

TRANSMEMBRANE HELIX AND TOPOLOGY PREDICTION USING HIERARCHICAL SVM CLASSIFIERS AND AN ALTERNATING GEOMETRIC SCORING FUNCTION

Allan Lo^{1,2}, Hua-Sheng Chiu³, Ting-Yi Sung³, 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*

Email: {allanlo, huasheng, tsung, hsu}@iis.sinica.edu.tw

Motivation: A key class of membrane proteins contains one or more transmembrane (TM) helices, traversing the membrane lipid bilayer. Various properties such as the length, arrangement and topology or orientation of TM helices, are closely related to a protein's functions. Although a range of methods have been developed to predict TM helices and their topologies, no single method consistently outperforms the others. In addition, topology prediction has much lower accuracy than helix prediction, and thus requires continuous improvements.

Results: We develop a method based on support vector machines (SVM) in a hierarchical framework to predict TM helices first, followed by their topology. By partitioning the prediction problem into two steps, specific input features can be selected and integrated in each step. We also propose a novel scoring function for topology models based on membrane protein folding process. When benchmarked against other methods in terms of performance, our approach achieves the highest scores at 86% in helix prediction (Q_2) and 91% in topology prediction ($TOPO$) for the high-resolution data set, resulting in an improvement of 6% and 14% in their respective categories over the second best method. Furthermore, we demonstrate the ability of our method to discriminate between membrane and non-membrane proteins, with higher than 99% in accuracy. When tested on a small set of newly solved structures of membrane proteins, our method overcomes some of the difficulties in predicting TM helices by incorporating multiple biological input features.

1. INTRODUCTION

Integral membrane proteins constitute a wide and important class of biological entities that are crucial for life, representing about 25% of the proteins encoded by several genomes¹⁻³. They also play a key role in various cellular processes including signal and energy transduction, cell-cell interactions, and transport of solutes and macromolecules across membranes⁴. Despite their biological importance, the proportion of available high-resolution structures is exceedingly limited at about 0.5% of all solved structures⁵, compared to that of globular proteins deposited in the Protein Data Bank (PDB)⁶. In the absence of a high-resolution structure, an accurate structural model is important for the functional annotation of membrane proteins. A membrane protein structural model defines the number and location of transmembrane helices (TMHs) and the orientation or topology of the protein relative to the lipid bilayer. However, experimental approaches for identifying membrane protein structural models are time-consuming⁷. Therefore, bioinformatics development in sequence-based prediction methods is valuable for elu-

cidating the structural genomics of membrane proteins.

Many different methods have been developed to predict structural models of transmembrane helix (TMH) proteins. Earlier approaches relied on physico-chemical properties such as hydrophobicity⁸⁻¹⁰ to identify TMH regions. Recently, more advanced methods using hidden Markov models^{3,11} and neural networks¹² have been developed, and they have achieved significant improvements in prediction accuracy. Although several methods are available, none of them have integrated multiple biological input features in a machine-learning framework. Furthermore, an evaluation study¹³ concluded that current accuracies were over-estimated, and topology prediction remained a major challenge.

In this paper, we propose a machine-learning approach called SVM tmh (SVM for *transmembrane helix* prediction) in a hierarchical classification framework to predict membrane protein structure. We divide the prediction task into two successive steps by using a tertiary classifier consisting of two hierarchical binary classifiers. The number and location of TMHs are predicted in the first step, followed by the prediction of the topology in the second step.

* Corresponding author.

Our key contributions are as follows: 1) By decomposing the prediction into two steps, we reduce the complexity involved in each step, and biological input features relevant to each classifier can be applied. 2) We select multiple input features, including those based on different structural parts of a TMH protein, and integrate them to predict helices. 3) For topology prediction, we propose a novel topology scoring function based on the current understanding of membrane protein insertion. To the best of our knowledge, the proposed topology scoring function is the first model to capture the relationship between topogenic factors and topology formation.

The performance of *SVMtmh* is compared with other methods across several benchmark data sets and *SVMtmh* achieves a marked improvement in both helix and topology prediction. Specifically, *SVMtmh* achieves the highest score at 91% for topology prediction (*TOPO*) and 86% for helix prediction (Q_2) in the high-resolution data set, an improvement of 14% and 6%, respectively, compared to the second highest score. In addition, *SVMtmh* yields the lowest false positive rate at 0.5% when tested for discrimination between membrane and non-membrane proteins. Finally, we apply *SVMtmh* to analyze a newly solved structure of bacteriorhodopsin (bR) and show that our method can

provide the correct structural model which is in close agreement with the structure obtained through X-ray crystallography. We also provide a detailed analysis of the comparison with other methods and conclude with a summary and directions for future work.

2. METHODS

2.1. System architecture

The proposed approach uses hierarchical binary classifiers to predict the helices and topology of an integral membrane protein. We represent the problem of membrane protein structure prediction as a multiple classification process and solve it in two steps using hierarchical SVM classifiers. The overall framework is described in this section.

Each residue of a TMH protein can be regarded as belonging to one of the three classes defined by its position with respect to the membrane: inner (*i*) loop, transmembrane helix (*H*), and outer (*o*) loop. The aim of predicting membrane protein structures is to identify the correct class of each residue. Since there are three classes for a protein sequence, we design a tertiary classifier, which consists of two binary classifiers in a hierarchical structure. An overview of the system architecture is shown in Fig. 1.

