Introduction to Bioinformatics
Regulatory sequences in DNA

I. Why regulation?

Here we are after only a few thousand years of recorded history, and we now know the secret of life -- DNA. We've figured out the complete genomic sequences of dozens of organisms, including humans, and can predict the amino acid sequences of almost every protein those genomes encode. In principle, though not yet in fact, we can also predict from the sequences of amino acids what functions the proteins will have and even change those functions to suit our wishes.

But don't feel smug: we still don't know how even the simplest living organism is formed.

Upon reflection, this should not surprise you. Suppose I could read every thought in your head, every thought you ever thought, even every thought you haven't thought yet. Everything you were capable of thinking. Would that tell me who you are? Not at all. If every possible thought went through your mind at once, there would be chaos, and you are not chaos.

What's missing is the regulation of your thoughts -- what relationships there are between what is around you and what is called to mind, how one thought connects to another. And that's what's missing from our understanding of genetics at this point: regulation.

At any given moment only a fraction of the genes an organism possesses are expressed as protein, and if they all turned on at the same time... certain death. You have genes that are turned on to protect you when you are overheated, when you are exposed to heavy metals, genes that are expressed only during early embryogenesis, and so forth. To understand how genes determine the form and function of an organism, we must understand not only what genes are but also what regulates their expression.

II. How regulation?

The flow of information from inactive DNA to active protein can be interrupted at any one of several points (**Fig. 1**). While there are many examples of control at each of the points shown, in most organisms regulation takes place primarily at the first step: the transcription from DNA to RNA. What this means is that if a gene is transcribed, the remaining steps leading to active protein proceed unhindered. Turn on the gene and you turn on the corresponding chemical reaction. So if we understood how transcription is controlled, we'd know a good deal about how a cell controls its capabilities.

SQ1: Why do you think that regulating initiation of transcription is so common as compared, say, to regulating the rate of protein degradation?

SQ2: And yet there are certainly many examples where enzyme activity is regulated in response to environmental influences. For example, there is the famous case of stress, represented within us by high adrenaline levels, causing an increase in the breakdown of glycogen by the activation of the enzyme responsible for the breakdown. Why do you imagine evolution might have preferred in this case regulation of enzyme activity rather than regulation of transcriptional initiation?

