Transcriptional Regulation

Statistics 246
Week 11, Lecture 1
Spring 2006

Greg Hather
References

• Terry wrote most of these slides.
• I borrowed pictures and other material from *Molecular Biology of the Cell, 4th* edition, by Bruce Alberts et al.
• Other sources include *Genes and Signals* by Mark Ptashne and Alexander Gann, and some reviews by Robert Tjian.
Why are we interested in this topic?

**Development**  In humans each cell (with some exceptions) contains the same set of genes. What differs is their expression, at different times, in different places, and in different combinations.

**Phenotype**  Differences in transcriptional regulation can cause differences in phenotype.

**Response to Environment**  Transcriptional regulation is one way that organisms respond to stimuli.
Why are we interested in this topic?

**Diseases and Drugs**  Some diseases are caused by problems with control of transcriptional regulation (for example, cancer). Some drugs target transcription factors.

**Evolution**  *Caenorhabditis elegans* has ~20,000 genes, but lacks the range of cell types and tissue of *Drosophila melanogaster*, which has < 14,000 genes. Humans probably have ~30,000 protein coding genes. What is the molecular basis of organismal complexity? One view is that changes in patterns of gene expression (rather than evolution of new genes) has an important, perhaps determinative role in generating diversity.
Transcription
Transcription

• Transcription requires:
  – an RNA polymerase enzyme
  – a DNA template, and plenty of
  – nucleotide triphosphates: ATP, CTP, GTP and UTP.
Transcription

The DNA template opens, and the RNA polymerase moves down one strand in the $3' \rightarrow 5'$ direction inserting ribonucleotides into a growing strand of RNA.

The ribonucleotides come from nucleotide triphosphates ATP, CTP, GTP and UTP, and follow the familiar rules of base pairing.

Synthesis of the RNA proceeds in the $5' \rightarrow 3'$ direction.

As each nucleoside triphosphate is added to the $3'$ end of the growing RNA strand, the two terminal phosphates are removed.
Transcription Factors

A transcription factor is a protein that binds to a DNA in a sequence specific manner to regulate transcription. Transcription factors regulate gene expression by helping (or inhibiting) transcription initiation. For example, a transcription factor could increase gene expression by helping to recruit RNA polymerase.

Transcription factors come in two flavors: specific and general.

Specific transcription factors only regulate certain genes.

General transcription factors are proteins found in eukaryotes but not in bacteria. General transcription factors are required for transcription initiation in all genes in eukaryotes. They assemble as part of the transcription initiation complex and then dissociate after transcription begins.
Transcription Regulation

Transcriptional regulation is about controlling how often a given gene is transcribed. Once an RNA polymerase starts to transcribe, it (almost always) continues to transcribe until it reaches the end of the gene.

By making conditions for transcription initiation more favorable or less favorable, transcription factors alter the rate at which transcripts are produced.

Note that transcription and translation are not fast processes. Thus, transcriptional regulation is not suitable for producing quick (ie less than 15 minutes) responses to environmental changes.
Where do transcription factors bind?

In bacteria, the transcription factor binding sites almost always occur close to the transcription start site (ie less than ~200 bp).

In eukaryotes, the transcription factor binding sites tend to occur further away from the transcription start site. In some cases, the transcription factor binding sites can be very far away (eg 10 kb).
Transcriptional regulation is just one type of regulation

- Other types of regulation include
  - RNA splicing
  - RNA degradation
  - Translational regulation
  - Posttranslational modifications
  - Protein localization
  - Allosteric regulation
  - Protein degradation
A DNA binding protein domain

Front view

Side view
Another DNA binding protein domain

Beta sheets can bind to DNA too.
A Heterodimer Binding to DNA
Gene expression and its regulation

In the 1950s, F. Jacob, J. Monod, A. Lwoff and others at the Institut Pasteur discovered through the study of two examples that *E. coli* does not express all its genes all the time. One involved its ability to grow on lactose. They discovered that the gene *lacZ* encoding β-galactosidase is silent until its substrate - lactose - is added. The gene is then transcribed (turned on) and the enzyme is synthesized.

