Introduction to Bioinformatics
How to find DNA sequences responsible for gene regulation

Outline:

I. Prelude: Practical pattern recognition
II. Why regulation?
III. How regulation?
IV. Searching for motifs
  A. A problem and a simple-minded solution
  B. A more nuanced solution: Position-specific scoring matrices

I. Prelude: Practical pattern recognition

Sitting next to me is our next guest, Giaccomo Fettucini,… is it fair to describe you as the world's foremost connoisseur of Italian pasta?

Well, I can only say that I enjoy my work.

It says here that you're able with a single taste to determine whether a plate of pasta was made by a true Italian chef. Is that right?

It's not as difficult as you make it sound. Anyone could do the same with an appreciation of the elements that make up true Italian pasta.

Hey, I'm anyone. Let's see if you're right. We didn't tell you this, but we arranged for three plates of pasta… Ed, could you bring them in? Up to the challenge, Giaccomo?

I never refuse a plate of good pasta.

Good, let's go. Here's the first… what do you think?

Ah! Delicious! Obviously the work of a master.

Let's see,… you're right! That plate came from *La Belle Noodle*, flown in from Firenze for this show. But how did you know?

Very simple. It has all the markings of a genuine Italian pasta: the red sauce, the hint of garlic, the meatballs that melt in your mouth.

I could do that, if that's all there is to it. Let's try the second plate.

Hmmm. I would place this somewhere in the south of Italy, though there's a hint of oriental influence.

I think we got you this time. That plate came from around the corner at Ming's Yum Yum Café… oh wait a second, I see here that the chef actually is from Naples. That's amazing! But this pasta uses a white sauce, so how could you tell,…

True, the sauce was white, not red, but all the other characteristics were there, so the source was quite obvious.

I get it. A single deviation from your list of requirements is still OK. Well, we have one final plate for you.

Very well… Che Diablo! Take it away!

I have to confess, that plate I made myself. But how did you know? I used a red sauce, added a hint of garlic, and the meatballs…

Yes but you murdered the linguini.

Maybe so, but that's still just one deviation.

I don't mind a different color sauce or some creativity with the spices, but no Italian chef could ever make pasta as limp as this!

Well folks, I hope you caught all that: pasta's Italian if it matches a consensus of characteristics, but one deviation is OK, unless it's in a characteristic that doesn't deviate. I guess that's why we need world famous connoisseurs.

