Bioinformatics and Bioengineering Summer Institute (2003)

Position-specific scoring matrices to search for repeated sequences

I. The Scenario: Search for a family of iterated sequences in the genome of *Nostoc*

Nitrogen-fixing cyanobacteria: Eat air and prosper!

Certain cyanobacteria, for example *Nostoc* PCC 7120, are among the only creatures on earth able to survive on CO_2 as a source of carbon, N_2 as a source of nitrogen, water as a source of electrons, and sunlight as a source of energy. This is quite a trick, because the process of fixing carbon with electrons from water necessarily produces O_2 as a byproduct, and the process of fixing N_2 is irreversibly inactivated by tiny amounts of oxygen (Fig. 1). *Nostoc* is able to protect the machinery of nitrogen-fixation from inactivation by producing specialized cells, called heterocysts, that rigorously exclude oxygen from within them (Fig. 2).

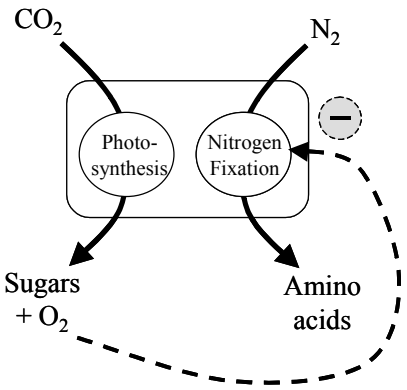


Fig. 1: Impossibility of simultaneous photosynthesis and N_2 fixation in same cell. O_2 produced by one inhibits the other.

Fixing nitrogen costs: How to pay only when needed?

Heterocysts are expensive to make and expensive to operate. More than half of the energy available to *Nostoc* is spent on heterocysts. So when an alternative source of nitrogen is present, like ammonia, *Nostoc* loses its heterocysts. And when that source is consumed or removed, vegetative cells differentiate into heterocysts within about 18 hours. This seems like intelligent behavior, but we're talking about a simple bacterium. How do the cells sense nitrogen-deprivation and translate that perception into the induction of the genes necessary for heterocyst differentiation? At present, the full answer to this question is not known.

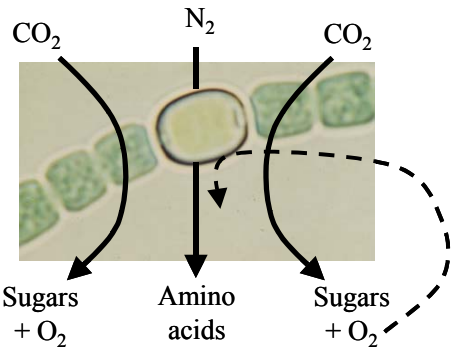


Fig. 2: Solution by *Nostoc*: Photosynthesis in vegetative cells and N_2 fixation in O_2-resistant heterocysts.

But we do know part of the answer: The absence of ammonia or any other source of nitrogen in the environment is sensed by a protein NtcA, which is activated by a metabolite that is present in high amounts only under conditions of nitrogen-starvation. Activated NtcA then turns on the transcription of many genes useful in fending off the effects of nitrogen-starvation.

This sounds like a satisfying story. Unfortunately, one of the most useful genes, *hetR*, is <u>not</u> turned on by NtcA (Fig. 3). This gene encodes a protein that is essential for heterocyst differentiation and appears to regulate the initiation of the process. NtcA is necessary for the transcription of *hetR* and the subsequent differentiation of heterocysts, but it is not <u>directly</u> involved in that transcription. There must be some unknown gene in between. This mystery gene must be turned on by NtcA activated by nitrogen starvation. Its product then turns on *hetR*, and heterocyst differentiation begins.

Complex behavior, Complex genome

Photosynthesis, nitrogen fixation, differentiation... these are things that most bacteria don't do, and no bacterium besides *Nostoc* and its relatives do all three. It may therefore come as no surprise to learn that the genome of *Nostoc punctiforme* is bigger than any bacterial genome known thus far. Most the large size is explained by a large number of genes relative to other bacteria. That's not all, however. *Nostoc* and its relatives also have thousands of instances of long, tandemly repeated DNA in its genome (Fig. 4). In this, *Nostoc* is like many eukaryotes and quite unlike other bacteria.

