THE USE OF FUNCTIONAL DOMAINS TO IMPROVE TRANSMEMBRANE PROTEIN TOPOLOGY PREDICTION

EMILY W. XU[†]

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

DANIEL G. BROWN

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

PAUL KEARNEY

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

Transmembrane proteins affect vital cellular functions and diseases, 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 [2,7], the current state-of-the-art method, has less than 52% accuracy on the prediction of transmembrane proteins collected by Moller *et al*. [1, 4]. Based on the assumption that there are functional domains that occur preferentially internal or external to the membrane, we have extended the model of TMHMM, incorporating functional domain information into it, using an approach originally used in gene finding [8]. Results show that our Augmented HMM, or AHMM, is better than TMHMM on both helix and sidedness prediction. This improvement is verified by both statistical tests as well as sensitivity and specificity studies. As prediction of functional domain improves, our system's prediction accuracy will likely improve as well.

1. Introduction

About 20% to 25% of proteins are membrane proteins [1, 2, 3]. These include both integral (transmembrane or TM) and peripheral membrane proteins. There are two known classes of integral membrane proteins: those with α-helical structure and those with β-barrel structure. Alpha-helical membrane proteins are the predominant type; thus, they are the focus of our modeling. Improvement in the sidedness (orientation) prediction of TM proteins remains a priority since the prediction accuracy for sidedness is even lower than the prediction accuracy for helix location. Figure 1 illustrates a model for the topology of a hypothetical TM protein**.**

[†] For correspondence. E-mail ewxu@ucalgary.ca or ewxu@monod.uwaterloo.ca.

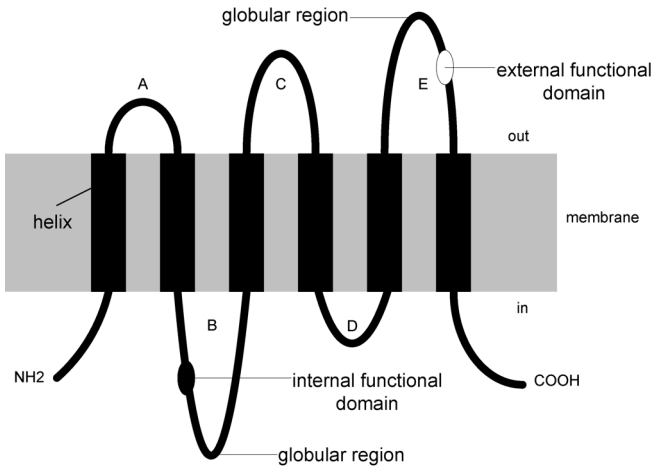


Figure 1: A model to illustrate the topology of a hypothetical transmembrane protein with six helices and three extracellular loops (A, C and E) and two intracellular loops (B and D). Both its N- and C-terminus are internal to the membrane.

2. Computational Prediction of TM Protein Topology

The basic problem in TM protein topology prediction is to find the location, number and orientation of the membrane spanning segments (helices). There are two kinds of approaches to predict TM protein topology: local and global. The local approach uses a sliding window to calculate scores of certain scales, for example, the hydrophobicity scale. The main weakness of the local approach is the lack of specificity. Compared with the global approach, for the same sequences, the local approach has higher false positives and lower true positives upon prediction. On the other hand, the global approach examines sequences as a whole and does not set any empirical cutoffs and rules [1]. The canonical example of the global approach is hidden Markov model (HMM)-based prediction methods.

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-Inside Rule) 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

An HMM is a probabilistic and generative model. It is a doubly embedded stochastic process. One is hidden and the other is observable. Only the sequence of output symbols is observed, but the states remain hidden [11]. HMMs are "decoded" when one uses an algorithm to predict the state sequence that give rise to a given output sequence. The most popular decoding algorithm is the Viterbi algorithm. The states of this sequence correspond to the features annotated by the HMM on a given sequence. HMMs easily model both sequence distributions and distributions in the lengths of features of sequence elements, such as loops and helices.

