BIOL591: Introduction to Bioinformatics
Position-Specific Scoring Matrices

Outline:

I. Prelude: Practical pattern recognition
II. Searching for motifs: simple minded approach vs PSSMs
III. Adjustments to PSSMs
    A. Adjustment to account for finite size of training set
    B. Normalization with respect to nucleotide composition
    C. Increase in size of training set through orthologs
    D. Decrease in window size through information analysis
IV. How to use PSSMs to find unknown conserved sequences

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. Searching for motifs: simple minded approach vs PSSMs

How would you go about looking for bona fide NtcA-binding sites? One approach is simply to scan the genome, looking for sequences that are identical to the identified consensus sequence (see Scenario 2) or differing from it by one nucleotide. This sounds reasonable, because this definition encompasses 19 out of 20 proven NtcA-binding sites (see Fig. 3 from notes for Sept 15).

Unfortunately, this method turns out to be less than successful, finding 8138 hits, exceeding the number of genes in the genome! Since I'm not prepared to believe that almost every gene in the genome is regulated by NtcA or any other regulatory protein, I conclude that my method doesn't work.

The dialog in **Section I** hints at the problem. An expert would not apply a strict consensus sequence, or 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.

A part of this expert process can be captured by what are called position-specific scoring matrices (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 in *Nostoc* (Table 1A). In Scenario 2, we 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.

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 close to the training sequences would tend to have higher scores at each position. The total score, i.e. the product of the scores at each position, should be higher that of most other sequences of similar length. Table 2 shows an example of how a sequence would be scored. The score of $9x10^{-7}$ does not have any meaning except in comparison with scores of other sequences of the same length, calculated using the same scoring table.

SQ1: What (on the basis of this small training set) would seem to be the most informative columns in predicting whether a sequence is an NtcA binding site?

SQ2: What does ".60" in the upper left corner of Table 1C mean?

SQ3: How was the score $9x10^{-7}$ in Table 2 obtained?

III. Adjustments to PSSMs

III.A. Adjustment to account for finite size of training set

If you reflect on the PSSM shown in Table 1C, you'll be struck by its unfairness. A single blemish in a sequence can knock the score down to zero without any hope of recovery. For example, if the sequence in Table 2 possessed a **T** in its first position, the elemental score at the first position would be zero, because none of the training sequences happen to have a corresponding **T**, and so the final product must also be zero. If you are confident that no real NtcA binding site has bases outside those in the training sequences, then a zero score is warranted, but the small number of sequences used to make the PSSM does not inspire such confidence. In practice, demanding total adherence of a test sequence to the position-dependent nucleotide content of a small set of training sequences renders the score almost meaningless.

To get around the problem of zero elemental scores, programs to calculate PSSMs have introduced what are known as pseudocounts. A certain number (**B**) of the total counts considered is set aside to reflect the overall composition of the sequences to be considered. There is evidently no theoretical justification to choose one value for **B** over another. One popular program (Gibbs Sampler) sets **B** to √N, the square root of the total number of training sequences. Another program (Meme) sets **B** to 0.1 regardless of the number of training sequences. In both cases, the influence of pseudocounts declines with the size of the training set (√N/N in the first case, 0.1/N in the other), which is just what you would want, since the purpose of pseudocounts is to diminish the distortions inherent in using a small training set. The higher the value of **B**, the more sequences will be found (including perhaps some real binding sites that might otherwise be missed) but at the cost of diluting the impact of what is known about the binding site. Lower values of **B** thus produce fewer false positives.

The score for a certain nucleotide at a certain position is then the observed counts plus pseudocounts all divided by the total number of possible counts:

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

C. Position-specific scoring matrix (B = 0)[b]

A	.60	.40	0	0	.20	0	0	1.0	.40	.20	.60	.80	.60	.40	.40	.20	.20	1.0	0	.40	.80	.40	.40	.20
C	.20	0	0	.40	0	0	0	0	.20	.80	0	0	.40	0	0	.40	0	0	1.0	.40	0	0	0	.40
G	.20	0	.20	0	0	1.0	0	0	.20	0	.20	0	0	.20	.20	0	0	0	0	0	0	0	0	0
T	0	.60	.80	.60	.80	0	1.0	0	.20	0	.20	.20	0	.40	.40	.40	.80	0	0	.20	.20	.60	.60	.40

