Eukaryotic Transcription Factors: Identification, Characterization and Functions\(^1,2\)

Vincent W. Yang

Departments of Medicine and Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

The regulation of tissue- and temporal-specific eukaryotic gene expression and the activation of genes in response to extracellular inducers are two fundamental processes that attract many a molecular biologist. The development of methods for cloning individual genes has provided the opportunity to study the mechanisms underlying these processes at a molecular level. What has been learned is that eukaryotic promoters consist of defined short stretches of DNA sequences, which are recognized by a variety of specific DNA-binding proteins that regulate transcription. The current challenges include an understanding of 1) which specific cis-acting DNA sequence elements and which trans-acting factors (transcription factors) are required for the expression of a given gene, 2) how a given set of DNA-protein interactions regulates the expression of a tissue-specific gene, and 3) how these interactions are integrated into the overall regulation of gene expression during development.

This article will address the current methodologies used to identify and characterize trans-acting eukaryotic transcription factors. In addition, some of the common classes of eukaryotic transcription factors studied to date and their functions in gene regulation will also be described. For a detailed description of methodologies, readers are referred to a number of recent publications (Hames and Higgins 1993, Latchman 1993, Revzin 1993). Readers are also encouraged to consult a number of recently published excellent textbooks on the molecular mechanisms of transcriptional regulation (Conaway and Conaway 1994, Goodburn 1996, Latchman 1995, McKnight and Yamamoto 1992, Wingender 1993).

DNA BINDING ASSAYS USED TO STUDY TRANSCRIPTION FACTORS

The principal strategy in identifying and characterizing transcription factors is based on their ability to recognize and interact with specific DNA sequences present in the promoters of eukaryotic genes. The detection of sequence-specific DNA binding activities from crude cell extracts is usually the first step that leads to the eventual purification and cloning of various transcription factors. Two techniques are often used in assessing DNA-protein interactions: the electrophoretic mobility shift assay (EMSA)\(^3\) (Fried and Crothers 1981, Garner and Revzin 1981) and the DNase I protection (footprinting) assay (Brenowitz et al. 1986, Galas and Schmitz 1978). Variations of these two methods include methylation interference (Brunelle and Schleif 1987), ultraviolet (UV) crosslinking (Chodosh et al. 1986), and Southwestern blotting (Kwast-Welfeld et al. 1993). All of these techniques are described below.

Electrophoretic mobility shift assay (EMSA)

In EMSA the binding of a sequence-specific DNA binding protein (present in a mixture of various proteins in a nuclear extract) to a radioactively labeled DNA fragment (or probe, usually a double-stranded, synthetic oligonucleotide) results in the formation of a DNA-protein complex with a reduced mobility of the DNA in a nondenaturing polyacrylamide gel. This DNA-protein complex can readily be distinguished electrophoretically from the unbound probe (Fig. 1A). To determine the sequence-specific nature of the interaction between the DNA and the protein in the complex, binding reactions are performed in the presence of excess amounts of unlabeled DNA fragments (competitors) that are identical (specific) or unrelated (nonspecific) to the radiolabeled probe. The presence of excess amounts of competitors with a sequence identical to that of the probe will inhibit or reduce the formation of the radiolabeled DNA probe-protein complex, whereas the presence of a nonspecific competitor will not affect the DNA-protein interaction (Fig. 1B, lanes 3–6). The identity of the protein that binds to the DNA can sometimes be established if an antibody directed against this protein is available. The addition of the antibody to the EMSA results in the formation of an even slower migrating complex that contains the DNA, protein and antibody in a process often referred to as supershift (Fig. 1B, lane 7). EMSA is simple to perform and is more sensitive than the DNase I footprinting assay in determining interaction between a protein and its cognate

\(^1\) This work was supported in part by grants from the National Institutes of Health (DK44484 and DK52230).

