

# Comprehensive Functional Identification of Prokaryotic Transmembrane Proteins by Binary Topology Pattern

Yoshiaki Sugiyama

gs01610@si.hirosaki-u.ac.jp

Masafumi Arai

gs01603@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, 3, Bunkyo-cho, Hirosaki 036-8561, Japan.

**Keywords:** transmembrane protein, transmembrane topology, binary topology pattern, functional identification, genome-wide analysis

## 1 Introduction

The functions of more than one half of proteins in proteome are not annotated yet. The functions of transmembrane (TM) protein, which corresponds to one fourth in a proteome, is known only a little, because of difficulty in determining TM protein structure experimentally. Accordingly, a lot of efforts have been made in an attempt to predict TM topology which is considered to correspond to the fold in the case of globular protein. Because, it is known that TM protein function can be identified by its TM topology, at least roughly.

We have developed a TM protein function identification method using the binary topology pattern based on the number of segments and loop length [7]. The topology pattern is expressed as a sequence of “1”, “0” and “\*”: “1” and “0” mean the long and short loop based on a defined threshold length, respectively, and “\*” means the binary loop length is not defined. The topology pattern of a query TM protein is associated to a particular function when it corresponds to the topology pattern of a TM protein having a known function.

In previous work, a common threshold length was used for all the loops. This time, we defined different threshold lengths for individual loops to improve identification accuracy [8]. By using this method, we identified comprehensively functions of putative TM proteins encoded within 39 microbial genomes, classifying the TM proteins into defined functional groups. We also compared the accuracy of our method in functional identification with one by BLAST.

## 2 Materials and Methods

We focused on TM proteins with up to 12-tms except 1- and 9-tms. The functional datasets of TM proteins were obtained from SWISS-PROT 38.0, by classifying the intact (not fragmental) sequences with defined TM topology into several functional groups according to the description in the DE, CC and KW lines. For example, the functions of 12-tms TM proteins were classified into 5 groups: “sodium transporter”, “sugar transporter”, “multidrug transporter”, “other transporters” and “others”. Then, we determined a topology pattern for each functional group using these datasets.

ORFs of the 39 complete microbial genomes were obtained from Genbank [1]. We first identified TM protein sequences of individual proteomes by SOSUI [2]. After applying the signal peptide detection method (92.7% accuracy) [5, 6] to remove the signal peptide region from the sequence, The TM topology was predicted by the consensus prediction method (67.7% accuracy with TM topology prediction) [3, 4]. Finally, we carried out the comprehensive functional identification of putative TM proteins by using the binary topology pattern.

## 3 Results and Discussion

We show here the results only for the case of 12-tms TM proteins. The binary topology patterns determined for the functional groups, “sodium transporter”, “sugar transporter”, “multidrug transporter”,

“other transporters” and “others” are (\*, 0, 1, 1, \*, \*, 0, 1, 1, \*, 1, 1, \*), (\*, \*, 0, 0, \*, 0, 0, 0, 0, \*, \*, \*, \*), (\*, \*, \*, \*, 1, \*, \*, 0, 1, \*, 1, \*, \*), (\*, \*, \*, \*, \*, \*, \*, \*, \*, \*, \*, \*, 0) and (\*, \*, 0, \*, \*, \*, \*, 0, 1, 1, 0, \*, 1), respectively, with the threshold lengths of (39, 12, 21, 13, 6, 28, 140, 29, 18, 13, 16, 14, 87). And, the accuracies of functional identification are 0.991, 0.846, 0.943, 0.533 and 0.781, respectively. The results of functional identification are summarized in Table 1 for 4 prokaryotic organisms. More than 90 % of the protein sequences are classified into 5 functional groups by our method.

Next, we searched SWISS-PROT 38 for the sequence most similar to a putative TM protein sequence to identify its function, by using BLAST (E-value,  $1.0 \times 10^{-5}$ ). As shown in Table 2, almost all the sequences were assigned the functions, in this case too. More than one half of the assigned functions, however, are hypothetical or putative ones. It should be noted that there are some discrepancies in functional identification between the two methods, e.g., with the number of sequences in “sugar transporter” and “others”. This is remaining for the future work to clarify.

Table 1: Functional identification of 12-tms TM proteins in 4 organisms.

	total	sodium transporter	sugar transporter	multidrug transporter	other transporters	others	not identified
<i>E. coli</i>	95.70 (89 / 93)	0.00 (0)	43.01(40)	9.68 (9)	43.01 (40)	0.00 (0)	4.30 (4)
<i>P. aeruginosa</i>	92.44 (110 / 119)	0.84 (1)	44.54 (53)	14.29 (17)	32.77 (39)	0.00 (0)	7.56 (9)
<i>B. subtilis</i>	95.83 (69 / 72)	0.00 (0)	43.06 (31)	13.89 (10)	38.89 (28)	0.00 (0)	4.17 (3)
<i>B. halodurans</i>	96.23 (51 / 53)	0.00 (0)	37.74 (20)	9.43 (5)	49.06 (26)	0.00 (0)	3.77 (2)

Table 2: Functional identification by BLAST

	total	sodium transporter	sugar transporter	multidrug transporter	other transporters	others	not identified
<i>E. coli</i>	100.00 (93 / 93) (54 / 93)*	0.00 (0) (0)*	16.13 (15) (5)*	1.08 (1) (0)*	59.14 (55) (30)*	23.66 (22) (19)*	0.00 (0)
<i>P. aeruginosa</i>	97.48 (116 / 119) (59 / 116)*	3.36 (4) (1)*	9.24 (11) (4)*	1.68 (2) (0)*	54.62 (65) (31)*	28.57 (34) (23)*	2.52 (3)
<i>B. subtilis</i>	95.83 (69 / 72) (34 / 69)*	8.33 (6) (2)*	9.74 (7) (3)*	4.17 (3) (0)*	45.83 (33) (13)*	27.78 (20) (16)*	4.17 (3)
<i>B. halodurans</i>	94.34 (50 / 53) (31 / 50)*	7.55 (4) (2)*	1.89 (1) (0)*	11.32 (6) (0)*	26.42 (14) (7)*	47.17 (25) (22)*	5.67 (3)

\*inside of parenthesis is the number of sequences of which function is hypothetical.

## References

- [1] Benson, D.A., Karsch-Mazrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.J., Genbank, *Nucleic Acids Res.*, 28:15–18, 2000.
- [2] Hirokawa, T., Boon-Chieng, S., and Mitaku, S., SOSUI: Classification and secondary structure prediction system for membrane protein, *Bioinformatics*, 14:378–379, 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] 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 Science - METMBS '01, CSREA Press*, 119–125, 2001.
- [6] Lao, D.M., Arai, M., Ikeda, M., and Shimizu, T., The presence of signal peptide significantly affects transmembrane topology prediction, *Bioinformatics*, *to be accepted*.
- [7] Natalia, P., Saito, K., and Shimizu, T., Transmembrane topology pattern and detection of transmembrane protein functions, *Genome Informatics* , 11:422–423, 2000.
- [8] Sugiyama, Y. and Shimizu, T., Detection of transmembrane protein function by a binary transmembrane topology pattern, *4th International Conference on Biological Physics*, 60, 2001.