Case study 2: Secondary structure prediction

Protein secondary structure prediction is a classical problem in bioinformatics. Here we focus on membrane proteins. Given that transmembrane stretches in these proteins are enriched with hydrophobic amino acids, the goal is to predict transmembrane segments. Construct an HMM for such a prediction starting from the sequence and known transmembrane regions of a well-characterized protein SecY from *E. coli*, see Akiyama & Ito (1987) in *EMBO J 6*: 3465–3470; see figure below). All amino acids are categorized into three classes: charged (C), neutral (N) and hydrophobic (H). The protein sequence (consequently abbreviated with C, H and H) together with the indicator for transmembrane regions (1) is given in the file "Sequence_case2.txt". Formulate a first-order HMM that can be used to predict transmembrane states based on the encoded amino acid sequence.

(1) Use occurrence counts to compute estimates for model parameters.
(2) Implement the Viterbi algorithm and use it to determine the prediction accuracy of your model (in terms of numbers of correctly and incorrectly annotated amino acids).
(3) Given the constructed HMM calculate the conditional probabilities (display the result in matrix-form: P (C I C), P(C I N), P(C I H), … P(H I H).
(4) Change the emission probabilities to radically different values. Is it easier or more difficult to annotate the hidden states with certainty? How could you change the emission probabilities in order to reduce the HMM to a simple Markov chain model? Define the resulting model.

https://en.wikipedia.org/wiki/Protein_domain

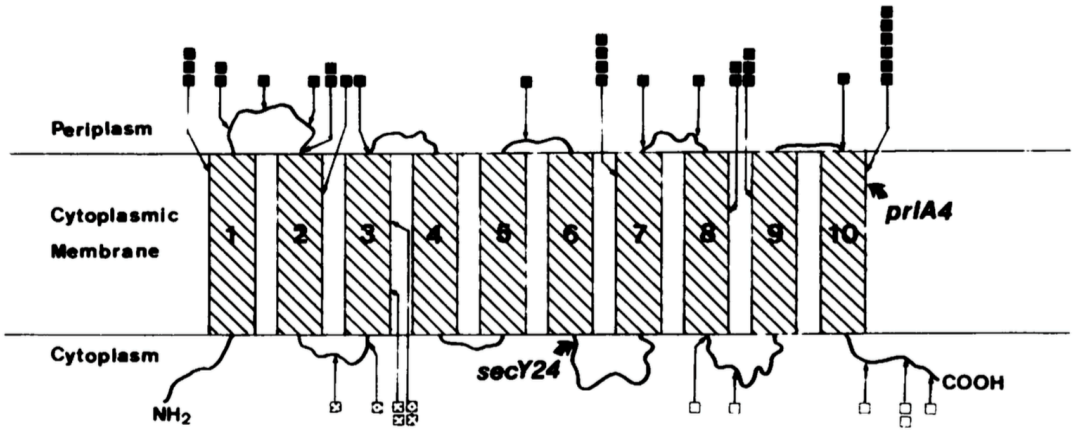

Fig. 6. A model for the orientation of the SecY protein in the membrane. Transmembrane segments are represented by hatched boxes. Filled squares with arrows indicate sites at which a highly active SecY−PhoA hybrid protein was generated, whereas dotted and open squares indicate those of intermediate and low enzyme activity respectively. Also shown are the locations of the amino acid alterations by the *secY24* (Shiba *et al.*, 1984) and the *prlA4* (Stader *et al.*, 1986) mutations.