Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function

*Allan Lo[1, 2], Hua-Sheng Chiu[3], Ting-Yi Sung[3], Ping-Chiang Lyu[2], and Wen-Lian Hsu[3, *]*

[1] Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan

[2] Department of Life Sciences, National Tsing Hua University, Hsinchu, Taiwan

[3] Bioinformatics Lab., Institute of Information Science, Academia Sinica, Taipei, Taiwan

Emails: {allanlo, huasheng, tsung}@iis.sinica.edu.tw, lslpc@life.nthu.edu.tw, hsu@iis.sinica.edu.tw

[*] To whom correspondence should be addressed. Tel: +886-2-27883799 ext. 1804; Fax: +886-2-27824814; Email: hsu@iis.sinica.edu.tw

ABSTRACT

The prediction of transmembrane (TM) helix and topology provides important information about the structure and function of a membrane protein. Due to the experimental difficulties in obtaining a high-resolution model, computational methods are highly desirable. In this paper, we present a hierarchical classification method using support vector machines (SVM) that integrates selected features by capturing the sequence-to-structure relationship and developing a new scoring function based on membrane protein folding. The proposed approach is evaluated on low- and high-resolution data sets with cross-validation and the topology (sidedness) prediction accuracy reaches as high as 90%. Our method is also found to correctly predict both the location of TM helices and the topology for 69% of the low-resolution benchmark set. We also test our method for discrimination between soluble and membrane proteins, and achieve very low overall false positive (0.5%) and false negative rates (0~1.2%). Lastly, the analysis of the scoring function suggests that the topogeneses of single- and multi-spanning TM proteins have different levels of complexity and the consideration of inter-loop topogenic interactions for the latter is the key to achieving better predictions. This method can facilitate the annotation of membrane proteomes to extract useful structural and functional information. It is publicly available at http://bio-cluster.iis.sinica.edu.tw/~bioapp/SVMtop.

KEYWORDS

membrane protein, topology prediction, transmembrane helix, support vector machines, structure prediction

INTRODUCTION

Integral membrane proteins represent a diverse class of proteins that constitute the key components of various cellular processes, including signal and energy transduction, the regulation and transport of ions and macromolecules across membranes, intercellular communication, and cell adhesion.[1] The biological importance of integral membrane proteins is also highlighted by their abundance, estimated at about 20-30% of all genes in both prokaryotic and eukaryotic organisms.[2,3] To gain a better understanding of how these proteins function, knowledge of their high-resolution structures is required. However, very little information about their high-resolution structures is available due to difficulties in applying experimental techniques in X-ray crystallography and NMR. Therefore, it is necessary to develop computational approaches for elucidating the structural genomics of integral membrane proteins.

Most integral membrane proteins belong to the helix-bundle class in which the basic architecture consists of one or more tightly packed transmembrane (TM) α-helices. These TM helices (TMHs) form independently stable folding units in a membrane environment. The goal of sequence-based prediction methods is to identify the location of TMHs and the topology or orientation of the N-terminus relative to the membrane. Studies of the amino acid propensities of TMHs led to the development of a plethora of methods based on simple hydrophobicity scales to search for potential

membrane-spanning segments.[4-6] Meanwhile, topology or sidedness prediction is accomplished by methods that incorporate the charge bias distribution between the inner and outer loops, which is also known as the 'positive-inside' rule.[7,8] Model-based approaches using hidden Markov models (HMM) to extract information from different regions of a TM protein are highly successful in predicting TMHs.[3,9-11] Another class of approaches using support vector machines (SVM) or neural networks combined with evolutionary information also improve the accuracy of both helix and topology prediction.[12-14] Consensus prediction methods based on multiple predictors have also been developed.[15,16]

Here, we present an approach called SVM*top* (SVM for transmembrane helix and *top*ology prediction), which is based on support vector machines. The task of helix and topology prediction is separated into two stages in a hierarchical framework, thus the complexity of each stage is reduced and relevant input features can be applied separately. For helix prediction in the first stage, we select and integrate multiple input features based on both the sequence and the structure of a TM helix for the first SVM classifier. In the second stage, topology (sidedness of the N-terminus) prediction is accomplished by the second classifier and a new scoring function called the *Alternating Geometric Scoring Function* (AGSF). To the best of our knowledge, the AGSF is the first attempt to model the relationship between the interactions of inter-loop topogenic signals and topogenesis.

In this paper, we first evaluate our approach on both low- and high-resolution data sets and show that it obtains high accuracy for helix and topology predictions and compares favorably with many existing methods. Second, we demonstrate that our method can discriminate between soluble and

membrane proteins at very low error rates (0% to 1.2%). Lastly, we compare the topology prediction accuracy of a well-established topology predictor, namely, the 'positive-inside' rule, and the AGSF for single-spanning and multi-spanning membrane proteins. Our analysis sheds light on the differences in topogenesis between single-spanning and multi-spanning membrane proteins and suggests that inter-loop topogenic interactions play an important role in the topogenesis of membrane proteins.

METHODS

Data sets

1) Low-resolution membrane proteins: A collection of low-resolution data sets compiled by Möller *et al.*[17] containing 145 proteins of good reliability (Trust Level A-C only) from a set of 148 non-redundant proteins is selected and manually validated using annotations from SWISS-PROT release 49.0.[18] Two proteins are removed because they have no membrane protein annotations. The final data set contains 143 proteins. The SVM models trained on these proteins are used for testing on the following data sets. These proteins are listed in Table 1S of the *Supporting Information*.

2) High-resolution membrane proteins: A set of high-resolution structures is obtained from MPtopo, a database of membrane protein topology.[19] The 3D_helix set contains 101 proteins with experimentally determined structures. Another set of 231 'alpha-non-redundant' proteins from the TMPDB database is also considered.[20] We merge these two data sets and use the Cd-hit program to remove redundancy at 30% sequence identity.[21] The final data set, which contains 258 non-redundant transmembrane proteins, is used for benchmarking with other methods and topology

prediction of single-spanning and multi-spanning membrane proteins. These proteins are listed in Table 2S of the *Supporting Information*.

3) Soluble proteins: A collection of 616 high-resolution soluble proteins from the Protein Data Bank, PDB[22], compiled by Chen *et al.*[23] is used to test for the discrimination between soluble and membrane proteins.

System Architecture

We represent transmembrane helix and topology prediction as a multi-classification problem and solve it in two stages using SVM classifiers. Figure 1 shows the overview of the proposed method. The SVM classification model uses a hierarchical framework in which a residue is predicted as helix (*H*) or non-helix (*~H*) in the first stage, followed by the prediction of the topology of the non-helix residues in the second. In the first stage, a binary SVM classifier is trained to predict helix and non-helix residues (*H/~H*). TMH candidates are determined and assembled from the predicted helix residues into physical segments. In the second stage, we predict the topology of the query protein. The non-helix residues (*~H*) are further discriminated by a second classifier into inner or outer loop residues (*i/o*). Since the residues in a loop segment can have different predicted topology labels (*i, o*), we apply AGSF to determine the final topology of the query protein.

A sliding window is used to partition a protein sequence into peptides and the class label (*H/i/o*) of the center residue of each peptide is predicted. The optimal length of the sliding window is incrementally searched from 3 to 41 for both classifiers. The optimal window size of w_1 for the first classifier and w_2 for the second classifier is 21 and 33, respectively. The SVM models are trained

using the Radial Basis Function (RBF) kernel in the LIBSVM package.[24] Ten-fold cross-validation is used to evaluate our method. The data set is first divided into ten equal-sized subsets, each of which is evaluated, in turn, using the classifier trained on the remaining nine subsets.

