Biochimica et Biophysica Acta 1818 (2012) 2679–2686

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Improving transmembrane protein consensus topology prediction using inter-helical interaction

Han Wang [a], Chao Zhang [b], Xiaohu Shi [a], Li Zhang [a], You Zhou [a,*]

[a] College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China
[b] Department of Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Article history:
Received 12 January 2012
Received in revised form 29 May 2012
Accepted 31 May 2012
Available online 6 June 2012

Keywords:
Transmembrane
Topology
Consensus prediction
Contact

ABSTRACT

Alpha helix transmembrane proteins (αTMPs) represent roughly 30% of all open reading frames (ORFs) in a typical genome and are involved in many critical biological processes. Due to the special physicochemical properties, it is hard to crystallize and obtain high resolution structures experimentally, thus, sequence-based topology prediction is highly desirable for the study of transmembrane proteins (TMPs), both in structure prediction and function prediction. Various model-based topology prediction methods have been developed, but the accuracy of those individual predictors remain poor due to the limitation of the methods or the features they used. Thus, the consensus topology prediction method becomes practical for high accuracy applications by combining the advances of the individual predictors. Here, based on the observation that inter-helical interactions are commonly found within the transmembrane helixes (TMHs) and strongly indicate the existence of them, we present a novel consensus topology prediction method for αTMPs, CNTOP, which incorporates four top leading individual topology predictors, and further improves the prediction accuracy by using the predicted inter-helical interactions. The method achieved 87% prediction accuracy based on a benchmark dataset and 78% accuracy based on a non-redundant dataset which is composed of polytopic αTMPs. Our method derives the highest topology accuracy than any other individual predictors and consensus predictors, at the same time, the TMHs are more accurately predicted in their length and locations, where both the false positives (FPs) and the false negatives (FNs) decreased dramatically. The CNTOP is available at: http://ccst.jlu.edu.cn/JCSB/cntop/CNTOP.html.

1. Introduction

Alpha helix transmembrane proteins (αTMPs) are found in all biological membranes and play a very important role in many critical life processes [1], such as the signaling of regulatory networks, cell-to-cell communication, and the transport of membrane-impermeable molecules. As the major category of integral membrane proteins, αTMPs are the prime targets for more than half of the drugs in the current market [2], for that reason, the conformations of αTMPs are indispensable. But transmembrane (TM) proteins are hard to crystallize outside of the biological membranes, high-resolution transmembrane protein structures remain scarce in comparison with globular proteins in the Protein Data Bank (PDB) [3], where they comprise less than 2% of the total number of proteins [4]. Therefore, with the improvement of next-generation sequencing technique, the sequence-based structure prediction for αTMPs becomes more and more useful.

Background info

Generally αTMPs contain one or more stretched helixes forming the bundles to cross the biological membrane, and these helixes are so called transmembrane helixes (TMHs). As the first step of structure prediction, topology prediction is used to predict the entire topology structure for a sequence. A predicted topology describes all the possible TMH(s) and their locations on the sequence, and the location of the N-terminal. With the accurate topology prediction, the protein structures can be better predicted, and the protein functions may be inferred. For decades, many methods have been developed for αTMP topology prediction. Early research mainly relied upon the hydrophobicity scales, which determine TMHs mostly by the hydrophobic properties of the residues, such as TopPred [5], DAS-TMfilter [6], SOSUI [7], and the one using the "positive-inside" rule [8]. Hidden Markov Model (HMM) based methods integrate many TM-specific features to identify TMHs, such as hydrophobic properties, residue polarity, e.g. HMMTOP [9], TMHMM [10], TMMOD [11], THUMBUP [12], Phobius [13], and PRODIV-TMHMM [14]. Limited by the HMM model, these methods are sensitive to the sequence-length when the TMH is too short (<16 residues) or too long (>35 residues) [15]. Many machine learning methods also have been employed. Among them, PHD [16] and MEMSAT [17] utilized neural networks (NN) in combination with evolutionary information; SVMTOP [18] and MEMSET-SVM [19] introduced the support vector machines (SVM) into prediction; MemBrain [15] combined numerous

* Corresponding author at: 2699 Qianjin St., B530 Computer Science Building of Jilin University Changchun 130021, China. Tel.: +86 13086869997 (mobile), +86 937 601 8886.
E-mail address: zyou@jlu.edu.cn (Y. Zhou).

0005-2736/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbamem.2012.05.030

machine learning methods together to improve accuracy. However, all of these methods have their limitations, and their prediction accuracy has been overestimated in whole-genome studies [20,21]. Compared with individual methods, consensus methods appeared to yield better results [22–25] by utilizing the advantages of integrated individual methods. However, they achieved limited improvements in the absence of additional guidance from the TM-specific properties.

TM-specific structural properties are considerably important for improving the prediction accuracy. As one of the most important structural properties, inter-helical interaction strongly influences the protein folding and stability [26], and it can be observed from residue–residue contacts. Contact prediction methods have been widely used in globular protein structure prediction and related research, but these methods did not perform very well for TM proteins. Notably, the contact prediction for TMHs is garnering increasing attention. The TMH residue contact has been certainly analyzed and classified, for instance, coevolving residue analysis [27] and TM environment knowledge-based potential energy matrix [28]. A number of TMH contact predictors are available, such as TMhit [29], TMHcon [30], MEMPACK [31], and TMhhcp [32], and many of them have been applied to the TMH folding prediction deriving more reliable structures [31,33,34], but none of them has been introduced to improve the topology prediction for αTMPs.

