Evaluation of Transmembrane Topology Prediction Methods by Using an Experimentally Characterized Topology Dataset

Masami Ikeda[1] gs99601@si.hirosaki-u.ac.jp

Masafumi Arai[2] si9701@si.hirosaki-u.ac.jp

Toshio Shimizu[1] slsimi@si.hirosaki-u.ac.jp

[1] Department of Information Science, Graduate School of Science, Hirosaki University, Hirosaki 036-8561, Japan

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

Keywords: transmembrane protein, transmembrane topology, topology prediction method, performance evaluation

1 Introduction

Transmembrane (TM) proteins have an extreme importance in biomedical field, and are known to share nearly 30% of genes in whole genomes [2, 4, 5]. TM protein function is considered to be identified from TM-topology: the number of transmembrane segments (TMSs), the position of TMS and the orientation of TMS to the membrane lipid bilayer. For this reason, high-performance prediction methods of TM-topology from amino acid sequence are becoming increasingly needed. Many of TM-topology prediction methods have been proposed already such as KKD [Klein, P. *et al.*, 1985], TopPred 2 [Claros, M. G. and von Heijne, G., 1994], TMpred [Rost, B. *et al.*, 1996], DAS [Cserzo, M. *et al.*, 1997], TMAP [Persson, B and Argos, P., 1997], MEMSAT 2 [Jones, D. T., 1998], SOSUI [Hirokawa, T. *et al.*, 1998], HMMTOP [Tusnady, G. E. and Simon, I., 1998], PRED-TMR [Pasquier, C. *et al.*, 1999] etc. In this study, we evaluated the prediction accuracy of these 9 methods using TM-topology data that we collected from papers reporting experimentally determined topology data of TM proteins.

2 Dataset and Methods

We have collected 794 references reporting TM-topology of proteins so far, and are continuing our efforts to increase this number. From these, we extracted 122 topology models that are experimentally determined, e.g. by gene fusion, which are all alpha helical TM proteins and have less than 30% sequence similarity each other [1, 3] (Fig. 2). Using this database as a test dataset, we evaluated the 9 TM-topology prediction methods mentioned above. In this dataset, we removed signal peptide sequences from the entries having signal peptides to avoid the possibility that they are predicted as 1st TMSs.

We sent the sequences in the test dataset to the respective prediction method's server on the World Wide Web in order to evaluate the number of TMSs, the position of TMS, and its orientation in the predicted results. In our evaluation, when the distance between the central positions of corresponding TMSs of dataset and predicted result is within 9 residues, we classify it the correct prediction with respect to TMS position.

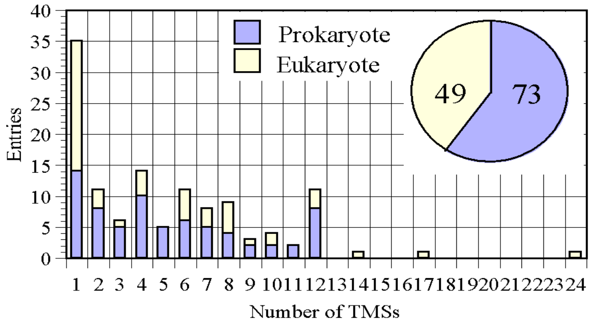

Figure 1. Transmembrane topology dataset

3 Results and Discussions

The evaluation results are summarized in Table 1. The 9 prediction methods predict the number of TMSs with the highest accuracy of 65.6% by HMMTOP and with the lowest of 38.5% by DAS. Moreover, the number of TMSs and position are correctly predicted at highest accuracy of 58.6% by MEMSAT 2 and at a lowest of 32% by DAS. In the case of the number of TMSs, position and orientation, the highest accuracy obtained is 46.7% by HMMTOP. For this case, only five methods except KKD, DAS, SOSUI and PRED-TMR are able to predict the sequence's orientation. We note that MEMSAT 2 could not perform the prediction for 6 entries. These results mean the topology prediction methods at this moment are not reliable enough to identify the functions of TM proteins from the sequences with such low prediction performances.

We are now trying to improve the prediction accuracies of the number of TMSs and the position by the optimal combination of the methods through the majority decision by using our topology dataset. Here, MEMSAT 2 and DAS should be excluded, since MEMSAT 2 sometimes results in prediction inability and DAS has the worst performance. By using this integrated method, we will expectedly be able to determine more precise preferences of TMS number and functions of TM proteins in various proteomes.

Table 1: Evaluation results.

prediction method	prediction accuracy (%) in				
	#TMS	#TMS & position	orientation	#TMS & orientation	#TMS & position & orientation
KKD	54.1	49.2	-	-	-
TopPred 2	59.0	50.8	72.1	49.2	41.0
TMpred	50.0	44.3	58.2	35.2	32.0
DAS	38.5	32.0	-	-	-
TMAP	54.9	43.4	54.1	34.4	27.0
MEMSAT 2	64.7*	58.6*	71.6*	48.4*	44.8*
SOSUI	56.6	50.8	-	-	-
HMMTOP	65.6	57.4	70.5	50.9	46.7
PRED-TMR	54.1	50.8	-	-	-

* MEMSAT 2 could not predict topology for 6 entries in the dataset.

References

[1] Kihara, D., Shimizu, T., and Kanehisa, M., Prediction of membrane proteins based on classification of transmembrane segments, *Protein Engineering*, 11(11):961–970, 1998.

[2] Mitaku, S., Ono, M., Hirokawa, T., Boon-Chieng, S., and Sonoyama, M., Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system, *Biophys. Chem.*, 82(2-3):165–171, 1999.

[3] Shimizu, T. and Nakai, K., Construction of a membrane protein database and an evaluation of several prediction methods of transmembrane segments, *Proc. Genome Informatics Workshop 1994*, 148–149, 1994.

[4] Stevens, T. J. and Arkin, I. T., Do more complex organisms have a greater proportion of membrane proteins in their genomes?, *Proteins*, 39(4):417–420, 2000.

[5] Wallin, E and von Heijne, G., Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Science*, 7(4):1029–1038, 1998.