Bacterium mutants showed that it was possible to separate the decision as to which gene to express, from the process of gene expression.
Lessons from bacteria

Let’s begin with *E. coli*’s RNA polymerase. The complex is called a holoenzyme, the catalytic core being made up of five polypeptides, one each of the beta (\(\beta\)), beta prime (\(\beta'\)) and omega (\(\omega\), not shown over) subunits, and two copies of the alpha (\(\alpha\)) subunit, which binds to the UP element.

There are 6-7 different \(\sigma\) subunits, each of which binds tightly to the core RNA polymerase and contributes promoter recognition and DNA melting properties to the enzyme. The most common are \(\sigma^{70}\) and \(\sigma^{54}\), the names reflecting size in kD.
E. coli RNA polymerase
Regulated recruitment

RNA polymerase bearing $\sigma^{70}$ is constitutively active, that is, it is expressed independently of environmental signals; activators work by recruiting that enzyme to specific genes, see next slide.

These activators often work in conjunction with repressors that, in their simplest mode of action, exclude the binding of polymerase to specific genes. We will illustrate with the lac genes, as in the following slides.
A generic *E. coli* pre-initiation complex

RE = regulatory element, a binding site
The *lac* genes

The enzyme β-galactosidase, the produce of the *lacZ* gene, cleaves lactose, the first step in the metabolism of that sugar. The gene is transcribed *if and only if* lactose is present in the medium. That physiological signal is almost entirely overridden by the simultaneous presence of glucose, a more efficient energy course than lactose.

The effects of these signals are mediated by two DNA binding proteins, each of which sense one of the sugars. *Lac repressor* binds the operator only in the absence of lactose, and *CAP activator* binds DNA only in the absence of glucose.

The story is told in the following slides.
Here CAP stands for catabolite activator protein, and is also known as the CRP (cyclic AMP receptor) protein. Operator here means the Lac repressor binding site. Both sites are ~ 20 bp.
Detecting physiological signals

Lac repressor undergoes a conformational (allosteric: other shape) change upon binding inducer, here allolactose, which greatly decreases DNA binding of the repressor.

In contrast, the allosteric change undergone by CAP upon binding cAMP increases its ability to bind DNA.

Glucose exerts its effect by somehow decreasing synthesis of cAMP.
Three states of the lac genes
Sequence elements recognized by the promoter

Genes and Signals, © 2002 by Cold Spring Harbor Laboratory Press, Chapter 1, Figure 5
DNA binding by a dimeric protein bearing a helix-turn-helix domain, used by both CAP and Lac repressor

R labels the “recognition helix. A single one contacts approx 6-8 bp.
Summary of the lesson from *E. coli*  

Even in this simple example, we have seen protein-DNA binding at several specific sites:

- RNA polymerase at the -10 and -35 sites and the UP element
- CAP at the CAP site
- Repressor at the operator.

We now turn to eukaryotes, at first the single celled yeast, and then (very briefly) fruit flies.
Next steps

Things can definitely get more complex and interesting with *E. coli*, see e.g. *A Genetic Switch: Third Edition Phage Lambda Revisited* by Mark Ptashne. However, we now leave prokaryotes and have a brief look at eukaryotes. Here things get more complicated in two distinct ways. First, the transcriptional pre-initiation complex is much more complicated, even for single-celled eukaryotes. We look briefly at it. Second, the cis-regulatory modules range from relatively simple, with yeast, to the more complex with fruit flies, and the very complex with humans. We begin with generalities and cartoons illustrating these points, and then turn to yeast.
Transcriptional regulation in eukaryotes: generalities

In eukaryotic cells, DNA is wrapped around proteins called histones, forming bead-like structures called nucleosomes, collectively called chromatin. Also, the chromosomes are sequestered in the nucleus.

