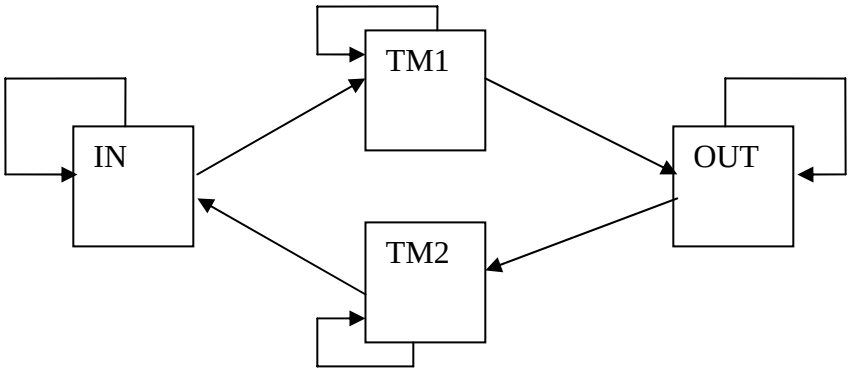
CISC 841 Bioinformatics
Spring 2006
Homework 2

Handed out: March 9, 2006
Due date: April 4, 2006

Hidden Markov Models.

You are about to build a hidden Markov model to predict the topology of transmembrane proteins. A transmembrane protein consists of three types of regions: the transmembrane domain (TM), the inside loop, and the outside loop. The transmembrane domains spanning the membrane are about 20 amino acids long, and tend to have more hydrophobic amino acids. The inside loops tend to have more positively charged residues. A transmembrane protein may contain multiple TM domains, going across the membrane inside and outside multiple times. The topology refers to the correct locations of these regions.

A hidden Markov model with a simple architecture depicted as follows is designed to capture the topological features of transmembrane proteins.



1. Implement the model training procedure for annotated data, which can downloaded from http://www.cis.udel.edu/~lliao/cis841s06/hw2_data. When you do counting for emission frequencies in TM domains, there is no need to differentiate TM1 (going outside) and TM2 (going inside), namely these two states will have the same emission frequencies for the 20 amino acids.

2. Implement the Viterbi algorithm. Make prediction on the sequence data and compare with the annotation. The performance is evaluated for both the location prediction and topology prediction.

3. Implement the model training procedure for non annotated data using Viterbi training.

a. Randomly initialize model parameters
b. Run Viterbi algorithm on each training sequence.
c. Update model parameters using the training procedure from part 1, treating the sequences as annotated.
d. Repeat step b and c until a stopping criterion is met. The stopping criterion suggested here is to check if the predicted TM domains change from the previous iteration. For example, when the average percentage of overlap between the predicted TM domains from the current iteration and the predicted TM domains from the previous domains is higher than a threshold, say 90%, then the iteration stops.
e. Evaluate the performance in the same way as in part 2 and compare the performance.

For more information, please read the paper by Kahsay, Gao & Liao and the references therein.

Reference:

Kahsay, Gao & Liao (2005). "An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes", Bioinformatics, vol. 21, pp. 1853-1858.