

# THE USE OF FUNCTIONAL DOMAINS TO IMPROVE TRANSMEMBRANE PROTEIN TOPOLOGY PREDICTION

EMILY W. XU

*Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Calgary, HS-1150,  
3330 Hospital Drive NW, Calgary, AB T2N 4N1, Canada*

*ewxu@ucalgary.ca*

PAUL KEARNEY

*Caprion Pharmaceuticals Inc., 7150 Alexander-Fleming, Montreal, QC H4S 2C8, Canada*

*pkearney@caprion.com*

DANIEL G. BROWN

*School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1,  
Canada*

*browndg@cs.uwaterloo.ca*

Transmembrane proteins affect vital cellular functions and pathogenesis, and are a focus of drug design. It is difficult to obtain diffraction quality crystals to study transmembrane protein structure. Computational tools for transmembrane protein topology prediction fill in the gap between the abundance of transmembrane proteins and the scarcity of known membrane protein structures. Their prediction accuracy is still inadequate: TMHMM, the current state-of-the-art method, has less than 52% accuracy in topology prediction on one set of transmembrane proteins of known topology. Based on the observation that there are functional domains that occur preferentially internal or external to the membrane, we have extended the model of TMHMM to incorporate functional domains, using a probabilistic approach originally developed for computational gene finding. Our extension is better than TMHMM in predicting the topology of transmembrane proteins. As prediction of functional domain improves, our system's prediction accuracy will likely improve as well.

*Keywords:* transmembrane protein topology prediction; functional domains; PROSITE; hidden Markov models; TMHMM.

## 1. Introduction

About 20% to 25% of the proteins encoded by a typical genome are membrane proteins.<sup>1, 2, 3</sup> These include both integral (transmembrane or TM) and peripheral membrane proteins. There are two known classes of integral membrane proteins: those with  $\alpha$ -helical structure and those with  $\beta$ -barrel structure. Alpha-helical membrane proteins are the predominant type; thus, they are the focus of our modeling. Figure 1 illustrates a model for the topology of a hypothetical TM protein. Typical computational programs often fail to infer the correct topology.<sup>1, 4, 5</sup> Our goal in this work is to improve a standard hidden Markov model approach for TM protein topology prediction, by incorporating probabilistically the presence of sequence tags that are preferentially internal to or external to membranes.

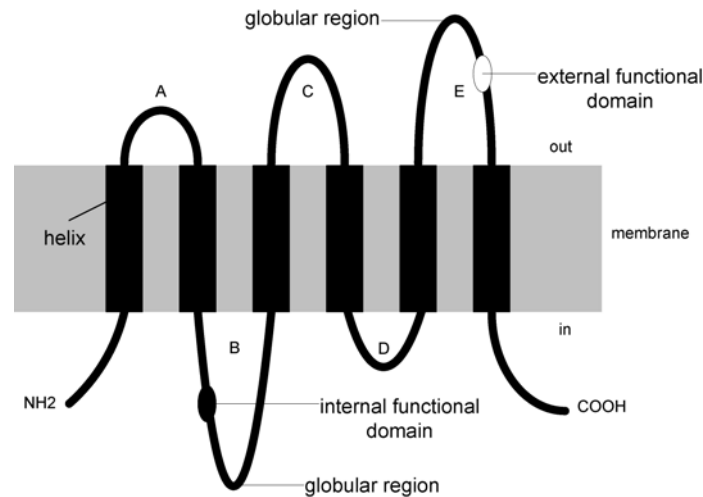


Figure 1: A model to illustrate the topology of a hypothetical transmembrane protein, with six helices, three extracellular loops (A, C and E) and two intracellular loops (B and D). On loop E, there is an external functional domain; on loop B, there is an internal functional domain. The N- and C-terminus are both internal to the membrane.

## 2. Computational Prediction of TM Protein Topology

The basic problem in TM protein topology prediction is to find the location and orientation (sidedness<sup>a</sup>) of the membrane spanning segments (helices). Local and global approaches are used for this problem. The local approach looks at local features of sliding windows of the amino acid sequence, such as hydrophobicity, and identifies likely helices and loops based on these properties. Global approaches, instead, look to identify all helices and loops as a group, optimizing a global criterion. A canonical example of this approach is the use of hidden Markov model (HMM)-based prediction methods. The main weakness of the local approach is the lack of specificity: it predicts too many false helices. On the other hand, the global approach examines sequences as a whole and does not set any empirical cutoffs and rules.<sup>1</sup>

### 2.1 Features of TM Proteins for *in silico* Modeling

Several features of TM proteins help in predicting topology. For example, helices are more hydrophobic than loops of TM proteins. The positively charged residues arginine (R) and lysine (K) are mainly found on the cytoplasmic side of TM proteins (the Positive-

<sup>a</sup> We define the “outside” of a TM protein as the part external to the membrane, while the “inside” of the TM protein is the region interior to the membrane.

