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? ***Quantitatively***, how much does each possible source (e.g. bigger genes) contribute to the difference?

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). Executing SEQUENCE-OF directly has a different effect than executing it within another function. To simulate its execution within LENGTH-OF, select the RESULT-ON option of SEQUENCE-OF.*
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.*

Probability

4. Basic probability:

4a. Consider a random gene from Anabaena variabilis, which has a GC-fraction of about 41%. Which of the following is more likely and why?

A. The gene is preceded by "GGGGGG"

B. The gene is preceded by "GGGGGG" and also begins with "ATG"

4b. Which of these two sequences is more likely to occur in a random genome with a GC-fraction of 30%? Why?

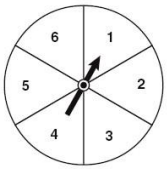
A. AAATTTCCCGGG

B. TCAAGGCTGCTA

4c. Which of the following is more likely and why?

A. You roll three dice and get one that is a 1, 2, or 3

B. You spin a spinner and get a 1, 2, or 3



4d. At the time I'm writing this, the record of the VCU basketball team is 16 wins and 4 losses. From this, presume that their probability of winning against a random opponent is 16/20. There are 11 games left in the regular season. If they are against random opponents, then what is the probability that VCU will win all the rest of their games?

5. Fuller et al (1984) considered the DnaA-binding sequence (TTAT[CA]CA[CA]A) and its role in the 300-bp region associated with the regulation of DNA synthesis in *E. coli.*

   a. What is the probability of finding this sequence in 9 specific nucleotides?

   b. What is the probability of finding an three instances of this sequence in a ***random*** 300-bp region of DNA with the characteristics of the E. coli genome?

6. Devise an algorithm (in theory – no need to implement it in BioBIKE or anywhere else) that will take a given piece of DNA and determine the 3-nt sequence that occurs within it at the greatest frequency. Use as few fancy functions as you can manage. The algorithm should work more efficiently than the one described by Compeau and Pevzner (see Section II of Computational detection of origins of replication). Would it still work well if you were looking for the most frequent 20-mer?

Molecular biology, genome analysis, probability... all at once

At the end of Computational detection of origins of replication, you took Compeau and Pevzner's reservations seriously and searched for TTAT[CA]CA[CA]A sites throughout the *E. coli* genome. Let's go further with this, to see whether such a search could find biologically interesting regions of the genome without prior knowledge of their existence.

7a. Generate a list of all matches to the DnaA-binding sequence in the *E. coli* genome, searching both strands (this is SQ30 of Computational detection of origins of replication). How many hits are there (COUNT-OF should be helpful), and how does this number compare to the theoretically expected number? (you might want to make use of your answer to Question 5a).

The output you get may not burst with meaning. The first line should be:

```
4537208   4537198   TTATACAAA   B
```

The first number is the starting point of the match and the second number is the ending point... note that the first number is larger than the second! Evidently the match goes backwards. It must have been found on the complementary strand. The "B" at the end confirms this interpretation.

7b. Bring up the genome sequence of *E. coli* and go to the coordinates of the first hit discussed above. What sequence is shown between those coordinates?

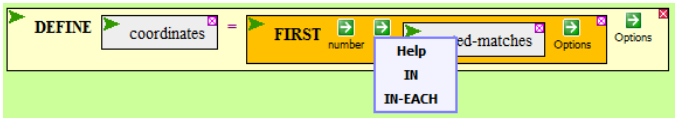
7c. Sort the hits by first coordinate by using BioBIKE's SORT function. You can see the results more conveniently by putting the SORT function in DISPLAY-LIST. Define a variable that contains this sorted list.

There are lots of purported DnaA-binding sequences. What makes the origin of replication special is not that it has such a sequence but that it has many, clustered together. Try looking for clusters. This is possible to do by hand, but not easy. Better to make a tool.

7d. Devise an algorithm (again in theory) that would use as input the results you obtained in Question 7c and would produce a list of ***distances*** between adjacent DnaA-binding sequences.

