Journal of Bioinformatics and Computational Biology

TOPTMH: Topology Predictor for Transmembrane α-Helices

REZWAN AHMED
Department of Computer Science & Engineering, University of Minnesota
Minneapolis, Minnesota 55455, USA
ahmed@cs.umn.edu

HUZEFA RANGWALA
Department of Computer Science, George Mason University
Fairfax, Virginia 22030, USA
rangwala@cs.gmu.edu

GEORGE KARYPIS
Department of Computer Science & Engineering, University of Minnesota
Minneapolis, Minnesota 55455, USA
karypis@cs.umn.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Alpha-helical transmembrane proteins mediate many key biological processes and represent 20%–30% of all genes in many organisms. Due to the difficulties in experimentally determining their high-resolution 3D structure, computational methods to predict the location and orientation of transmembrane helix segments using sequence information are essential. We present, TOPTMH a new transmembrane helix topology prediction method that combines support vector machines, hidden Markov models, and a widely-used rule-based scheme. The contribution of this work is the development of a prediction approach that first uses a binary SVM classifier to predict the helix residues and then it employs a pair of HMM models that incorporate the SVM predictions and hydropathy-based features to identify the entire transmembrane helix segments by capturing the structural characteristics of these proteins. TOPTMH outperforms state-of-the-art prediction methods and achieves the best performance on an independent static benchmark.

Keywords: Membrane Protein; Secondary Structure; Classification; Support Vector Machines; Hidden Markov Model.

1. Introduction

Transmembrane helical (TMH) proteins play a crucial role in several cellular functions, such as cell-to-cell communication, cell signaling, and transportation of ions and small molecules[3], and are of key interest to the pharmaceutical industry as approximately 50% of all existing drugs are targeting transmembrane proteins[20]. Experimental determination of the three dimensional structure of TMH

proteins is challenging, because they are difficult to crystallize and are too large for NMR studies[27]. As such, TMH proteins represent only 1% of known 3D protein structures[2], even though they account for about 20%–30% of the encoded proteins in several organisms[37]. Computational methods that can accurately predict the topology of TMH proteins by identifying the helical segments along with their orientation relative to the interior of the cell (also called cytoplasm) are currently the only high-throughput approach to characterize structural aspects of transmembrane proteins.

Over the years, a number of different methods have been developed for predicting the topology of TMH proteins. In general, these methods need to predict the following items: (i) the type of each residue (e.g., helix, loop, etc.), (ii) the TMH segments, and (iii) their orientation. The various methods developed differ on the number of distinct steps that they use to predict the above items. Some methods predict each item individually, others utilize predictors that combine some of these steps, and others predict all three items in a single step. The residue types are predicted by either relying on the fact that membrane segments contain primarily hydrophobic residues (e.g., TopPred[35]) or by utilizing machine-learning approaches (e.g., neural networks, support vector machines) using as features the amino acid sequence of the protein or evolutionary information in the form of sequence profiles (e.g., PHDhtm[31], MEMSAT3[13], SVMTop[25]). The segments are identified using simple hydrophobicity plots[22] to ascertain probable helical segments and then employ various rules based on the expected lengths of the TMH segments to either accept, reject, or break long segments [35,25,38]. The segment orientation is often determined by relying on the fact that the regions between TMH segments that are positively charged tend to reside in the intracellular regions of the membrane (*positive-inside* rule[35]). The approaches that combine segment identification with orientation determination (e.g., MEMSAT3) employ dynamic programming methods to determine the different segments of a TMH protein and its orientation relative to the cytoplasm. Finally, the approaches that predict all of the above items in a single step utilize hidden Markov models (HMM) that capture the different structural components of a TMH protein (e.g., TMH segment, inside loop, outside loop, signal peptide, etc.) as separate modules. These models are trained on the amino acid sequence of the proteins (e.g., TMHMM[32] or on sequence profiles (e.g., Phobius[14]) and predict the topology by determining its most probable path through that model using Viterbi decoding[28].

This paper focuses on improving the accuracy of HMM-based approaches by combining them with an SVM-based approach that predicts the types of each residue. Specifically, we developed a TMH topology prediction algorithm, called TOPTMH, that solves the residue-type prediction, segment identification, and orientation determination in three distinct steps. The type of each residue is annotated via an SVM-based approach utilizing a window-based encoding of the residues' profile information and a second order exponential kernel function[30,29,17]. The segments are identified by using a pair of HMMs that model the different structural compo-

nents of TMH proteins. The first HMM uses as input the SVM predictions for each residue, whereas the second HMM uses as input hydropathy information as measured by a recently introduced hydrophobicity scale[9]. Finally, the orientation of the predicted segments is determined by applying the positive-inside rule.

The advantages of this approach are three-fold. First, by using a discriminative approach to learn a residue-type prediction model, the accuracy of these predictions are higher than those obtained (indirectly) by the HMM model. Second, by encoding the protein sequences via the SVM predictions, whose signal is significantly higher than that of the raw sequence profile, the demands imposed during HMM parameter estimation are substantially reduced allowing it to better focus on learning how to correctly identify the different segments. Third, by combining the outputs of the HMM models trained on the SVM predictions and on the hydrophobicity scores, it allows TOPTMH to correctly identify the TMH segments that have an amino acid composition that is similar to that of signal peptides.

