中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-07-010

ON THE ACCURACY OF TRANSMEMBRANE HELIX PREDICTION METHODS USING AN UPDATED BENCHMARK

Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, Wen-Lian Hsu

Sep. 14, 2007 || Technical Report No. TR-IIS-07-010

http://www.iis.sinica.edu.tw/LIB/TechReport/tr2007/tr07.html

ON THE ACCURACY OF TRANSMEMBRANE HELIX PREDICTION METHODS USING AN UPDATED BENCHMARK

Allan Lo[1,2], Hua-Sheng Chiu[3], Ting-Yi Sung[3], 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

Abstract

The prediction of transmembrane (TM) helix and topology is an important field of bioinformatics owing to the difficulties in obtaining high-resolution structures of membrane proteins. Many methods have been developed and several evaluations have compared the performance of individual methods using benchmarks from various sources. We present an analysis of a popular evaluation method by Kernytsky and Rost, which is created using data sets from more than six years ago. Our analysis shows that the benchmark contains data that have substantial disagreements in comparison with the current annotations in SwissProt Release 54.1. Furthermore, the benchmark also contains issues such as annotations of low reliability, sequence redundancy, and presence of signal peptides. We perform updating and cleansing of the above issues in the benchmark, and evaluate eleven widely used methods, including SVM*top*, a hierarchical classification method based on support vector machines (SVM). The results show that SVM*top* is ranked highly among the top-performing methods for helix prediction, correctly predicting the location of helices for more than 80% of the updated benchmark. Given the discrepancies and noises in the original benchmark, it should be used with discretion for assessing the performance of TM helix predictions. The analysis also implies that there is an urgent need for creating a new benchmark for an accurate and objective comparison. The updated benchmark is available for public use at http://bio-cluster.iis.sinica.edu.tw/~bioapp/SVMtop/dataset.htm.

Keywords

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

I. 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 genomes [1-3]. 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 membranes [4]. Despite their biological importance, the proportion of available high-resolution structures is exceedingly limited at about 0.5% of all solved structures [5], 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-consuming [7]. Therefore, bioinformatics development in sequence-based prediction methods is valuable for elucidating the structural genomics of membrane proteins.

Many different methods have been developed to predict structural models of TM proteins. Earlier approaches relied on physico-chemical properties such as hydrophobicity [8-10] to identify TMH regions. Recently, more advanced methods using hidden Markov models [3, 11], neural networks [12] and support vector machines (SVM) [13] have been developed, and they have achieved significant improvements in prediction accuracy.

* Corresponding author.
§ This work was supported in part by the thematic program of Academia Sinica under grant AS94B003 and AS95ASIA02.

Concurrently, several evaluation studies have examined the performance of different methods [14-18]. However, no method is consistently ranked as the top-performing method because 1) the data sets used for evaluation are quite different; and 2) no consistent criteria for measuring performance are used. One evaluation, in particular, assesses the TM helix prediction accuracy and is available as an online server [17]. For a total of 2247 proteins, this benchmark consists of four groups of data sets including high- and low-resolution TM proteins, signal peptides, and soluble proteins. It offers a static evaluation by allowing users to upload their predictions onto the server, and the results will be benchmarked against other methods. Essentially, this evaluation is the first of its kind, and it is valuable for benchmarking purposes when developing a new method given its timeliness and usability. However, since the TM protein data sets are compiled from more than six years ago, it is very likely that they have undergone substantial modifications [19]. Therefore, we design this study to evaluate the suitability of the benchmark.

First, we use the benchmark to evaluate a method of our previous work, SVM*top*, and the initial results appear to be contradictory to the results of our own assessment [13, 20]. The poor results achieved by SVM*top*, particularly in the TM helix prediction, prompted us to carry out a detailed analysis. As the first step, we question the validity and correctness of the data sets by checking if the annotations are updated with the current annotations in SwissProt Release 54.1 [21]. If the data sets are indeed updated, they should have very high agreement with the current annotations, and the annotations evaluated by the benchmark should in turn, achieve very high performance. Since the original benchmark is used, the identity of each protein (high- or low- TM protein, signal peptide, or soluble protein) is not given. We identify each of the 2247 proteins in the benchmark by matching it with a data set from Chen *et al*. [16] and using the BLASTP program [22]. Next, the annotation of each TM protein is extracted from SwissProt Release 54.1. The TM proteins of current annotation are then uploaded to the server to test for consistency. In the second step, we also re-assess SVM*top*, among other existing methods for comparison using the updated annotations.

