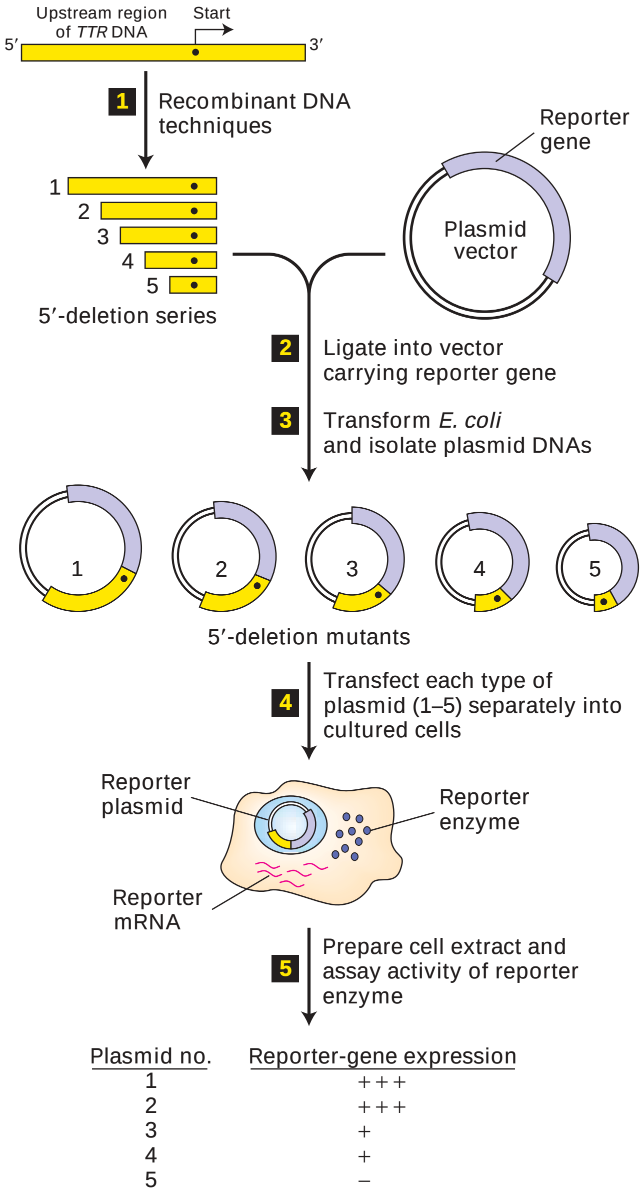
Regulatory Elements in Eukaryotic DNA Often Are Many Kilobases from Start Sites

In eukaryotes, as in bacteria, a DNA sequence that specifies where RNA polymerase binds and initiates transcription of a gene is called a **promoter.** Transcription from a particular promoter is controlled by DNA-binding proteins, termed **transcription factors,** that are equivalent to bacterial *repressors* and *activators*. However, the DNA control elements in eukaryotic genomes that bind transcription factors often are located much farther from the promoter they regulate than is the case in prokaryotic genomes. In some cases, transcription factors that regulate expression of protein-coding genes in higher eukaryotes bind at regulatory sites tens of thousands of base pairs either **upstream** (opposite to the direction of transcription) or **downstream** (in the same direction as transcription) from the promoter. As a result of this arrangement, transcription from a single promoter may be regulated by binding of multiple transcription factors to alternative control elements, permitting complex control of gene expression.

For example, alternative transcription-control elements regulate expression of the mammalian gene that encodes transthyretin (*TTR*), which transports thyroid hormone in blood and the cerebrospinal fluid that surrounds the brain


Upstream region of TTR DNA
Start
5′
3′
1 Recombinant DNA techniques
1
2
3
4
5
5′-deletion series
Reporter gene
Plasmid vector
2 Ligate into vector carrying reporter gene
3 Transform E. coli and isolate plasmid DNAs
1
2
3
4
5
5′-deletion mutants
4 Transfect each type of plasmid (1–5) separately into cultured cells
Reporter plasmid
Reporter enzyme
Reporter mRNA
5 Prepare cell extract and assay activity of reporter enzyme
Plasmid no.
Reporter-gene expression
1 +++
2 +++
3 +
4 +
5 –

▲ EXPERIMENTAL FIGURE 11-3 **5′-Deletion analysis can identify transcription-control sequences in DNA upstream of a eukaryotic gene.** Step 1: Recombinant DNA techniques are used to prepare a series of DNA fragments that extend from the 5′-untranslated region of a gene various distances upstream. Step 2: The DNA fragments are ligated into a reporter plasmid upstream of an easily assayed reporter gene. Step 3: The DNA is transformed into *E. coli* to isolate plasmids with deletions of various lengths 5′ to the transcription start site. Step 4: Each plasmid is then transfected into cultured cells (or used to prepare transgenic organisms) and expression of the reporter gene is assayed (step 5). The results of this hypothetical example (*bottom*) indicate that the test fragment contains two control elements. The 5′ end of one lies between deletions 2 and 3; the 5′ end of the other lies between deletions 4 and 5.

and spinal cord. Transthyretin is expressed in hepatocytes, which synthesize and secrete most of the blood serum proteins, and in choroid plexus cells in the brain, which secrete cerebrospinal fluid and its constituent proteins. The control elements required for transcription of the *TTR* gene were identified by the procedure outlined in Figure 11-3. In this experimental approach, DNA fragments with varying extents of sequence upstream of a start site are cloned in front of a **reporter gene** in a bacterial plasmid using recombinant DNA techniques. Reporter genes express enzymes that are easily assayed in cell extracts. Commonly used reporter genes include the *E. coli lacZ* gene encoding β-galactosidase; the firefly gene encoding luciferase, which converts energy from ATP hydrolysis into light; and the jellyfish gene encoding green fluorescent protein (GFP).

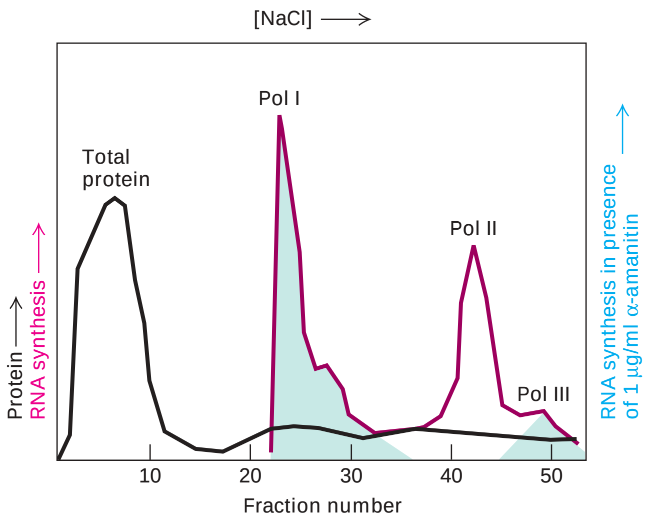
By constructing and analyzing a *5′-deletion series* upstream of the *TTR* gene, researchers identified two control elements that stimulate reporter-gene expression in hepatocytes, but not in other cell types. One region mapped between ≈2.01 and 1.85 kb upstream of the *TTR* gene start site; the other mapped between ≈200 base pairs upstream and the start site. Further studies demonstrated that alternative DNA sequences control *TTR* transcription in choroid plexus cells. Thus, alternative control elements regulate transcription of the *TTR* gene in two different cell types. We examine the basic molecular events underlying this type of eukaryotic transcriptional control in later sections.

Three Eukaryotic Polymerases Catalyze Formation of Different RNAs

The nuclei of all eukaryotic cells examined so far (e.g., vertebrate, *Drosophila*, yeast, and plant cells) contain three different RNA polymerases, designated I, II, and III. These enzymes are eluted at different salt concentrations during ion-exchange chromatography and also differ in their sensitivity to α-amanitin, a poisonous cyclic octapeptide produced by some mushrooms (Figure 11-4). Polymerase I is very insensitive to α-amanitin; polymerase II is very sensitive; and polymerase III has intermediate sensitivity.

Each eukaryotic RNA polymerase catalyzes transcription of genes encoding different classes of RNA. *RNA polymerase I*, located in the nucleolus, transcribes genes encoding precursor rRNA (**pre-rRNA**), which is processed into 28S, 5.8S, and 18S rRNAs. *RNA polymerase III* transcribes genes encoding tRNAs, 5S rRNA, and an array of small, stable RNAs, including one involved in RNA splicing (U6) and the RNA component of the signal-recognition particle (SRP) involved in directing nascent proteins to the endoplasmic reticulum (Chapter 16). *RNA polymerase II* transcribes all protein-coding genes; that is, it functions in production of mRNAs. RNA polymerase II also produces four of the five small nuclear RNAs that take part in RNA splicing.