Inside Rule of von Heijne<sup>6</sup>), and play a major role in determining orientation. Hydrophobicity and the Positive-Inside Rule have been used widely in TM protein topology prediction methods.

## **2.2 Hidden Markov Model**

A hidden Markov model is a probabilistic generative finite automaton. Widely used in bioinformatics,<sup>7</sup> it allows the representation of sequence features by states in the automaton, where each state emits residues based on a probabilistic distribution, and then transitions to another state based on another state-specific distribution. HMMs produce sequences for which the generating state is hidden, but the standard Viterbi decoding algorithm<sup>7</sup> can identify for a sequence the maximum probability sequence of states that could give rise to the sequence. The states of this sequence correspond to the features annotated by the HMM on a given sequence. HMMs easily model both the lengths of features in TM protein sequences, such as loops and helices, and their sequence content.

## **2.3 Review of Existing HMM Models**

### **2.3.1 Transmembrane HMM (TMHMM)**

TMHMM's model contains seven different types of states: one for the core of transmembrane helices, two for caps on either side, one for loops on the cytoplasmic side, two for short and long loops on the non-cytoplasmic side, and one for 'globular domains' in the middle of a loop. Because of the limited number of proteins of known topology for training, for each state type, many states of that type have the same emission probability distribution, to avoid overfitting. The transition matrix is a sparse matrix. There is no difference in the models of TMHMM 1.0<sup>8</sup> and TMHMM 2.0<sup>2</sup> (collectively known as TMHMM in this paper), but TMHMM 2.0 was retrained on the same data set, and has higher prediction accuracy.<sup>1</sup> TMHMM's HMM can model helices 15–35 residues long, which is the longest among current HMM models.

### **2.3.2 HMM for topology prediction (HMMTOP)**

HMMTOP<sup>9</sup> is based on the idea that sequences found in different parts of the cell should have dramatically different residue contents, but the resultant model is largely a straightforward HMM with five types of states: inside and outside loops, inside and outside tails and transmembrane helices. Two tails between adjacent helices form a short loop and tail-loop-tail form a long loop. Since it has only five types of states, rather than TMHMM's seven, its performance may be poorer, as seven types of states may give greater sensitivity to the variation of the amino acid compositions.<sup>10</sup>

HMMTOP's creators found that short loops with lengths between 5 and 30 amino acid residues appeared significantly more often than expected (a different distribution than geometric distribution).<sup>9</sup> Consequently, they modeled the length of a tail of 1–15 residues.

## 2.4 Newly developed techniques

Two recently developed methods, ENSEMBLE<sup>11</sup> and Phobius<sup>12</sup>, and PRODIV-TMHMM, a recent expansion of TMHMM to include sequence profiles<sup>13</sup> combine multiple approaches and add more information into the system to aid prediction. These more contemporary methods parallel methods in other areas of sequence analysis, such as gene finding, where external information such as alignments or repeats are incorporated into probabilistic gene finders, or where multiple gene finders have their results joined.<sup>14, 15, 16</sup>

### 2.4.1 ENSEMBLE

ENSEMBLE explicitly combines a neural network and two HMMs to make predictions. One HMM models the hydrophobicity of TM helices, and the other models the amphipathicity of TM helices. ENSEMBLE stresses the amphipathicity of TM helices since some TM helices in multispanning TM proteins are not entirely exposed to the lipid bilayer. Instead, they are partially or completely shielded by other TM helices.<sup>11, 12, 17</sup> The neural network uses a local approach with a certain size of window to make predictions. ENSEMBLE takes the local average of the results of all three approaches to predict each residue's state, unlike consensus methods<sup>18, 19</sup> which simply take a majority vote from a pool of selected algorithms.