Helix prediction

Feature selection and encoding

We select multiple input features based on both the sequence and the distinct structural parts of a TM helix and integrate them in the first SVM classifier (*H/~H*). Figure 1S of the *Supporting Information* shows the feature selection based on the structure of a TM helix. The total dimension of each peptide encoded by a single feature is the length of the vector multiplied by w_1 and the values of the feature vectors are scaled in the range of [0, 1]. The encoding of each feature is described below:

1) Amino acid (AA) composition: this feature contains sequence information about a TM helix. A vector of length 20 encodes the feature in which each residue of a peptide is represented by one in the position corresponding to the type of amino acid, and zeros otherwise.

2) Di-peptide (DP) composition: this feature is represented by the pair-residue helix propensity, $P(X, Y)$, where (X, Y) is an ordered pair of amino acids of X followed by Y. Each residue of a peptide is represented by a vector of length 20 in which the position corresponding to the type of amino acid of X holds a real value of its pair-residue helix propensity, $P(X, Y)$, and zeros otherwise.

3) Helix core feature: hydrophobicity is the main driving force behind protein folding in the membrane, and the helix core is populated by mainly hydrophobic residues.[25] We choose a

hydrophobicity scale (HS) recently determined by membrane insertion experiments.[26] Each residue of a peptide is represented by a vector of length 20 in which the position corresponding to the type of amino acid has a real value of its hydrophobicity, and zeros otherwise.

4) Helix caps feature: the helix-capping regions near the membrane-water interface show a preference for aromatic and polar residues.[25] These amino acids are characterized by an amphiphilic (AM) index that describes the local information at the helix-capping ends.[27] Each residue of a peptide is represented by a vector of length 20 in which the position corresponding to the type of amino acid has a real value of its amphiphilicity, and zeros otherwise.

5) Helix face feature: a TM helix can be described in terms of its surroundings, depending on whether it is buried or exposed to the lipids.[28] The predicted relative solvent accessibility (RSA) is obtained using the SABLE II server.[29] Each residue of a peptide is encoded by two vectors, each of length of 20. The positions in the vectors corresponding to the type of amino acid are represented by its predicted RSA and confidence, and zeros otherwise.

Determination of TMH candidates

Potential TMH segments are identified by determining if they are TMH candidates and subsequently assembling them into physical TMH segments. The algorithm proposed in the THUMBUP[30] program is modified as follows:

Step 1: Filtering Define a cut-off value, l_{min}, as the minimal length for a TMH candidate. A predicted helix segment is a TMH candidate if its length is no shorter than l_{min}. Otherwise, it is converted to a non-helix segment. A TMH candidate is considered for assembly in Steps 2 and 3.

Step 2: Extension The optimal TMH length, l_{opt}, is set at 21, to reflect the thickness of the hydrocarbon core of a lipid bilayer.[30] If the length of a TMH candidate is between l_{min} and l_{opt}, it is extended to l_{opt} from its N- and C-termini. Two or more TMH candidate helices are merged, if they overlap after the extension.

Step 3: Splitting Define l_{max}, as the cut-off value for the length of a TMH candidate to be split. A TMH candidate is split into two helices in the center if its length is greater than or equal to l_{max}. We also define l_{gap} as the number of loop residues to be inserted between the split helices.

We optimize l_{min}, l_{max}, and l_{gap} on the training data set. The optimized values of l_{min} and l_{max} for best prediction performance are 7 and 33, respectively. The optimized value of l_{gap} is 2.

Topology prediction

Feature selection and encoding

We use evolutionary profiles as the input feature to the second SVM classifier. PSI-BLAST[31] search is performed on the NCBI non-redundant database.[32] The E-value threshold is chosen as 10^{-5} and we allow 5 iterations for the generation of a Position Specific Scoring Matrix (PSSM). Each residue of a peptide is represented by a vector composed of 20 log-odds scores indicated by the PSSM. The total dimension of each peptide encoded by this feature is the length of the vector multiplied by w_2. The values of the feature vectors are scaled in the range of [0, 1].

Alternating geometric scoring function

In most cases, an integral membrane protein follows special constraints on its topology such that it must start with an inner (*i*) or outer (*o*) loop which alternates in order to connect the TM helices.

Therefore, the problem of predicting the topology of an integral membrane protein is reduced to predicting the orientation of the N-terminal loop. A number of membrane protein topology studies indicate that topogenesis is the result of various topogenic signals distributed among the loop regions, in which their overall effect is likely additive.[33,34] Furthermore, the two-stage membrane protein folding model suggests that topogenesis is established soon after the insertion of the membrane-spanning segments.[35,36] Based on these observations, we formulate the following assumptions about topogenesis in the AGSF: 1) the topology of the N-terminal loop is affected by the topogenic signals present in the loop regions along the entire protein sequence; and 2) the topogenic signals near the N-terminus are more likely a factor in the overall orientation of the N-terminal loop since they are inserted at an earlier stage. We incorporate these two assumptions in the development of AGSF in order to consider the diminishing contribution of topogenic signals in each downstream loop segment, the further away it is from the N-terminus.

In the AGSF, the contribution of topogenic signals is inversely proportional to the order of the loop segments in which they are located relative to the N-terminus in a geometric series: given a transmembrane protein that has n non-helical segments s_j ($1 \le j \le n$ and $n, j \in \mathbf{N}$) predicted in the first step. For each s_j of length $|s_j|$, we define two ratios, R_i and R_o, to represent the predicted ratios of the topology labels i and o, respectively:

$$R_i(j) = (\#\text{ of inner loop residues in } s_j / |s_j|) \times 100\% \quad (1)$$

$$R_o(j) = (\#\text{ of outer loop residues in } s_j / |s_j|) \times 100\%, \quad (2)$$

where $R_i + R_o = 100\%$. To determine the protein topology, we also define two topology scores, TS_i and TS_o, which represent, respectively, the contribution from topogenic signals for the N-terminal loop on the inside and the outside of the membrane:

$$TS_i = \sum_{j \text{ is odd}} W(j) \times R_i(j) + \sum_{j \text{ is even}} W(j) \times R_o(j) \quad (3)$$

$$TS_o = \sum_{j \text{ is even}} W(j) \times R_i(j) + \sum_{j \text{ is odd}} W(j) \times R_o(j) \quad (4)$$

$$W(j) = 1/b^{(j-1)\times EI}, b \text{ and } EI \in \mathbf{R}, \quad (5)$$

where b and EI denote the base and the exponent increment, respectively. $W(j)$ is a geometric function which assigns weights to the $R_i\ (j)$ and $R_o\ (j)$ terms. If $TS_i \geq TS_o$, then the predicted topology is inside (i); otherwise, it is outside (o). In Figure 2, we show the calculation of the topology scores TS_i and TS_o using AGSF as an example.

Positive-inside rule

The positive-inside rule is a strong topology determinant as established by several studies[7,8,33,34,37] and it is compared to AGSF. To predict the topology using the positive-inside rule, we calculate the number of positive residues (Arg and Lys) within 10 residues from the helix start or end and 5 residues into the helices. The positive residues are summed for both the odd- and the even-numbered loops. If the sum of the positive charges is greater in the odd- numbered of loops, then the predicted topology for the N-terminus is inside (i); otherwise, it is outside (o). In the case when the charge difference is zero, the prediction becomes undetermined and thus, we count it as a false prediction.

