BIOL591: Introduction to Bioinformatics
Regulatory protein

Reading in text: Nothing in either book is really pertinent

Outline:

A. Why regulation?
B. How regulation?
C. Regulation of cyanobacterial genes by environmental nitrogen
D. Simulation: A tool to assess likelihood

A. 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 what regulates their expression.

The differentiation of specialized heterocysts by *Nostoc* in response to nitrogen-deprivation (see *Our Story* for this unit) is an excellent example of gene regulation in action. Somehow, the biochemical sensing of a metabolic imbalance triggers a carefully orchestrated sequence of gene expression at specific times and in specific cells leading to the appearance of an enzyme capable of nitrogen fixation and of a complex machinery to support it. Our task in this section is to understand how gene expression might be regulated.

SQ1: In what respect do you think muscle cells differ from liver cells? Their DNA? Their RNA? Their protein?

SQ2: What's so bad about having all your genes turned on at once? Why not have your heavy metal-protection genes turned on BEFORE you encounter trouble?

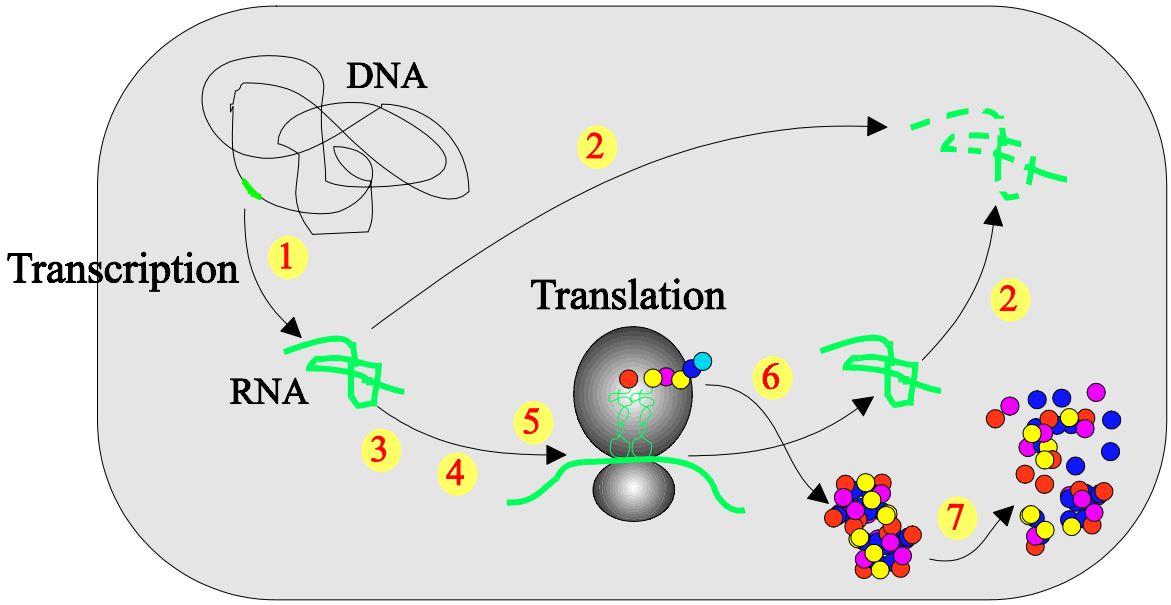

Fig. 1: Control points over gene expression. Choke points in the route from DNA through RNA to active protein (not all shown): **1.** Binding of RNA polymerase/Initiation of transcription, **2**. Degradation of RNA, **3.** Processing of RNA, **4.** Availability of RNA, **5.** Binding of RNA to ribosome/Initiation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

B. 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.

SQ3: Why do you think that regulating initiation of transcription is so common as compared, say, to regulating the rate of protein degradation?

Let's turn for the moment to a well-studied example of gene regulation: the transcriptional regulation of the gene *lacZ* in *E. coli*. This gene encodes the enzyme ß-galactosidase, which catalyzes the breakdown of milk sugar (lactose) into simple sugars that the bacterium can digest. Transcription of *lacZ* is tightly regulated, so that the gene is transcribed and ß-galactosidase is made only when *E. coli* needs to digest lactose, e.g. when its preferred sugar, glucose, is not available but lactose is.

