Introduction to Bioinformatics

Accompaniment to *Discovering Genomics…*

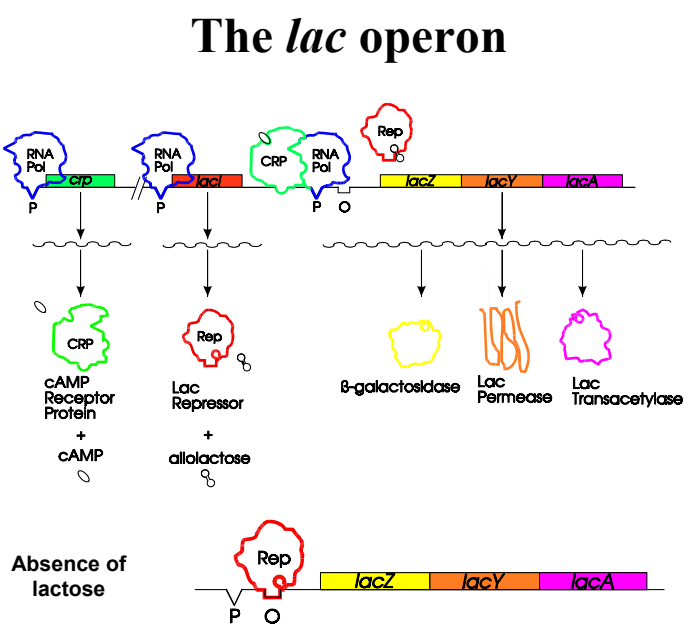
Reading in *Discovering Genomics, Proteomics, and Bioinformatics* by Campbell & Heyer
(henceforth referred to as DGPB by C&H)

pp.107-124, Discovery Questions 1-28, Math Minutes 4.1-4.3 (left over from Feb 14)
pp.124-128, Discovery Questions 29-44

Much of the reading may be pretty mysterious if you are not familiar with mechanisms of gene regulation. I've provided below descriptions of two models of gene regulation: (a) the *lac* operon, which has some regulatory elements typical of gene regulation by bacteria, and (b) an idealized picture of eukaryotic gene regulation.

A. The *lac* operon

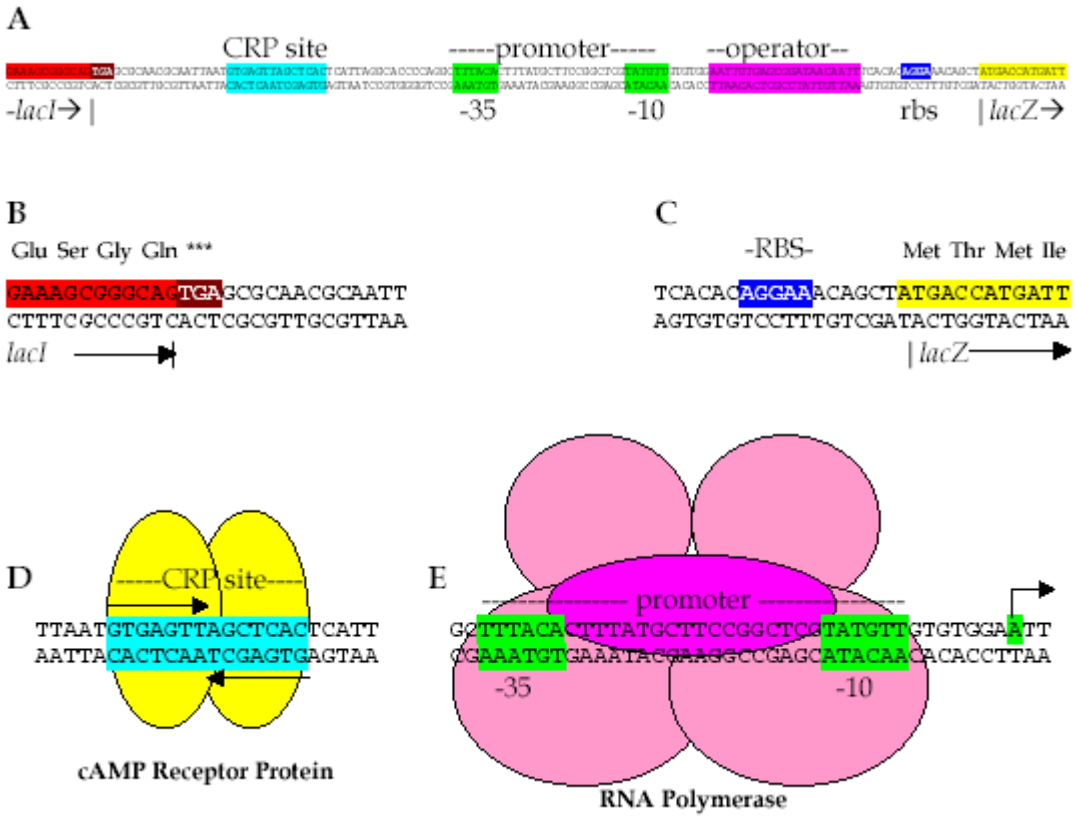
Bacteria typically transcribe genes several at a time, with one molecule of RNA containing information from multiple genes. The genes transcribed together are called an operon. The *lac* operon (arguably the best-studied operon in existence) consists of three genes, all related to the metabolism of the sugar lactose: *lacZ*, encoding an enzyme that breaks down the lactose, *lacY*, encoding a protein that transports lactose into the cell, and *lacA,* encoding a protein whose function is not clear. Ribosomes bind at different points on the transcribed message and translate the three proteins.