Evaluation metrics

To assess the prediction accuracy, we follow the evaluation measures as described by Chen *et al*.[23] There are three types of measures: per-protein, per-segment, and per-residue accuracy as listed in Table 1. We define an additional per-protein score, Q_{TM}, as the percentage of correctly predicted topology models (both the location of helices and topology predicted correctly). For per-residue measures, we also use Matthew's correlation coefficient (*MCC*), which is a more robust measure than using recall or precision alone.[38] In addition, we have used an overlap of at least 9 residues for a correctly predicted TMH segment, whereas many other methods have used a more relaxed criterion of 3 overlapping residues.[3,9,14] A correctly predicted TMH segment is defined as a one-to-one overlap with the true TMH segment.

RESULTS

Cross-validation results and comparison with other methods

The prediction accuracy of SVM*top* is validated on two data sets, including those experimentally verified by low- and high-resolution methods. Table 2 shows the cross-validation accuracy of SVM*top*, which performs very well in terms of the per-segment score $Q_{htm}^{\%obs}$, correctly predicting at least 93% of all observed TM helices. The percentage of proteins with all helices predicted correctly (Q_{ok}) reaches 73% for the low-resolution set. Most notably, topology prediction is achieved by SVM*top* with a high level of accuracy over 90% for the low-resolution data as well. The most stringent and relevant score, Q_{TM}, which calculates the percentage of proteins with both the location

of helices and topology predicted correctly, has a success rate of 69% and 63% for low- and high-resolution sets, respectively.

To allow an extensive comparison, we evaluated the benchmark data sets on ten widely used methods, namely TMHMM2[3], HMMTOP2[9], PHDhtm v.1.96[14], MEMSAT3[39], TopPred2[8], SOSUI 1.1[40], SPLIT4[41], ConPred II[15], Phobius[42], and PolyPhobius[43], as shown in Table 2. Since we did not have the cross-validation results of all the methods compared, we evaluated the benchmarks based on their online implementations. As a result, their accuracy may be over-estimated. The most important observation is that SVM*top* achieves the highest Q_{TM} for both data sets. This compares with the next best method, the recently updated MEMSAT3, which obtains 68% and 57% in Q_{TM} for low- and high- resolution data sets, respectively. Specifically, SVM*top* improves MEMSAT3 in Q_{TM} by 6% for the high-resolution proteins.

With respect to topology prediction, SVM*top* also performs competitively against all methods assessed, as reflected by *TOPO*. SVM*top* and MEMSAT3 are among the highest scoring methods for topology prediction with accuracy better than 80% for both data sets. For helix prediction, SVM*top* consistently achieves accuracy over 70% in Q_{ok}. This compares favorably with ConPred II, a consensus method that obtains a Q_{ok} score at about 75% for the low-resolution data set, but only correctly predicts 69% of the high-resolution set. Slight improvements are seen in per-segment scores ($Q_{htm}^{\%obs}$ and $Q_{htm}^{\%prd}$) by SVM*top*, which achieves accuracy in the range of 93%-95%. Lastly, in terms of per-residue accuracy, SVM*top* yields one of the highest scores, as measured by Q_2 (91%)

and *MCC* (0.81) in the high-resolution set, and ranks as a close second behind ConPred II and PolyPhobius for the low-resolution set.

Analysis of AGSF on topology prediction accuracy

The relationship between AGSF and topology prediction accuracy is analyzed in detail. Figure 3 shows the topology prediction accuracy achieved by applying different parameter values on the low-resolution data set. The topology prediction accuracy is indicated by colors and the white circles represent the best accuracy (90%). The parameter values of each circle are listed in Table 3S of the *Supporting Information*. The set of values of (*b*, *EI*) we used to calculate the topology scores in AGSF is (1.6, 1). The black dashed line at $b = 1$ divides the figure into two blocks, I on the left, and II on the right. Three interesting observations can be made from this figure. First, when $b < 1$ in Block I, the topology prediction accuracy is very low, especially for the upper-left region (<70%; grey). In this region, AGSF evaluates the downstream terms with an augmenting weight, which is equivalent to applying an increasing contribution of topogenic signals that are further away from the N-terminus. Second, in Block II, low topology accuracy (~81%; navy) occurs in the left-vertical and the lower-horizontal regions. In these two regions, AGSF is simplified to assigning an equal weight to all downstream signals in the loop regions. Thus, in such cases, AGSF considers the contribution from all topogenic determinants in the downstream sequence equally. Third, also in Block II, a slightly lower topology accuracy (87%; green) is observed in the upper-right region when both *b* and *EI* are large. In this region, AGSF assigns a very small weight to the signals present in the downstream loop segments. The only contribution considered in the calculation of the topology

scores in this case is dominated by the first N-terminal loop because the downstream signals are negligible.

The poor accuracy in these regions suggests that for better topology prediction accuracy, the inclusion of the following two assumptions in AGSF may be important: 1) topogenesis of membrane proteins is influenced by topology determinants distributed along the protein sequence and they should be considered collectively; and 2) the contribution of each downstream topology determinant to the topology of the N-terminal loop is not equal and diminishes as a function the further away it is from the N-terminus.

Discrimination of soluble and membrane proteins

An important function of a TM helix prediction program is the ability to discriminate between soluble and membrane proteins. We apply our method to a data set of 616 soluble proteins taken from Chen *et al.*[23] A cut-off length is established as the minimum length of a predicted helix. Any protein that does not have at least one helix greater than or equal to the minimum length is classified as a soluble protein. The cut-off length is chosen as the function that minimizes the sum of false positive and false negative rates for soluble and membrane proteins. In Figure 4, we show the selection process of the cut-off length that minimizes the sum of the error rates at 17 residues. A false positive represents that a soluble protein has been classified as a membrane protein. Conversely, a false negative indicates that a membrane protein has been classified as a soluble protein. There is a trade-off between the two depending on the chosen cut-off length.

We compare the error rates for several approaches in Table 3. Of all the methods compared, SVM*top* yields the lowest overall false positive rate of less than 0.5%, while keeping the overall false negative rates at less than 1.2%. Some methods, such as TopPred2 and ConPred II, have very low false negative rates (0.0%) at the expense of higher false positive rates (>8.0%). The highly accurate results achieved by SVM*top* demonstrate that it can effectively identify data sets containing a mixture of soluble and membrane proteins. Hence, this rapid and accurate method can characterize a given proteome by screening out membrane proteins.

Topology prediction accuracy of single-spanning and multi-spanning membrane proteins

To further characterize the effect of AGSF and the positive-inside rule on topology prediction, we compare their performance on single- and multi-spanning membrane protein classes in the high-resolution data set. Any protein predicted by SVM*top* with two or more TM helices is counted as multi-spanning. In the high-resolution data set, single-spanning membrane proteins account for approximately 1/3 of all proteins.

In Table 4, two interesting observations can be made from the above comparison. First, for the multi-spanning membrane proteins, the highest topology prediction accuracy (*TOPO*) is achieved by AGSF when $EI = 1$ (92%), followed by $EI = 0$ (90%), and lastly by $EI = \infty$ (87%). However, for the single-spanning proteins, the highest accuracy is achieved when $EI = \infty$ (83%). When $EI = \infty$, the scoring function is dominated by the first loop because the downstream signals are negligible. However, when *EI* is 0 or 1, the scoring function considers the topogenic signals in all loop segments with an equal weight or a diminishing weight, respectively. The results suggest that for

multi-spanning TM proteins, the assumptions we made about AGSF, the equivalent of *EI* = 1, may be important for topology prediction accuracy. However, for single-spanning TM proteins, this trend is less pronounced.