However, ENSEMBLE must make *ad hoc* compromises in the presence of loop clashes, or helix deletions. In addition, ENSEMBLE combines only advanced approaches (HMM and neural networks) whereas some previous consensus methods also use some simple methods such as TopPredII.<sup>20, 21</sup>

### 2.4.2 Phobius

Current TM protein topology prediction methods are effective at discerning globular proteins from TM proteins and transit peptides from TM helices.<sup>2, 4, 22</sup> However, discernment between signal peptides and signal anchors (TM helices) is far from ideal.<sup>2, 4, 5, 11, 12, 22</sup> Lao *et al.* pointed out that the presence of signal peptides significantly degrades the performance of TM protein topology prediction methods.<sup>22</sup> SignalP<sup>23</sup> is an HMM which discriminates signal peptides from signal anchors. Recently, a method called Phobius combined TMHMM with SignalP to try to overcome the problem.

The rationale behind Phobius is that if a TM protein contains a signal peptide that is predicted by SignalP, the sidedness of the N-terminal of the mature protein will be predicted correctly as well. Phobius makes some small modifications on both TMHMM and SignalP, and incorporates some newly trained transition probabilities to join the two models together. In essence, with a transition from the last state of SignalP to the outer loop state in the TMHMM, Phobius connects two HMMs sequentially, one for signal peptide prediction, and one for the following TM helix prediction.

### 2.4.3 Multiple sequence alignment information in new methods

Evolutionary information is also being incorporated into TM protein topology prediction via multiple alignments. ENSEMBLE uses sequence profile information from multiple sequence alignments in training all predictors. More specifically, it modifies the emission probabilities of an HMM by the use of a position specific score matrix obtained from multiple sequence alignments. This seems to help to improve TM protein topology prediction accuracy. Phobius also makes attempt to include homolog information in the prediction. PRODIV-TMHMM calculates the geometric mean of the joint probability of the aligned sequences. This makes the amino acids in each column of the profile be emitted by the same state. The authors noted that the prediction accuracy of PRODIV-TMHMM on a set of multispanning sequences increased by approximate 10% compared with methods based on single sequences. Nevertheless, the authors also indicated that the performance of a profile-based HMM method depends on the quality of the profile.

To incorporate multiple sequence alignment information into TM protein topology prediction is not a new idea. PHDhtm<sup>24</sup> has implemented this approach, and achieves better performance than methods with no multiple sequence alignment information. However, recently the author of PHDhtm also claimed that membrane helices are not entirely conserved among species. The divergence could cause the methods which force maintenance of this property to fail.<sup>5</sup>

### 2.4.4 Helix length

Most methods (including TMHMM 2.0) do not model helices longer than 35 residues. However, helix lengths longer than 35 residues have been seen in high- resolution structures of TM proteins.<sup>21, 25</sup> ENSEMBLE has addressed this issue by making the maximum length unbounded. Phobius still has not.

In addition, prediction preference for certain TM segment lengths is shown in methods such as TMHMM2.0 and PHDhtm.<sup>11, 12, 22</sup> Käll *et al.* claimed that the bias towards certain length of TM helix resulted from overtraining.<sup>12</sup> Phobius seems to rectify this problem by training on nearly twice as many sequences as TMHMM 1.0. ENSEMBLE avoids this overtraining by postprocessing the prediction results with a dynamic programming algorithm.

### 2.4.5 Current programs do not incorporate functional domains

HMMTOP 2.0<sup>9</sup> added some preliminary experimental information (including pattern predictors) on top of the HMMTOP 1.0 to help improve prediction accuracy. It allows the user to localize one or more sequence segments in any of the five structural regions used in HMMTOP. Möller *et al.* also suggested using additional information such as protein domains or post-translational modifications when the prediction from TMHMM is in doubt.<sup>1</sup> However, information on protein domains or post-translational modifications has not been automatically implemented into any of current programs.

### 3. Adding Functional Domains to TMHMM to Improve the Prediction Accuracy

Here, we introduce AHMM, our extension of TMHMM. It uses a probabilistic technique to incorporate pattern and domain predictors externally into TMHMM's HMM by adjusting the probabilities of certain topologies at certain positions in a sequence. We hypothesize that this incorporation will improve the prediction accuracy of TMHMM.

#### 3.1 Method

There are exponentially many state paths  $\pi$  through the hidden Markov model that correspond to a given sequence  $x$ . We use the Viterbi algorithm to find the most probable state path  $\pi^*$  for a given sequence, which maximizes  $P(x, \pi)$  over all paths.