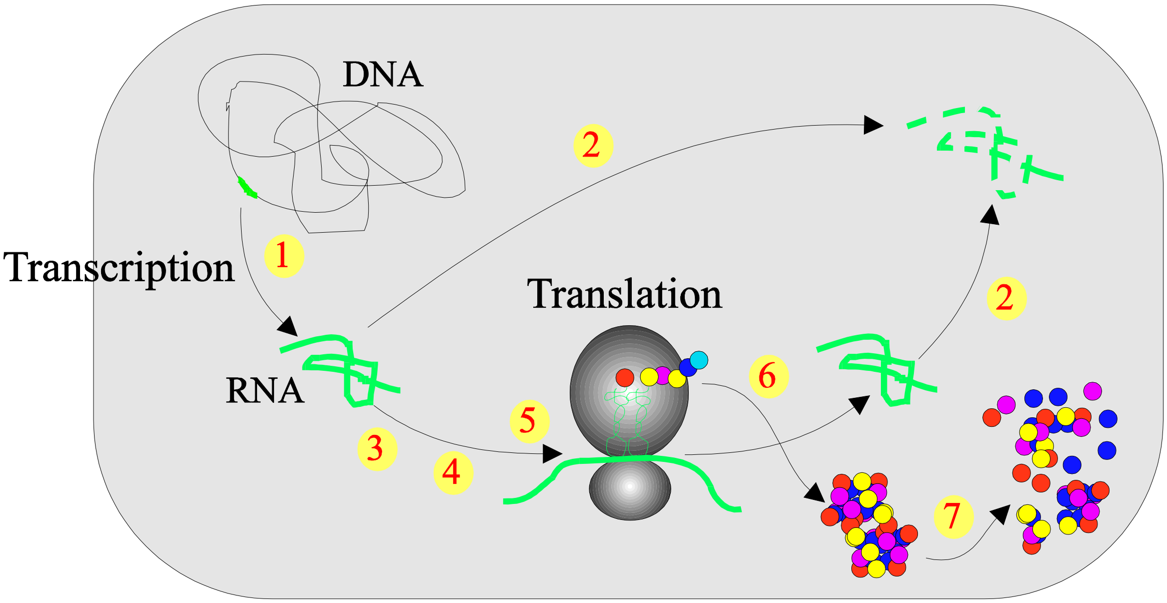


Fig. 1: Control points over gene expression. Choke points in the route from DNA through RNA to active protein (not all shown): **1.** Binding of RNA polymerase/Initiation of transcription, **2**. Degradation of RNA, **3.** Processing of RNA, **4.** Availability of RNA, **5.** Binding of RNA to ribosome/Initiation of translation, **6.** Modification of protein, **7.** Degradation of protein.

And a good deal ***is*** known about the regulation of some genes. A case in point is the regulation of the *E. coli* *lac* operon, genes that encode proteins important in the utilization of the sugar lactose. An operon is a group of contiguous genes transcribed together, presumably because their encoded proteins are needed under the same conditions. The three genes of the *lac* operon are *lacZ*, which encodes β-galactosidase, which breaks the disaccharide lactose down to the monosaccharides glucose and galactose; *lacY*, which encodes the Lac permease, a protein that transports lactose into the cell; and *lacA*, which encodes lactose acetyltransferase, an enzyme whose function in lactose metabolism is not clear.

The three genes are expressed (produce protein) so long as RNA polymerase, the enzyme that synthesizes RNA, finds its binding site on the DNA next to *lacZ*, the **promoter**, and begins synthesis. All the regulatory mechanisms centers around that basic question: Does RNA polymerase bind or doesn't it? If it does, then transcription of the operon occurs, and the transcript is translated into the three proteins.

Fig 2 illustrates the mechanisms governing whether RNA polymerase binds to the *lac* 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.

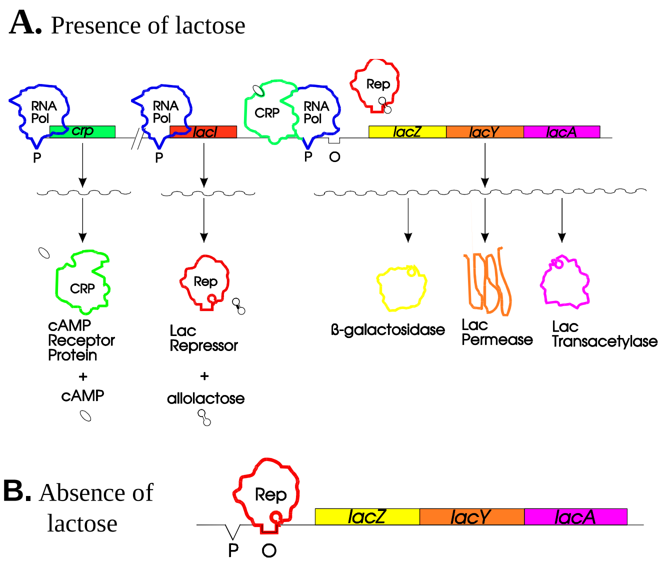


Fig. 2. Regulation of the *lac* operon.

The players in this drama are shown in greater detail in **Figure 3**. Note that the repressor and CRP protein both are dimeric (with two identical subunits) and both bind at palindromic sequences. This is typical of many DNA-binding proteins. Binding of proteins to palindromic DNA makes a good deal of evolutionary sense. It's important that proteins bind specifically and not to many random site in the genome. It is helpful if the binding site is relatively rare. Requiring that TWO copies of a protein binds doubles the specificity of the binding without requiring that evolution figure out how to make a protein that can recognize so long a DNA sequence.

SQ3: Suppose that CRP were a monomeric protein. How many sites would it find at random in the 4,639,675 nt-genome of *E. coli*? (As it happens, the genome has about equal frequencies of the four nucleotides)

SQ4: By what factor is the expected number of recognition sites in *E. coli* decreased if one presumes that binding of CRP requires a dimeric protein? Does doubling the number of nucleotides in the recognition site half the number of expected binding sites?