Second, the positive-inside rule successfully predicts both classes of membrane proteins with an accuracy of 73%-76%. In comparison, the AGSF (b = 1.6, *EI* = 1) achieves an accuracy (*TOPO*) of 81% and 92% for single-spanning and multi-spanning membrane proteins, respectively. Most notably, the AGSF improves the positive-inside rule by 16% in the multi-spanning class. This also suggests that under the assumptions in the AGSF, the topology of multi-spanning TM proteins can be predicted with higher confidence. One limitation of the positive-inside rule becomes apparent when the charge difference is zero and the prediction is thus undetermined. From our analysis, the undetermined proteins account for about 7% and 9% of our low- and high-resolution data sets, respectively.

DISCUSSION

From the results of several benchmarks, it is clear that SVM*top* is among the top-performing predictors for membrane protein topology, particularly in overall prediction accuracy on a per-protein basis (Q_{TM}). A fundamental difference between SVM*top* and the HMM-based methods (TMHMM2, HMMTOP2, Phobius, and PolyPhobius) is the architecture of prediction models. SVM*top* uses a hierarchical scheme that separates helix and topology prediction into two stages, whereas the HMM-based methods predict one of the three states of a residue directly. Although the latter tend to be less computationally expensive, using a hierarchical scheme may be more intuitive

for membrane biologists. This is supported by both experimental and theoretical considerations. Experimentally, TM domains can be distinguished from non-TM domains on the basis of hydrophobicity.[4-6] In two-stage membrane folding, the TM helices first partition laterally into the lipid bilayer, followed by their association and possible re-orientation of TM and non-TM domains.[35,36] Each stage of the folding process is governed by different factors. By using a hierarchical scheme, factors determining helix formation or topogenesis can be applied separately to improve the prediction in both stages. In fact, the two-stage architecture has been used in earlier methods such as TopPred2 and PHDhtm, before the HMM-based methods became the mainstream. Although SVM*top* employs a prediction framework that is highly similar to that used in TopPred2 and PHDhtm, the biological features used in each stage are quite different. The success of SVM*top* in predicting TM helices may be attributed to the combination of both sequence and structural features, whereas both TopPred2 and PHDhtm only use hydrophobicity or sequence information in the form of single residue compositions or a profile. SVM*top* uses several structural features corresponding to the helix core, caps, and faces, in which each feature is summarized by hydrophobicity, amphiphilicity, and relative solvent accessibility, respectively. In particular, relative solvent accessibility may play an important role in predicting the TM helices for multi-spanning membrane proteins because individual TM helices can be tightly packed into a bundle and, as a result, exhibit different patterns of exposure to lipids.[28,44] In contrast, single-spanning membrane proteins contain TM helices that are generally in contact with lipids; thus hydrophobicity is a determining factor. Amphiphilicity as a helix-capping feature may contribute to the helix prediction

at the membrane-water interface, as this region contains different amino acid preferences and side-chain snorkeling behavior.[45,46]

For the prediction of topology or the sidedness in the second stage, SVM*top* uses evolutionary information derived from PSL-BLAST profiles and discriminates inside and outside loop segments by calculating topology scores through AGSF. The analysis of AGSF's performance on topology prediction accuracy lends support to the assumptions we incorporated in AGSF, which are also particularly important for multi-spanning membrane proteins. It can be inferred that the topogeneses for single- and multi-spanning membrane proteins are likely influenced by different magnitudes of topogenic factors. Remarkably, our results are consistent with topogenesis studies of Type I membrane proteins in which cleavable signal sequences cause the N-terminal loop to translocate on the exoplasmic side.[33] In contrast, multi-spanning membrane proteins are more likely to have multiple topogenic signals; hence their final topology is a dynamic process in which changes in orientation depend on the contribution of the topogenic signals.

In applying the positive-inside rule to the prediction of sidedness, there is a significant portion (7-9%) of the data sets that remains undetermined as a result of the zero charge difference. Such a prediction outcome would be a rare event for SVM*top* because the calculation of topology scores is not a frequency count. Recently, controversies about membrane proteins that adopt 'dual' topologies have been resolved and the work by Rapp *et al*. provided a direct link between the positive-inside rule and the duality of membrane topology.[47] They showed that the topology of the multi-drug resistance protein, *EmrE*, can be altered by subtly manipulating the positive charges. Thus, the

positive-inside rule may be suitable for identifying such dual-topology membrane proteins. However, they are generally limited in number and have been estimated at less than 0.1% of the *E. coli* inner membrane proteome.[48] In fact, there are no dual-topology proteins in our benchmark data sets, as their topologies have been determined by various low- and high-resolution methods. In this respect, SVM*top* is able to compensate for the insufficiency of topogenic information when using the positive-inside rule solely, by increasing the coverage of prediction for those proteins carrying a zero charge bias.

One of the central concepts of AGSF is the topogenic contribution of each downstream loop segment given a weight as a function of its distance from the N-terminus. The contribution of individual residues to the overall topology is not directly observed because their topogenic information is diluted at the segment level. Furthermore, the topology score is calculated as the sum of all topogenic signals in their respective loop segments. Therefore, it is difficult to pin-point a specific residue responsible for the topology. The positive-inside rule, for example, is a statistical rule that correlates the sidedness with the difference of positive charges across the membrane. However, it is unlikely that every positively charged residue contributes equally to the topology. From the raw output of the SVM classifier, there are no over-represented patterns that correspond to strongly predicted topogenic features. One possible interpretation is that the SVM classifier is able to detect some as yet uncharacterized topogenic signals with strongly predicted topologies across the sequence and their contribution is considered collectively through AGSF.

In addition to a robust feature selection criterion, another ingredient for good performance in any kernel-based methods such as the SVM is an appropriate choice of kernel function. A formal discussion of this topic is beyond the scope of this work, but it is worth mentioning that choosing a suitable kernel function is largely context-dependent. In the development of SVM*top*, we favored a non-linear kernel function such as the RBF (radial basis function) over the linear kernel function because the relationship between the class labels (*H*/~*H*; *i*/*o*) and the features is likely non-linear. Second, since the number of instances (Stage 1: >50K; Stage 2: >37K) is much larger than the number of features in their respective stages (Stage 1: 80×w_1 ; Stage 2: 20×w_2), using a non-linear kernel may lead to better results.[49] As for polynomial kernel function, it has been noted that there are more hyperparameters in the polynomial kernel which influences the complexity of model selection.[49] One important aspect about SVM is that, regardless the type of kernel function used, the results do not allow direct interpretation on the sequence level because input sequences are first transformed into vectors before mapping via the kernel function in the feature space. Other discriminative machine-learning algorithms such as neural networks also share this property in which direct interpretations of results in terms of the input is difficult.