To transcribe a typical eukaryotic gene, a host of chromatin remodeling (e.g. the ATP-dependent SWI/SNF coactivators) and chromatin-modifying factors (histone acetyl transferases: HATs) must first establish an environment favorable to transcription. These chromatin targeted events allow upstream promoter and enhancer binding transcription factors to gain access to the DNA template. Much is relevant to our task, but we can only sketch the basics.
Chromatin
Gene activator protein

Histone acetylase

Chromatin remodeling complex

Specific pattern of histone acetylation

Remodeled nucleosomes

General transcription factors and RNA polymerase holoenzyme recruited

TRANSCRIPTION ACTIVATION
Eukaryotic pre-initiation complex

TATA = TATA element, TBP = TATA binding protein
RE = a generic response element
The Pol II enzyme contains roughly 12 subunits, including homologues of the $\beta$, $\beta'$, $\omega$ and $\alpha$ subunits. Promoter recognition and DNA melting are shared among the general transcription factors, which includes TFIIA, TFIIB, TFIID (not shown), TFIIE, TFIIF and TFIIF. TFIIA, TFIIB and TFIID recognize promoter proximal DNA elements including the TATA box, the BRE (TFIIB recognition element), the initiator (INR) element and the downstream promoter element (DPE). TFIIF binds to Pol II and helps recruit the enzyme to the promoter region. TFIIE binds to TFIIF and helps recruit it to the promoter region. TFIIF melts the DNA and more.
IWSI (initiation switch) and histone acetylase transferases (HATs) work to remodel and modify the chromatin structure, and can be present in numerous multi-subunit complexes. SWI/SNF complexes also function to remodel chromatin, whereas CRSP/MED is thought to mediate signals between enhancer bound factors and the transcriptional machinery.
A generic yeast transcription unit

TATA = the TATA element, a binding site for TBP.

UAS = upstream activating sequences, usually two or three closely linked binding sites for one or two different sequence specific transcription factors.
Generic complex metazoan transcriptional control modules

INR = initiator

DPE = downstream promoter element
Lessons from yeast

Some general features of activation in yeast:
- the transcriptional machinery is not bound to the promoter prior to activation
- yeast activators have separable activating and DNA-binding domains
- an activating region must be tethered to DNA near a promoter to activate transcription.

We turn now to the GAL genes, those required for the metabolism of galactose. Cells growing in the absence of galactose express these genes at very low levels, and the addition of galactose induces them more than 1000-fold. An additional regulator of the GAL genes is the repressor Mig1.
The four binding sites to which Gal4 binds comprise the galactose Upstream activating sequence (UAS\textsubscript{G}) and is about 118 bp. The Gal4 sites are each 17 bp. The binding site of Mig1 is 13 bp, see later.
Regulation of Gal1

The Gal1 gene encodes galactokinase, which catalyzes the conversion of galactose into glucose. In addition, the Gal1 gene is also involved in a negative feedback mechanism. In the presence of glucose, the regulatory protein Mig1 binds to the corresponding Mig1 site on the promoter, which results in the repression of transcription of the Gal1 gene and galactose is not converted into glucose.

Mig1 does not work alone: it recruits a “repressing complex,” a key component of which is Tup1. (Mig1 sites are also found in front of many other genes repressed by glucose.)

In the absence of galactose and glucose, Gal4 is bound to the UAS$_G$ (see next slide) but it does not activate transcription because it is inhibited by Gal80.
The three states of the Gal1 gene

In a, cells are grown in a “neutral” sugar raffinose. For simplicity, Only a single Gal4 dimer is shown bound to the $\text{UAS}_G$. The dotted lines Indicate that Mig1 also represses the gene encoding Gal4.
A dimer of Gal4 bound to a 17-bp site
From M. Bulyk, Genome Biology 2003
Detecting physiological signals

A protein called Gal3 senses galactose and conveys that signal to the gene; upon binding the sugar and ATP, Gal3 binds Gal80, freeing Gal4’s activating region. In the absence of glucose, Mig1 is held in the cytoplasm in phosphorylated form. Glucose inhibits the kinase responsible for that phosphorylation, and thereby allows Mig1 to enter and remain in the nucleus. There it binds DNA and recruits a repressing complex, as already seen.