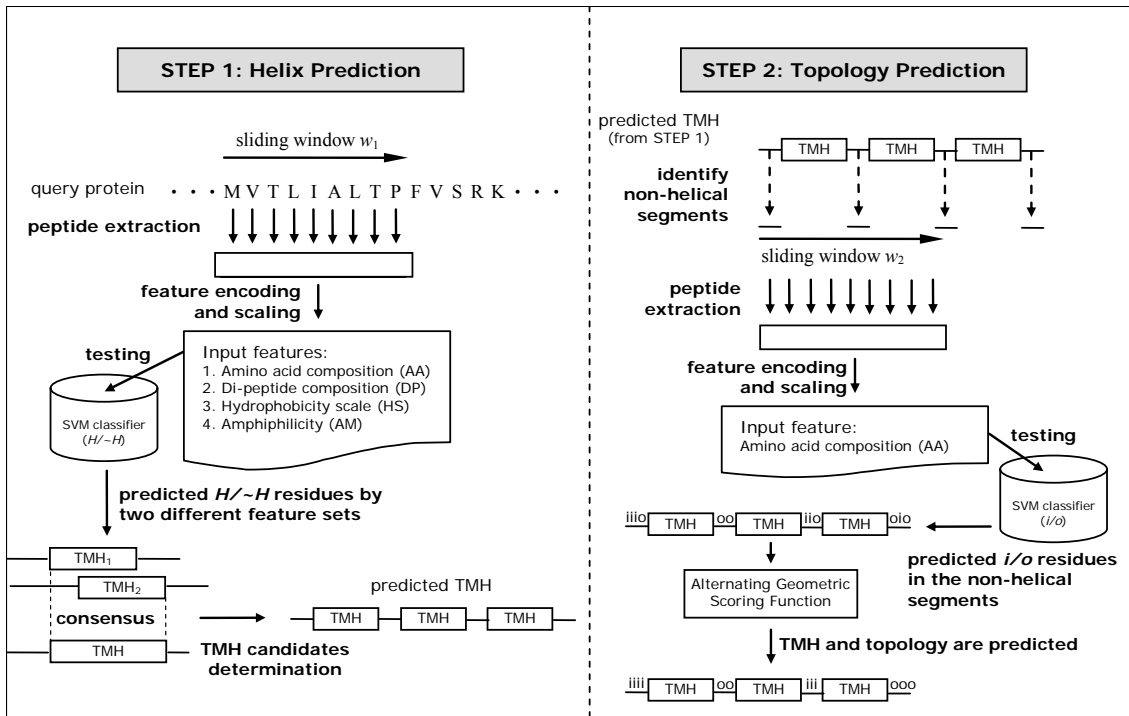


Fig. 1. Overview of the *SVMtmh* system architecture.

In Step 1, TM and non-TM residues ($H/\sim H$) are predicted. We use different feature sets (Section 3.3) to train our SVM classifiers and then combine the results from the best two combinations into a consensus prediction, which is screened for TMH candidates and subsequently assembled into physical TMH segments. In Step 2, the remaining non-helix residues ($\sim H$) from Step 1 are classified as either inner or outer (i/o) residues. To determine if the topology of the protein is an inner (i) or an outer loop (o), we apply the proposed alternating geometric scoring function. Briefly, the classification framework is performed in two steps, each of which uses an associated binary classifier ($H/\sim H$, i/o).

We use sliding windows to partition a protein sequence into peptides. The optimal length of the sliding window, w , is incrementally searched from 3 to 41 for both classifiers. The optimal window sizes, w_1 for the first classifier and w_2 for the second classifier, are found to be 21 and 29, respectively.

2.2. Training and testing

We train our classifiers with the LIBSVM package¹⁴ and Radial Basis Function (RBF) is chosen as the kernel function. The associated parameters (C , γ) are optimized at (1.8661, 0.1250). The cost weight is adjusted to avoid under-prediction in unbalanced data sets. Since the helix and non-helix classes make up about 30% and 70%

of the data set respectively, we set the cost weight at 7/3 for the first classifier. Similarly, we set the cost weight at 1/1 for the second classifier to reflect the proportion of the inner and outer loop classes in the data set. Ten-fold cross-validation is used to evaluate our method. The data set is first divided into ten subsets of equal size. Each subset is in turn tested using the classifier trained on the remaining nine subsets. Since each residue of the whole data set is only predicted once, the overall prediction accuracy is the percentage of correctly predicted residues. The values in the feature vectors are scaled in the range of [0, 1].

2.3. Helix prediction

2.3.1. Feature selection and extraction

The choice of relevant features is critical in any prediction models. Thus, in the present study we select features that capture important relationships between a sequence and the structure. TMH proteins are subject to global constraints of the lipid bilayer since they contain membrane-spanning helices¹⁵. Additionally, TM helices can be divided into distinct local structural parts, including the core and end regions based on the propensity of amino acids³. Fig. 2 shows the selection of features to capture both the global and local information of a TM helix. The representation of each feature is described below:

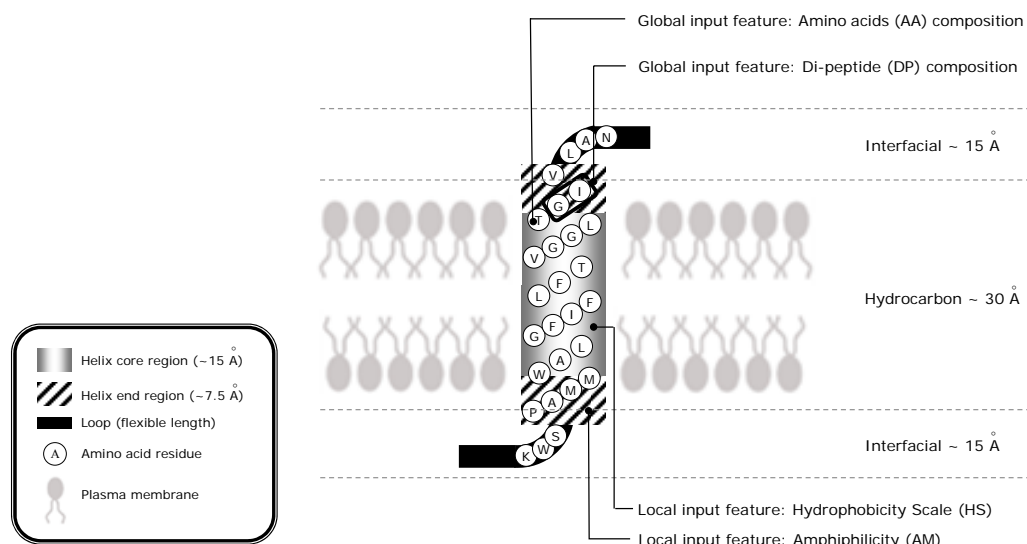


Fig. 2. Transmembrane (TM) helix structure in the lipid bilayer: helix core and end regions. Loops connect between the adjacent TM helices. We select global and local input features to capture information contained in a TM helix. Global input features: amino acid (AA) and di-peptide (DP) compositions. Local input features: hydrophobicity scale (HS)¹⁶ and amphiphilicity (AM)¹⁷. The helix core region is surrounded by an aliphatic hydrocarbon layer about 30 Å in thickness. The helix end regions are embedded in the water-membrane interface of about 15 Å.