2.3 Review of Existing HMM Models

2.3.1 HMM for topology prediction (HMMTOP)

HMMTOP [5] contains five types of states: inside loop, inside tail, membrane helix, outside tail and outside loop. Two tails between adjacent helices form a short loop and tail-loop-tail form a long loop. Tusnady and Simon 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) [5]. Consequently, they modeled the length of a tail of 1–15 residues. The design of HMMTOP's model is similar to TMHMM on helix and loop structure.

2.3.2 Transmembrane HMM (TMHMM)

TMHMM contains seven different types of states: one for the helix core, 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. This slight expansion in number of state varieties may give greater sensitivity to the variation of the amino acid compositions than five states [6]. Because of the limited number of proteins of known topology for training, for each state type, there are a number of states joined with its emission probability to avoid overfitting. The transition matrix is a sparse matrix. There is no difference in the models of TMHMM 1.0 and TMHMM 2.0 (collectively known as TMHMM in this paper), but TMHMM 2.0 was retrained on the same data set. TMHMM 2.0 has higher prediction accuracy over TMHMM 1.0 [1]. TMHMM models helices 15–35 residues long, the longest among current HMM models. Sonnhammer *et al*. set the loop ladder length to 10 amino acids long based on observation made during prediction [7].

2.3.3 Current programs do not incorporate functional domains

HMMTOP 2.0 [5] 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. Moller *et al.* also suggested using additional information such as

protein domains or post-translational modifications when the prediction from TMHMM is in doubt [1]. However, information on protein domains or post-translational modifications has not been automatically implemented into any of these programs.

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

Here, we introduce AHMM, which incorporates pattern and domain predictors externally into TMHMM, to adjust the probabilities of certain topologies at certain positions in a sequence. We predict that incorporation of internal and external functional domains can augment prediction accuracy of TMHMM.

3.1 Viterbi Algorithm

There are exponentially many state paths π corresponding to a given sequence x. We use the Viterbi algorithm [13] to find the most probable state path π_i to be the optimal state path for a given sequence, i.e. the state path that maximizes $P(x,\pi)$.

3.2 Method

We have changed the way TMHMM computes the Viterbi probability of the possible topologies of an input sequence, by taking advantage of signature and domain predictors found in the sequence. We boost the probability of topologies that predict internal functional domains as internal, and external functional domains as external to the membrane. We decrease the probability of other topologies accordingly. Our functional domains will be described in full detail in Section 3.3.

For a signature, the probability of topologies is modified only at its start position. For a domain, the probabilities of topologies 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 [8] to modify the HMM probabilities when a signature or domain predictor is encountered. For example, for an internal signature: for sequence *x*, $P(\pi_i, x \mid H)$ is the probability of topology π_i at the position of the signature given that it is internal. H denotes the signature is internal. P_H denotes the probability that the signature is internal. Φ_H denotes the set of topologies that identify the protein as internal at the position of the signature. $P(\Phi_H)$ denotes the unaugmented probability that the site is predicted to be internal at the position of the signature. $P(\pi_i, x)$ is the probability of topology π_i, as calculated by Viterbi algorithm for decoding HMM sequences. Since $P(\Phi_H) < 1$, $\frac{P_H}{P(\Phi_H)} > P_H$, therefore $\frac{P_H}{P(\Phi_H)} + (1 - P_H)$ is always greater than 1, and $(1 - P_H)$ is always less than 1. Specifically,

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

For example, from position 240 to position 440 of sequence ENVZ_ECOLI, there exists a HIS_KIN (Histidine kinase domain profile) domain. It is supposed to be internal. Since TMHMM predicts this region as external, it gives the wrong prediction. However, AHMM boosts the probability for topologies being internal at both position 240 and 440 by using the first part of the formula $\left(\frac{P_H}{P(\Phi_H)} + (1 - P_H) \right) \bullet P(\pi_i, x)$, if $\pi_i \in \Phi_H$). On the other hand, it lowers the probability for topologies being external at the two positions by using the second part of the formula ($(1 - P_H) \bullet P(\pi_i, x)$, if $\pi_i \notin \Phi_H$). It gives the correct prediction (Figure 2).