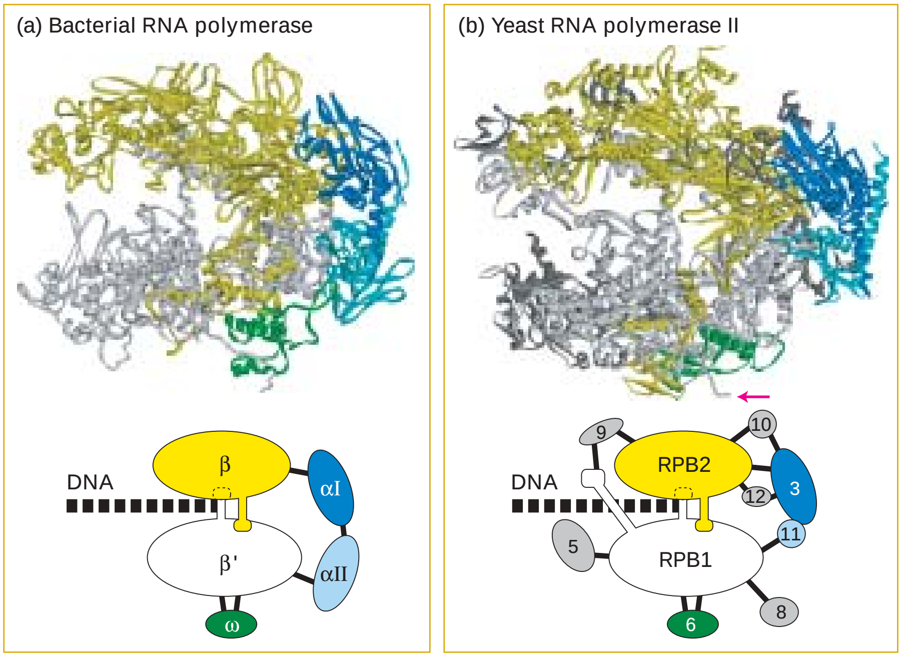
Each of the three eukaryotic RNA polymerases is more complex than *E. coli* RNA polymerase, although their structures are similar (Figure 11-5). All three contain two large



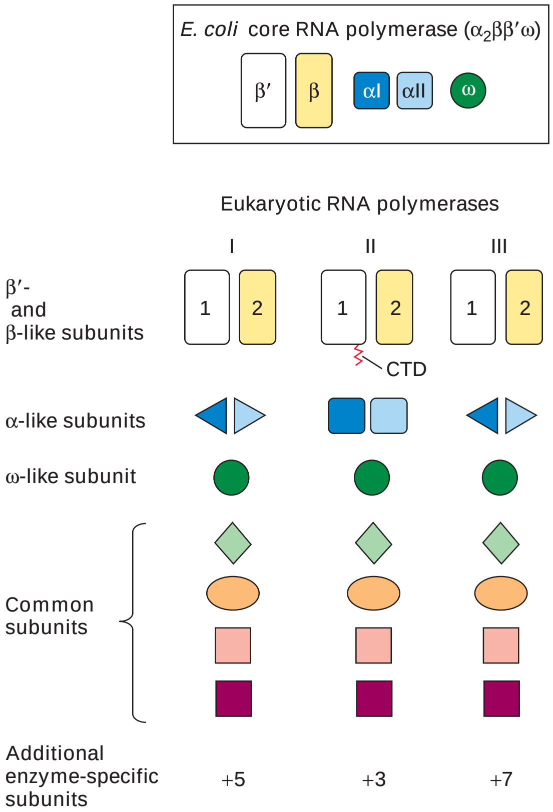
◀ EXPERIMENTAL FIGURE 11-4 **Column chromatography separates and identifies the three eukaryotic RNA polymerases, each with its own sensitivity to α-amanitin.** A protein extract from the nuclei of cultured eukaryotic cells is passed through a DEAE Sephadex column and adsorbed protein eluted (black curve) with a solution of constantly increasing NaCl concentration. Three fractions from the eluate subsequently showed RNA polymerase activity (red curve). At a concentration of 1 μg/ml, α-amanitin inhibits polymerase II activity but has no effect on polymerases I and III (green shading). Polymerase III is inhibited by 10 μg/ml of α-amanitin, whereas polymerase I is unaffected even at this higher concentration. [See R. G. Roeder, 1974, *J. Biol. Chem.* **249**:241.]

subunits and 10–14 smaller subunits, some of which are present in two or all three of the polymerases. The best-characterized eukaryotic RNA polymerases are from the yeast *S. cerevisiae*. Each of the yeast genes encoding the polymerase subunits has been cloned and sequenced and the effects of gene-knockout mutations have been characterized. In addition, the three-dimensional structure of yeast RNA polymerase II missing two nonessential subunits has been determined (see Figure 11-5). The three nuclear RNA polymerases from all eukaryotes so far examined are very similar to those of yeast.

The two large subunits (RPB1 and RPB2) of all three eukaryotic RNA polymerases are related to each other and are similar to the *E. coli* β′ and β subunits, respectively



▲ FIGURE 11-5 **Comparison of three-dimensional structures of bacterial and eukaryotic RNA polymerases.** These C_α trace models are based on x-ray crystallographic analysis of RNA polymerase from the bacterium *T. aquaticus* and RNA polymerase II from *S. cerevisiae.* (a) The five subunits of the bacterial enzyme are distinguished by color. Only the N-terminal domains of the α subunits are included in this model. (b) Ten of the twelve subunits constituting yeast RNA polymerase II are shown in this model. Subunits that are similar in conformation to those in the bacterial enzyme are shown in the same colors. The C-terminal domain of the large subunit RPB1 was not observed in the crystal structure, but it is known to extend from the position marked with a red arrow. (RPB is the abbreviation for "*R*NA *p*olymerase *B*," which is an alternative way of referring to RNA polymerase II.) [Part (a) based on crystal structures from G. Zhang et al.,1999, *Cell* **98**:811. Part (b) from P. Cramer et al., 2001, *Science* **292**:1863.]


E. coli core RNA polymerase ($\alpha_2\beta\beta'\omega$)
β′
β
αI
αII
ω
Eukaryotic RNA polymerases
I
II
III
β′- and β-like subunits
1
2
1
2
1
2
CTD
α-like subunits
ω-like subunit
Common subunits
Additional enzyme-specific subunits
+5
+3
+7

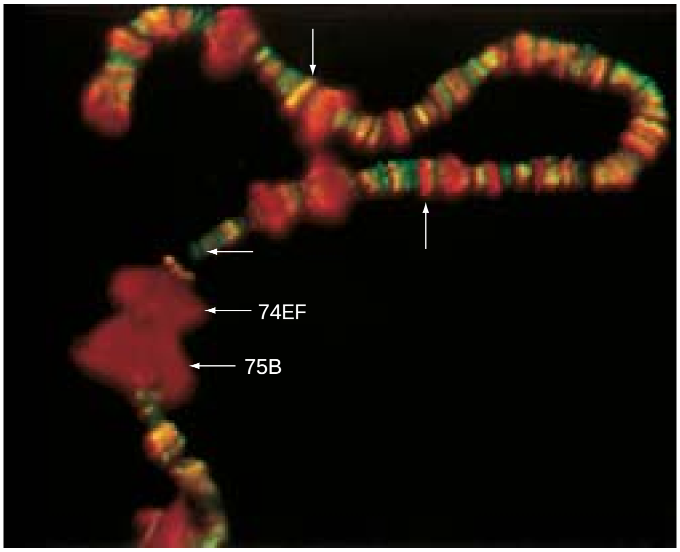
▲ FIGURE 11-6 **Schematic representation of the subunit structure of the *E. coli* RNA core polymerase and yeast nuclear RNA polymerases.** All three yeast polymerases have five core subunits homologous to the β, β′, two α, and ω subunits of *E. coli* RNA polymerase. The largest subunit (RPB1) of RNA polymerase II also contains an essential C-terminal domain (CTD). RNA polymerases I and III contain the same two nonidentical α-like subunits, whereas RNA polymerase II contains two other nonidentical α-like subunits. All three polymerases share the same ω-like subunit and four other common subunits. In addition, each yeast polymerase contains three to seven unique smaller subunits.

(Figure 11-6). Each of the eukaryotic polymerases also contains an ω-like and two nonidentical α-like subunits. The extensive similarity in the structures of these core subunits in RNA polymerases from various sources indicates that this enzyme arose early in evolution and was largely conserved. This seems logical for an enzyme catalyzing a process so basic as copying RNA from DNA.

In addition to their core subunits related to the *E. coli* RNA polymerase subunits, all three yeast RNA polymerases contain four additional small subunits, common to them but not to the bacterial RNA polymerase. Finally, each eukaryotic polymerase has several enzyme-specific subunits that are not present in the other two polymerases. Gene-knockout experiments in yeast indicate that most of these subunits are essential for cell viability. Disruption of the few polymerase subunit genes that are not absolutely essential for viability nevertheless results in very poorly growing cells. Thus it seems likely that all the subunits are necessary for eukaryotic RNA polymerases to function normally.

