Bioinformatics and Bioengineering Summer Institute (2003)

Position-specific scoring matrices to search for repeated sequences
Applications of PSSMs

I. Discovery of new motifs

In the examples we've considered thus far, the beginning point is a set of aligned sequence, either repeated sequences that had been found by another program or NtcA binding sites that had been found by experiment. By far the more frequent situation is that no set of aligned sequences exists, but we wish it did.

For example, suppose you were interested understanding the immune response, in particular how certain genes are turned on specifically in response to immune challenge. You've collected several such genes and reason that there must be ***something*** in common in the regulatory regions preceding these genes, but what? It's an awful lot of DNA to eyeball successfully, so you're looking for electronic help.

PSSMs may be of service, even though you don't have an alignment of conserved motifs (that's precisely what you're hoping to discover). First, let's talk a bit about regulatory regions, by way of the well studied example, the *lac* operon in *E. coli*.

I.A. Environmental control over transcription of *lacZ*: A paradigm for gene regulation

Like most genes, expression of the gene *lacZ* is regulated primarily by whether RNA polymerase initiates transcription of it. This gene encodes the enzyme ß-galactosidase, which catalyzes the breakdown of milk sugar (lactose) into simple sugars that the bacterium can digest. If you think about it, *E. coli* has a problem. It would definitely like to be able to metabolize lactose when it finds itself in the gut of a milk-drinking baby, but making ß-galactosidase in most adults is just a waste of good metabolic energy. Also, when the simple sugar glucose is in plentiful supply, it makes no sense to bother with lactose.

Fig. 1A shows the region of the *E. coli* chromosome near the *lacZ* gene, and a closer look at the region (Fig. 1B-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. 1B). 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 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. 1D. 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* has run out of glucose and could use an alternative source of energy, CRP, a protein sensitive to the state of the cell (and indirectly to the presence/absence of glucose), binds nearby the promoter. Now RNA polymerase can bind *both* to the promoter and to CRP, very stably, and *lacZ* may be well transcribed.

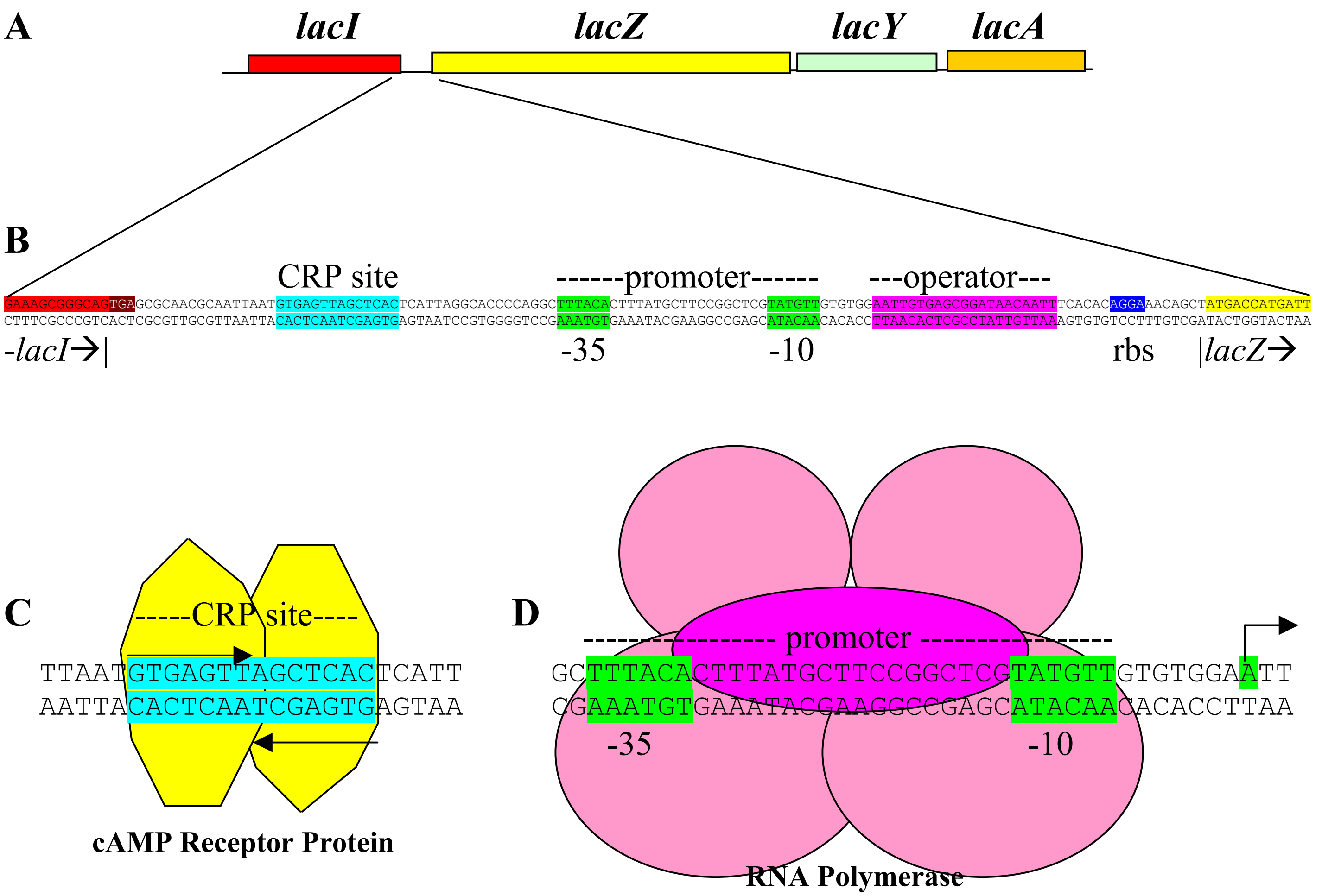
A
lacI
lacZ
lacY
lacA
B
CRP site
------promoter------
---operator---
GAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATT
CTTTCGCCCGTCACTCGCGTTGCGTTAATTACACTCAATCGAGTGAGTAATCCGTGGGGTCCGAAATGTGAAATACGAAGGCCGAGCATACAACACACCTTAACACTCGCCTATTGTTAAAGTGTGTCCTTTGTCGATACTGGTACTAA
-lacI→|
-35
-10
rbs
|lacZ→
C
-----CRP site----
TTAATGTGAGTTAGCTCACTCATT
AATTACACTCAATCGAGTGAGTAA
cAMP Receptor Protein
D
---------------- promoter ----------------
GCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATT
CGAAATGTGAAATACGAAGGCCGAGCATACAACACACCTTAA
-35
-10
RNA Polymerase