II. 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.

III. 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?

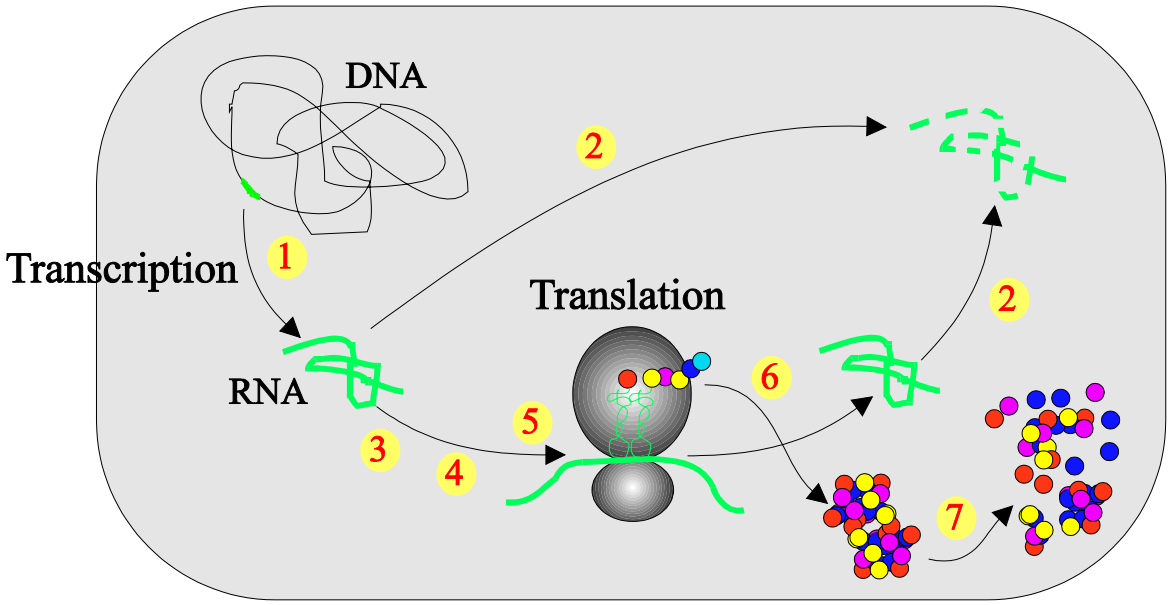

Fig. 1: Control points over gene expression. Choke points in the route from DNA through RNA to active protein (not all shown): **1.** Binding of RNA polymerase/Initiation of transcription, **2**. Degradation of RNA, **3.** Processing of RNA, **4.** Availability of RNA, **5.** Binding of RNA to ribosome/Initiation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

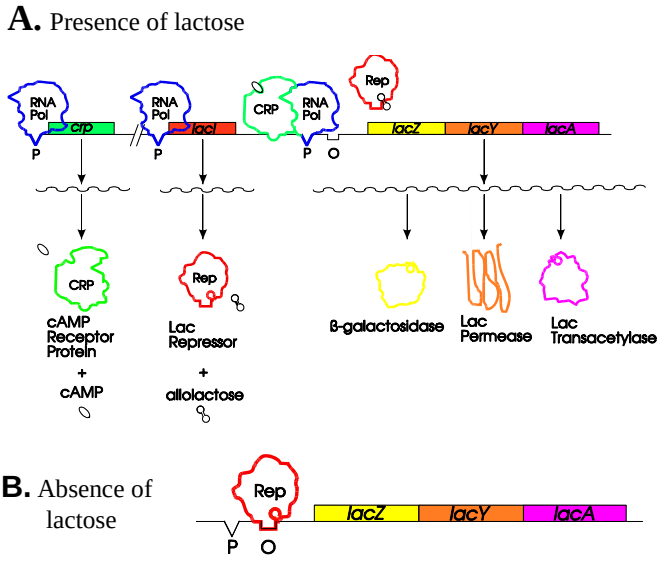

Fig. 2. Regulation of the *lac* operon.

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.

SQ2: 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.

SQ3: 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?