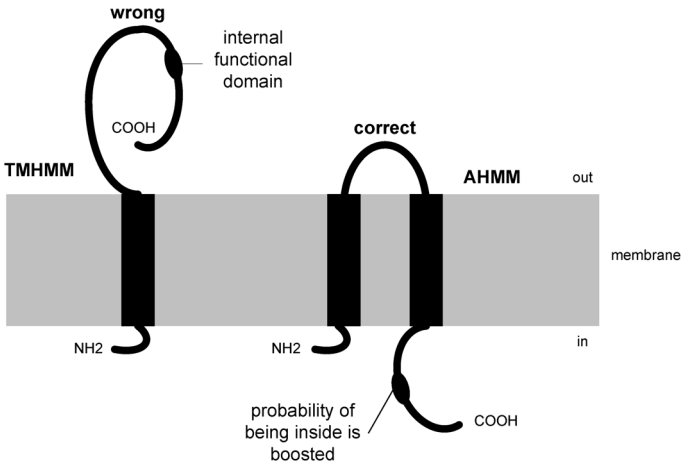


Figure 2: Topologies of ENVZ_ECOLI predicted by TMHMM and AHMM respectively.

3.3 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 [9]. It represents a conserved region of several proteins. In this paper, we use "signature" to emphasize a PROSITE 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 [13], also known as profiles). We use the term "functional domains" in this paper to refer to PROSITE signature and domain predictors.

3.4 Selection of Pattern and Domain Predictors

We use computational approach to choose specific signatures and domains that are located preferentially internal or external to the membrane. We identified them as follows:

1. Use ps_scan, a perl program in PROSITE, to run the training sequences against PROSITE database to obtain the corresponding signature(s) and/or domain(s) for each sequence with profile cut-off level L = 0 (trusted cut-off for positive matches).
2. For each PROSITE signature or domain detected in the training sequences, check to see where it resides with respect to the membrane and how many non-redundant sequences contain it.
3. If a signature or domain appears exclusively on one side of the membrane at least twice, it is selected for further test.
4. Incorporate all signatures and domains selected from step 3) into Viterbi algorithm and exclude all signatures and domains that cause an error during the prediction on training sequences. The remaining signatures and domains are the potential predictors. They are then tested on the test sequences. In this experiment, we arbitrarily set P_H as 0.6 because we do not know its true value.

4. Experimental Results

After computationally extracting functional domains from training data and testing on the test data, we also conducted experiments to test the robustness of AHMM as well as its sensitivity and specificity on helix and sidedness prediction.

4.1 Data Sets

We used two data sets for our experiments. One is a 157 protein data set from the TMHMM training set [7] and the other is a 72 protein data set from the collection of Moller *et al*.. They are all TM proteins with experimentally known topology.

The 157 protein data set is chosen as training data to extract potential pattern and domain predictors for AHMM. The data set includes both eukaryotic, prokaryotic and organelle TM proteins.

The test data is from the Moller *et al*. collection, though we excluded organelle and all membrane proteins that have not been completely annotated and those present in the 157 protein data set. Thus, only 72 protein sequences were used as test data.

The prediction accuracy (the percentage of correctly predicted sequences) for TMHMM on the 157 protein data set is approximately 79%, whereas on the 72 data set is 55.56%, or 40 out of 72 sequences.

4.2 Test for the Robustness of AHMM

We incorporated the potential signature and domain predictors extracted from the 157 sequences into Viterbi algorithm and tested them on the 72 sequences. With profile cut-off level L = 0, we found one sequence (CPXA_ECOLI) predicted wrongly by TMHMM 2.0 but correctly by AHMM.