1. *Amino acid composition (AA)*: This basic feature enables us to capture the global information of a TM helix. Each residue of a peptide is represented by a vector of length of 20 which is indicated by 1 in the position corresponding to the amino acid type of the residue, and 0 otherwise.
2. *Di-peptide composition (DP)*: We consider the coupling effect of two adjacent residues that contain global information along the sequence. This feature is represented by the pair-residue occurrence probability, $P(X,Y)$ where (X,Y) is an ordered pair of amino acids of X followed by Y. The vector space of this feature input comprises 400 dimensions.
3. *Helix core feature*: Hydrophobicity is used to capture local information within the core region of a TM helix where it is a major stabilizing factor¹⁶. We select a hydrophobicity scale (HS) recently determined by membrane insertion experiments¹⁶. Each residue is represented by a vector of length 20 that has a real value corresponding to its hydrophobicity.
4. *Helix ends feature*: The end regions of a TM helix near the membrane-water interface exhibit a preference for aromatic and polar residues, as shown in amino acid propensity studies^{16,17}. We select an amphiphilicity (AM) index¹⁷ as a input feature to capture the local information contained in the helix-capping ends. Each residue is represented by a vector of length 20 that has a real value corresponding to its amphiphilicity.

2.3.2. Determination of TMH candidates

To identify potential TMH regions, it is necessary to determine if there are any TMH candidates among our initial prediction results. We do this by modifying the algorithm proposed in the THUMBUP program¹⁸ to determine TMH candidates and assemble them into physical TMH segments.

Step 1: Filtering

We 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 at least l_{min} ; otherwise, it is converted to a non-helix segment. Steps 2 and 3 describes the assembly of a TMH candidate.

Step 2: Extension

An optimal TMH length, l_{opt} , is set at 21 to reflect the thickness of the hydrocarbon core of a lipid bilayer¹⁹. 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 candidates are merged if they overlap after the extension.

Step 3: Splitting

We define l_{max} , as the cut-off value for the length of a TMH candidate to be split. A TMH candidate whose length is greater than or equal to l_{max} is split into two helices, starting from its N- and C-termini with the loop in the center.

We optimize l_{min} and l_{max} on the training data set (Section 3.1). The optimized values for l_{min} and l_{max} for the best prediction performance are 9 and 38, respectively.

2.4. Topology prediction

2.4.1. Input feature

Using the second classifier, we predict the topology label (*i/o*) of each non-helix residues from the results of the first classifier (*H/~H*). Amino acid composition is employed as the input feature. The encoding scheme follows the same procedure outlined in the helix prediction section.

2.4.2. Alternating geometric scoring function

The purpose of predicting of the topology of a TMH protein is to determine the orientation of the protein with respect to the membrane. A TMH protein follows special constraints on its topology such that it always starts with an inner (*i*) loop or outer (*o*) loop that must *alternate* in order to connect the TM helices. Therefore, the problem of predicting the topology of a TMH protein is reduced to predicting the topology of the first loop located at the N-terminus.

There is growing body of evidence that the final topology is influenced by multiple signals distributed along the entire protein in the loop segments, including the charge bias, loop size, and folding of the N-terminal loop domain²⁰. Furthermore, the widely accepted two-stage model for membrane protein folding suggests that the final topology of a membrane protein is established in the early stages of membrane insertion²¹. These biological phenomena form the basis of our assumptions about topology models. First, we assume that topology formation is a result of contributing signals present in

the various loop segments. Second, signals embedded in the loop segments near the N-terminus are more likely to be a factor in the formation of topology since they are inserted in the membrane at an earlier time. Based on these assumptions, we develop a novel topology scoring function that considers the topogenic contribution from all loop segments that diminishes over a distance away from the N-terminus.

In the proposed topology scoring function, the contribution of signals in the loop segments varies inversely proportional to their distance from the N-terminus in a geometric series: Given a transmembrane protein that has n non-helical segments s_j ($1 \leq j \leq n$ and $n, j \in \mathbb{N}$) 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 topology labels i and o , respectively.

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

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

where $R_i + R_o = 100\%$. To determine the protein topology, we define two topology scores, TS_i and TS_o , where TS_i is for the N-terminal loop on the inside of membrane and TS_o is for the outside.

$$TS_i = \sum_{1 \leq j \leq n} W(j) \times [\alpha R_i(j) + (1 - \alpha) R_o(j)] \quad (3)$$

$$TS_o = \sum_{1 \leq j \leq n} W(j) \times [(1 - \alpha) R_i(j) + \alpha R_o(j)] \quad (4)$$

$$\alpha = \begin{cases} 1, & \text{if } j \text{ is odd} \\ 0, & \text{if } j \text{ is even} \end{cases} \quad (5)$$

$$W(j) = 1/b^{(j-1) \times EI}, \quad b \text{ and } EI \in \mathbb{R}, \quad (6)$$

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 topology of the N-terminal loop is

inside; otherwise, the topology is outside.

For the calculation of topology scores, the geometric scoring function alternates between the inner (i) and outer (o) loops to take into account the alternating nature of the connecting loops. Fig. 3 illustrates the calculation of alternating geometric scoring function for an example protein.

3. RESULTS AND DISCUSSION

3.1. Data sets

1. *Low-resolution TMH proteins*: We train and perform ten-fold cross-validation on a collection of low-resolution data set compiled by Möller *et al.*²². We select 145 proteins of good reliability from a set of 148 non-redundant proteins. We manually validate this data set using annotations from SWISS-PROT release 49.0²³ and further remove two proteins because they have no membrane protein annotations. The final data set contains 143 proteins for which low-resolution topology models are available. This entire data set is also used to train our model for testing on the following three data sets.
2. *High-resolution TMH proteins*: We use a collection of 36 high-resolution TMH proteins from PDB compiled by Chen *et al.*¹³ and obtain topology information for 35 out of 36 proteins. We validate this data set using annotations from SWISS-PROT release 49.0²³ and update the topologies of two proteins.
3. *Soluble proteins*: A collection of 616 high-resolution soluble proteins from PDB compiled by Chen *et al.*¹³ is used to test for discrimination between membrane and soluble proteins.
4. *Newly solved TMH proteins*: Four newly solved high-resolution TMH proteins²⁴ are used as an independent test set.

