19: Transcriptional regulation in eukaryotes

Promoters
1. Eukaryotic genes differ in their state of expression

![Diagram of expression states of promoters for RNA polymerase II](https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_IV%3A_Regul...)

Figure 4.5.1. Expression states of promoters for RNA polymerase II. Each of these states has been described for particular genes, but it is not clear that all states are in one obligatory pathway. For instance, it possible that some gene activation events could go from silent chromatin to basal transcription without passing through open but repressed and paused transcription.

a. **Basal transcription**

1. Is frequently studied by *in vitro* transcription, using defined templates and either extracts from nuclei or purified components.
2. Requires RNA polymerase with general transcription factors (e.g. TFIID, TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH for RNA polymerase II), as previously covered in Part Three.

b. **Activated transcription**

1. Occurs via transcriptional activators interacting directly or indirectly with the general transcription complex to increase the efficiency of initiation.
2. The transcriptional activators may bind to specific DNA sequences in the upstream promoter elements, or they may bind to enhancers (see Section B below).
3. The basic idea is to increase the local concentration of the general transcription factors so the initiation complex can be assembled more readily. The fact that the activators are bound to DNA that is close to the target (or becomes close because of looping of the DNA) means that the local concentration of that protein is high, and hence it can boost the local concentration of the interacting general transcription factors.
3. Stalled polymerases

RNA polymerase will transcribe about 20 to 40 nucleotides at the start of some genes and then stall at a pause site. The classic example are heat-shock genes in Drosophila, but other cases are also known. These genes are activated by release of stalled polymerases to elongate. In the case of the heat shock genes, this requires heat shock transcription factor (HSTF). The mechanism is still under study; some interesting ideas are:

1. Phosphorylation of the CTD of the large subunit of RNA polymerase II causes release to elongation ("promoter clearance"). One candidate (but not the only one) for the CTD kinase is TFIIH.
2. Addition of a processivity factor (analogous to *E. coli* Nus A?), maybe TFIIIS.

B. Silencers

Figure 4.5.2. Transcriptionally silent chromatin, mediated by Rap1 and SIR proteins.

C. Enhancers

1. Enhancers are **cis-acting regulatory sequences that increase level of expression of a gene**, but they operate **independently of position and orientation**. These last two operational criteria distinguish enhancers from promoters.
2. Examples

   a. **SV40 control region**

   1. SV40 (simian virus 40) infects monkey kidney cells, and it will also cause transformation of rodent cells. It has a double stranded DNA genome of about 5 kb. Because of its involvement in tumorigenesis, it has been a favorite subject of molecular virologists. The early region encodes tumor antigens (T-Ag and t-Ag) with many functions, including stimulating DNA replication of SV40 and blocking the action of endogenous tumor suppressors like p53 (the 1993 "Molecule of the Year"). The late region encodes three capsid proteins called VP1, VP2 and VP3 (viral protein n). A region between the early and late genes controls both replication and transcription of both classes of genes.
   2. The control region has an origin of replication with binding sites for T-Ag.
a. Wild type SV40 expresses T-Ag upon infection of monkey cells and lyses infected cells. However, a viral strain lacking the 72 bp repeats shows a highly reduced level of T-Ag and rarely lyses infected cells.

b. If the 72 bp repeats are added back to the mutant SV40 genome, except they are placed between the ends of the early and late genes (180° from their wild-type position), T-Ag is expressed at a high level and one obtains productive infections.

c. If the orientation of the 72 bp repeats is reversed, one still gets high level expression of viral genes and productive infection. In fact, it is needed for expression of the late genes in the wild-type, which are transcribed in the opposite direction from the early genes.

d. One concludes that the enhancer is needed for efficient transcription of the target promoters, but it can act in either orientation and at a variety of different positions and distances from the targets.

e. Work done virtually concurrently with that described above showed that the 72 bp repeats work on other "heterologous" genes, so that, for example b-globin genes could be expressed in nonerythroid cells. In fact this was one of the key observations in the discovery of the enhancer.

f. One copy of the 72 bp region will work as an enhancer, but two copies work better.

b. Immunoglobulin genes

1. This was the first enhancer of a cellular gene discovered. Researchers noted that a region of the intron was exceptionally well conserved among human, rabbit and mouse sequences, and subsequent deletion experiments showed that the intronic enhancer was required for expression.

