

Comprehensive Analysis of Transmembrane Protein Sequences in 39 Microbial Genomes

Masafumi Arai¹ Keisuke Noto¹ Demelo Madrazo Lao¹
gs01603@si.hirosaki-u.ac.jp s498046@si.hirosaki-u.ac.jp gs00620@si.hirosaki-u.ac.jp

Masami Ikeda² Toshio Shimizu¹
srikeda@cc.hirosaki-u.ac.jp slsimi@si.hirosaki-u.ac.jp

- ¹ Department of Electronic Information System Engineering, Faculty of Science and Technology, Hirosaki University, 3, Bunkyo-cho, Hirosaki 036-8561, Japan
² Department of Science of Bioresources, The United Graduate School of Agricultural Sciences, Iwate University, 18-1, Ueda 3-chome, Morioka 020-8550, Japan

Keywords: transmembrane protein, transmembrane topology, signal peptide, genome-wide analysis, consensus transmembrane topology prediction, prokaryotic genome

1 Introduction

The genome-wide analysis of transmembrane (TM) proteins (the TM protein proportion in proteome, the distribution of TM topology, comprehensive functional identification of TM proteins, etc.) has been tried recently by using TM topology prediction methods. The performance of these prediction methods is, however, not so high enough, as confirmed by our re-assessment of the prediction performance by using a dataset of experimentally-characterized TM topologies (202 entries): HMMTOP 2.0 [8] is the highest with only 57.3% accuracy in predicting TM topology for prokaryotic sequences [3, 4]. In these analyses, the straight treatment of signal peptides (SPs) is usually avoided: e.g., the sequences with probable SP are excluded from the analysis [5]. This prevents us from estimating accurately the TM protein proportions in proteomes. And also, soluble protein sequences with SP are easily predicted as single-spanning TM protein.

In this study, we carried out genome-wide analysis of prokaryotic TM proteins (39 genomes) by using the “consensus TM topology prediction method” [3, 4] (67.7% accuracy with TM topology prediction) and by treating the SPs properly [6, 7].

2 Dataset and Methods

We used ORFs of 39 microbial genomes downloaded from GenBank [1]: proteobacteria (12 species), gram-positive bacteria (11), archaea (10) and others (6).

We first discriminated TM and soluble protein sequences for individual proteomes by SOSUI [2]. After applying the SP detection method (92.7% accuracy) [6, 7] to remove the SP region from the sequence, we predicted the TM topology of TM proteins by the consensus prediction method [3, 4].

3 Results and Discussion

TM protein proportions in proteomes obtained in our study are around 22%, e.g., 21.8% for *E. coli*, which is considerably lower than the previous reports, i.e., 25–30%. And, the distribution of the number of TM segments (TMSs) in our results is largely different from previous ones, in the range of 1–6 TMSs, in particular. This may be ascribed not only to the higher accuracy of the topology prediction method but also to the appropriate treatment of SPs used in our study. And, we found out that the secretory protein proportions in proteomes are largely different from category to category: proteobacteria (17.6%), gram-positive bacteria (12.6%), archaea (7.9%) and others (16.0%), as shown in Table 1. It is interesting that archaea have remarkably lower proportions of secretory proteins in

Table 1: Averaged proportions of TM proteins in proteomes and of proteins with SP.

category	TM protein proportion in proteome	proportion of proteins with SP		
		proteome	TM	soluble
proteobacteria	21.4	17.3	16.3	17.6
gram-positive bacteria	22.7	13.3	15.4	12.6
archaea	21.6	9.8	16.8	7.9
others	21.3	16.1	16.3	16.0
total	21.8	14.3	16.2	13.7

their proteomes. It is also notable that the TM proteins with odd number of TMSs show a higher tendency to have SP.

Much higher fraction of TM proteins of type I (with SP, N-out) is having a long N-tail (≈60 residues) comparing to type II (without SP, N-in) and type III (without SP, N-out). From this observation, it could be concluded that the existence of SP in TM protein sequence is helping the protein with having a long N_{out}-tail in type I.

As is reported elsewhere [9], the averaged overall sequence length increases proportionally to the number of TMSs, by 31 residues in our results. This linear increase is observed for only TM proteins (≠3-TMS). While, both single- and double-spanning TM proteins have the same level of averaged overall sequence length as one of the soluble proteins. This length of 31 residues corresponds to one TMS (ca. 21 residues) with adjoining loop regions (5 residues, both sides), which is considered to be a constructing module. This phenomenon suggests the possibility that the multi-spanning (≠4-TMS) TM topologies have been evolved by the “duplication” of the constructing module of 31 residues comprising a TMS. TM proteins of 1- and 2-TMS might be evolved by another way, i.e., the occurrence of new TMSs by mutation.

References

- [1] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.J., GenBank, *Nucleic Acids Res.*, 28(1):15–18, 2000.
- [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., Okuno, T., and Shimizu, T., The prediction accuracy of transmembrane topology is improved by a consensus method: An application to genome-wide analysis, *4th International Conference on Biological Physics*, 60, 2001.
- [4] Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T., Transmembrane topology prediction methods: A re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topology, *In Silico Biol.*, 2001, *in press*.
- [5] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. Mol. Biol.*, 305(3):567–580, 2001.
- [6] Lao, D.M. and Shimizu, T., A method for discriminating a signal peptide and a putative 1st transmembrane segment, *Proc. the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences - METMBS '01*, CSREA Press, 119–125, 2001.
- [7] Lao, D.M., Arai, M., Ikeda, M., and Shimizu, T., The presence of signal peptide significantly affects transmembrane topology prediction, *Bioinformatics*, *to be accepted*.
- [8] Tusnady, G. E. and Simon, I., The HMMTOP transmembrane topology prediction server, *Bioinformatics*, 17, 2001, *in press*.
- [9] Wallin, E. and von Heijne, G., Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Sci.*, 7(4):1029–1038, 1998.