\(^2\) The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

Abbreviations used: AD, activation domain; DBD, DNA binding domain; DMS, dimethyl sulfate; DNase I, deoxyribonuclease I; EMSA, electrophoretic mobility shift assay; HLH, helix-loop-helix; TAF, TBP-associated factors; TBP, TATA-binding protein; UV, ultraviolet.

Manuscript received 1 June 1998. Revision accepted 3 August 1998.
DNA sequence. Furthermore, this assay provides a quantitative measure of the amount of a particular DNA binding activity. However, EMSA does not give a direct readout of the DNA nucleotides that the protein recognizes. For the latter information a technique with a higher resolution such as DNase I footprinting is necessary.
DNase I protection (footprinting) assay

In DNase I footprinting the binding of a protein to a specific region within a singly end-labeled DNA fragment protects this region from digestion by DNase I. This results in a region of DNase I protection (footprint) when the digested DNA products are resolved on a denaturing polyacrylamide gel (Fig. 2). This technique allows the determination of a short stretch of protein-binding site within a relatively large DNA fragment. The exact nucleotide sequence in the protected region can readily be determined by concurrently running Maxam and Gilbert sequencing reactions of the same DNA fragment alongside the DNase I digestion products. The disadvantage of this procedure is that it is less sensitive than the gel mobility shift assay and is technically more difficult to perform. However, if cloned transcription factors are available and are produced in large amounts in bacteria, DNase I footprinting can frequently provide quantitative information on the binding activities of the proteins.

Other commonly employed DNA binding assays

(i) Methylation interference assay. This assay is based on the fact that methylation of specific guanine or adenine residues within the target DNA sequence inhibits the binding of a transcription factor to that site (Fig. 3). A singly end-labeled DNA probe is first partially methylated with dimethyl sulfate (DMS) and incubated with the nuclear extract of interest. The protein-DNA complex is then separated from the free DNA using EMSA. Both protein-bound and free DNA are eluted from the gel, cleaved at the site of modification with piperidine and resolved by denaturing polyacrylamide gel electrophoresis. If methylation occurs at a particular guanine or adenine residue that is critical for the DNA-protein interaction, the binding of the protein to that DNA will be inhibited, resulting in the recovery of that DNA only from the free DNA fraction. The presence of particular guanine and adenine bands in the free DNA fraction and their concomitant absence from the bound DNA fraction are indicative of those nucleotides being the contact points of the protein.

(ii) UV crosslinking. Irradiation of DNA with ultraviolet light produces pyrimidine free radicals that are chemically active and can form covalent bonds such as thymidine dimers. This reactive property of UV-irradiated DNA can
be used to link transcription factors to their respective recognition sites. When a protein-DNA complex is irradiated with UV light, it causes the formation of covalent bonds between pyrimidines and certain amino acid residues in the transcription factor that are in close proximity to the DNA. The labeling of a transcription factor in this fashion allows for the easy and rapid determination of its approximate molecular weight in a denaturing polyacrylamide gel even in crude extracts. Frequently, halogenated analogues of thymidine (for example, bromodeoxyuridine or BrdU) are incorporated into the DNA enzymatically to enhance the crosslinking between protein and DNA.

(iii) Southwestern blotting. As the name implies, Southwestern blotting is a variation of the Western blotting technique. Cell extracts containing the DNA-binding protein of interest are resolved by denaturing polyacrylamide gel electrophoresis followed by electrophoretic transfer to a nitrocellulose membrane. The membrane is then probed with a radioactively labeled DNA fragment bearing the recognition site, preferably in the form of tandem repeats. The protein that interacts with the probe can be visualized by autoradiography after nonspecifically bound DNA is first washed away from the membrane.

PURIFICATION AND CLONING OF TRANSCRIPTION FACTORS

To characterize the biochemical properties of transcription factors, it is often necessary to study them in pure (cloned) forms. Several different approaches have recently been developed to achieve the cloning of cDNAs encoding various transcription factors. They generally fall into two major categories: protein purification by conventional biochemical methods, followed by peptide sequencing or antibody generation for cDNA library screening, and expression cloning of transcription factors based on their sequence-specific DNA recognition.