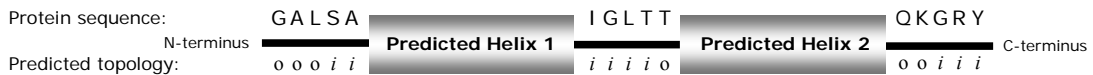


Fig 3. An example of evaluating a TMH protein's topology with alternating geometric scoring function. Helices are predicted in the first step (Section 2.3). A predicted loop segment can have more than one type of topology. We use the proposed alternating geometric scoring function to determine the final topology. In 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 for $(b, EI) = (1.6, 1.0)$ indicated in Section 3.6, $TS_i = 1 \times R_i(1) + 1/(1.6^{1 \cdot 0}) \times R_o(2) + 1/(1.6^{2 \cdot 0}) \times R_i(3) \approx 0.7594$. Similarly, $TS_o \approx 1.2563$. $TS_o > TS_i$, therefore, the final topology for the N-terminal loop is outside (o).

3.2. Evaluation metrics

There are two sets of evaluation measures for the TMH prediction: per-segment and per-residue accuracies¹³. Per-segment scores indicate how accurately the location of a TMH region is predicted and per-residue scores report how well each residue is predicted. Table 1 lists the per-segment and per-residue metrics used in this paper.

In the calculation of per-segment scores, two issues must be addressed when counting a helix as correctly predicted. First, a minimal overlap of observed helix segments must be defined. For this, we use a less relaxed criterion which requires at least 9 overlapping residues. An evaluation study by Chen *et al.*¹³ used a more relaxed minimal overlap of only 3 residues. Second, we do not allow an overlapping observed helix to be counted twice. We use the following examples to illustrate these two issues (H = Helix):

```

Observation:  - - HHHHHHHHHHHHHH - - HHHHHHHHHHHH - -
Prediction 1:  - - - - - HHH - - - - - HHHHHHHHHH - - -
Prediction 2:  - - - - HHHHHHHHHHHHHHHHHHHHHHHH - - -
Prediction 3:  - - HHH - HHHHHHHHHH - - HHHHHHHHHH - - -

```

Prediction 1 achieves 100% accuracy if the minimal overlap is 3 residues. If the minimal overlap is 9 resi-

dues, Prediction 1 achieves 50% accuracy. Prediction 2 achieves 50% accuracy because it already overlaps with the first observed helix. Prediction 3 achieves 100% accuracy if the minimal overlap is 3 residues, but the second predicted helix is an over-prediction since we only count an overlapping observed helix only once. Prediction 3 achieves 50% accuracy if the minimal overlap is 9 residues because the first predicted helix does not satisfy the minimal overlap requirement. In addition, the second predicted helix is also an over-prediction, thus it is not counted.

3.3. Performance of input feature combinations for helix prediction

We test the performance of different input feature combinations for the first classifier. The following combinations are considered: 1) AA only; 2) AA and any one of DP, HS, and AM; 3) AA and any two of DP, HS, and AM; and 4) all four features. We also construct a consensus prediction from the two top-performing combinations through probability estimation using LIBSVM²⁵. The value of the estimated probability for each residue corresponds to the confidence given for its predicted class. In the case of disagreement between the predicted classes, the consensus prediction takes the result of a prediction that has the highest probability.

Table 1. Evaluation metrics used in this work. Per-segment metrics include Q_{ok} , $Q_{htm}^{%obs}$, $Q_{htm}^{%prd}$ and $TOPO$. Per-residue metrics include Q_2 , $Q_{2T}^{%obs}$, and $Q_{2T}^{%prd}$. N_{prot} is the number of proteins in a data set. We follow the same performance measures proposed by Chen *et al.*¹³

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
$Q_{htm}^{%obs}$	$\frac{\text{number of correctly predicted TM in data set}}{\text{number of TM observed in data set}} \times 100\%$	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
$TOPO$	$\frac{\text{number of proteins with correctly predicted topology}}{N_{prot}} \times 100\%$	percentage of correctly predicted topology
Q_2	$\frac{\sum_i^{N_{prot}} \text{number of residues predicted correctly in protein } i}{\sum_i^{N_{prot}} \text{number of residues in protein } i}} \times 100\%$	percentage of correctly predicted TMH residues
$Q_{2T}^{%obs}$	$\frac{\text{number of residues correctly predicted in TM helices}}{\text{number of residues observed in TM helices}} \times 100\%$	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

Table 2 shows the performance of combinations of input features and the consensus prediction. Combination 5 achieves the highest score for Q_{ok} at 71.9% and performs consistently well in other per-segment and per-residue measures. Combination 6 has a strikingly high $Q_{2T}^{%obs}$ score of 85.9%. The purpose of consensus prediction is to maximize the benefits of both combinations. In fact, the consensus approach increases the Q_{ok} score of Combination 6 by 1.5%, while the $Q_{2T}^{%obs}$ score only decreases by 0.3%. Compared to Combination 5, the consensus has a decrease in Q_{ok} of 1.4%, but an increase in $Q_{2T}^{%obs}$ of 3.8%. In addition, the consensus approach also scores the highest for Q_2 at 89.1%. The consensus approach is selected as our best model for comparison with other approaches.

3.4. Performance on high- and low-resolution data sets

SVMtmh is compared to other methods for high and low-resolution data sets in Table 3. For the low-resolution set, *SVMtmh* ranks the highest among all the

compared methods for per-segment measures in *TOPO*, Q_{ok} , and $Q_{htm}^{%pred}$ at 84%, 71%, and 95%, respectively. Specifically, *SVMtmh* improves *TOPO* by 5% over the second best method for the low-resolution data set. For the high-resolution set, most notably, *SVMtmh* has the highest score at 91% for *TOPO*, a 14% improvement over the second best method. Another marked improvement is also observed for the high-resolution set in

Table 2. Performance of input feature combinations and the consensus method. Input features: AA (amino acid composition), DP (di-peptide composition), HS (hydrophobicity scale)¹⁶ and AM (amphiphilicity)¹⁷.

