A Consensus Transmembrane Topology Prediction Method of High-Reliability

Jun-Xiong Xia srka@si.hirosaki-u.ac.jp
Toshio Shimizu slsimi@si.hirosaki-u.ac.jp

Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan

Keywords: transmembrane protein, transmembrane topology prediction, consensus approach, high reliability, genome-wide analysis

1 Introduction

It has been revealed that the function of transmembrane (TM) proteins (20-30% in most genomes [1]) can be classified and identified with the information of its TM topologies, i.e., the number of TM segments (TMSs), the position of TMS and the orientation of the TMS to the membrane lipid bilayer [6]. Therefore, developing the TM topology prediction method with high reliability is critical task for the elucidation of TM protein functions. Although many TM topology prediction methods have been proposed, the prediction accuracies of these methods are still not high enough, i.e., at most 50-60% as to whole TM topology [3]. In this study, we propose a new consensus approach (ConPred_elite) with reliabilities of 0.98 and 0.95 for prokaryotic and eukaryotic TM protein sequences, respectively, by combining the results from five currently used TM topology prediction methods. We applied this method to TM proteins extracted from 87 prokaryotic and 12 eukaryotic proteomes.

2 Materials and Methods

As a training dataset in this study, we used TMPDB_alpha_non-redundant dataset, which includes 138 prokaryotic and 93 eukaryotic sequences with experimentally-characterized TM topologies [4]. Completely sequenced genome data of 87 prokaryotic (15 archaebacterial and 72 bacterial) and 12 eukaryotic proteomes were also obtained from the public databanks.

The combinations of five methods from the six selected TM topology prediction methods, i.e., TMpred, TopPred II, TMAP, MEMSAT 1.8, TMHMM 2.0 and HMMTOP 2.0, are used for consensus TM topology prediction. We target the sequence that all the five methods could both predict the N-tail location and the number of TMSs (=1). Then, we check each position of the corresponding TMSs. The maximum distance of the center positions between each corresponding TMS is calculated and compared with the defined value of the "allowable deviation", n (residues). If the distance is within n, then the average of the five center positions is calculated. Only when the consensus TMS prediction is made for all the TMSs by repeating this process toward the C-terminus, can the consensus TM topology prediction be completed for the target sequence. The consensus TM topology prediction is categorized into two modes. One is that all the five predictions agree on one TM topology model (*agree_one*). The other is the predictions split into two TM topology models, i.e., all the five predictions agree on the number of TMSs and TMS positions but disparity in N-tail location (*split_two*).

If the center positions of all the predicted TMSs coincide with their corresponding TMSs of experimentally-characterized data within 11 residues and also the N-tail locations are consistent with each other, the consensus TM topology model is considered as correct.

3 Results and Discussion

By using TMPDB_alpha_non-redundant dataset, we determined the best combination and the optimal value of n (7-20 residues) for prokaryotic and eukaryotic sequences separately. The combination, "TMpred, TMAP, MEMSAT 1.8, TMHMM 2.0 and HMMTOP 2.0" was obtained as optimal for both prokaryotic and eukaryotic sequences with ns of 15 and 11 residues, respectively (ConPred_elite). ConPred_elite can achieve the prediction reliabilities of 0.98 for prokaryotic and 0.95 for eukaryotic sequences, although the prediction efficiencies which could be called "yield", are estimated at 30.4% and 21.5%, respectively.

We applied ConPred_elite to the TM protein sequences predicted by SOSUI [2] from 87 prokaryotic and 12 eukaryotic proteomes. The signal peptide regions detected by DetecSig [5] were removed before the TM topology prediction. The finally obtained TM topology data compose of 12,345 prokaryotic and 7,404 eukaryotic sequences, i.e., corresponding to yields of 22.9% and 13.3%, respectively. By using BLAST (Altschul *et al.*, 1997) and ALIGN (Myers and Miller, 1998), these sequences were classified into three types, i.e., "known", "putative" (similar to known), "unknown", according to the SWISS-PROT annotations. As show in Table 1, 82.2% and 75.5% of prokaryotic and eukaryotic sequences are of "unknown" function, respectively. In these function-unknown sequences, not small number of functionally important novel TM proteins are expected to be contained. Given the reliable TM topology information by ConPred_elite, these TM proteins would be characterized rather easily in near future.

Table 1: The state of functional identification against TM protein sequences predicted by ConPred_elite from 87 prokaryotic and 12 eukaryotic proteomes.

categories	ORFs	TM proteins	sequences predicted by ConPred_elite	functional identification		
				known[a] (%)	putative[b] (%)	unknown[c] (%)
prokaryotes	239,359	61,221	12,345	367 (3.0)	1,830 (14.8)	10,148 (82.2)
eukaryotes	214,593	55,799	7,404	682 (9.2)	1,129 (15.2)	5,593 (75.5)

[a] E-value $< 10^{-5}$ (by BLAST search against SWISS-PROT release 41.00 (Boeckmann *et al.*, 2003)) and sequence identity $\geq 95\%$ (by ALIGN).

[b] E-value $< 10^{-5}$ and $30\% \leq$ sequence identity $< 95\%$.

[c] E-value $\geq 10^{-5}$, or E-value $< 10^{-5}$ and sequence identity $< 30\%$.

References

[1] Arai, M., Ikeda, M., and Shimizu, T., Comprehensive analysis of transmembrane topologies in prokaryotic genomes, *Gene*, 304:77-86, 2003.

[2] Hirokawa, T., Boon-Chieng, S., and Mitaku, S., SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, 14(4):378-379, 1998.

[3] Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T., Transmembrane prediction methods: a reassessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, *In Silico Biol.*, 2(1):19-33, 2002.

[4] Ikeda, M., Arai, M., and Shimizu, T., TMPDB: a database of experimentally-characterized transmembrane topologies, *Nucleic Acids Res.*, 31(1):406-409, 2003.

[5] Lao, D.M. and Shimizu, T., Methods for detecting the signal peptide in transmembrane and globular proteins, *Genome Informatics*, 12:340-342, 2001.

[6] Sugiyama, Y., Natalia, P., and Shimizu, T., Identification of transmembrane protein functions by binary topology patterns, *Protein Eng.*, 16(7):479-488, 2003.