Introduction to Bioinformatics
Problem Set 2: Molecular Biology Investigations

1. Display the first 2000 nucleotides of the ss120. Not very illuminating!
    a. 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.
    b. Do the same with the ss120 chromosome from 10000 to 12000. How might you account for the differences between these two regions?
    c. For one gene of your choice, place under the sequence of the gene (within the chromosome) the sequence of the protein it encodes. You can get that sequence in a variety of ways. The simplest way is to make use of the BioBIKE convention that a gene-name preceded by "p-" represents the corresponding protein.

2. Why are genome sizes different?
    a. 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).

    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?

3. Is there really 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 **SPLIT** (STRINGS-SEQUENCE, STRING-PRODUCTION menu)
        - Try **SPLIT**ting "123456789"
        - Try **SPLIT**ting "123456789" using the AT 2 option
        - You might even try plowing through the Help screen for **SPLIT**. You can find this by mousing over the green action triangle at the upper left of the **SPLIT** box, clicking on Help, and clicking on Full Documentation.

- Replace "123456789" with the ***sequence*** of a gene (e.g. pro0029). How can **SPLIT** help you extract the codons of a gene?
- How would you describe what **SPLIT** does?

d. Investigate **COUNTS-OF** (STRINGS-SEQUENCE, STRING-ANALYSIS menu)
- Try getting the **COUNTS-OF** "A" in the sequence of a gene.
- Replace "A" with *nucleotides*. Notice that the result now is a ***list*** of counts. Does any number in the list correspond to the first result you got (with "A")?
- What about the rest of the numbers? It might help to know what *nucleotides* means. To do this, execute just the box with *nucleotides* in it. From the result, form a hypothesis as to what the four 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. Investigate **ALL-DNA-SEQUENCES** (STRINGS-SEQUENCES, STRING-PRODUCTION menu)
- The function evidently calls for a number governed by the keyword LENGTH-OF. Give it a number and execute the resulting function.
- How would you describe what **ALL-DNA-SEQUENCES** does?

f. Combine the elements above with the needs you formulated in Steps ***a*** and ***b*** to determine codon usage in a single gene. You may choose to do this in multiple steps or to combine the elements into one humongous function. That's a matter of style. I would advise that at least at first you use the multiple-step approach so that you can see the results of each operation.

g. Replace the single gene with all coding genes of the organism ss120 and go through the same steps. Did it work? Of course it worked! The computer accomplished what you asked it to do. It always does (almost). But you may realize now that what you asked for is not exactly what you wanted.

h. Analyze the problem and formulate a plan. How might you intervene (in theory) in the sequence of events so that the result would be more to your liking?

i. Investigate **SIMPLIFY-LIST** (LIST-TABLES, LIST-PRODUCTION menu)
- Give this function the complicated result of Step ***g***. How do you interpret the product of the function?
- ***Test*** your hypothesis. You can do so readily by giving the function ***literal lists***. A literal list consists of a single quote (' = "Interpret what follows literally, without trying to evaluate it") followed by a list within parentheses. For example `'(1 2 a bc)` is a literal list. Make a complex literal list (lists within a list) and feed it to **SIMPLIFY-LIST**.
- How would you describe what **SIMPLIFY-LIST** does?

j. Add **SIMPLIFY-LIST** to your procedure (or insert it into your humongous complex function) to achieve a count of the codons in the coding genes of ss120.

k. **SORT** the results to make them easier to compare to the codon table. Find **SORT** on the LIST-TABLES, LIST -PRODUCTION menu, and supply it with the result of

SIMPLIFY-LIST. Note that since the list is not simple (there are lists within lists), you need to specify by which position you wish to sort. Choose BY-POSITION from the options, and if you wish to sort by codon, enter 1 (the first position). If you wish to sort by count, enter 2.

l. 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.