SQ4: 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?**

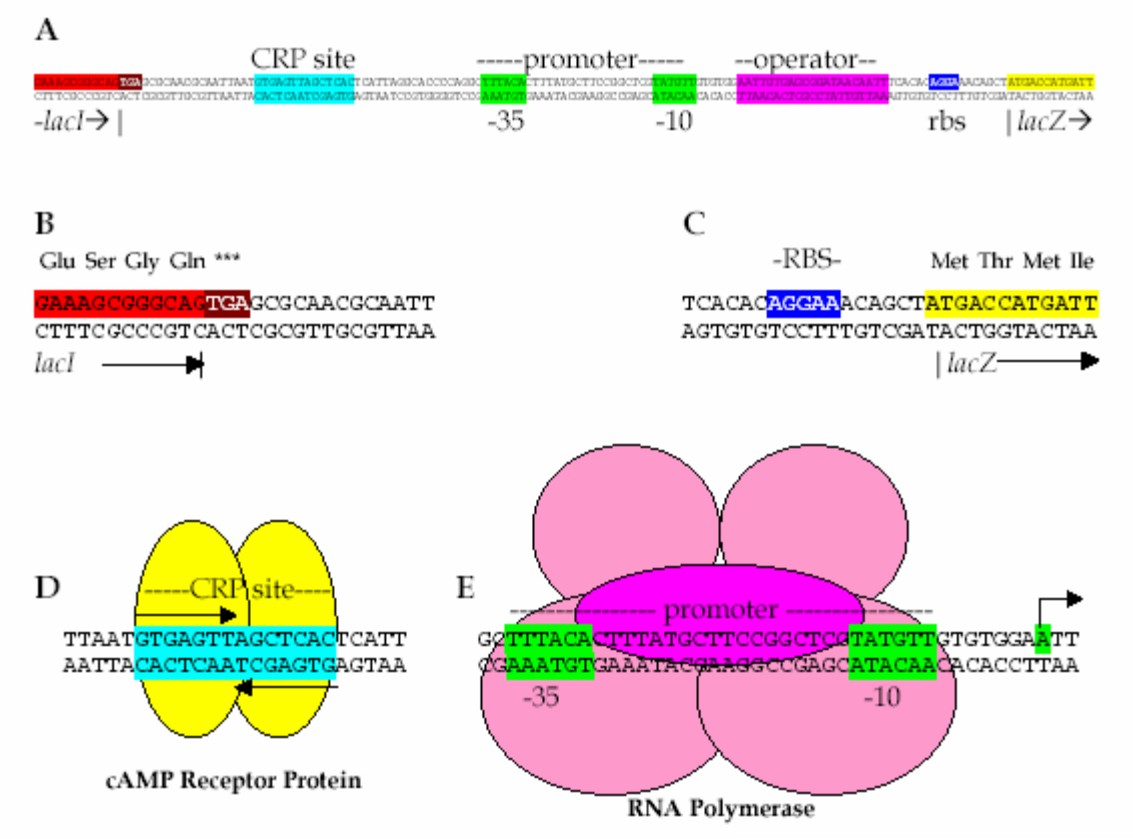

Fig. 3. Nucleotide sequence of the regulatory region of the Lac operon. Sites colored on both strands indicate DNA binding sites for protein. Sites colored on only one strand indicate features of interest on the transcribed RNA. Panel **A** shows the nucleotide sequence of the region between *lacI* and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

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.

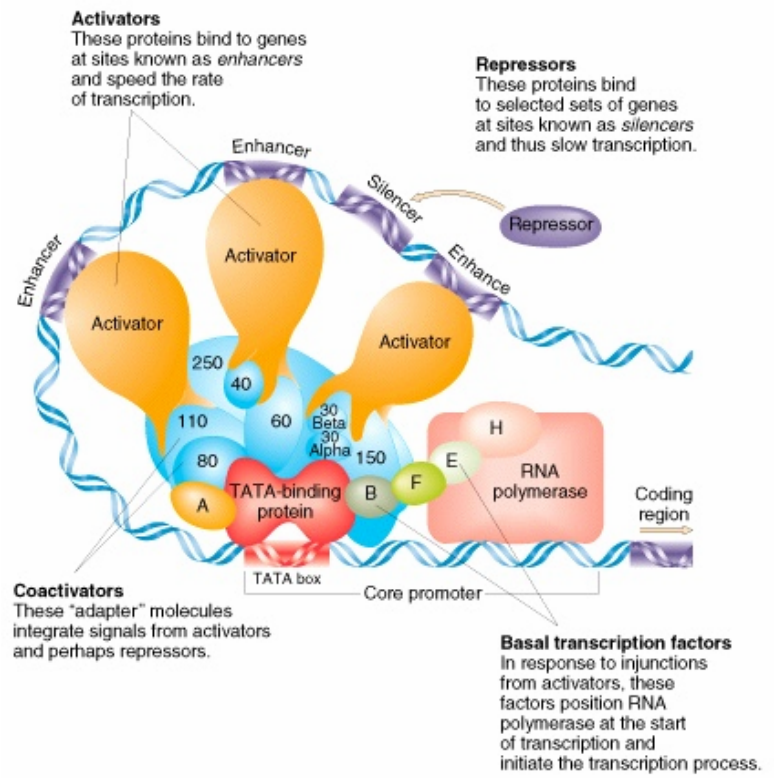

Figure 4. Binding of RNA polymerase to the promoter of a typical eukaryotic gene. Figure from Griffiths et al (1996) *Introduction to Genetic Analysis*, 6th ed., WH Freeman and Co.

.

IV. Searching for motifs

IV.A. A problem and a simple-minded solution

All organisms make metabolic adjustments depending on the environment they find themselves in. For example, cyanobacteria express on set of genes when they are growing on ammonia as a nitrogen source and another when they're deprived of ammonia and forced to use an alternative source. We do not completely understand how cyanobacteria sense nitrogen-deprivation, but one important element is known: the protein NtcA responds directly to nitrogen-deprivation, changing its conformation so that it becomes able to bind to specific sequences upstream from nitrogen-regulated genes. Many sites recognized by NtcA protein have been determined by cutting DNA to which NtcA has been bound and determining what sequence NtcA protects. Some sites are shown in **Fig. 5**.

Don't be fooled into thinking the problem of how nitrogen regulates gene expression has been solved! There are far more genes regulated by nitrogen than shown in the table. How can we find out what they are? One way is to repeat the NtcA-binding experiments with all sequences upstream from genes. With as many as 8000 genes in a cyanobacterium, this is far from practical! An alternative approach made possibly by the availability of genomic sequences is to look computationally for sites that may bind NtcA.