The results indicate that 1) the original benchmark does contain a substantial amount of inconsistencies with the current annotations (>20%), in addition to sequence redundancy and signal peptides; and 2) the performance of SVM*top* is under-estimated using the original benchmark due to the issues outlined above. The prediction accuracy for helix location of SVM*top* is well over 80% when evaluated using the updated data sets for both high- and low-resolution TM proteins, and it compares favorably with ten other methods. This agrees well with our previous work of evaluations from two benchmarks consisting of high- and low-resolution TM proteins. An important implication is that evaluations containing old annotations should be used with caution and the results should be carefully interpreted, if not avoided entirely. This also calls for a new benchmark of updated annotations for a fair assessment of TM prediction methods.

II. METHODS

A. SVMtop methodology

SVM*top* represents a support vector machine method for transmembrane helix and *top*ology prediction, using two-stage hierarchical classification. 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). Figure 1 shows the system architecture of the method. Detailed feature selection and encoding as well as the calculation of AGSF are described in Lo *et al.* [13, 20].

B. Evaluating SVMtop using the original benchmark

We evaluated SVM*top* using the original benchmark as published online by Kernytsky and Rost [17] by uploading the prediction of SVM*top* onto the server at http://cubic.bioc.columbia.edu/tmh_benchmark. The benchmark contains four different types of proteins: high-resolution TM proteins, low-resolution TM proteins, signal peptides, and soluble proteins. Therefore, this benchmark also assesses the power to discriminate signal peptides and soluble proteins in addition to helix predictions.The prediction is accepted in two formats: 1) helix start and end numbers after the protein number (e.g. >0 11, 35 denotes that the helix starts at 11^{th} position and ends at the 35^{th} amino acid); and 2) per-residue prediction (e.g. >1 LLHHHHLL), where L stands for loop residues and H represents helix residues after the protein number. We uploaded the prediction in Format 1 onto the server and the results are returned with a

comparison with other methods as shown in Section III-A.

C. *Identifying the proteins in the benchmark and retrieving the current annotations*

Each of the proteins in the original benchmark must first be identified before we can retrieve their annotations in the SwissProt. Therefore, we performed extensive work to decode the identity of all 2247 proteins by first matching it with another data set used in Chen *et al.* [16] which is available at http://cubic.bioc.columbia.edu/papers/2002_htm_eval/data/. We will refer to this data set as Chen-2002. It contains sequences and protein IDs for 36 high-resolution TM proteins, 165 low-resolution TM proteins, 616 soluble proteins, and 1418 signal peptides. We tried to first identify each of the 2247 proteins in the benchmark by one-against-all matching with the sequences given in Chen-2002. If a protein sequence matched 100% with the sequence as given in Chen-2002, it was assigned as one of the high/low/soluble/signal proteins as labeled in Chen-2002. If it was not matched, we used blastp searching against SwissProt v.54.1 to identify them. Each of the 2247 proteins was identified according to its category and Protein Name or PDB ID. Examples identified by matching with Chen-2002, are shown in the order of protein number in benchmark, group, database|protein name or PDB ID:

>1 HIGH_RES,PDB|1afo_A; >36 LOW_RES,SWISS|1b27_human;
>201 SIGNAL,11S3_HELAN; >1620 SOLUBLE,1a1d;

An example of a protein identified by BLASTP, is shown in the order of protein number in benchmark, blastp-identified, group of protein, database|protein name or PDB ID:

>25 BLASTP,SOLUBLE,SWISS|MEL_APICE;

