An improved hidden Markov model for transmembrane topology prediction

Robel Kahsay[1], Li Liao[1,2,3], and Guang Gao[1]

1. Delaware Biotechnology Institute, Newark, DE 19715
2. Department of Computer and Information Sciences
University of Delaware, Newark, DE 19716

Abstract

In this work, we proposed a hidden Markov model for transmembrane protein sequences. The architecture of the model, based on an existing model TMHMM by Sonnhammer et al, contains 7 types of states including helix core, helix caps, loops on both the cytoplasmic side and non-cytoplasmic side, and a globular domain state embedded in the middle of loops. The model differs from TMHMM by how to treat the loops on the both sides, and by use of Dirichlet priors. Using Maximum Likelihood, the model was trained and cross-validated on a set of 160 sequences with known topology. The prediction accuracy for membrane domain location and topology are %89 and %84 respectively, both surpassing significantly these of the best existing model TMHMM (%83 and %77).

1 Introduction

Accurate prediction of transmembrane topology in integral membrane proteins has been an active subject in bioinformatics and remains as a challenge. Membrane proteins, as gateway between the cell and its environment, transport specific metabolites, drugs, ions, and also act as receptors for a variety of signaling molecules such as hormones, nucleotides, and odorants. On average, about 25% proteome of an organism are membrane proteins. Because of their functional roles, membrane proteins are important targets of pharmacological agents, and identification and classification of membrane proteins are essential. Correct prediction of the transmembrane helical topology can help identify binding sites, and may provide insights to functional inferences for different parts of the sequences.

Most computational approaches rely on the compositional bias of amino acids at different regions of the sequence. For example, there is a high propensity of hydrophobic residues in transmembrane alpha helices due to the hydrophobic environment in lipid membranes. Because such bias is quite noticeable and consistent, the location of transmembrane domains can often be easily identified, with high accuracy, even by a simple method such as applying a threshold on the hydrophobic propensity curve. Another compositional signal is the abundance of positively charged residues in the segments that are located on the cytoplasmic side of the membrane, and therefore is referred to as "the positive inside rule" for predicting the orientation of a transmembrane helix [11, 12, 13]. Unlike the hydrophobic signal for transmembrane helices, the "positive inside rule" is a weaker signal and often confused by significant presence of positively charged residues in globular domains of the protein on the non-cytoplasmic side. As a result, it is a more difficult task to correctly predict a protein's overall topology – all transmembrane segments and their orientation.

There are basically two ways for improving a method's prediction accuracy: by enhancing

[3]To whom the correspondance should be made: lliao@cis.udel.edu.

the signal/noise ratio for those weak signals or by identifying new signals and associating them with the target to be predicted.

For example, significant improvements of prediction accuracy were reported [7] by applying multiple sequence alignment to proteins with similar topology so that the positive residue content in the inside loops may become obvious as patterns (motifs) emerging in the aligned columns. It shall be noted, however, that multiple sequence alignment may not always be suitable either due to insufficient number of homologues or the length variation of these inside loops. Other methods have been attempted at exploring more subtle signals, e.g., correlation of compositional biasses at different positions. The best perfermance attained so far is by using artificial neural networks [8], a method known for its capability of capturing complex nonlinear signals. Despite its improvement at prediction accuracy, the neural network, notorious for its black-box property, provides fewer insights to those signals that the network is designed to capture.

A hidden Markov model, TMHMM, has recently been used for transmembrane topology prediction [10, 5]. Hidden Markov models, as a probabilistic framework, have been widely applied in computational biology, with remarkable success [2, 3, 4, 6]. Unlike artificial neural networks, the architecture of an HMM corresponds closely to the biological entities being simulated by the model. The model consists a set of states, each corresponding to a region in the protein sequence being modelled. Each state has an associated probability distribution over the 20 amino acids (called emission frequency) charactering the compositional bias (e.g., hydrophobicity and positive charge) of amino acids in the corresponding region. The architecture of the model specifies how these states are connected to one another, so the transitions from state to state reflect how biologically the different regions are assembled to form the entire protein. For example, the "grammar" that requires inside loops and outside loops to alternate each other can be incorporated in the architecture. Much of the learning and predictive power of HMMs lies in the stochastic transitions between states, i.e., there is a transition probability associated with each transition. The transition probabilities, along with emission frequencies, enable the models to capture correlations among signals. For instance, a transmembrane helix that normally would be missed due to its poor hydrophobicity by a method using a fixed threshold may still be picked by the model if the surrounding topogenic signals strongly support it. Although the performance of TMHMM still cannot compete with the neural network method, it compares favorably with most of the previous methods. More importantly, the success of TMHMM demonstrates the potential of HMMs as a powerful framework to transmembrane topology prediction.