Biochemical purification of transcription factors

A major difficulty in the purification of transcription factors is their low abundance (ranging between 10^3 and 10^7 molecules per cell). Assuming 100 pmol (5 μg) for a protein with a molecular weight of 50,000 kDa) of pure protein are required for the production of antibody or peptides for sequencing, it is estimated that 10^11 cells are needed as starting materials for a transcription factor averaging 10^4 molecules per cell and a 5% recovery of protein. Another requirement is that a DNA sequence with high affinity and specificity for the transcription factor should be identified so that the DNA binding activity of the protein can be monitored during each step of the purification process.

The biochemical purification of a transcription factor begins with the preparation of nuclear extracts from appropriate cells or tissues. This step gives between 10- and 100-fold enrichment of the nuclear-localizing protein. Ammonium sulfate fractionation is frequently used to further concentrate the protein in the nuclear extracts. Following dialysis to remove the ammonium sulfate, the protein mixtures are then subjected to fractionation by conventional column chromatography. A number of different chromatographic procedures can be employed. Examples include gel filtration chromatography such as Sephacryl S-300, heparin-agarose affinity chromatography and DNA-cellulose chromatography. The latter two are based on the ability of most transcription factors to bind to negatively charged heparin and to interact with random DNA sequences, respectively. After washing the loaded column, proteins adsorbed to the column are eluted with appropriate buffers and collected in fractions. Each fraction is then assayed for DNA binding by EMSA or DNase I footprinting. The fractions that contain the highest activity are pooled and subjected to the next step of purification, which frequently involves sequence-specific DNA affinity chromatography (Kadonaga and Tjian 1986, Rosenfeld and Kelly 1986). To construct the column, multiple tandem repeats of a double-stranded oligonucleotide bearing the high affinity-binding sequence for the transcription factor are covalently attached to a matrix support such as cyanogen bromide-activated Sepharose beads. After passing the proteins through this column, the transcription factor will bind to the matrix, which can then be eluted with a salt gradient. The procedure can be repeated to increase the protein purity. Many transcription factors can be purified up to 1000-fold by two sequential DNA affinity chromatographic steps.

Expression cloning of transcription factors

Two recently described techniques can be used in expression cloning, which is based primarily on the sequence-specific interaction of a transcription factor with its high affinity-binding site. One involves the screening of an expression cDNA library with a radiolabeled oligonucleotide probe containing the recognition site for the transcription factor (Singh et al. 1989, Vinson et al. 1989). The other is a genetic selection using the yeast one-hybrid system (Wang and Reed 1993). Both will be briefly discussed below.

(i) In situ detection of transcription factors. The principal of this technique is similar to that of Southwestern blotting. An expression cDNA library is constructed using bacteriophage vectors, such as Agt11, in which the cloned cDNA is expressed as a fusion protein with β-galactosidase upon addition of an inducer of the lac operon. The induced proteins are transferred to nitrocellulose membranes followed by an optional step of denaturation and stepwise renaturation. The membranes are then probed with radiolabeled tandem repeats of an oligonucleotide bearing the high affinity-binding sequence. After washing the membranes to remove any nonspecifically attached probes, autoradiography is performed to identify the cDNA clones that bind to the probe. The authenticity of these clones can subsequently be verified by binding crude extracts prepared from lysogens of the recombinant phage to the radiolabeled probe using EMSA or DNase I footprinting assay.

(ii) The yeast one-hybrid selection system. The yeast one-hybrid system stems from the important observation that most eukaryotic transcription factors are modular by nature, composed of a target-specific DNA-binding domain and a target-independent transcription activation domain. In this procedure (Fig. 4) candidates of cDNA clones encoding the transcription factor of interest (TF X) are expressed in the yeast as fusion proteins that contain a target-independent activation domain (AD) of the potent yeast transcription factor, GAL4. A reporter construct is generated that consists of multiple copies of the target sequence (X) adjacent to a low activity promoter (Pmin) directing HIS3 gene expression. The presence of this reporter allows sufficient growth for the selection of a stably integrated yeast strain when it is introduced into a his^3' parent yeast strain. The further introduction of a GAL4-candidate fusion protein capable of binding to the target sequence present in the reporter will strongly activate
transcription and increase levels of HIS3. This allows rapid growth and subsequent positive identification of candidate clones in media that lack histidine.