No.	Input Feature (s)	Per-segment (%)			Per-residue (%)		
		Q_{ok}	$Q_{htm}^{%obs}$	$Q_{htm}^{%pred}$	Q_2	$Q_{2T}^{%obs}$	$Q_{2T}^{%pred}$
1	AA	71.2	93.8	93.9	89.1	82.9	83.0
2	AA+DP	69.8	94.0	93.8	88.9	81.9	83.2
3	AA+HS	71.2	92.8	94.2	89.1	81.9	84.0
4	AA+AM	70.5	93.6	93.6	89.1	83.0	82.9
5	AA+DP+HS	71.9	93.6	94.2	89.0	81.8	83.7
6	AA+DP+AM	69.0	93.4	94.0	89.0	85.9	80.6
7	AA+HS+AM	68.3	93.3	94.2	88.8	79.8	84.4
8	AA+DP+HS+AM	69.1	92.3	95.4	89.0	80.9	84.3
9	Consensus (5+6)	70.5	93.2	94.9	89.1	85.6	81.4

Table 3. Performance of prediction methods for low- and high-resolution data sets. Per-segment and per-residue scores of all methods compared are taken from an evaluation by Chen *et al.*¹³. *TOPO* scores for the high-resolution data set are re-evaluated due to the update of topology information. The shaded area outlines the four top-performing methods. Note that we do not have cross-validation results for all other methods. Therefore, their accuracies might be over-estimated. In addition, we use a minimal overlap of 9 residues whereas Chen *et al.*¹³ used only 3 residues. Methods are sorted by their Q_{ok} values for the low-resolution data set.

Methods	Low-resolution							High-resolution						
	Per-segment (%)				Per-residue (%)			Per-segment (%)				Per-residue (%)		
	Q_{ok}	$Q_{htm}^{%obs}$	$Q_{htm}^{%pred}$	<i>TOPO</i>	Q_2	$Q_{2T}^{%obs}$	$Q_{2T}^{%pred}$	Q_{ok}	$Q_{htm}^{%obs}$	$Q_{htm}^{%pred}$	<i>TOPO</i>	Q_2	$Q_{2T}^{%obs}$	$Q_{2T}^{%pred}$
<i>SVMtmh</i>	71	93	95	84	89	86	81	83	96	98	91	86	82	90
TMHMM2	68	91	94	77	89	82	84	75	92	96	66	80	72	88
PHDpsiHtm08	67	95	94	67	89	87	77	84	99	98	57	80	76	83
HMMTOP2	66	94	93	79	90	85	83	83	99	99	77	80	69	89
PRED-TMR	58	92	93		90	78	86	61	84	90		76	58	85
PHDhtm08	57	86	86	68	87	83	75	64	77	76	60	78	76	82
PHDhtm07	56	85	86	72	87	83	75	69	83	81	69	78	76	82
SOSUI	49	88	86		88	79	72	71	88	86		75	66	74
TopPred2	48	84	79	59	88	74	71	75	90	90	57	77	64	83
DAS	39	93	81		86	65	85	79	99	96		72	48	94
Ben-Tal	35	79	90		87	67	83	65	94	89		67	79	66
Wolfenden	29	56	82		80	47	76	64	97	90		71	74	72
WW	27	90	75		81	83	59	60	79	89		72	53	80
GES	23	93	68		78	87	53	58	95	89		69	77	68
Eisenberg	20	90	63		72	89	47	56	93	86		62	80	61
KD	13	88	59		63	91	42	54	95	91		71	71	72
Heijne	11	89	55		51	91	35	52	93	83		60	83	58
Hopp-Wodds	11	87	58		54	90	36	52	94	83		58	83	58
Sweet	11	87	59		58	88	38	48	91	84		59	80	58
Av-Cid	10	87	58		53	89	36	47	95	83		58	80	56
Roseman	9	89	56		48	91	34	45	93	82		61	85	58
Levitt	9	88	56		49	91	35	45	92	82		55	85	55
Nakashima	9	88	56		50	90	35	43	90	83		63	83	60
A-Cid	8	87	57		52	89	35	40	93	79		56	85	55
Lawson	8	86	57		43	89	32	39	88	83		60	84	58
Radzicka	6	87	56		41	91	32	36	92	80		56	84	56
Bull-Breese	6	86	56		40	91	32	33	86	79		55	84	54
EM	5	89	56		41	91	32	31	92	77		57	85	55
Fauchere	5	87	56		43	91	33	28	43	62		62	28	56

which *SVMtmh* obtains the highest score for Q_2 at 86%, compared to the second best methods at 80%. Generally, *SVMtmh* performs 3% to 12% better for the high-resolution set than for the low-resolution in terms of per-segment scores. Meanwhile, for per-residue scores, the accuracy for the high- and low-resolution data sets is similar in the range of 81% to 90%. The shaded area in Table 3 denotes the four top-performing approaches, which are selected to further predict newly solved membrane protein structures (Section 3.7).

3.5. Discrimination between soluble and membrane proteins

To assess our method’s ability to discriminate between soluble and membrane proteins, we apply *SVMtmh* to the soluble protein data set. A cut-off length is chosen as the minimum TMH length. Any protein that does not have at least one predicted TMH exceeding the minimum length is classified as a soluble protein. We calculate the false positives (FP) rates for the soluble protein set, where a false positive represents a soluble protein being falsely classified as a membrane protein. Similarly, we also calculate the false negatives (FN) rates for both high- (FN_{high}) and low-resolution (FN_{low}) membrane protein sets using the chosen cut-off length. Clearly, the cut-off length is a trade-off between the FP and FN rates. Therefore, the cut-off length selected must minimize $FP + FN_{high} + FN_{low}$. Fig. 4 shows the FP and FN rates as a function of cut-off length. The cut-off length at 18, which minimizes the sum of all errors is used to discriminate between soluble and membrane proteins. Table 4 shows the results of our method compared to the other methods. *SVMtmh* is capable of distinguishing soluble and membrane proteins at FP and FN_{low} rates at less than 1% and FN_{high} rate at 5.6%. In general, most advanced methods such as TMHMM2³ and PHDpsiHtm08¹² achieve better accuracies than simple hydrophobicity scale methods including Kyte-Doolittle (KD)⁸ and White -Wimley (WW)¹⁰.