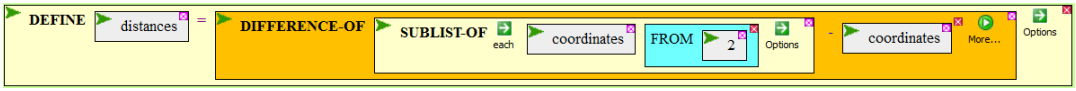
7e. You almost are not familiar with the BioBIKE functions that would enable you to implement your algorithm, so I'll provide a trick:

(i) DEFINE a variable (say coordinates) as the FIRST elements of the result you obtained in 7c (executed without DISPLAY-LIST). Your form should look something like this:



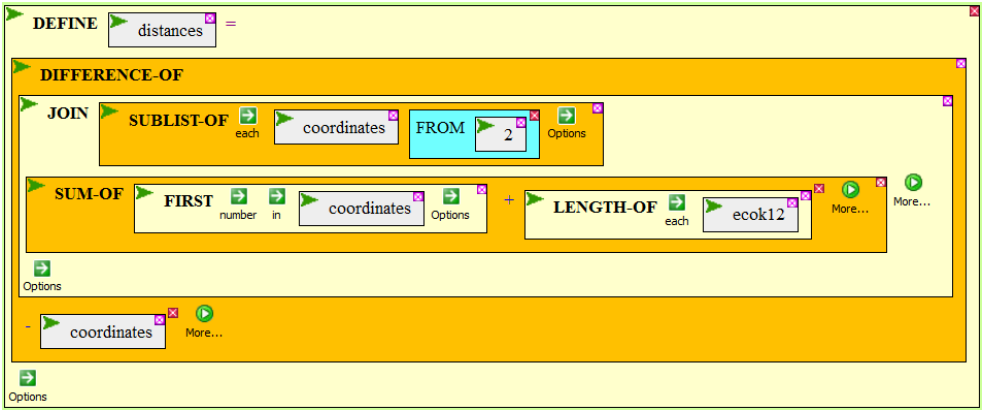
Note that mousing over the **in** icon allows you to specify that you want the first item IN-EACH of the lines of the sorted matches. Execute this function to obtain a list of sorted beginning coordinates.

(ii) My strategy is to DEFINE a variable that is the DIFFERENCE-OF each coordinate from the one before. Here's how:



Execute the function. What do you get?

(iii) Of course you get an error,* because you're subtracting a list that starts at the second position from the same list, starting from the first position. The second list is one item smaller than the second, and BioBIKE doesn't know what to do with the last element. To fix this, add an appropriate extra coordinate. Since the *E. coli* genome is circular, the extra coordinate will be the FIRST coordinate. The LENGTH-OF the genome is added to this to avoid a negative distance. In short, do this:



Execute this and check the first couple of distances by hand, using your result from Question 7c. Correct?*

(iv) The last new function in this trick is INTERLEAVE. Play with it a bit – bring it into the workspace and type (1 2 3 4) in the first *list* box and type (5 6 7 8) in the second. Then execute the function. Do you understand what it does?

* If you don't see why, try setting `coordinates` to (1 2 4 7) and execute each part of the function.

Now delete those test values and put in the first box the variable containing distances and put in the second box the variable containing the full sorted matches. Display a sorted list of the result as before, something like this:

7f. Now, finally, we're ready to look at biological implications. The list you obtained in Question 7e should have only a few lines that indicate a close spacing of DnaA-binding sites. What are their coordinates?

7g. Where in the genome do these coordinates lie? Is there anything biologically interesting about these locations? Bring up the genome of *E. coli*, and go to the coordinates. Is any of the regions familiar? In the case of new regions, does the coordinate lie inside of a gene or outside? If outside, what gene is it upstream from? What significance, if any, does that gene have? If you're unfamiliar with the protein encoded by the gene, find out something about it.

7h. Comment on the effectiveness of looking for clusters of DnaA-binding sites in *E. coli* as a means of finding interesting regions of the genome. What does it do well? What might be its limitations?