*Sequence analysis*

# An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes

Robel Y. Kahsay[1], Guang Gao[1] and Li Liao[1,2,∗]

[1]Delaware Biotechnology Institute, Newark, DE 19715, USA and [2]Department of Computer and Information Sciences, University of Delaware, 103 Smith Hall, Newark, DE 19716, USA

## ABSTRACT

**Motivation:** Knowledge of the transmembrane helical topology can help identify binding sites and infer functions for membrane proteins. However, because membrane proteins are hard to solubilize and purify, only a very small amount of membrane proteins have structure and topology experimentally determined. This has motivated various computational methods for predicting the topology of membrane proteins.

**Results:** We present an improved hidden Markov model, TMMOD, for the identification and topology prediction of transmembrane proteins. Our model uses TMHMM as a prototype, but differs from TMHMM by the architecture of the submodels for loops on both sides of the membrane and also by the model training procedure. In cross-validation experiments using a set of 83 transmembrane proteins with known topology, TMMOD outperformed TMHMM and other existing methods, with an accuracy of 89% for both topology and locations. In another experiment using a separate set of 160 transmembrane proteins, TMMOD had 84% for topology and 89% for locations. When utilized for identifying transmembrane proteins from non-transmembrane proteins, particularly signal peptides, TMMOD has consistently fewer false positives than TMHMM does. Application of TMMOD to a collection of complete genomes shows that the number of predicted membrane proteins accounts for ∼20–30% of all genes in those genomes, and that the topology where both the N- and C-termini are in the cytoplasm is dominant in these organisms except for *Caenorhabditis elegans*.

**Availability:** http://liao.cis.udel.edu/website/servers/TMMOD/

**Contact:** lliao@cis.udel.edu

## INTRODUCTION

Membrane proteins serve as highly active mediators between the cell and its environment or between the interior of an organelle and the cytosol. They catalyze specific metabolites and ions across the 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 act as signal receptors and transduce signals such as neurotransmitters, growth factors and hormones across the membrane. 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 they have both hydrophobic and hydrophilic regions on their surfaces, and thus it is difficult to determine their structure experimentally. Such a situation has motivated the development of various computational methods for predicting the topology of membrane proteins. Most of these computational approaches rely on the compositional bias of amino acids at different regions of the sequence (von Heijne, 1994). For example, there is a high propensity of hydrophobic residues in transmembrane alpha helices due to the hydrophobic environment in lipid membranes. Because such a 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 in membrane proteins is the abundance of positively charged residues in the segments (loops) 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 (von Heijne, 1986, 1992). Unlike the hydrophobicity 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. Consequently, it is more difficult to correctly predict the overall topology of a given protein, i.e. the orientation of all transmembrane segments.

There are basically two ways for improving the prediction accuracy of any given model: by enhancing the signal/noise ratio for those weak signals or by identifying new signals and associating them with the topology. For example, significant improvements of prediction accuracy were reported (Persson and Argos, 1994) by applying multiple sequence alignment to proteins with similar topology so that the positive residue content in the cytoplasmic loops may become obvious in the aligned motifs. A more recent work along this line is PRODIV–TMHMM, a profile-based hidden Markov model (Viklund and Elofsson, 2004), where a 10% increase in performance is reported with the use of homologous sequences. However, it shall be noted that multiple sequence alignment may not always be suitable, either due to insufficient number of homologs or due to the length variations in these cytoplasmic loops. Other methods have been attempted at exploring more subtle signals such as correlation of compositional bias at different positions. The best performance attained so far is by using artificial neural networks (Rost *et al*., 1996), a method known

---

*To whom correspondence should be addressed.

for its capability of capturing complex nonlinear signals. Despite its improvement at prediction accuracy, the artificial neural network method, well known for its black-box property, provides little insight into those signals that the network is designed to capture.