We experimentally evaluated the performance of TOPTMH on three widely used datasets. Our evaluation was performed in two phases. First, we evaluated the gains obtained by TOPTMH by comparing it against an approach that uses a rule-based scheme to identify the TMH segments from the SVM predictions and another that uses just a single HMM model trained on the SVM predictions. Our evaluation showed that the HMM-based segment identification outperforms the rule-based approach by at least 50% in terms of the Q_{ok} score, which measures per-segment accuracy, and that by combining both the SVM- and the hydrophobicity-based HMM models, a further 3%–19% improvements can be obtained. Second, we evaluated its performance by comparing it against Phobius[14] and MEMSAT3[13]. Our evaluation showed that TOPTMH outperforms both of them across the different datasets. We also evaluated the performance of TOPTMH on an independent static benchmark[19]. The results on this blind evaluation showed that TOPTMH achieves the highest scores on high-resolution sequences (Q_2 score of 84% and Q_{ok} score of 86%) against existing state-of-the-art systems while achieving low signal peptide error.

2. Background and Definitions

2.1. *Transmembrane Helical Proteins*

The structure of a TMH protein of a series of helical segments passing through the cell's membrane (bilipid layer) separated by loop segments that are either on the inside or the outside side of the membrane. TMH segments can have two orientations: they can be going from the inside to the outside or from the outside to the inside of the cell. This orientation is relative to the location of N-terminus of the TMH protein. The TMH topology prediction problem involves predicting the residues that make up the helical segments and their orientation.

2.2. *Position Specific Scoring Matrices*

The position specific scoring matrix (PSSM) of a protein is obtained from a multiple sequence alignment of that protein and a set of other proteins that have a statistical significant sequence similarity (i.e., they are expected to be homologs). For a sequence X of length n, its PSSM is represented by a $n \times 20$ matrix $\mathcal{P}_X$. The n rows of this matrix correspond to the various positions in X and the columns correspond to the 20 distinct amino acids. The position specific scoring matrices used by TOPTMH were generated using the latest version of the PSI-BLAST algorithm[1] (available in NCBI's blast release 2.2.13), and were derived from the multiple sequence alignment constructed after five iterations using an e value of 10^{-2} for initial and subsequent sequence inclusions (i.e., we used `blastpgp -j 5 -e 0.01 -h 0.01`). The PSI-BLAST was performed against the SWISS-PROT[4] database release 53.0 that contains 269,293 sequences. A post processing step was performed to extract the log-odds scores ($n \times 20$ matrix) of each protein sequence from the PSI-BLAST output to use as the input feature for residue classification.

2.3. *Hydrophobicity Scale*

A hydrophobicity (HP) scale assigns a value to each of the 20 standard amino acids based on its hydrophobicity. In the context of TMH prediction methods, the Kyte and Doolittle[22] and the GES[7] HP scales are commonly used. These scales are based on biophysical or statistical analysis of high-resolution membrane protein structures and do not fully capture the cellular context of the membrane proteins[9]. For this reason, TOPTMH uses a recently published[9] HP scale (ΔG^{aa}_{app} scale) that captures the energetics of the protein-lipid interaction in biological contexts and thus is more biologically relevant. It has been shown that this scale is able to determine the topology of membrane proteins with higher precision than other scales[36].

3. TOPTMH Algorithm

The TOPTMH algorithm solves the TMH prediction problem by first assigning a score to each residue based on its likelihood to be in a helix state (residue annotation step), then using these scores it determines the protein's TMH segments (segment identification step), and finally using the positive-inside rule it determines their orientation (orientation determination step). These steps are described in the rest of this section.

3.1. *Residue Annotation Step*

We developed an SVM-based TMH residue annotation approach that uses features obtained from the protein's PSSM. Its overall structure is similar to that used by existing methods for SVM-based structural and functional annotation of protein residues using position specific scoring matrices (e.g., secondary structure

for globular proteins[17], solvent accessible surface area[30], disorder prediction[30], and DNA-binding[30]).

TOPTMH formulates the residue annotation problem as a binary classification problem whose goal is to predict if a residue belongs to a helix state or not. For each residue i of a protein sequence X, the input to the SVM is a $(2w+1)$-length subsequence (w`mer`) of X centered at position i. Each w`mer` is represented by a vector x_i of length $(2w+1) \times 20$ that is obtained by concatenating the rows of the PSSM for each position of the w`mer`. This w`mer`-based input is used for both training and prediction. The parameter w determines the length of the local environment around the ith sequence position used while building and applying the model and its optimal value is determined experimentally.

TOPTMH uses SVMlight[12] to learn the actual SVM model and utilizes the second order exponential function (*soe*)[17] as its kernel function. The *soe* kernel has been shown to produce better results than the traditional radial basis function (*rbf*) kernel for various sequence annotation prediction problems[17,30,29]. In the context of TOPTMH, the *soe* kernel function is given by

$$\mathcal{K}^{soe}(x_i, y_j) = \exp\left(1 + \frac{\mathcal{K}^2(x_i, y_j)}{\sqrt{\mathcal{K}^2(x_i, y_j)\,\mathcal{K}^2(x_i, y_j)}}\right), \tag{1}$$

where x_i and y_j are the vector representations of two w`mers`, $\mathcal{K}^2$ is given by

$$\mathcal{K}^2(x_i, y_j) = \langle x_i, y_j \rangle + \langle x_i, y_j \rangle^2, \tag{2}$$

and $\langle x_i, y_j \rangle$ denotes the dot-product of the x_i and y_j vectors.

