Discriminating Transmembrane Proteins From Signal Peptides Using SVM-Fisher Approach

Robel Y. Kahsay
Delaware Biotechnology Inst.
University of Delaware
Newark, DE 19715, USA

Guang R. Gao
Delaware Biotechnology Inst.
University of Delaware
Newark, DE 19715, USA

Li Liao *
Dept. of Computer & Information Sciences, University of Delaware
Newark, DE 19716, USA

Abstract

Most computational methods for transmembrane protein topology prediction rely on compositional bias of amino acids to locate those hydrophobic domains in transmembrane proteins. Because signal peptides also contain hydrophobic segments, these computational prediction methods often misidentify signal peptides as transmembrane proteins. Here, we present a new approach that combines the SVM-Fisher discrimination method and TMMOD – a hidden Markov model based predictor for transmembrane proteins. While TMMOD alone has already outperformed most existing methods in both identification and topology prediction, this new approach further improves the ability of TMMOD to discriminate between transmembrane proteins and signal peptide containing proteins, reducing mis-prediction of signal peptides by more than 30% in our test data.

1. Introduction

Membrane proteins have diverse functional roles in cellular activity. In transport mechanism, they are active mediators between the cell and its environment or the interior of an organelle and the cytosol. As enzymes, they catalyze specific metabolites and ions across membrane barriers, convert the energy of sunlight into chemical and electrical energy and couple the flow of electrons to the synthesis of ATP. Furthermore, they serve as signal receptors and transduce signals such as neurotransmitters, growth factors and hormones across the membrane. On average, about 25% of the proteome of an organism are membrane proteins [1, 3]. Because of their vast functional roles, membrane proteins are important targets of pharmacological agents.

Unfortunately, membrane proteins are hard to solubilize and purify in their native conformation because of their hydrophobic nature. Thus, a very small number of them have experimentally determined structure and topology. This has motivated various computational methods for identification and topology prediction of membrane proteins. Most of these computational approaches rely on the compositional bias of amino acids at different regions of the sequence.

Recently, we have developed a hidden Markov model based predictor (TMMOD) that has a more accurate treatment of the loops structure at both cytoplasmic and non-cytoplasmic sides of the membrane and utilizes a Bayesian based optimization for effective training [1]. The performance of TMMOD is shown to be superior to other existing reference methods. For topology prediction, TMMOD has a success rate at 89%, which is around 10% higher than that of TMHMM[5, 3] – a state-of-the-art transmembrane protein predictor. For identification, TMMOD has generally less false positives in comparison to TMHMM, especially in discriminating signal peptides from transmembrane proteins. However, though to a lesser extent when compared to TMHMM, TMMOD has an inherent problem in discriminating transmembrane proteins from signal peptides.

In this work, we propose to combine the SVM-Fisher discrimination approach with TMMOD to further improve its ability to identify integral membrane proteins from proteins containing signal peptide. Using the SVM-Fisher discrimination method, we are able to reduce mis-prediction of signal peptides by more than 30% in a widely used dataset. Moreover, we present results for topology prediction accuracy of TMMOD using a newly compiled data in comparison with other reference methods including Phobius[2], the most recent successor of TMHMM.

2. The TMMOD predictor

The architecture of the four submodels of TMMOD is illustrated in Figure 1. The overall skeleton has kept that of TMHMM, which reflects the three-component basic structure of transmembrane protein sequences: transmembrane helix, cytoplasmic and non-cytoplasmic loops. The transmembrane region is modeled with two cap regions of 5 residues each surrounding a core region of variable length 5-25 residues (Figure 1A). Therefore, the total length for

*Corresponding author: lliao@cis.udel.edu

helices varies from 15 to 35 residues, covering the actual size range observed for transmembrane domains. Although the submodel contains two chains of transmembrane states to model paths going inwards and outwards, all their parameters are estimated collectively.