A hidden Markov model, TMHMM, has recently been used for transmembrane topology prediction (Sonnhammer *et al.*, 1998; Krogh *et al.*, 2001). Hidden Markov models, as a probabilistic framework, have been widely applied in computational biology with remarkable success (Durbin *et al.*, 1998). Unlike artificial neural networks, the architecture of hidden Markov models corresponds closely to the biological entities being simulated by the model. In TMHMM, the model comprises seven sets of states, with each set corresponding to a type of regions in the protein sequence. Each set of states has an associated probability distribution over the 20 amino acids characterizing the compositional bias in the corresponding regions. In addition, the model architecture specifies the interconnection of states within each set or submodel and also specifies how these submodels are connected to one another. Transitions among states within a given submodel determine the length distribution of the corresponding regions whereas transitions from one submodel to another reflect how the different regions are arranged to form the entire protein. The transition probabilities, along with emission frequencies, enable the model to capture correlations among signals.

In this work, we present an improved hidden Markov model, TMMOD, for predicting transmembrane topology and identifying transmembrane proteins from soluble proteins. The TMMOD differs from TMHMM in both model architecture and training procedure. The architectural differences are on the cytoplasmic and the non-cytoplasmic loop submodels. For the training procedure, we adopt the Bayesian based approach where the model parameters are set by posterior mean estimator (PME). In cross-validation experiments using the datasets which were used by TMHMM, our model outperformed TMHMM in both topology prediction and identification of membrane proteins.

## MATERIALS AND METHODS

### The TMMOD model architecture

The overall skeleton of the TMMOD's architecture is a linear structure of three two-way connected submodels for cytoplasmic loop, transmembrane helix and non-cytoplasmic loop. The two-way connections between the cytoplasmic loop and the transmembrane helix and between the transmembrane helix and the non-cytoplasmic loop, plus a self return connection in the loop submodels, allow a path cycling through the three components of transmembrane proteins: cyto-loop, helix and noncyto-loop. A path can start with either a cyto-loop or a noncyto-loop, reflecting the fact that a transmembrane protein can have its N-terminus either inside or outside the cell. The architectures of the submodels for these three components are illustrated in Figure 1.

The submodel for transmembrane helix, identical to that of TMHMM, has two cap regions each of five residues surrounding a core region of variable length 5–25 residues (Fig. 1A). Therefore, the model can represent helices of size 15–35 residues long, a range that covers the actual sizes observed for transmembrane domains. This submodel contains two chains of transmembrane states, with one chain going inwards and the other going outwards, as a mechanism to enforce the structural constraint, i.e. a transmembrane helix has to span the membrane. Since there are no observed differences in amino acid composition and length distributions between 'inwards' helices and 'outwards' helices, the emission and transition parameters for these two chains are estimated collectively.

The architecture of TMMOD differs from that of TMHMM by how the loops are modeled (Fig. 1B, C and D). In order to capture the known biases of
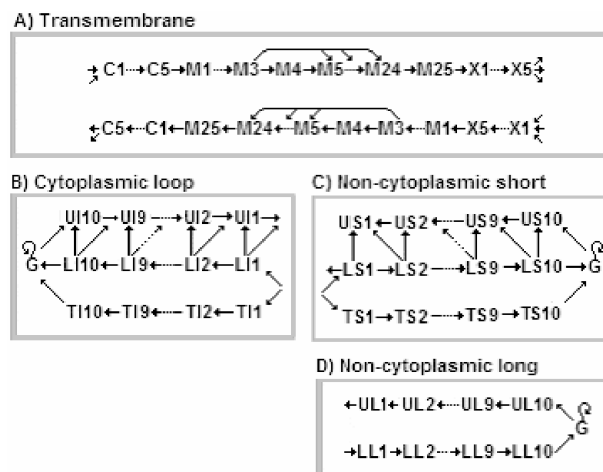