3.2. *Segment Identification Step*

In order to determine the best approach for identifying the TMH segments we developed and studied three different approaches. The first approach utilizes a simple scheme based on empirical rules and the other two predict the topology by employing hidden Markov models (HMM)[28]. The first HMM-based approach uses a single HMM based solely on the SVM scores, whereas the second uses two HMMs—one based on SVM scores and one based on hydrophobicity scales.

3.2.1. *Rule-Based*

The rule-based segment identification approach post-processes the SVM-based residue annotations and identifies the segments by applying some heuristics rules that take into account the minimum and maximum lengths of the TMH segments. Specifically, for each protein, this approach traverses the SVM annotated residues and identifies all maximal contiguous segments that were annotated as TMHs by the SVM. Any TMH segment whose length l is shorter than the minimum length of L_{min} residues is rejected (i.e., converted into non-helix residues). If any of the remaining segments have $l > L_{max}$, they are split into two separate segments as

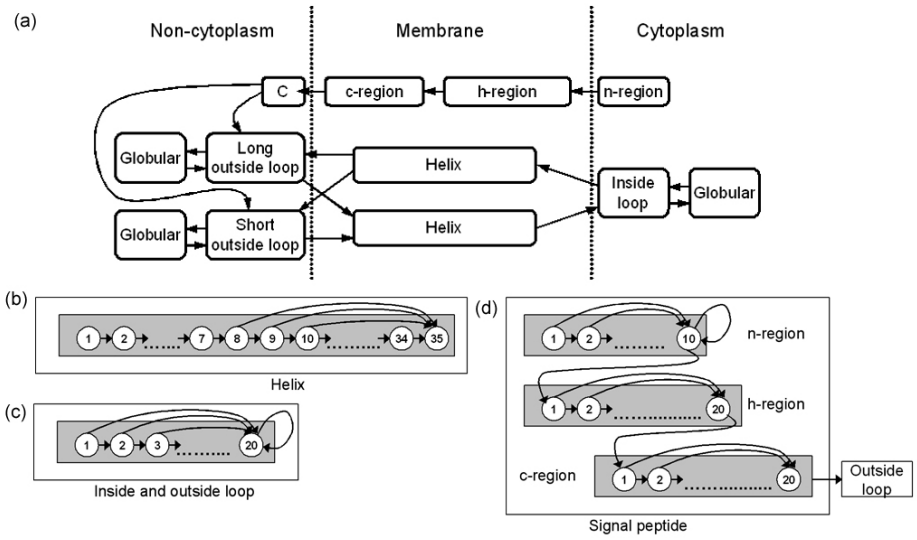

Fig. 1. The layout of the HMM model used in TOPTMH.

follows. For the segments with $l \leq 2L_{opt} + C$, the segment is split by changing the middle C residues into loops. For segments with $l > 2L_{opt} + C$, the segment is split by creating two helical segments consisting of the first and last L_{opt} residues and converting the remaining central residues into loops. The threshold values L_{min}, L_{max}, L_{opt} and C are set as 9, 38, 19 and 6 respectively. These values were initially chosen based on a literature review[35,3,38] and then optimized to provide the best results given the SVM-based annotations produced by TOPTMH.

3.2.2. *HMM-Based*

The HMM-based segment identification approaches determine the segments of a TMH protein by *threading* the sequence into an HMM model that is designed to capture the various structural components of a TMH protein. These approaches were motivated by recent studies which showed that HMM-based TMH prediction methods are well-suited for predicting the topology of TMH proteins as they can directly learn from the data the various structural constraints associated with TMH protein segments and their relations to the protein's underlying sequence and/or PSSM[3,15,6]. However, unlike these methods, the HMM-based approaches that we developed take into account the SVM-scores produced by the residue annotation step, which provide better per-residue predictions for the helix/non-helix states than the maximum likelihood approaches used by HMMs.

The architecture of our HMM model, shown in Figure 1, is designed to capture the known structural information of TMH proteins and is similar to that employed by Phobius[14]. The model contains four major compartments: (i) helix, (ii) inside

loop, (iii) outside loop, and (iv) signal peptide. The helix compartment is composed of two submodels each containing 35 states. One submodel is used for modeling helix segments that go from inside towards the outside, and the other for the helix segments that go from outside towards the inside. In each of these submodes, states 1–8 contain transitions to only the next state, whereas states 9–34 can transition to the next state or to state 35 (last state). Thus, any predicted helix segment will be of length 9–35 residues long. The outside loop compartment is divided into two submodels to represent long and short non-cytoplasmic loops. Each of these submodels contains 20 states to model loops that are at least 1–20 residues long. Each submodel also has a state with self-transition to represent long cytoplasmic loops. The inside loop compartment also contains 20 states to allow it to model loops that are 1–20 residues long. The signal peptide compartment was designed based on Phobius model and it has three regions: the n-region (10 states), the h-region (20 states), and the c-region (20 states). The last state of the c-region represents a cleavage site transitioning to a outside loop state.

The HMM models were built using the UMDHMM[16] package (version 1.02), which was modified to take as input annotated protein sequences. The threading of a sequence through the HMM model was done using the Viterbi[28] algorithm.

HMM Based on SVM Scores (HMM-SVM). This approach builds an HMM model that only takes into account the per-residue SVM scores produced by the annotation step. To construct the training set, the SVM score for each residue is computed. Since, HMMs are primarily designed to operate on finite size alphabets, the raw SVM scores are discretized into a finite number of bins with each bin corresponding to a distinct symbol. The final training set for the HMM corresponds to a set of proteins with known TMH topology represented as sequences of SVM-score based bins. A similar SVM-based prediction followed by discretization is performed when this model is used to predict the topology of a test protein. We discretized the SVM scores into equal-size intervals, and assigned all residues with scores ≤ -3 and ≥ 3 into the first and last bin, respectively.