The Largest Subunit in RNA Polymerase II Has an Essential Carboxyl-Terminal Repeat

The carboxyl end of the largest subunit of RNA polymerase II (RPB1) contains a stretch of seven amino acids that is nearly precisely repeated multiple times. Neither RNA polymerase I nor III contains these repeating units. This heptapeptide repeat, with a consensus sequence of Tyr-Ser-Pro-Thr-Ser-Pro-Ser, is known as the *carboxyl-terminal domain (CTD).* Yeast RNA polymerase II contains 26 or more repeats, the mammalian enzyme has 52 repeats, and an intermediate number of repeats occur in RNA polymerase II from nearly all other eukaryotes. The CTD is critical for viability, and at least 10 copies of the repeat must be present for yeast to survive.



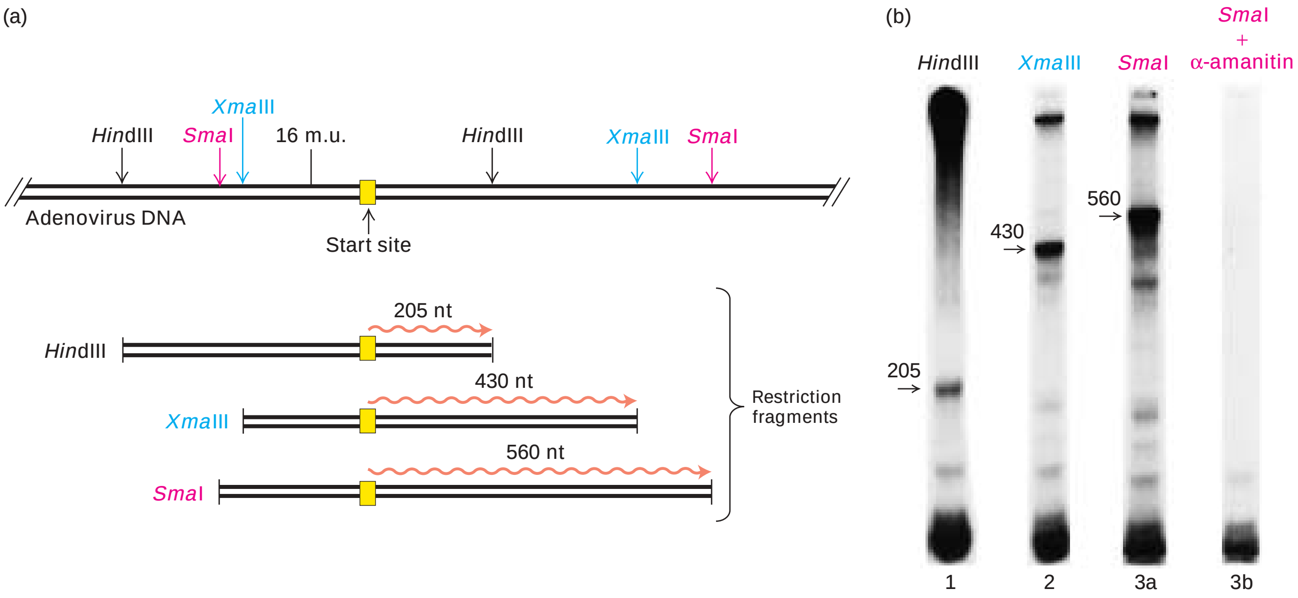
▲ EXPERIMENTAL FIGURE 11-7 **Antibody staining demonstrates that the carboxyl-terminal domain (CTD) of RNA polymerase II is phosphorylated during in vivo transcription.** Salivary gland polytene chromosomes were prepared from *Drosophila* larvae just before molting. The preparation was treated with a rabbit antibody specific for phosphorylated CTD and with a goat antibody specific for unphosphorylated CTD. The preparation then was stained with fluorescein-labeled anti-goat antibody (green) and rhodamine-labeled anti-rabbit antibody (red). Thus polymerase molecules with an unphosphorylated CTD stain green, and those with a phosphorylated CTD stain red. The molting hormone ecdysone induces very high rates of transcription in the puffed regions labeled 74EF and 75B; note that only phosphorylated CTD is present in these regions. Smaller puffed regions transcribed at high rates also are visible. Nonpuffed sites that stain red (up arrow) or green (horizontal arrow) also are indicated, as is a site staining both red and green, producing a yellow color (down arrow). [From J. R. Weeks et al., 1993, *Genes & Dev.* **7**:2329; courtesy of J. R. Weeks and A. L. Greenleaf.]

In vitro experiments with model promoters first showed that RNA polymerase II molecules that initiate transcription have an unphosphorylated CTD. Once the polymerase initiates transcription and begins to move away from the promoter, many of the serine and some tyrosine residues in the CTD are phosphorylated. Analysis of polytene chromosomes from *Drosophila* salivary glands prepared just before molting of the larva indicate that the CTD also is phosphorylated during in vivo transcription. The large chromosomal "puffs" induced at this time in development are regions where the genome is very actively transcribed. Staining with antibodies specific for the phosphorylated or unphosphorylated CTD demonstrated that RNA polymerase II associated with the highly transcribed puffed regions contains a phosphorylated CTD (Figure 11-7).

RNA Polymerase II Initiates Transcription at DNA Sequences Corresponding to the 5′ Cap of mRNAs

Several experimental approaches have been used to identify DNA sequences at which RNA polymerase II initiates transcription. Approximate mapping of the transcription start site is possible by exposing cultured cells or isolated nuclei to ^{32}P-labeled ribonucleotides for very brief times, as described earlier. After the resulting labeled nascent transcripts are separated on the basis of chain length, each size fraction is incubated under hybridization conditions with overlapping restriction fragments that encompass the DNA region of interest. The restriction fragment that hybridizes to the shortest labeled nascent chain, as well as all the longer ones, contains the transcription start site. One of the first transcription units to be analyzed in this way was the major late transcription unit of adenovirus. This analysis indicated that transcription was initiated in a region ≈6 kb from the left end of the viral genome.

The precise base pair where RNA polymerase II initiates transcription in the adenovirus late transcription unit was determined by analyzing the RNAs synthesized during in vitro transcription of adenovirus DNA restriction fragments that extended somewhat upstream and downstream of the approximate initiation region determined by nascent-transcript analysis. The rationale of this experiment and typical results are illustrated in Figure 11-8. The RNA transcripts synthesized in vitro by RNA polymerase II from the



▲ EXPERIMENTAL FIGURE 11-8 **In vitro transcription of restriction fragments and measurement of the RNA lengths localize the initiation site of the adenovirus major late transcription unit.** (a) The top line shows restriction sites for *Hin*dIII (black), *Xma*III (blue), and *Sma*I (red) in the region of the adenovirus genome where the transcription-initiation site was located by nascent-transcript analysis (near 16 map units). The *Hin*dIII, *Xma*III, and *Sma*I restriction fragments that encompass the initiation site were individually incubated with a nuclear extract prepared from cultured cells and ^{32}P-labeled ribonucleoside triphosphates. Transcription of each fragment begins at the start site and ends when an RNA polymerase II molecule "runs off" the cut end of the fragment template, producing a run-off transcript (wavy red lines). (b) The run-off transcripts synthesized from each fragment were then subjected to gel electrophoresis and autoradiography to determine their exact lengths. Since the positions of the restriction sites in the adenovirus DNA were known, the lengths of the run-off transcripts in nucleotides (nt) produced from the restriction fragments precisely map the initiation site on the adenovirus genome, as diagrammed in part (a). In the gels shown here, the bands at the top and bottom represent high- and low-molecular-weight RNA transcripts that are formed under the conditions of the experiment. The sample in lane 3b is the same as that in lane 3a, except that α-amanitin, an inhibitor of RNA polymerase II, was included in the transcription mixture. See the text for further discussion. [See R. M. Evans and E. Ziff, 1978, *Cell* **15**:1463, and P. A. Weil et al., 1979, *Cell* **18**:469. Autoradiogram courtesy of R. G. Roeder.]

start site determined in this experiment contained an RNA cap structure identical with that present at the 5′ end of nearly all eukaryotic mRNAs (see Figure 4-13). This 5′ cap was added by enzymes in the nuclear extract, which can add a cap only to an RNA that has a 5′ tri- or diphosphate. Because a 5′ end generated by cleavage of a longer RNA would have a 5′ monophosphate, it could not be capped. Consequently, researchers concluded that the capped nucleotide generated in the in vitro transcription reaction must have been the nucleotide with which transcription was initiated. The finding that the sequence at the 5′ end of the RNA transcripts produced in vitro is the same as that at the 5′ end of late adenovirus mRNAs isolated from cells confirmed that the capped nucleotide of adenovirus late mRNAs coincides with the transcription-initiation site.

Similar in vitro transcription assays with other cloned eukaryotic genes have produced similar results. In each case, the start site was found to be equivalent to the capped 5′ sequence of the corresponding mRNA. Thus synthesis of eukaryotic precursors of mRNAs by RNA polymerase II begins at the DNA sequence encoding the capped 5′ end of the mRNA. Today, the transcription start site for a newly characterized mRNA generally is determined simply by identifying the DNA sequence encoding the 5′ end of the mRNA.