2. After rearrangement of the immunoglobulin gene to fuse VDJ regions, one is left with a large intron between this combined variable region gene and the constant region. An enhancer is found in that intron, and another enhancer is found 3’ to the polyA addition site.
Figure 4.5.5. Enhancers in the intron and 3' flank of an immunoglobulin gene.

(3) The enhancers have multiple binding sites for transcriptional regulatory proteins

(a) Several of these sites are named for the enhancer they were discovered in. E.g. mE1, mE2, etc. are binding sites for enhancer proteins identified in the gene for the immunoglobulin heavy chain m (mu).

The protein YY1 (ying yang 1) binds to the mE1 site (CCAT is the core of the consensus) and bends DNA there.

The octamer site (ATTTGCAT) is bound by two related proteins. Oct1 is found in all tissues examined, whereas Oct2 is lymphoid specific - the first example of a tissue-specific transcription factor. Transcriptional activators that do not have their own DNA binding sequence, like VP16 from Herpes virus, will bind to Oct proteins, which bind to DNA, and the complex can activate transcription.

(b) Some proteins will bind to sites both in the promoter and the enhancer, e.g. Oct proteins. Remember Oct1 also acts at the SV40 enhancer.

c. Summary

1. The position of the enhancer can be virtually anywhere relative to the gene, but the promoter is always at the 5' end.
2. Examples are known of enhancers 5' to the gene (upstream), adjacent to the promoter (like in SV40), downstream from the gene (some globin genes), within the gene (immunoglobulins) or far upstream within a locus control region (globin genes, see Chapter 20.)

Enhancers can occur in a variety of positions with respect to genes

- Enhancer
- Upstream
- Adjacent
- Downstream
- Internal
- Distal

Figure 4.5.6.
3. Multiple binding sites for transcriptional activators

a. All enhancers characterized thus far have multiple binding sites for activator proteins.

b. Multiples of binding sites are needed for function of the enhancer.

1. In experiments with the SV40 enhancer, it was noted that some mutations that decreased the infectivity of the virus caused a mutation of one of the domains of the enhancer, e.g. domain A. When these mutants were then selected for pseudo-revertants to wild-type, with infectivity largely restored, it was found that the pseudo-revertants had duplicated one of the remaining domains. Subsequently, multimers of the various protein-binding sites were shown to be active, but monomers had little activity.

2. The domain (e.g. A, C and B in the SV40 enhancer) with at least two binding sites is called an enhanson. Multiple enhansons make up an enhancer.

D. DNA binding domains

Computer-assisted three-dimensional views of several transcription factors, illustrating many of the domains described here, can be viewed as Chime tutorials at

- [www.bmb.psu.edu/pugh/514/mdls](http://www.bmb.psu.edu/pugh/514/mdls)
- [www.clunet.edu/BioDev/OMM/cro/cromast.htm](http://www.clunet.edu/BioDev/OMM/cro/cromast.htm)
1. **Helix-turn-helix, homeodomain**

1. The sequence of the "homeodomain" forms three helices separated by tight turns.
2. Helix three occupies the major groove at the binding site on the DNA. It is the recognition helix, forming specific interactions (H-bonds and hydrophobic interactions) with the edges of the base pairs in the major groove.
3. Helices one and two are perpendicular to and above helix three, providing alignment with the phosphodiester backbone. The N-terminal tail of helix interacts with the minor groove of the DNA on the opposite face of the DNA.
4. Helix two + helix three is comparable to the helix-turn-helix motif first identified in the I Cro and repressor system.

![Figure 4.5.12. Helix-turn-helix in the “homeodomain”](image)

1. **Examples**

(a) Homeotic genes and their relatives.

All these are involved in regulating early developmental events in *Drosophila*. They are transcription factors (regulating the genes that determine the next developmental fate), and they have this same protein motif for their DNA binding domains.