**Fig. 1.** Architecture of the four submodels: (**A**) transmembrane submodel (C, M and X state types), (**B**) cytoplasmic loop submodel (G and I state types), (**C**) non-cytoplasmic short loop submodel (S and G state types), (**D**) non-cytoplasmic long loop submodel (L and G state types). To assemble the model, panel (B) shall be attached to the left of panel (A), with UI1 and LI1 in (B) pointing to C1 in (A), and C5 in (A) pointing to LI1 and TI1 in B. Both panels (C) and (D) shall be attached to the right of panel (A).

amino acid compositions at near the border between loops and helices, the first and last 10 residues of a loop region are explicitly modeled, i.e. each residue corresponds to an individual state in the model. These 20 states are marked as LI1–LI10 and UI1–UI10 in Figure 1B for loops inside the cytoplasm and as LS1–LS10 and US1–US10 in Figure 1C for short loops in the non-cytoplasm. As shown in Figure 1B and C, a ladder-like structure is formed to allow for loop length to vary from just one residue, by traversing only state LI1 or LS1, to 20 residues, by traversing all 20 states. All other residues in the middle of a loop longer than 20 are collectively represented by one 'globular' state which has a transition back to itself and thus can repeat as many times as the loop length dictates. Following TMHMM, since non-cytoplasmic loops longer than 100 appear to have compositional characteristics different from those of the short non-cytoplasmic loops, two separate non-cytoplasmic loop submodels are used for representing them, as depicted in Figure 1C and D.

Inspired by TMHMM's design of using a separate submodel for long loops, we studied the length distribution of the loops in the training sequences (Fig. 2). The length distribution shows that, ~90% of the loops are shorter than 40 residues, and the rest are quite spread out, as indicated by a long tail. Similar findings with respect to the loop length distribution were reported in Wallin and von Heijne (1998) and Liu and Rost (2001). To capture this distribution more effectively, we introduced a separate chain of states (Figure 1B and C) in parallel to the ladder-like structure in cytoplasmic and short non-cytoplasmic loop submodels. As such, we want the transition parameters in the ladder-like part of the submodels to explicitly model the length distribution of the loops that are <40 amino acids, while the longer loops are directed through the bypass. More specifically, the transition probability $LI_k \to LI_{k+1}$ or $LSI_k \to LS_{k+1}$ now reflects the likelihood of loops with length $>2k$ but $<40$; whereas the same transition parameter in a submodel without the bypass would have to reflect the likelihood of all loops with lengths $>2k$. We expect that such an effective estimation of the distribution of loop lengths would enhance the signal-to-noise ratio of the topogenic signal.

Another architectural difference from TMHMM is the use of a simpler submodel, as shown in Figure 1D, for non-cytoplasmic loops with lengths $>100$ amino acids. These long loops do not require a ladder-like structure since all of them are $>100$ amino acids and thus there is no need for an early exit. Overall, TMHMM reported 83 transition and 133 emission parameters, whereas our model has 66 transition and 133 emission parameters.
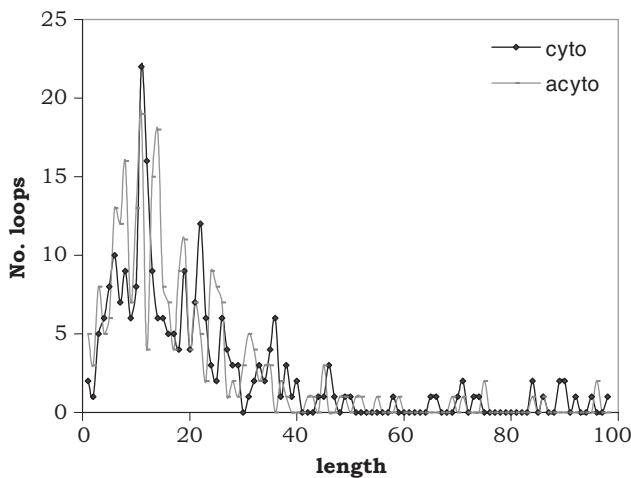
**Fig. 2.** Length distribution of cytoplasmic and short non-cytoplasmic loops.

