Introduction to Bioinformatics
Problem Set 3: Genome Sequencing

1. Assemble a sequence with your bare hands! You are trying to determine the DNA sequence of a very (very) small plasmids, which you estimate by gel electrophoresis to be about 200 nt. You have made a shotgun library of the miniplasmid and have generated reads of about 20 nt each (you must be using a very early technology!). Your objective now is to assemble those reads into the full sequence.

   a. Take a look at the sequences you will assemble. Within BioBIKE, click on the FILES menu, then click on Files. Click on the Shared-files subdirectory and then locate and click on a file called `mini-plasmid-reads.txt`. This file is in what is called FastA format, each read consisting of a one-line label preceded by ">" and then the DNA sequence. Approximately how many reads are there? Write down the label and the sequence of the first read.

   b. Return to BioBIKE and load the sequences you will assemble. Bring down **READ** (used to read from files) from the INPUT-OUTPUT menu.
      - The file-name is "mini-plasmid-reads.txt"
      - It is in the SHARED subdirectory (choose the SHARED flag from Options)
      - It is in FastA format (chose the FASTA flag from Options)
      - Execute, and don't be alarmed by the format of the result.

      Can you find the label and the sequence of the first read?

   c. **DEFINE** a variable (perhaps something like `reads`) as the contents of the file you read. Either cut and paste the **READ** function into the value hole of **DEFINE**, or fill that hole with **PREVIOUS-RESULT**.

   d. Make a first-pass assembly of the reads, using **ALIGNMENT-OF** to find overlaps (you can locate the function on the STRING/SEQUENCE menu, Bioinformatics Tools submenu).

   e. ALIGNMENT-OF was not designed to assemble reads, and it isn't very good at it. Copy the results into your favorite word processor, and continue the assembly by hand. Warning! It's easy to just sit and stare blankly at the sequences. If you find yourself making little progress, step back and ask yourself what kinds of overlaps are you trying to find. Then devise a systematic approach towards finding them, making use of the search capabilities of the word processor.

   f. Did you get a complete plasmid sequence from these reads? Probably not. Why so many separate pieces?

   g. Investigate **INVERSION-OF** (found in the STRING-SEQUENCES menu, String-production submenu. Bring down the function, and click on Help (in the green action arrow menu). Then click on Full Documentation. From the examples, do you understand what **INVERSION-OF** does? Try it out. Put in a sequence (in quotes of course) into the argument hole, predict what it should produce, and see if you're right.

   h. Why would **INVERSION-OF** be useful in analyzing reads? Which strand of a genome being sequenced gets read by sequencing reactions?

i. Use **INVERSION-OF** to produce the opposite strands of all your reads. Then align as before the **JOIN**ed reads and inverted reads. Copy the alignment into a word processor and join together as many reads as possible. Of course you should speed up the process by using the knowledge gained from step **1.E**.

j. How many contigs do you get now, and how do you interpret them?

k. What fraction of the plasmid have you covered with your assembled reads? Use in the calculation the total number of nucleotides in your contigs and orphan reads.

l. In what ways was the process you went through to assemble the reads similar to the process used to assemble the *Drosophila* genome. In what ways did the latter process differ from yours?

2. Reconsider Problem 1, looking at the data as a whole.

   a. How many reads are there? (You might use **COUNT-OF** on the variable you defined in Problem 1.C)

   b. How many nucleotides are there in the reads? (You might get a **SUM-OF** the **LENGTHS-OF** the reads)

   c. What is the average read length?

   d. What is the calculated ***coverage*** of the mini-plasmid?

   e. What fraction of the plasmid do you expect is represented by the reads? This is a very common type of question in bioinformatics but not at all easy to answer the first time you encounter it. So let me break it down.

      i. The fraction of the plasmid you expect is represented by the reads is equal to one minus what?

      ii. The fraction of the plasmid you expect is NOT represented by the reads is equal to the probability that a specific nucleotide is not found in any of the reads. This may be the most difficult connection to make, so let's dwell on this a bit. If the probability is 50% that the nucleotide at coordinate 29 is not found in any read and (since there's nothing special about coordinate 29) 50% that ***any*** specific nucleotide is not found in any read, then on average, half of the nucleotides will be represented by the reads and half won't be. Draw pictures, visualize, but don't just accept the words. Get the idea into your head as a picture.

      iii. The probability that the nucleotide at coordinate 29 is not found in any read may be calculated from the probability that it isn't found in the first read AND that it isn't found in the second read AND … all the way to the last read. How do you combine these probabilities? Are the reads independent of one another? If I told you that the nucleotide is not found in the first read, would you know any more than before as to whether it is found in the second read? If not, then they're independent. How do you combine independent probabilities to calculate a ***joint*** probability (i.e., the probability that all the events occur)?

iv. What is the probability that coordinate 29 is not found in a particular read? Consider the read to have the length of an average read. What fraction of the entire miniplasmid is taken up by an average read? If I threw a dart at the plasmid, what is the probability that I would hit a nucleotide within an average read?

v. If you were able to arrive at a number for 2.E.iv, then you can use that number to work backwards through steps iii → i and get the desired probability.