3.6. Effect of alternating geometric scoring function on topology accuracy

We characterize the dependency of topology accuracy (*TOPO*) on the values of the base (b) and the exponent increment (EI) used in the alternating geometric scoring function for the low-resolution data set. Fig. 5 shows

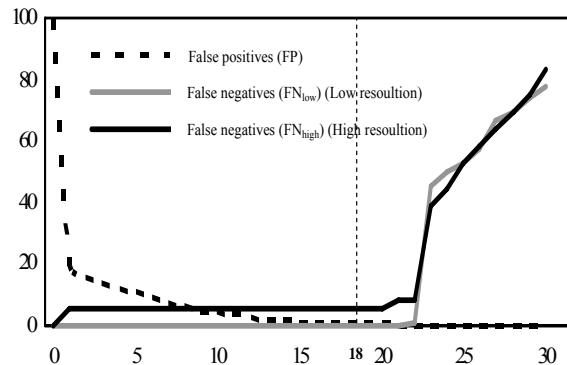


Fig. 4. The false positive and false negative rates as a function of cut-off length. The x-axis: cut-off length; the y-axis: false positive and false negative rates (%). Discrimination between soluble proteins and membrane proteins is based on the cut-off length chosen. The cut-off length at 18 (dashed line) is chosen to minimize the sum of all three error rates ($FP + FN_{low} + FN_{high}$).

Table 4. Confusion between soluble and membrane proteins. The results of all compared methods are taken from Chen *et al.*¹³. False positive rates for soluble proteins are calculated in the second column. In the third and fourth columns, false negative rates for membrane proteins are reported. Methods are sorted by false positive rates.

Methods	False positives (%)	False negatives (%)	
		Low-resolution	High-resolution
<i>SVMtmh</i>	0.5	0	5.6
TMHMM2	1	4	8
SOSUI	1	4	8
PHDpsiHtm08	2	8	3
PHDhtm08	2	23	19
Wolfenden	2	13	39
Ben-Tal	3	4	11
PHDhtm07	3	16	14
PRED-TMR	4	1	8
HMMTOP2	6	1	0
TopPred2	10	11	8
DAS	16	0	0
WW	32	0	0
GES	53	0	0
Eisenberg	66	0	0
KD	81	0	0
Sweet	84	0	0
Hopp-Woods	89	0	0
Nakashima	90	0	0
Heijne	92	0	0
Levitt	93	0	0
Roseman	95	0	0
A-Cid	95	0	0
Av-Cid	95	0	0
Lawson	98	0	0
FM	99	0	0
Fauchere	99	0	0
Bull-Breese	100	0	0
Radzicka	100	0	0

the relationships between topology accuracy coded by colours and the variables in the scoring function. The white circles indicate the highest topology accuracy at about 84% and their corresponding values for b and EI . The region in which half of the white circles (8/16) occur falls in the ranges for b and EI between [1.5, 2.5]

and $[0.5, 1.5]$, respectively. The set of values for (b, EI) we choose for the scoring function is $(1.6, 1.0)$. An interesting observation is that low topology accuracy (80%: blue and 79%: navy) occurs in the vertical-left, lower-horizontal, and upper-right regions. In the vertical-left ($b = 1$) and the lower-horizontal ($EI = 0$) regions, the scoring function is simplified to assigning an equal weight of 1 to all loop signals regardless of their distance from the N-terminus. Conversely, in the upper-right region, when both b and EI are large, the scoring function assigns very small weights to the loop signals downstream of the N-terminus. The poor accuracy in the vertical-left and the lower-horizontal region is a result of considering the contribution of every signal in the loop segments equally. On the other hand, in the upper-right region, the poor performance is due the contribution from downstream signals made negligible by the scoring function. Therefore, our analysis supports the assumptions we have made about our scoring function: 1) topology formation is a result of contributing signals distributed along the protein sequence, particularly in the loop regions; and 2) the contribution of each downstream loop segment on the first loop segment is not equal and diminishes as a function of distance away from the N-terminus. Our results suggest that the inclusion of both assumptions in modeling membrane protein topology is a key factor in achieving the best topology accuracy.

3.7. Performance on newly solved structures and analysis of bacteriorhodopsin

To illustrate the performance of the top four methods on the high and low-resolution data sets as shown in Table 3, we test four recently solved membrane protein structures not included in the training set. The results are shown in Table 5. The best predicted protein is a photosynthetic reaction center protein (PDB ID: 1umx_L), for which all methods predict all helices correctly ($Q_{ok} = 100\%$). On the other hand, only two methods are capable of predicting all the helices from a bacteriorhodopsin (bR) structure (PDB ID: 1tn0_A) correctly ($Q_{ok} = 100\%$). In terms of topology prediction, most methods predict correctly for all four proteins. We devote our analysis to bR to illustrate that TMH prediction is by no means a trivial task and continuous development in this area is indispensable in advancing our understanding of membrane protein structures.

Fig. 6(a) displays the high-resolution structure of bR from PDB. Bacteriorhodopsin (bR) is a member of the rhodopsin family, which is characterized with seven distinct transmembrane helices that can be indexed from Helix A to G. Studies of synthetic peptides of each of the seven TM helices of bR have shown that Helix A to Helix E can form independently stable helices when inserted into a lipid bilayer²⁶. However, Helix G does

Table 5. Performance of top four approaches shaded in Table 3 for newly solved membrane proteins. Proteins are indicated by their PDB codes and their observed topologies. Topology terms N_{in} : N-terminal loop on the inside of membrane; N_{out} : N-terminal loop on the outside of membrane. PRED_TOPO: predicted topology.