SQ5: Binding sites are often found by mutation.

a. What is the expected level of expression of the *lac* operon if the operator is mutated so that it no longer binds the Lac repressor?

b. What if the CRP binding site is mutated so that it no longer binds CRP protein?

c. What if BOTH the operator and the CRP binding sites are mutated?

d. Suppose that there were two genes in *E. coli* encoding the Lac repressor. One was wild-type (made good repressor) and one was mutant (made non-functional repressor). What would be the resulting phenotype? Which allele would you call "dominant"?

e. Suppose that there were two *lac* operons in *E. coli*. One had a wild-type operator (bound repressor) and the other had a mutant operator (did not bind repressor). What would be the resulting phenotype? Which allele would you call "dominant"?

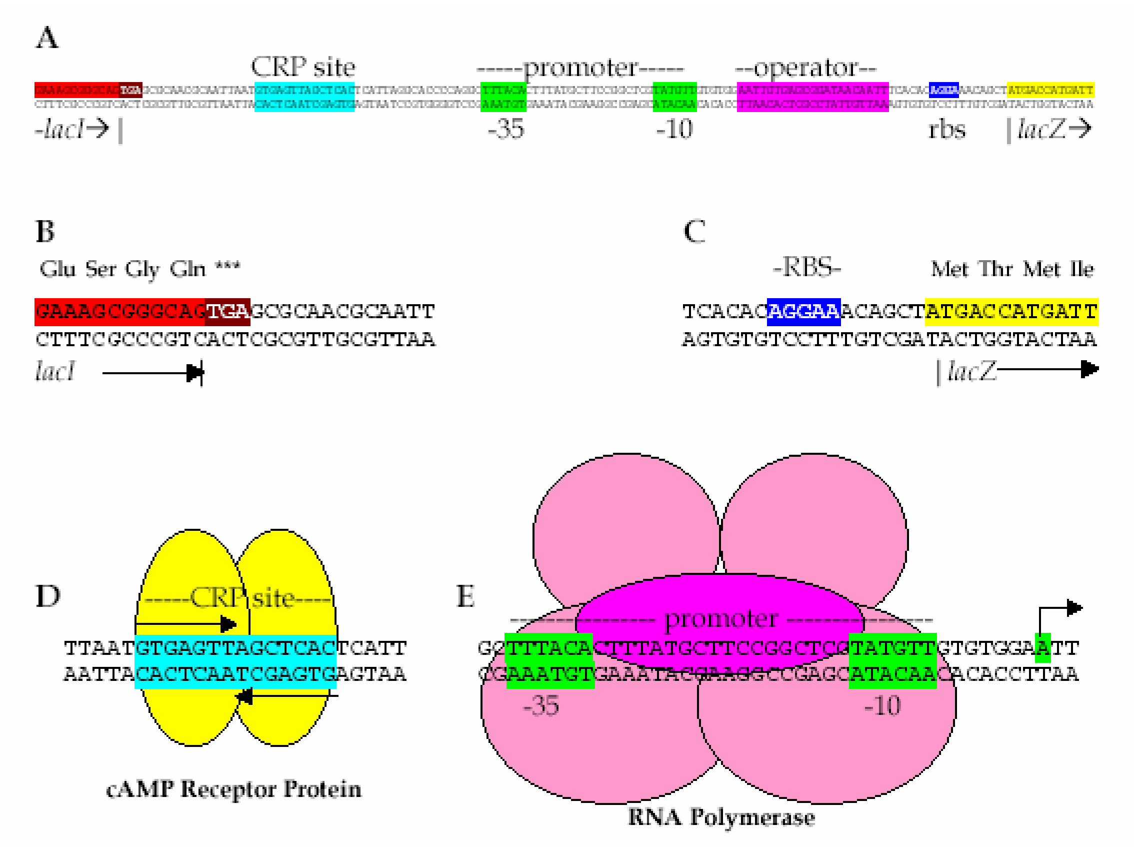
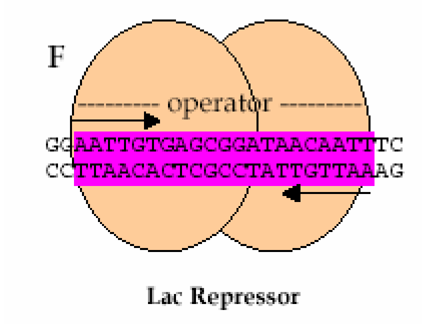