Some specific examples are the products of these genes:

- the pair-rule gene *eve* = *even skipped*
- the segment polarity gene *en* = *engrailed*
- the homeotic gene *Antp* = *antennapedia*
(d) In a protein with 3 adjacent Zn fingers, e.g. Sp1 (remember this protein from the SV40 early promoter), each finger binds in the major groove to contact three adjacent base pairs. For the high affinity binding site, one finger contacts GGG, the next finger contacts GCG, and the remaining finger contacts GGG. So the three fingers curve along to contact the major groove for most of one turn of the helix.

(e) Members of this class of Zn finger proteins have multiple fingers, usually in a tandem array. Examples include TFIIIA (the motif was discovered in this protein) with 9 fingers, a CAC-binding protein (related to some extent to Sp1) with 3 fingers, and Drosophila ADR1 with 2 fingers.

(2) Cys2Cys2

(a) Consensus sequence:

Cys-X2-Cys-X1-3-Cys-X2-Cys

(b) Forms a distinctly different structure from the Cys2His2 Zn fingers.

1. Note that the number of amino acids between the 2 "halves" of the finger (1 to 3 in this case) is much less than the 12 that separate the two halves of a Cys2His2 Zn finger.
2. The Cys2Cys2 fingers are not interchangable with Cys2His2 Zn fingers in domain swap experiments.
3. The proteins do not have extensive repetitions of the motif, in contrast to proteins with Cys2His2 Zn fingers.
Figure 4.5.15. Basic helix-loop-helix proteins

(3) Examples include heterodimers that can exchange partners

(a) MyoD is a key protein in commitment of mesodermal tissues to muscle differentiation. Other relatives, such as myogenin and myf5, are equally important and provide redundant functions. All are muscle-specific and have a similar binding domain. MyoD is active when it has E12 or E47 as its heterodimeric partner; when active it will stimulate transcription of muscle specific genes such as the one encoding creatine kinase. E12 and E47 were initially discovered as proteins that bound to enhancers of immunoglobulin genes, but are found in virtually all cell-types. Another protein, called Id, can also bind to E12 or E47 by its HLH domain. However, Id lacks a basic domain, so heterodimers with Id are not active. So the activity of bHLH proteins can be regulated by exchange of partners.

(b) A developing theme is that one of partners of a bHLH heterodimer is ubiquitous (e.g. E12, E47 in mammals, da = daughterless in Drosophila) and the other is tissue-specific (MyoD or AC-S = achaete-scute, a regulator of neurogenesis in Drosophila). The ubiquitous components may be involved in regulating a variety of other tissue-specific proteins with bHLH domains.

(c) Myc, one of many regulators of the cell cycle, is a bHLH protein. It forms partners with Max, and it is possible that this is important in regulation of the cell cycle.

E. Transcriptional activation domains

1. Acidic

This domain has been postulated to be an "acid blob" or an amphipathic helix with acidic residues on one face. Recent physico-chemical studies of GAL4 have shown b-sheet structure. At this point no single structure has been established. Examples:

GAL4 protein, VP16, GCN4, glucocorticoid hormone receptor, AP1, and the I repressor (activation of PRM).
2. Gln-rich

This domain is rich in glutamine, as its name implies. Examples of proteins containing the domain are Sp1, Antp, Oct1 and Oct2.

3. Pro-rich

Again, the domain is rich in proline. Examples include CTF/NF1 (involved in regulation of replication as nuclear factor 1, and proposed to be one of many proteins binding to CCAAT motifs).

4. Work so far has not established well-defined secondary or tertiary structures for these domains.

One possibility is that the activation domains assume their proper structure after binding to its target, i.e. an induced fit model.