Protein (observed topology)	Methods	PRED_TOPO	Per-segment (%)			Per-residue (%)		
			Q_{ok}	$Q_{hm}^{%obs}$	$Q_{hm}^{%prd}$	Q_2	$Q_{2T}^{%obs}$	$Q_{2T}^{%prd}$
1tn0_A (N_{out})	SVMtmh	N_{out}	100	100	100	85	84	94
	TMHMM2	N_{out}	0	86	100	71	68	87
	PHDpsiHtm08	N_{out}	0	71	100	76	77	87
	HMMTOP2	N_{out}	100	100	100	73	69	90
1vfp_A (N_{in})	SVMtmh	N_{in}	0	70	100	87	57	74
	TMHMM2	N_{in}	0	70	100	86	54	72
	PHDpsiHtm08	N_{in}	0	50	50	86	52	72
	HMMTOP2	N_{in}	0	80	89	85	58	63
1umx_L (N_{in})	SVMtmh	N_{in}	100	100	100	90	91	89
	TMHMM2	N_{in}	100	100	100	85	78	89
	PHDpsiHtm08	N_{out}	100	100	100	82	92	75
	HMMTOP2	N_{in}	100	100	100	83	78	83
1xfh_A (N_{in})	SVMtmh	N_{in}	0	70	78	60	62	60
	TMHMM2	N_{in}	0	70	88	63	57	65
	PHDpsiHtm08	N_{in}	0	50	56	53	69	53
	HMMTOP2	N_{in}	0	90	90	71	73	71

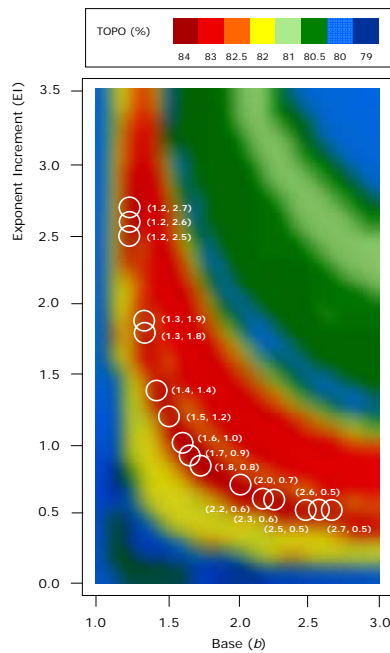


Fig. 5. The relationship between base (b) and exponent increment (EI) in the alternating geometric scoring function and topology accuracy. The x-axis: base (b); the y-axis: exponent increment (EI). The accuracy of topology prediction ($TOPO$) for low-resolution data set is divided into 8 levels, each indicated by a colour. The best accuracy (84%) and its associated (b , EI) values occur within the white circles.

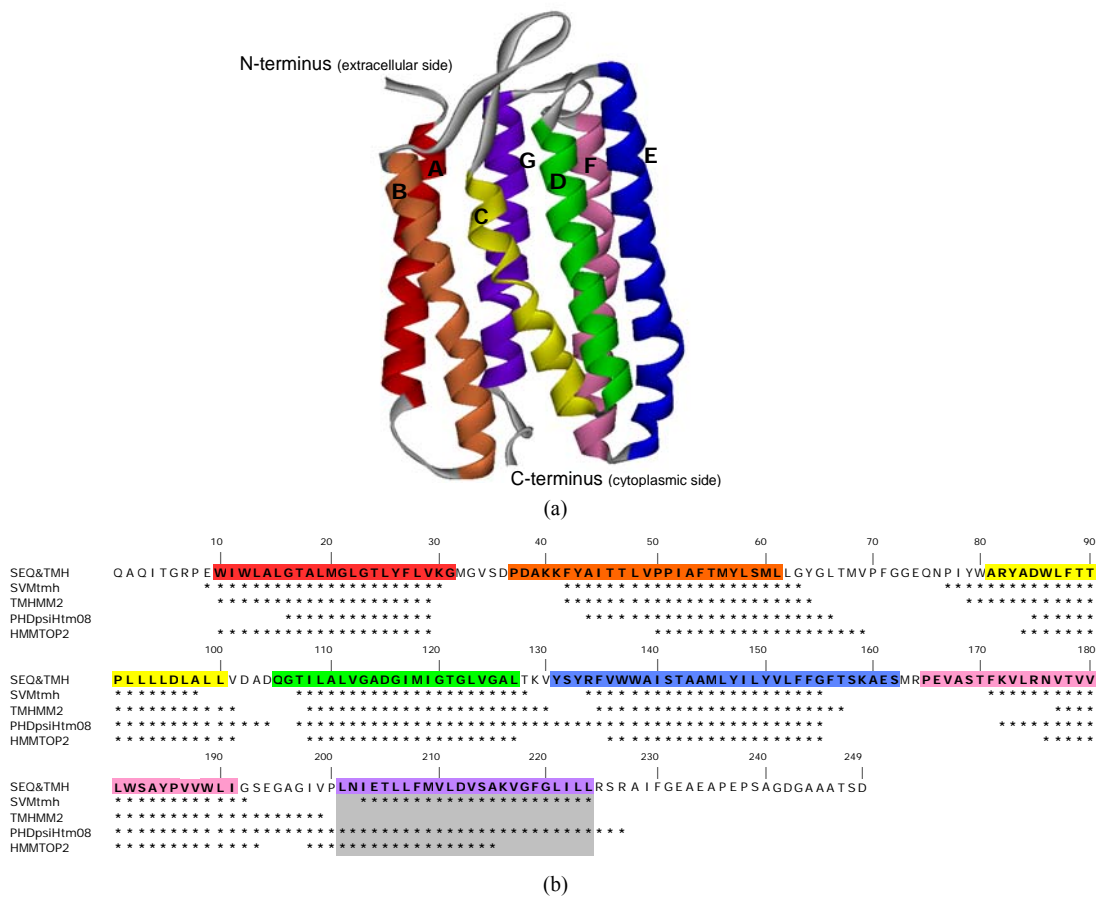


Fig. 6(a). The structure of a bacteriorhodopsin (bR) (PDB ID: 1tn0_A). Each helix is coloured and indexed from A to G. Figure is prepared with ViewerLite²⁹. **Fig. 6(b).** Prediction results of bR by the top four methods (* = predicted helix). The observed helices are indicated by colour boxes. The region of Helix G (purple) and its predictions are highlighted in grey.

not form a stable helix in detergent micelles²⁷ and exhibits structural irregularity at Lys216 by forming a π -bulge²⁸. However, despite its atypical structure, Helix G is important in the function of bR, as it binds to retinal and undergoes conformation change during the photosynthetic cycle²⁸.