Fig. 2A shows the region of the *E. coli* chromosome near the *lacZ* gene, and a closer look at the region (Fig. 2B-D) tells us how the regulation of *lacZ* transcription is achieved. Most of *E. coli*'s genome is comprised of genes encoding protein, but some of it lies between genes (e.g., between *lacI* and *lacZ*; Fig. 2B). These intergenic regions are necessary for the control of transcription. For a gene to be transcribed, it needs to possess a binding site for the enzyme RNA polymerase, which catalyzes the synthesis of RNA (transcription), and that binding site must lie

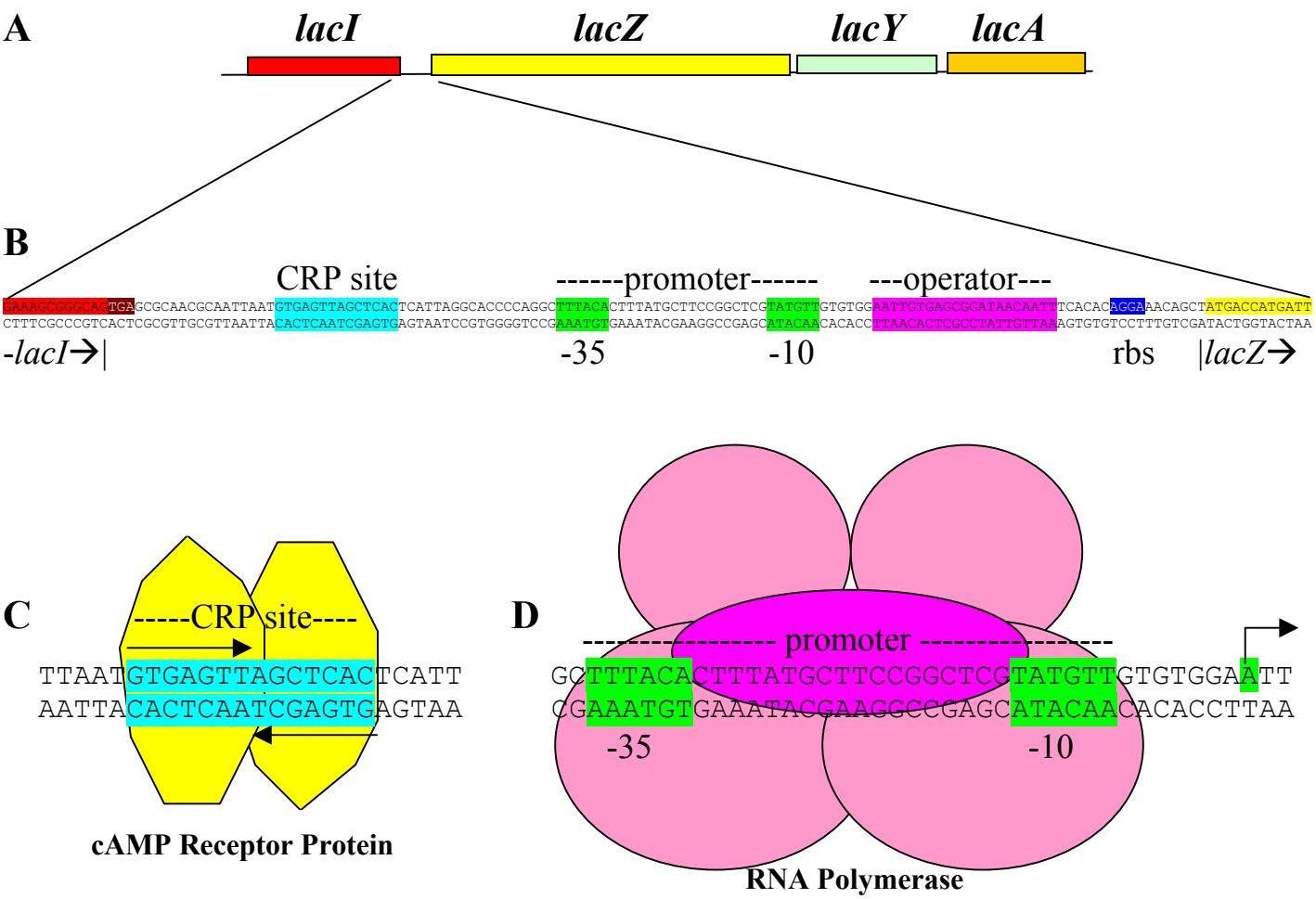

Fig. 2. Nucleotide sequence of the regulatory region of the Lac operon. Sites colored on both strands indicate DNA binding sites for protein. Sites colored on only one strand indicate features of interest on the transcribed RNA. **A.** The region of the *E. coli* genome surrounding the *lacZ* gene (total about 6000 nucleotides). **B.** The nucleotide sequence of the region upstream of *lacZ*, containing some of sites important in the regulation of the gene. The promoter is the binding site for RNA polymerase. The CRP site is the binding site for the transcriptional regulatory protein CRP. The operator and ribosome binding site (rbs) lie outside the scope of the present discussion. **C.** cAMP Receptor Protein (CRP) binding to its binding site. CRP is a dimeric protein, each subunit recognizing 5'-GTGAGTT-3' (shown by arrows). Note that the binding site is palindromic. **D.** 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).

before the gene so that the entire gene is transcribed. Binding sites for protein on DNA are no more than specific sequences of nucleotides. RNA polymerase binds to the *E. coli* genome at two specific sequences separated by about 25 nucleotides, as shown in Fig. 2D. The site at which RNA polymerase binds to DNA to initiate transcription is called the **promoter**.

This binding is not very stable in the case of *lacZ* DNA, however, and little transcription of the gene would take place if the *lacZ* promoter were the only means by which RNA polymerase found the proper place to initiate RNA synthesis. The weak binding of RNA polymerase to the *lacZ* promoter provides an opportunity for regulation. When *E. coli* is starving and could use an alternative source of energy, CRP, a protein sensitive to the state of the cell, binds nearby the

promoter. Now RNA polymerase can bind *both* to the promoter and to CRP, very stably, and *lacZ* is well transcribed.

Using **transcriptional regulatory proteins**, such as CRP, to modulate the binding of RNA polymerase is a clever way to run a cell. The cell can use the same RNA polymerase for transcription but modify its efficiency with different regulatory proteins sensitive to different environmental conditions.