Fig. 1. 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)

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.

SQ1: What would be the result of a mutation that altered or deleted several of the nucleotides shown in green in Fig. 2D?

SQ2: What fraction of genes do you think are preceded by promoters? What fraction is preceded by CRP-binding sites?

The binding of regulatory protein to DNA nearby the promoter to stabilize the binding of RNA polymerase is a strategy used by both prokaryotes and eukaryotes to control the expression of genes. Knowing what protein binding sites precede a gene could, in principle, tell us under what conditions the gene should be expressed. And this is a lot, because the primary difference between mice and humans is not what genes they have but rather the regulation of their expression.

I.B. Strategies for identifying binding sites for regulatory protein

So how can we find and identify binding sites for regulatory protein? You'll see one clue by examining the sequence bound by CRP (Fig. 1C). Notice that there is a 7-bp sequence on one strand that is almost identical to a 7-bp sequence on the opposite strand. Another way of saying this is that the sequence is self-complementary, a palindrome.[1]

SQ3: 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?

A palindromic sequence (or a direct repeat) gives you two protein binding sites. If the protein has evolved to interact with itself productively, then the repeated structure of the binding site gives you a complex binding site (good for specificity) recognized by a protein with simpler recognition requirements. Nature thus gets more for less. Li et al (2002) exploit the tendency of Nature to use dimeric protein to regulate gene expression to find their binding sites in genomes.

But that will come later. There are also strategies to find regulatory protein binding sites that make no assumptions as to the quaternary structure of the protein. Let's return to the problem of identifying common regions upstream from genes that are turned on specifically in response to immune challenge. Consider the following procedure:

1. Construct a set of DNA sequences, consisting of a 3000 nucleotides upstream of each of the genes you've collected that are induced by immune challenge.
2. Choose at random a short sequence within one of the upstream regions.
3. Find the closest match to that short sequence in each of the other upstream regions.
4. From this set of short sequences, construct a PSSM (you knew I'd get there sometime).
5. Use the PSSM to score each of the short sequences, relative to a score based on background nucleotide frequencies
6. Adjust the length and positions of the short sequence to find the sequence in that region that gives the best possible relative score. Repeat this step until further changes worsen the score. For example, if you by chance hit on a sequence that has similarity to sequences in other upstream regions only in the leftmost five nucleotides, you might be

[1] A literary description of palindromes is provided on the web site.