We have changed the way TMHMM computes the Viterbi probability of the possible topologies of an input sequence, by taking advantage of the presence of sequence signatures and predicted domains in the sequence. We boost the probability of topologies that predict common internal functional domains as internal, and common external functional domains as external to the membrane, and decrease the probability of other topologies accordingly. The functional domains we use are described in detail in Section 3.2.

For a signature, the topology probabilities are modified only at its start position. For a domain, the topology probabilities are modified at both the start position and end position of the domain.

Our augmented model uses a technique first implemented in the gene finder GenomeScan<sup>14</sup> to modify the HMM probabilities when a signature or predicted domain is encountered. Consider a sequence  $x$  that has a signature  $H$  typically found internal to the membrane. Suppose  $P_H$  is a supplied estimate of the probability that the signature is found internally. Let  $\Phi_H$  be the set of all topologies through the HMM where  $H$  is found internally, and  $P(\Phi_H)$  be the probability given  $\Phi_H$  by the HMM: it is the sum of the probabilities of the topologies  $\pi$  in  $\Phi_H$ , where  $P(\pi, x)$  is the probability of a given topology  $\pi$ , as calculated by the Viterbi decoding algorithm. Following the procedure initiated in GenomeScan, we boost the probability of the topologies in  $\Phi_H$  by

the factor  $\frac{P_H}{P(\Phi_H)} + (1 - P_H)$ , which is always greater than 1, and reduce the

probability of the other topologies, by multiplying their probability with  $(1 - P_H)$ . Specifically, after our changes, if we are given a signature  $H$ , then the posterior probability of a topology  $\pi$  is:

$$P(\pi, x | H) = \begin{cases} \left( \frac{P_H}{P(\Phi_H)} + (1 - P_H) \right) \bullet P(\pi, x), & \text{if } \pi_i \in \Phi_H \\ (1 - P_H) \bullet P(\pi, x), & \text{if } \pi_i \notin \Phi_H \end{cases}$$

This change to the probabilities is easily incorporated into the decoding process, with negligible increase in runtime.

For example, from position 240 to position 440 of sequence ENVZ\_ECOLI, there exists a predicted histidine kinase domain. Such domains are typically found internal to membranes, and this is in fact the correct topology for this protein. TMHMM predicts this region as external, and gives the wrong prediction. However, AHMM boosts the probability of topologies that are internal at both position 240 and 440, using the first part of the formula, and lowers the probability of topologies that are external at either of the two positions by using the second part of the formula. With this change, AHMM gives the correct prediction (Figure 2).

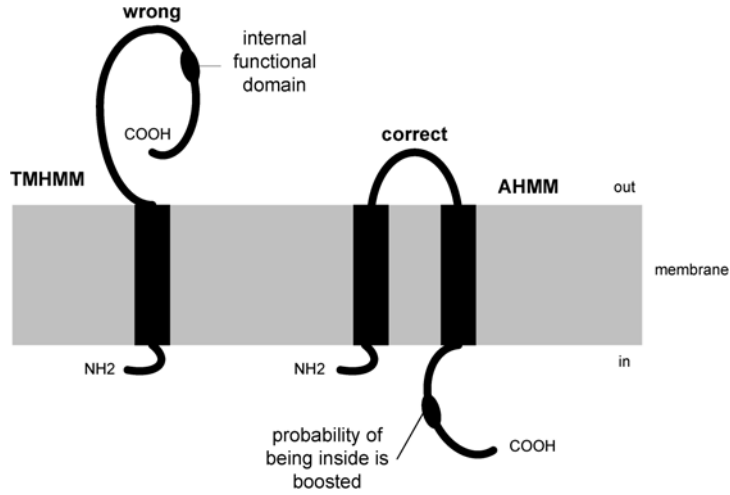


Figure 2: Topologies of ENVZ\_ECOLI predicted by TMHMM and AHMM respectively.

### 3.2 Definition of Pattern and Domain Predictors

A particular cluster of amino acid types in a protein sequence is known as a pattern, motif, signature, or fingerprint. It represents a conserved region of several proteins. In this paper, we use “signature” to emphasize a PROSITE<sup>26</sup> specific pattern versus its consensus pattern.

Domains refer to functional or structural domains that are not detected by patterns because of their extreme sequence divergence. PROSITE identifies domains with

position specific score matrices (PSSM<sup>7</sup>, also known as profiles). We use the term “functional domains” in this paper to refer to PROSITE signature and domain predictors.