SQ4: What would be the result of a mutation that altered or deleted several of the nucleotides shown in green in Fig. 2D?

SQ5: What fraction of genes do you think are preceded by promoters? What fraction are preceded by CRP-binding sites?

C. Regulation of cyanobacterial genes by environmental nitrogen

The cyanobacterium *Nostoc* needs multiple layers of regulation to govern the expression of genes related to nitrogen utilization. This is because in addition to the usual responses bacteria make in response to nitrogen deprivation, *Nostoc* has a last ditch plan in case all else fails: fixing atmospheric N_2. Since nitrogen fixation is expensive, and even more so the heterocysts required to support it, *Nostoc* must respond differently to mild nitrogen starvation (e.g. absence of ammonia, the preferred nitrogen source) and total nitrogen starvation. The regulation connecting nitrogen starvation to the expression of genes necessary for nitrogen fixation is not known.

As described in the Scenario, you suspect you may have discovered the connection. Fig. 3 shows the known binding sites of the regulatory protein NtcA upstream from cyanobacterial promoters known to be regulated by nitrogen deprivation. Others[1] have noted the strong tendency for certain nucleotides to appear at a relatively fixed distance upstream from the promoter, and you have found the same pattern upstream from *hetQ*. Since the product of *hetQ* is involved indirectly in the control over heterocyst differentiation, you reason that the NtcA binding site may be the point at which nitrogen regulation is exerted. . . if it *IS* an NtcA binding site (see Programming notes).

SQ6: Palindromes are sequences that read the same backwards and forwards (e.g. Napoleon's lament, "*Able was I ere I saw Elba*"). When referring to DNA, the term takes on a special meaning: the nucleotide sequence of one strand is the same as that of its complementary strand read backwards.[2] The CRP-binding site shown in Fig. 2C is palindromic as are many protein binding sites. Examine the figure carefully. Given the structure of CRP, why is it that its binding site is palindromic?

SQ7: Is the binding site of NtcA palindromic? How does NtcA bind to its binding site?

SQ8: What do you think would be the effect if the region surrounding the NtcA binding site of *hetQ* were somehow attached to *lacZ*?

---

[1] Herrero et al (2001). Nitrogen control in cyanobacteria. J Bacteriol 183:411-425.

[2] A complementary sequence is one in which each nucleotide is replaced by the nucleotide it pairs with (A with T, G with C, and vice versa). Thus, ATGAC has a complement TACTG and a reverse complement of GTCAT. The sequence GATC is the reverse complement of itself… it's a palindrome!