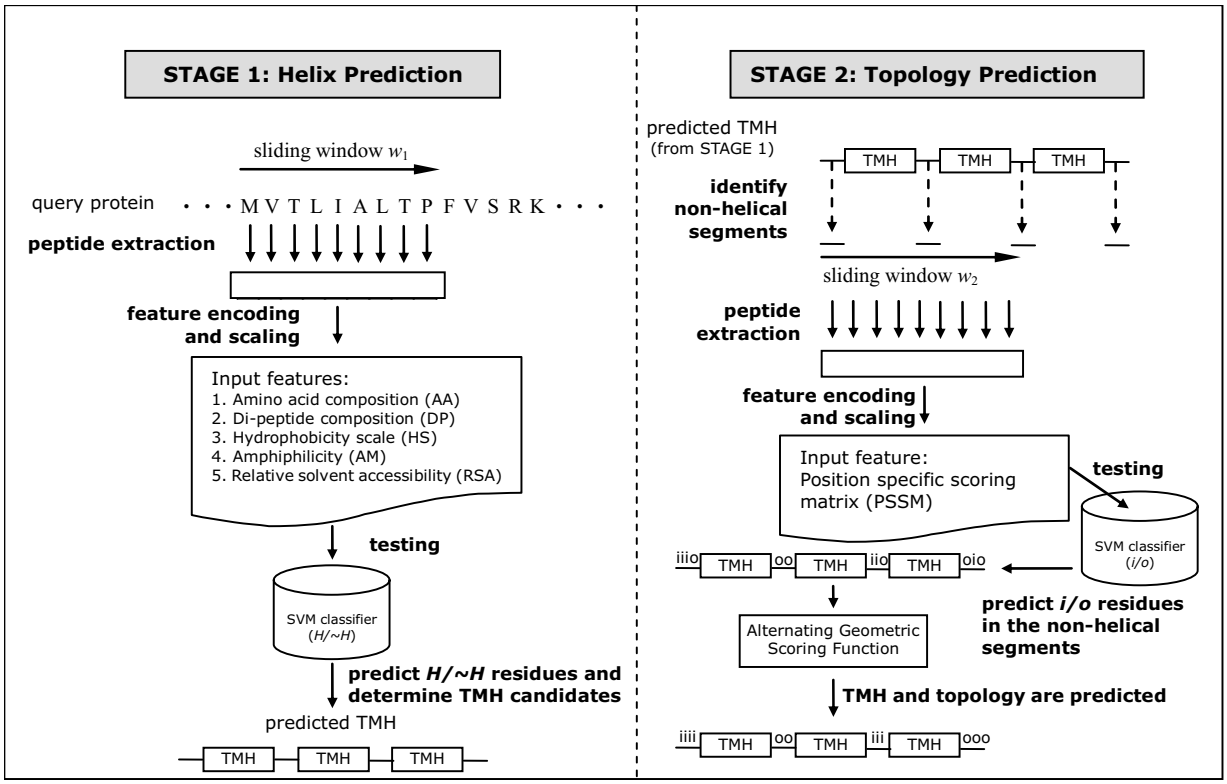

Fig. 1. Flowchart of SVM*top*. The left panel describes Helix Prediction of Stage 1 in the order of: peptide extraction by sliding windows; feature encoding and scaling; prediction of helix (*H*) and non-helix (*~H*) residues; determination of TMH candidates. The right panel describes Topology Prediction 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 shows the flowchart of the abovementioned procedure. For each of the low- and high-resolution TM proteins, we obtained their PDB IDs or SwissProt protein names. We then downloaded the corresponding annotations from SwissProt Release 54.1 and obtained the topology information from the FT lines of the raw text format. Some high-resolution TM proteins do not have complete SwissProt topology annotations, thus we took the annotations from MPtopo database (Last update, Aug. 30, 2007) [23]. The high-resolution TM proteins with MPtopo annotations are 1bgy_C, 1bgy_D, 1bgy_G, 1bgy_J, 1bl8_A, and 1ehk_A. The TM proteins with updated annotations were then uploaded as predictions onto the server, to test for any inconsistency with the server's annotation. The results are shown in Section III-C.

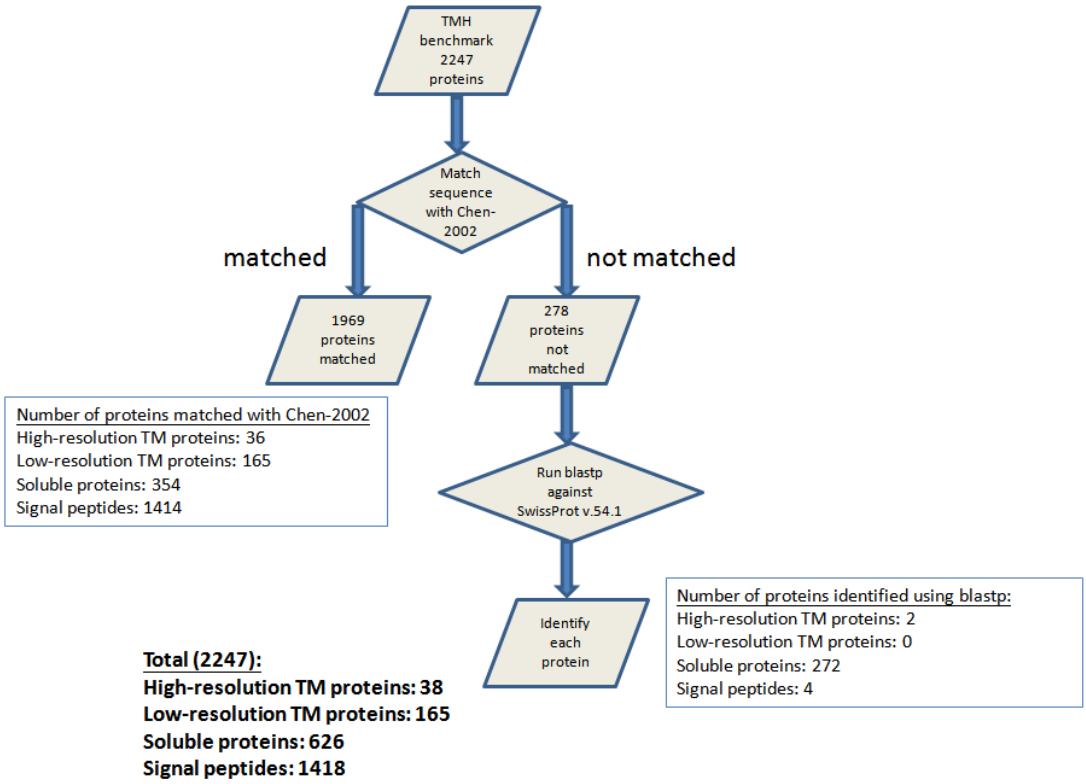

Fig. 2. The procedure taken to identify the proteins in the benchmark.