### 3.3 Selection of Pattern and Domain Predictors

We use a computational approach to choose specific signatures and domains that are located preferentially internal or external to the membrane. We identify them in a three-step process.

First, we use `ps_scan`<sup>27</sup>, a perl program found in the PROSITE package, to find PROSITE signatures and domains in each training sequence with profile cut-off level  $L = 0$  (trusted cut-off for positive matches). Next, for each PROSITE signature or domain detected in the training sequences, we check which side of the membrane it falls upon, and how many non-redundant sequences contain it. If a signature or domain appears exclusively on one side of the membrane, and at least three times, we select it for further test. Finally, we incorporate all the signatures and domains selected into the Viterbi algorithm using the probabilistic method described above, and use the augmented Viterbi algorithm to re-predict the topology of the training sequences; we exclude any signatures and domains that cause a prediction error.

The remaining signatures and domains are then used to predict the topology of test sequences. In this experiment, we arbitrarily set  $P_H$  to 0.6 because we do not know its true value for an arbitrary predictor. The minimum number of times a signature or domain must appear, which in our case is three, is also empirically chosen; we expect it will increase as we have more sequences of known topology.

## 4. Experimental Results

We conducted the following experiments to test the robustness of AHMM as well as its sensitivity and specificity on helix and sidedness prediction.

### 4.1 Data Sets

We obtained our data set from three sources: the TMHMM training set<sup>8</sup>, the non-D trust level sequences from the collection of Möller *et al.*<sup>28</sup> and sequences from 3D\_Helix and 1D\_Helix sets of the Membrane Protein Topology (MPtopo) database<sup>25</sup> of May 17, 2005. Most of them have experimentally known topology. We excluded all sequences from the TMHMM training set and the non-D trust level sequences of Möller *et al.* collection that are present in MPtopo sets.

The TMHMM training set includes both eukaryotic, prokaryotic and organelle TM proteins. From the Möller *et al.* collection, we excluded organelle membrane proteins (due to the annotation issue) and incompletely annotated ones. We filtered the data set following the approach of Hobohm *et al.*<sup>29</sup>, which guarantees no more than 30% identity among any sequences in the test set; this was done using the needle program from the EMBOSS 3.0.0<sup>30</sup> package. This left us with 245 sequences.



The topology prediction accuracy (the percentage of correctly predicted complete topologies) for TMHMM 2.0 on the 245 proteins is approximately 55%.

#### 4.2 Robustness test for AHMM

In order to test the robustness of the method, we re-sampled and evaluated the 245 sequences twenty times. That is, we randomly selected 165 sequences from the 245 sequences as training sequences and used the rest 80 sequences as test sequences. Then, we conducted the computational selection of signatures and domains from the training sequences, using PROSITE<sup>31</sup> data file release 19.0 of 26-Apr-2005 with profile cut-off level  $L = 0$ , and tested them on the test sequences. We repeated this twenty times. We evaluated the performance of AHMM by looking at both the per-residue level, which computes the fraction of amino acids that are predicted in the same position as in the reference topology, and the whole topology level, where a prediction is counted as correct if the N-terminus sidedness is correct and if every helix in the reference topology overlaps in at least five amino acids with the corresponding predicted helix. The test results are shown in Table 1. The comparison between AHMM and TMHMM is made only on the test sequences with identified PROSITE functional domains.

Table 1. Comparison between AHMM and TMHMM at both per-residue and per-sequence levels for test sequences with functional domains from 20 resamplings. The first group of columns shows prediction accuracy of the three methods at the per-residue level. The second group of columns considers the sequence as a whole and shows comparison between AHMM and TMHMM 2.0 only. The final line gives the weighted average of the twenty tests, with each row weighted proportional to the number of amino acids or sequences found to contain a functional domain.