HMM Based on SVM Scores and Hydrophobicity Scores (HMM-SVM+HP). This model builds a pair of HMM models—one based on SVM scores (HMM-SVM) and one based on the hydrophobicity values (HMM-HP) of known TMH sequences and combines the topology predictions from both HMM models. This approach was motivated by the fact that in certain cases, the SVM-based residue annotation may fail to identify certain hydrophobic TMH segments. This is further discussed in Section 5.

The HMM-SVM model is identical to that described in the previous section. The HMM-HP model is built by first encoding the amino acids of each TMH protein as a sequence of discretized hydrophobicity values. Table 1 shows the scheme used to discretize the hydrophobicity values for each amino acid. Both the HMM-SVM and HMM-HP models are used independently to predict the TMH segments. The

Table 1. Discretization of Hydrophobicity values.

Labels	Amino Acids	HP Values
1	R, E, K, D	$2.5 < h$
2	N, H, P, Q	$1.0 < h < 2.5$
3	T, Y, G, S	$-0.1 < h < 0.9$
4	F, V, C, A, M, W	$-0.4 < h < -0.1$
5	I, L	$h < -0.5$

HP Values denotes a range of hydrophobicity values decided based on[9]

final set of predictions consists of the segments predicted by HMM-SVM and those segments predicted by HMM-HP that do not overlap with any of the segments of HMM-SVM. Two segments are considered to overlap if they have more than five residues in common. Since this approach combines both the SVM- and HP-based HMM models, we will refer to it as HMM-SVM+HP.

3.3. *Orientation Determination Step*

Once the TMH segments have been identified, their orientation relative to the N-terminus is determined by applying the positive-inside rule[35] using the technique introduced in THUMBUP[38]. In this approach, each protein is first coded into a binary sequence by assigning a one to the first protein residue and all the arganine and lysine residues and a zero to the remaining residues. Then, a score is computed for each loop by adding the values of its 15 neighboring residues on each side. If the total score for odd-numbered loops is greater than or equal to that of even loops, the N-terminus is inside the membrane, otherwise it is outside.

4. Experimental Design

4.1. *Datasets*

We evaluated the prediction performance of the TOPTMH method on datasets used by the Phobius and MEMSAT3 methods, on a high-resolution dataset, and by participating on the static benchmark[18].

The datasets obtained from the Phobius study included a set of 247 transmembrane proteins and a set of 45 transmembrane proteins that contained signal peptide residues with transmembrane helix segments. We will denote the first dataset as TM-Only and the second as TM-SP. The dataset obtained from MEMSAT3 consisted of a set of 184 non-homologous transmembrane proteins denoted as Möller that also contained a few signal peptide proteins.

The high-resolution dataset consists of 176 transmembrane protein sequences that have experimentally determined 3D structures. These proteins were collected from the OPM[26] and MPtopo[11] databases. All the 3D TMH sequences from both databases were collected and homology reduced at 40% sequence identity by using cd-hit[23]. The final dataset denoted as 3D is a combination of 68 OPM sequences and 108 MPtopo sequences.

The static benchmark consists of a set of 2247 sequences whose true annotations are not given to the public. A method predicts the annotations for these sequences and uploads them to the evaluation server. The server assesses the quality of the predictions and compares them to that obtained by other methods. The 2247 sequences contain four distinct subsets. The first is the high-resolution subset which contains sequences of proteins whose high resolution structure is available, the second is the low-resolution subset that includes membrane proteins detected using low resolution structures, the third subset is the globular protein subset which includes globular protein sequences and the fourth is the signal peptide subset that includes proteins sequences with signal peptide residues. The sequences provided to the public is not grouped in the above mentioned subsets, but the results published on the evaluation server is presented accordingly.

4.2. *Training & Testing Methodology*

For each of the TM-SP and TM-Only datasets, the different methods were evaluated using a standard 10-fold cross validation protocol by splitting the proteins into 10 different parts. The percent sequence identity between the different folds were at most 30% and 35% for the TM-Only and TM-SP datasets, respectively. The ten folds were identical to that used by Phobius making it possible to directly compare our results with those obtained by Phobius.

The two-level HMM-SVM model was trained as follows. The training set was further split into 10 different folds $\{f_1, \ldots, f_{10}\}$. For each fold f_i, the other nine folds were used to train the SVM model and then used to predict the residues for the proteins in f_i. At the end of this step all the residues of the proteins in the training set have SVM predictions. These predictions are then used to train the HMM model for the training set. In addition, the entire training set is used to build an SVM residue prediction model. Note that the test set is not used anywhere during training. During testing, the residues of each test protein are first predicted using the SVM model built on the entire training set, and these predictions are provided as input to the HMM model to predict the TMH segments.

The predictions for the 3D dataset and the static benchmark were obtained by training the SVM and HMM models using all the sequences from TM-SP and TM-Only datasets.

4.3. *Evaluation Metrics*

The performance of TMH prediction is evaluated on a per-residue and on a per-segment basis using well-established metrics[3]. The per-residue evaluation measures the ability of a method to correctly annotate the different residues into helices or non-helices (two classes). We used three per-residue metrics denoted by $Q_{2T}^{\%obs}$, $Q_{2T}^{\%prd}$, and Q_2. $Q_{2T}^{\%obs}$ is the percentage of observed TMH residues that are predicted correctly (helix recall), $Q_{2T}^{\%prd}$ is the percentage of predicted TMH residues

that are predicted correctly (helix precision), and Q_2 is the percentage of correctly predicted residues (both helix and non-helix).