**COMMON CLASSES OF TRANSCRIPTION FACTORS AND THEIR FUNCTIONS**

Transcription factors are often classified based on the structural motifs that constitute their DNA-binding domains. In most cases the protein makes a large number of contacts with the DNA, involving hydrogen bonds, ionic bonds and hydrophobic interactions. Although each individual contact is weak, the 20 or so contacts that typically form at the protein-DNA interface ensure that the interaction is both specific and strong. In this section several major classes of eukaryotic transcription factors (Fig. 5) and their functions will be briefly reviewed. Readers are reminded that transcription of eukaryotic genes requires the participation of many additional regulatory proteins other than the sequence-specific DNA-binding transcription factors addressed below. Examples of these important proteins involved in transcriptional regulation include the TATA-binding protein (TBP), TBP-associated factors (TAF) and the recently identified family of transcription co-activators such as p300/CBP. Several recent articles provide excellent reviews of these topics (Glass et al. 1997, Goodrich and Tjian 1994, Sauer and Tjian 1997, Shikama et al. 1997, Verrijzer and Tjian 1996).

**FIGURE 4** Cloning of transcription factors by the yeast one-hybrid system. A reporter plasmid is constructed in a yeast vector that contains the HIS3 reporter driven by a minimal promoter (Pmin). This vector is linked to three tandem copies of the binding sequence (X) to which the transcription factor X (TF X) binds. This plasmid is stably introduced into an Urac1−, His3− yeast strain using uracil2−, histidine2− media (although expression of HIS3 from the minimal promoter is low, it is sufficient to allow growth and to select for integration of the reporter plasmid). The stably transformed yeast strain is then used to screen a library containing the AD of GAL4 fused to a library of cDNA sequences obtained from a given cell or tissue. Clones that contain a fusion protein between GAL4 AD and the DBD of TF X will strongly activate HIS3 expression through binding to the tandem repeats of DNA sequence X in the HIS3 reporter, allowing the positive selection of rapidly growing clones in histidine− media.

**FIGURE 5** Major classes of eukaryotic transcription factors. (A) & (B) A homeodomain bound to its specific DNA sequence. The homeodomain is folded into three α helices packed tightly together by hydrophobic interaction (A). The part containing helices 2 and 3 closely resembles the helix-turn-helix motif with the recognition helix (helix #3) making important contacts with the major groove (B). (C) & (D) One type of zinc finger protein. This protein belongs to the Cys-Cys-His-His (C2H2) family of zinc finger proteins, named after the amino acids that grasp the zinc. The three-dimensional structure of the zinc finger is constructed from an antiparallel β sheet followed by an α helix (C). The zinc finger transcription factors often contain multiple zinc fingers that are contiguous to each other and contact the DNA in similar ways. In (D) a small sphere represents the zinc atom in each finger. (E) A leucine zipper dimer bound to DNA. Two α helical DNA-binding domains (bottom) dimerize through their α-helical leucine zipper region (top) to form an inverted Y-shaped structure. (F) A helix-loop-helix dimer bound to DNA. The two monomers are held together in a four-helix bundle; each monomer contributes two α helices connected by a flexible loop of protein. (Reproduced with permission from Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J.D.: Molecular Biology of the Cell, 3rd ed., ©1994, by Garland Publishing, Inc., New York, NY).
The helix-turn-helix motif (including the homeodomain proteins)

The first DNA-binding protein motif to be recognized was the helix-turn-helix. It is constructed from two α helices connected by a short chain of amino acids, which constitutes the turn. The more carboxyl-terminal helix is called the recognition helix because it fits into the major groove of DNA. As is the case with many sequence-specific DNA binding proteins, helix-turn-helix proteins bind as symmetrical dimers to a DNA composed of two symmetrically arranged half sites with similar sequences. Examples of this class of proteins include the bacterial tryptophan repressor and the phage λ repressor.