run	sequences	per-residue accuracy			per-sequence accuracy			
		TMHMM2.0	TMHMM1.0	AHMM	no change	made correct	made incorrect	functional domains
1	7	0.7572	0.7620	0.8705	6	1	0	7
2	2	0.0611	0.0571	0.9817	0	2	0	2
3	6	0.7322	0.4870	0.9859	3	3	0	6
4	2	0.7580	0.7585	0.7585	2	0	0	2
5	2	0.6133	0.6124	0.9845	1	1	0	2
6	4	0.7528	0.5245	0.9955	2	2	0	4
7	4	0.4333	0.4312	0.9744	2	2	0	4
8	1	0.0844	0.3511	0.9756	0	1	0	1
9	4	0.7834	0.7853	0.9857	3	1	0	4
10	2	0.9764	0.4532	0.9924	1	1	0	2
11	2	0.6605	0.0485	0.9838	0	2	0	2
12	3	0.9838	0.5625	0.9967	2	1	0	3
13	6	0.5191	0.5088	0.8608	4	2	0	6
14	3	0.9901	0.9901	0.9901	3	0	0	3
15	4	0.7355	0.7362	0.9944	3	1	0	4
16	5	0.4534	0.4369	0.9835	3	2	0	5
17	4	0.8128	0.8141	0.9803	3	1	0	4
18	2	0.9973	0.9973	0.9973	2	0	0	2
19	5	0.9581	0.9796	0.9796	5	0	0	5
20	3	0.9632	0.6956	0.9909	2	1	0	3
wavg		0.7311	0.6245	0.9561		33.8%	0 %	

At the per-residue level, we computed the weighted average for sequences with functional domains at each run and over all twenty runs (Table 1). We calculated the mean of the differences between AHMM and TMHMM for 20 runs and its confidence interval (C.I.) at both the per-residue and whole sequence levels. On average, AHMM predicted correctly 22.5% more residues, with 95% C.I. = (22.09%, 22.91%) and 38.96% more topologies correct, with 95% C.I. = (24.11%, 53.82%) than TMHMM, for sequences that include predicted functional domains. AHMM also has smaller standard deviation (SD) than TMHMM (data not shown) for prediction at the per-residue level.

We conducted statistical tests to test the results from our 20 runs of resampling at the per-residue level, to test the significance of the difference between AHMM and TMHMM. Since the population of TM proteins might not be normally distributed, we used the non-parametric sign and Wilcoxon Matched-Pairs Signed-Ranks test to compare the weighted averages between TMHMM and AHMM for sequences with functional domains, using SPSS 13, which found that the results are significant at the 0.01 probability level, and that AHMM gives better results at the per-residue level than both versions of TMHMM for sequences with functional domains.

#### **4.3 Sensitivity and Specificity of TMHMM and AHMM on Helix and Sidedness Prediction**

In addition to the experiments above, we further tested the sensitivity and specificity of TMHMM2.0 and AHMM on helix and sidedness prediction on test sequences with functional domains from the twenty-time resampling (Table 2). We define the sensitivity to be the fraction of true helices/sidedness correctly identified, and the specificity to be the fraction of predicted helices/sidedness that are true helices/sidedness.

Table2. Comparison of mean of weighted average and standard deviation of sensitivity and specificity between TMHMM2.0 and AHMM on helix and outsidedness prediction for test sequences with functional domains from 20 resamplings.

run	SEH		SPH		SEO		SPO	
	TMHMM	AHMM	TMHMM	AHMM	TMHMM	AHMM	TMHMM	AHMM
mean	.9205	1.000	.9080	.9742	.8156	.9887	.6569	.9630
SD	.0974	.0000	.1258	.0779	.2882	.0108	.2999	.0779

SEH: sensitivity for helix at per-sequence level

SPH: specificity for helix at per-sequence level

SEO: sensitivity for outsidedness at per-residue level

SPO: specificity for outsidedness at per-residue level

We computed the weighted average of the performance for all 20 runs; the mean of weighted average improvement of AHMM over TMHMM for sequences with PROSITE functional domains is 7.95% in sensitivity (95% C.I. = (3.39%, 12.51%)) and 6.63% in specificity (95% C.I. = (1.27%, 11.98%)) for helix prediction, and 17.31% in sensitivity (95% C.I. = (4.15%, 30.47%)) and 30.61% in specificity (95% C.I. = (16.24%, 44.98%))

for sidedness prediction. AHMM has smaller standard deviation than TMHMM for all the tests.

## 5. Discussions and Conclusion

AHMM can improve TM protein topology prediction accuracy at both per-residue and per-sequence levels. Furthermore, it improves both sensitivity and specificity on helix and sidedness prediction. It fixes errors in the prediction of the orientation of the membrane protein, and also fixes helix number errors. Following are some discussions on  $P_H$  of the formula, the scope of AHMM, and functional domains.