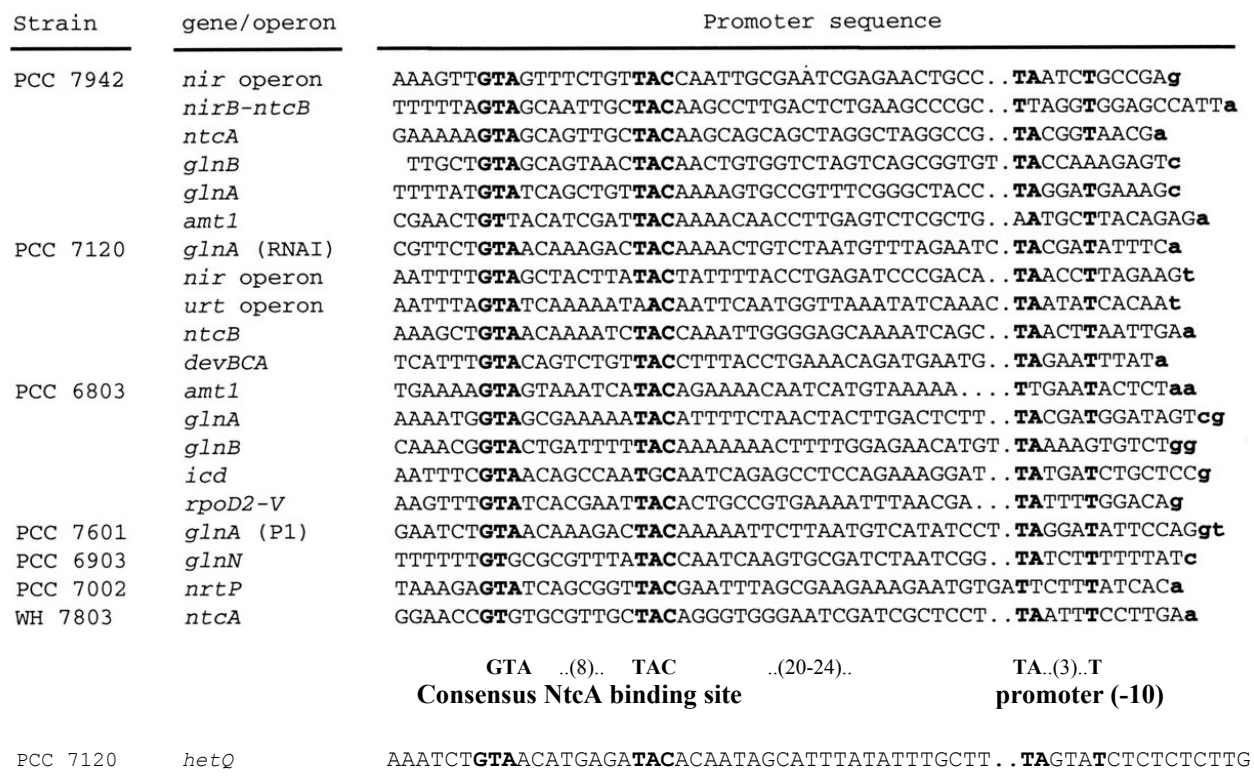

Fig. 3: Alignment of known NtcA binding sites upstream from cyanobacterial genes regulated by nitrogen deprivation. The accepted consensus binding sequence is given below, along with the sequence you noticed upstream from *hetQ*.

D. Simulation: A tool to assess likelihood

The scenario poses the problem of how to assess the likelihood that a putative NtcA binding site that you happened to stumble over is real. You've decided that to declare the likelihood high if the criterion you used to identify the site (given at the bottom of Fig. 3) is sufficiently stringent that you are unlikely to have encountered such a sequence by chance. Although I worded this last sentence very carefully, it is still uncomfortably vague. What do you mean "by chance"? We'll leave this question unanswered for the moment.

Simulations may provide insight into such questions. The program *DiceRoll* (available at the web site) illustrates how they work. *DiceRoll* attempts to answer the question of how likely it is to find at least one match in a roll of five dice (a question of interest to *Yahtzee* players, for example). The roll 6-2-4-1-2 would be counted as a successful event, while the roll 2-4-1-5-3 would be counted as an unsuccessful event.