In this study, we propose a SVM-based consensus method, CNTOP, which firstly introduces the inter-helical interactions into topology prediction of αTMPs. In order to quantify the inter-helical interactions, the contact strength of TMH residue has been calculated. Then it has been used to comprise a five dimension vector for each possible TMH residues with the topology prediction results from four top leading predictors. The method utilizes the vectors to identify the TMH residues, and then predicts the entire topology structure. The CNTOP takes the advantages of the consensus method such that all the potential TMHs are possibly to be found, and more importantly, the prediction is guided by the TM-specific structural characteristic, which enriches the feature for the prediction of the sequence patterns. Compared with any individual topology predictor and other consensus predictors, the CNTOP achieved the best prediction accuracy on two benchmark datasets, and outperformed all its counterparts against our non-redundant testing dataset, especially in accurately predicting the locations of TMHs.

2. Materials and methods

2.1. Data sets

Two benchmark datasets were used to compare the performance of CNTOP against other available methods. The first one containing 184 sequences, is a subset of the Möller set [7,35–37], which annotates the sequences with both crystal structures and biochemical characterization. The other one, Topology Data Bank of Transmembrane Proteins (TOPDB) [38] has 1452 αTMP sequences, including 510 bitopic sequences and 942 polytopic ones. It is the most complete and comprehensive collection of transmembrane protein datasets containing experimentally validated topology information. For the purpose of conducting a large-scale test on the method, we further selected the database Protein Data Bank of Transmembrane Proteins (PDBTM) [39,40], which identifies and annotates the TMPs from the Protein Data Bank (PDB) [3] by their 3D structures. The PDBTM includes 1302 αTMPs (released on 9/30/2011), from which 5779 sequences were parsed, and then 2879 sequences were left after removing the bitopic αTMPs which have no inter-helical interactions. To avoid the influence of homologous sequences, we clustered the 2879 sequences with 30% identity, so the sequences from different clusters are non-redundant. The two biggest non-redundant clusters were collected respectively as testing dataset and training dataset. The training dataset contains 153 sequences, and the testing dataset has 223 sequences. There are no overlaps between the training and test datasets (Support Table S1, S2).

2.2. Contact strength

Although TMH topology prediction is continuously improving with diverse methods and sequence-based features, the individual predictors still cannot yield the expected accuracy. To date, many features derived from amino acid sequences have been taken into consideration, such as sequence profiles, residue substitution matrixes, statistics-based TMH residue frequency, and the TMH-specific residues. However, as an important structural characteristic, TMH contact has not been used. The αTMPs forming the stable inter-membrane structures highly depend on the helix–helix interactions, and the residue contacts are the essential driver. Thus, it can be inferred that the sequences with more contact-active residues have bigger chances to be TMHs. The strength of contact activity can be represented as the contact strength. Only the predicted contact can be used to obtain the contact strength in this study, but even the predicted contact strength of the TMH residue shows the potential to improve the accuracy of topology prediction theoretically.

For those known-structure proteins, the residue contact has been clearly defined. There are three definitions for the existing contact between a pair of residues: 1) 8 Å as a maximal distance between their C-beta atoms (C-alpha for glycines) [41–43]; 2) the distance between any two atoms from the pair is less than the sum of their van der Waals radii plus 0.6 Å [29,44]; 3) the minimal distance between side chain or backbone heavy atoms in an the pair is less than 5.5 Å [30]. The TMH residue contact predictor MEMPACK [31], which was used in our method, optimizes the prediction results based on the three definitions above. It employs the PSI-BLAST [45] profiles and lipid layer exposure SVM prediction scores as features to predict the residue–residue contact. Several other strategies are also applied to improve the prediction accuracy, such as a 7 residue slide window is used to detect the contacts rather than a single residue which detects the interactions between TMH packing motifs [26]. In addition, the residues with various sequence distances are taken into consideration whether they have contact or not, and many other TMH-specific features, such as the TMH lengths are used as constraints to adjust the prediction results [46].

MEMPACK predicts the contacts for all TMH residue pairs on a given topology structure, where the residues of each pair belong to the different TMHs. The predicted contacts are real number scores produced by a SVM model, where a nonzero value of the score reveals that the pairs of residues are all TMH residues, while a zero value infers that at least one residue is not the a residue. Notably, the existence of contact can be identified by a positive score or denied by a negative score, and the bigger absolute value of the score indicates the prediction is more reliable. To further describe the contact activity of a particular TMH residue, the contact strength $ConS_i$ of the TMH residue at position i is defined as follows:

$$ConS_i = \frac{\sum_j \ln(|contact(i,j)| + \varepsilon) \cdot pair(i,j)}{\sum_j pair(i,j)}, \quad (1)$$

where $pair(i,j) = \begin{cases} 1 & \text{if } i,j \notin \text{same } TMH \\ 0 & \text{otherwise} \end{cases}$, and $\varepsilon = 1.0e{-}50$.

$contact(i,j)$ is the contact score for the residue pair at position i and j predicted by MEMPACK, ε is a positive constant used as a pseudocount. By Eq. (1), the contact strength of the residue at position i is defined as the mean contact with all residues in other TMHs. The bigger the $ConS_i$ value is, the more the residue at position i is considered to be a TMH residue. Further usage of contact strength will be introduced in Section 2.4.2.