In this work, we proposed an improved hidden Markov model that gives better performance, and may also offer new insights into the mechanisms involved with translocations of short and long loops across the membrane. The model differs from TMHMM in both model architecture and model training procedure. The architectural modification is on the inside and outside loop submodels. Our model parameter estimation is a single step Maximum Likelihood estimation with the inclusion of prior knowledge. Using a test set of 160 sequences with known topology, a set which was used to report the performance of TMHMM, our model's prediction accuracy for membrane domain location and topology are 89% and 84% respectively, both surpassing significantly those of the best existing model TMHMM (83% and 77%).

2 Method

In this section, we give detailed descriptions of the model architecture, training and priors, and the dateset.

Model

The architecture of the model is illustrated in Figures 1 to 5. The overall skeleton has kept that of TMHMM, which reflects the three-component basic structure of transmembrane protein sequences: transmembrane helix, inside and outside loops. Because of the existence of different amino acid frequencies within a transmembrane region, a transmembrane helix is split into three parts: the helix core, and two caps on both ends of the helix core. For similar reasons, a loop's first and last 10 residues are explicitly modelled, i.e., each residue corresponds to an individual state in the model. All other residues in the middle of a loop are collectively represented by one "globular" state, which has a transition back to itself thus can repeat as many times as the loop length dictates. Since long loops outside the membrane appear to have different properties than short loops, two separate chains of states are introduced, as depicted in Figure 1, to represent long loops and short loops respectively.

The transmembrane helix is modelled, exactly as in TMHMM, by two cap regions of 5 residues each, surrounding a core region of variable length 5-25 residues. Therefore the total length for helices varies from 15 to 35 residues, covering the actual size range observed for transmembrane domains. The flanking cap regions – one for the cytoplasmic side and one for the non-cytoplasmic side – have their own amino acid distributions. The state diagram for the helix core is shown in Figure 5. Although the model contains two sets of transmembrane states to model paths going inwards and outwards, all their parameters are mirrored and tied to each other, i.e., they are estimated collectively.

The architecture here differs from that of TMHMM mainly by how these loops are actually modelled (Figures 2-4). For inside loops and short outside loops, the first 10 residues are critical and contain most of the topogenic signals embodied as distinct usage of amino acids at each individual residues and correlation among different residues. Because those loops may have less than 20 residues, a ladder-like configuration is adopted to allow for early exit from loop states. Take an inside loop sub model (Figure 2) for example, in this case, each of the first 9 LI states has three transitions out: to the next LI state, to the UI state on the top and to the preceeding UI state. The last LI state has only two outgoing transitions. However, when the length of a loop is greater than 20, the first 10 states entering the loop shall all have been traversed to reach the globular state. In other words, transition to the next LI state in this second case is certain. Therefore, a separate chain of 10 TI states, in parallel to the chain of 10 LI states, is introduced in the architecture to differentiate the two cases. In TMHMM, these two cases are accommodated by just one chain of LI states. As transitions between states reflect correlation between residues, we expect the refined treatment here to enhance the performance of the model. Our model also treats the long outside loops differently; each state in the long loop has only one transition in and one transition out, except for the globular state, which has two ins and two outs. Because the length of long loops is larger than 100 residues, the above configuration makes intuitive sense; there is no need to have the "ladder-like" transition to bypass the globular state. This reduces the number of free transition parameters from 83 (that of TMHMM) to 62. The total number of free parameters for our model, including 7 x 19 emission parameters from seven states, is 195 whereas TMHMM has 216 free parameters.