Table 4.5.1. Selected eukaryotic transcription factors and their properties

<table>
<thead>
<tr>
<th>Name</th>
<th>System</th>
<th>Binding site (top strand)</th>
<th>Quaternary structure</th>
<th>DNA binding domain</th>
<th>Activation domain</th>
<th>Other comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engrailed</td>
<td>early development</td>
<td></td>
<td></td>
<td>homeodomain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp1</td>
<td>SV40, cellular housekeeping genes</td>
<td>GGGGCGGGG</td>
<td>monomer</td>
<td>3 Zn fingers</td>
<td>Gln-rich</td>
<td>phosphoprotein</td>
</tr>
<tr>
<td>AP1</td>
<td>SV40, cellular enhancers</td>
<td>TGASTCA</td>
<td>heterodimer, Jun-Fos, Jun2, others</td>
<td>basic region + Leu zipper</td>
<td>acidic</td>
<td>regulated by phosphorylation</td>
</tr>
<tr>
<td>Oct1</td>
<td>lymphoid and other genes</td>
<td>ATTTGCAT</td>
<td>monomer, but can bind VP16</td>
<td>POU domain + homeodomain (HTH)</td>
<td>Gln-rich, also binds VP16</td>
<td>Oct1 is ubiquitous, Oct2 is lymphoid specific</td>
</tr>
<tr>
<td>GAL4</td>
<td>yeast galactose regulon</td>
<td>CGGAGCAGGT</td>
<td></td>
<td></td>
<td>Zn2Cys6, binuclear cluster</td>
<td>acidic</td>
</tr>
<tr>
<td>Protein</td>
<td>Function</td>
<td>Sequence</td>
<td>Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucocorticoid receptor</td>
<td>Responsive genes</td>
<td>TGGTACAAATGTTCT</td>
<td>2 Zn fingers, close to Zn finger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MyoD</td>
<td>Determination of myogenesis</td>
<td>CAGCTG</td>
<td>Heterodimer with E12/E47: active; basic-helix-loop-helix</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMG(I)Y</td>
<td>Interferon gene and others</td>
<td>Minor groove</td>
<td>Homodimer, inactive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP16</td>
<td>Herpes simplex virus</td>
<td>Not bind tightly to DNA</td>
<td>Binds to proteins like Oct1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**With Enhancer**

- Polymerase density and amount of transcription increases in all cells in a population

**Without Enhancer**

- Polymerase & transcript

---

**Figure 4.5.18.**

- a. In looping models, the activators bound to the enhancer are brought in close proximity to their targets at the promoter.
by forming loops in the DNA.

1. The activators can make direct contact with their target (perhaps the pre-initiation complex), or they may operate through an intermediary called a co-activator or mediator.
2. If a loop is formed, in principle it does not matter how large the loop is or if the activator binding site is 5' or 3' to the target. This could explain the ability of enhancers to operate independently of position.

![Tracking for activation](image)

Figure 4.5.19.

c. The looping model is favored at this time. However, it has been difficult to design experiments that definitely rule out tracking. Several observations show that DNA can form loops *in vitro*, allowing contact between proteins at the enhancer and those at the promoter. For instance:

1. Using electron microscopy, one can visualize loops of DNA held together by interactions between enhancer-bound activator proteins and proteins bound to the promoter.
2. The biochemical approaches show that the activation domains of transcription factors can bind to components of the pre-initiation complex, such as TFIID (see Section H).

![Promoter/Enhancer for IFNB](image)
b. E.g. the enhancer for the interferon-beta gene, which is located just upstream from the promoter, has binding sites for three dimeric "conventional" transcription factors: NFκB (p50 + p65), IRF, and a heterodimer of ATF2 + Jun (a relative of AP1). In addition, there are three specific binding sites for HMGI(Y).

1. HMGI(Y) is a member of the "high mobility group" of nonhistone chromosomal proteins. Most HMG proteins are abundant in the nucleus, albeit not as abundant as histones.
2. HMGI(Y) binds in the minor groove of DNA and bends the DNA.
3. It also makes specific protein-protein contacts with IRF, ATF2 and NFκB, even in the absence of DNA.
4. By bending the DNA at precise positions by a defined amount, and by aiding the binding of other proteins, HMGI(Y) seems to play a critical role in assembly of the enhancer complex in juxtaposition with the promoter.
5. In general, proteins that bend DNA can be the agents that cause the looping to bring the enhancer-binding proteins in proximity to their targets.

c. Other proteins that bend DNA

cAMP-CAP (recall this from catabolite repression in E. coli), IHF = integration host factor (required for integration of λ DNA to form a prophage, via a large complex called an intasome), and YY1 (ying yang 1) which has either negative or positive effects on a large variety of genes in mammals.