2.3. Incorporated topology predictors

There are several available TMH topology predictors, among which various αTMP-specific patterns and computational methods

are employed. Based upon their performances and the models, four top leading predictors were selected for CNTOP, namely, TMHMM 2.0 (TM) [10], TMMOD 3.0 (TD) [11], MEMSAT3 (MS) [17], and MEMSAT-SVM (MV) [19]. The TM, a successful HMM-based predictor, incorporates hydrophobicity, charge bias, helix lengths, and grammatical constraints into one model, and it is possible to model the helix length [10,47]. Another HMM-based method, TD, differs from TM in the architecture by using submodels for loops on both sides of the membrane, and the model parameters are also different. The MS firstly introduces the sequence conservation information to the topology prediction and uses a NN to score each possible TMH, and determinates the final topology by searching all possible topological models with a dynamic programming algorithm. The MV uses the evolutionary information as the key feature to find all kinds of segments in the sequences, including TM helix/none-TM helix, inside loop/outside loop, reentrant helix/none-reentrant helix and signal peptide/none-signal peptide, and four corresponding SVMs are adopted to identify those segments.

These four predictors are the typical representatives of the major methods in this field, and cover most of the sequence-based features which can discriminate the TMPs from the globular proteins, so that, they can capture the sequence's characteristics and patterns from diverse perspectives. Whereas, it guarantees that most native TMHs can be found by them, at least one of them. In order to use the predicted topologies, the topology types are unified to the same format, among which the TMH residues are denoted 'H', and outside parts (in the extracellular) and inside parts (in the cytoplasm) are respectively denoted 'o' and 'i', while the unknown ones are 'U'.

2.4. CNTOP topology prediction

The CNTOP predicts the topology for an αTMP according to the steps shown in Fig. 1. As the inputs of step 1 and step 3, the topologies of the target sequence must be obtained by running above four incorporated predictors. Step 1 collects all the possible TMHs from four predicted topologies, and then produces a decoy topology for the target. The goal of this step is to collect all possible TMHs, whether the prediction is true or not, because when one native TMH residue is missed, it has no chance to be identified as a TMH residue any more in our method. In the procedure, the N-terminal location is initialized using a voting mechanism from the four predicted locations, and the details will be described in Section 2.4.1. Step 2 uses the decoy topology as input to calculate the contact strength for each TMH residue on the decoy topology, which has been introduced in Section 2.2 and defined in Eq. (1). In step 3, each TMH residue on the decoy topology will be identified whether or not it is a convincible TMH residue by a SVM, which uses the contact strength and topology type (H, o, i, or U) of the residue as input vector. The SVM will be introduced in Section 2.4.2. In order to collect all possible TMH residues into the decoy topology, many uncorrected predictions are included, but they will be recognized by the SVM. Then their topology types will be corrected on the original decoy topology, and it will result into the non-TMHs and the N-terminal location changing. The final output topology is produced based on the readjustment of the decoy topology, which is the responsibility of step 4, and is discussed in Section 2.4.3.

2.4.1. TMH collection for decoy topology

In the procedure of producing a decoy topology, three cases are taken into consideration. Fig. 2(a) shows the simplest and most common case. All predictors predicted the same TMH with only minor differences in their locations, and the predicted TMH is clearly independent from the other TMHs. Thus, the corresponding TMH starts from the left most and ends at the right most end of all the predicted locations. Sometimes some predictors predict TMHs that are inconsistent with other predictors. As shown in Fig. 2(b), the fourth predictor predicted a single TMH overlapped with two TMHs predicted by other predictors, and this makes it difficult to tell where the native TMH is, such that the TMH will be ignored. The last case, illustrated in Fig. 2(c), one separated TMH was predicted by the fourth topology, while the other three predictors have not predicted a TMH in the corresponding location. In order to get all possible TMHs, these two regions are both collected on the decoy topology.

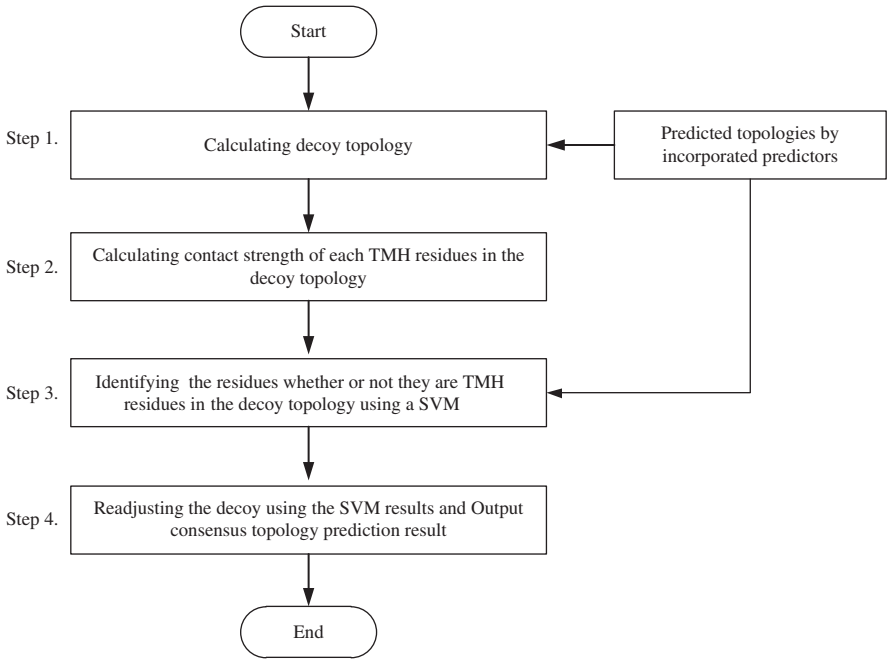

Fig. 1. The prediction procedure of CNTOP. Each box corresponds to a step executed by CNTOP. All the last three steps use the outputs of their previous step as input, and the four incorporated individual predictors provide the predicted topologies as the input for step 1, as well the additional input for step 3.