Strain	gene/operon	Promoter sequence
PCC 7942	*nir* operon	AAAGTT**GTA**GTTTCTGT**TAC**CAATTGCGAATCGAGAACTGCC..**TA**ATC**T**GCCGA**g**
	nirB-ntcB	TTTTTA**GTA**GCAATTGC**TAC**AAGCCTTGACTCTGAAGCCCGC..**T**TAGG**T**GGAGCCATT
	ntcA	GAAAAA**GTA**GCAGTTGC**TAC**AAGCAGCAGCTAGGCTAGGCCG..**TA**CGG**T**AACG**a**
	glnB	TTGCT**GTA**GCAGTAAC**TAC**AACTGTGGTCTAGTCAGCGGTGT.**TA**CCAAAGAGT**c**
	glnA	TTTTAT**GTA**TCAGCTGT**TAC**AAAAGTGCCGTTTCGGGCTACC..**TA**GGA**T**GAAAG**c**
	amt1	CGAACT**GT**TACATCGAT**TAC**AAAACAACCTTGAGTCTCGCTG..A**A**TGC**T**TACAGAG**a**
PCC 7120	*glnA* (RNAI)	CGTTCT**GTA**ACAAAGAC**TAC**AAAACTGTCTAATGTTTAGAATC.**TA**CGA**T**ATTTC**a**
	nir operon	AATTTT**GTA**GCTACTTA**TAC**TATTTTACCTGAGATCCCGACA..**TA**ACC**T**TAGAAG**t**
	urt operon	AATTTA**GTA**TCAAAAATA**AC**AATTCAATGGTTAAATATCAAAC.**TA**ATA**T**CACAA**t**
	ntcB	AAAGCT**GTA**ACAAAATC**TAC**CAAATTGGGGAGCAAAATCAGC..**TA**ACT**T**AATTGA**a**
	devBCA	TCATTT**GTA**CAGTCTGT**TAC**CTTTACCTGAAACAGATGAATG..**TA**GAA**T**TTAT**a**
PCC 6803	*amt1*	TGAAAA**GTA**GTAAATCA**TAC**AGAAAACAATCATGTAAAAA....**T**TGAA**T**ACTCT**aa**
	glnA	AAAATG**GTA**GCGAAAAA**TAC**ATTTTCTAACTACTTGACTCTT..**TA**CGA**T**GGATAGT**cg**
	glnB	CAAACG**GTA**CTGATTTT**TAC**AAAAAAACTTTTGGAGAACATGT.**TA**AAAGTGTCT**gg**
	icd	AATTTC**GTA**ACAGCCAA**T**G**C**AATCAGAGCCTCCAGAAAGGAT..**TA**TGA**T**CTGCTCC**g**
	rpoD2-V	AAGTTT**GTA**TCACGAAT**TAC**ACTGCCGTGAAAATTTAACGA...**TA**TTT**T**GGACA**g**
PCC 7601	*glnA* (P1)	GAATCT**GTA**ACAAAGAC**TAC**AAAAATTCTTAATGTCATATCCT.**TA**GGA**T**ATTCCAG**gt**
PCC 6903	*glnN*	TTTTTT**GT**GCGCGTTTA**TAC**CAATCAAGTGCGATCTAATCGG..**TA**TCT**T**TTTTAT**c**
PCC 7002	*nrtP*	TAAAGA**GTA**TCAGCGGT**TAC**GAATTTAGCGAAGAAAGAATGTGA**T**TCTT**T**ATCAC**a**
WH 7803	*ntcA*	GGAACC**GT**GTGCGTTGC**TAC**AGGGTGGGAATCGATCGCTCCT..**TA**ATT**T**CCTTGA**a**
		GTA ..(8).. **TAC** ..(20-24).. **TA**..(3)..**T**
		Consensus NtcA binding site **promoter (-10)**

Fig. 5: Alignment of known NtcA binding sites upstream from cyanobacterial genes regulated by nitrogen deprivation. The accepted consensus binding sequence is given below. The organisms are: *Synechococcus* PCC 7942, *Anabaena* PCC 7120, *Synechocystis* PCC 6803, *Tolypothrix* PCC 7601, *Pseudanabaena* PCC 6903, *Synechococcus* PCC 7002, and *Synechococcus* WH 7803. Taken from Herrero et al [J Bacteriol (2001) 183:411-425.