This program, like all simulations, can be divided into two separate parts: the model of an event, and the analysis of an event. Computers can't really roll dice, never mind what kind of graphics you might see on the screen. This program models the rolling of five dice as the generation of five random integers from 1 to 6. Seen this way, some of the assumptions of the model are clear:

1. Each die produces only 1, 2, 3, 4, 5, or 6 (I'm comfortable with this one)
2. The probabilities of producing any given number is the same
   (I think that's close to the truth, but do physical dice actually behave this way?)
3. Each die acts independently of the others, and each roll is independent of the last
   (This is certainly a simplification of reality and one that might be significant. I can imagine a human picking up the dice, shaking them about a bit to make a satisfying noise, and then dropping them on the table, all without really moving them around enough to lose the biasing influence of the previous roll.)

The analysis of the virtual dice is more straightforward: score the five random integers, in much the same way you'd tabulate the faces of physical dice.

How can you simulate the process leading up to your discovery of the putative NtcA-binding site? One idea is to model the process by generating millions of random sequences 44 nucleotides in length (the maximum size of the consensus NtcA sequence) and to ask how many of these sequences contain a sequence of the pattern shown in Fig. 3. There are a number of objections one could raise to the model. First, you didn't look at just 44 nucleotides. You looked at the entire region upstream from *hetQ*. Presumably, if you had found a putative NtcA binding site *anywhere* in that sequence you would have been equally amazed. So a better model would be to generate random sequences of several hundred nucleotides, the same size as the region you examined.

Second, how do you simulate real DNA sequences? Should you assume that all four bases (G, A, T, and C) occur with equal frequencies, the analogous assumption made by *DiceRoll*? This assumption clearly flouts reality: the composition of *Nostoc* DNA is about 20% G, 20% C, 30% A, and 30% T. Indeed, few organisms have genomes where equal frequencies of bases is a good approximation. We can readily adjust the model to use the observed *un*equal frequencies. Is this sufficient?

Is a random sequence of bases with a composition matching that of natural DNA a good model of the real thing? It goes without saying that real DNA is not random (if it were, we would not be having this conversation), but maybe it is random *enough* for our purposes. Unfortunately not. The sequence **GTA** (the first three bases of the consensus NtcA binding sequence) when found positioned appropriately in a gene encodes the commonly found amino acid glycine. One might thus expect it to occur more frequently than, say, the sequence **TAG**, which does not encode an amino acid. A random sequence would not take this bias of natural DNA into account and would thus underestimate the number of NtcA binding sites.

That's no problem. Just as we can make a random sequence that incorporates natural base frequencies. So can we incorporate the natural frequencies of three-base sequences. Is *that* enough? Maybe not. Amino acids aren't strung together at random. One might well imagine that long sequences of bases that encode certain amino acid sequences are more commonly found in a genome than one would expect by chance. If such a DNA sequence contains part of the NtcA consensus sequence, then again our random model fails. To make matters worse, you didn't examine just any DNA sequence: you looked specifically at a region of DNA that lies upstream from genes. Could such regions have a different base composition and other properties different from those of the genome as a whole? You bet.

I hope you're getting the idea that it is surprisingly difficult to model natural DNA. When using any simulation, and particularly one that attempts to capture the essence of natural DNA, you need to be on high alert to the underlying assumptions. Simulations can be extremely powerful, often providing the only practical way of gaining insight into complex systems, however, one must never lose sight of the difficulties in extracting from nature the essential features necessary to fit a model into a computer. We'll return to this issue in a few weeks when we discuss Markov models.

Fortunately, we have a way of avoiding all the concerns surrounding simulations: count the actual number of NtcA-consensus sequences in the actual genomic sequence. Of course, this is possible only when the actual genomic sequence is known, and you now see one of a multitude of reasons why people are scrambling over each other to get the genome of their favorite organism sequenced. You'll be working with a second program, *SequenceSearch*, that counts the number of NtcA-consensus sequences in the genome of *Nostoc.* It is important to keep firmly in mind that the analysis component of the simulation program (which we never wrote) and the counting portion of *SequenceSearch* does *not* determine actual NtcA-binding sites (determined by NtcA) but rather NtcA-consensus sites (invented by humans), which we *think* has something to with the site recognized by the NtcA protein. In reality, the protein might have somewhat different ideas.

SQ10. What are some issues to consider in modeling DNA sequences? Apart from those discussed, try to think of some others.

SQ11. If simulations of DNA sequences are so problematic, why would anyone ever do them?