KEY CONCEPTS OF SECTION 11.1

Overview of Eukaryotic Gene Control and RNA Polymerases

- The primary purpose of gene control in multicellular organisms is the execution of precise developmental decisions so that the proper genes are expressed in the proper cells during development and cellular differentiation.

- Transcriptional control is the primary means of regulating gene expression in eukaryotes, as it is in bacteria.

- In eukaryotic genomes, DNA transcription control elements may be located many kilobases away from the promoter they regulate. Different control regions can control transcription of the same gene in different cell types.

- Eukaryotes contain three types of nuclear RNA polymerases. All three contain two large and three smaller core subunits with homology to the β′, β, α, and ω subunits of *E. coli* RNA polymerase, as well several additional small subunits (see Figure 11-6).

- RNA polymerase I synthesizes only pre-rRNA. RNA polymerase II synthesizes mRNAs and some of the small nuclear RNAs that participate in mRNA splicing. RNA polymerase III synthesizes tRNAs, 5S rRNA, and several other relatively short, stable RNAs.

- The carboxyl-terminal domain (CTD) in the largest subunit of RNA polymerase II becomes phosphorylated during transcription initiation and remains phosphorylated as the enzyme transcribes the template.

- RNA polymerase II initiates transcription of genes at the nucleotide in the DNA template that corresponds to the 5′ nucleotide that is capped in the encoded mRNA.

11.2 Regulatory Sequences in Protein-Coding Genes

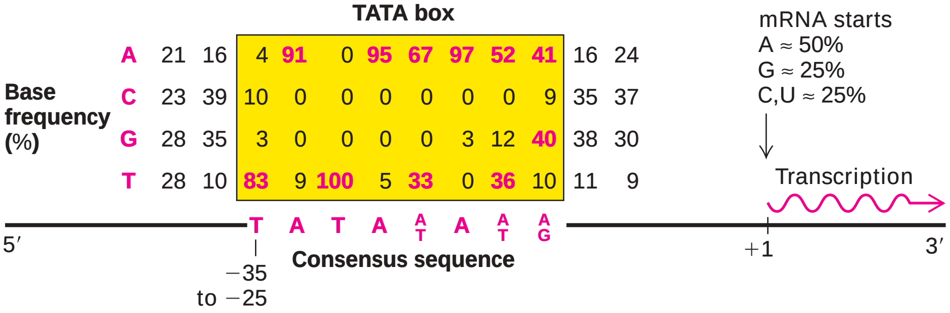
As noted in the previous section, expression of eukaryotic protein-coding genes is regulated by multiple protein-binding DNA sequences, generically referred to as **transcription-control regions.** These include promoters and other types of control elements located near transcription start sites, as well as sequences located far from the genes they regulate. In this section, we take a closer look at the properties of various control elements found in eukaryotic protein-coding genes and some techniques used to identify them.

The TATA Box, Initiators, and CpG Islands Function as Promoters in Eukaryotic DNA

The first genes to be sequenced and studied in in vitro transcription systems were viral genes and cellular protein-coding genes that are very actively transcribed either at particular times of the cell cycle or in specific differentiated cell types. In all these rapidly transcribed genes, a conserved sequence called the **TATA box** was found ≈25–35 base pairs upstream of the start site (Figure 11-9). Mutagenesis studies have shown that a single-base change in this nucleotide sequence drastically decreases in vitro transcription by RNA polymerase II of genes adjacent to a TATA box. In most cases, sequence changes between the TATA box and start site do not significantly affect the transcription rate. If the base pairs between the TATA box and the normal start site are deleted, transcription of the altered, shortened template begins at a new site ≈25 base pairs downstream from the TATA box. Consequently, the TATA box acts similarly to an *E. coli* promoter to position RNA polymerase II for transcription initiation (see Figure 4-11).

Instead of a TATA box, some eukaryotic genes contain an alternative promoter element called an *initiator.* Most naturally occurring initiator elements have a cytosine (C) at the −1 position and an adenine (A) residue at the transcription start site (+1). Directed mutagenesis of mammalian genes with an initiator-containing promoter has revealed that the nucleotide sequence immediately surrounding the start site determines the strength of such promoters. Unlike the conserved TATA box sequence, however, only an extremely degenerate initiator consensus sequence has been defined:

$$(5') \text{ Y-Y-A}^{+1}\text{-N-T/A-Y-Y-Y } (3')$$



A	21	16	4	91	0	95	67	97	52	41	16	24
C	23	39	10	0	0	0	0	0	0	9	35	37
G	28	35	3	0	0	0	0	3	12	40	38	30
T	28	10	83	9	100	5	33	0	36	10	11	9

▲ FIGURE 11-9 **Determination of consensus TATA box sequence.** The nucleotide sequences upstream of the start site in 900 different eukaryotic protein-coding genes were aligned to maximize homology in the region from −35 to −26. The tabulated numbers are the percentage frequency of each base at each position. Maximum homology occurs over an eight-base region, referred to as the *TATA box*, whose consensus sequence is shown at the bottom. The initial base in mRNAs encoded by genes containing a TATA box most frequently is an A. [See P. Bucher, 1990, *J. Mol. Biol.* **212**:563, and http://www.epd.isb-sib.ch/promoter_elements.]

where A^{+1} is the base at which transcription starts, Y is a pyrimidine (C or T), N is any of the four bases, and T/A is T or A at position +3.

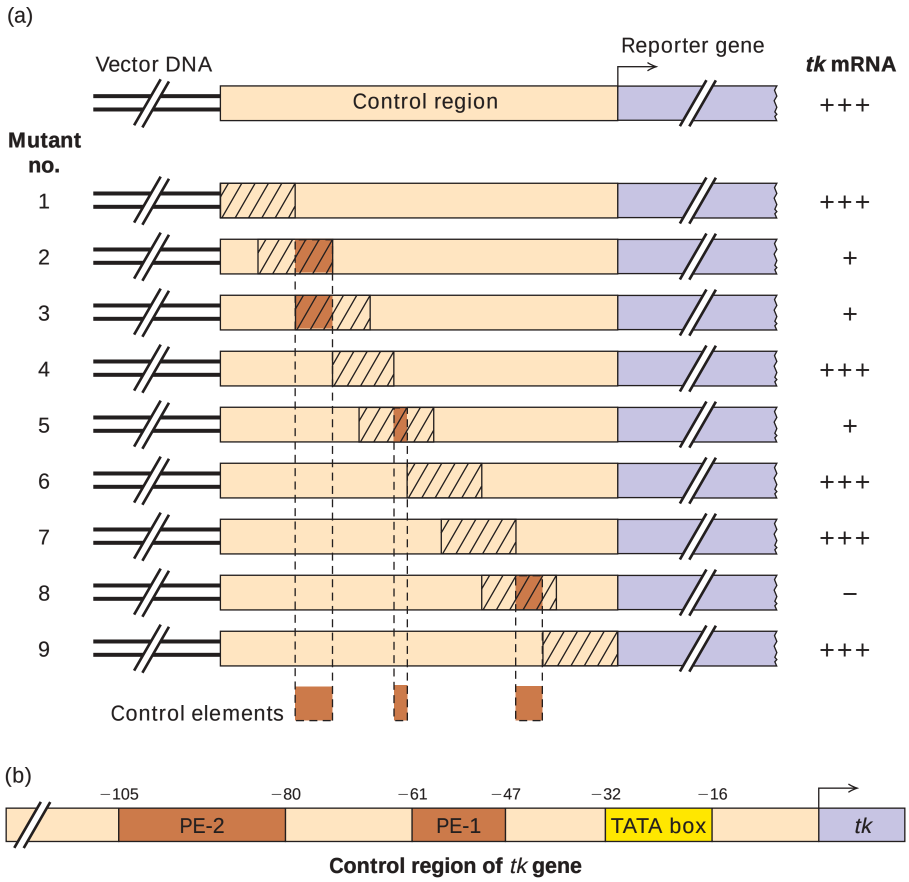
Transcription of genes with promoters containing a TATA box or initiator element begins at a well-defined initiation site. However, transcription of many protein-coding genes has been shown to begin at any one of multiple possible sites over an extended region, often 20–200 base pairs in length. As a result, such genes give rise to mRNAs with multiple alternative 5′ ends. These genes, which generally are transcribed at low rates (e.g., genes encoding the enzymes of intermediary metabolism, often called "housekeeping genes"), do not contain a TATA box or an initiator. Most genes of this type contain a CG-rich stretch of 20–50 nucleotides within ≈100 base pairs upstream of the start-site region. The dinucleotide CG is statistically underrepresented in vertebrate DNAs, and the presence of a CG-rich region, or *CpG island*, just upstream from a start site is a distinctly nonrandom distribution. For this reason, the presence of a CpG island in genomic DNA suggests that it may contain a transcription-initiation region.

