Bioinformatics Computing
COSC494/594
Homework #3

Reading assignment:

Read Chapter 4 and Durbin handout (review this week's material)
Read Chapter 2 and Glimmer paper (review for this next week's material)
Read Chapter 5 and 6 (final material for midterm)

Reminder: midterm on Thurs, 2/27 but we'll discuss

Homework problems: (due 2/20)

1. Implement the dishonest casino Hidden Markov Model (HMM) similar to the one described on page 54 of Durbin. Assume the casino is always fair at the start (i.e., at t=0, Pr(F) = 1 & Pr(L) = 0), and shifts to another state with probability 0.05. Generate a random sequence of 300 rolls based on this model. Please submit the code and instructions how to run in your drop box; a particular instance of 300 rolls in your submission is optional (8 points).

2. Download the two particular instances available from the course website.

3. Determine the probability of the benchmark files given the dishonest casino model using an implementation of the Forward algorithm. Please report the probabilities in your report and include your source code in your submission with instructions how to run it (10 points).

4. Determine the most likely state sequence of the benchmark files given the dishonest casino model of #2 using an implementation of the Viterbi algorithm. Please save the result as files "viterbi.1.txt" and "viterbi.2.txt" and label the states "F" and "L" as used in Figure 3.5 in Durbin. Please also include these result files and the source code in your submission (10 points).

5. Download two Anthrax strains from GenBank, the gold standard "Ames ancestor" that is virulent (NC_007530) and the non-virulent lab strain "Ames" (NC_003997).

6. Visit the GLIMMer website (http://ccb.jhu.edu/software/glimmer/index.shtml). Read the release notes, information, and download the latest version.

7. Run Glimmer3 on both genomes and include the output gene calls in your submission (4 pts each). Please do not include the genomes as they are large, but hold on to them wherever you are doing your work; these specific genomes will be the starting point for homework #4 after the midterm.