In order to test the robustness of the method, we re-sampled and evaluated a total of 229 sequences (the 157 training plus 72 test sequences) twenty times at both amino acid level[1] and sequence level[2] . That is, we randomly selected 157 non-redundant sequences from the 229 sequences as training sequences and the rest as test sequences. Then, we conducted the computational selection of signatures and domains from the training sequences and tested them on the test sequences. We repeated this twenty times. Only the test results are shown below (Table 1). The comparison between AHMM and TMHMM is made only on the test sequences with identified PROSITE functional domains.

[1] The percentage of overlap of amino acids with the reference topology.
[2] For each helix in the reference topology, if at least 5 amino acids in the prediction overlap with it, we believe at sequence level the helix prediction is correct. If the N-terminus orientation is also correct, then the topology prediction is correct.

Table 1. Comparison between TMHMM and AHMM at amino acid level and at sequence level for test sequences with functional domains from 20 resamplings. Column TMHMM2.0 is the percentage of correctly predicted amino acids by TMHMM 2.0 over sequences with potential PROSITE functional domain predictors; similarly, column TMHMM1.0 is the percentage of correctly predicted amino acids by TMHMM 1.0 and column AHMM is the percentage of correctly predicted amino acids by AHMM.

run	72 (amino acid level)			72 (sequence level)				
	TMHMM2.0	TMHMM1.0	AHMM	same	better[1]	worse[2]	ISD[3]	# of seqs
1	0.8588	0.8547	0.9435	14	3	0	22	17
2	0.8266	0.7246	0.8654	10	2	0	12	12
3	0.8911	0.7923	0.9775	16	3	0	29	19
4	0.8605	0.8282	0.9775	13	4	0	26	17
5	0.9614	0.8856	0.9410	14	2	2	27	18
6	0.7429	0.7661	0.9819	7	5	0	20	12
7	0.8456	0.7391	0.9799	8	3	0	16	11
8	0.9083	0.9085	0.9754	12	1	0	20	13
9	0.9663	0.9660	0.9660	15	0	0	17	15
10	0.7730	0.7560	0.9563	11	4	0	21	15
11	0.8936	0.8894	0.9670	10	2	0	17	12
12	0.8291	0.7357	0.9401	12	4	0	23	16
13	0.9279	0.9253	0.9790	10	2	0	12	12
14	0.9454	0.8445	0.9813	12	2	0	20	14
15	0.8722	0.8369	0.9792	11	2	0	23	13
16	0.7557	0.7329	0.9818	7	4	0	15	11
17	0.8285	0.7584	0.9459	11	4	0	24	15
18	0.7136	0.5807	0.9382	6	4	0	18	10
19	0.8164	0.8125	0.9281	13	3	0	19	16
20	0.9749	0.9353	0.9744	16	0	0	18	16
wavg[4]	0.8607	0.8137	0.9587		0.19 01	0.007		

[1]better—number of sequences where TMHMM 2.0 predicted wrongly but AHMM predicted correctly

[2]worse—number of sequences where TMHMM 2.0 predicted correctly but AHMM predicted wrongly

[3]ISD—number of signatures and domains identified

[4]wavg (weighted average) = total number of correctly predicted amino acids / total number of amino acids of all sequences with functional domains.

Column "# of seqs" lists the actual number of sequences for comparison between TMHMM and AHMM at each run. The actual number of sequences for comparison depends on the number of sequences containing identified functional domains.

At amino acid 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.). We also calculated the mean of the differences between "better" and "worse" for 20 runs and its confidence interval. On average, AHMM predicted correctly 9.94% more with 95% C.I. = (6.43%, 13.45%) at amino acid level and 20.35% more with 95% C.I. = (14.43%, 26.27%) at sequence level than TMHMM for sequences with functional domains. AHMM also has smaller standard deviation (SD) than TMHMM (data not shown) for prediction at amino acid level. This result is verified by a four-time 5-fold cross-validation.

