Computer Science 2300: Lab 6

Due: May 2, 2012

In this lab you will implement an algorithm for sequence alignment.

1 Edit Distance

Write a program that finds the "distance" between two different nucleotide (genetic) sequences. The distance is defined as the edit distance, as described in class and the DPV textbook, with the gap penalty and mismatch penalty both equal to 1. You may want to test your code on some small sequences, but it should work on all pairs of the sequences (formatted as plain text) provided to you. Compute the edit distance between each pair of these sequences, leaving out e.coli. Also compute the distance between e. coli and salmonella. Report the normalized edit distance per nucleotide (the edit distance divided by the sum of the lengths of the two sequences) and time taken to run your code on each pair. Which pair is most similar by this measure? You may be interested in looking up the relationships among the pairs afterwards (try Wikipedia).
Note: Be careful about your code's space and time usage. Inefficient solutions may work on smaller problems, but not scale to cases like e. coli and salmonella. Your code has to eventually work on these cases!

2 Extra Credit

For two points of extra credit (i.e. you get 5/3 on this lab, and it carries over to the rest of your grade), also store and print the actual best alignment (hint: Problem 6.24 in DPV may be useful).

3 Final Results

To receive full credit, show the TA the pairwise normalized edit distances as well as the amount of time the algorithm took to run on each of the pairs. To receive two points of extra credit, show your TA the best alignment between salmonella and e. coli (don't print this out, just show it to the TA!).
Note: We do not provide you with any support code for this assignment.