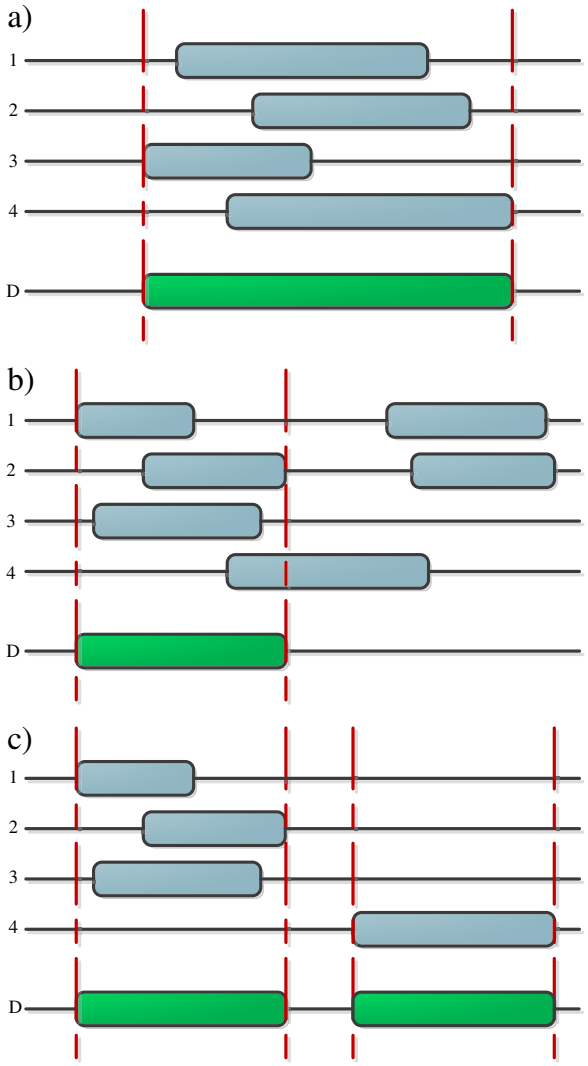

Fig. 2. The TMH collection for the decoy topology. Each line with a label represents a topology structure predicted by corresponding predictors, where the topologies predicted by individual predictors are labeled as numbers, and the decoy topology is labeled as D. The boxes on the line correspond to the predicted TMHs. To outstand the TMHs collected by the decoy topology, the corresponding boxes use the color green. (a) The collection strategy for the TMH consistently predicted by all the individual predictors. (b) The collection strategy for the inconsistently predicted TMHs, where the inconsistent one has overlaps with TMHs predicted by the other predictors. (c) The collection strategy for the TMHs that are absent or present only on a few topologies.

2.4.2. SVM model

SVM has been widely used in bioinformatics solving the classification problem [48]. It can form a non-linear higher dimension separating hyperplanes to separate the vectors from the data sets. Here, the elements in the feature vectors are required to be non-linear dependent. In this study, we used the SVM model as a binary classifier to identify whether each residue on the decoy belongs to the TMH, or not. For each of them, the four predicted topology types and the contact strength are non-linear dependent. They compose a five dimensional vector $V_i = (TM_i, TD_i, MS_i, MV_i, ConS_i)$, where i refers to the residue position in the target entry, TM_i, TD_i, MS_i, MV_i respectively represent the residue topology types (TMH residue or not) predicted by the four predictors, and $ConS_i$ is the contact strength. TM_i, TD_i, MS_i, MV_i are respectively assigned to 1 when the corresponding predicted topology types are the TMH, otherwise they are assigned to 0. The contact strength uses the value calculated in Eq. (1). The SVM will denote topology type 'H' to the residues which have been identified as TMH residues and 'U' to those fake TMH residues on the decoy, the rest of the non-TM residues keep the original topology type.

We generalized the model using the LIBSVM [49] toolkit version 3.11. The radial basis function (RBF) kernel was used as the kernel function, and the optimal parameters (C = 8.0, g = 0.0078125) were obtained by using a grid-search with the training data set. The residues that have been identified as non-TMH residues are then marked.

2.4.3. Readjusting decoy topology

Based on the results of the SVM, we scan the whole to produce a final prediction result. Firstly, we record the position i of the first TMH residue as the start point of the TMH. Then we keep moving the window until it reaches position j, when both residues at positions $j+1$ and $j+2$ are not the TMH residues. The sequence from position i to j is determinated as a TMH if the distance between the two positions is larger than 12 residues. Since the TM helix segments are normally between 17 and 25 residues [50], using 12 residues as threshold is reasonable to indicate the existence of TMHs. We can search the next TMHs by using the same method from the position $j+3$ till reaching the end of sequence.