Promoter-Proximal Elements Help Regulate Eukaryotic Genes

Recombinant DNA techniques have been used to systematically mutate the nucleotide sequences upstream of the start sites of various eukaryotic genes in order to identify transcription-control regions. By now, hundreds of eukaryotic genes have been analyzed, and scores of transcription-control regions have been identified. These control elements, together with the TATA-box or initiator, often are referred to as the *promoter* of the gene they regulate. However, we prefer to reserve the term *promoter* for the TATA-box or initiator sequences that determine the initiation site in the template. We use the term **promoter-proximal elements** for control regions lying within 100–200 base pairs upstream of the start site. In some cases, promoter-proximal elements are cell-type-specific; that is, they function only in specific differentiated cell types.

One approach frequently taken to determine the upstream border of a transcription-control region for a mammalian gene involves constructing a set of 5′ deletions as discussed earlier (see Figure 11-3). Once the 5′ border of a transcription-control region is determined, analysis of *linker scanning mutations* can pinpoint the sequences with regulatory functions that lie between the border and the transcription start site. In this approach, a set of constructs with contiguous overlapping mutations are assayed for their effect on expression of a reporter gene or production of a specific mRNA (Figure 11-10a). One of the first uses of this type of analysis identified promoter-proximal elements of the thymidine kinase (*tk*) gene from herpes simplex virus (HSV). The results demonstrated that the DNA region upstream of the HSV *tk* gene contains three separate transcription-control sequences: a TATA box in the interval from −32 to −16, and two other control elements farther upstream (Figure 11-10b).

To test the spacing constraints on control elements in the HSV *tk* promoter region identified by analysis of linker scanning mutations, researchers prepared and assayed constructs containing small deletions and insertions between the elements. Changes in spacing between the promoter and promoter-proximal control elements of 20 nucleotides or fewer had little effect. However, insertions of 30 to 50 base pairs between a promoter-proximal element and the TATA box was equivalent to deleting the element. Similar analyses of other eukaryotic promoters have also indicated that considerable flexibility in the spacing between promoter-proximal elements is generally tolerated, but separations of several tens of base pairs may decrease transcription.

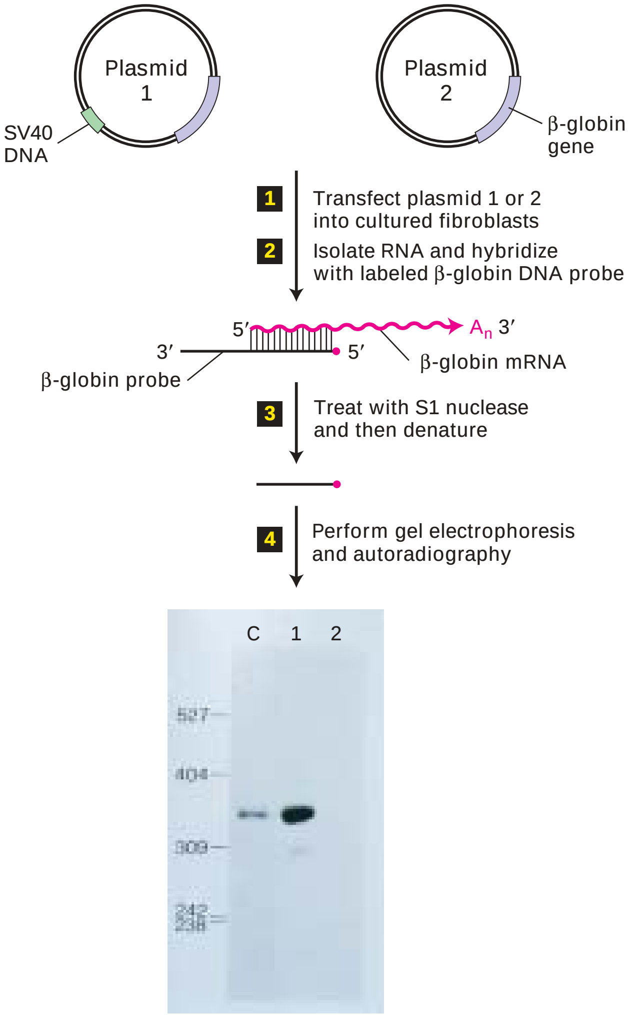


▲ EXPERIMENTAL FIGURE 11-10 **Linker scanning mutations identify transcription-control elements.** (a) A region of eukaryotic DNA (orange) that supports high-level expression of a reporter gene (light blue) is cloned in a plasmid vector as diagrammed at the top. Overlapping linker scanning (LS) mutations (crosshatch) are introduced from one end of the region being analyzed to the other. These mutations result from scrambling the nucleotide sequence in a short stretch of the DNA. After the mutant plasmids are transfected separately into cultured cells, the activity of the reporter-gene product is assayed. In the hypothetical example shown here, LS mutations 1, 4, 6, 7, and 9 have little or no effect on expression of the reporter gene, indicating that the regions altered in these mutants contain no control elements. Reporter-gene expression is significantly reduced in mutants 2, 3, 5, and 8, indicating that control elements (brown) lie in the intervals shown at the bottom. (b) Analysis of LS mutations in the transcription-control region of the thymidine kinase (*tk*) gene from herpes simplex virus (HSV) identified a TATA box and two promoter-proximal elements (PE-1 and PE-2). [Part (b) see S. L. McKnight and R. Kingsbury, 1982, *Science* **217**:316.]

Distant Enhancers Often Stimulate Transcription by RNA Polymerase II

As noted earlier, transcription from many eukaryotic promoters can be stimulated by control elements located thousands of base pairs away from the start site. Such long-distance transcription-control elements, referred to as **enhancers,** are common in eukaryotic genomes but fairly rare in bacterial genomes. The first enhancer to be discovered that stimulates transcription of eukaryotic genes was in a 366-bp fragment of the simian virus 40 (SV40) genome (Figure 11-11). Further analysis of this region of SV40 DNA revealed that an ≈100-bp sequence lying ≈100 base pairs upstream of the SV40 early transcription start site was responsible for its ability to enhance transcription. In SV40, this enhancer sequence functions to stimulate transcription from viral promoters. The SV40 enhancer, however, stimulates transcription from all mammalian promoters that have been tested when it is inserted in either orientation anywhere on a plasmid carrying the test promoter, even when it is thousands of base pairs from the start site. An extensive linker scanning mutational analysis of the SV40 enhancer indicated that it is composed of multiple individual elements, each of which contributes to the total activity of the enhancer. As discussed later, each of these regulatory elements is a protein-binding site.

Soon after discovery of the SV40 enhancer, enhancers were identified in other viral genomes and in eukaryotic


Plasmid 1
SV40 DNA
Plasmid 2
β-globin gene
1 Transfect plasmid 1 or 2 into cultured fibroblasts
2 Isolate RNA and hybridize with labeled β-globin DNA probe
5′
A_n 3′
3′
5′
β-globin probe
β-globin mRNA
3 Treat with S1 nuclease and then denature
4 Perform gel electrophoresis and autoradiography
C
1
2
527
404
309
242
238

▲ EXPERIMENTAL FIGURE 11-11 **Plasmids containing a particular SV40 DNA fragment showed marked increase in mRNA production compared with plasmids lacking this enhancer.** Plasmids containing the β-globin gene with or without a 366-bp fragment of SV40 DNA were constructed. These plasmids were transfected into cultured cells, and any resulting RNA was hybridized to a β-globin DNA probe (steps 1 and 2). The amount of β-globin mRNA synthesized by cells transfected with one or the other plasmid was assayed by the S1 nuclease–protection method (step 3). The restriction-fragment probe, generated from a β-globin cDNA clone, was complementary to the 5′ end of β-globin mRNA. The 5′ end of the probe was labeled with ^{32}P (red dot). Hybridization of β-globin mRNA to the probe protected an ≈340-nucleotide fragment of the probe from digestion by S1 nuclease, which digests single-stranded DNA but not DNA in an RNA-DNA hybrid. Autoradiography of electrophoresed S1-protected fragments (step 4) revealed that cells transfected with plasmid 1 (lane 1) produced much more β-globin mRNA than those transfected with plasmid 2 (lane 2). Lane C is a control assay of β-globin mRNA isolated from reticulocytes, which actively synthesize β-globin. These results show that the SV40 DNA fragment in plasmid 1 contains an element, the enhancer, that greatly stimulates synthesis of β-globin mRNA. [Adapted from J. Banerji et al., 1981, *Cell* **27**:299.]