## The TMMOD model training

Model parameters are estimated by Bayesian approach (PME) using single Dirichlet and substitution matrix mixtures priors (Durbin *et al.*, 1998). For each of the seven types of states as shown in Figure 1, the substitution matrix mixtures is given by

$$\beta_{ja} = A \sum_b c_{jb} P(a|b) \tag{1}$$

where $\beta_{ja}$ is the pseudocount for amino acid '$a$' in state type $j$, $c_{jb}$ is the observed frequency (or count) for amino acid '$b$' in state type $j$, $P(a|b)$ is the conditional probability of amino acid '$a$' given amino acid '$b$' (derived from BLOSUM50 matrix), and $A$ is a constant.

The technique of using Dirichlet prior assumes that the observed frequencies of 20 amino acids in each of the seven types of states were stochastically generated from a distribution $\vec{p} = (p_1, \ldots, p_{20})$, which itself is chosen from a distribution specified by a parametric Dirichlet density $\rho(\vec{p})$,

$$\rho(\vec{p}) = \frac{\prod_{a=1}^{20} p_i^{\alpha_a - 1}}{Z} \tag{2}$$

where $Z$ is the normalizing constant. Each of the training sequences, with their topology known, is partitioned into segments according to the state types such that all residues in a segment are emitted from the same type of states. An observed count vector over amino acids is found for each of these segments, and these count vectors are grouped into seven classes according to the state types. For each class, the parametric Dirichlet density function ($\alpha$ parameters) is estimated from the observed count vectors by following a procedure outlined in Brown *et al.* (1993) and Sjolander *et al.* (1996). Then, a pseudocount for amino acid '$a$' in states of type $j$ is given as

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

The above equation for deriving pseudocounts differs from the standard by the constant $A$, which is introduced to tighten the Dirichlet density without affecting its mean (Durbin *et al.*, 1998). The final emission frequency of amino acid '$a$' from state type $j$ after adding both types of pseudocounts is then given as follows:

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

We also produce a single component Dirichlet pseudocount vector to regularize transitions in the ladder-like part of the submodels by taking the three outgoing transition counts from each of the lower chain of states as vectors in three dimensions. A detailed description of our model training procedure is described in Kahsay *et al.* (2004).

The topology of a membrane protein is predicted using Viterbi algorithm. We also compute the three posterior probabilities that a given residue is in a transmembrane helix, on the cytoplasmic side or on the periplasmic side. This additional information, which at times can be even more informative than the single most probable state path, shows where the prediction is certain and what alternatives there might be.

## Datasets

The two datasets used to validate the model on topology prediction were downloaded from the TMHMM website (http://www.cbs.dtu.dk/services/TMHMM). The first dataset contains 83 transmembrane sequences of known topology, with 45 of them being single spanning. The second dataset has 160 transmembrane sequences, with 52 of them being single spanning. The topology of most proteins in these datasets is determined experimentally.

We adopted the same 10-fold cross-validation for topology prediction as in Sonnhammer *et al.* (1998). Both datasets are divided into 10 subsets. The subsets from the first dataset contain either eight or nine sequences, and all the subsets of the second dataset have exactly 16 sequences each. To make the learning task more challenging, the subsets are prepared in such a way that 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 subset. This is repeated 10 times, and each time a different subset is selected as the test set. The prediction accuracy is the average over the 10 runs.

For discrimination or identification experiments, test datasets contain the set of 160 transmembrane proteins (positives) and other non-transmembrane proteins (negatives). The test datasets for discrimination experiments are the same as used in Krogh *et al.* (2001). These datasets include 645 soluble proteins, six porins and a set of signal peptides from different classes of organisms. For whole genome analysis, all genes annotated in the Genbank entry of the genomes and chromosomes used were downloaded from ftp://ncbi.nlm.nih.gov/genbank/genomes/, except for *Caenorhabditis elegans*, which was downloaded from the URL: ftp://genome.wustl.edu/pub/

## RESULTS

### Topology prediction

The accuracy is measured by the number of sequences from the test sets whose topology and location of all transmembrane helices are correctly predicted. Following the same criterion used in Sonnhammer *et al.* (1998), a predicted helix is counted as correct if it overlaps by at least five residues with a true helix. The performance is also measured by the sensitivity and specificity for identifying individual transmembrane domains.