The per-segment evaluation measures the ability of a method to correctly identify the actual TMH segments. We used three per-segment metrics denoted by $Q_{htm}^{\%obs}$, $Q_{htm}^{\%prd}$, and Q_{ok}. $Q_{htm}^{\%obs}$ is the percentage of observed TMH segments that are predicted correctly (TMH segment recall), $Q_{htm}^{\%prd}$ is the percentage of predicted TMH segments that are predicted correctly (TMH segment precision), and Q_{ok} is the percentage of proteins for which all the TMH segments are predicted correctly. Note that Q_{ok} is a very strict metric as each protein contributes either a zero or an one. In the above metrics, a predicted TMH segment is considered to be correctly identified if there is an overlap of ten residues between the predicted and observed helix segments.In addition, a predicted helix segment is counted only once. This is illustrated by considering the following examples:

```
Obs1 : TTTTTTTTTTTTTTT------TTTTTTTTTTTT
Pred1: -----TTTTTTTTTTTTTTTTTTTTTTTTTT---

Obs2 : ---TTTTTTTTTTTTTTTTTTTTTTTTTTTT--
Pred2: TTTTTTTTTTTTT------TTTTTTTTTTTTTT
```

In this example, Obs1 and Pred1 are the observed and predicted TMH segments for a particular protein sequence. During evaluation, the second segment of the Obs1 sequence will not be considered as correctly predicted, since the only segment predicted in Pred1 is already accounted for in the first segment of the Obs1 sequence. On the other hand, the second segment of the Pred2 sequence will be considered as incorrectly predicted as the first segment will be considered for the only segment in Obs2 sequence.

Although, the per-residue measures capture the accuracy of a method to predict the annotation label for a residue, it is not able to assess the ability of the method to identify the TMH segments separated by loop regions of different lengths. Hence, TMH prediction algorithms are mostly evaluated using per-segment metrics.

5. Results

5.1. *Residue Annotation Performance*

The performance achieved by the SVM-based residue annotation for different values of w is shown in Table 2. This table shows the per-residue performance metrics (Q_2, $Q_{2T}^{\%obs}$ and $Q_{2T}^{\%prd}$) for a subset of the TM-ONLY dataset. We observe that in terms of the various metrics, the performance achieved for different values of w is rather similar. The only exception is $w = 2$, where the performance is substantially lower than the rest. Overall, the best performance was obtained using w`mer` of length seven. For this reason, all the remaining experiments presented in this study use $w = 7$.

Table 2. Residue Annotation Performance with varying $w$mer length.

$w$mer	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$
2	86.6	78.1	76.9
5	88.2	85.3	75.5
7	**88.3**	84.7	**77.4**
11	88.3	**85.5**	76.6

The numbers in bold show the best performing $w$mer for that metric. $Q_{2T}^{\%obs}$ denotes the helix recall and $Q_{2T}^{\%prd}$ denotes the helix precision for the per-residue based performance evaluation.

5.2. *Segment Identification Performance*

Table 3 presents the per-residue and per-segment based results of different TMH segment identification approaches on the TM-Only and TM-SP datasets. For the SVM-HMM approach, Table 3 shows three different sets of results that were obtained by binning the SVM scores into 5, 7, and 12 bins (HMM-SVM-D5, HMM-SVM-D7, and HMM-SVM-D12). The row labeled "Raw-SVM" shows the results obtained by using as TMH segments the maximal contiguous segments that were predicted as TMHs by the SVM (i.e., the set of segments that form the input to the rule-based segment identification approach). Comparing the per-residue performance achieved by the various approaches we see that Raw-SVM achieves very good per-residue two-state accuracy (Q_2). It has the highest Q_2 value for TM-Only and the second highest for TM-SP. However, focusing on this metric alone is misleading because most of the residues in transmembrane proteins are non-helix[24] and relatively high Q_2 values can be obtained by simply predicting most of the residues as being in a non-helix state. Consequently, high Q_2 values represent good perfor-

Table 3. TMH Segment Identification Performance.

	TM-SP			TM-Only		
	Per-Residues Scores					
Methods	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$
Raw-SVM	96.73	71.10	86.60	90.64	84.30	83.10
Rule	95.16	59.56	95.89	89.19	79.65	87.36
HMM-SVM-D5	96.28	76.39	84.87	89.40	85.54	82.25
HMM-SVM-D7	96.45	76.85	87.72	89.34	85.61	82.23
HMM-SVM-D12	96.24	77.56	84.45	89.31	86.13	81.35
HMM-SVM-D7+HP	97.08	84.80	88.50	89.46	86.21	82.04
	Per-Segment Scores					
Methods	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$
Raw SVM	35.55	85.23	70.09	38.86	94.34	74.33
Rule	64.44	75.00	100.00	70.85	92.88	94.96
HMM-SVM-D5	64.44	84.09	87.05	71.66	95.39	93.73
HMM-SVM-D7	71.11	85.23	92.59	72.06	95.63	93.52
HMM-SVM-D12	60.00	85.22	85.22	70.04	95.80	92.87
HMM-SVM-D7+HP	84.44	93.18	93.18	73.68	96.12	93.33