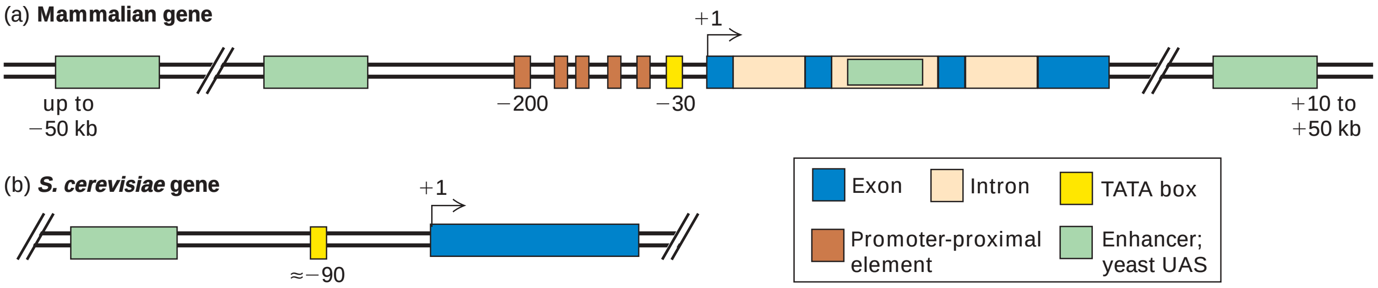
cellular DNA. Some of these control elements are located 50 or more kilobases from the promoter they control. Analyses of many different eukaryotic cellular enhancers have shown that they can occur upstream from a promoter, downstream from a promoter within an intron, or even downstream from the final exon of a gene. Like promoter-proximal elements, many enhancers are cell-type-specific. For example, the genes encoding antibodies (immunoglobulins) contain an enhancer within the second intron that can stimulate transcription from all promoters tested, but only in B lymphocytes, the type of cells that normally express antibodies. Analyses of the effects of deletions and linker scanning mutations in cellular enhancers have shown that, like the SV40 enhancer, they generally are composed of multiple elements that contribute to the overall activity of the enhancer.

Most Eukaryotic Genes Are Regulated by Multiple Transcription-Control Elements

Initially, enhancers and promoter-proximal elements were thought to be distinct types of transcription-control elements. However, as more enhancers and promoter-proximal elements were analyzed, the distinctions between them became less clear. For example, both types of element generally can stimulate transcription even when inverted, and both types often are cell-type-specific. The general consensus now is that a spectrum of control elements regulates transcription by RNA polymerase II. At one extreme are enhancers, which can stimulate transcription from a promoter tens of thousands of base pairs away (e.g., the SV40 enhancer). At the other extreme are promoter-proximal elements, such as the upstream elements controlling the HSV *tk* gene, which lose their influence when moved an additional 30–50 base pairs farther from the promoter. Researchers have identified a large number of transcription-control elements that can stimulate transcription from distances between these two extremes.

Figure 11-12a summarizes the locations of transcription-control sequences for a hypothetical mammalian gene. The start site at which transcription initiates encodes the first (5′) nucleotide of the first exon of an mRNA, the nucleotide that is capped. For many genes, especially those encoding abundantly expressed proteins, a TATA box located approximately 25–35 base pairs upstream from the start site directs RNA polymerase II to begin transcription at the proper nucleotide. Promoter-proximal elements, which are relatively short (≈10–20 base pairs), are located within the first ≈200 base pairs upstream of the start site. Enhancers, in contrast, usually are ≈100 base pairs long and are composed of multiple elements of ≈10–20 base pairs. Enhancers may be located up to 50 kilobases upstream or downstream from the start site or within an intron. Many mammalian genes are controlled by more than one enhancer region.

The *S. cerevisiae* genome contains regulatory elements called **upstream activating sequences (UASs),** which function



▲ FIGURE 11-12 **General pattern of control elements that regulate gene expression in multicellular eukaryotes and yeast.** (a) Genes of multicellular organisms contain both promoter-proximal elements and enhancers, as well as a TATA box or other promoter element. The promoter elements position RNA polymerase II to initiate transcription at the start site and influence the rate of transcription. Enhancers may be either upstream or downstream and as far away as 50 kb from the transcription start site. In some cases, enhancers lie within introns. For some genes, promoter-proximal elements occur downstream from the start site as well as upstream. (b) Most *S. cerevisiae* genes contain only one regulatory region, called an *upstream activating sequence (UAS),* and a TATA box, which is ≈90 base pairs upstream from the start site.

similarly to enhancers and promoter-proximal elements in higher eukaryotes. Most yeast genes contain only one UAS, which generally lies within a few hundred base pairs of the start site. In addition, *S. cerevisiae* genes contain a TATA box ≈90 base pairs upstream from the transcription start site (Figure 11-12b).

KEY CONCEPTS OF SECTION 11.2

Regulatory Sequences in Protein-Coding Genes

- Expression of eukaryotic protein-coding genes generally is regulated through multiple protein-binding control regions that are located close to or distant from the start site (Figure 11-12).

- Promoters direct binding of RNA polymerase II to DNA, determine the site of transcription initiation, and influence transcription rate.

- Three principal types of promoter sequences have been identified in eukaryotic DNA. The TATA box, the most common, is prevalent in rapidly transcribed genes. Initiator promoters are found in some genes, and CpG islands are characteristic of genes transcribed at a low rate.

- Promoter-proximal elements occur within ≈200 base pairs upstream of a start site. Several such elements, containing ≈10–20 base pairs, may help regulate a particular gene.

- Enhancers, which contain multiple short control elements, may be located from 200 base pairs to tens of kilobases upstream or downstream from a promoter, within an intron, or downstream from the final exon of a gene.

- Promoter-proximal elements and enhancers often are cell-type-specific, functioning only in specific differentiated cell types.

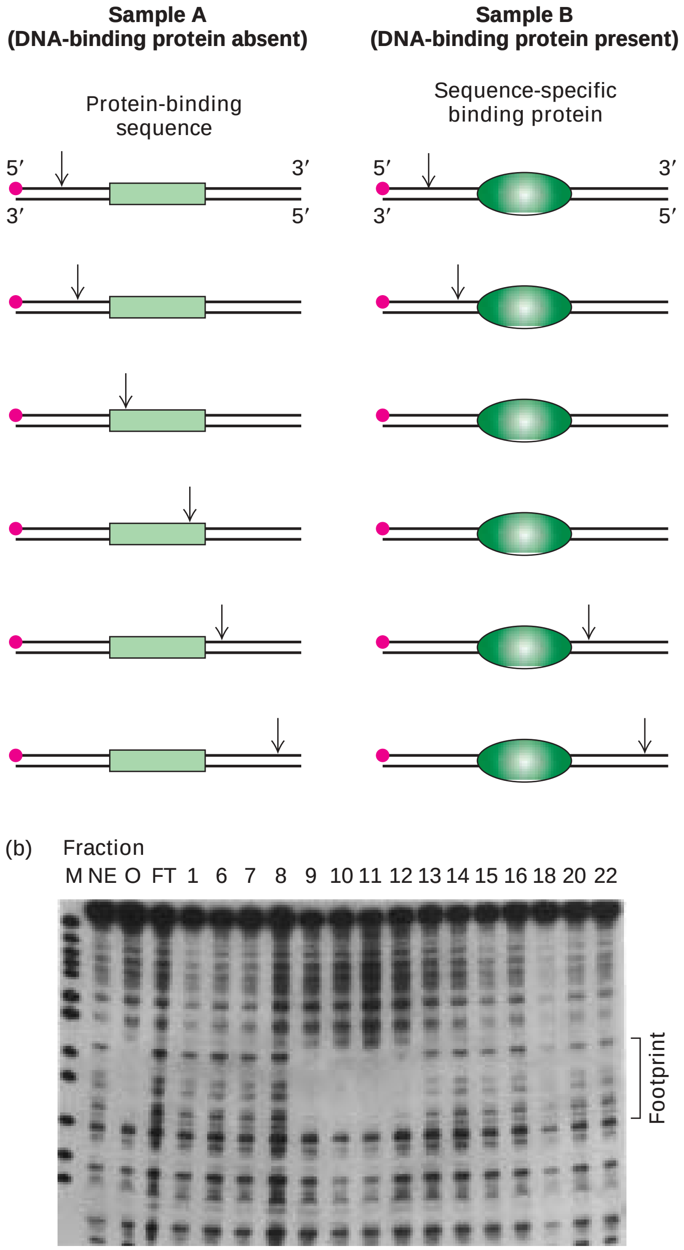
11.3 Activators and Repressors of Transcription

The various transcription-control elements found in eukaryotic DNA are binding sites for regulatory proteins. In this section, we discuss the identification, purification, and structures of these transcription factors, which function to activate or repress expression of eukaryotic protein-coding genes.