How to predict protein binding sites? A simple minded approach would be to take the consensus sequence (at the bottom of Fig. 5) and search for that sequence throughout the genome of a cyanobacterium. Let's try it:

SQ5. Find all sequences in the chromosome of *Anabaena* PCC 7120 that match the consensus sequence. To do this use the following BioBIKE forms:

```
(SEQUENCES-LIKE-PATTERN "GTA.{8}TAC.{20,24}TA...T"
    IN (SEQUENCE-OF A7120.chromosome))
```

which captures all instances in the chromosome of the given pattern. The pattern is read this way:

1. **Starts with** `GTA`
2. **Followed by exactly 8 characters of any type (`.` matches anything)**
3. **Followed by** `TAC`
4. **Followed by anywhere from 20 to 24 characters of any type**
5. **Followed by** `TA`**, then three characters, then** `T`

How many such sequences are there? Each element of the list is one instance (giving you the start and stop coordinates, the sequence, and the direction). Get a count of the list, using either `COUNT-OF` **or** `LIST`**. How many are there?**

SQ6. From the consensus sequence, do you think that NtcA binds to DNA as a monomer or a dimer?

Study Question 5 illustrates one shortcoming with the approach: There are too many sequences that match the consensus pattern. There aren't that many genes regulated by NtcA! A second shortcoming can be appreciated by reinspecting Fig. 5:

SQ7. How many of the proven NtcA-binding sites shown in Fig. 3 match the consensus pattern?

So we have a problem: Finding sequences by matching the consensus pattern gives too many false positives and too many false negatives. We need another approach.

IV.B. A more nuanced solution: Position-specific scoring matrices

The dialog in **Section I** hints at the problem. An expert would not apply a strict consensus sequence nor apply a strict rule (e.g. one mismatch allowed) but instead would consider a sequence in light of his accumulated experience. He would look at many characteristics, perhaps some subconsciously, and allow candidates the same kinds of imperfections as he has observed with real sequences, but only those kinds.

The ultimate expert is NtcA itself. Short of an in depth interview with a cooperative protein, the best we can do is to try to extrapolate from our own experience. Here's an analogous situation. Suppose you want to find all ways that people spell the word "color". You might look for all words that differed from only one letter, e.g. "coler", "color", "kolor". Unfortunately, this procedure would also give you "polor" and "colox", which are not likely spelling errors. If you wanted to limit your set to those instances where people *mean* color, then you could collect a training set of words where by context you're convinced the intent was "color" and see what kinds of mistakes were made. You'd probably find that the vowels showed some variability but the consonants were seldom missed. Learning from this, you might accept a word even with two errors (e.g. culer) but not one that replaced "l" with some other consonant.

SQ8. Looking at Fig. 5, where in the region where NtcA binds are deviations from the consensus sequence not tolerated? Where are they tolerated a little? Where are they tolerated a lot?

IV.B.1. A straightforward PSSM and how to score a sequence with it

A part of this expert process can be captured by what are called ***p***osition-***s***pecific ***s***coring ***m***atrices (PSSMs). Given an aligned set of sequences, it is very easy to construct a PSSM. Let's consider again the sequences surrounding the proven NtcA binding sites, confining ourselves for the moment just to the binding sites found in *Anabeana* PCC 7120 (**Table 1A**). The consensus sequence used only the six most highly conserved nucleotides within the NtcA-binding site: GTA...(N_8)...TAC. Ignoring the other positions tosses out a good deal of potentially useful information, as can be seen from the table of occurrences (**Table 1B**) and the PSSM derived from it (Table 1C). The latter is taken directly from the former by dividing the number of occurrences by the total number of sequences.

SQ9. Explain the source of the four fractions found in the ninth column of Table 1C (the first fraction is .500).