To help us understand and assess how the model architecture and use of different regularizers have contributed to the performance, three variations of the architecture, including the one shown in Figure 1, and three regularization schemes are tested. Model M1 is the architecture of TMHMM with our training. Model M2 has two bypassed ladder-like submodels on each side of the membrane, a design intended to see if differentiating long and short loops on the cytoplasmic side as well will perform better. Model M3 is the architecture shown in Figure 1. Regularizer scheme (a) uses Dirichlet prior based pseudocounts; (b) uses substitution matrix based pseudocounts; and (c) uses both. The performance for these variations on the two datasets is given in Table 1. It is shown that the model depicted in Figure 1, using both Dirichlet and substitute matrix based regularizers, has achieved the best performance: 89% accuracy for both topology and location on the first dataset (83 sequences), and 84% accuracy for topology and 89% accuracy for locations on the second dataset (160 sequences). The performance improvement of TMMOD

**Table 1.** Prediction accuracy for the cross-validation experiments

| Model | Regularizer scheme | Dataset | Correct topology | Correct location | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|---|
| M1 | (a) | S-83 | 65 (78.3%) | 67 (80.7%) | 97.4 | 97.4 |
|  | (b) |  | 51 (61.4%) | 52 (62.7%) | 71.3 | 71.3 |
|  | (c) |  | 64 (77.1%) | 65 (78.3%) | 97.1 | 97.1 |
| M2 | (a) | S-83 | 61 (73.5%) | 65 (78.3%) | 99.4 | 97.4 |
|  | (b) |  | 54 (65.1%) | 61 (73.5%) | 93.8 | 71.3 |
|  | (c) |  | 54 (65.1%) | 66 (79.5%) | 99.7 | 97.1 |
| M3 | (a) | S-83 | 70 (84.3%) | 71 (85.5%) | 98.2 | 97.4 |
|  | (b) |  | 64 (77.1%) | 65 (78.3%) | 95.3 | 71.3 |
|  | (c) |  | **74 (89.2%)** | **74 (89.2%)** | **99.1** | **97.1** |
| TMHMM |  | S-83 | **64 (77.1%)** | **69 (83.1%)** | **96.2** | **96.2** |
| PHDtm |  | S-83 | **(85.5%)** | **(88.0%)** | **98.8** | **95.2** |
| M1 | (a) | S-160 | 117 (73.1%) | 128 (80.0%) | 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 |
| M2 | (a) | S-160 | 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 |
| M3 | (a) | S-160 | 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 |  | S-160 | **123 (76.9%)** | **134 (83.8%)** | **97.1** | **97.7** |

Numbers in bold represent the best results from different methods.

over that of TMHMM (77% topology and 83% locations on the first dataset, and 77% topology and 84% locations on the second dataset) is significant. It is noted that, on the first dataset where the results for PHDtm are also available, the TMMOD's performance even slightly exceeds the performance (86% for topology and 88% for locations) of PHDtm, the best existing method, which utilizes multiple alignments—a data source that carries extra information. It is also worth noting that, because proteins with known topologies (ones applied for the training) constitute a biased set, the expected accuracy when applied to entire proteomes may be significantly lower, as was shown by Melen *et al.* (2004).