Training and Priors

Model parameters are estimated from observed frequencies with significant regularisation. This single step maximum likelihood (ML) estimation[2] is feasible because the state path is known in training sequences, as shown in Figure 2. For example, the emission frequency of state j is

$$e_j(a) = \frac{c_{ja}}{\sum_{a'} c_{ja'}}$$

where c_{ja} is the number of times that residue "a" is seen at state j in training sequences. One advantage of such single step ML estitmation is its speed and simplicity; unlike the Baum-Welch training used in TMHMM, this does not require iterative reestimation.

However, ML estimation is susceptible to overfitting when there are insufficient training data. For instance, if residue "a" is not ever seen in state j in the training sequences, then $e_j(a)$ will be estimated as zero, which may not reflect the true frequency, because residue "a" may be observed in state j as more training sequences become available. A widely used approach to address the issue of overfitting is the use of regularisers derived from prior information. One approach we adopted to come up with such regularisers is to use substitution matrix mixtures(see [3]) as given by

$$\beta_{ja} = A \sum_b f_{jb} P(a|b) \tag{2}$$

where β_{ja} and f_{jb} are pseudocount and observed frequency for amino acid a in state j, $P(a|b)$ is conditional probability of amino acid a given amino acid b (derived from BLOSUM50 matrix), and A is a constant. This type of priors is used only for emission frequency estimations.

Another regularization scheme we adopted is the widely used regularization scheme based on Dirichlet mixiture priors. In our case, for each segment in the labeled training sequences, an observation count vector $\vec{c} = (c_1, \cdots, c_{20})$ over 20 amino acids is found. These observed count vectors are clustered into seven groups according to the type of segment (referred to as G, X, I, M, C, S, or L. See Figures 1-5) from which they were calculated. For each of these groups, a single component Dirichlet prior vector is found. That is, we have Dirichlet vectors $\vec{\alpha^j}$ where $j = G$, X, I, M, C, S, L. We also produce a single component Dirichlet vector to regularise transtions in the "ladder-like" sub models. The lower chain of states for these "ladder-like" submodels (Figures 2 and 3) have three outgoing transitions each. If we take each of these three transition counts as vectors, we have 20 observation vectors from which we can derive a single component Dirichlet vector.

Each $\vec{\alpha^j}$ vector is normalized and then multiplied by a constant A to be used as pseudocount vector.

$$\sigma_{ja} = A \frac{\alpha_a^j}{\sum_{a'} \alpha_{a'}^j} \tag{3}$$

where σ_{ja} is pseudocount for amino acid a in state j. The final emission frequency after being regularized with priors is

$$e_j(a) = \frac{c_{ja} + \sigma_{ja} + \beta_{ja}}{\sum_{a'} (c_{ja'} + \sigma_{ja'} + \beta_{ja'})} \tag{4}$$

The constant A is the same for both β and σ. For emission frequencies, the constant A is set at a value of 30000 when the priors are used for estimations at state G, and at a value of 3000 for other states. For transitions, only σ pseudocounts are added as transition priors to the raw estimation, and A is set at a value of 1.

Once all transition probabilities and emission frequencies are fixed according to the above procedure, the model is ready to be used for predicting a sequence's topology. The prediction corresponds to an annotation of the sequence using the states in the model, in other words, corresponds to a state path – each residue in the sequence is marked with the state where it is emitted. The predictions are made by using standard Viterbi algorithm [3], which gives the most probable state path.

Dataset

The dateset used in this work was downloaded from the TMHMM website, and it contains 160 proteins. The topology of most proteins in this dataset is determined experimentally. There are 108 multi-spanning and 52 single-spanning proteins.

We adopted the same 10-fold cross validation in Sonnhammer *et al* [10]. The dataset is divided into 10 subsets. Each subset contains about 16 sequences mostly similar to one another. Precisely,

sequences from different subsets are no more than 25% identical to each other. The model is trained on nine subsets and then is used to make predictions on the remaining set. This is repeated for 10 times, and each time a different subset is selected as testing set. The prediction accuracy is the average over the 10 runs.