As more high-resolution structures continue to accumulate, more complex architectures in TM helices are unveiled. Short helices that do not fully traverse the lipid bilayer are frequently observed in many transporters and channels.[50] Since these helices are shorter on average, they are easily missed by current prediction methods. Furthermore, short helices as re-entrant regions that return to the same side of the membrane are exceptions to the alternating topological connectivity in

extra-membranous loops. A typical example can be observed in the structure of aquaporin, in which the NPA domain is linked by a short helix and the connecting loops exit on the same side of the membrane.[51] Interestingly, the prediction by SVM*top* detects all TM helices, and the C-terminal loop is predicted to be localized on the extracellular side. However, the high-resolution structure shows that both termini are localized on the cytoplasmic side. One possible explanation is that SVM*top* enforces the predicted topology to obey the alternating connectivity for loop domains through AGSF. The majority of topology prediction methods that embed this biological grammar as part of their model architecture do not cope well with the violations as in the case of re-entrant regions. This can be explained with the recently modified two-stage membrane protein folding model, which includes a third stage when short helices or loops are inserted in the membrane.[52] The three-stage model is supported by the folding of potassium channel KcsA, which contains a short pore (P) helix and a selectivity filter in each subunit of the tetramer.[53] It is possible that the more hydrophobic TM segments (S1 and S2) insert in the membrane first and associate through helix-helix interactions, thereby creating internal spaces for the subsequent folding of the less hydrophobic P-helix. For the prediction of re-entrant structures, we believe that current methods would benefit from including a third stage by relaxing the constraints built into the prediction models or developing a rule-based approach. In addition, a study on re-entrant regions concluded that they carry a higher content of aromatic residues.[54] This information may be useful in developing algorithms for those irregular structures.

A common issue with many TM predictors is the presence of signal peptides (SPs). It has been shown that due to the hydrophobic h-region in SPs, they are often falsely predicted as TM helices.[55,56] Therefore, using a signal peptide prediction method such as SignalP as a pre-screening step would likely remedy the situation.[57] Alternatively, TM predictors that integrate an SP prediction method can also identify potential SPs from the sequence. One particularly successful example is Phobius, which has shown a significant reduction in false classification rates.[43] In the future development of SVM*top*, we hope to address this problem by incorporating a signal peptide predictor. The inclusion of such an extension would undoubtedly result in better discrimination between SPs and TM helices.

CONCLUSION

Although integral membrane proteins are abundant in many genomes, relatively little is known about their three-dimensional structures. Therefore, an accurate topology model that serves as an intermediate step is highly valuable. We have presented a new approach using hierarchical SVM classifiers and achieved marked improvements in both stages of helix and topology prediction. Our method can also discriminate between soluble and membrane proteins with a high degree of accuracy and sensitivity. From the analysis of AGSF, we observe that, for multi-spanning TM proteins in particular, better accuracy can be obtained by considering the contribution of downstream topogenic signals in a diminishing fashion. This finding indicates that topogenesis is governed by several topogenic determinants that should be considered collectively when predicting the topology of a multi-spanning TM protein. Our work adds to an expanding collection of TM helix and

topology prediction methods, which are essential for the determination and modeling of the three-dimensional structure of integral membrane proteins. The predictions will provide information for further structural and functional studies on membrane proteomes.

ACKNOWLEDGEMENT

We are grateful to Dr. Einer Rødland for his critical reading of the manuscript; and to Jia-Ming Chang, Hsin-Nan Lin, and Wen-Chi Chou for stimulating discussions and providing computational assistance. This work was supported in part by the thematic program of Academia Sinica under grant AS94B003 and AS95ASIA02.

SUPPORTING INFORMATION AVAILABLE

The table listing of the membrane proteins used in this work; figure showing feature selection for helix prediction; the table listing of optimal AGSF parameter values. This material is available free at http://pubs.acs.org.

REFERENCES

1. Ubarretxena-Belandia, I.; Engelman, D. M. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* **2001**, 11(3), 370-6.

2. Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **1998**, 7(4), 1029-38.

3. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001**, 305(3), 567-80.

4. Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, 157(1), 105-32.

5. Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U S A* **1984**, 81(1), 140-4.

6. White, S. H.; Wimley, W. C. Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, 28, 319-65.

7. von Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *Embo J.* **1986**, 5(11), 3021-3027.

8. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **1992**, 225(2), 487-94.

9. Tusnady, G. E.; Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **1998**, 283(2), 489-506.

10. Viklund, H.; Elofsson, A. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* **2004**, 13(7), 1908-17.

11. Martelli, P. L.; Fariselli, P.; Casadio, R. An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* **2003**, 19 Suppl 1, i205-11.

12. Yuan, Z.; Mattick, J. S.; Teasdale, R. D. SVMtm: support vector machines to predict transmembrane segments. *J. Comput. Chem.* **2004**, 25(5), 632-6.

13. Lo, A.; Chiu, H.S.; Sung, T.Y.; Hsu, W.L. Transmembrane helix and topology prediction using hierarchical SVM classifiers and an alternating geometric scoring function. *Proceedings of the 2006 IEEE Computational Systems Bioinformatics Conference* **2006**, 31-42.

14. Rost, B.; Fariselli, P.; Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **1996**, 5(8), 1704-18.

15. Arai, M.; Mitsuke, H.; Ikeda, M.; Xia, J. X.; Kikuchi, T.; Satake, M.; Shimizu, T. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.* **2004**, 32, W390-3.

16. Amico, M.; Finelli, M.; Rossi, I.; Zauli, A.; Elofsson, A.; Viklund, H.; von Heijne, G.; Jones, D.; Krogh, A.; Fariselli, P.; Luigi Martelli, P.; Casadio, R. PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res.* **2006**, 34, W169-72.

17. Moller, S.; Kriventseva, E. V.; Apweiler, R. A collection of well characterised integral membrane proteins. *Bioinformatics* **2000**, 16(12), 1159-60.

18. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* **1997**, 75(5), 312-6.

19. Jayasinghe, S.; Hristova, K.; White, S. H. MPtopo: A database of membrane protein topology. *Protein Sci.* **2001**, 10(2), 455-8.

20. Ikeda, M.; Arai, M.; Okuno, T.; Shimizu, T. TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.* **2003**, 31(1), 406-9.

21. Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, 22(13), 1658-9.

22. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28(1), 235-42.

23. Chen, C. P.; Kernytsky, A.; Rost, B. Transmembrane helix predictions revisited. *Protein Sci.* **2002**, 11(12), 2774-91.

24. Chang, C.C.; Lin, C.J. LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

25. Ulmschneider, M. B.; Sansom, M. S.; Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **2005**, 59(2), 252-65.

26. Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **2005**, 433(7024), 377-81.

27. Mitaku, S.; Hirokawa, T.; Tsuji, T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **2002**, 18(4), 608-16.

28. Beuming, T.; Weinstein, H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics* **2004**, 20(12), 1822-35.

29. Adamczak, R.; Porollo, A.; Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **2004**, 56(4), 753-67.

30. Zhou, H.; Zhou, Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci.* **2003**, 12(7), 1547-55.

31. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, 25(17), 3389-402.

32. Wheeler, D. L.; Chappey, C.; Lash, A. E.; Leipe, D. D.; Madden, T. L.; Schuler, G. D.; Tatusova, T. A.; Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2000**, 28(1), 10-4.

33. Goder, V.; Spiess, M. Topogenesis of membrane proteins: determinants and dynamics. *FEBS Lett.* **2001**, 504(3), 87-93.

34. Kida, Y.; Morimoto, F.; Mihara, K.; Sakaguchi, M. Function of positive charges following signal-anchor sequences during translocation of the N-terminal domain. *J. Biol. Chem.* **2006**, 281(2), 1152-8.

35. Popot, J. L.; Engelman, D. M. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* **1990**, 29(17), 4031-7.

36. Popot, J. L.; Engelman, D. M. Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* **2000**, 69, 881-922.

37. Goder, V.; Junne, T.; Spiess, M. Sec61p contributes to signal sequence orientation according to the positive-inside rule. *Mol. Biol. Cell.* **2004**, 15(3), 1470-8.

38. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **1975**, 405(2), 442-51.

39. Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007,** 23(5), 538-44.

40. Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **1998**, 14(4), 378-9.

41. Juretic, D.; Zoranic, L.; Zucic, D. Basic charge clusters and predictions of membrane protein topology. *J. Chem. Inf. Comput. Sci.* **2002**, 42(3), 620-32.

42. Kall, L.; Krogh, A.; Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004,** 338(5), 1027-36.

43. Kall, L.; Krogh, A.; Sonnhammer, E. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **2005,** 21 Suppl 1, i251-7.

44. Adamian, L.; Liang, J. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct. Biol.* **2006,** 6, 13.

45. Chamberlain, A. K.; Lee, Y.; Kim, S.; Bowie, J. U. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *J. Mol. Biol.* **2004,** 339(2), 471-9.

46. Granseth, E.; von Heijne, G.; Elofsson, A. A study of the membrane-water interface region of membrane proteins. *J. Mol. Biol.* **2005,** 346(1), 377-85.

47. Rapp, M.; Seppala, S.; Granseth, E.; von Heijne, G. Emulating membrane protein evolution by

rational design. *Science* **2007,** 315(5816), 1282-4.

48. Daley, D. O.; Rapp, M.; Granseth, E.; Melen, K.; Drew, D.; von Heijne, G. Global topology analysis of the Escherichia coli inner membrane proteome. *Science* **2005**, 308(5726), 1321-3.

49. Hsu, C. W.; Chang, C. C.; Lin, C. J. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

50. Cuthbertson, J. M.; Doyle, D. A.; Sansom, M. S. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.* **2005**, 18(6), 295-308.

51. Sui, H.; Han, B. G.; Lee, J. K.; Walian, P.; Jap, B. K. Structural basis of water-specific transport through the AQP1 water channel. *Nature* **2001**, 414(6866), 872-8.

52. Engelman, D. M.; Chen, Y.; Chin, C. N.; Curran, A. R.; Dixon, A. M.; Dupuy, A. D.; Lee, A. S.; Lehnert, U.; Matthews, E. E.; Reshetnyak, Y. K.; Senes, A.; Popot, J. L. Membrane protein folding: beyond the two stage model. *FEBS Lett.* **2003,** 555(1), 122-5.

53. Cordero-Morales, J. F.; Cuello, L. G.; Zhao, Y.; Jogini, V.; Cortes, D. M.; Roux, B.; Perozo, E. Molecular determinants of gating at the potassium-channel selectivity filter. *Nat. Struct. Mol. Biol.* **2006,** 13(4), 311-8.

54. Viklund, H.; Granseth, E.; Elofsson, A. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J. Mol. Biol.* **2006**, 361(3), 591-603.

55. Moller, S.; Croning, M. D.; Apweiler, R. Evaluation of methods for the prediction of membrane

spanning regions. *Bioinformatics* **2001,** 17(7), 646-53.

56. Lao, D. M.; Arai, M.; Ikeda, M.; Shimizu, T. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics* **2002,** 18(12), 1562-6.

57. Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **2004,** 340(4), 783-95.

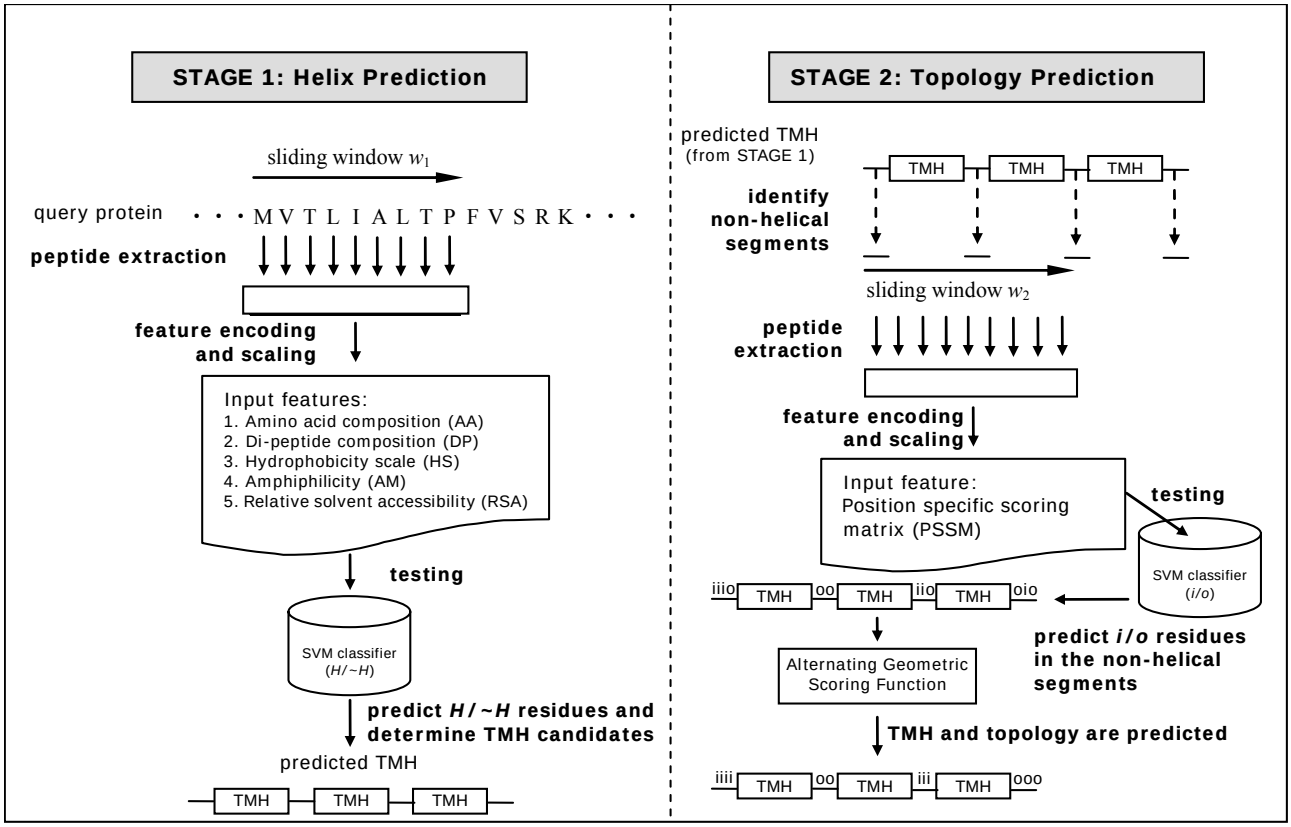

Figure 1. Flowchart of SVM*top*. The left panel describes Helix Prediction of Stage 1 in the order of: peptide extraction by sliding windows; feature encoding and scaling; prediction of helix (*H*) and non-helix (~*H*) residues; determination of TMH candidates. The right panel describes Topology Prediction of Stage 2 in the order of: identify non-helical segments; peptide extraction by sliding windows; feature encoding and scaling; prediction of inside (*i*) and outside (*o*) residues; applying AGSF to obtain the final topology.