An important class of helix-turn-helix proteins are the homeodomain proteins (Fig. 5A, B). The homeodomain is a stretch of about 60 amino acids that is highly conserved in a class of genes called homeotic genes, which play a critical role in the development of organisms. In Drosophila at least 60 homeodomain proteins have been identified. Homologues of these proteins have been identified in virtually all eukaryotes from yeasts to man.

The zinc finger motifs

As the name zinc finger implies, this family of proteins contains a domain that utilizes zinc as an important component of the DNA-binding region. Amino acid residues such as cysteine and histidine tetrahedrally coordinate the zinc atom. Together they form a finger-like projection that is in close contact with the DNA. A common type of zinc finger protein is exemplified by the Xenopus transcription factor TFIIB that is involved in the transcription of ribosomal genes. The finger is a simple structure that consists of an α helix and a β sheet held together by the zinc (Fig. 5C). This type of zinc finger is often found in a cluster with additional fingers, arranged one after the other so that the α helix of each can contact the major groove of DNA, forming nearly a continuous stretch of α helix along the groove (Fig. 5D).

A second type of zinc finger is found in a larger family of intracellular steroid receptor proteins. It forms a different type of structure that is in fact more similar to the helix-turn-helix motif in which two α helices are packed together with two zinc atoms. Like helix-turn-helix proteins, these proteins form dimers and allow one of the two α helices of each subunit to interact with the major groove of the DNA.

The leucine zipper motif

The leucine zipper proteins bind to DNA as dimers. Although in many other proteins the dimerization and the DNA-binding domains are distinct, the leucine zipper motif combines both functions. Two α helices, one from each monomer, are joined together to form a short coiled-coil. The helices are held together by interactions between hydrophobic amino acid side chains (usually leucines). Just beyond the dimerization interface, the two α helices separate to form a Y-shaped structure, allowing the side chains (often basic residues) to contact the major groove of DNA (Fig. 5E).

Regulatory proteins with the leucine zipper motif can form either homodimers or heterodimers. Because heterodimers are typically formed from two different proteins with distinct DNA-binding sequences, the ability of leucine zipper proteins to form heterodimers greatly increases the repertoire of DNA-binding specificity that these proteins display. Examples of leucine zipper proteins capable of forming heterodimers include the FOS and JUN families of transcription factors that are important in growth regulation. Also, the C/EBP family of transcription factors falls into this category.

The helix-loop-helix (HLH) motif

An HLH motif consists of a short α helix connected by a loop to a second longer α helix. The flexibility of the loop allows the two helices to pack against each other. This motif is involved in both dimer formation and DNA contact (Fig. 5F). As with the leucine zipper proteins, HLH proteins are capable of forming either homodimer or heterodimer. An important example of the HLH protein is myo D1, a regulatory protein that is essential for the formation of muscle cells.

SUMMARY

Eukaryotic transcription factors are modular proteins that utilize distinct domains for transcriptional activation (or repression) and DNA binding. The highly specific interaction between a given transcription factor and its cognate binding sequence forms the basis for the biochemical characterization and eventual purification of these important regulatory proteins. Commonly used techniques in the assessment of DNA binding of a transcription factor include EMSA and DNase I protection (footprinting) assay. Transcription factors are often purified and cloned based on their specific binding sequences. Finally, several important classes of structural motifs present in many eukaryotic transcription factors are presented. The understanding of the structure and function relationship between transcription factors and the genes that they regulate should provide the basis for understanding the overall molecular mechanisms controlling gene expression.

LITERATURE CITED


EUKARYOTIC TRANSCRIPTION FACTORS