In addition to the outstanding performance for TMMOD, several other observations can also be made from Table 1 about the effect of different variations on model architecture and regularization. First, we notice that the Dirichlet prior based regularizer is consistently more effective than the substitution matrix mixture based regularizer for all three different architectures. Second, we noticed that combining the Dirichlet and the substitute matrix mixture based priors enhanced the model performance, but not always; indeed the performance was even decreased in some cases. In the contrast, we notice that M3 attained the best performance among the three architectures in all three variations of regularizers, suggesting that the model architecture played a more decisive role for better performance. Another observation is that M2, which has two bypassed ladder-like loop submodels on each side of the membrane, has better performance than M1 which is the original TMHMM architecture; it is reasonable to believe that the better performance is probably due to the bypass introduced. However, the best performance is achieved by model M3, which has the bypassed ladder-like loop submodels on both sides of the membrane, but has an extra, simple submodel for loops (longer that 100) only on the non-cytoplasmic side. This observation further validates the hypothesis made in

TMHMM that differentiation of short and long loops only applies to the non-cytoplasmic side.

## Discrimination between non-membrane and membrane proteins

In addition to predicting the transmembrane protein topology, TMMOD can also be used for identifying/discriminating helical membrane proteins from other proteins. In general, this can be done by using Forward algorithm to calculate the model likelihood for a given sequence (Durbin *et al.*, 1998). For comparison reasons, we adopted the three more refined measures proposed in Krogh *et al.* (2001). The first measure, abbreviated as 'pred. no. tmh', is simply the number of helices in the most likely structure found by the model. The other two measures are the expected number of residues in transmembrane ('exp. no. aa') and the expected number of transmembrane helices ('exp. no. tmh'), which are computed from the posterior probabilities. The probability that a given sequence is a membrane protein is higher when the expected number of residues in any of the predicted helices is high. Since the shortest transmembrane helices are ~18 residues long, a cut-off value should be set at ~18. If the expected number of transmembrane helices in a protein is ≥1, the protein is likely to be a helical transmembrane protein.

In a discrimination experiment designed to identify the 160 membrane proteins from the 645 water-soluble proteins, the measures described above were calculated using the 10-fold cross-validation models. This means, the measures for the 16 sequences in a given subset are calculated using a model that was trained on the remaining nine subsets (144 sequences). For the non-membrane proteins, the averages over the 10 cross-validation models were calculated. Even though all the three measures give discrimination with high accuracy, we have used the 'exp. no. aa' as our standard measure. Figure 3 shows the fraction of false positives and negatives at different
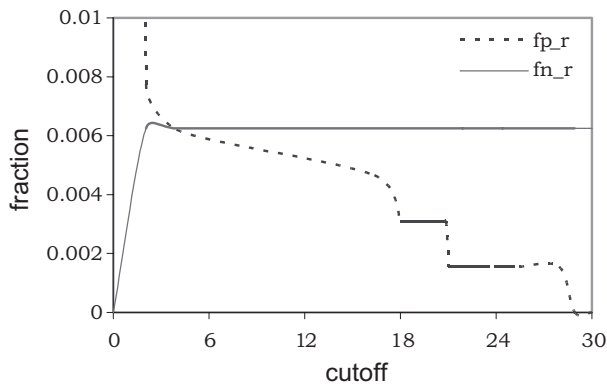
**Fig. 3.** Discrimination between transmembrane proteins and soluble proteins. A decision is made based on the expected number of residues in transmembrane helices ('exp. no. aa'). The fraction of false negative (continuous line) and false positive predictions (broken line) as a function of the cut-off value of 'exp. no. aa'.

**Table 2.** False positives by TMMOD and TMHMM

| Model | PDB entries | Description | Expected number aa in membrane | SD |
|---|---|---|---|---|
| TMHMM | 1RDZ (A) | Fructose 1,6-bisphosphatase | 24.3 | 3.6 |
| | 1KVD (A) | Smk toxin | 24.7 | 0.5 |
| | 1NOX | Nadh oxidase | 21.0 | 1.1 |
| | 1CIY | CryIA (a) insecticidal toxin | 20.6 | 1.6 |
| | 1ENO | Enoyl acyl carrier protein reductase | 18.9 | 5.7 |
| TMMOD | 1KVD (A) | Smk toxin | 27.9 | 0.2 |
| | 1CIY_ | CryIA (a) insecticidal toxin | 29.9 | 0.1 |

The PDB entries all have a known 3D structure. The PDB identifiers with the chain in parenthesis are listed in Column 2. Column 4 contains the expected number of residues in helices ('exp. no. aa') averaged over the ten cross-validation models, and the last column is the standard deviation.

cut-off values of this measure. At a cut-off value of 18, TMMOD has two false positives (~0.3%) whereas TMHMM reported five false positives (~0.6%) (Table 2). As it was the case with TMHMM, the chlorophyll *a–b* binding protein ab96 (Swiss-Prot entry CB21_PEA) is the only membrane protein that is classified as a non-membrane protein (i.e. as false negative).