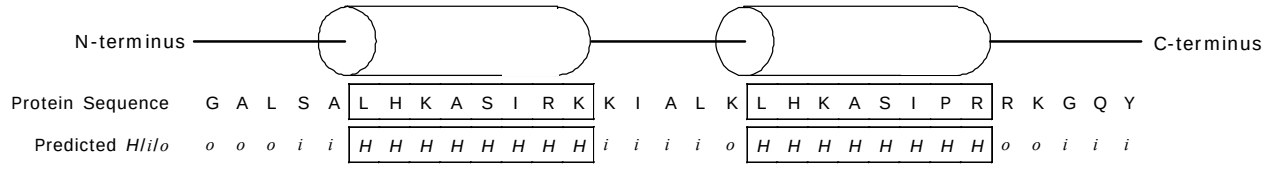

Figure 2. A hypothetical protein as an example for topology prediction using AGSF. Helix residues (*H*, enclosed by cylinders) are predicted by SVM*top* before the topology prediction (*i*/*o*). Predicted loop segments can have more than one type of topology (*i*, *o*) from the raw output of prediction on a residue basis. First, we calculate the ratios of predicted topology labels (*i*/*o*) in each loop segment. From this example, $R_i(1) = 2/5$, $R_o(1) = 3/5$, $R_i(2) = 4/5$, $R_o(2) = 1/5$, $R_i(3) = 3/5$, and $R_o(3) = 2/5$. Given a set of optimal values of (b, EI) = (1.6, 1), the Topology Scores (TS_i and TS_o) are calculated using AGSF as follows: $TS_i = [1/(1.6^0)]\times R_i(1) + [1/(1.6^1)]\times R_o(2) + [1/(1.6^2)]\times R_i(3) \fallingdotseq 0.7594$. Similarily, $TS_o \fallingdotseq 1.2563$. Here, $TS_o > TS_i$, therefore, the final topology for the N-terminal loop is predicted as outside (*o*).

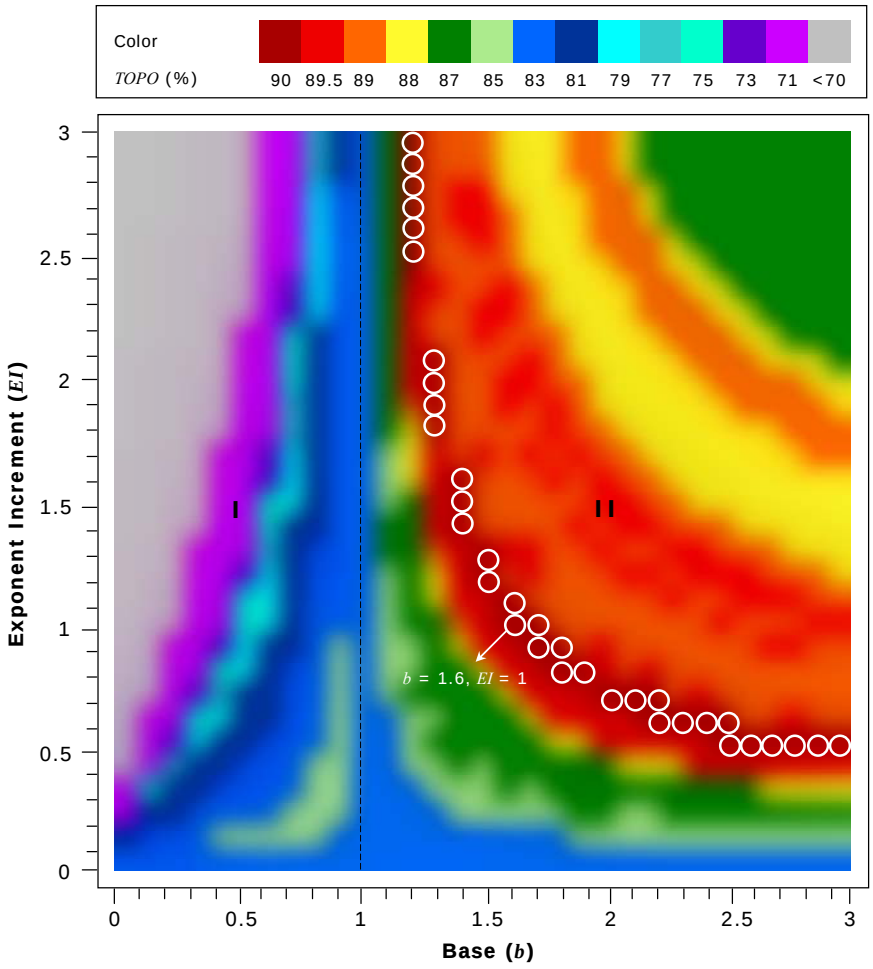

Figure 3. The relationship of base (*b*) and exponent increment (*EI*) in the AGSF and topology prediction accuracy (*TOPO*) in the low-resolution data set. The horizontal-axis: base (*b*); the vertical-axis: exponent increment (*EI*). Topology prediction accuracy is divided into 14 levels and each is coded by a color. A black dashed line at $b = 1$ divides the figure into Block I and II. In Block I, low topology prediction accuracy is observed, especially in the upper-left regions (70-71%). In Block II, the best accuracy (90.21%) occurs within the white circles along the region in dark red. The value of (*b*, *EI*) chosen in the AGSF for topology prediction is (1.6, 1).

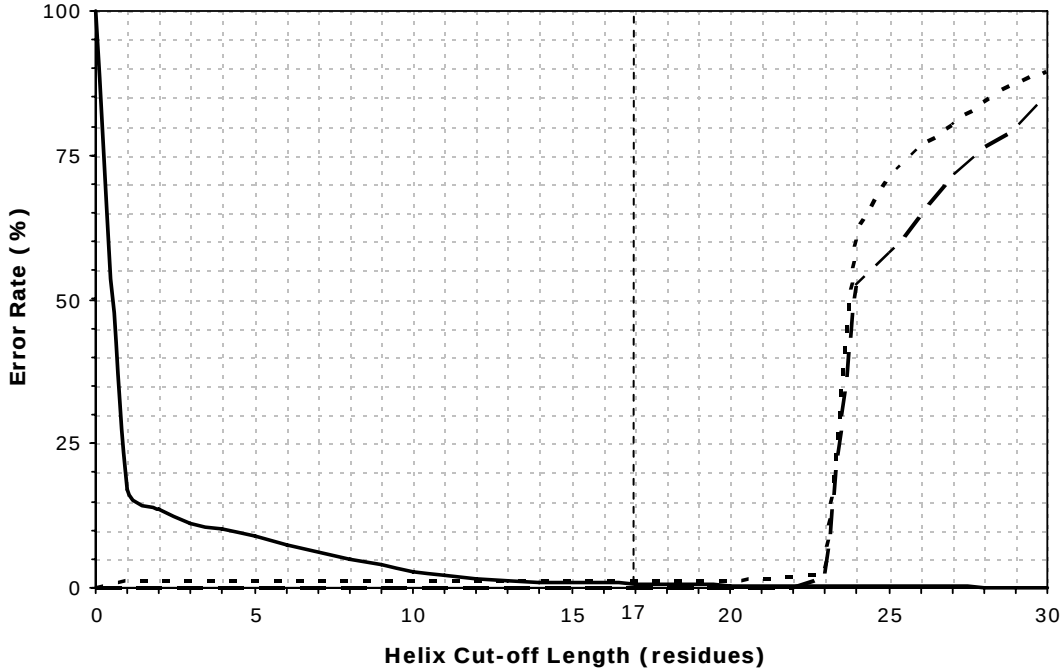

Figure 4. Determining the helix cut-off length for the discrimination of soluble and membrane proteins. The horizontal-axis: helix cut-off length; the vertical-axis: error rate (%). False positive rate is indicated by the black line. False negative rates for low- and high-resolution data sets are indicated by the dashed and the dotted lines, respectively. The vertical dashed line at 17 residues is chosen to minimize the sum of all three errors.

