Introduction to Bioinformatics
Making and using Position-Specific Scoring Matrices

We left our heroes struggling to identify p53-binding sites near human genes, hoping to find out which genes might be regulated by this known to be important in preventing the unrestrained cellular growth typical of tumors. Hoh et al (2002) first constructed a table of counts for each position of aligned p53-binding sites. Rather than follow their efforts from a distance, we will jump in and do the deed ourselves, within BioBIKE. Since BioBIKE is at its best with cyanobacterial sequences, we'll switch to an analogous cyanobacterial problem: the identification of sites bound by the cyanobacterial regulatory protein, NtcA. You've encountered NtcA already (see notes for March 1), but NtcA… p53… the biology doesn't matter for the moment. The techniques to identify binding sites are the same.

I. Making a PSSM

We'll make the PSSM in two stages: (1) Construct a table of counts of each nucleotide at each position, and (2) Convert the table of counts to a table of modified frequencies, the PSSM. There is a preceding stage of gathering and aligning proven binding sites, but we'll presume that has already done for us, as it has been in the case of NtcA-binding sites.

I.A. Construct a table of counts

The first task is to bring into BioBIKE the set of aligned NtcA-binding sites published by Hererro et al (2001) and given on Fig. 5 (p.7) of the notes of March 1. Go to those notes to refresh your memory of what we're after. Then:

1. Log onto BioBIKE (VCU public cyanobacterial edition)
2. Enter BBL-MODE if you're not there already*
3. Click on **Browse Files**, then …**/shared-files/**, and finally **ntca-binding-sites-gapped.txt**

This is our starting point. We have all of the sequences shown in Fig. 5, each in FastA format. FastA format specifies that a sequence is preceded by a one-line description, beginning with the character ">".

SQ1. What are key differences between Fig. 5 and the file accessible to BioBIKE?

SQ2. Why is it important that the sequences within the file have the same lengths?

Use the browser back button to get back to the web listener. We need to read in the file we were just looking at. To do this:

4. Type `(READ-FASTA-FILE "ntca-binding-sites-gapped.txt" SHARED)`
5. Put this in a variable: `(DEFINE ntca-sites AS *)`
6. Examine what `ntca-sites` looks like by typing the name of the variable
7. Notice that there are NtcA-binding sites from several organisms. Make a subset of the list `ntca-sites` that consists only of binding sites from *Synechocystis* PCC6803 (S6803). Call it `S6803-ntcA-sites`.

* By typing `(ENSURE-BBL-MODE)` you make sure that you're always in BBL-MODE now and forever.

We want to go through each position in each sequence of `S6803-sites` counting how many A's, C's, G's, and T's there are, thereby filling in the counts table that we can visualize as shown below:

	1	2	3	4	5	6	7	8	9	10	. . .
A											
C											
G											
T											

Make up a table, called test-table, and set the cell at position 3, nucleotide "G", to 1, like so:

(DEFINE test-table[1 "G"] = 1)

and:

(DEFINE test-table[2 "A"] = 1)

Finally display the table:

(DISPLAY-TABLE test-table)

You can also change the value of a table by incrementing individual cells:

(INCREMENT test-table[2 "A"])

Now we the tools to make a strategy: Go through each column and each sequence. Increment the cell of the letter from the sequence at the current column. Do this for every sequence and every column.

SQ3. Complete the loop below to create a counts table for `S6803-ntcA-sites`.

```
(DEFINE number-of-columns AS (LENGTH-OF S6803-ntcA-sites[1][2]))
(FOR-EACH column FROM 1 TO number-of-columns
     (FOR-EACH (label sequence) IN S6803-sites
          AS letter = fill in
          (INCREMENT counts-table[ fill in ])))
```

Then display `counts-table`.

SQ4. What is the significance of `(LENGTH-OF S6803-ntcA-sites[1][2])` in the first line of the code shown in SQ3? Why [1][2]?

SQ5. Do the values in the counts table correspond to what you'd expect from the set of aligned sequences?

I.B. Construct a table of frequencies

The counts table cannot be used directly to evaluate unknown sequences. Recall from the notes for March 1 that the evaluation process entails multiplying fractions corresponding to the frequency of a given nucleotide at a given position. Fortunately, it's not difficult to transform a table of counts to a table of frequencies. You just need to create a new table (call it `freq-table`) where each cell is the corresponding cell of `counts-table` divided by the total number of letters in the column.

SQ6. Modify the code you completed in SQ4 to define `freq-table`. Hints:

a. Define `number-of-rows` in a much as you defined `number-of-columns`

b. Figure out the relationship between `freq-table` and `counts-table`. How would you define a specific cell of `freq-table`?

c. Loop through every cell of `counts-table` (not each element of S6803-sites)

Display `freq-table` as follows:

```
(DISPLAY-TABLE freq-table COLUMN-WIDTH 3)
```

SQ7. Do the values in the frequency table correspond to what you'd expect from the set of aligned sequences and from the counts table? What about column 41? Does it add up to 100%? (It may or may not, depending on your method)

SQ8. Notice the many zeros in the table. What problems do these pose?

We now have to solve those problems. We do so by adding one count to the total of each column of the counts table and a fractional count to each cell before calculating the frequency for that cell. Ideally, the fractional count should be related to the background frequency of the nucleotide, but for now, make the fractional count 0.25.

SQ9. Modify the code you completed in SQ6, changing the calculation of each cell to reflect the formula:

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

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*

Then display the table.

SQ10. Do the values in the modified frequency table make sense?

Risking a reaction that may range from fury to exhilaration, I must tell you know that all you have done is already built into BioBIKE. Try this:

```
(DEFINE freq-table2 AS (MAKE-PSSM-FROM S6803-sites))
```

Then display the table.

SQ11. How does the table you constructed by hand compared to the table constructed by `MAKE-PSSM-FROM`?

SQ12. Don't you wish you had read ahead? [Ans: No! You're a better person by being able to make a PSSM by hand!]