D. Performance evaluation using the updated benchmark

After we updated the benchmark using the current annotations from SwissProt Release 54.1, we re-evaluated SVM*top* and also ten other popular TM helix prediction methods, including TMHMM2 [3], HMMTOP2 [11], PHDhtm v1.96 [12], MEMSAT3 [24], TopPred 2 [25], SOSUI 1.1 [26], SPLIT4 [27], ConPred II [28], Phobius [29], and PolyPhobius [30]. We used the default settings for each online server compared, wherever applicable.

E. Evaluation metrics

To assess the prediction accuracy, we followed the evaluation measures as described by Chen *et al*. [16]. There are three types of measures: per-protein, per-segment, and per-residue accuracy as listed in Table I. For per-residue measures, we also used Matthew's correlation coefficient (*MCC*), which is a more robust measure than using recall or precision alone [31]. 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 [3,11,12]. A correctly predicted TMH segment is defined as a one-to-one overlap with the true TMH segment.

TABLE I
EVALUATION METRICS USED IN THIS WORK

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
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

We only evaluate the per-protein measure, Q_{ok} for helix prediction because no topology information can be retrieved for some proteins. 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.

III. Results and Discussion

A. Performance of SVMtop evaluated by the original benchmark

The performance of SVM*top* assessed by the original benchmark for high- and low-resolution TM proteins is listed in Table II and Table III, respectively. From the first glimpse of the results, although SVM*top* (indicated as "YOU") achieves relatively good per-residue accuracy for both data sets, it does not perform favorably against other methods such as HMMTOP, TMHMM, and PHDhtm in per-protein score. Particularly, the Q_{ok} (percentage of proteins with all helices predicted correctly) score is very low for low-resolution set, at 57%. This is significantly lower than that of TMHMM, at 75%. These results are very different from our previous evaluations using more expanded data sets [13, 20]. Another peculiar observation is that, SVM*top* has the best Q_2, good per-segment recall $Q_{htm}^{\%obs}$ and precision $Q_{htm}^{\%prd}$ (>90%), but the Q_{ok} score is significantly lower (57%) than that of Q_2 (92%) in the low-resolution set. Therefore, we decided to examine the benchmark more closely. Since SVM*top* was trained on an updated version of Möller's data set [19] using SwissProt, we already found a lot of modifications (unpublished results). We suspected

that the benchmark which also used Möller's data set would likely contain some old information. Hence, we identified each of the 2247 proteins, and downloaded their annotations from SwissProt.

In Table IV, SVM*top*'s performance on soluble proteins is listed. It is clear that SVM*top* achieves the best discrimination among all methods compared by achieving 0% false positive and negative rates. In addition, SVM*top* is also assessed in terms of signal peptide discrimination in Table V. Since SVM*top* was not trained to discriminate signal peptides, it is expected that many of the signal peptides could be wrongly predicted as TM helices (93%). From this preliminary evaluation, we decided to examine the TM proteins more closely, by addressing the poor performance in Q_{ok}.

TABLE II
HIGH-RESOLUTION ACCURACY OF SVM*top* BEFORE UPDATING THE BENCHMARK

Method ?	Qok ?	Qhtm %obs ?	Qhtm %prd ?	Q2 ?	Q2T %obs ?	Q2T %prd ?
PHDpsihtm08	84	99	98	80	76	83
HMMTOP2	83	99	99	80	69	89
DAS	79	99	96	72	48	94
YOU	75	96	93	86	81	90
TopPred2	75	90	90	77	64	83
TMHMM1	71	90	90	80	68	81
SOSUI	71	88	86	75	66	74
PHDhtm07	69	83	81	78	76	82
KD	65	94	89	67	79	66
PHDhtm08	64	77	76	78	76	82
GES	64	97	90	71	74	72
PRED-TMR	61	84	90	76	58	85
Ben-Tal	60	79	89	72	53	80
Eisenberg	58	95	89	69	77	68
Hopp-Woods	56	93	86	62	80	61
WW	54	95	91	71	71	72
Roseman	52	94	83	58	83	58
Av-Cid	52	93	83	60	83	58
Levitt	48	91	84	59	80	58
A-Cid	47	95	83	58	80	56
Heijne	45	93	82	61	85	58
Bull-Breese	45	92	82	55	85	55
Sweet	43	90	83	63	83	60
Radzicka	40	93	79	56	85	55
Nakashima	39	88	83	60	84	58
Fauchere	36	92	80	56	84	56
Lawson	33	86	79	55	84	54
EM	31	92	77	57	85	55
Wolfenden	28	43	62	62	28	56