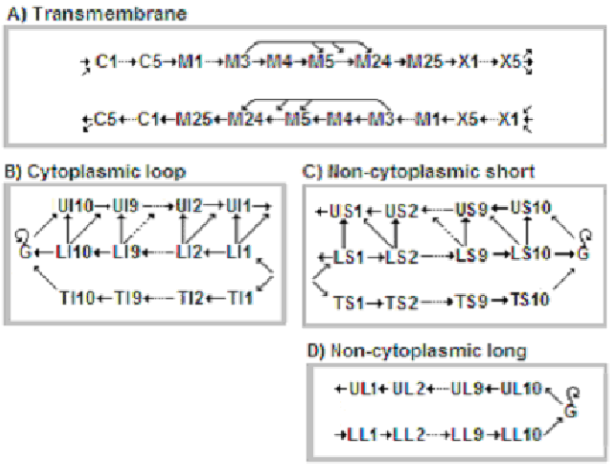

Figure 1: **Architecture of the four submodels of TMMOD. The outgoing and incoming arrows from each submodel show the interconnection between submodels.**

The architecture of TMMOD differs from that of TMHMM mainly by how the loops are modeled (Figure 1B, 1C and 1D). The length distribution of the cytoplasmic and non-cytoplasmic loops or the spacing between transmembrane domains in the training sequences has a long tail. To learn this distribution effectively, we modified the ladder-like cytoplasmic and short non-cytoplasmic loop submodels of TMHMM by introducing additional chain of states (T-states) that bypass the ladder arrangement. The idea behind this modification is that, 90% of the sample contains loops with lengths less than 40 amino acids with only 10% representing all other lengths. For this reason, we want the transition parameters of the ladder-like submodel to explicitly model the length distribution of those loops that are less than 40 amino acids long while longer loops are directed through the bypass.

Model parameters were estimated using Bayesian approach (PME) with single component Dirichilet and substitution matrix mixture based regularizers [6]. The protein topology was predicted from the state labeling of the sequence residues using the Viterbi algorithm (1-best). For discrimination, the measure *exp-no-aa* [3] with a threshold was used to determine whether a sequence is a transmembrane protein or not. The *exp-no-aa* measure is basically the sum of posterior probabilities of the states in the transmembrane submodel for sequence positions in a predicted TM helix.

3. The SVM-Fisher discrimination

A generative probability model such as a profile hidden Markov model assigns a likelihood score to any given sequence. The model is supposed to assign higher likelihood score to the set of sequences it is trying to model. The likelihood score for a sequence x is computed using the standard forward-backward algorithm. In addition to the generative likelihood score, the forward-backward algorithm gives sufficient statistics for all parameters of the model. The sufficient statistics for a parameter tells how the parameter was involved when scoring the sequence. Therefore, the dependence of the likelihood score on each model parameter can be systematically represented by taking gradients of the likelihood score with respect to each parameter. These gradients are components of the so called Fisher vector which is given as,

$$\vec{U_x} = \nabla_\theta \ logP(x|\theta) \tag{1}$$

In the SVM-Fisher discrimination method [4], sequences are mapped into such Fisher vectors which are then used for Support Vector Machine (SVM) based classification. As reported in [4], such approach for membership discrimination gives better performance as compared to a generative probability model. In this work, we modified the SVM-Fisher method so that it can be combined in tandem with TMMOD.

In TMMOD, unlike with the task of family membership prediction using profile hidden Markov models, we are primarily interested in finding the most probable path of hidden states or label s for a given sequence x. Thus, our quantity of interest is not the likelihood the model assigns to sequences $P(x|\theta)$, but the conditional probability $P(s|x,\theta)$ of the label for a given observation sequence x [6, 8]. Using Bayes rule, the Fisher vector in this case should be of the form,

$$\begin{aligned} \vec{U_{s|x}} &= \nabla_\theta \ logP(s|x,\theta) \\ &= \nabla_\theta \ logP(s,x|\theta) \ - \ \nabla_\theta \ logP(x|\theta) \end{aligned} \tag{2}$$

where the joint probability $P(s,x|\theta)$ is calculated using the forward-backward algorithm by summing over only those valid paths that result in the given label s, whereas the total probability $P(x|\theta)$ is found using the standard forward-backward algorithm summing over all possible paths.