The training sequences are labeled three states: M for transmembrane regions, I for inside loops, and O for outside loops. In order to use these labeled sequences for training our model, we need to first translate their 3-state labeling to a 7-state labeling, which corresponds to states in the model. The translation is very straightforward. For example, given a "M" region (which shall be 15 residue long or longer), we simply replace the first 5 labels on the inside side from "M" to "C", and last 5 labels to "X". The details are given in Figures 6 and 7. It is noted that this translation is preprocessing of the training data, used in both this model and TMHMM, and should by no means be misunderstood as the relabeling used in TMHMM, which is part of their training method.

3 Results

The accuracy performance of the model is measured by the number of sequences (and as the percentage of 160 sequences in the dateset) whose topology and locations of transmembrane helices are correctly predicted. The performance is also measured by the sensitivity and specificity for identifying single individual transmembrane domains. Following the same criterion in TMHMM, a predicted helix is counted as correct if if overlaps by at lease 5 residues with a true helix.

In Tabel 1, the performance of the model with different variations on architecture and use of priors is listed. For comparison, the results of TMHMM from Sonnhammer et al 1998 are included. The model depicted in Figure 1 along with use of combined priors from both Dirichlet and Substitute matrix has achieved the best performance at both topology (84%) and locations (89%). The improvement over that of TMHMM (77% topology and 84% locations) is significant.

To help us understand how the architecture and use of different priors in this method contribute to its best performance, three variations of architecture and three variations of priors, as defined in the caption of Tabel 1, are tested. Several observations can be made from the performance of these variations. First, we noticed that Dirichlet prior is consistently more effective than substitution matrix mixiture based prior for all three different architectures. Combining Dirichlet and substitute matrix mixiture based priors enhanced the model performance, but not always; indeed performance was even decreased in some cases. In the contrast, we noticed that model-3 attained best performance among the three architectures in all three variations of priors suggesting that the model architecture played a decisive role for better performance. Another observation is that model-2, which has two symmetric loops on each side of the membrane, has better performance than model-1 of the original TMHMM architecture, probably due to the refined treatment of each individual loops. However, model-3, which has two alternative loop paths on the non-cytoplasmic side, has achieved the best performance. This observation further validates the hypothesis made in TMHMM that differentiation of short and long loops only applies to the non-cytoplamic side.

4 Discussion

We presented a hidden Markov model that significantly outperformed the existing model for predicting transmembrane topology in single sequences. The improved accuracy of the model is attributed to the refined treatments of inside and outside loops, and the use of combined priors. Based on the experiments it is shown that the model architecture plays a more crucial role than the use of priors. Because the model architecture closely corresponds to the topology of transmembrane proteins, a

Model	Prior	Correct topologies	Correct locations	Single TM sensitivity	Single TM specificity
model-1	(a)	117 (73.1%)	128 (80%)	97.4%	97.0%
	(b)	92 (57.5%)	103 (64.4%)	77.4%	80.8%
	(c)	117 (73.1%)	126 (78.8%)	96.1%	96.7%
model-2	(a)	120 (75.0%)	132 (82.5%)	98.4%	97.2%
	(b)	97 (60.6%)	121 (75.6%)	97.7%	95.6%
	(c)	118 (73.8%)	135 (84.4%)	98.4%	97.2%
model-3	(a)	120 (75.0%)	133 (83.1%)	97.8%	97.6%
	(b)	110 (68.8%)	124 (77.5%)	94.5%	98.1%
	(c)	**135 (84.4%)**	**143 (89.4%)**	**98.3%**	**98.1%**
TMHMM		123 (76.9%)	134 (83.8%)	97.1%	97.7%

Table 1: model-1: Original TMHMM; model-2: Symetric two loops in each side; model-3: Our new architecture. (a) Single component Dirichilet prior; (b) Substitution matrix based prior; (c) Both.

good performance of the model can serve as validation of the hypotheses that are made in the model. In our case, the architecture's special treatment of loops at different length scales may suggest that the mechanisms of translocating a loop are more sensitive to the length of the loop than what is previously understood.

Our model was trained by using maximum likelihood estimator, which is simple, single-run, and very fast. In contrast, the previous model TMHMM adopted a complex three-stage learning procedure that involved Baum-Welch EM iterations, use of relabeling, and "discriminative" training. It is reasonable to believe that higher accuracy may be obtained by some sophisticated training, which can be the future work in this line. Another possible improvement is to adopt a mixture of Dirichlet priors instead of a single component one. As an application, there are plans to use the method for predicting integral transmembrane proteins in complete genomes.