The raw output of the benchmark server is shown. The question mark (?) denotes a hyperlink for explanation in the original output. SVM*top* is indicated as "YOU". (Sorting on column 1 - highlighted)

TABLE III
LOW-RESOLUTION ACCURACY OF SVM*top* BEFORE UPDATING THE BENCHMARK

Method ?	Qok ?	Qhtm %obs ?	Qhtm %prd ?	Q2 ?	Q2T %obs ?	Q2T %prd ?
TMHMM1	72	91	92	90	83	80
PHDpsiHtm08	67	95	94	89	87	77
HMMTOP2	66	94	93	90	85	83
PRED-TMR	58	92	93	90	78	86
YOU	57	95	85	92	89	82
PHDhtm08	57	86	86	87	83	75
PHDhtm07	56	85	86	87	83	75
SOSUI	49	88	86	88	79	72
TopPred2	48	84	79	88	74	71
DAS	39	93	81	86	65	85
Ben-Tal	35	79	90	87	67	83
Wolfenden	29	56	82	80	47	76
WW	27	90	75	81	83	59
GES	23	93	68	78	87	53
Eisenberg	20	90	63	72	89	47
KD	13	88	59	63	91	42
Sweet	11	87	59	58	88	38
Hopp-Woods	11	87	58	54	90	36
Heijne	11	89	55	51	91	35
Av-Cid	10	87	58	53	89	36
Roseman	9	89	56	48	91	34
Nakashima	9	88	56	50	90	35
Levitt	9	88	56	49	91	35
Lawson	8	86	57	43	89	32
A-Cid	8	87	57	52	89	35
Radzicka	6	87	56	41	91	32
Bull-Breese	6	86	56	40	91	32
Fauchere	5	87	56	43	91	33
EM	5	89	56	41	91	32

The raw output of the benchmark server is shown. The question mark (?) denotes a hyperlink for explanation in the original output. SVM*top* is indicated as “YOU”. (Sorting on column 1 - highlighted)

TABLE IV
CONFUSION WITH SOLUBLE PROTEINS BEFORE UPDATING THE BENCHMARK

Method	False Positives	High-resolution	Low-resolution
YOU	0	0	0
SOSUI	1	8	4
TMHMM1	1	8	4
PHDhtm08	2	19	23
PHDpsihtm	2	3	8
Wolfenden	2	39	13
Ben-Tal	3	11	4
PHDhtm07	3	14	16
PRED-TMR	4	8	1
HMMTOP2	6	0	1
TopPred2	10	8	11
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
A-Cid	95	0	0
Av-Cid	95	0	0
Roseman	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 raw output of the benchmark server is shown.
SVM*top* is indicated as “YOU”. (Sorting on column 1 - highlighted)

TABLE V
CONFUSION WITH SIGNAL PEPTIDES BEFORE UPDATING THE BENCHMARK

Method	**% of proteins with signal proteins**
PHDpsihtm08	23
PHDhtm08	24
Wolfenden	26
TMHMM1	34
PRED-TMR	41
PHDhtm07	45
HMMTOP2	48
Ben-Tal	57
SOSUI	61
TopPred2	82
WW	90
YOU	93
DAS	97
GES	98
A-Cid	99
Av-Cid	99
Bull-Breese	99
EM	99
Eisenberg	99
Fauchere	99
Heijne	99
Hopp-Woods	99
KD	99
Lawson	99
Levitt	99
Nakashima	99
Radzicka	99
Roseman	99
Sweet	99

The raw output of the benchmark server is shown.
SVM*top* is indicated as “YOU”. (Sorting on column 1 - highlighted)

B. *Issues of the original benchmark*

In the process of updating the topological information using the annotations from the latest version of SwissProt Release 54.1, we discovered that there are many issues with the original benchmark:

1. It contains protein annotations of low reliability: This benchmark contains 49 proteins that are taken from Trust Level D of Möller's data set. According to the original publication by Möller [19], "annotations assigned the trust level D should not be used for training or testing purposes." These proteins annotated with low reliability level may lead to a biased evaluation. The 49 proteins are (identified by the protein number as given in the benchmark): 37, 40, 42, 48, 49, 54, 55, 56, 58, 59, 71, 81, 85, 86, 87, 89, 94, 99, 100, 101, 103, 104, 105, 110, 111, 112, 113, 117, 118, 119, 120, 122, 132, 137, 139, 146, 149, 151, 152, 156, 163, 164, 165, 171, 173, 178, 180, 184, and 191.
2. Signal peptides are not removed from some low-resolution TM proteins: When we closely examined the TM proteins individually, we found 28 proteins that have sequences containing signal peptides as annotated by SwissProt Release 54.1. They should have been removed for testing purposes to avoid biases. The proteins that contain signal peptides are (identified by the protein number as given in the benchmark): 36, 41, 43, 57, 69, 74, 86, 87, 96, 98, 100, 101, 102, 103, 117, 118, 119, 122, 139, 145, 146, 149, 152, 154, 160, 173, 180, and 191. We kept the signal peptides (SPs) to evaluate the agreement between the benchmark and the annotations from SwissProt to test for consistency between the two. However, we removed the SPs for a fair evaluation of all methods in the updated benchmark.
3. It contains redundant proteins: We found two pairs of proteins that have 100% sequence identity (high-resolution TM proteins: 21 and 2243; soluble proteins: 2078 and 2079). This contradicts with the paper by Kernytsky and Rost, in which the authors claimed that all redundant proteins were filtered out [17].
4. Others: Incorrect protein IDs; Two protein IDs were incorrect when we matched the sequences using Chen-2002. According to Chen-2002, protein 166 is PTMA_ECOLI, but the correct ID should be PTM3C_ECOLI. Similarly, protein 36 should be 1B27_HUMAN, not 1B14_HUMAN.

C. *Inconsistency with SwissProt Release 54.1*

We identified 38 high-resolution and 165 low-resolution TM proteins from the original benchmark and retrieved their respective annotations from SwissProt. To evaluate the consistency between the benchmark's annotations and the SwissProt's, we uploaded the annotations from SwissProt as our own prediction onto the server (http://cubic.bioc.columbia.edu/tmh_benchmark.). If there is a high agreement between the two, the accuracy should be very high. If not, it is very likely that the annotations in the benchmark is not consistent with the SwissProt, and could be out of date. We kept all the redundant proteins and signal peptides in this evaluation; signal peptides were treated as 'loops' when we uploaded the SwissProt annotations. The results for high- and low-resolution TM proteins are listed in Table VI and Table VII, respectively.

From the results of Table VI and Table VII, it is clear that there is some significant disagreement between the annotations in SwissProt Release 54.1 and that of the benchmark. The Q_{ok} of both data sets are around 77-78%, and this means that there is more than 20% in each data set that has some inconsistencies. This means that more than 20% of the data sets contain outdated information. Furthermore, given the many issues discovered while updating the data set described above, it is of no surprise that the original benchmark could be biased. Therefore, we not only found the original benchmark full of unreliable information and noises, but also outdated annotations.

D. *Comparison of SVMtop with other methods using the updated benchmark*

We updated the benchmark using the current annotations from SwissProt, and removed the issues such as redundancy and signal peptides. This resulted in a total of 37 high-resolution and 165 low-resolution proteins. We re-evaluated ten top-performing approaches using the updated benchmark. The results are shown in Table VIII and Table IX.

From the results of the evaluation using the updated benchmark, SVM*top* performs competitively against other methods in terms of all of the per-protein, per-segment, and per-residue scores. A dramatic improvement is seen in Q_{ok} by evaluating using the updated benchmark; low-resolution data set before updating (57%) and after (83%). Since the number of samples in high-resolution is small (37), it is probably not convincing by drawing conclusion based on it. However, we also observed some improvement in Q_{ok} as well, from 75% (before updating) to 86% (after updating). The improvement is likely a result of removing the noises and updating the correct topological information in the

benchmark. Since SVM*top* was trained on a more recent version of SwissProt, its performance is less likely to suffer from data inconsistency when evaluated by benchmarks of updated annotations. This could also mean that for those methods that were trained using Möller's data set but did not renew the topological annotations, the assessment may not truly reflect the predictive power of individual methods. Another highly ranked method, MEMSAT3, obtains 89% in Q_{ok} for high-resolution data set, and 78% for low-resolution set. Phobius and its homologous version, PolyPhobius, also rank as the top methods in both data sets. These results are consistent with our evaluations [13, 20].

TABLE VI
HIGH RESOLUTION ACCURACY OF SWISSPROT RELEASE 54.1

Method ?	Qok ?	Qhtm %obs ?	Qhtm %prd ?	Q2 ?	Q2T %obs ?	Q2T %prd ?
PHDpsihtm08	84	99	98	80	76	83
HMMTOP2	83	99	99	80	69	89
DAS	79	99	96	72	48	94
YOU	77	94	91	89	85	88
TopPred2	75	90	90	77	64	83
TMHMM1	71	90	90	80	68	81
SOSUI	71	88	86	75	66	74
PHDhtm07	69	83	81	78	76	82
KD	65	94	89	67	79	66
PHDhtm08	64	77	76	78	76	82
GES	64	97	90	71	74	72
PRED-TMR	61	84	90	76	58	85
Ben-Tal	60	79	89	72	53	80
Eisenberg	58	95	89	69	77	68
Hopp-Woods	56	93	86	62	80	61
WW	54	95	91	71	71	72
Roseman	52	94	83	58	83	58
Av-Cid	52	93	83	60	83	58
Levitt	48	91	84	59	80	58
A-Cid	47	95	83	58	80	56
Heijne	45	93	82	61	85	58
Bull-Breese	45	92	82	55	85	55
Sweet	43	90	83	63	83	60
Radzicka	40	93	79	56	85	55
Nakashima	39	88	83	60	84	58
Fauchere	36	92	80	56	84	56
Lawson	33	86	79	55	84	54
EM	31	92	77	57	85	55
Wolfenden	28	43	62	62	28	56