4. Calculation of Fisher gradients

To calculate the k^{th} component of the Fisher vector or the derivative with respect to the model parameter θ_k, let us consider the second term in Eq(2) first. The probability $P(x|\theta)$ can be written as a sum over all possible paths

through the model

$$P(x|\theta) = \sum_{\pi} P(x, \pi|\theta) \quad (3)$$

where $P(x, \pi|\theta)$ is the probability from a single path π and can be written as

$$P(x, \pi|\theta) = \prod_{i=1}^{N} \left(\frac{\theta_i}{\sum_a \theta_a} \right)^{n_i(\pi,x)} \quad (4)$$

The value $n_i(\pi, x)$ is the number of times the model parameter θ_i is used in the path π for the observation sequence x, and $\sum_a \theta_a = 1$ is the distribution θ_i is drawn from. The derivative of the likelihood score with respect to the k^{th} parameter is then,

$$\begin{aligned} \frac{\partial log(P(x|\theta))}{\partial \theta_k} &= \sum_{\pi} \frac{\partial P(x, \pi|\theta))}{\partial \theta_k} \frac{1}{P(x|\theta)} \\ &= \sum_{\pi} \left(\frac{1}{\theta_k} - 1 \right) n_k(\pi, x) \frac{P(x, \pi|\theta)}{P(x|\theta)} \\ &= \sum_{\pi} \left(\frac{1}{\theta_k} - 1 \right) n_k(\pi, x) P(\pi|x, \theta) \\ &= \frac{n_k(x)}{\theta_k} - n_k(x) \end{aligned} \quad (5)$$

where $n_k(x)$ is the expected number of times the k^{th} parameter is used by the observation sequence x given as a weighed average over all possible paths. This is basically the posterior probability for the parameter and can be easily computed from forward-backward matrices which are used to find $P(x|\theta)$ [6]. In the same manner, the first term in Eq(2) is differentiated and the final expression is given by

$$\frac{\partial log(P(s, x|\theta))}{\partial \theta_k} = \frac{m_k(x)}{\theta_k} - m_k(x) \quad (6)$$

Again, $m_k(x)$ is the expected number of times the k^{th} parameter is used by the observation sequence x by considering only those paths that result in the correct label s. The $m_k(x)$ posterior probability values are computed using the forward-backward matrices used to find $P(s, x|\theta)$ [6]. Finally, the complete expression for our modified Fisher gradient with respect to k^{th} parameter is,

$$\frac{\partial log(P(s|x, \theta))}{\partial \theta_k} = \frac{n_k(x) - m_k(x)}{\theta_k} + m_k(x) - n_k(x) \quad (7)$$

This expression for the gradient is slightly different from what is reported in [8], but is in agreement with that of [4].

5. Feature selection

Once the sequences are mapped into the Fisher score vectors, the next step is to use these vectors to train an SVM classifier which can be used to classify other sequences. To avoid overfitting when training an SVM classifier, it is desirable to train the SVM classifier using only those feature values that vary considerably between membrane proteins and signal peptides. In other words, we need to reduce the dimension of our Fisher vectors by selecting those discriminative components, i.e., the parameters that are believed to be used differently by TM proteins and signal peptides. For example, we know that signal peptides start in the cytoplasmic side and traverse the membrane just once with the cleavage site being located in the non-cytoplasmic side. This means we should look at Fisher gradients with respect to transition parameters for effective SVM based discrimination between the two classes of proteins.

To do this selection, Fisher gradients with respect to transition parameters in the loop submodels of TMMOD were computed for a set of 247 TM proteins and a set of 1275 proteins containing signal peptides which were compiled in [2]. The components of the normalized resultant Fisher vector for each class of proteins is shown in Figure 2. For each resultant gradient value, the corresponding state transition in TMMOD is shown on the x-axis.