Fig. 3: Connection between nitrogen source and regulation of differentiation of heterocysts

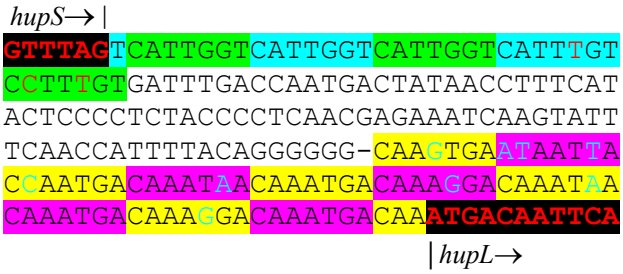


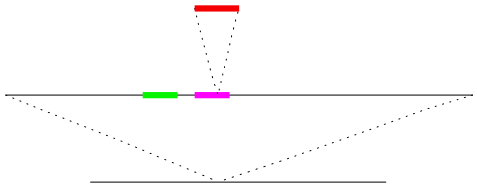
Fig. 4: Typical heptameric repeats in *Nostoc*. Shown is the intergenic region between genes *hupS* and *hupL* (both highlighted in black with red letters). The intergenic region is bounded by two sets of heptameric repeats. The first has as its unit CATTGGT and the second CAAATGA. Note that the two sets are well over 80% complementary to each other.

You are interested in the heptameric repeats that lie within the intron interrupting the gene encoding leucine-tRNA (Fig. 5). Yes, that's right. Bacteria aren't supposed to have introns, but cyanobacteria do -- this intron at least. While all cyanobacteria (and chloroplasts) appear to carry the intron within the leu-tRNA gene, *Nostocs* go a step further and interrupt the intron with a pair of heptameric repeats. The number of repeated units varies from strain to strain within the range of 2 to 5 units per set.

You're struck by an exceptional case: *Nostoc* strain Nos37 (and one other strain) (Fig. 5). Its intron carries is interrupted at still another level: a 24-bp sequence is interpolated within the second set of heptameric repeats. Tandem repeats are one thing. There are known mechanisms to create tandem repeats, but where could 24 bp have come from?

Transposable elements (also called transposons or insertion sequences) insert themselves at nearly random positions in genomes. However, natural transposons carry transposases, enzymes that catalyze the process of transposition. Transposases are no smaller than about 200 amino acids, which translates to genes no smaller than 600 nucleotides. A transposon carrying such a gene must be at least several hundred nucleotides in length, and most transposons have lengths over a 1000 nucleotides. Is this 24-bp sequence a novel microtransposon? Does it represent an unrecognized means by which DNA may be propagated?

This is important stuff for a few reasons. First of all,



Nostoc (*E. vellosus*) Nos23

AAAAGTTCTGAGTGCTGAGTGGTGAGTGGTGGATA
AACTAAAAATTCATAACTCCTAATTCATAACTCCT
AACTATTCGG
AA

Nostoc (*P. britannica*) Nos37

AAAAGTCGTGAGTACTGAGTAGTGAGTAAATTAA
AACTCTTAACTCTT
GTACAGACGTGATTAATCGCGTCT
AACGATTCGG

Fig. 5: Repeated sequences within intron of tRNA-leu(UAA). Structure of tRNA-leu(UAA) in cyanobacteria is shown. In all cyanobacteria tested, the gene contains an intron. In *Nostocs*, the intron contains paired heptameric repeats (green and magenta). In rare *Nostoc*, the second set of heptameric repeats is interrupted with a 24-bp sequence (red). Two example introns are shown. The intron from *Nostoc* (*E. vellosus*) Nos 23 is typical of *Nostocs*. The intron from *Nostoc* (*P. britannica*) Nos 37 is a rare instance where the second heptameric repeat contains an interrupting sequence.

understanding the ways of transposons has enabled us to understand the way genomes evolve and also to manipulate genomes. If this is a new, equally important phenomenon,…. Furthermore, repeated sequences, however they may arise, are suspected to lead to genomic rearrangements. The genomes of *Nostocs* rearrange at a very rapid rate, as shown in Fig. 6. Gene order is poorly conserved in the genomes of the two *Nostocs* shown. The largest segments that have retained contiguity are no larger than about 10 kb, i.e. about 0.1% of the total genome size!