The raw output of the benchmark server is shown. The question mark (?) denotes a hyperlink for explanation in the original output. SwissProt Release 54.1 is indicated as "YOU". (Sorting on column 1 - highlighted)

TABLE VII
LOW RESOLUTION ACCURACY OF SWISSPROT RELEASE 54.1

Method ?	Qok ?	Qhtm %obs ?	Qhtm %prd ?	Q2 ?	Q2T %obs ?	Q2T %prd ?
YOU	78	96	94	97	95	91
TMHMM1	72	91	92	90	83	80
PHDpsiHtm08	67	95	94	89	87	77
HMMTOP2	66	94	93	90	85	83
PRED-TMR	58	92	93	90	78	86
PHDhtm08	57	86	86	87	83	75
PHDhtm07	56	85	86	87	83	75
SOSUI	49	88	86	88	79	72
TopPred2	48	84	79	88	74	71
DAS	39	93	81	86	65	85
Ben-Tal	35	79	90	87	67	83
Wolfenden	29	56	82	80	47	76
WW	27	90	75	81	83	59
GES	23	93	68	78	87	53
Eisenberg	20	90	63	72	89	47
KD	13	88	59	63	91	42
Sweet	11	87	59	58	88	38
Hopp-Woods	11	87	58	54	90	36
Heijne	11	89	55	51	91	35
Av-Cid	10	87	58	53	89	36
Roseman	9	89	56	48	91	34
Nakashima	9	88	56	50	90	35
Levitt	9	88	56	49	91	35
Lawson	8	86	57	43	89	32
A-Cid	8	87	57	52	89	35
Radzicka	6	87	56	41	91	32
Bull-Breese	6	86	56	40	91	32
Fauchere	5	87	56	43	91	33
EM	5	89	56	41	91	32

The raw output of the benchmark server is shown. The question mark (?) denotes a hyperlink for explanation in the original output. SwissProt Release 54.1 is indicated as “YOU”. (Sorting on column 1 - highlighted)

TABLE VIII
ACCURACY OF HIGH-RESOLUTION TM PROTEINS BASED ON UPDATED BENCHMARK

Method	Per-protein (%)	Per-segment (%)		Per-residue (%)			
	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	*MCC*
SVM*top*	**86.4865 (32/37)**	**94.6565**	**99.2000**	**93.4627**	**88.3428**	**94.0309**	**0.8399**
TMHMM2	78.3784 (29/37)	90.8397	98.3471	84.8769	76.1185	89.5146	0.6984
HMMTOP2	78.3784 (29/37)	92.3664	96.0317	83.6246	73.8500	87.0082	0.6569
PHDhtm v1.96	59.4565 (22/37)	85.4962	94.9153	79.8252	63.5161	90.0402	0.6109
MEMSAT3	89.1892 (33/37)	95.4198	98.4252	86.5978	81.3170	88.8468	0.7317
TopPred2	75.6757 (28/37)	89.3130	97.5000	84.0827	71.0775	89.7018	0.6626
SOSUI 1.1	75.6757 (28/37)	91.6031	97.5610	82.3361	76.2445	86.0597	0.6650
SPLIT4	83.7838 (31/37)	93.1298	93.3871	84.4134	80.8759	86.1409	0.7007
ConPred II	81.0811 (30/37)	93.1298	96.8254	86.2441	74.9842	90.3569	0.6979
Phobius	78.3784 (29/37)	92.3664	98.3740	84.4991	77.6938	89.1218	0.7065
PolyPhobius	83.7838 (31/37)	94.6565	97.6378	86.1955	80.2773	90.5151	0.7402

The total number of high-resolution proteins is 37; one redundant protein (protein number: 2243; PDB ID: 1prc_h) is removed.