D. Position-specific scoring matrix (B = √N = 2.2)[c]

A	.51	.38	.099	.099	.24	.099	.099	.79	.38	.24	.51	.65	.51	.38	.38	.24	.24	.79	.099	.38	.65	.38	.38	.24
C	.19	.056	.056	.33	.056	.056	.056	.056	.19	.61	.056	.056	.33	.056	.056	.33	.056	.056	.75	.33	.056	.056	.056	.33
G	.19	.056	.19	.056	.056	.75	.056	.056	.19	.056	.19	.056	.056	.19	.19	.056	.056	.056	.056	.056	.056	.056	.056	.056
T	.099	.51	.65	.51	.65	.099	.79	.099	.24	.099	.24	.24	.099	.38	.38	.38	.65	.099	.099	.24	.24	.51	.51	.38

E. Position-specific scoring matrix (B = 0.1)[c]

A	.59	.40	.006	.006	.20	.006	.006	.99	.40	.20	.59	.79	.59	.40	.40	.20	.20	.99	.006	.40	.79	.40	.40	.20
C	.20	.004	.004	.40	.004	.004	.004	.004	.20	.79	.004	.004	.40	.004	.004	.40	.004	.004	.98	.40	.004	.004	.004	.40
G	.20	.004	.20	.004	.004	.98	.004	.004	.20	.004	.20	.004	.004	.20	.20	.004	.004	.004	.004	.004	.004	.004	.004	.004
T	.006	.59	.79	.59	.79	.006	.99	.006	.20	.006	.20	.20	.006	.40	.40	.40	.79	.006	.006	.20	.20	.59	.59	.40

F. Position-specific scoring matrix: Log-odds form (B = 0.1)[c,d]

A	0.2	0.4	2.2	2.2	0.7	2.2	2.2	0.0	0.4	0.7	0.2	0.1	0.2	0.4	0.4	0.7	0.7	0.0	2.2	0.4	0.1	0.4	0.4	0.7
C	0.7	2.5	2.5	0.4	2.5	2.5	2.5	2.5	0.7	0.1	2.5	2.5	0.4	2.5	2.5	0.4	2.5	2.5	0.0	0.4	2.5	2.5	2.5	0.4
G	0.7	2.5	0.7	2.5	2.5	0.0	2.5	2.5	0.7	2.5	0.7	2.5	2.5	0.7	0.7	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
T	2.2	0.2	0.1	0.2	0.1	2.2	0.0	2.2	0.7	2.2	0.7	0.7	2.2	0.4	0.4	0.4	0.1	2.2	2.2	0.7	0.7	0.2	0.2	0.4

[a] Alignment of proven NtcA-binding sites, as discussed in Scenario 1. Boxes shaded in red are the positions of the conserved sequence used in Scenario 1 to search for putative NtcA-binding sites.

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

[c] 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 *Nostoc* PCC 7120. Table 1D was calculated with the default scoring system used by the Gibbs Sampler, and Table 1E used the default scoring system of Meme.

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

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]	.60	.60	.80	.60	.20	1.0	1.0	1.0	.20	.80	.60	.80	.60	.40	.40	.40	.20	1.0	1.0	.40	.80	.60	.60	.40
w/ps'counts[b]	.51	.51	.65	.51	.24	.75	.79	.79	.24	.61	.51	.65	.51	.38	.38	.38	.24	.79	.75	.38	.65	.51	.51	.33
Normal'd[c]	1.6	1.6	2.0	1.6	.75	4.2	2.5	2.5	.75	3.4	1.6	2.0	1.6	1.2	1.2	1.2	.75	2.5	4.2	1.2	2.0	1.6	1.6	1.8

[a] Scoring matrix from Table 1C used. The product of the elemental scores is $9\text{x}10^{-7}$.

[b] Scoring matrix from Table 1D used.

[c] Scoring matrix from Table 1D used, correcting for background nucleotide frequencies by dividing the raw score (with pseudocounts) by the frequency of the given nucleotide. The product of the elemental scores is $3.2\text{x}10^{5}$.

Score(position,nucleotide) = (**q** + **p**) / (**N** + **B**)

where **q** = observed counts for the nucleotide at the given

p = pseudocounts = **B** (overall frequency of nucleotide)

N = total number of sequences (= maximum number of observed counts)

B = total number of allocated pseudocounts

In the example shown in Table 1D, the score for an adenine in position one is calculated:

Score(position 1, **A**) = $[3+\sqrt{5}\,(0.32)]\ /\ [5+\sqrt{5}] = 0.51$

(where 0.32 is the fraction of nucleotides in the intergenic sequences of *Nostoc* that are adenines). This score isn't much different from the score without using pseudocounts. The main difference is in scores that would otherwise be zero, e.g. the change from 0 to 0.99 in the case of thymine in the first position.

SQ4: What is the practical effect of a very small value of B? A very large value of B?