### 5.1 The Value of $P_H$

There is subjectivity in the choice of the value of the prior probability  $P_H$  used for functional domains in the HMM extension formula. We set  $P_H = 0.6$  for all functional domains incorporated into AHMM, and also tried  $P_H = 0.9$ , which made no difference compared to 0.6. This might suggest that the functional domains in the experiment are fairly specific.

### 5.2 The Scope of AHMM

Patterns and domains studied in AHMM were derived from native integral membrane protein, and as such, AHMM is not valid for predicting the topology of artificial membrane proteins, which are used to study membrane protein biogenesis<sup>32</sup> and design artificial membrane protein receptors<sup>33</sup>. By redistributing positively charged amino acids in the loops, the topologies of artificially engineered membrane proteins are altered, so functional domains that typically reside on one side of the membrane could end up on the different side of the membrane. For example, the fusion protein LEP-LEP, which is constructed from *E.coli* inner membrane leader peptidase (LEP), demonstrates this limitation.

LEP has two TM segments and a N<sub>out</sub>-C<sub>out</sub> topology (both N- and C-terminus reside on the periplasmic side of the TM protein). The loop containing the PROSITE signature SPASE\_I\_3 (Signal peptidases I signature 3) of LEP is on the external side of the membrane. However, by introducing 3 lysines (K) to the second loop of LEP-LEP, the topology changes to one where the loop containing this signature now appears on the internal side of the membrane.<sup>32</sup>

### 5.3 Functional Domains and Prediction Accuracy

Using the Sequence Retrieval System SRS Release 7.1.1, there are 29488 entries in UniProtKB/Swiss-Prot<sup>34</sup> Release 47.7 of 16-Aug-2005 and 175405 entries in UniProtKB/TrEMBL<sup>34</sup> Release 30.7 of 16-Aug-2005 with keyword “transmembrane” search. We found 6.5% of Swiss-Prot entries and 0.65% of TrEMBL entries having

signatures and domains extracted from the 245 sequences without counting the amino acid RICH domains.

Only a fraction of sequences have PROSITE functional domain predictors. As more and more sequences with known topology are available, we expect that more useful predictors (including those which were filtered out at present) could be found in the future. We also would expect that as more and more signatures and domains are available, the prediction accuracy would be further improved with more potential predictors.

In summary, even with the current updates in TM protein topology prediction, incorporation of protein functional domain information remains a viable approach that could still be incorporated on top of an HMM, such as Phobius, to assist prediction. For example, we may search a query sequence for possible functional domain predictor and add available functional domain predictor information into its HMM to make better predictions. We could also incorporate domains selected in SMART<sup>35</sup> by Bernsel and von Heijne<sup>36</sup> to AHMM. The prediction accuracy can be further improved if the domains are very specific. The AHMM server is available at <http://genome.math.uwaterloo.ca/ahmm/>.

### Acknowledgments

We especially thank Broňa Brejová from the University of Waterloo to help set up the AHMM server and the helpful discussion with her. We also thank Ming Li, John Tsang, Mike Hu, and Tomáš Vinař from the University of Waterloo; Peter Ehlers and Tak Shing Fung, from the University of Calgary for their helpful discussions; and Michel Dominguez from Caprion for his opinion on functional domain sidedness. The research of all authors was supported by the Natural Science and Engineering Research Council of Canada, and the research of the last author was also supported by the Human Frontier Science Program.

### References

1. Möller S, Croning MDR, and Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17 (7): 646–653, 2001.
2. Krogh A, Larsson B, Heijne GV and Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305: 567–580, 2001.
3. Tusnady GE and Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17(9): 849–850, 2001.
4. Chen CP, Kernytsky A, and Rost B. Transmembrane helix predictions revisited. *Protein Sci.*, 11: 2774–2791, 2002.
5. Chen CP and Rost B. State-of-the-art in membrane protein prediction. *Applied Bioinformatics*, 1(1): 21–35, 2002.
6. von Heijne G. The Distribution of Positively Charged Residues in Bacterial Inner Membrane Proteins Correlates With the Trans-Membrane Topology. *EMBO J.*, 5: 3021-3027, 1986.