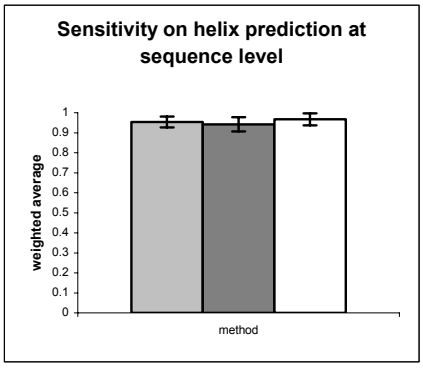
Only two sequences were predicted correctly by TMHMM but wrongly by AHMM in the twenty resamplings. This occurred because a particular signature (the EGF-like domain signature 2) appeared on the different side of the membrane in the test data than it was in the training data.

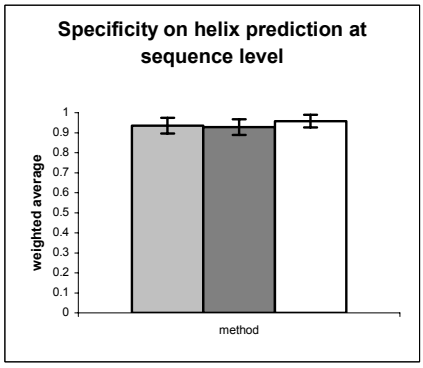
Functional domains for the above experiment were obtained from PROSITE release 18.9 of 4-Oct-2003 with profile cut-off level L = 0.

We conducted statistical tests to test the results from 20 runs of resampling at amino acid level. The hypothesis is that there is no difference between AHMM and TMHMM. Since the population of TM proteins might not be normally distributed, we conducted non-parametric tests, sign test and Wilcoxon Matched-Pairs Signed-Ranks Test over 20 runs to compare weighted averages between TMHMM and AHMM for sequences with functional domains. We ran all statistical tests with SPSS for UNIX release 6.1. 1-tail Ps of all the statistical tests are less than 0.01. This indicates that if the null hypothesis is true, the chance of getting such sample difference in Table 1 is $P < 0.01$. Therefore, we reject the null hypothesis and conclude that AHMM is better than both versions of TMHMM for sequences with functional domains.

4.3 Sensitivity[3] and Specificity[4] of TMHMM and AHMM on Helix and Sidedness Prediction

In addition to the experiments above, we further tested the sensitivity and specificity of TMHMM and AHMM on helix and outsidedness prediction on test sequences with functional domains from the twenty-time resampling (Figure 3).





[3] Sensitivity = true positives / (true positives + false negatives) (the number of correct predictions out of the reference number).

[4] Specificity = true positives / (true positives + false positives) (the number of correct predictions out of the total number of predictions).

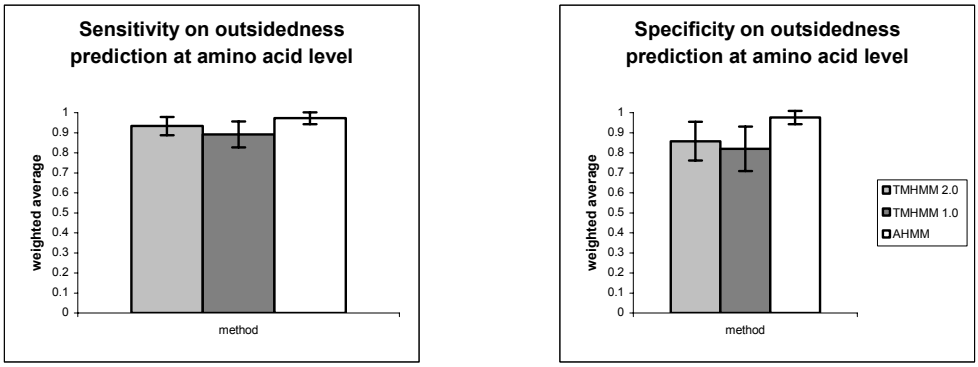


Figure 3. Comparison of weighted average and standard deviation of sensitivity and specificity between TMHMM and AHMM on helix and outsidedness prediction for test sequences with functional domains from 20 resamplings.

