for activation are likely to be surfaces used for interactions with other proteins of the transcription machinery. It seems unlikely that these short acidic regions of limited homology could encode catalytic activities (e.g., topoisomerases, nucleases, methylases) that might be involved in transcription. In this regard, transcriptional activation regions might be analogous to signal sequences that are used for targeting proteins to particular locations within the cell. Mitochondrial import sequences, for example, are generally defined by short hydrophobic regions (with occasional basic residues) of variable sequence, which are believed to interact with some components of the membrane (von Heijne, EMBO J. *5*, 1335–1342, 1986).

If the activation region reflects an interaction site, with what proteins does it interact? The obvious choices are RNA polymerase II, a TATA-binding protein, or histones. Although basic histones might be expected to associate with acidic activation regions, the following observation suggests that the critical interaction is more likely to be with a TATA binding protein. In the case of the *HIS3* promoter region, which contains different kinds of TATA elements (T_R and T_C), activation by either GCN4 or GAL4 is observed only in combination with T_R (Struhl, MCB, op. cit.). As this restriction does not depend on the distance between the GCN4 or GAL4 binding sites and the *HIS3* TATA region, it probably reflects a functional distinction between the two different classes of TATA elements. These two TATA elements have different sequences, and almost certainly interact with different DNA-binding proteins. Thus, I suggest that the GCN4 and GAL4 activation regions can associate with a protein that binds T_R to stimulate transcription, whereas they are unable to interact with a protein that recognizes T_C (Figure 2). In contrast, transcriptional activation through T_C occurs with poly(dA-dT) upstream elements, possibly by a different molecular mechanism involving the unusual structure of poly(dA-dT).

Figure 2.



A view of the transcriptional initiation process in yeast (and other eukaryotes) is that RNA polymerase II initiates mRNA synthesis at discrete sites upon recognizing a transcription complex composed of activator proteins (GAL4, GCN4), general transcription factors (such as TATA proteins), and the DNA (Figure 2). The complex is formed/stabilized by specific interactions between the proteins and cognate DNA sequences, and by protein–protein interactions between specific activators and general transcription factors. In this sense, the DNA serves as a specific scaffold for the assembly of an active transcription complex. The striking observation that activator proteins function at long and variable distances from TATA elements can be explained by looping out of the intervening DNA to bring the proteins closer together (reviewed by Ptashne, Nature *322*, 697–701, 1986), and/or by variable conformations of the large nonessential parts of the activator protein such that its critical acidic surface can interact with the other protein(s). Finally, transcriptional regulation can be accomplished in several ways, via: specific DNA sequences that allow only certain proteins to bind; environmentally regulated cofactors that affect the synthesis or specific DNA binding of one of the proteins; and rules of interaction between specific and general transcription factors. Using just these basic parameters, one can easily imagine how the transcription of a single gene can be controlled in a complex manner.