

Bioinformatics Computing
CSE40532/60532
Homework #3

Reading assignment:

Read Chapter 4 and Durbin handout (review of last week's material)
Read Chapter 2 and Glimmer paper (review for this week's material)
Read Chapter 5 and 6 (week of 9/25)

Reminder: midterm on Thurs, 10/5

Homework problems: (due 9/28)

1. Implement the dishonest casino Hidden Markov Model (HMM) described on page 54 of Durbin. Assume the casino is always fair at the start (i.e., at $t=0$, $\text{Pr}(F) = 1$ & $\text{Pr}(L) = 0$), and shifts to another state with probability 0.05 as described. 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 the drop box 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 submit your source code in the drop box 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 submit these result files and the source code in your drop boxes (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://www.cbcb.umd.edu/software/glimmer/>). Read the release notes, information, and download the latest version.
7. Run Glimmer3 on both genomes and place the output genes in your dropbox (4 pts each). This and these specific genomes will be the starting point for homework #4 that you will work on after the midterm.