3. Find genes with your bare hands!

a. Use **READING-FRAMES-OF** (in the GENES-PROTEINS menu, Translation submenu) to display the translation of the first 1500 nucleotides of the ss120 genome. Why are there six lines labeled "translation-frames"?

b. Notice that there is a DNA sequence on the top line. What is it?

c. Notice that there is a DNA sequence on the fourth line. What is that?

d. What are the letters on the second line? How frequently do these letters occur relative to the DNA sequence? Why?

e. What is the significance of the first letter, K, of the second line? How does it relate specifically to the letters of the DNA sequence? If you have a hypothesis, test it.

f. What is the significance of the first letter, K, of the third line and the first letter, S, of the fourth line? How do they relate to the letters of the DNA sequence?

g. What about the first letters, L, F, and A, of the fifth through seventh lines? If you have an idea, be sure to test it.

h. Notice that there are some asterisks on the lines. What do they mean, and how do they relate to the DNA sequences?

i. Print out the results of **READING-FRAMES-OF** (or copy it into a word processor). With a highlighter (or highlighting within the word processor. Highlight every segment in a translation frame between asterisks. Use different colors for each line. I've started you off below:

```
           Sequence     1 AAAGCTAGAT GGCAGAAAGG TTTTTGAATA ATTTCCACAG ATTCCACAAG
Translation-Frame-1     1 K  A  R  W   Q  K  G   F  *  I   I  S  T  D   S  T  R
Translation-Frame-2     1  K  L  D   G  R  K  V   F  E  *   F  P  Q  I   P  Q  D
Translation-Frame-3     1   S  *  M   A  E  R   F  L  N  N   F  H  R   F  H  K
         Complement     1 TTTCGATCTA CCGTCTTTCC AAAAACTTAT TAAAGGTGTC TAAGGTGTTC
Translation-Frame-4     1    L  *  I   A  S  L   N  K  F   L  K  W  L   N  W  L
Translation-Frame-5     1   F  S  S   P  L  F   T  K  S  Y   N  G  C   I  G  C
Translation-Frame-6     1     A  L   H  C  F  P   K  Q  I   I  E  V   S  E  V  L

           Sequence    51 ACCTACTACT ACTGTATTAA TTTCATATAA TTAAATTAGA ATTACTAGAA
Translation-Frame-1    51  P  T  T   T  V  L  I   S  Y  N   *  I  R   I  T  R  R
Translation-Frame-2    51   L  L  L   L  Y  *   F  H  I  I   K  L  E   L  L  E
Translation-Frame-3    51 T  Y  Y  Y   C  I  N   F  I  *   L  N  *  N   Y  *  K
         Complement    51 TGGATGATGA TGACATAATT AAAGTATATT AATTTAATCT TAATGATCTT
Translation-Frame-4    51  V  *  *   *  Q  I  L   K  M  Y   N  F  *   F  *  *  F
Translation-Frame-5    51 S  R  S  S   S  Y  *   N  *  I   I  L  N  S   N  S  S
Translation-Frame-6    51   G  V  V   V  T  N   I  E  Y  L   *  I  L   I  V  L
```

j. What do you notice about the colored lines? What does this signify? Check your hypothesis through BioBIKE.

4. Find genes with someone else's bare hands!

   a. Get the same sequence you used in Problem 3, with **DISPLAY-SEQUENCE-OF**. Choose the FastA option, so that the coordinates don't appear.

   b. Copy the sequence (with or without the FastA header, it doesn't matter) and go to the following web site: *http://exon.gatech.edu/GeneMark/* GeneMark is a *Gene*-finding program developed by *Mark* Borodovsky. In brief, it analyzes sequences it's pretty sure are real genes, extracts features from these sequences and then asks whether anything in the sequences you supply match the set of extracted features. It also looks for start and stop codons.

      i. Click on GeneMark.hmm in the gray bar to the right

      ii. Click on GeneMark.hmm for prokaryotes

      iii. Paste your sequence into the big white box

      iv. In the species box, choose Prochlorococcus CCMP1375 (i.e. ss120)

      v. At the bottom, click boxes for Generate PDF graphics (screen), Translate predicted genes into proteins, and Sequences of predicted genes

      vi. Click Start GeneMark.hmm

   c. Examine the section that appears, entitled Gene Prediction in Text Format. Did GeneMark consider the correct sequence? Check this. Is the length correct? Is the G+C content correct? The latter means the percent of G in the sequence plus the percent of C. You know how to do this (see Problem 1), but a simple way is to use the **GC-FRACTION** function (STRINGS-SEQUENCES menu, String-analysis submenu).

   d. How many genes did GeneMark identify? Do they correspond to your idea (from Problem 3) of where they are?

   e. Check the first gene found by comparing the nucleotide or protein sequence predicted by GeneMark with the sequence you get from BioBIKE.

   f. Scroll back to the top of the page and click on View PDF Graphical Output. Scroll to the second page. How many lines of graphical output do you get? Why that number? Ignore the dotted red lines and focus on the solid black lines. The height of the line corresponds to the prediction by GeneMark that the region at that point has features typical of ss120 genes. Where are such regions? Why do they occur on lines 1 and 3?

   g. This method is a lot faster than highlighting reading frames, no?