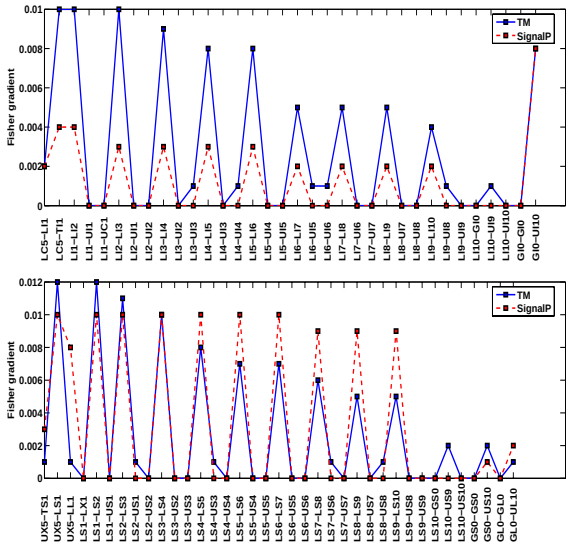

Figure 2: **Normalized resultant vectors. Fisher vectors that correspond to the 247 TM proteins are added and the resultant is normalized. The same thing is done for vectors that correspond to the 1275 signal peptides. The first panel is for gradients w.r.t transitions in the cytoplasmic loop submodel while the second is for gradients w.r.t transitions in the non-cytoplasmic loop submodel.**

As expected, the resultant vectors vary in the components corresponding to transitions that are used differently by the two sets of proteins. For the cytoplasmic loops,

Method	False SP/True TM	False TM/True SP
exp-no-aa	15/247 (6.1%)	185/1275 (14.5%)
SVM-Fisher	15/247 (6.1%)	117/1275 (9.2%)
Phobius	19 /247 (7.7%)	45/1275 (3.5%)
SignalP-NN	106/247 (42.9%)	29/1275 (2.3%)
SignalP-HMM	47 /247 (19.0%)	18/1275 (1.4%)

Table 1: **The second column reports the fraction (and percentage) of the 247 TM proteins falsely identified as signal peptides. The third column reports the fraction (and percentage) of the 1275 signal peptides falsely identifies as TM proteins. Results for the reference methods are from [2].**

these components correspond to outgoing transitions from the LC_5 state and the $LI_k - LI_{k+1}$ transitions. These transitions are much less used by signal peptides because signal peptides begin in the G or UI states and do not use those transitions. For non-cytoplasmic loops, the $LS_k - LS_{k+1}$ transitions for k greater than 3 and the $UX_5 - LL_1$ transition are more frequently used by signal peptides since signal peptides tend to have longer non-cytoplasmic loops. Only these parameters, therefore, will be used for generating Fisher vectors in our SVM-Fisher based discrimination.

6. Discrimination results

The performance of this SVM-Fisher discrimination approach was measured using 10-fold cross validation experiments. Each of the positive and negative datasets described earlier was partitioned into ten groups in such a way that the maximum sequence similarity between subsets is 40% [2]. A model is trained using nine subsets of the positive set and is used to transform these nine positive subsets and the remaining positive subset into positive training vectors and positive test vectors respectively. The same model is used to transform the nine negative subsets and the remaining negative subset into negative training vectors and negative test vectors respectively. These vectors are then used as input to an SVM classifier with a polynomial kernel. The training and testing of the SVM classifier is conducted using the SVM-Light package [7] with the package default setting. The whole procedure was repeated 10 times by rotating the testing subset among the 10 subsets. The discrimination performance results in comparison to our old discrimination method using the *exp-no-aa* measure and other reference methods are given in Table 1.

