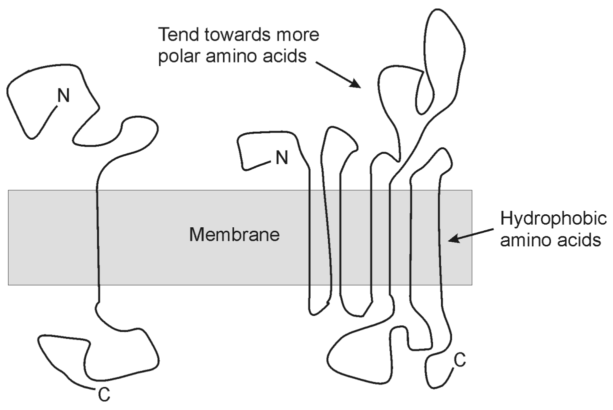
Problem Set #2 BCH339N Systems Biol/Bioinformatics Marcotte due 3/8/2016

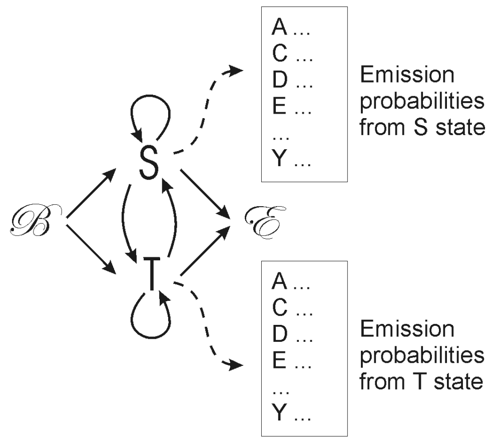
A Hidden Markov Model for finding transmembrane protein sequences

1. Proteins in a cell are often synthesized by a ribosome at one location in a cell (for example, in the cytosol or attached to the endoplasmic reticulum), then shipped someplace else in the cell where the proteins actually do their work. About a fourth to a third of the proteins get attached to membranes, often with a portion of the protein on one side of a membrane, a portion penetrating the membrane, and a portion on the other side of the membrane:



In this problem, we'll train a hidden Markov model to detect the regions of these so-called "integral membrane proteins" that end up buried in the membrane. The reason we can do this is because the amino acids in these regions tend to be much greasier in character than the amino acids in the non-membrane-associated part of the protein.

We'll construct a very simple hidden Markov model of the following architecture:


A ...
C ...
D ...
E ...
...
Y ...
Emission probabilities from S state
S
T
ℬ
ℰ
A ...
C ...
D ...
E ...
...
Y ...
Emission probabilities from T state

where ℬ is the beginning state, S is the hidden state for an amino acid in a water soluble segment of the protein, T is the hidden state for an amino acid in a transmembrane segment, and ℰ is the end state. In reality, we won't model the end state, so you can essentially ignore it for this exercise. So, the protein sequence on the left in the example above would be something like SSSSSSSSSSSSSSSTTTTTTTSSSSSSSSSSSS and the protein sequence on the right would have seven runs of T's, flanked and interspersed with runs of S's.

From the course web page, download 3 data sets:
soluble_sequences, transmembrane_sequences, and state_sequences
These files have, respectively, a set of amino acid sequences from soluble segments of proteins, a set of amino acid sequences from trans-membrane regions of proteins, and a set of state sequences for all of the known integral membrane proteins of yeast.

Calculate the frequencies of amino acids in the first two files (remember your programs from the first problem set?). Use these frequencies to generate the emission probability tables of the S and T states of the HMM.

Calculate the frequencies of beginning a sequence in each of the 2 states in the third file. Use these to calculate the transition probabilities from the beginning state of the HMM.

Calculate the number of occurrences of each *digram* in the state file. (e.g., SS or TT or ST or TS. A digram is a 2 character string. In general, a string of *n* characters is an *n*-gram.) You can use your other program from problem set#1 for this. From the numbers of observations, you should be able to construct the transition probability matrix for the HMM.

Now, take the amino acid sequence `KNSKIIFFFFLIII` and calculate the most likely state sequence using the Viterbi algorithm.

Turn in the transition probability matrix, emission probability matrix, the Viterbi matrix, and the most likely state sequence.

2. How many transmembrane amino acid segments (*i.e.*, segments of all T's) does the average integral membrane protein have? Show your calculation.

3. Draw a reasonable alternate architecture for the transmembrane sequence predicting HMM. Why might you prefer one of the two HMM architectures over the other?

4. Draw a reasonable architecture for an HMM that would identify protein secondary structures within protein sequences. If you wanted to take into account the fact that some particular mixtures of secondary structures were more common than others (for example, there are many proteins which have only beta sheets, many which have only alpha helices, many which are mixed alpha/beta, and some which aren't so easily categorized), how would you modify your model to incorporate this data? Draw the modified model as well.

5. Suggest an application (ANY application!) of HMM's to a biological problem other than the applications we have discussed (*i.e.*, not gene finding, CpG islands, protein secondary structure or transmembrane segment prediction, sequence alignment). Suggest a possible HMM architecture that could be applied to that problem.