(22 Jan 2013)

Introduction to Bioinformatics
Problem Set 2: Molecular Biology Investigations

NOTE WELL: Every question should be answered with a tangible demonstration and/or argument of the validity of your answer. Always seek to give specific examples and numerical arguments, when appropriate.

1. Where are the genes in a genome? Examine the genome sequence of *Prochlorococcus marinus* ss120 (nicknamed ss120) in CyanoBIKE.
   a. First of all, note that the genome sequence is decorated with several colors. What is the significance of those colors? (Have a hypothesis? ***Test*** it and provide numerical evidence for or against it)
   b. More colors… how many are there? What is the significance of the ***specific*** colors? Why is one red used in one place and blue in another? (Give specific examples to bolster your explanation)
   c. What are the first three nucleotides for genes of different colors? Revisit **1b**?
   d. What are the coordinates for genes of different colors? What's the significance of the arrow? Revisit **1b** and **1c**?
   e. Note that there are regions of the chromosome that are black. What is their significance? Why are some black regions very small and others much larger?

2. How are gene sequences related to protein sequences? Keep examining the ss120 genome.
   a. Consider the DNA sequence of the gene pro0001. While considering it, go back to BioBIKE and display the sequence of its corresponding protein. You can do this in either of two ways:
      i. SEQUENCE-OF p-pro0001 [putting "p-" in front of a gene signifies its protein]
      ii. SEQUENCE-OF PROTEIN-OF pro0001
      [*You can find PROTEIN-OF on the GENES-PROTEINS menu*]

      Using the genetic code, make sure that the DNA sequence of pro0001 in fact encodes at least the beginning of p-pro0001.

      [*If the one-letter amino acid symbols are mysterious, try the **Resources & Links** page on the course web site and click **Abbreviations***]
   b. Scroll down the ss120 genome sequence until you reach pro0007. Repeat **2a** for this gene, convincing yourself that at least the beginning of the gene sequence, when translated through the genetic code, in fact encodes p-Pro0007.

3. Why are genome sizes different?
   a. What is the length of the genome of *Prochlorococcus marinus* ss120? How about the genome of *Synechocystis* PCC 6803 (nicknamed S6803)? The LENGTH-OF function may prove useful, putting the name of the organism as the entity.

You can think of a genome as consisting of coding regions (the genes) and the sequences in between (the intergenic sequences).

b. Does S6803 have more genes than ss120?

c. Does S6803 have bigger genes than ss120?

d. Does S6803 have bigger intergenic sequences than ss120?

e. Summarize what you've found. Why are the genome sizes different?

f. Suppose in **2a** you chose to get the LENGTH-OF the SEQUENCE-OF each organism (if you didn’t do this, try it now!). How do you explain the results?

Here are two strategies to consider when you get a mysterious result:

1. *Execute complex functions bit by bit, from the inside out (in this case, executing first SEQUENCE-OF organism and then the entire function)*
2. *See what HELP for the function has to say. To reach help, mouse over the green action icon (the green wedge) and click Help. Or click a question mark (?) next to the name of the function on a menu.*

4. Is there a correlation between the number of codons that encode an amino acid and how common that amino acid is in proteins? If such a relationship exists, is it quantitative? In other words, are serine and leucine, with 6 codons apiece, 6-times more common than tryptophan and methionine, with only 1 codon apiece? Find out, using as a test the coding genes of the organism *Prochlorococcus marinus* ss120.

a. What information do you need to know in order to answer this question?

b. What kinds of functions do you need in order to gather that information?

Here are some functions that might be of use to you:

c. Investigate **COUNTS-OF** (STRINGS-SEQUENCE, STRING-ANALYSIS menu)

- Try getting the **COUNTS-OF** "G" (glycine) in the sequence of some protein.
- Replace "G" with `*amino-acids*` (which you can bring down from the DATA menu). Notice that the result now is a ***list*** of counts. Does any number in the list correspond to the first result you got (with "G")?
- What about the rest of the numbers? It might help to know what `*amino-acids*` means. To do this, execute just the box with `*amino-acids*` in it (by clicking **Execute** from the action menu). From the result, form a hypothesis as to what the numbers mean. ***Test*** that hypothesis.
- It might be easier on you if you could label each result with the name of the thing that was counted. You can! Try out the LABELED keyword.
- How would you describe what **COUNTS-OF** does?
- Another complication,... **COUNTS-OF** has a IN option and an IN-EACH option. What's the difference Try the following cases:

```
COUNTS-OF "G" IN "GTPGR"
COUNTS-OF "G" IN ("G" "T" "P" "G" "R")
COUNTS-OF "G" IN ("G" "GG" "GGG" "G")
```

Now try the same functions but using IN-EACH in place of IN (if its legal).