TABLE IX
ACCURACY OF LOW-RESOLUTION TM PROTEINS BASED ON UPDATED BENCHMARK

Method	Per-protein (%)	Per-segment (%)		Per-residue (%)			
	Q_{ok}	$Q_{htm}^{\%obs}$	$Q_{htm}^{\%prd}$	Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	*MCC*
SVM*top*	**83.0303 (137/165)**	**96.3989**	**98.0282**	**94.8582**	**92.8303**	**89.0381**	**0.8810**
TMHMM2	75.1515 (124/165)	90.5817	94.6454	91.0567	83.8260	83.9784	0.7904
HMMTOP2	76.3636 (125/165)	94.3213	92.4016	91.4575	84.5456	84.1148	0.7958
PHDhtm v1.96	61.8182 (102/165)	90.4432	90.4432	89.5834	77.1295	84.8826	0.7552
MEMSAT3	77.5758 (128/165)	92.7978	94.4993	90.4748	85.6217	80.9363	0.7800
TopPred2	51.5152 (85/165)	87.1191	89.8571	89.6116	78.0954	81.9970	0.7418
SOSUI 1.1	61.8182 (102/165)	87.2576	93.7500	88.6461	79.9430	80.2708	0.7411
SPLIT4	76.3636 (125/165)	93.3518	93.0939	89.6165	88.1628	77.6921	0.7717
ConPred II	76.3636 (125/165)	94.4598	93.8102	92.1497	85.2716	86.0076	0.8132
Phobius	75.1515 (124/165)	91.8283	95.3957	90.6124	84.5845	84.8375	0.8010
PolyPhobius	72.1212 (119/165)	93.0748	92.6897	91.2168	87.1321	84.2010	0.8122

The total number of low-resolution proteins is 165.

IV. Conclusion

From our analysis of the benchmark by Kernytsky and Rost [17], it is evident that there are some critical issues about the benchmark that would adversely affect how the results are interpreted for an objective comparison. A large portion of low-resolution TM proteins (49/165) contains annotations of low reliability and signal peptides (28/165). Another issue with this benchmark is its annotations that are inconsistent with the current information in SwissProt Release 54.1. We have shown that, by updating the annotations and removing the critical issues inherent in the original data, a more objective evaluation that agrees with previous results can be obtained. In summary, the original benchmark should not be used for a fair comparison; if used, the results should be carefully interpreted because the results may be affected to a large extent due to the existing biases in the data. With the continual updates in SwissProt, we believe that it is of utmost importance to update the topological information whenever possible. This has a significant impact on developing TM topology prediction methods because the training and testing data must be updated in order to achieve better performance. In addition, the results also suggest a strong need for development of a new benchmark of annotations, free of inconsistencies and noises (manuscript in preparation). We also provide the updated benchmark for public use. It is available at http://bio-cluster.iis.sinica.edu.tw/~bioapp/SVMtop/dataset.htm.

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 topology 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] Lo A, Chiu HS, Sung TY, and Hsu WL. Transmembrane helix and topology prediction using hierarchical SVM classifiers and an alternating geometric scoring function. Proceedings of IEEE Computational Systems Bioinformatics Conference, 2006.

[14] Ikeda M, Arai M, Lao DM and Shimizu T. Transmembrane topology prediction methods: A re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol*. 2002, **2**: 19-33.

[15] Möller S, Croning MDR, and Apwiler R. Evaluations of methods for the prediction of membrane spanning regions. *Bioinformatics* 2001; **17**: 646-653.

[16] Chen CP, Kernytsky A, and Rost B. Transmembrane helix predictions revisited. *Protein Sci* 2002; **11:** 2774-2791.

[17] Kernytsky A and Rost B. Static benchmarking of membrane helix predictions. *Nucleic Acids Research*, 2003; **31**(13): 3642-3644.

[18] Cuthbertson JM, Doyle DA, and Sansom MS. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.* 2005, **18**(6): 295-308.

[19] Möller S, Kriventseva EV, and Apweiler R. A collection of well characterised integral membrane proteins. *Bioinformatics* 2000, **16**(12): 1159-1160.

[20] Lo A, Chiu HS, Sung TY, Lyu PC, and Hsu WL. Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J. Proteome Res*. (submitted)
[21] Bairoch A, and Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* 1997, **75**(5): 312-316.
[22] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 1990, **215**:403-410.
[23] Jayasinghe S, Hristova K, and White SH. MPtopo: A database of membrane protein topology. *Protein Sci.* 2001, **10**(2): 455-458.
[24] Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007**, 23**(5): 538-544.
[25] von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 1992, **225**(2): 487-494.
[26] Hirokawa T, Boon-Chieng S, and Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 1998, **14**(4): 378-379.
[27] Juretic D, Zoranic L, and Zucic D. Basic charge clusters and predictions of membrane protein topology. *J. Chem. Inf. Comput. Sci.* 2002, **42**(3): 620-632.
[28] Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, and Shimizu T. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.* 2004, **32**: W390-393.
[29] Kall L, Krogh A, and Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 2004**, 338**(5): 1027-1036.
[30] Kall L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 2005, **21** Suppl 1, i251-257.
[31] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* 1975, **405**(2), 442-451.