Footprinting and Gel-Shift Assays Detect Protein-DNA Interactions

In yeast, *Drosophila,* and other genetically tractable eukaryotes, numerous genes encoding transcriptional activators and repressors have been identified by classical genetic analyses like those described in Chapter 9. However, in mammals and other vertebrates, which are less amenable to such genetic analysis, most transcription factors have been detected initially and subsequently purified by biochemical techniques. In this approach, a DNA regulatory element that has been identified by the kinds of mutational analyses described in the previous section is used to identify *cognate* proteins that bind specifically to it. Two common techniques for detecting such cognate proteins are **DNase I footprinting** and the **electrophoretic mobility shift assay.**

DNase I footprinting takes advantage of the fact that when a protein is bound to a region of DNA, it protects that DNA sequence from digestion by nucleases. As illustrated in Figure 11-13, when samples of a DNA fragment that is labeled at one end are digested in the presence and absence of a DNA-binding protein and then denatured, electrophoresed, and the resulting gel subjected to autoradiography, the region protected by the bound protein appears as a gap, or "footprint," in the array of bands resulting from digestion in the absence of protein. When footprinting is performed

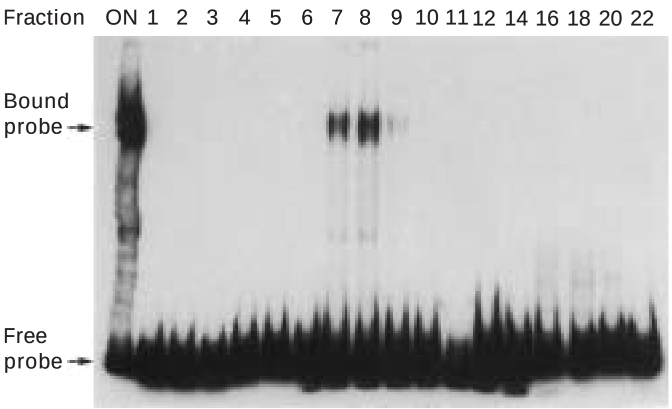

Sample A
(DNA-binding protein absent)
Sample B
(DNA-binding protein present)
Protein-binding sequence
Sequence-specific binding protein
5′
3′
3′
5′
5′
3′
3′
5′
(b)
Fraction
M NE O FT 1 6 7 8 9 10 11 12 13 14 15 16 18 20 22
Footprint

▶ EXPERIMENTAL FIGURE 11-13 **DNase I footprinting reveals control-element sequences and can be used as an assay in transcription factor purification.** (a) DNase I footprinting can identify control element sequences. A DNA fragment known to contain the control-element is labeled at one end with ^{32}P (red dot). Portions of the labeled DNA sample then are digested with DNase I in the presence and absence of protein samples thought to contain a cognate protein. DNase I randomly hydrolyzes the phosphodiester bonds of DNA between the 3′ oxygen on the deoxyribose of one nucleotide and the 5′ phosphate of the next nucleotide. A low concentration of DNase I is used so that on average each DNA molecule is cleaved just once (vertical arrows). If the protein sample does not contain a cognate DNA-binding protein, the DNA fragment is cleaved at multiple positions between the labeled and unlabeled ends of the original fragment, as in sample A on the left. If the protein sample contains a cognate protein, as in sample B on the right, the protein binds to the DNA, thereby protecting a portion of the fragment from digestion. Following DNase treatment, the DNA is separated from protein, denatured to separate the strands, and electrophoresed. Autoradiography of the resulting gel detects only labeled strands and reveals fragments extending from the labeled end to the site of cleavage by DNase I. Cleavage fragments containing the control sequence show up on the gel for sample A, but are missing in sample B because the bound cognate protein blocked cleavages within that sequence and thus production of the corresponding fragments. The missing bands on the gel constitute the footprint. (b) A protein fraction containing a sequence-specific DNA-binding protein can be purified by column chromatography. DNase I footprinting can then identify which of the eluted fractions contain the cognate protein. In the absence of added protein (NE, *n*o extract), DNase I cleaves the DNA fragment at multiple sites, producing multiple bands on the gel shown here. A cognate protein present in the nuclear extract applied to the column (O, *o*nput) generated a footprint. This protein was bound to the column, since footprinting activity was not detected in the flow-through protein fraction (FT). After applying a salt gradient to the column, most of the cognate protein eluted in fractions 9–12, as evidenced by the missing bands (footprints). The sequence of the protein-binding region can be determined by comparison with marker DNA fragments of known length analyzed on the same gel (M). [Part (b) from S. Yoshinaga et al., 1989, *J. Biol. Chem.* **264**:10529.]

with a DNA fragment containing a known DNA control element, the appearance of a footprint indicates the presence of a transcription factor that binds that control element in the protein sample being assayed. Footprinting also identifies the specific DNA sequence to which the transcription factor binds.

The electrophoretic mobility shift assay (EMSA), also called the *gel-shift* or *band-shift* assay, is more useful than the footprinting assay for quantitative analysis of DNA-binding proteins. In general, the electrophoretic mobility of a DNA fragment is reduced when it is complexed to protein, causing a shift in the location of the fragment band. This assay can be used to detect a transcription factor in protein fractions incubated with a radiolabeled DNA fragment containing a known control element (Figure 11-14).

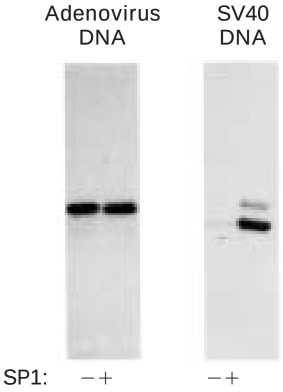
In the biochemical isolation of a transcription factor, an extract of cell nuclei commonly is subjected sequentially to several types of column chromatography (Chapter 3). Fractions eluted from the columns are assayed by DNase I footprinting or EMSA using DNA fragments containing an identified regulatory element (see Figures 11-13 and 11-14). Fractions containing protein that binds to the regulatory element in these assays probably contain a putative transcription factor. A powerful technique commonly used for the final step in purifying transcription factors is *sequence-specific DNA affinity chromatography,* a particular type of



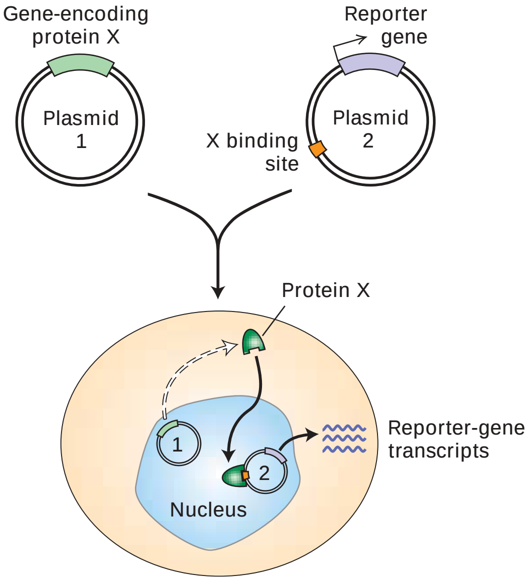
▶ EXPERIMENTAL FIGURE 11-14 **Electrophoretic mobility shift assay can be used to detect transcription factors during purification.** In this example, protein fractions separated by column chromatography were assayed for their ability to bind to a radiolabeled DNA-fragment probe containing a known regulatory element. After an aliquot of the protein sample loaded onto the column (ON) and successive column fractions (numbers) were incubated with the labeled probe, the samples were electrophoresed under conditions that do not denature proteins. The free probe not bound to protein migrated to the bottom of the gel. A protein in the preparation applied to the column and in fractions 7 and 8 bound to the probe, forming a DNA-protein complex that migrated more slowly than the free probe. These fractions therefore likely contain the regulatory protein being sought. [From S. Yoshinaga et al., 1989, *J. Biol. Chem.* **264**:10529.]

affinity chromatography in which long DNA strands containing multiple copies of the transcription factor–binding site are coupled to a column matrix. As a final test that an isolated protein is in fact a transcription factor, its ability to modulate transcription of a template containing the corresponding protein-binding sites is assayed in an in vitro transcription reaction. Figure 11-15 shows the results of such an assay for SP1, a transcription factor that binds to GC-rich sequences, thereby activating transcription from nearby promoters.

Once a transcription factor is isolated and purified, its partial amino acid sequence can be determined and used to clone the gene or cDNA encoding it, as outlined in Chapter 9. The isolated gene can then be used to test the ability of the encoded protein to activate or repress transcription in an in vivo transfection assay (Figure 11-16).



▲ EXPERIMENTAL FIGURE 11-15 **Transcription factors can be identified by in vitro assay for transcription activity.** SP1 was identified based on its ability to bind to a region of the SV40 genome that contains six copies of a GC-rich promoter-proximal element and was purified by column chromatography. To test the transcription-activating ability of purified SP1, it was incubated in vitro with template DNA, a protein fraction containing RNA polymerase II and associated general transcription factors, and labeled ribonucleoside triphosphates. The labeled RNA products were subjected to electrophoresis and autoradiography. Shown here are autoradiograms from assays with adenovirus and SV40 DNA in the absence (−) and presence (+) of SP1. SP1 had no significant effect on transcription from the adenovirus promoter, which contains no SP1-binding sites. In contrast, SP1 stimulated transcription from the SV40 promoter about tenfold. [Adapted from M. R. Briggs et al., 1986, *Science* **234**:47.]