[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!

able to find a more extensive match by sliding the sequence over to the left in each of the upstream regions. In this way you can reach a local maximum of possible score.

7. If the score is amongst the best you've gotten so far, then save the short sequence and the score. Otherwise toss it.
8. Repeat steps 2 through 7 many times.

In this way, you reverse the procedure we've followed the past few days: instead of using aligned sequences to construct a PSSM, you use PSSMs to find well aligned sequences.

Two popular programs are available to try to sift through sequences you think have something in common, as described above. The two, MEME (http://meme.sdsc.edu/meme/website/)[3] and Gibbs Sampler (http://bayesweb.wadsworth.org/gibbs/gibbs.html), work very similarly, but, like Blast, neither one guarantees that it will find the optimal solution to the problem, and in practice, they very often fail to find sequences that you think ought to be there. To use either, you need to supply a training set in a single file (FastA format works). You can specify a number of things, but the most important is to say whether you demand that the motif(s) sought must occur in each of the submitted sequences (oops option), must occur in either zero or one of the submitted sequences (zoops option), or may occur in any number.

PSSMs are a basic tool of bioinformatics that is used in a wide variety of other applications. One flavor of Blast (Psi-Blast: position-specific iterated Blast) allows you to align sequences you specify that were found by conventional Blast in order to make a PSSM that is used to sharpen subsequent searches. Phi-Blast (pattern-hit initiated Blast) works much the same way, except that the initial set of hits are found not by a conventional Blast query but by a submitted sequence pattern.

II. Case study: Identification of the binding sites of regulatory proteins… Li et al (2002)

First let us be clear. It is almost ***always*** difficult to understand a research article, unless you wrote it. Perhaps I could understand somewhat more of this article than you at first reading (or maybe not), but that's only a question of degree. The main advantage I may have is that I ***know*** that given enough effort, the shroud of confusion that surrounds this or any other research article will lift and all will become clear. It may take a while, though, and require reading lots of other articles.

Ordinarily, I read articles for specific reasons. I want to answer a question that the article can help with. Thus, I seldom read the article like a novel but instead take control, interrogating the article about matters important to me. I read very non-linearly, looking for answers to my questions and passing readily over matters I judge to be extraneous to that quest. In this case, I've set as my goals to understand the principle behind the method the authors used to find putative regulatory protein binding sites and to assess whether the method is effective.

[3] Meme is also implemented on the Hercules server here at VCU, a much more convenient way to use the program since output is virtually instantaneous. To connect to it, use your terminal emulator to go to hercules.vcu.edu, and enter your name and password. Then (after a few preliminaries), type meme.

II.A. Introduction

With that in mind, I read the Introduction more for general interest, not worrying about things I didn't understand [except one: is PSWM = PSSM? Answer: yes]. Sometimes I read Abstracts, but often they're too dense to get much out of. A few points and translations:

Par. 2, beginning: *One commonly used approach… is to delineate a group of coregulated genes….* They're talking in particular about the use of microarrays to identify genes that are turned on under the same conditions (hence might have the same regulatory sequences).

Par. 2, line 5: *An alternative approach is to compare the regulatory regions of orthologous genes in different species….* For "orthologous" read "similar". It stands to reason that if two species have genes encoding the same protein, they may be regulated similarly, and the regulatory protein binding sites may have been conserved over evolution.

Par. 3, Sentence 1: *…categorized as either direct search… or… guess a pattern and improve it iteratively….* You might recognize the latter category as that described in Section I.B above. The references cite descriptions of Meme and Gibbs sampler.

Par. 4, Line 5: *This bipartite character… either dimerization of the transcription factor or the presence of two DNA-binding domains… as in the case of sigma factors.* See Fig. 1. For the first case, see CRP. For the second, see RNA polymerase (of which sigma factor is a subunit).

II.B. Remainder of the article

I wish I could go further with this collaborative reading of the article, but time has run out. Instead, I propose that we split the article in the following way. Everyone is invited to read the entire article of course, but please be sure to understand as fully as you can the specific sections given below:

Methods: The algorithm

Danielle, Gaurav, Peter, Robert

Results: Deriving PSWMs from Overrepresented Dimer Patterns (including Table 1)

Emily, Josh

Results: PSWMs that identify known transcription factor-binding sites (including Table 2)

Greg, Lyndsey

Results: PSWMs that predict regulons of uncharacterized transcription factors (except Fig. 1)

David, Luke, Rebecca, Tom

Results: PSWMs that predict regulons of uncharacterized transcription factors (specifically Fig. 1)

Chris, Mark