Q_2 and Q_{ok} represent the overall prediction accuracy for per-residue based and per-segment based evaluation. $Q_{2T}^{\%obs}$ denotes the helix recall and $Q_{2T}^{\%prd}$ denotes the helix precision for the per-residue based performance evaluation. $Q_{htm}^{\%obs}$ denotes the helix segment recall and $Q_{htm}^{\%prd}$ denotes the helix segment precision for the per-segment based performance evaluation.

mance only if they are accompanied with high helix recall ($Q_{2T}^{\%obs}$) values. In light of this discussion, we see that the HMM-based segment identification approaches tend to achieve considerably better recall values (especially for TM-SP) while their helix precision ($Q_{2T}^{\%prd}$) is in some cases better than that of the Raw-SVM approach. Among the different schemes, the rule-based approach achieves the best precision results, whereas the approach that combines the SVM- and HP-based HMMs (HMM-SVM-D7+HP) achieves the best recall. However, unlike the high precision achieved by the HMM-SVM-D7+HP approach, the rule-based scheme achieves the lowest recall leading to the worst Q_2 values.

Comparing the per-segment performance, we see that the Raw-SVM approach achieves Q_{ok} scores that range from 35%–40%, which are by far the lowest among the different approaches. These results indicate that even though Raw-SVM can correctly predict a large fraction of the helical residues, it fails to predict correctly large contiguous portions of each helical segment. On the other hand, the per-segment performance achieved by the other segment identification approaches are considerably higher. Both the rule- and HMM-based approaches are able to significantly improve over Raw-SVM for both the TM-SP and TM-Only datasets. Among them, the approaches based on HMM-SVM outperform the rule-based approach by 2%–12%, even though the latter achieved the highest $Q_{htm}^{\%prd}$ scores (100% and 96.44% for TM-SP and TM-Only, respectively).

The overall best Q_{ok} results were obtained by the HMM-SVM-D7+HP approach. In particular, the Q_{ok} values achieved by HMM-SVM-D7+HP are 19% and 3% better than the next best performing scheme (HMM-SVM-D7) on the TM-SP and TM-Only datasets, respectively. The large performance advantage of HMM-SVM-D7+HP over HMM-SVM-D7 on the TM-SP dataset are primarily due to increases in recall ($Q_{htm}^{\%obs}$). HMM-SVM-D7+HP achieves a $Q_{htm}^{\%obs}$ of 93.18% compared to the 85.23% achieved by HMM-SVM-D7. A possible explanation for the relatively poor performance of HMM-SVM-D7 is that due to the signal peptide segments present in some of the sequences in the TM-SP dataset, the SVM model fails to identify some of the TMH residues. However, these residues can be correctly identified when hydrophobicity scores are considered, and as such the combined HMM-SVM-D7+HP approach leads to better overall results.

5.3. *Performance Comparison with Previous Methods*

In this section we compare the performance achieved by the TOPTMH method that uses the HMM-SVM-D7+HP topology prediction approach against that achieved by some of the previously developed TMH prediction methods on various datasets.

5.3.1. TOPTMH *Performance Comparison with Phobius and MEMSAT3.*

Tables 4 and 5 compares the performance achieved by TOPTMH against that achieved by Phobius and MEMSAT3, which are two of the best TMH prediction

Table 4. Performance Comparison with Phobius.

	TM-SP	TM-Only
Method	Accuracy	Accuracy
TOPTMH	93.18	75.71
Phobius	91.10	63.60

Accuracy is the percentage of correctly predicted proteins. A prediction is correct when all predicted TMH segments overlap all observed TMH segments over a five residue stretch and loops were located correctly.

methods currently available. Phobius uses a sophisticated HMM to mark the TMH and signal peptide regions and MEMSAT3 uses a combination of neural network and dynamic programming to identify the TMH segments. To facilitate direct comparisons between TOPTMH and these methods, the performance metrics in these tables are similar to those used in the evaluations of Phobius and MEMSAT3.

Comparing TOPTMH's performance against Phobius (Table 4) we see that TOPTMH achieves accuracies that are 2% and 10% higher than those achieved by Phobius on the TM-SP and TM-Only datasets, respectively. The performance advantage of TOPTMH over Phobius also holds for the Möller dataset (Table 5) as well. TOPTMH performed better in all three categories by correctly predicting 162, 149, and 134 proteins compared to the 152, 134, and 126 proteins predicted by Phobius, respectively.

Comparing TOPTMH's performance against MEMSAT3 (Table 5) we see that TOPTMH was able to predict the correct number of TMH segments for more proteins (162 *vs* 156) and predict the correct topology for a similar number of proteins (149 *vs* 150). However MEMSAT3 was able to predict more proteins with both correct topology and location than TOPTMH (147 *vs* 134). We believe that this is primarily due to the fact that due to the binary classification of the protein sequences in helix and non-helix residues, TOPTMH was not able to effectively differentiate between inside and outside loops and thus could not perform similar to MEMSAT3.

Table 5. Performance Comparison with MEMSAT3 on the Möller dataset.

Method	# TM SEG	# TOPO	# TOPO+LOC	# TOPO+LOC(10)
TOPTMH	162 (88.04%)	149 (80.98%)	134 (72.83%)	131 (71.20%)
Phobius	152 (82.60%)	134 (72.80%)	126 (68.40%)	120 (65.20%)
MEMSAT3	156 (84.80%)	150 (81.50%)	147 (79.90%)	141 (76.60%)

TM SEG denotes the number of predicted proteins that had correct number of TMH segments irrespective of topology or location. # TOPO denotes the number of proteins for which the orientation of the protein (N-terminus is inside or outside of the cytoplasm) was predicted correctly. # TOPO+LOC denotes the number of proteins for which the topology and the TMH segment locations were predicted correctly. This score was calculated based on five residue segment overlap. # TOPO+LOC(10) shows the # TOPO+LOC scores for ten residue segment overlap.