The PSSM gives us a tool to score how close any sequence is to the collected sequences used to create the scoring matrix (also called the training sequences). You would expect that a sequence functionally related to the training sequences would tend to have higher scores at each position,

Table 1: Examples of position-specific scoring matrices from sequence alignment

A. Sequence alignment[a]

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
glnA-71	G	T	T	C	T	G	T	A	A	C	A	A	A	G	A	C	T	A	C	A	A	A	A	C
nirA-71	A	T	T	T	T	G	T	A	G	C	T	A	C	T	T	A	T	A	C	T	A	T	T	T
ntcB-71	A	A	G	C	T	G	T	A	A	C	A	A	A	A	T	C	T	A	C	C	A	A	A	T
hetC-71	A	A	T	C	T	G	T	A	A	C	A	T	G	A	G	A	T	A	C	A	C	A	A	T
devBCA-71	C	A	T	T	T	G	T	A	C	A	G	T	C	T	G	T	T	A	C	C	T	T	T	A

B. Table of occurrences[a]

A	4	3	0	0	1	0	0	6	3	1	4	4	3	3	2	2	1	6	0	3	4	3	3	1
C	1	0	0	3	0	0	0	0	1	5	0	0	2	0	0	2	0	0	6	2	1	0	0	2
G	1	0	1	0	0	6	0	0	1	0	1	0	1	1	2	0	0	0	0	0	0	0	0	0
T	0	3	5	3	5	0	6	0	1	0	1	2	0	2	2	2	5	0	0	1	1	3	3	3

C. Position-specific scoring matrix (no pseudocounts; B = 0)[b]

A	.667	.500	.000	.000	.167	.000	.000	1.00	.500	.167	.667	.667	.500	.500	.333	.333	.167	1.00	.000	.500	.667	.500	.500	.167
C	.167	.000	.000	.5	.000	.000	.000	.000	.167	.833	.000	.000	.333	.000	.000	.333	.000	.000	1.00	.333	.167	.000	.000	.333
G	.167	.000	.167	.000	.000	1.00	.000	.000	.167	.000	.167	.000	.167	.167	.333	.000	.000	.000	.000	.000	.000	.000	.000	.000
T	.000	.500	.833	.500	.833	.000	1.00	.000	.167	.000	.167	.333	.000	.333	.333	.333	.833	.000	.000	.167	.167	.500	.500	.500

D. Position-specific scoring matrix (with pseudocounts; B = 1)

A	.617	.474	.046	.046	.189	.046	.046	.903	.474	.189	.617	.617	.474	.474	.331	.331	.189	.903	.046	.474	.617	.474	.474	.189
C	.169	.026	.026	.454	.026	.026	.026	.026	.169	.740	.026	.026	.311	.026	.026	.311	.026	.026	.883	.311	.169	.026	.026	.311
G	.169	.026	.169	.026	.026	.883	.026	.026	.169	.026	.169	.026	.169	.169	.311	.026	.026	.026	.026	.026	.026	.026	.026	.026
T	.046	.474	.760	.474	.760	.046	.903	.046	.189	.046	.189	.331	.046	.331	.331	.331	.760	.046	.046	.189	.189	.474	.474	.474

E. Position-specific scoring matrix: Log-odds form (with pseudocounts B = 1)[c]

A	0.21	0.32	1.34	1.34	0.72	1.34	1.34	0.04	0.32	0.72	0.21	0.21	0.32	0.32	0.48	0.48	0.72	0.04	1.34	0.32	0.21	0.32	0.32	0.72
C	0.77	1.59	1.59	0.34	1.59	1.59	1.59	1.59	0.77	0.13	1.59	1.59	0.51	1.59	1.59	0.51	1.59	1.59	0.05	0.51	0.77	1.59	1.59	0.51
G	0.77	1.59	0.77	1.59	1.59	0.05	1.59	1.59	0.77	1.59	0.77	1.59	0.77	0.77	0.51	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59
T	1.34	0.32	0.12	0.32	0.12	1.34	0.04	1.34	0.72	1.34	0.72	0.48	1.34	0.48	0.48	0.48	0.12	1.34	1.34	0.72	0.72	0.32	0.32	0.32

[a]Alignment of proven *Anabaena* NtcA-binding sites, as shown in Figure 5. Boxes shaded in red are the positions of the accepted consensus sequence.

[b]Shading indicates fraction of occurances for that base at that position: red (1.0), orange (0.8), yellow (0.6).

[d]Each element of the table is equal to the negative $\log_{10}$ of the corresponding element of Table 1D.

Table 2: Example of scoring a sequence with a PSSM

urt-71	A	T	T	T	A	G	T	A	T	C	A	A	A	A	A	T	A	A	C	A	A	T	T	C
Score[a]	.67	.50	.83	.50	.17	1.0	1.0	1.0	.17	.83	.67	.67	.50	.50	.33	.33	.17	1.0	1.0	.50	.67	.50	.50	.33
Background[b]	.32	.32	.32	.32	.32	.18	.32	.32	.32	.18	.32	.32	.32	.32	.32	.32	.32	.32	.18	.32	.32	.32	.32	.18

[a]Scoring matrix from Table 1C used.

