CSE 5290: Algorithms for Bioinformatics
Assignment 4 (Released Nov 17, 2009)
Submission deadline: 1 pm, Dec 1, 2009

1. The assignment can be handwritten or typed. It MUST be legible.
2. You must do this assignment individually.

Question 1

Implement keyword trees (as described in the book) in C/C++/Java. For extra credit integrate your code in R (it is enough to be able to invoke your code from inside R). This is likely easier if you use C/C++ but there is a new project called rJava that will help if you wish to use Java.
Download the complete genome of Mycobacterium tuberculosis (accession number NC_000962) and Streptococcus mutans (accession number NC_004350). Use your keyword tree implementation to build keyword trees of both organisms for all words of soze $l = 20$. Note that you should build one keyword tree not two, and that each leaf node in the keyword tree should record the number of times n_1, n_2 that the word corresponding to the node occurs in each organism (respectively). List 50 words that have the maximum values of $n_1 - n_2$ and 50 words that have the maximum values of $n_2 - n_1$.
Submit your code and these results.

Question 2

Clustering microarray data: use the dataset given the course directory. Using the `cluster` package in R, generate hierarchical top-down and bottom-up clustering of the data using Euclidean distance. Using the output generate 8-10 clusters in each dataset.
Generate a confusion matrix of the two cluster sets produced. Plot a fluctuation plot for this matrix using the `ggfluctuation` method in the `ggplot2` package in R.
Submit both your code and the fluctuation plot.