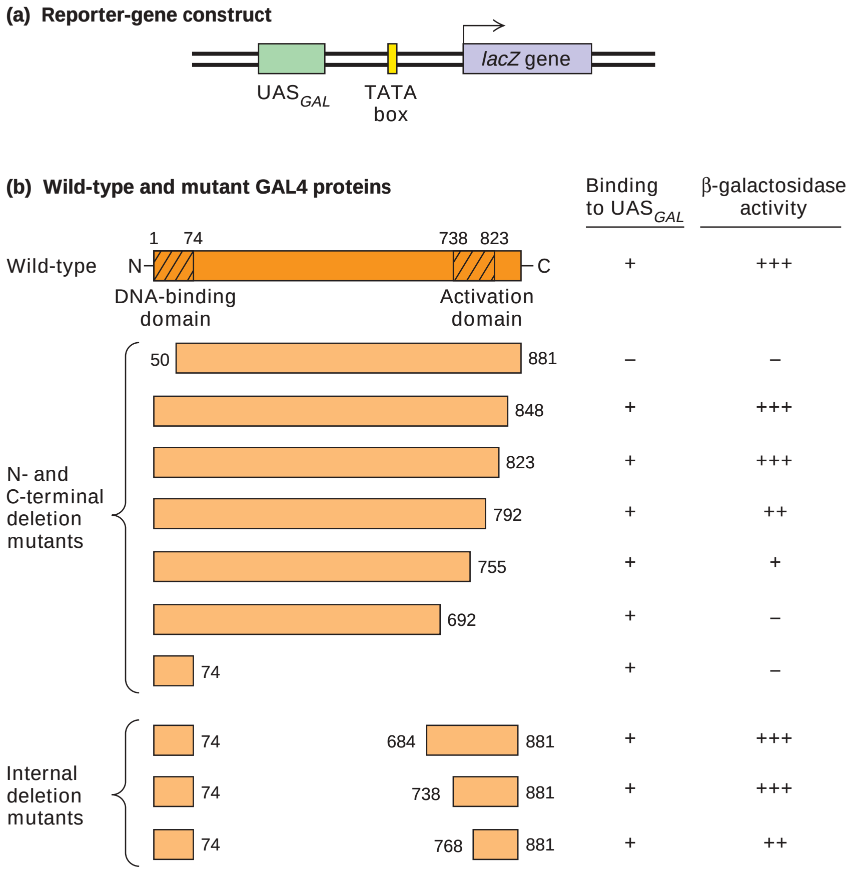


▲ EXPERIMENTAL FIGURE 11-16 **In vivo transfection assay measures transcription activity to evaluate proteins believed to be transcription factors.** The assay system requires two plasmids. One plasmid contains the gene encoding the putative transcription factor (protein X). The second plasmid contains a reporter gene (e.g., *lacZ*) and one or more binding sites for protein X. Both plasmids are simultaneously introduced into cells that lack the gene encoding protein X. The production of reporter-gene RNA transcripts is measured; alternatively, the activity of the encoded protein can be assayed. If reporter-gene transcription is greater in the presence of the X-encoding plasmid, then the protein is an activator; if transcription is less, then it is a repressor. By use of plasmids encoding a mutated or rearranged transcription factor, important domains of the protein can be identified.

Activators Are Modular Proteins Composed of Distinct Functional Domains

Studies with a yeast transcription activator called GAL4 provided early insight into the domain structure of transcription factors. The gene encoding the GAL4 protein, which promotes expression of enzymes needed to metabolize galactose, was identified by complementation analysis of *gal4* mutants (Chapter 9). Directed mutagenesis studies like those described previously identified UASs for the genes activated by GAL4. Each of these UASs was found to contain one or more copies of a related 17-bp sequence called UAS$_{GAL}$. DNase I footprinting assays with recombinant GAL4 protein produced in *E. coli* from the yeast GAL4 gene showed that GAL4 protein binds to UAS$_{GAL}$ sequences. When a copy of UAS$_{GAL}$ was cloned upstream of a TATA box followed by a *lacZ* reporter gene, expression of *lacZ* was activated in galactose media in wild-type cells, but not in *gal4* mutants. These results showed that UAS$_{GAL}$ is a transcription-control element activated by the GAL4 protein in galactose media.

A remarkable set of experiments with *gal4* deletion mutants demonstrated that the GAL4 transcription factor is composed of separable functional domains: an N-terminal **DNA-binding domain,** which binds to specific DNA sequences, and a C-terminal **activation domain,** which interacts with other proteins to stimulate transcription from a nearby promoter (Figure 11-17). When the N-terminal DNA-binding domain of GAL4 was fused directly to various

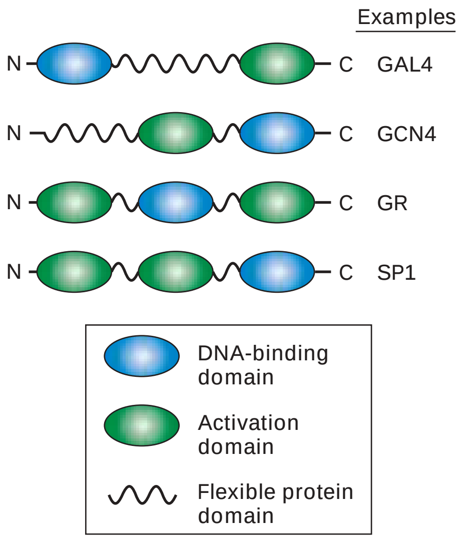


▲ EXPERIMENTAL FIGURE 11-17 **Deletion mutants of the GAL4 gene in yeast with a UAS$_{GAL}$ reporter-gene construct demonstrate the separate functional domains in an activator.** (a) Diagram of DNA construct containing a *lacZ* reporter gene and TATA box ligated to UAS$_{GAL}$, a regulatory element that contains several GAL4-binding sites. The reporter-gene construct and DNA encoding wild-type or mutant (deleted) GAL4 were simultaneously introduced into mutant (*gal4*) yeast cells, and the activity of β-galactosidase expressed from *lacZ* was assayed. Activity will be high if the introduced *GAL4* DNA encodes a functional protein. (b) Schematic diagrams of wild-type GAL4 and various mutant forms. Small numbers refer to positions in the wild-type sequence. Deletion of 50 amino acids from the N-terminal end destroyed the ability of GAL4 to bind to UAS$_{GAL}$ and to stimulate expression of β-galactosidase from the reporter gene. Proteins with extensive deletions from the C-terminal end still bound to UAS$_{GAL}$. These results localize the DNA-binding domain to the N-terminal end of GAL4. The ability to activate β-galactosidase expression was not entirely eliminated unless somewhere between 126–189 or more amino acids were deleted from the C-terminal end. Thus the activation domain lies in the C-terminal region of GAL4. Proteins with internal deletions (*bottom*) also were able to stimulate expression of β-galactosidase, indicating that the central region of GAL4 is not crucial for its function in this assay. [See J. Ma and M. Ptashne, 1987, *Cell* **48**:847; I. A. Hope and K. Struhl, 1986, *Cell* **46**:885; and R. Brent and M. Ptashne, 1985, *Cell* **43**:729.]

of its C-terminal fragments, the resulting truncated proteins retained the ability to stimulate expression of a reporter gene in an in vivo assay like that depicted in Figure 11-16. Thus the internal portion of the protein is not required for functioning of GAL4 as a transcription factor. Similar experiments with another yeast transcription factor, GCN4, which regulates genes required for synthesis of many amino acids, indicated that it contains an ≈60-aa DNA-binding domain at its C-terminus and an ≈20-aa activation domain near the middle of its sequence.

Further evidence for the existence of distinct activation domains in GAL4 and GCN4 came from experiments in which their activation domains were fused to a DNA-binding domain from an entirely unrelated *E. coli* DNA-binding protein. When these fusion proteins were assayed in vivo, they activated transcription of a reporter gene containing the cognate site for the *E. coli* protein. Thus functional transcription factors can be constructed from entirely novel combinations of prokaryotic and eukaryotic elements.

Studies such as these have now been carried out with many eukaryotic activators. The structural model of eukaryotic activators that has emerged from these studies is a modular one in which one or more activation domains are connected to a sequence-specific DNA-binding domain through flexible protein domains (Figure 11-18). In some cases, amino acids included in the DNA-binding domain also contribute to transcriptional activation. As discussed in a later section, activation domains are thought to function by binding other proteins involved in transcription. The presence of flexible domains connecting the DNA-binding domains to activation domains may explain why alterations in the spacing between control elements are so well tolerated in eukaryotic control regions. Thus even when the positions of transcription factors bound to DNA are shifted relative to each other, their activation domains may still be able to interact because they are attached to their DNA-binding domains through flexible protein regions.



▲ FIGURE 11-18 **Schematic diagrams illustrating the modular structure of eukaryotic transcription activators.** These transcription factors may contain more than one activation domain but rarely contain more than one DNA-binding domain. GAL4 and GCN4 are yeast transcription activators. The glucocorticoid receptor (GR) promotes transcription of target genes when certain hormones are bound to the C-terminal activation domain. SP1 binds to GC-rich promoter elements in a large number of mammalian genes.