The performance of discriminating between signal peptides and transmembrane proteins has increased by identifying 68 signal peptides which were otherwise classified as membrane proteins by the *exp-no-aa* measure. We also compare the results for our SVM-Fisher approach with that of Phobius [2], the most recent successor of TMHMM, that is based on a hybrid hidden Markov model specialized to discriminate signal peptides from TM proteins. As shown in Table 1, Phobius has wrongly classified 4 more TM proteins as signal peptides and our SVM-Fisher has wrongly classified 72 more signal peptides as TM proteins. Although this combibination of TMMOD and SVM-Fisher is not as discriminative as Phobius, as shown in next section, TMMOD is still superior to Phobius and other methods for topology prediction.

Method	Dataset	1-best	5-best
TMMOD	set-all	178 (72.1%)	201 (81.4%)
Phobius		157 (63.6%)	
TMHMM2.0		161 (65.2%)	
HMMTOP2.1		165 (66.8%)	
TMMOD	set-new	75 (58.6%)	90 (69.7%)
Phobius		69 (53.9%)	
TMHMM2.0		57 (44.5%)	
HMMTOP2.1		65 (50.8%)	

Table 2: **Topology prediction accuracy. The third column gives the number (and percentage) of proteins whose topology is correctly predicted, by TMMOD using *1-best* scheme and by other methods. The last column gives the results when TMMOD is used with the *5-best* scheme.**

7. On topology prediction accuracy

Recently, Phobius[2], a combined transmembrane topology and signal peptide predictor has been reported. The transmembrane topology prediction accuracy of Phobius and other reference methods was validated using two data sets: *set-all* consisting 247 TM proteins and its subset *set-new* which excludes those sequences from *set-160* in [5]. A prediction was counted as correct when all predicted TM helices overlap all annotated TM helices by at least 5 residues and the predicted locations of the loops are correct. We performed a 10-fold cross validation experiment for TMMOD using the same data and the resulted accuracy along with those of Phobius and other reference methods is given in Table 2. As shown in the table, in both validation datasets, TMMOD is the most accurate topology predictor.

In TMMOD, topology is predicted using the Viterbi algorithm which finds the label or path of hidden states π^* that maximizes the probability $P(x, \pi|\theta)$ of the observation sequence x. For a model architecture such as that of TMMOD in which many states are labeled with the same symbol, it is very probable that many paths different from

π^* would give probabilities that are close to the maximum $P(x, \pi^*|\theta)$. Thus, the correct path we are looking for might be one of these highly probable paths instead of the optimal path π^*. To see if this is the case, our Viterbi algorithm was modified to give the 5 best decoded paths (*5-best*). So, if we were to be given five chances to predict the topology, the corresponding prediction accuracy results would improve, and this is shown in the last column of Table 2.

8. Conclusions

We have presented an application of the SVM-Fisher discrimination approach in combination with our hidden Markov model based predictor (TMMOD) to further improve the accuracy of discriminating integral membrane proteins from signal peptides. Using the SVM-Fisher discrimination method, we are able to reduce mis-prediction of signal peptides by more than 30%. Although a recent method Phobius has better performance in discriminating signal peptides from transmembrane proteins, TMMOD has better results for topology prediction accuracy, and our approach of combining TMMOD and SVM-Fisher offers a good tradeoff between the two.

Acknowledgments

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 and Engineering grant.

References

[1] R. Kahsay, L. Liao, G. Gao "An Improved Hidden Markov Model for Transmembrane Protein Detection and Topology Prediction and Its Applications to Complete Genomes," *Bioinformatics*, 21, pp. 1853-1858.

[2] L. Kall, A. Krogh, E. Sonnhammer "A combined transmembrane topology and signal peptide prediction method," *Journal of Molecular Biology*, 338, pp. 1027-1036.

[3] 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.

[4] T. Jaakkola, M. Diekhans, D. Haussler "A discriminative framework for detecting remote protein homologies" *Computational Biology*, 7, pp. 95-114.

[5] 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.

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

[7] T. Joachims *Making large-Scale SVM Learning Practical*, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, Cambridge

[8] A. Krogh "Hidden Markov models for labeled sequences," *In Proc. of the 12th IAPR International Conference on Pattern Recognition, IEEE Computer Society Press, Los Alamitos, CA*, pp. 140-44.