7. Durbin R, Eddy S, Krogh A and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.
8. Sonnhammer ELL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology*, 175–182, AAAI Press, 1998.
9. Tusnady GE and Simon I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.*, 283: 489–506, 1998.
10. Tusnady GE and Simon I. Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.*, 41: 364–368, 2001.
11. Martelli PL, Fariselli P, and Casadio R. An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, 19 (Suppl.1): i205–i211, 2003.
12. Käll L, Krogh A, and Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, 338:1027–1036, 2004.
13. Viklund H and Elofsson A. Best  $\alpha$ -helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, 13(7): 1908–17, 2004.
14. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Res.*, 11(5): 803–806, 2001.
15. Korf I, Flicek P, Duan D and Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 (Suppl. 1): S140–S148, 2001.
16. Brejová B, Brown DG, Li M and Vinař T. ExonHunter: A Comprehensive Approach to Gene Finding. *Bioinformatics*, 21(Suppl 1): i57–i65, June 2005.
17. Mueckler M. and Makepeace C. Analysis of Transmembrane Segment 8 of the GLUT1 Glucose Transporter by Cysteine-scanning Mutagenesis and Substituted Cysteine Accessibility. *J. Biol. Chem.*, 279: 10494–10499, 2004.
18. Nilsson J, Persson B, and von Heijne G. Consensus predictions of membrane protein topology. *FEBS Letters*, 486: 267–269, 2000.
19. Ikeda M, Arai M, Okuno T, and Shimizu T. TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucl. Acids. Res.*, 31(1): 406–409, 2003.
20. Claros MG and von Heijne G. TopPred II: An improved software for membrane protein structure predictions. *CABIOS Appl. Notes*, 10(6): 685–686, 1994.
21. Chen CP and Rost B. Long membrane helices and short loops predicted less accurately. *Protein Sci.*, 11: 2766–2773, 2002.
22. Lao, DM, Arai M, and Shimizu T. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics*, 18 (12): 1562–1566, 2002.
23. Nielsen H and Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, 122–130, AAAI Press, 1998.
24. Rost B, Fariselli P and Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, 5(8): 1704–1718, 1996.
25. Jayasinghe S, Hristova K, and White SH. MPtopo: A database of membrane protein topology. *Protein Sci.*, 10: 455–458, 2001.
26. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3: 265–274, 2002.
27. Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*, 1(2): 107–108, 2002.
28. Möller S, Kriventseva EV, and Apweiler R. A collection of well characterized integral membrane proteins. *Bioinformatics*, 16(12): 1159–1160, 2000.

29. Hobohm U, Scharf M, Schneider R, and Sander C. Selection of representative protein data sets. *Protein Science*, 1, 409-417, 1992.
30. Rice P, Longden I, and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6): 276—277, 2000.
31. Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A. Recent improvements to the PROSITE database. *Nucleic Acids Res.*, 32: 134-137, 2004.
32. Gafvelin G and von Heijne G. Topological “frustration” in multispanning *E.coli* inner membrane proteins. *Cell*, 77: 401–412, May 6, 1994.
33. Pule M, Finney H and Lawson A. Artificial T-cell receptors. *Cytotherapy*, 5(3): 211-26, 2003.
34. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). *Nucl. Acids. Res.*, 33:D154-159, 2005.
35. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, and Bork P. SMART 4.0: Towards genomic data integration. *Nucleic Acids Res.*, 32: D142 – D144, 2004.
36. Bernsel A and von Heijne G. Improved membrane protein topology prediction by domain assignments. *Protein Sci.*, 14:1723 – 1728, 2005.



**Emily W. Xu** received her B. S. degrees in Biology and Computer Science from the University of Regina in 1998 and 1999 respectively. She received her M.Math. in Bioinformatics from the School of Computer Science, University of Waterloo in 2004. She is currently working at the Sun Center of Excellence for Visual Genomics, University of Calgary.



**Paul Kearney** received his undergraduate degree in mathematics from Queen's University in 1992, and his Ph.D. in computer Science from the University of Toronto in 1997. Previously, Paul was Vice-President of Research at Bioinformatics Solutions Inc.. He was also the Director of the Bioinformatics Program at the University of Waterloo, and most recently a Bioinformatics Research Consultant for Bristol-Meyers Squibb. Currently he is the executive director of the bioinformatics group at Caprion Pharmaceuticals Inc. and an adjunct assistant professor at the University of Waterloo.



**Daniel Brown** received his undergraduate degree in mathematics with computer science from the Massachusetts Institute of Technology in 1995, and his Ph.D. in computer science from Cornell University in 2000. He then spent a year at the Whitehead Institute/MIT Center for Genome Research, working on the Human and Mouse Genome Projects. Since 2001, he has been an assistant professor in the School of Computer Science at the University of Waterloo.