Acknowledgement

This publication was made possible by NIH Grant Number P20 RR-15588 from the COBRE Program of the National Center for Research Resources, and by a DuPont Science & Engineering grant.

References

[1] M. Brown, R. Hughey, A. Krogh, I.S. Mian, K. Sjolander and D. Haussler, "Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families", *Proc. First International Conference on Intelligent Systems for Molecular Biology*, Washington D.C., July, 1993.

[2] G.A. Churchill, "Hidden Markov chains and the analysis of genome structure", *Computers and Chemistry*, **16**, pp. 107-115.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, "Biological sequence analysis: probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.

[4] S.R. Eddy, "Hidden Markov models", *Current Opinion in Structural Biology*, **6**, pp. 361-365.

[5] A. Krogh, B. Larsson, G. von Heijne, and E. Sonnhammer, "Prediction Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes", Journal of Molecular Biology, **305**, pp. 567-580.

[6] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling", *Journal of Molecular Biology*, **235**, pp. 1501-1531.

[7] B. Persson, and P. Argos, "Prediction of transmembrane protein topology utilizing multiple sequence alignments", Journal of Protein Chem. **16**, pp. 453-457.

[8] B. Rost, P. Fariselli, and R. Casadio, "Topology prediction for helical transmembrane proteins at 86% accuracy", Protein Sci. **5**, pp. 1704-1718.

[9] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, D. Haussler, (1996) "Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology," *Comput Appli Biosci* 1996 **12**, pp. 327-45.

[10] E. Sonnhammer, G. von Heijne, and A. Krogh, "A hidden Markov model for predicting transmembrane helices in protein sequences", Proceedings of ISMB 6, 1998, pp 175-182.

[11] G. von Heijne, "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology", *EMBO Journal*, **5**, pp. 3021-3027.

[12] G. von Heijne, "Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule." *Journal of Molecular Biology*, **225**, pp. 487-494.

[13] G. von Heijne, "Membrane proteins: from sequence to structure", *Annu. Rev. Biophys. Biomol. Struct.*, **23**, pp. 167-192.

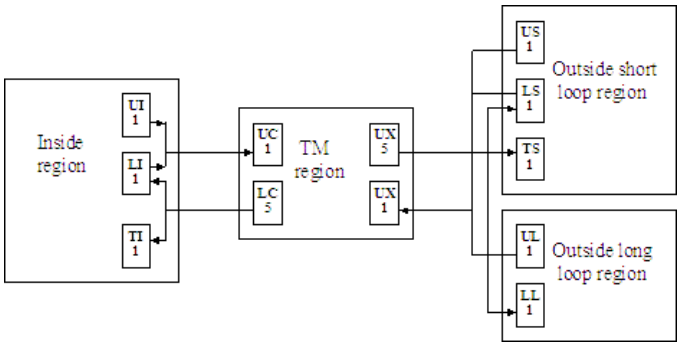

Figure 1: Interconnection between the sub-models.

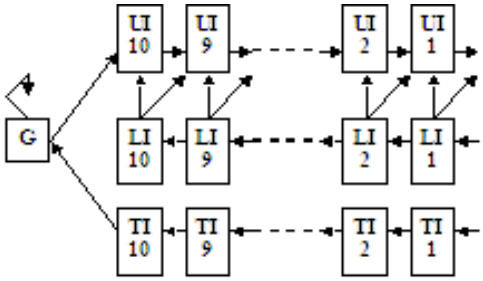

Figure 2: Inside loop sub-model. If loop length is less than 20 the path through LI states is traversed. Otherwise the path is through the TI states.

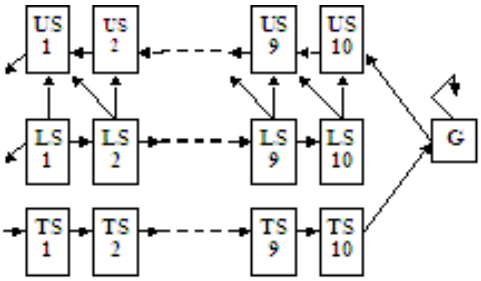

Figure 3: Outside short loop sub-model. If loop length is less than 20 the path through LS states is traversed. Otherwise the path is through the TS states.