- Translate each of the six cases above into an English sentence.
- From the results, of the above experiment, describe the difference between the IN and IN-EACH options.

d. Investigate **CODONS-OF** (GENES-PROTEINS, TRANSLATION menu)
   - Type "G" (glycine) in the *amino-acid* box and execute the function.
   - Do the same, replacing “G” with `*amino-acids*`.
   - How would you describe what **CODONS-OF** does? Are you sure?

e. Investigate **PROTEINS-OF** (GENOME menu)
   - Put in your favorite organism as the entity and execute the function.
   - What is the result (in the Result pane)?
   - How would you describe what **PROTEINS-OF** does?

f. Combine the elements above with the needs you formulated in Steps **4a** and **4b** to determine number of codons for each amino acid and the number of amino acids in all proteins of ss120.

   You may well encounter a technical difficulty in calculating the amino acid counts for all proteins. There are many ways to solve this problem, but here are two tools you may find useful.

g. Investigate **JOIN** (STRINGS-SEQUENCE, STRING-PRODUCTION menu)
   - Try bringing `*amino-acids*` into the *list-or-string* box, delete the *item* box (using the red x inside that box), and execute the function. How do the results differ from what you get by executing the `*amino-acids*` box?
   - Try bringing **PROTEINS-OF** `ss120` into the *list-or-string* box of **JOIN** and execute the function.
   - Ooooh, that wasn’t what you wanted. What could you do to join what you want joined?

h. Investigate **SUM-OF** (ARITHMETIC, AGGREGATE-ARITHMETIC menu)
   - Type in your two favorite numbers into the *number* boxes and execute the function to make sure it does what you think it does.
   - Try the following:

```
SUM-OF (1 2 3) + 5
SUM-OF 5 + (1 2 3)
SUM-OF (1 2 3) + (3 4 5)
SUM-OF ((1 2 3) (3 4 5))
```
     [you'll have to X out the unused box]

   - Now something more interesting. Delete the contents of the input box of **SUM-OF** in the last example, drag/drop or copy/paste into the empty box the answer you got in **4f** to **COUNTS-OF CODONS-OF** `*amino-acids*`, and execute the function.

     [*You can copy something by mousing over the action icon of the box you want to copy and clicking copy. You can paste by mousing over the action icon of the empty target box and clicking paste.*]

   What does the answer mean?

- How can you apply **SUM-OF** to get the counts of all amino acids in all proteins of ss120?

Bottom line: Is there a correlation between the number of codons encoding an amino acid and the number of times that amino acid appears in proteins? You can squint at the numbers, but there is an easier way

j. Investigate **PLOT** (INPUT-OUTPUT menu)

- In the *list-or-table* box, type `(1 2 4 8)`, or something like that (note: Don't be tempted to include commas!), and execute the function. The X axis is a bit peculiar, but you shouldn't otherwise be too surprised by the resulting graph.
- Now try putting in the *list-or-table* box ***pairs*** of numbers, something like `((1 1)(2 4)(3 9))`
- Do you see that you'd be able to plot number of codons vs number of amino acids, if you could only get the two lists of numbers into the required format – pairs of numbers:

  ((*number of codons for* aa_1) (*total number of* aa_1 *in all proteins*)
  (*number of codons for* aa_2) (*total number of* aa_2 *in all proteins*)
  . . .
  (*number of codons for* aa_{20}) (*total number of* aa_{20} *in all proteins*))

  Well, you can! (see below)

k. Investigate **INTERLEAVE** (LISTS-TABLES, LIST-PRODUCTION menu)

- In the first *list* box, type (1 2 3 4), type (5 6 7 8) in the second, and execute the function.
- How would you describe what **INTERLEAVE** does?
- Clear the two *list* boxes and drag/drop or copy/paste the list of the number of codons in the first entry box and the list of the number of amino acids in the second. Execute it to make sure the function does what you hope it does
- Drag the **INTERLEAVE** box into the **PLOT** function and plot number of codons vs number of amino acids

l. Is there a correlation between the number of codons that encode an amino acid and how common that amino acid is in proteins? Is it quantitative?