The three genes are expressed (produce protein) so long as RNA polymerase, the enzyme that synthesized RNA, finds its binding site on the DNA next to *lacZ* and begins synthesis. The binding site for RNA polymerase is called the **promoter**. As it happens, the *lac* promoter is not the optimal sequence for binding RNA polymerase, and the protein does not attach to the promoter stably, unless another protein, cAMP Receptor Protein (CRP), attaches to ***its*** nearby binding site. The combined presence of CRP and the weak promoter make stable binding of RNA polymerase much more likely. CRP binds to its binding site only if the bacterium's favorite sugar, glucose, is not present. If it ***is*** present, then there's no sense making the proteins encoded by the *lac* operon, just as there's no sense preparing the barbeque if you've decided to eat pizza.

All this is true ***if*** lactose is present in the surrounding medium (there's no sense deciding to eat pizza if there's no pizza to be had). If lactose is ***not*** present, then a protein called the *lac* repressor binds near the promoter blocking the action of RNA polymerase. Lactose prevents this by binding to the repressor and changing its shape so that it cannot attach to DNA. All of this is good: lactose present means that the repressor does not block RNA polymerase from transcribing

the *lac* operon; lactose absent means that RNA polymerase will not waste time making RNA for protein that won't be used. The players in this drama are shown in greater detail below:



Nucleotide sequence of the regulatory region of the Lac operon. Sites colored on both strands indicate DNA binding sites for protein. Sites colored on only one strand indicate features of interest on the transcribed RNA. Panel **A** shows the nucleotide sequence of the region between *lacI* and *lacZ*, containing some of sites important in the regulation of the Lac operon. Panel **B** and **C** show the end of *lacI* and the beginning of *lacZ*, respectively. The ribosomal binding site (RBS) preceding *lacZ* is highlighted. Panel **D** shows cAMP Receptor Protein (CRP) binding to its binding site. CRP is a dimeric protein, each subunit recognizing 5'-GTGAGTT-3' (shown by arrows). Panel **E** shows RNA polymerase binding to the Lac promoter at two sites: approximately 10 and 35 nucleotides upstream from the start of base at which transcription begins (shown by an arrow pointing in the direction of transcription). Panel **F** shows the Lac repressor binding to the operator. The repressor is a dimeric protein, each subunit recognizing 5'-AATTGT-3' (shown by arrows).