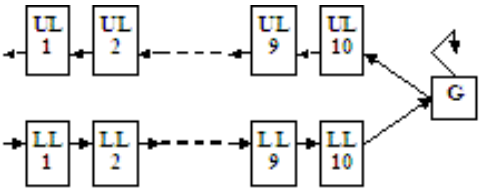

Figure 4: Outside long loop sub-model. Loops with length greater than 100 are deemed as long and modeled by this sub-model. In all these LL or UL states, there is only one transition in the forward direction.

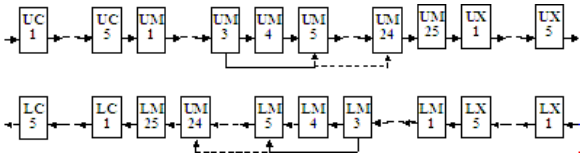

Figure 5: Transmembrane region sub-model. The two chain of states represent paths in each direction. There are 5 LX (lower external cap) states and 5 UX (upper external cap) states. Similarly, in the cytoplasmic side, there are 5 LC (lower cytoplasmic cap) states and 5 UC (upper cytoplasmic cap) states. The helical core is represented by 25 states and length distribution is captured in terms of transition probabilities from the 3rd state.

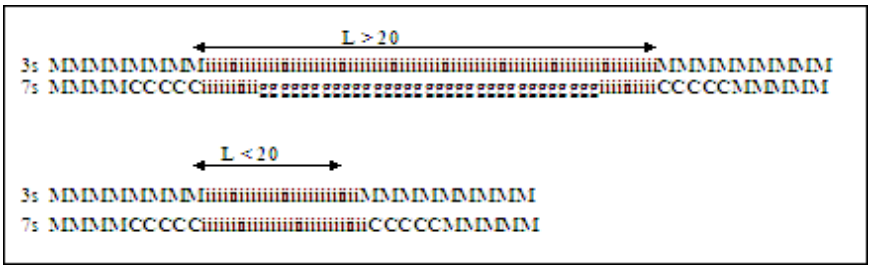

Figure 6: Translation of the 3-state labeling of outside loops into 7-state labeling according to the architecture. Loops with length greater than 100 are translated per the submodel for long loops whereas those with length less than 100 are translated per the submodel for short loops. For short loops, if the length is less than 20, the label of the first 10 residues is translated into the LS states, otherwise is translated into the TS states, per the architecture shown in Figure 3.

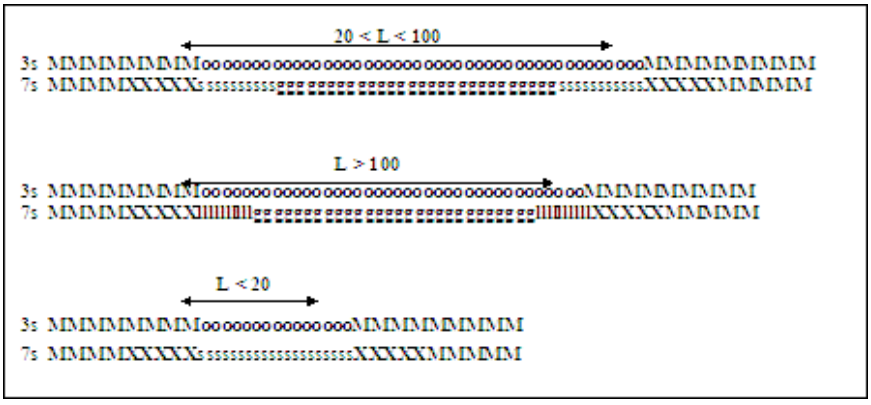

Figure 7: Translation of the 3-state labeling of inside loop into 7-state labeling according to the architecture. If the length of a loop is less than 20, the label of the first 10 residues is translated into the LI states, otherwise is translated into the TI states, per the architecture shown in Figure 2.