We conducted weighted average for each run as well as for all twenty runs. We calculated the mean of the differences between AHMM and TMHMM for 20 runs and its confidence interval. Results show that AHMM is 1.79% more sensitive with 95% C.I. = (0.53%, 3.04%) and 2.58% more specific with 95% C.I. = (1.49%, 3.68%) than TMHMM on helix prediction and 4.08% more sensitive with 95% C.I. = (2.34%, 5.82%) and 11.89% more specific with 95% C.I. = (7.18%, 16.61%) on sidedness prediction for sequences with PROSITE functional domains. Figure 3 illustrates that AHMM is especially more specific and sensitive than TMHMM on sidedness prediction. Except that AHMM has slightly bigger SD than TMHMM for sensitivity on helix prediction, AHMM has smaller SD for all the other tests.

5. Discussions and Conclusion

AHMM can improve TM protein topology prediction accuracy at both sequence level and amino acid levels. Furthermore, it improves both sensitivity and specificity on helix and sidedness prediction. It fixes not only sidedness errors, but also helix number errors. Sidedness and helix position are not two independent issues. Therefore, topology should be examined as a whole. 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 certain subjectivity in the choice of the value of the probability or weight P_H for functional domains in the GenomeScan formula. As mentioned earlier, we set $P_H = 0.6$ for all functional domains incorporated into AHMM. We 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

We also have an important observation on AHMM. Patterns and domains studied in AHMM were derived from native integral membrane proteins. Thus, AHMM is not valid for predicting artificial membrane proteins. By redistributing positively charged amino acids in the loops, the topologies of artificially engineered membrane proteins are altered. Functional domains reside on one side of the membrane could end up on the different side of the membrane. One example of the artificial membrane proteins is the fusion protein LEP-LEP, which is constructed from *E.coli* inner membrane leader peptidase (LEP).

LEP has two TM segments and a N_{out}-C_{out} toplogy (both N- and C-terminus reside on the cytoplasmic 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 2^{nd} loop of LEP-LEP, the mutant adopts "leave one out" topology and the loop containing signature SPASE_I_3 appears on the internal side of the membrane [10].

5.3 Functional Domains and Prediction Accuracy

Using the Sequence Retrieval System SRS Release 7.1.1, there are 23146 entries in Swiss-Prot [12] and 57496 entries in TrEMBL with keyword "transmembrane" search. We found 12% of Swiss-Prot entries and 4.7% of TrEMBL entries having signatures and domains extracted from the 229 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 would expect 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. With profile cut-off level L = 0, PROSITE release 18.9 of 4-Oct-2003 was compared with release 17.4 of May 2002. We found more functional domains (i.e. IG_LIKE Ig-like domain profile) and predicted one more sequence (MYP0_HUMAN) correctly.

Acknowledgments

We thank Ming Li, Brona Brejova, Tomas Vinar, John Tsang and Mike Hu from the University of Waterloo, and 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 second author was also supported by the Human Frontier Science Program.

References

1. Moller 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. Moller S, Kriventseva EV, and Apweiler R. A collection of well characterized integral membrane proteins. *Bioinformatics*, 16(12): 1159–1160, 2000.
5. 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.
6. Tusnady GE and Simon I. Topology of membrane proteins. *J. Chem. Inf. Comput. Sci*., 41: 364–368, 2001.
7. 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.
8. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Res.*, 11(5): 803–806. 2001.
9. 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.
10. Gafvelin G and Von Heijne G. Topological "frustration" in multispanning *E.coli* inner membrane proteins. *Cell*, 77: 401–412, May 6, 1994.
11. Rabiner, LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2): 257–286, Feb. 1989.
12. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M: The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids. Res*., 31: 365–370, 2003.
13. Durbin R, Eddy S, Krogh A and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.