B. Eukaryotic genes

The *lac* operon may seem confusing at first, but once you get used to it, it displays a certain simple logic. Eukaryotic gene regulation remains complicated no matter how long you stare at it. The basic idea is the same: Control the binding of RNA polymerase and you control the expression of the gene.

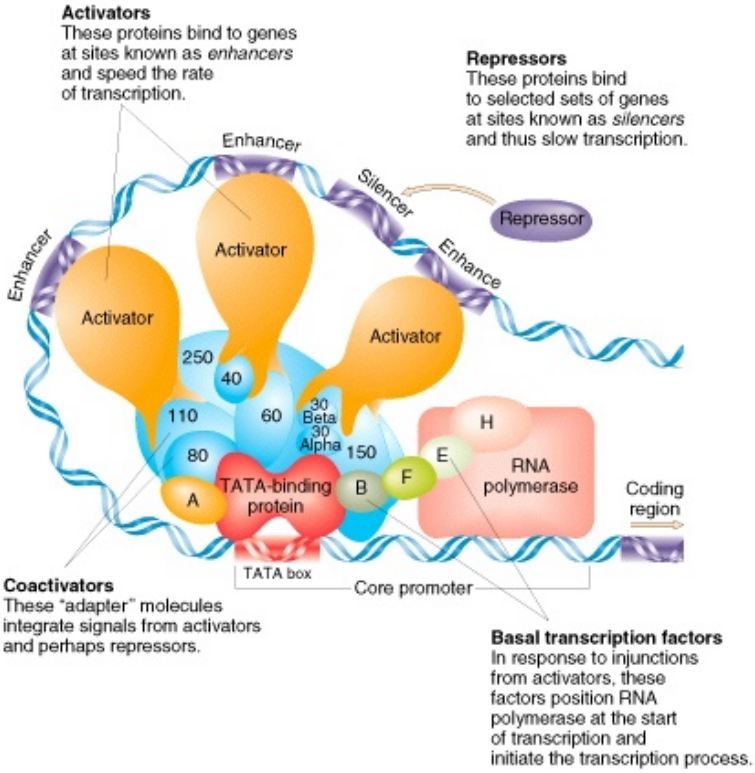


Figure from Griffiths et al (1996) *Introduction to Genetic Analysis,* 6th ed. WH Freeman and Co.

In eukaryotes, the idea seen in the *lac* operon of increasing weak binding of RNA polymerase to a promoter has been taken to the ultimate extreme. RNA polymerase does not bind at all to the promoter. Rather, it binds to a complex of proteins that bind to the promoter, called a TATA box, because the sequence of the typical promoter contains the sequence TATA. The binding of the protein complex to the promoter is modulated by an army of transcriptional activator proteins that collectively form a nest into which the protein complex rests. The activators bind to their binding sites (enhancers), which may be quite distant from the promoter, but binding may be affected by a variety of environmental conditions, e.g. the presence or absence of a certain hormone. Binding of activator proteins may also be prevented by the binding of repressor proteins.

C. Relation to microarrays

Transcriptional factors, like CRP, or one of the activator proteins in the cartoon above, may be used by the cell to affect the expression of many genes. For example, CRP regulates the expression not only of the *lac* operon but (reasonably enough) the expression of genes involved in the metabolism of other non-favorite sugars. Likewise, the same hormone-responsive activator protein may regulate a host of genes related to the response to that hormone. For example, estrogen, through its binding to an activator protein, affects the expression of a number of genes whose products are required for the production of chicken eggs.

Since transcriptional regulators may affect multiple genes, mutants lacking a transcriptional regulator (or overexpressing one) may have aberrant expression of many genes. Microarrays can tell us what genes have aberrant expression in such mutants, providing clues as to what gene products are important in a given cellular response.