

CS196-1 HW5

Due: Thursday, March 13th 2:30pm

Reading

- *How To Lie with Statistics*, Chapter 5
- *An Introduction to Bioinformatics Algorithms*, Section 6.10: Multiple Alignment
- *Whole-genome shotgun assembly and comparison of human genome assemblies*
<http://www.pnas.org/cgi/reprint/101/7/1916>
- *FASTA file format*
<http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>

Overlaps (35)

We are working toward building our own set of tools for genome analysis and assembly. This problem is our first clear step towards assembly. We have talked about the importance of being able to find overlaps between two sequences. Write a program that, given a set of sequences, calculates the maximum overlap between each pair of sequences. We have provided, at <http://www.cs.brown.edu/courses/cs196-1/hw5.html>, a file which contains the result of breaking one long sequence into several pieces with overlap.

Your program should account for 3 special cases:

- DNA Inversions that reverse the orientation of the DNA sequence
- Overlaps that are completely contained in each other.
- Allow for an error rate threshold for each overlap

Your program should have two inputs: the name of FASTA file containing the overlaps and the rate of missense mutations.

Use your program to combine these pieces to find the original sequence.

Dot Plot (25)

In Homework 2 we introduced Dot Plots as a powerful means of visualizing sequences. In this homework, we apply Dot Plot techniques to start working towards understanding the more complex sequences and corresponding 3-dimensional structure present in proteins.

1. Modify your dot plot program to render the dot plot as a graph, not just an ASCII display, if it does not already. If your chosen programming language is not amenable to graphical display, try generating output which can be used by gnuplot (<http://www.gnuplot.info/>) to draw a plot. Send your program to cs196-1tas@cs.brown.edu
2. Run the program for the first 100 amino acids of two homologs (of your choosing) of BRCA1 available at <http://www.cs.brown.edu/courses/cs196-1/hw5.html> and show the dot plot.
3. Amino acids within the same family share some of the same properties, and may behave in similar ways. Extend your program so that it generates dot plots to compare amino acid sequences by family. Specifically, replace each amino acid in the input string with its family and draw dots when the families match. Assign each family a distinct color and draw each dot in its corresponding family's color. Run this program on the first 100 amino acids of the same two sequences and show the dot plot.

Finding Known Genes (25)

At <http://www.cs.brown.edu/courses/cs196-1/hw5.html> we have provided two sequences for you: one contains a simulated bacterial plasmid, and the other contains a list of resistance factors (genes which give bacteria the ability to protect themselves against antibiotics).

Your job is to find out which resistance factors are present in the plasmid. Use local alignment to find all the present factors through a single run of the alignment algorithm on the plasmid and a concatenation of the resistance factors.

1. How many factors are in the plasmid?
2. Which ones are they?
3. How confident are you that you found them all?

Explain your methods, and include any visualizations that you used.

Time Complexity (15)

Prove or otherwise demonstrate the big-O time complexity for the following algorithms:

- Smith-Waterman local alignment
- Needleman-Wunsch global alignment
- Global alignment with an affine gap penalty

You must show work.