[b]The background frequencies used to calculate the scores are **A** = **T** = 0.32; **C** = **G** = 0.18. These are the observed average nucleotide frequencies in intergenic sequences of *Anabaena* PCC 7120.

as they should be more similar to the training sequences. Each score in the PSSM can be considered a probability, and so multiplying the probabilities at each position indicates the probability that the entire sequence would arise within a true binding site. This procedure is easier to see than to explain, so take a look at **Table 2**. Notice that at each position, the score is just that determined by the sequence of urt-71 at the given position (the sequence of urt-71 may be found in Table 1A). Multiplying the individual scores together gives the joint probability that the urt-71 sequence would be produced, given the frequencies of Table 1C.

SQ10. Calculate the joint probability, by multiplying together all the individual probabilities. No, don't grind your fingers to pulp punching a calculator, copy the line and paste it into BioBIKE:

(* *paste numbers here* **)**

SQ11. Why is the product so small? How do you interpret the number? Is the urt-71 sequence very unlikely to be an NtcA-binding site?

That joint probability does look pretty low, but consider: What's the probability that DNA from your parents could have produced the genome sequence you actually possess? The probability is obviously very, very low, but still, yours is a much more likely result than, say, the genome sequence of a mongoose. We need some reasonable point of comparison.

A logical probability to compare is a random sequence. The urt-71 demonstrably exists. Is it more likely to be the result of the frequencies given in Table 1C or the frequencies of the nucleotides as they exist in general in the genome between genes? Table 2 gives these frequencies (called ***background frequencies***) as well.

SQ12. Calculate the joint probability of the background frequencies. What is the ratio of the joint probability calculated in SQ12 to the joint probability of the background frequencies?

SQ13. How do you interpret the results? Construct a relatively short English sentence that makes use of the ratio calculated in SQ12 and says something meaningful.

SQ14. Why did I use background frequencies based on the nucleotide frequencies in intergenic regions? Why not the nucleotide frequencies in the entire genome? When might I want to do that?

IV.B.2. A not-so-straightforward PSSM: Pseudocounts

Suppose that the urt-71 sequence began with a T rather than an A.

SQ15. Recalculate the ratio of SQ12 using this sequence. What do you get and what does it mean.

Is this fair? Is it fair that you should get SO different a ratio just by changing one nucleotide in what seems at first glance to be a relatively unimportant position? I should say not. This injustice

comes about because our sample size is small. The PSSM was constructed based on only six sequences. Maybe if we had a few dozen proven NtcA binding sites, we'd find some with A in the first position (in fact, see Fig. 5).

The way the problem of small sample size is addressed is to somewhat arbitrarily add mythical counts to the count totals for each nucleotide at each position. These added counts are called pseudocounts. There is no accepted theory to suggest how many counts to add. I've chosen to add 1 pseudocount, distributed to each nucleotide according to its background frequency. Since A occurs in intergenic sequences at a frequency of 0.32, I add 0.32 to the A-count at each position. The score for a given cell is then given as:

$$\text{Score for nucleotide at given position} = \frac{\text{(counts for nucleotide at that position) + (pseudocounts for nucleotide)}}{\text{Total counts at that position + Total pseudocounts}}$$

or symbolically:

$$S_{i,n} = \frac{q_{i,n} + p_n}{N + B}$$

where $q_{i,n}$ = observed counts at position ***i*** for the nucleotide ***n***

p_n = pseudocounts for the nucleotide ***n***

N = total number of sequences (= total counts at any position)

B = total number of allocated pseudocounts

$S_{i,n}$ = score at position ***i*** for the nucleotide ***n***

SQ16. Calculate $S_{i,n}$ for the first position of the PSSM for the nucleotide T, with the total number of pseudocounts taken to be 1 and distributing that one pseudocounts according to the background nucleotide frequency.

SQ17. Recalculate the ratio of SQ12 using modified urt-71 sequence (first nucleotide is T instead of A), using a value of 1 for total pseudocounts and distributing the pseudocounts B). It's too painful to look up each of the $S_{i,n}$ scores in Table 1D, so just estimate the ratio, noting that the pertinent values in Table 1C and and Table 1D are very similar to each other, except for one.

SQ18. Interpret your result from SQ17.