Table 1. Evaluation metrics used in this work. Per-protein measures include Q_{ok}, *TOPO*, and Q_{TM}. Per-segment measures include $Q_{htm}^{\%obs}$ and $Q_{htm}^{\%prd}$. Per-residue measures include Q_2, $Q_{2T}^{\%obs}$, $Q_{2T}^{\%prd}$, and *MCC*. N_{prot} is the number of proteins in a data set; *TP*: true positive; *TN*: true negative; *FP*: false positive; *FN*: false negative.

Symbol	Formula	Description
Q_{ok}	$\frac{\sum_{i}^{N_{prot}} \delta_i}{N_{prot}} \times 100\%$, with $\delta_i = \begin{cases} 1, \text{ if } Q_{htm}^{\%obs} \wedge Q_{htm}^{\%prd} = 100\% \text{ for protein } i \\ 0, \text{ otherwise} \end{cases}$	percentage of proteins in which all its TMH segments are predicted correctly
TOPO	$\frac{\text{number of proteins with correctly predicted topology}}{N_{prot}} \times 100\%$	percentage of correctly predicted topology
Q_{TM}	$\frac{\text{number of proteins with all TMH segments and topology predicted correctly}}{N_{prot}} \times 100\%$	percentage of proteins in which both TMH segments and topology are predicted correctly
$Q_{htm}^{\%obs}$	$\frac{\text{number of correctly predicted TM in data set}}{\text{number of TM observed in data set}} \times 100\%$	TMH segment recall
$Q_{htm}^{\%prd}$	$\frac{\text{number of correctly predicted TM in data set}}{\text{number of TM predicted in data set}} \times 100\%$	TMH segment precision
Q_2	$\frac{\sum_{i}^{N_{prot}} \frac{\text{number of residues predicted correctly in protein } i}{\text{number of residues in prtoein } i}}{N_{prot}} \times 100\%$	averaged percentage of correctly predicted TMH residues of all proteins
$Q_{2T}^{\%obs}$	$\frac{\text{number of residues correctly predicted in TM helices}}{\text{number of residues observed in TM helices}} \times 100\%$	TMH residue recall
$Q_{2T}^{\%prd}$	$\frac{\text{number of residues correctly predicted in TM helices}}{\text{number of residues predicted in TM helices}} \times 100\%$	TMH residue precision
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, where TP : number of correctly predicted helix residues TN : number of correctly predicted non-helix residues FP : number of incorrectly predicted helix residues FN : number of incorrectly predicted non-helix residues	Matthew's correlation coefficient for TMH residues

Table 2. Performance of prediction methods for low- and high-resolution data sets. The value of (*b*, *EI*) used for SVM*top* is (1.6, 1). SOSUI 1.1 does not predict the topology of an integral membrane protein, thus *TOPO* and Q_{TM} are not available.

Method	Low-resolution								
	Per-protein (%)			Per-segment (%)		Per-residue (%)			
	Q_{ok}	*TOPO*	Q_{TM}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	*MCC*
SVM*top*	**73.29**	**90.21**	**69.23**	**94.76**	**93.94**	**89.23**	**87.50**	**80.35**	**0.77**
TMHMM2	68.53	76.22	58.74	90.39	93.52	89.23	82.82	83.03	0.76
HMMTOP2	64.34	72.03	55.94	89.96	93.78	87.89	79.36	84.37	0.75
PHDhtm v.1.96	39.86	67.83	29.37	76.27	85.76	85.35	81.71	76.59	0.71
MEMSAT3	70.63	88.11	67.83	91.56	90.24	87.91	84.54	77.63	0.73
TopPred2	57.34	66.43	42.66	86.75	91.13	88.00	76.85	82.90	0.72
SOSUI 1.1	63.64	-	-	88.36	91.55	87.00	80.41	78.66	0.71
SPLIT4	72.73	83.22	64.34	93.45	91.32	88.07	87.56	76.88	0.74
ConPred II	74.83	83.92	65.04	94.76	92.21	90.07	84.37	84.13	0.78
Phobius	72.03	75.52	60.84	92.87	93.14	88.92	83.92	82.57	0.77
PolyPhobius	71.33	77.62	61.54	94.47	91.54	89.75	86.84	83.11	0.79
Method	High-resolution								
	Per-protein (%)			Per-segment (%)		Per-residue (%)			
	Q_{ok}	*TOPO*	Q_{TM}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	*MCC*
SVM*top*	**72.09**	**82.17**	**62.79**	**92.78**	**94.46**	**90.90**	**87.84**	**84.36**	**0.81**
TMHMM2	59.30	68.99	46.12	86.93	93.78	87.70	78.59	83.55	0.74
HMMTOP2	65.89	75.97	52.71	90.34	89.98	87.68	78.30	82.30	0.73
PHDhtm v.1.96	38.37	60.47	25.58	74.43	84.59	84.55	78.28	78.03	0.70
MEMSAT3	64.84	81.40	56.64	87.67	91.09	87.16	79.64	78.84	0.71
TopPred2	50.39	65.12	37.21	84.50	90.05	86.96	74.06	82.47	0.71
SOSUI 1.1	56.98	-	-	85.06	92.17	86.15	76.88	80.02	0.71
SPLIT4	65.12	79.07	54.65	89.77	91.56	87.12	83.84	78.00	0.73
ConPred II	69.14	77.34	55.43	90.94	91.31	88.63	79.99	84.17	0.75
Phobius	67.05	70.93	54.65	88.72	93.58	87.81	79.42	83.76	0.75
PolyPhobius	67.44	72.48	55.81	90.91	91.28	88.79	82.66	83.34	0.77

Table 3. Error rates between soluble and membrane proteins of all methods compared. False positive (FP): a soluble protein classified as a membrane protein; false negative (FN): a membrane protein classified as a soluble protein.

Method	False Positive Rate(%)	False Negative Rates(%)	
		Low-resolution	High-resolution
SVM*top*	**0.49**	**0.00**	**1.16**
TMHMM2	1.00	3.50	6.20
HMMTOP2	6.98	6.30	1.94
PHDhtm v.1.96	2.00	4.90	5.43
MEMSAT3	2.11	5.59	6.98
TopPred2	24.35	0.00	0.00
SOSUI 1.1	0.97	4.90	6.59
SPLIT4	7.47	0.00	1.55
ConPred II	8.77	0.00	0.39
Phobius	2.11	3.50	5.81
PolyPhobius	3.73	1.40	3.10

Table 4. Comparison between the AGSF and the positive-inside rule on topology (sidedness) prediction accuracy (*TOPO*) for single- and multi-spanning membrane proteins. The numbers inside the parentheses indicate the number of single- and multi-spanning proteins out of a total of 258.

Membrane protein type	Method			
	AGSF			Positive-inside rule
	$EI = 1$	$EI = 0$	$EI = \infty$	
Single-spanning (86/258)	81.40	79.07	82.56	73.26
Multi-spanning (172/258)	91.86	89.53	87.21	75.58

* The value of base (*b*) in the calculation of AGSF is 1.6 for all values of *EI*.

SYNOPSIS TOC

SVM*top* uses a hierarchical classification scheme of support vector machines to predict the transmembrane helix and topology. Helix prediction is accomplished by integrating biological features that describe a transmembrane helix in a lipid environment. We also develop a novel scoring function which models inter-loop topogenic interactions to predict the topology. Standard benchmarks show improved accuracy compared to existing methods in both helix and topology predictions, as well as soluble and membrane protein discrimination.