5.3.2. TOPTMH *Performance on the Static Benchmark.*

Table 6 shows the performance achieved by TOPTMH on the static benchmark. From these results we see that TOPTMH achieved the highest Q_{ok} score of 86% for the high-resolution sequences and the highest Q_2 scores of 84% and 90% for the high- and low-resolution sequences, respectively. Moreover, TOPTMH has performed about 7% better in TMH prediction than both MEMSAT3 and Phobius. Note that even though HMMTOP2 achieved $Q_{htm}^{\%obs}$ and $Q_{htm}^{\%prd}$ scores that were higher than the corresponding scores achieved by TOPTMH, its Q_{ok} score of is lower than that achieved by TOPTMH. This is due to the fact that even though HMMTOP2 identified more TMH segments in total than TOPTMH, it was not as successful in predicting proteins for which all of the TMH segments were identified correctly.

5.3.3. TOPTMH *Performance on the* 3D *Dataset.*

Table 7 shows the per-segment based results achieved by TOPTMH, Phobius, TMHMM2[21] (HMM based method), and MEMSAT3 on the 3D dataset. Two sets of results are presented. The first shows the performance achieved on the entire 3D dataset, whereas the second, shows the performance achieved on a subset (3D-Sub) that contains only sequences that have less than 40% sequence identity to the set of sequences used to train TOPTMH's model. The 3D-Sub contains 118 sequences. The results for MEMSAT3 were obtained by running it locally, whereas the results for Phobius and TMHMM2 were obtained by querying their respective web-servers.

Comparing TOPTMH against Phobius and TMHMM2 we see that it produces predictions whose accuracy in terms of Q_{ok} is 17%–26% better for both 3D and 3D-Sub. The superior result of TOPTMH over Phobius and TMHMM2 shows that the hybrid SVM-HMM based method can predict with better accuracy than HMM only methods. The performance advantage of TOPTMH over MEMSAT3 is

Table 6. TMH Benchmark Results.

	High Resolution Accuracy						Low Resolution Accuracy					
	Per-segment			Per-residue			Per-segment			Per-residue		
Method	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$
TOPTMH	**86**	95	96	**84**	75	90	66	92	88	**90**	84	80
PHDpsihtm08	84	99	98	80	76	83	67	95	94	89	87	77
HMMTOP2[33]	83	99	99	80	69	89	66	94	93	90	85	83
MEMSAT3	80	98	97	83	78	88	63	92	87	88	86	76
Phobius	80	92	93	80	69	84	65	90	88	90	81	79
DAS[5]	79	99	96	72	48	94	39	93	81	86	65	85
TopPred2[8]	75	90	90	77	64	83	48	84	79	88	74	71
TMHMM1[32]	71	90	90	80	68	81	**72**	91	92	90	83	80
SOSUI[10]	71	88	86	75	66	74	49	88	86	88	79	72
PHDhtm07	69	83	81	78	76	82	56	85	86	87	83	75

The results for TOPTMH, MEMSAT3, and Phobius were obtained by collecting predictions for the test set of the TMH static benchmark[18] and submitting the results to the benchmark server. All the other results were provided by the TMH static benchmark evaluation web-site.

Table 7. Performance comparisons for 3D dataset.

	3D Dataset (176)			3D-Sub Dataset (118)		
Method	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$
TOPTMH	**77.27**	**90.93**	97.21	**70.34**	87.08	95.80
Phobius	61.36	87.43	96.05	54.24	84.71	95.00
TMHMM2	61.93	87.43	86.78	59.32	85.14	96.39
MEMSAT3	73.86	89.77	**97.45**	**70.34**	**87.26**	**97.16**

The 3D-Sub is the subset of 3D dataset that contains only sequences that have less than 40% sequence identity to the set of sequences used to train TOPTMH's model. The results for MEMSAT3 were obtained by running it locally; the results for Phobius and TMHMM2 were obtained from their web-servers.

less consistent. TOPTMH outperforms MEMSAT3 on the 3D dataset by 4% (Q_{ok}) but it performs comparably on the 3D-Sub dataset.

6. Discussion

In order to understand the cases that TOPTMH fails to predict correctly, we analyzed TOPTMH's results for the 3D dataset. TOPTMH predicted the TMH segments with very high precision ($Q_{htm}^{\%prd}$ of 97.21%), but failed to identify about 9% of the TMH segments ($Q_{htm}^{\%obs}$ of 90.93%) (Table 7). In analyzing the incorrectly predicted sequences, we found that the errors fall under four well-defined categories. The first category contains errors in which TOPTMH merged two consecutive short TMH segments and thus it failed to predict the second segment (≈37% of the errors). Many of these merged short consecutive segments correspond to *reentrant regions*[34] (they enter and exit the membrane on the same side), that according to the work of Viklund et al[34] they should have been annotated as a single segment. The second category contains errors in which TOPTMH failed to predict short TMH segments (≈15% of the errors). This includes TMH segments that are shorter than the minimum length of nine residues that TOPTMH's HMM model is designed to capture (Section 3.2.2). In addition, it includes TMH segments that are mostly nine residues long for which the signals captured by SVM from the protein sequences were too weak for HMM to identify these regions as TMH segments. The third category contains errors that are due to SVM's failure to correctly predict the types of some of the residues in the TMH segments (≈30% of the errors). Finally, the fourth category contains errors that are due to over-prediction from the hydrophobicity-based HMM model (≈15% of the errors).