The above procedure may remove a few fake TMHs from the original decoy topology, and thereby results in the non-TM segments becoming irregular with the topology type, so that the non-TM segments have to be readjusted to produce an output topology. For the purpose, we firstly scan the current decoy for those non-TM segments which include the 'U' type residues, and denote the topology type 'U' to the whole segments; then seek the 'i' type non-TM segment nearest the 'U' type segments. According to the direction from the 'i' type segment to the 'U' type segments, all the non-TM segments are denoted to topology types 'i' and 'o' alternatively, so that a compete topology structure is produced. Here, because the inside segments are more reliably detected compared with the outside parts by means of the positive-inside rule, the nearest inside segment is used to decide the topology type for those unknown segments.

2.5. Availability

The CNTOP is implemented using Java language, the executable program is available for free downloading at: http://ccst.jlu.edu.cn/JCSB/cntop/CNTOP.html, there are no restrictions to use by academics. The program runs on Linux and Windows with Java Runtime Environments supported (version 1.6 and up).

3. Results and discussion

3.1. A sample of CNTOP

To further understand how CNTOP works, the details of prediction are shown in Fig. 3 step by step using the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complex (PDB ID: 1AR1:B) [51] as a sample. The outputs of all steps are presented and discussed as follows:

In the first step, we produced the decoy topology using four incorporated predictors. Comparing the native topology of 1AR1_B, three out of four predictors (TM, MD and MS) predicted the wrong numbers of TMH and the N-terminal location. Those three predictors failed in the first TMH, so they treated N-terminal as inside, and only the MV predicted the correct topology entirely. According to the TMH collection strategy introduced in Section 2.4.1, three TMHs were collected to the decoy topology including the incorrect one located at (11, 34), and the other two native TMHs were lengthened respectively from (65–84) to (65–89) and (104–126) to (105–137). Meanwhile, the N-terminal of the decoy was also incorrectly labeled as inside, according to the voting mechanism. In this step, CNTOP collected all of the possible TMHs and denoted

Fig. 3. The CNTOP prediction processes and their performances. Each line represents a topology structure predicted by the predictors labeled in front of the line, where TM: TMHMM 2.0, TD: TMMOD 3.0, MS: MEMSAT3, MV: MEMSAT-SVM, the length of sequence is showed at the end of the line, and the boxes are the same as shown in Fig. 2, while the numbers in the boxes mark the locations of the TMHs. The output of each step is listed under the corresponding arrows, among which the bars shown after step 2 present the contact strength of each TMH residue on the decoy topology, and the bars after step 3 represent the corresponding residues that have been identified as TMH residues. The final predicted topology of CNTOP is compared with the native one and showed after step 4.

an initial N-terminal for the decoy topology with a fake TMH and the wrong N-terminal were predicted, but the next three steps will correct these "mistakes".

Then, the pairwise contacts of predicted TMH residues on the decoy topology were calculated by using MEMPACK. Shown as the result of step 2, the regions with enriched contact activity residues perfectly matched the native topology, where the incorrect TMH did not have contact activity residues. There was a non-TMH residue appeared contact activity in position 87 and a native TMH residue missed contact activity in position 126, however, most native TMH residues on the decoy topology can be determined by the mean of the residue contact strength.

Step 3 is the most important step to validate each predicted TMH by using the SVM discussed in Section 2.4.2. The identification processing only applies to all residues predicted to be TMH residues, and that is the reason why we try to collect all the possible TMHs and TMH residues to the decoy topology. Found in the results of step 3, those identified TMH residues show bars at corresponding positions, while non-TMH residues have no bars. It illuminates the fact that the SVM has a capacity to identify most TMH residues from an

inaccurate topology, although there still are few residues that were identified incorrectly. Though a fake TMH had been collected on the decoy topology in the previous step, there was only one residue incorrectly identified as TMH in its location. And based on it, the fake TMH will be recognized and corrected in this step.

Resulting from step 3, the locations of TMHs on the decoy topology had been adjusted when the topology types of many residues were changed, and so were the adjacent non-TMHs. In this example, the readjustment step removed the fake TMH from the decoy topology and lengthened the second TMH in positions 67 and 83. Although the residue in position 87 is a TMH residue, it had to be excluded from the TMH for standing far from other contact active residues. Then the topology types of the non-TM segments had to be reassigned upon removing the fake TMH. According to our readjustment strategy, the N-terminal was predicted to be located outside. Finally, the CNTOP predicted the correct topology of 1AR1_B, including the TMH number, TMH locations, and the N-terminal location. There were only 2 residues that had been incorrectly predicted by CNTOP, respectively in positions 65 and 126, but there were 10 incorrect residues predicted by the MV, and the results of the other three predictors were even worse.

3.2. Accuracy of topology prediction

Generally, the accuracy of the topology prediction can be accessed from three perspectives: 1) the number of TMHs; 2) the locations of those predicted TMHs; and 3) the N-terminal location. As the most common evaluation criteria, the topology accuracy has been used by many methods [10,11,13,17,52], and it counts for the topologies that both the TMHs and the N-terminal location have been correctly predicted, where the TMHs is considered to be correct when the TMH number and their locations are all correct, and a TMH is counted as a correct one when it has at least five residues overlapping with the native TM segment.

