Introduction to Bioinformatics
Problem Set 4: Genome Assembly, Statistics

Genome assembly and gene identification

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 CyanoBIKE, click on the **FILES** menu, then on **Shared-files**, 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" (be sure to include quotation marks)
 - 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. Bring down a **DEFINE** box from the DEFINITION menu, type `reads` into the *var* box (and close the box, of course), and then choose one of the following methods, any of which should work:

i. Drag the entire **READ** function you just executed into the *value* box of **DEFINE**. Then execute the **DEFINE** box.

ii. OR... Drag the result from Q1.b into the *value* box of **DEFINE**. Then execute the **DEFINE** box.

iii. OR... Copy the entire **READ** function, using the Copy function on the Action Menu (green wedge) to the upper left of **READ** and paste it into the *value* box of **DEFINE,** using the Paste function on the Action Menu of the *value* box).

iv. OR... Copy the box containing the result obtained from the **READ** function (in the Result Pane), using the Copy function on the Action Menu (green wedge) of the result and paste it into the *value* box of **DEFINE**.

v. OR... Click the *value* box of **DEFINE**, type *, press Enter, and execute the **DEFINE** box. The asterisk (*) represents the last result in the Result Pane.

vi. OR... Click the *value* box of **DEFINE**, and bring down the function **PREVIOUS-RESULT** from either the ALL menu or OTHER-COMMANDS menu.

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

e. Those spaces every 10 nucleotides make it easier to read large sequences, but they will be a nuisance here. Get rid of them by selecting the GROUP-LENGTH option in the **ALIGNMENT-OF** function, and enter a large number for the group-length (or enter 0 if you like). Then reexecute the function.

f. 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.

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

h. 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.

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

j. 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**.

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

l. 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.

m. 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 that you expect is represented by the reads is equal to 100% minus what? (just the obvious... no need for deep thought here)

   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?

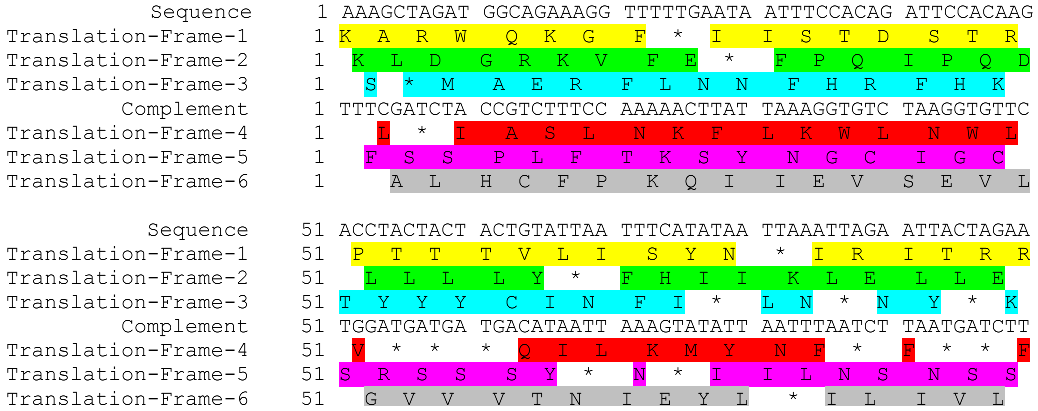
e. What is the significance of the first letter, K, of the second line? How does it relate specifically to the letters of the DNA sequence? If you have a hypothesis, test it.

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.

Was Mendel Right? and DEFINE-FUNCTION

4. What is the code of your function to simulate Mendel's experiment, probably named MENDELS-EXPERIMENT? Use the **Show-code** option on the action menu for the function to generate the code in text format.

5. Use your function and some trial-and-error to find the number closest to Mendel's actual number of white flowers that you would NOT consider close enough to the number predicted by a 3:1 ratio. Define a number as "not close enough" if that number or numbers even more deviant from expectation appears in fewer than 5% of trials of MENDELS-EXPERIMENT.

6. Find the same number as in Problem 5 using a chi-squared table.

7. Define a function that accepts a number and returns its square in the Results Pane (***not*** displays it in a popup window).

8. In which of the following cases would a **chi-squared test** be useful? To answer this question, consider whether you could answer the question by means of a repeat-experiment-many-times simulation of the kind you did in MENDELS-EXPERIMENT.

8a. Are the genes of *Anabaena* PCC 7120 significantly longer, on average, than the genes of ss120?

8b. Your unidentified viral sequence has dinucleotide counts of {AA = 60, AC = 72, AG = 52, …}. Is there good reason to believe that the fragment could have been derived from the virus Mx8?

8c. Is level of transcription of the gene encoding melanin induced by ultraviolet radiation? I've measured the expression of the gene 12 times: 6 times with UV and 6 times without.

8d. Are genes in *Synechococystis* PCC 6803 (S6803) that are annotated as “hypothetical” biased towards small genes?