SQ5: Calculate yourself the value of 0.59 in the upper left corner of Table 1D.

III.B. Normalization with respect to nucleotide composition

The overall score derived from a PSSM can be deceptive, because a PSSM derived from a set of training sequences with a base composition similar to the overall base composition will give an arbitrary sequence a higher score than a PSSM derived from a set of training sequences with a base composition deviating from the norm. To eliminate this bias, each elemental score is compared to the frequency with which the given nucleotide occurs in the greater population of sequences to be considered (not the training set but, in this example, all intergenic sequences within *Nostoc*). :

Normalized score = raw score / (overall frequency of given nucleotide)

For example, the score for adenine in position 1 (not shown on any table) would be normalized:

Normalized score(position 1, **A**) = 0.51 / 0.32 = 1.6

Traditionally, scoring tables are given as logs of the scores or the negative logs of the scores, because the addition of logs is a faster operation on the computer than the equivalent multiplying of the elemental scores. Table 1F shows an example of a PSSM in log-odds form.

SQ6: Calculate the normalized value of the usual upper left hand square of Table 1E, presuming that all nucleotides occur with equal frequency (i.e. background frequencies are all 0.25).

SQ7: Calculate the upper left value of Table 1F.

III.C. Increase in size of training set through orthologs

Now we're prepared to attack the problem at hand: what genes in *Nostoc* are preceded by sequences that look like NtcA binding sites according to a PSSM derived from sequences of bona fide NtcA binding sites, and a problem in the problem set will be to do just that with the material at hand. The results will not turn out to be very useful, however, for two reasons. First, five sequences is not a very big training set, and accidental similarities amongst the sequences will bias the results, giving spurious hits of sequences that happen to have the same accidental similarities. With a larger training set, the impact of these accidents will be diluted and the similarities necessitated by the binding of NtcA will be further accentuated. But how do we get those sequences?

Our resources in *Nostoc* PCC 7120 have been exhausted – there are no more known NtcA binding sites. However, it is reasonable to suppose that the same genes known to possess NtcA binding sites in this organism will be regulated in the same way in closely related organisms. It seems legitimate, therefore, to expand the list of training sequences with orthologous genes in other *Nostoc*s. I've done this using sequences from *Nostoc punctiforme*, another strain whose genome has been sequenced, yielding the training set shown in Table 3. Doubling the size of the training set in this way should increase the accuracy of prediction of NtcA binding sites.

III.D. Decrease in window size through information analysis

Table 3 makes clear a second problem. The bases that are highly conserved are only a small fraction of the size of each training sequence. At one extreme, we could confine our attention to only the six conserved sequences of the NtcA-binding site. If we do this, we will learn nothing that we did not already know. The benefit of PSSMs lies in going beyond highly conserved nucleotides… but how far? If we use all 76 nucleotide positions of the training set to score candidate sequences in the *Nostoc* genome, then accidental matches may swamp out similarities that are actually important in NtcA binding. How can we strike a balance between specificity and comprehensiveness?

Fortunately, there's a way of having a good bit of both. Upon inspection of the PSSM, it is clear that some positions are more informative than others. For example, the first position of the sequences in Table 3 has instances of each of the four nucleotides, with none clearly predominating. We wouldn't expect that this position would contribute much to discriminating true from false matches. On the other hand, the shaded GTA is highly discriminating. A true NtcA binding site is not likely to have any bases there besides GTA. We would like to give greater weight to those sites that are more discriminatory. How?

Table 3: Training set including sequences from two *Nostoc*s[a]

```
71-devB  CATTACTCCTTCAATCCCTCGCCCCTCATTTGTACAGTCTGTTACCTTTACCTGAAACAGATGAATGTAGAATTTA
Np-devB  CCTTGACATTCATTCCCCCATCTCCCCATCTGTAGGCTCTGTTACGTTTTCGCGTCACAGATAAATGTAGAATTCA
71-glnA  AGGTTAATATTACCTGTAATCCAGACGTTCTGTAACAAAGACTACAAAACTGTCTAATGTTTAGAATCTACGATAT
Np-glnA  AGGTTAATATAACCTGATAATCCAGATATCTGTAACATAAGCTACAAAATCCGCTAATGTCTACTATTTAAGATAT
71-hetC  GTTATTGTTAGGTTGCTATCGGAAAAAATCTGTAACATGAGATACACAATAGCATTTATATTTGCTTTAGTATCTC
71-nirA  TATTAAACTTACGCATTAATACGAGAATTTTGTAGCTACTTATACTATTTTACCTGAGATCCCGACATAACCTTAG
Np-nirA  CATCCATTTTCAGCAATTTTACTAAAAAATCGTAACAATTTATACGATTTTAACAGAAATCTCGTCTTAAGTTATG
71-ntcB  ATTAATGAAATTTGTGTTAATTGCCAAAGCTGTAACAAAATCTACCAAATTGGGGAGCAAAATCAGCTAACTTAAT
Np-ntcB  TTATACAAATGTAAATCACAGGAAAATTACTGTAACTAACTATACTAAATTGCGGAGAATAAACCGTTAACTTAGT
71-urt   ATTAATTTTTATTTAAAGGAATTAGAATTTAGTATCAAAAATAACAATTCAATGGTTAAATATCAAACTAATATCA
Np-urt   TTATTCTTCTGTAACAAAAATCAGGCGTTTGGTATCCAAGATAACTTTTTACTAGTAAACTATCGCACTATCATCA
```