Is this 24-bp sequence transposable? Wait a second -- let's not get ahead of ourselves. First of all, how many times does this sequence appear in the genome? If it's just once, then in the absence of a possibility of comparison, we'll probably never know how the sequence got there.

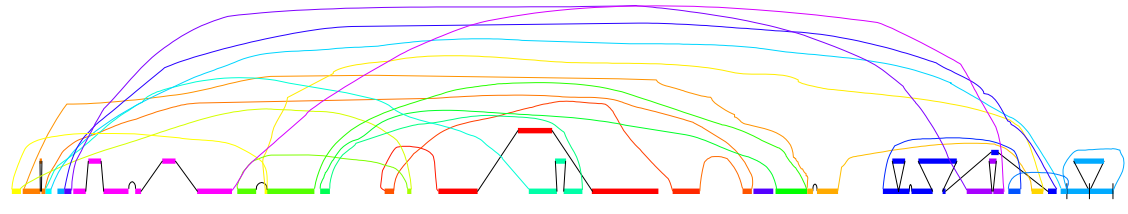


Fig. 6: Alignment of the genomes of *Nostoc* PCC 7120 and *Nostoc punctiforme*. A small portion (148,196 nucleotides) of the 9 megabase genome of *Nostoc punctiforme* is represented as a gray line. On top of it is a representation of the entire 6.4 megabase circular genome of *Nostoc* PCC 7120. Each color represents less than 20% of the genome. The placement of a colored segment next to the gray line indicates that these portions of the two genomes are similar to each other.

The simplest way to check on copy number is to use BLAST. We'll learn a good deal more about this program, but for now, suffice to say that it compares a sequence you give it with a large database, typically a genome or an even larger set of DNA sequences. When you use BLAST to compare the sequence to the genomic DNA of *Nostoc* PCC 7120, you find 25 exact copies. That's a lot! Very few transposons occur in organisms with higher copy number than that.

Still, you're not satisfied. BLAST found exact copies, but you know that transposons can still hop even when there are minor defects in their sequences. Perhaps there are more (maybe ***many*** more) copies of the 24-bp sequence besides those found by BLAST.

How can you get a complete catalog of sequences similar to the 24-bp sequence?

SQ1: Why doesn't everybody fix atmospheric N_2. Sounds like a free lunch, no?

SQ2: Two kinds of repeated sequences have been introduced: tandem and dispersed. The the latter class was not named or defined. How do you think it differs from tandem repeats? Which class do you think transposons fall into?

SQ3: Fig. 4 makes the claim that two sets of tandem repeats are highly complementary to each other. Do you agree?

II. Interlude

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.

III. Position-specific scoring matrices (PSSMs)

III.A. Simple-minded strategy

How would you go about looking for nearly identical copies of the 24-bp interpolation? One approach is simply to scan the genome, looking for sequences that are identical to the interpolation or differ from it by one nucleotide, or two, or three....

One problem is deciding where you draw the line. This problem can be addressed statistically (and ***is*** addressed in the Problem Set), so let's go… OK, you've now written a program to scan the *Nostoc* genome for all sites with no more than 4 differences from the original 24-bp sequence. You can see the results in Fig. 7.

There certainly are more plausible copies of the 24-bp sequence than the 25 exact copies found by BLAST. However, another worry arises. You're really looking not for sequences that you as a human think are similar to the 24-bp interpolation but those that are ***functionally*** similar, as judged by *Nostoc*. Some differences may seem minor to you but major to the cyanobacterium. How can you judge importance from a cyanobacterial perspective?

Short of an in depth interview with a cooperative cell, the best we can do is to try to extrapolate from 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 consonent.

We don't have such a set, since we have no way of assessing the function the 24-bp sequence may (or may not) possess. Fortunately, there's a reasonable alternative.

III.B. Random vs systematic change

An expert human would not apply a strict consensus sequence, or apply a strict rule (e.g. one mismatch allowed) but instead consider a sequence in light of his accumulated experience. He would look at many characteristics, perhaps some subconsciously, and allow candidates the same kind of imperfections as he has observed with real sequences.

