Introduction to Bioinformatics
Problem Set 2: Molecular Biology Investigations

1. Display the first 2000 nucleotides of the ss120. Not very illuminating!

    e. Copy the display into a word processor and modify it by highlighting to make evident where are the genes in this region of the chromosome and where are the intergenic sequences. You may find helpful the FROM and TO keywords in the **GENES-OF** function. *Verify that the nucleotides you have highlighted are indeed the nucleotides of the gene* (which, of course, you can get using DISPLAY-SEQUENCE-OF for each gene).

    f. Do the same with the ss120 chromosome from 10000 to 12000. How might you account for the differences between these two regions?

2. Why are genome sizes different?

    e. What is the length of the genome of *Prochlorococcus marinus* ss120 (ss120)? How about the genome of *Synechocystis* PCC 6803 (S6803)?

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

    f. Does S6803 have more genes than ss120?

    g. Does S6803 have bigger genes than ss120?

    h. Does S6803 have bigger intergenic sequences than ss120?

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

3. Examine a codon table. Notice that two amino acids (methionine and tryptophan) are encoded by only one codon while three other amino acids (serine, arginine, and leucine) are encoded by six! 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, arginine, and leucine 6-times more common than tryptophan and methionine? Find out, using as a test the coding genes of the organism *Prochlorococcus marinus* ss120.

    Strategy

    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?

    It will be useful to test your methods on a single protein before ramping up to the entire set of proteins of ss120.

    c. DEFINE a variable to represent any arbitrary protein in ss120. You can get a list of proteins in the same way you've gotten a list of genes, except using PROTEINS-OF instead of GENES-OF.

    You've counted oligonucleotides in nucleotide sequences before, using COUNT-OF. Here's a short-cut variation. It will be simpler if you use COUNTS-OF (notice the plural).

d. Investigate **COUNTS-OF** (STRINGS-SEQUENCE, STRING-ANALYSIS menu)
   - Try getting the **COUNTS-OF** "L" in the sequence of your arbitrary protein.* Replace "L" with *amino-acids* (obtainable from the DATA menu). Notice that the result now is a ***list*** of counts. How many? Does any number in the list correspond to the first result you got (with "L")?
   - 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. From the result, form a hypothesis as to what the 20 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 LABEL keyword.
   - How would you describe what **COUNTS-OF** does?

e. Get the amino acid counts ***all*** proteins in ss120.
   - Remember that you can replace a single protein with a list of proteins and generate a list of results. You had a list of proteins a moment ago…
   - What do the results mean? ***Test*** that hypothesis.
   - If you used the LABEL keyword in the expression, delete that option and execute the expression again.

f. The result from the previous expression is way too much information. You don't want to know amino acid counts for each separate protein but rather the ***sum*** of all those counts, for each amino acid. The SUM-OF (ARITHMETIC, AGGREGATE-ARITHMETIC menu) function may be useful.
   - Copy and paste the result from the expression of **2.e**, and paste it into the first argument box of SUM-OF. Ignore the second argument box, or if it irritates you, delete it by clicking it's X icon. Execute the function.
   - What do the results mean? You may have an idea, but it's difficult to be sure, since you don't know what is the sum of those 1000's of numbers. You can test your hypothesis by using SUM-OF with a list of numbers you ***do*** know the answer for. Try copying the following `((1 2 3)(2 3 4)(3 4 5))`, pasting it into the argument box of SUM-OF, then executing the expression.

g. Unfortunately, you've lost the labels. It would be nice to get them back.
   - To put them back, use the INTERLEAVE function (LIST-TABLES, LIST-PRODUCTION menu), putting the result of SUM-OF as the first argument and putting *amino-acids* as the second. Execute the function.
   - What does INTERLEAVE do?

h. You could scan the list to see which amino acid has the highest counts, but that's what computers are for.
   - Use the SORT function (LIST-TABLES, LIST-PRODUCTION menu), and copy/paste the previous result into the argument box. Click the Options icon and select DESCENDING.

---

* If you're not familiar with the names or amino acids or their one-letter symbols, check http://www.people.vcu.edu/~elhaij/IntroBioinf/Links/GenCode.html

- Execute the function. What is the result?

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

4. By the end of the third section of ***What is a gene?*** you probably noticed what may have seemed like an anomalous number of certain nucleotides at certain positions upstream from the genes of *Synechococcus* PCC 7942. It may be that you've discovered a signal of biological importance. Or you might have been fooled by random fluctuation. After all, the differences aren't huge. Maybe they just arose by chance. How can you test whether the deviations observed in the table might have arisen by chance?

   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?

   c. Investigate **RANDOM-DNA-SEQUENCE**
   (STRINGS-SEQUENCE, STRING-PRODUCTION menu)
      - What do you get if you run the function directly? If you run it again?
      - Explore the LIKE keyword. Supply it with the value "GAGAGAGA" and run the function again.
      - How would you describe what **RANDOM-DNA-SEQUENCE** does?

   d. Provide **RANDOM-DNA-SEQUENCE** with the list of upstream sequences you generated in ***What is a gene?*** Run the function, define a variable that contains the resulting list. Then redo the steps you took in ***What is a gene?*** to create the table of nucleotide frequencies.

   e. Do you believe that the higher frequencies of certain nucleotides upstream from the coding genes of *Synechococcus* PCC 7942 are likely to have arisen by chance? How confident are you of your answer. ***Quantitatively***, how confident are you? If you're not prepared to give an answer containing a probability, consider what you would need to do so that you ***would*** be ready to do so.