## Signal peptides and porins

The signal peptides that target a protein for export contain a hydrophobic region and can easily be mistaken as a transmembrane region. TMMOD was tested on a set of signal peptides by measuring how many of the signal peptides were predicted to be membrane proteins by using the measure as described above. As shown in Table 3, TMMOD performed much better than TMHMM at identifying signal peptides as non-membrane proteins. A substantial improvement over TMHMM at discriminating signal peptides from TM is also reported in Phobius, an integrated hidden Markov model that can model both transmembrane topology and signal peptide (Kall *et al.*, 2004). We also tested TMMOD on the six porins from Krogh *et al.* (2001), all of which were correctly predicted as not containing transmembrane helices.

**Table 3.** The number of signal peptides mistakenly predicted as transmembrane proteins

| Class | No. of signal peptides | No. predicted as tm by TMHMM | No. predicted as tm by TMMOD |
|---|---|---|---|
| Eukaryotes | 1011 | 209 (21%) | 87 (9%) |
| Gram-negatives | 266 | 60 (23%) | 33 (12%) |
| Gram-positives | 141 | 85 (60%) | 60 (43%) |

The first column is the organism type, and the second column is the total number of signal peptides in that class. The last two columns are the number of false positives for TMHMM and TMMOD respectively.

**Table 4.** The number of predicted transmembrane proteins in complete genomes

| Organism | Number of genes | Exp. no. aa ≥18 (%) | Pred. tmh ≥1 (%) |
|---|---|---|---|
| *Treponema pallidum* | 1031 | 20.37 (23.4) | 20.5 (23.7) |
| *Borrelia burgdorferi* | 850 | 25.5 (28.7) | 27.8 (28.7) |
| *Chlamydia pneumoniae* | 1052 | 25.9 (27.9) | 26.1 (27.8) |
| *Chlamydia trachomatis* | 894 | 21.7 (23.3) | 22.4 (24.5) |
| *Aquifex aeolicus* | 1522 | 18.9 (20.3) | 19.9 (20.7) |
| *Synechococcus sp* | 3169 | 23.98 (25.8) | 24.1 (25.8) |
| *Thermotaga maritima* | 1846 | 21.99 (22.9) | 22.6 (24.1) |
| *Methanococcus jannaschii* | 1715 | 18.1 (18.5) | 18.4 (18.9) |
| *Methanobacterium thermoautotrophicum* | 1869 | 19.8 (21.8) | 20.4 (21.8) |
| *Archaeoglobus fulgidus* | 2407 | 19.1 (20.3) | 19.3 (20.4) |
| *Pyrococcus abyssi* | 1765 | 21.4 (22.6) | 22.2 (22.9) |
| *Pyrococcus horikoshii* | 2064 | 23.21 (27.5) | 24.4 (25.9) |

For each organism, the number of annotated genes, the percentages of membrane proteins predicted using the two measures described in text.

## Genome-wide analysis of membrane proteins

For genome annotation purpose, it is desirable to have an accurate estimate of the number of membrane proteins as well as an accurate estimate of the frequency for proteins of different topologies to be expected in a given genome. The outstanding performance of TMMOD on both discriminating membrane proteins from soluble proteins and predicting the transmembrane topology has motivated us to apply it to estimate the number of membrane proteins in a collection of organisms with fully sequenced genomes. A similar work of using TMHMM is reported in Krogh *et al.* (2001). Due to space limitation, we only report estimates for the genomes that do not develop signal peptides. The complete results for 21 genomes are available at the TMMOD webserver.