The topology accuracy can roughly describe the performance of a topology predictor, but it cannot further evaluate the prediction precision. As shown in the above example, CNTOP and MV have similar performances in terms of topology accuracy, but MV predicted the incorrect residues four times more than the CNTOP, which cannot be presented by the topology accuracy. Furthermore, the number of correct TMHs will decrease when the overlap rises [17], but the lengths of the TM helix segments are normally between 17 and 25 residues [50], the reasonable choice of the overlap has not been discussed. To evaluate the prediction accuracy more accurately, we reference the false positive (FP) rate and the false negative (FN) rate of TMH residues. The FPs are those native TMH residues which have been predicted as non-TMH residues; on the contrary, the FNs are those wrongly predicted non-TMH residues. Lower FP rate and FN rate indicate that more residues have been predicted to the correct topology type, thereby the predicted TMHs are more likely to be located in the correct places and with the proper length. In addition, both the over-prediction and under-prediction can be detected by the FP rate and FN rate. The over-prediction predicts more TMHs or enlarges those TMHs, by which it can increase the correct rate of topology by sacrificing the FN rate. And the under-prediction is too conserved in predicting leading to increases in the FPs. Therefore, the overall improvement of topology prediction should satisfy two constraints: 1) improving the topology accuracy; 2) decreasing both the FPs and FNs at the same time.

3.3. Comparison with other predictors

3.3.1. Accuracy rates of topology on benchmark datasets

For the comparison of the prediction accuracy with other top leading topology predictors, we used the benchmark datasets, Möller dataset and TOPDB dataset. To clearly present the performances, the comparison was made using topology accuracy, where the correct predicted TMH was defined as having 5 residues overlap with a native one. As shown in Table 1, for the Möller dataset, CNTOP achieved the best topology accuracy (87%) among all predictors, including six individual predictors and one consensus predictor, among which, MT obtained a prediction accuracy (80%) better than all the individual predictors and MV was the best individual predictor which obtained the highest accuracy (78%). The MT also used SVMs to predict the TMHs and N-terminal locations, but did not use any additional structural information. CNTOP outperformed MT, because it incorporated the contact strength to identify the TMH residues, and it is also the most important contribution of CNTOP. Meanwhile, CNTOP achieved the best accuracy (75%) on the TOPDB dataset, which surpassed the consensus method MT by 6%, and the margin enlarged to 8% compared with the best individual predictor MV.

Table 1
Numbers of TM segments predicted by different topology predictors using benchmark datasets.

Method	Algorithm	Topology Acc. (%)	
		Möller dataset	TOPDB dataset
CNTOP	Consensus	**87**	**75**
MT	Consensus	80	69
TM[a]	HMM	60	56
TD	HMM	62	65
MS[a]	NN	77	66
MV[a]	SVM	78	67
PRODIV[a]	HMM	46	37
SVMTOP[a]	SVM	70	42

The abbreviations of the methods are the same as described in Fig. 3. The accuracy of topology is the correct rate of the predicted TMH, where the correct topology means that all its TMHs are correctly predicted as well as the N-terminal location. The best prediction accuracy of each data set is marked using bold text.

[a] The corresponding prediction accuracies were previously reported by Nugent et al. [19].

For each predictor, the prediction accuracy on the Möller dataset is obviously higher than that on the TOPDB dataset, which is caused by the different sizes of the datasets and the errors that existed in them. The TOPDB dataset is almost eight times larger than its counterpart, it is reasonable that the statistical features of topology drop down against such a big sample space. Although it was reported that only 69% of the original Möller topologies are correct [19], the Möller dataset collected the sequences from the previously used datasets, or literature-derived, the proteins that have no biochemical characterization available were excluded, while the TOPDB determinates the topologies using the 3D structures. Thus the Möller dataset will lead to a higher prediction accuracy.

3.3.2. Overall performance on non-redundant dataset

To further access the performance of CNTOP, we compared our results with those of four incorporated individual methods and one of the other consensus methods using the same dataset. Here, the topology accuracy is used as a basic criterion, while two other criteria also have been adapted, the TMH prediction accuracy and the N-terminal prediction accuracy. Furthermore, the FN rate and the FP rate are used as additional criterion to present the prediction accuracy of the TMH locations, where, the decrease in the both rates indicates that the TMH locations are better predicted. Differing from the Möller dataset and TOPDB dataset, our testing dataset is a non-redundant and polytopic-protein-only dataset. The existence of homologous sequences will increase the correct rate of topologies when the method was trained to be familiar with them, but it is the opposite for the other methods, while the non-redundant dataset can present the performance more comprehensively. The CNTOP advances the prediction for polytopic proteins profited by utilizing the contact strength, while the bitopic proteins are easier to

be accurately predicted by all the predictors. To completely present the improvement, our testing dataset excluded the bitopic proteins, which accounts for about 19% of the Möller dataset, and 35% of the TOPDB dataset.

As shown in Table 2, CNTOP achieved the best topology accuracy (77.6%), best FP (11.9%) and best FN (6.9%) rates. Among the individual predictors, TM and TD obtained the lowest topology accuracy since they have the highest FP and FN rates than other predictors. MS and MV have similar performances at overall accuracy. MV is better in topology accuracy, but cannot surpass MS in TMH location accuracy. But the slight superiority of MV in topology accuracy is mostly contributed by the higher N-terminal accuracy, while its TMH accuracy is about 2% lower than MS. However, the much lower FN rate of MV indicates that it predicted less fake TMH than MS. No doubt that the CNTOP is superior to any individual with respect to all the aspects, it has 7% improvement of topology accuracy compared with that of the best individual predictor. However, the more important fact is that CNTOP derives such an improvement based on the condition that both FP rate and FN rate are decreased, which means the increased TMH prediction accuracy did not come from over-prediction, the TMHs were more accurately predicted in either the number or their locations.