7. Conclusions

In this paper we developed the TOPTMH method to predict the transmembrane α-helix topology using sequence information. TOPTMH uses PSI-BLAST constructed profiles and hydrophobicity information within a hybrid SVM- and HMM-based framework. This novel hybrid method captures the power of SVM-based models to discriminate between the helical and non-helical residues with the power of HMMs to identify length-dependent topological structures. Experiments on the

Phobius and MÖLLER datasets showed that TOPTMH achieves high per-residue and per-segment accuracies and that on an independent static benchmark it outperforms existing state-of-the-art methods such as PHDpsihtm08[31], HMMTOP2[33], MEMSAT3[13], Phobius[14], and TopPred2[8].

Acknowledgments

This work was supported by NSF IIS-0431135, NIH RLM008713A, and by the University of Minnesota's Digital Technology Center and Minnesota Supercomputing Institute.

References

1. S. F. Altschul, L. T. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–402, 1997.
2. H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The protein data bank. *Nucleic Acids Res*, 28:235–242, 2000.
3. C. P. Chen and B. Rost. State-of-the-art in membrane protein prediction. *Appl Bioinformatics*, 1(1):21–35, 2002.
4. The UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Res.*, 35:D193–D197, 2007.
5. M. Cserzö, E. Wallin, I. Simon, G. von Heijne, and A. Elofsson. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng*, 10(6):673–676, Jun 1997.
6. A. Elofsson and G. von Heijne. Membrane protein structure: Prediction versus reality. *Annu. Rev. Biochem*, 76:125–140, 2007.
7. D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 15:321–353, 1986.
8. D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353, 1986.
9. T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, 2005.
10. T. Hirokawa, S. Boon-Chieng, and S. Mitaku. Sosui: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14:378–379, 1998.
11. S. Jayasinghe, K. Hristova, and S. H. White. Mptopo: A database of membrane protein topology. *Protein Sci*, 10(2):455–458, Feb 2001.
12. T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press, 1999.
13. D. T. Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, 2007.
14. Lukas Käll, Anders Krogh, and Erik L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–1036, May 2004.

15. Lukas Käll and Erik L L Sonnhammer. Reliability of transmembrane predictions in whole-genome data. *FEBS Lett*, 532(3):415–418, Dec 2002.
16. Tapas Kanungo. *UMDHMM: Hidden Markov Model Toolkit*. Cambridge University Press, 1999.
17. George Karypis. Yasspp: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins*, 64(3):575–586, Aug 2006.
18. A. Kernytsky and B. Rost. Static benchmarking of membrane helix predictions. *Nucl Acids Res*, 31(13):3642–3644, 2003.
19. Andrew Kernytsky and Burkhard Rost. Static benchmarking of membrane helix predictions. *Nucleic Acids Res*, 31(13):3642–3644, Jul 2003.
20. T. Klabunde and G. Hessler. Drug design strategies for targeting g-protein-coupled receptors. *ChemBioChem*, 3:928–944, 2002.
21. A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580, Jan 2001.
22. J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.
23. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 2006.
24. J. Liu and B. Rost. Comparing function and structure between entire proteomes. *Protein Sci.*, 10:1970–1979, 2001.
25. Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, Ping-Chiang Lyu, and Wen-Lian Hsu. Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J Proteome Res*, 7(2):487–496, Feb 2008.
26. Mikhail A Lomize, Andrei L Lomize, Irina D Pogozheva, and Henry I Mosberg. Opm: orientations of proteins in membranes database. *Bioinformatics*, 22(5):623–625, Mar 2006.
27. Amit Oberai, Yungok Ihm, Sanguk Kim, and James U Bowie. A limited universe of membrane protein families and folds. *Protein Sci*, 15(7):1723–1734, Jul 2006.
28. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.
29. Huzefa Rangwala and George Karypis. frmsdpred: Predicting local rmsd between structural fragments using sequence information. *Proteins*, 72(3):1005–1018, Aug 2008.
30. Huzefa Rangwala, Christopher Kauffman, and George Karypis. A generalized framework for protein sequence annotation. In *Proceedings of the NIPS Workshop on Machine Learning in Computational Biology*, 2007.
31. B. Rost, P. Fariselli, and R. Casadio. Topology prediction for helical transmembrane proteins at 86accuracy. *Protein Sci*, 5(8):1704–1718, Aug 1996.
32. E. L. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth Intl. Conf. on Intelligent Systems for Molecular Biology*, pages 175–82, 1998.
33. G. E. Tusnády and I. Simon. The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, Sep 2001.
34. Hkan Viklund, Erik Granseth, and Arne Elofsson. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J Mol Biol*, 361(3):591–603, Aug 2006.
35. G. von Heijne. Membrane protein structure prediction. hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology*, 225(2):487–494, 1992.

36. Gunnar von Heijne. Formation of transmembrane helices in vivo–is hydrophobicity all that matters? *The Journal of general physiology*, 129(5):353–356, 2007.
37. E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, 7(4):1029–38, 1998.
38. H. Zhou and Y. Zhou. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-markov-model-based method. *Protein Sci*, 12:1547–1555, 2003.