The results of the predictions by all four approaches are shown in Fig. 6(b). Interestingly, all approaches are successful in identifying the first six helices (Helix A – E) with good accuracy. However, most methods do not predict with the same level of success for Helix G. In particular, TMHMM2 misses Helix G entirely and PHDpsihtm08 merges predictions for Helix F and Helix G into one long helix. SVMtmh and HMMTOP2¹¹ are the only two out of all four methods that can correctly identify the presence of Helix G. Furthermore, upon a closer examination of Helix G, HMMTOP2 over-predicts by 3 residues at the N-terminus and severely under-predicts by 9 residues at the C-terminus. SVMtmh only under-predicts by 2 residues at the N-terminus of Helix G. The poor prediction results may be due to the intrinsic structural irregularity as described earlier, which adds another level of complexity into the TMH prediction problem. Despite the difficulties involved in predicting the correct location of Helix G, SVMtmh is successful in producing a prediction for the bR structure, which is in close agreement with the experimental approach. One possible reason for our success in this case could be the integration of multiple biological input features that encompass both global and local information for TMH prediction. TMHMM2 and HMMTOP2 rely solely on amino acid composition as sequence information, while PHDpsiHtm08 only uses sequence information from multiple sequence alignments. In contrast, SVMtmh incorporates a combination of both physico-chemical and sequence-based input features for helix prediction.

4. CONCLUSION

We have proposed an approach based on SVM in a hierarchical framework to predict transmembrane helix and topology in two successive steps. We demonstrate that by separating the prediction problem using two classifiers, specific biological input features associated with individual classifiers can be applied more effectively. By integrating both the sequence and structural

input features and using a novel topology scoring function, SVMtmh achieves comparable or better per-segment and topology accuracy for both high- and low-resolution data sets. When tested for confusion between membrane and soluble proteins, SVMtmh discriminates between them with the lowest false positive rate compared to the other methods. We further analyze a set of newly solved structures and show that SVMtmh is capable of predicting the correct helix and topology of bacteriorhodopsin as derived from a high resolution experiment.

With regard to future work, we will continue to enhance the performance of our approach by incorporating more relevant features in both stages of helix and topology prediction. We will also consider some complexities of TM helices, including helix lengths, tilts, and structural motifs, as in the case of bacteriorhodopsin. Supported by the results we achieved, our approach could prove valuable for genome-wide predictions to identify potential integral membrane proteins and their topologies.

While obtaining high-resolution structures for membrane proteins presents itself as a major challenge in the field of structural biology, the need for accurate prediction methods is highly demanded. We believe that the continuous development of computational methods with the integration of biological knowledge in this area will be immensely fruitful.

Acknowledgments

We gratefully thank Jia-Ming Chang, Hsin-Nan Lin, Wei-Neng Hung, and Wen-Chi Chou for providing helpful discussions and computational assistance. This work was supported in part by the thematic program of Academia Sinica under grant AS94B003 and AS95ASIA02.

References

1. Wallin E and von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* 1998; **7**: 1029-1038.
2. Stevens TJ and Arkin IT. The effect of nucleotide bias upon the composition and prediction of transmembrane helices. *Protein Sci* 2000; **9**: 505-511.
3. Krogh A, Larsson B, von Heijne G, and Sonnhammer EL. Predicting transmembrane protein topol-

- ogy with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; **305**: 567-580.
4. Ubarretxena-Belandia I and Engelman DE. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr Op in Struc Bio* 2001; **11**: 370-376.
 5. White SH. The progress of membrane protein structure determination. *Protein Sci* 2004; **13**: 1948-1949.
 6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000; **28**: 235-242.
 7. van Geest M and Lolkema JS. Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiol Mol Biol Rev* 2000; **64**: 13-33.
 8. Kyte J and Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982; **157**: 105-132.
 9. Eisenberg D, Weiss RM, and Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A* 1984; **81**: 140-144.
 10. White SH and Wimley WC. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* 1999; **28**: 319-365.
 11. Tusnady GE and Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 1998; **283**: 489-506.
 12. Rost B, Fariselli P, and Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* 1996; **5**: 1704-1718.
 13. Chen CP, Kernytsky A, and Rost B. Transmembrane helix predictions revisited. *Protein Sci* 2002; **11**: 2774-2791.
 14. Chang CC and Lin CJ. LIBSVM: A library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
 15. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 1992; **225**: 487-494.
 16. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, and von Heijne G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 2005; **433**: 377-381.
 17. Mitaku S, Hirokawa T, and 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**: 608-616.
 18. Zhou H and Zhou Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci* 2003; **12**: 1547-1555.
 19. Jayasinghe S, Hristova K, and White SH. Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* 2001; **312**: 927-934.
 20. Goder V and Spiess M. Topogenesis of membrane proteins: determinants and dynamics. *FEBS Letters* 2001; **504**: 87-93.
 21. Popot JL and Engelman DM. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* 1990; **29**: 4031-4037.
 22. Moller S, Kriventseva EV, and Apweiler R. A collection of well characterised integral membrane proteins. *Bioinformatics* 2000; **16**: 1159-1160.
 23. Bairoch B, Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* 1997; **5**: 312-316.
 24. Cao B, Porollo A, Adamczak R, Jarrell M, and Meller J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. *Bioinformatics* 2006; **22**: 303-309.
 25. Wu TF, Lin CJ, and Weng RC. Probability estimates for multi-class classification by pairwise coupling. *JMLR* 2004; **5**: 975-1005.
 26. Booth PJ. Unravelling the folding of bacteriorhodopsin. *Biochim Biophys Acta* 2000; **1460**: 4-14.
 27. Hunt JF, Earnest TN, Bousche O, Kalghatgi K, Reilly K, Horvath C, Rothschild KJ, and Engelman DM. A biophysical study of integral membrane protein folding. *Biochemistry* 1997; **36**: 15156-15176.
 28. Luecke H, Schobert B, Richter HT, Cartailler JP, and Lanyi JK. Structure of bacteriorhodopsin at 1.55 Å resolution. *J Mol Biol* 1999; **291**: 899-911.
 29. ViewerLite for molecular visualization. Software available at <http://www.jaici.or.jp/sci/viewer.htm>.