CNTOP outperformed MT in terms of predicting the TMH number and their location and is also far better at detecting the TMHs surpassing MT by almost 7%. With a similar residue FP rate, CNTOP has a much better FN rate, resulting in predicting more accurate locations of the TMHs. The results proved that the TMH contact strength played a very important role in the prediction. Unlike the features used in model based methods, contact strength brings TM-specific structural characteristics into the SVM model, and makes TMH residue identification more reliable and decreasing the FN. CNTOP was slightly worse than MT in N-terminal prediction, the reason mainly because of the average N-terminal prediction abilities of the integrated predictors. CNTOP derives the best overall accuracy of topology prediction compared with all the incorporated individual predictors, and it is better than another consensus predictor MT in predicting TMHs and their locations.

4. Conclusions

This paper describes a novel consensus-based topology prediction method for αTMPs, namely CNTOP, which incorporates four top leading individual topology predictors, and also introduces the contact strength of residues to identify the TMHs and their locations to improve the topology prediction accuracy, especially for the polytopic αTMPs. The performance of CNTOP was compared to six other individual predictors and one consensus method using two commonly used benchmark datasets. Our method achieved an 87% prediction accuracy, that is 9% better than that obtained from the best individual predictors, and 7% better than that obtained from a consensus method. A more challenging comparison has been proposed based on our non-redundant testing dataset and evaluated using more criteria. The CNTOP remains the best predictor with the lowest FP and FN rates.

Table 2
The topology prediction precision comparison.

Method	Topology Acc. (%)	TMH Acc. (%)	N-terminal Acc. (%)	FP rate (%)	FN rate (%)
CNTOP	**77.6**	**78.3**	**84.3**	**11.9**	**6.9**
MT	70.9	71.2	82.1	11.8	13.7
TM	49.2	51.1	72.6	18.4	15.3
TD	54.4	58.9	71.5	19.4	14.4
MS	69.7	73.6	76.4	12.7	15.9
MV	70.6	71.5	80.2	15.9	8.7

The abbreviations of methods are the same as described in Fig. 3. The prediction accuracy is described from the three basic aspects: correct rate of predicted TMHs, N-terminal location and the accuracy of TMH locations which is descripted by the FP rate and FN rate together. The lower of both rates indicates the higher prediction accuracy of the TMH locations derived. The best result of each criterion is marked using bold text.

The results demonstrate that the inter-helical interactions of αTMPs are helpful for the identification of the TMHs using contact strength, which can more accurately locate the TMHs. CNTOP utilizes the advances of the consensus methods and the TM-specific structural property, so it can improve not only the prediction accuracy of the TMH number, but also the prediction accuracy of TMH locations, which is even more challenging for other predictors, so it improves the overall performance of the topology prediction. The performance of our method still has room for improvement with better development of TMP contact prediction and individual topology predictors, and it also can shed light on the studies of TMP structure prediction and function prediction.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 61175023, 60973092, 60903097, the Science-Technology Development Research Project from Jilin Province of China No. 201215022, and the Ph.D. Program Foundation of MOE of China (20090061120094). The support from the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China is also acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbamem.2012.05.030.

References

[1] J. Liang, H. Naveed, D. Jimenez-Morales, L. Adamian, M. Lin, Computational studies of membrane proteins: Models and predictions for biological understanding, Biochim. Biophys. Acta 1818 (2012) 927–941.
[2] T. Klabunde, G. Hessler, Drug design strategies for targeting G-protein-coupled receptors, Chembiochem 3 (2002) 928–944.
[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[4] H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, The Protein Data Bank and the challenge of structural genomics, Nat. Struct. Biol. 7 (2000) 957–959 (Suppl.).
[5] M.G. Claros, G. von Heijne, TopPred II: an improved software for membrane protein structure predictions, Comput. Appl. Biosci. 10 (1994) 685–686.
[6] M. Cserzo, F. Eisenhaber, B. Eisenhaber, I. Simon, TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter, Bioinformatics 20 (2004) 136–137.
[7] T. Hirokawa, S. Boon-Chieng, S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, Bioinformatics 14 (1998) 378–379.
[8] G. Heijne, The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology, EMBO J. 5 (1986) 3021–3027.
[9] G.E. Tusnady, I. Simon, The HMMTOP transmembrane topology prediction server, Bioinformatics 17 (2001) 849–850.
[10] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[11] R.Y. Kahsay, G. Gao, L. Liao, An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes, Bioinformatics 21 (2005) 1853–1858.
[12] H. Zhou, Y. Zhou, Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method, Protein Sci. 12 (2003) 1547–1555.
[13] L. Kall, A. Krogh, E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, J. Mol. Biol. 338 (2004) 1027–1036.
[14] H. Viklund, A. Elofsson, Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information, Protein Sci. 13 (2004) 1908–1917.
[15] H. Shen, J.J. Chou, MemBrain: improving the accuracy of predicting transmembrane helices, PLoS One 3 (2008) e2399.
[16] B. Rost, R. Casadio, P. Fariselli, Refining neural network predictions for helical transmembrane proteins by dynamic programming, Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB, International Conference on Intelligent Systems for Molecular Biology, 4, 1996, pp. 192–200.
[17] D.T. Jones, Improving the accuracy of transmembrane protein topology prediction using evolutionary information, Bioinformatics 23 (2007) 538–544.
[18] A. Lo, H.S. Chiu, T.Y. Sung, P.C. Lyu, W.L. Hsu, Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function, J. Proteome Res. 7 (2008) 487–496.