A model (M3) was trained using all the 160 sequences in the second training set as described above. For each genome, transmembrane proteins were predicted based on the 'pred. no. tmh' and 'exp. no. aa' measures described earlier. For these predicted transmembrane proteins, their topology was also predicted.

Table 4 summarizes the predictions by TMMOD on 12 complete genomes, including the number of genes that are predicted to encode integral membrane proteins. In general, the number of predicted integral membrane proteins comprises between ~20 and ~30% of

the total number genes for a genome. For almost all organisms in Table 4, TMHMM, whose results are listed in parentheses, has predicted more proteins as integral membrane proteins than TMMOD does. This finding, together with the results from the previous discrimination experiments where TMHMM was shown to have had more false positives, leads us to speculate that TMHMM may suffer from problems of over-prediction. As for the occurrence frequencies of different topologies, we found that multispanning proteins with both N- and C-termini inside cytoplasm are strongly preferred in all organisms with the exception of *C.elegans*. This is in agreement with the predictions from TMHMM.

## DISCUSSION

We presented here an improved hidden Markov model TMMOD for transmembrane topology prediction. In the cross-validation experiments on membrane proteins with known topology, TMMOD outperforms not only a similar method, TMHMM, on which our model is prototyped, but also the previously best method (PHDhtm) which utilizes presumably more information from multiple alignments. TMMOD also surpassed TMHMM in identifying integral membrane proteins from other proteins, particularly, signal peptides.

By running TMMOD on a group of 21 complete genomes, we estimate that integral membrane proteins account for ∼20–30% of all genes in all genomes, and that $N_{in}$–$C_{in}$ topology of transmembrane proteins, namely with both the N- and C-termini inside cytoplasm, is preferred in all organisms except *C.elegans*. This result is in general agreement with what is reported in Krogh *et al.* (2001).

By experimenting with different variations of model architecture and training regularizers, we concluded that the model architecture is a more decisive factor for better performance. It is of further interest to refine the model architecture, particularly in such a way that long range correlations across different regions of integral membrane proteins can be better captured.

It is worth noting that substantial improvements in accuracy over TMHMM are also reported in some recent works (Arai *et al.*, 2004; Kall *et al.*, 2004). These methods achieved better performance either via a 'consensus' of various individual methods (Arai *et al.*, 2004) or by a more integrated way (Kall *et al.*, 2004). Although it is difficult to directly compare TMMOD to these methods due to the use of different datasets, these methods can benefit from TMMOD by either weighing in TMMOD for the 'consensus' or incorporating TMMOD's loop treatments into the architecture.

## REFERENCES

Arai,M. *et al.* (2004) ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.*, **32**, W390–W393.

Brown,M., Hughey,R., Krogh,A., Mian,I.S., Sjolander,K. and Haussler,D. (1993) Using dirichlet mixture priors to derive hidden markov models for protein families. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 47–55.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998). *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Kahsay,R., Liao,L. and Gao,G. (2004) An improved hidden markov model for transmembrane topology prediction. In *The Proceedings of the 16th IEEE International Conference for Tools with Artificial Intelligence*, IEEE Computer Society, pp. 634–639.

Kall,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Krogh,A. *et al.* (2001) Prediction transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.

Melen,K. *et al.* (2004) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.

Persson,B. and Argos,P. (1997) Prediction of transmembrane protein topology utilizing multiple sequence alignments. *J. Protein Chem.*, **16**, 453–457.

Rost,B. *et al.* (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.

Sjolander,K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.

Sonnhammer,E. *et al.* (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. ISMB* **6**, 175–182.

Viklund,H. and Elofsson,A. (2004) Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.

von Heijne,G. (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. *EMBO J.*, **5**, 3021–3027.

von Heijne,G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.*, **225**, 487–494.

von Heijne,G. (1994) Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167–192.

Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029–1038.