More generally, in the absence of their respective inducing signal, transcriptional regulators (such as Mig1) tend not to be found in the nucleus with their activating regions free to work. Rather, their activating regions are masked (as for Gal4), or, as with Mig1, they are maintained outside the nucleus until the inducing signal is detected.
Where have we got to?

Understanding how Gal4 works would take us very far afield. We would have to consider how it interacts with other aspects of the transcription pre-initiation complex, including the mediator, TBP. TFIIBm TFIIF, TAFs, as well as nucleosome modifying complexes SWI/SNF and more. You can read more about this in Ptashne and Gann. We’ll make do with the following:

**Summary.** In yeast, activators work by recruiting polymerase, but they do so in the face of a vastly more complicated transcriptional machinery. However, new functions have been accommodated in this scheme. For example, where nucleosome modifiers are required, the same activating regions that recruits proteins directly associated with polymerase recruits these enzymes as well.
Signal integration and combinatorial control

These are two very important features of transcriptional regulation for higher eukaryotes, but which exist in bacterial and yeast. **Signal integration** ensures that certain genes are switched on (or off) only when the proper combination of two or more signals is present. For example, we have seen the combination of an activator and a repressor used to integrate two signals in bacteria and yeast.

**Combinatorial control** is present in bacteria. For example, CAP works on some 200 genes in addition to the lac genes. CAP works with the Gal repressor to control transcription of the gal genes in *E. coli*, the products of which metabolize galactose.
Here Mcm1 works both as an activator and a repressor. It bears an activating region that is exposed when it binds with $\alpha_1$, but covered when it is bound (at a different site) with $\alpha_2$. $\alpha_2$ when bound either with Mcm1 or $a_1$, recruits the Tup1 repressing complex.
Note on higher eukaryotes

Three strategies are exploited to a greater extent here than in yeast:

- **There is more signal integration**: for many genes, activation depends on the simultaneous presence of multiple activators, each of which might be controlled by a separate signal.

- **There is more combinatorial control**: many activators work in various combinations with other activator (and repressors) to regulate different sets of genes.

- **Genes are often controlled by alternative sets of regulators**: this allows the same gene to be expressed in response to different sets of signals.
Notes on higher eukaryotes

A typical gene in higher eukaryotes is associated with a greater number of regulatory protein binding sites than its yeast counterpart. These sites are often dispersed over thousands of base pairs, so the regulators must often “work at a distance”.

A set of contiguous regulatory binding sites is called an enhancer (see e.g. the stripe enhancers in the next 2 slides) and these are sometimes found to work many thousands of base pairs upstream or downstream from genes.
The yeast genome encodes about 300 transcription factors.

By contrast, the genome sequences of C. elegans and Drosophila have at least 1,000 transcription factors. There may be as many as 3,000 in the human genome.

Given the combinatorial nature of transcriptional regulation, these modest differences in numbers could produce a dramatic expansion in regulatory complexity. We close with a hint of this.
The muscle heart enhancer responds in a highly restricted region of the late embryo to the confluence of three signals sent by other cells. Moreover, only some cells have, by virtue of their history, regulatory proteins required for the activity of this enhancer. Thus specificity is achieved: only certain cells have the potential to respond to the signals; of these cells only those in the right place at the right time, receive all three signals.
The eve stripe 2 enhancer

A *Drosophila* embryo is shown at the top, with the seven stripes where *eve* is expressed. Stripe 2 is provided by the stripe 2 enhancer, consisting of two activators (Bcd, Hb) and two repressors (Kr, Gnt). Their distribution across the embryo is shown in the center of the figure.