[19] T. Nugent, D.T. Jones, Transmembrane protein topology prediction using support vector machines, BMC Bioinformatics 10 (2009) 159.
[20] K. Melen, A. Krogh, G. von Heijne, Reliability measures for membrane protein topology prediction algorithms, J. Mol. Biol. 327 (2003) 735–744.
[21] L. Kall, E.L. Sonnhammer, Reliability of transmembrane predictions in whole-genome data, FEBS Lett. 532 (2002) 415–418.
[22] J. Nilsson, B. Persson, G. von Heijne, Consensus predictions of membrane protein topology, FEBS Lett. 486 (2000) 267–269.
[23] M. Ikeda, M. Arai, D.M. Lao, T. Shimizu, Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, In Silico Biol. 2 (2002) 19–33.
[24] P.D. Taylor, T.K. Attwood, D.R. Flower, BPROMPT: a consensus server for membrane protein prediction, Nucleic Acids Res. 31 (2003) 3698–3700.
[25] M. Klammer, D.N. Messina, T. Schmitt, E.L. Sonnhammer, MetaTM—a consensus method for transmembrane protein topology prediction, BMC Bioinformatics 10 (2009) 314.
[26] S.E. Harrington, N. Ben-Tal, Structural determinants of transmembrane helical proteins, Structure 17 (2009) 1092–1103.
[27] A. Fuchs, A.J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, D. Frishman, Co-evolving residues in membrane proteins, Bioinformatics 23 (2007) 3312–3319.
[28] A.N. Jha, S. Vishveshwara, J.R. Banavar, Amino acid interaction preferences in helical membrane proteins, Protein Eng. Des. Sel. 24 (2011) 579–588.
[29] A. Lo, Y.Y. Chiu, E.A. Rodland, P.C. Lyu, T.Y. Sung, W.L. Hsu, Predicting helix–helix interactions from residue contacts in membrane proteins, Bioinformatics 25 (2009) 996–1003.
[30] A. Fuchs, A. Kirschner, D. Frishman, Prediction of helix–helix contacts and interacting helices in polytopic membrane proteins using neural networks, Proteins 74 (2009) 857–871.
[31] T. Nugent, S. Ward, D.T. Jones, The MEMPACK alpha-helical transmembrane protein structure prediction server, Bioinformatics 27 (2011) 1438–1439.
[32] X.F. Wang, Z. Chen, C. Wang, R.X. Yan, Z. Zhang, J. Song, Predicting residue-residue contacts and helix–helix interactions in transmembrane proteins using an integrative feature-based random forest approach, PLoS One 6 (2011) e26767.
[33] S.R. McAllister, C.A. Floudas, Alpha-helical topology prediction and generation of distance restraints in membrane proteins, Biophys. J. 95 (2008) 5281–5295.
[34] A. Fuchs, D. Frishman, Structural comparison and classification of alpha-helical transmembrane domains based on helix interaction patterns, Proteins 78 (2010) 2587–2599.
[35] P.G. Bagos, T.D. Liakopoulos, S.J. Hamodrakas, Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins, BMC Bioinformatics 7 (2006) 189.
[36] S. Moller, M.D. Croning, R. Apweiler, Evaluation of methods for the prediction of membrane spanning regions, Bioinformatics 17 (2001) 646–653.
[37] S. Moller, E.V. Kriventseva, R. Apweiler, A collection of well characterised integral membrane proteins, Bioinformatics 16 (2000) 1159–1160.
[38] G.E. Tusnady, L. Kalmar, I. Simon, TOPDB: topology data bank of transmembrane proteins, Nucleic Acids Res. 36 (2008) D234–D239.
[39] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, Nucleic Acids Res. 33 (2005) D275–D278.
[40] G.E. Tusnady, Z. Dosztanyi, I. Simon, Transmembrane proteins in the Protein Data Bank: identification and classification, Bioinformatics 20 (2004) 2964–2972.
[41] M. Punta, B. Rost, PROFcon: novel prediction of long-range contacts, Bioinformatics 21 (2005) 2960–2968.
[42] J. Cheng, P. Baldi, Improved residue contact prediction using support vector machines and a large feature set, BMC Bioinformatics 8 (2007) 113.
[43] A. Vullo, I. Walsh, G. Pollastri, A two-stage approach for improved prediction of residue contact maps, BMC Bioinformatics 7 (2006) 180.
[44] R.F. Walters, W.F. DeGrado, Helix-packing motifs in membrane proteins, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 13658–13663.
[45] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[46] T. Nugent, D.T. Jones, Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm, PLoS Comput. Biol. 6 (2010) e1000714.
[47] E.L. Sonnhammer, S.R. Eddy, E. Birney, A. Bateman, R. Durbin, Pfam: multiple sequence alignments and HMM-profiles of protein domains, Nucleic Acids Res. 26 (1998) 320–322.
[48] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (2006) 1565–1567.
[49] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27.
[50] C.P. Chen, A. Kernytsky, B. Rost, Transmembrane helix predictions revisited, Protein Sci. 11 (2002) 2774–2791.
[51] C. Ostermeier, A. Harrenga, U. Ermler, H. Michel, Structure at 2.7 A resolution of the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complexed with an antibody FV fragment, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 10547–10553.
[52] E.L. Sonnhammer, G. von Heijne, A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences, Proceedings/… International Conference on Intelligent Systems for Molecular Biology ; ISMB, International Conference on Intelligent Systems for Molecular Biology, 6, 1998, pp. 175–182.