Fig. 3. Nucleotide sequence of the regulatory region of the Lac operon. Sites colored on both strands indicate DNA binding sites for protein. Sites colored on only one strand indicate features of interest on the transcribed RNA. Panel **A** shows the nucleotide sequence of the region between *lacI* and *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).


F
--------- operator ---------
GGAATTGTGAGCGGATAACAATTTC
CCTTAACACTCGCCTATTGTTAAAG
Lac Repressor

What about 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.

In eukaryotes, however, 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, illustrated in **Fig. 4**. 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.

Mammals differ very little from one another in protein-encoding genes. The genetic basis for the differences between humans and other primates lies primarily in the regulation of our genes, not in the proteins they encode.

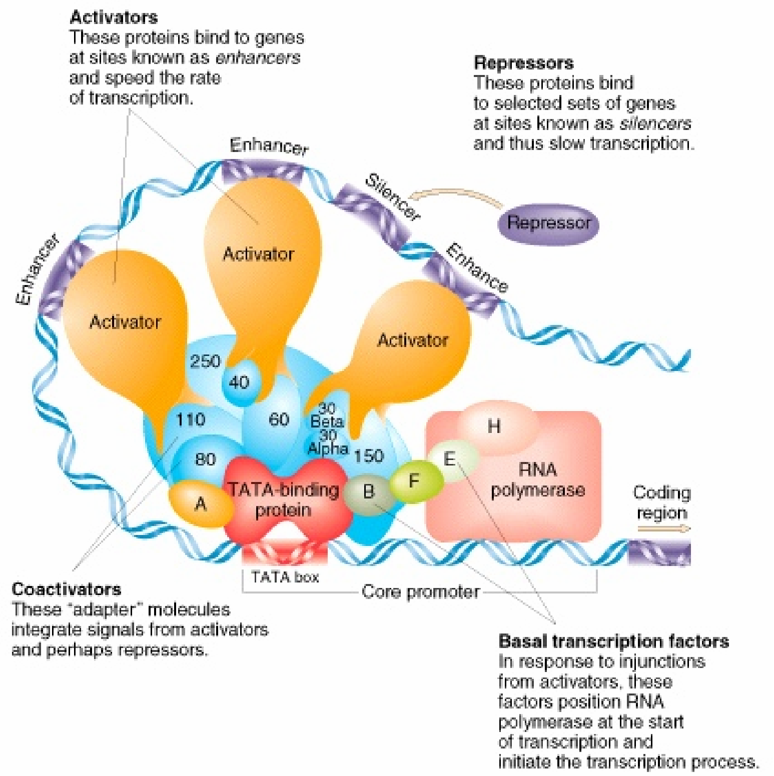

Activators
These proteins bind to genes at sites known as *enhancers* and speed the rate of transcription.
Repressors
These proteins bind to selected sets of genes at sites known as *silencers* and thus slow transcription.
Enhancer
Silencer
Repressor
Enhancer
Enhancer
Activator
Activator
Activator
250
40
110
60
30
Beta
30
Alpha
150
80
H
E
RNA polymerase
A
TATA-binding protein
B
F
Coding region
TATA box
Core promoter
Coactivators
These "adapter" molecules integrate signals from activators and perhaps repressors.
Basal transcription factors
In response to injunctions from activators, these factors position RNA polymerase at the start of transcription and initiate the transcription process.

Figure 4. Binding of RNA polymerase to the promoter of a typical eukaryotic gene. Figure from Griffiths et al (1996) *Introduction to Genetic Analysis*, 6th ed., WH Freeman and Co.

SQ6: If you were given a mysterious DNA sequence and wanted to discover its biological function, would it be enough to identify the protein it encodes?

SQ7: Where else would you look for other clues?