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. As an example, let's consider sequences surrounding the proven NtcA binding sites in *Nostoc* (Table 1A). In the program FindNtcA.TRU (used in Introduction to Computer Science), we looked for binding sites in the genome using 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 9×10^{-7} does not have any meaning except in comparison with scores of other sequences of the same length, calculated using the same scoring table.

SQ4: 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?

SQ5: What does ".60" in the upper left corner of Table 1C mean?

SQ6: How was the score 9×10^{-7} in Table 2 obtained?

III.C. Adjustment in PSSMs to account for finite size of the 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

GTACAGACGTGATTAATCGCGTCT
GTAGAGACGCGATTTATCGCGTCT
GTACGGACGCAATTTATCGCGTCT
GTGGAGACGCGAGCAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTGAAGACGCGATTCATCGCGTCT
GTAGAGACGCGATGAATCGCGTCT
GTAGAGACGCGATGAGTCGCGTCT
GTAGAGACGCGATTCATCGCGTCT
GTAGAGACGCGATTCATCGCGTCT
GTAGAGACGCGACGAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTAGAGACGCGATTCATCGCGTCT
GTAGAGACGCGATGAATCGCGTCT
GTAGAGACGCGATTTATCGCGTCT
GTACAGAAGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTAGACAGGCGATTCATCGCGTCT
TTAGAGGCGCGATTTATCGCGTCT
GTAGAGACGCGATGAATCGTGTCT
GTACAGACGCGATTAATCGCGTCT
TTAGATATATGATTAATCGCGTCT
GTACAGACATGATTAATCGGGTCT
GTAAAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
TTAGAGACGCGATTAATCGCTCGG
GTAGAGACTCAGTTCATCGTGTCT
GTAGAGACGCGATTCATCGCGTCT
GTAGAGACGCGATTCATCGCGTCT
GGACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACTCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
AAAAAGACGCGATGTATCGCGTCT
GTAGATACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCATCT
GTACAGACGCGATTAATCGCGTCT
GTACAGATGCGATTCATCGGATGT
GTAGAGACGCGATGAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACTCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTACAGACGCGATTAATCGCGTCT
GTAGAGACGCGATTCATCGCGTCT

Fig. 7: Sequences in genome of *Nostoc* PCC 7120 similar to original 24-bp sequence. Only sequences with 4 or fewer defects were retained. The most common nucleotide at each position is colored yellow. In position 4, the relatively common guanine residues are colored cyano.

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'count	.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[b]	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 9×10^{-7}.

[a] 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 3.2×10^{5}.

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:

$$\text{Score(position,nucleotide)} = (\mathbf{q} + \mathbf{p}) / (\mathbf{N} + \mathbf{B})$$

where

q = observed counts for the nucleotide at the given position

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:

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

(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.

SQ7: What is the practical effect of a very small value of B? A very large value of B?

SQ8: Calculate yourself the value of 0.59 in the upper left corner of Table 1D.

III.D. 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, for the normalized score for adenine in position 1 of Table D (not shown on any table) would be:

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.

SQ9: 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).

SQ10: Calculate the upper left value of Table 1F.

III.E. Decrease in window size through information analysis

If we were going to seek NtcA binding sites by means of a PSSM, we'd have to do a lot better than the meager sequences shown in Table 2. Let's expand the size of the training set and also the length of each sequence considered (Table 3). Expanding the length makes it possible to discern conserved positions that we might have missed from the smaller sequences. However, doing this poses a question: how far should we go in this direction? 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 from the previous program. 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,

Table 3: Training set including sequences from two *Nostocs*[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.

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?

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 within 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 (terms with zero logs are not considered). 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.

Mirroring 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 rank the various 24-bp sequences in the genome of *Nostoc*. The program *PSSM.TRU* (available from the web site) does that.

SQ11: What is the information content at a position where all nucleotides are identical? What is the uncertainty?

SQ12: Which position(s) in Table 1 would you expect to have the highest information content? The lowest?

SQ13: Calculate the information content of the 36th column (the second column after GTA).