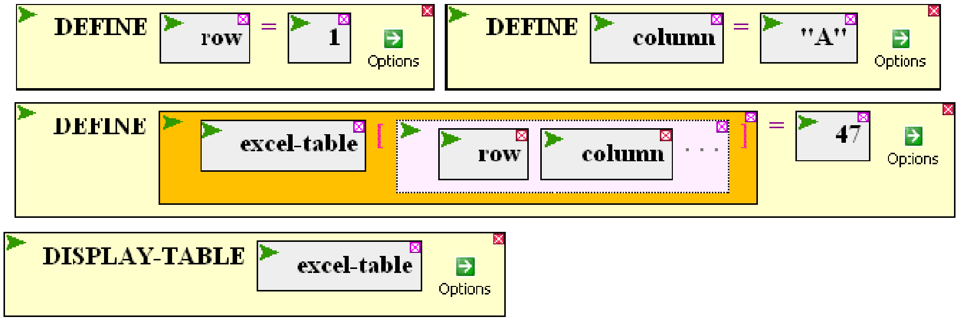
Introduction to Bioinformatics

Problem Set 8: Tables and Motif Discovery Through PSSM's

Tables and the [ ] notation

1. Create a table of the form letter[number], where letter[1] = "A", letter[2] = "B", etc.

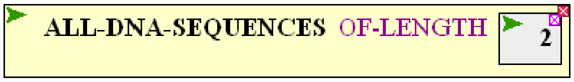
2. Make and display an x * y multiplication table for x and y, each going from 1 to 15. The [] function (in the List/Tables menu) will be essential here. You can use it to define elements of a two-dimensional table, as shown by example below:



```
DEFINE row = 1   Options        DEFINE column = "A"   Options
DEFINE excel-table [ row column . . . ] = 47   Options
DISPLAY-TABLE excel-table   Options
```

3. Make and display a table containing information about the organisms known to BioBIKE. The row labels should be the names of the organisms. The column labels should genome size, number of genes, and GC-fraction.

4. Determine the frequency of each dinucleotide in the genome of the cyanobacterium ss120. Put the values in a table and then display it. The following function will be useful:



```
ALL-DNA-SEQUENCES OF-LENGTH 2
```

5. The [ ] notation can be used for other types of data besides tables. Think of `a[b]` as "*a sub b*". Define a list of some sort, and use the [ ] function to pull out its third element. Try the same thing with a string or a sequence.

6. Use tables within a loop that counts every instance of 13-mers (13-nucleotide sequences) within the che12 genome. Sort the resulting table, sorting the most abundant 13-mers to be t the top of the list, and display it. Compare your list to the sequences found by Gomathi et al.

Using PSSM's to find sequence motifs

7. In last week's lab you found motifs of iron-sulfur proteins – now use it!

7a. Redo the steps to obtain the motifs, and identify a motif with the characteristic C**C**C***C pattern.

7b. Extract the alignment of sequences from that motif (output by MOTIFS-IN). One of the FIRST, SECOND,… functions will be useful for this purpose, once you know which of the functions is the one you want.

7c. With that alignment in hand, run APPLY-PSSM-TO the PROTEINS-OF any bacterium, using the alignment to fill in the box governed by the WITH-PSSM-FROM option. The output will be of the form: (protein amino-acid-coordinate FORWARD score).

7d. Extract the proteins from the list produced by APPLY-PSSM-TO. The FIRST function will be useful, if you use the IN-EACH pre-option.

7e. Display the descriptions of the proteins. Are any/all iron-sulfur proteins?

8. If your research project is protein-oriented, use MOTIFS-IN and APPLY-PSSM-TO to look for instances of your favorite protein in different phage genomes.