Project #3 – Hidden Markov Models (HMMs)

Biological Computation   
Spring 2017  
Due: May 2nd

For this project there are two general options:

Option 1) Test the limits of the Viterbi algorithm’s ability to determine which state in an HMM generated a given symbol using a known HMM.

Option 2) Use the Baum-Welch algorithm to find the parameters of HMMs of know structure.

For Option 1 there are several parameters that effect how well the Viterbi algorithm works: the transmission probabilities (between states in the HMM), the emission probabilities, the number of states in the HMM, and the length of the sequence generated. For example, if the transmission probabilities are close to 0.5 the HMM will change between states too rapidly for the changes to be detected. If the emission probabilities between the states are too similar (e.g. if the “loaded” die is only very slightly loaded) it is very difficult to distinguish between states. More states (especially if they loop back on themselves) can it difficult to determine which state the HMM. If the sequence being analyzed is very short there may not be enough data to determine which state a given symbol was generated from.

Because our Viterbi code is written for HMMs with two states for this project option it will be easiest to focus on the transmission and emission probabilities. If you want to analyze HMMs with more than two states the Viterbi code will need to be expanded to handle additional states (which is not too difficult to do for a limited number of states). Pick at least two sets of transmission and two sets of emission probabilities. Use those to create at least four HMMs. For each HMM generate sample sequence data sets and for each data set use the Viterbi algorithm and measure how well the algorithm determines state. Note that unless the sequences are fairly short our current Viterbi algorithm will suffer from underflow errors and so will need to be converted to using log probabilities.

For Option 2 the idea is similar, create at least 4 different HMMs with different emission and transition probabilities. For each HMM create a set of sequences. Then use those sequences and the Baum-Welch (or another optimization) algorithm to try to recreate the probabilities in the HMMs.

## Groups

You may work in groups of two or three on this project, which may be helpful in balancing the coding and analysis portions of the project.

## Write-up

The write-up should be about 3 pages long. Include an abstract, a description of the HMMs, including the emission and transmission probabilities, the data and analysis of the data, and a discussion/conclusion. Include the code of both the HMM and the analysis algorithm as an appendix.