BNFO 301: Introduction to Bioinformatics
Problem Set: Motif Search Using PSSM's

1. Bring in the NtcA sites:
   (LOAD-SHARED-FILE "ntca-sites")

2. Bring in information about the ntcA genes to which the sites are attached:
   (LOAD-SHARED-FILE "ntca-genes")

3. Extract the NtcA sequences as a list. Do this by making a loop that goes through ntca-sites, collecting out the second element of each list.

4. [Watch as I bring the list over to another version of BioLingua (where MEME works) and run the sequences through MEME]

5. Make a list called ntca-genes consisting of the genes contained in ntca-gene-info. Do this by making a loop that goes through ntca-gene-info, extracting the second element of each list (the gene name), and collecting the (GENE-NAMED gene-name).

6. Find the orthologs of these *Anabaena* PCC 7120 (A7120) genes in the closely related cyanobacterial strains *Anabaena variabilis* (Avar) and *Nostoc punctiforme* (Npun). Make use of the function:
   (ORTHOLOGS-OF gene-list IN organism-list)
   (it won't seem obvious why now, but include A7120 in the list as well).

7. Find the sequences upstream of the orthologs that are similar to the NtcA sites in *Anabaena*. A quick and dirty way is Blast:

```
(FOR-EACH set in ntca-orthologs
      FOR item IN ntca-sites
      AS motif = (SECOND item)
      AS label = set
      AS target = (INTERLEAVE label
                              (SEQUENCES-UPSTREAM-OF set LENGTH 800))
      AS hits = (BLAST motif target)
      COLLECT hits)
```

8. You'll find that the results are less than satisfactory. Too many missing hits! Make Blast less strict by modifying one of the lines to:

```
AS hits = (BLAST motif target :WORD-SIZE 9 :CUT-OFF 1)
```

9. Now too many hits! Take only the top three (one for each organism) by modifying one of the lines to:

```
COLLECT (FIRST-N 3 hits))
```

10. Remove the internal structure of the list (note the parentheses within parentheses) by flattening it:

```
(FLATTEN *)
```