[a] Sequences with proven NtcA-binding sites from *Nostoc* PCC 7120 (71) and similar sequences upstream from orthologous genes from *Nostoc punctiforme* (Np). The *hetC* gene from *Nostoc* PCC 7120 has no obvious ortholog in the portion of the genome of *Nostoc punctiforme* that has been sequenced thus far. The standard conserved nucleotides are shaded in red. Each sequence constitutes the region between the conserved GTA and TAC, 31 nucleotides upstream and 31 nucleotides downstream.

We can quantitate the discriminatory power of each position through the concept of uncertainty and information content. Uncertainty, in words, means something like how many yes/no questions you'd have to ask in order to determine the information under consideration. For example, the column with all **G**'s requires no questions to determine the nucleotide at that position in the training set: it's **G**. However, the first column, one would have to ask some questions. How many? One might think four: Are you **A**? **C**? **G**? **T**? but you can do better than that. Just two questions are sufficient: Are you a purine (**A** or **G**)? Do you participate in only two hydrogen bonds (**A** or **T**)? So the uncertainty lies somewhere between 0 (perfect information) and 2 (no information). The formula for uncertainty for in a column (*c*) is:

Uncertainty ($\mathbf{H_c}$) = - Sum [p_{ic} $\log_2(p_{ic})$]   (summed over all four nucleotides)

Where p_{ic} is the fraction of nucleotides in column *c* that is nucleotide *i*. So, for the first column, the uncertainty is (calculating for A, then, C, then G, then T):

$$\mathbf{H_1} = -\{[4/11 \log_2(4/11)] + [3/11 \log_2(3/11)] + [1/11 \log_2(1/11)] + [3/11 \log_2(3/11)]\}$$
$$= 1.87$$

pretty close to 2, while the uncertainty for the column to the left of GTA is:

$$\mathbf{H_{31}} = -\{[1/11 \log_2(1/11)] + [1/11 \log_2(1/11)] + [1/11 \log_2(1/11)] + [8/11 \log_2(8/11)]\}$$
$$= 1.28$$

By setting a suitable threshold, we can filter out the poorly discriminating columns and focus only on those that can help us.

Converse to the concept of uncertainty is the concept of information content: the greater the uncertainty, the less the information content. It is the information content that is generally reported by programs that generate PSSMs. It can be thought of as how far away in uncertainty a sequence is from maximal uncertainty and can be calculated:

Information content = Sum ($\mathbf{H_{max}} - \mathbf{H_c}$)   (summed over all columns)

So the first term in this sum would be (2 – 1.87) and the 31st term would be (2 – 1.28).

Now, finally, we can examine the program that might help us find the mystery gene that is regulated by NtcA and in turn regulates the expression of *hetR* to determine heterocysts differentiation. The program *FindMotif* (available from Scenario 5) does that.

SQ8: What is the information content at a position where all nucleotides are identical? What is the uncertainty?

SQ9: Which position(s) in Table 1 would you expect to have the highest information content? The lowest?

SQ10: Calculate the information content of the 36th column (the second column after GTA).

IV. How to use PSSMs to find unknown conserved sequences

In the present case we have the advantage of knowing already in some cases where NtcA binds. More often, we have only a collection of genes that are possibly coregulated – may they turn on at the same time and place during development or turn on in response to the same environmental stimulus (e.g. heat, hormone, etc.). From physiological and genetic information, we might collect genes whose upstream regions should share some regulatory sequence in common. But how to find that sequence? How can we build a PSSM when we have no idea how to align the very different upstream regions?

Two main programs are available to try to sift through sequences that the user feels has something in common. The two, MEME (http://meme.sdsc.edu/meme/website/) 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 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.