# A comparison of transmembrane topologies greatly improves the comprehensive functional classification and identification of prokaryotic transmembrane proteins

**Masafumi Arai[1,2,*], Kosuke Okumura[1], Masanobu Satake[2,3], Toshio Shimizu[1]**

**Keywords:** transmembrane protein, transmembrane topology, functional classification and identification, prokaryotic species, genome-wide analysis

## 1 Introduction.

Many of proteins have not yet been annotated, with about one half of all proteome sequences being classified as functionally "putative" or "unknown" at best [1]. Such is the case, in particular, for transmembrane (TM) proteins, which account for as much as 20-30% of proteomes in individual species [2]. This is partly because TM protein sequences of known function are much less compared with soluble proteins. Recent studies, however, revealed that TM protein functions are closely correlated to their TM topologies, i.e., the number of TM segments (TMSs), positions of TMSs and N-tail location [3]. In this study, we propose a new method for the comprehensive classification and identification of TM protein functions by a clustering approach based on TM topology similarity. Prior to performing the clustering, we first investigate the current status of the functional identification of TM proteins based on sequence similarity.

## 2 Materials and Methods.

Out of 239,359 protein sequences of 87 sequenced prokaryotic (72 bacterial and 15 archaean) species in the GenBank database, 51,044 sequences were extracted as TM protein and their TM topologies (1-12 TMSs) (~21%) were predicted, by using SOSUI [4] (TM protein sequence prediction, ≥98% accuracy), DetecSig (signal peptide prediction and removal, 88% accuracy) [5] and ConPred (TM topology prediction, 69.6% and 83.3% accuracies for the number of TMSs & TMS positions and N-tail location, respectively) [6]. The procedures and the genome-wide analysis of TM topologies are described in detail in our previous paper [2].

The obtained TM protein sequences were classified into three categories, i.e., "known", "putative" and "unknown", according to the level of functional annotations in the SWISS-PROT database by homology search and sequence similarity comparison (details not shown here). Then, these annotated sequences were clustered by the single-linkage method based on TM topology similarity between sequences with the same number of TMSs. The TM topology similarity between sequences 1 and 2, $S_{1,2}$ is calculated as:

$$S_{1,2}\,(\%) = 100 \times \sum_{i=1}^{n+1} \min(l_{1,i},\, l_{2,i}) / \sum_{i=1}^{n+1} \max(l_{1,i},\, l_{2,i}),$$

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

[2] Department of Developmental Biology and Neuroscience, Graduate School of Life Sciences, Tohoku University, Sendai 980-8577, Japan

[3] Department of Molecular Immunology, Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan

[*] E-mail address: d01603@si.hirosaki-u.ac.jp

where, $n$, $l_{1,i}$ and $l_{2,i}$ are the number of TMSs, the length of the $i$-th loop in sequences 1 and 2, respectively, and $\min(l_{1,i}, l_{2,i})$ and $\max(l_{1,i}, l_{2,i})$ are the lengths of the shorter and longer loops in $l_{1,i}$ and $l_{2,i}$, respectively. The thresholds of TM topology similarity were determined so that the sequences included in the representative clusters (with ≥10 sequences) would occupy over 50% out of all the sequences.

# 3 Results and Discussion.

Using our clustering approach, the functionally classified and identified TM proteome sequences was increased from 24.3% to 60.9%. Almost half of them used to be "unknown" sequences before applying the clustering method. Additional analysis of the TM topologies in the clusters provided important information regarding TM protein functions that cannot be ascertained from sequence similarity.

Table 1: The results of the functional classification and identification of TM proteins (1-12 TMSs) from the 87 prokaryotic species based on sequence similarity and TM topology similarity.

| | | Based on sequence similarity | | | | Based on TM topology similarity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TMSs | Total sequences | Functionally annotated sequences | | | Identified[1] | Threshold TM topology similarity | In the representative clusters (with ≥10 sequences) | | | | | Classified and identified[2] |
| | | "Known" | "Putative" | "Unknown" | | | Clusters | Sequences | Functionally annotated sequences | | | |
| | | | | | | | | | "Known" | "Putative" | "Unknown" | |
| 1 | 14,590 | 584 | 2,191 | 11,815 | 19.0% | 98% | 74 | 7,337 | 332 | 1,295 | 5,710 | 58.2% |
| 2 | 6,928 | 229 | 785 | 5,914 | 14.6% | 92% | 46 | 3,660 | 157 | 534 | 2,969 | 57.5% |
| 3 | 4,059 | 105 | 602 | 3,352 | 17.4% | 85% | 32 | 2,281 | 75 | 426 | 1,780 | 61.3% |
| 4 | 4,493 | 130 | 813 | 3,550 | 21.0% | 84% | 41 | 2,515 | 97 | 561 | 1,857 | 62.3% |
| 5 | 3,643 | 131 | 923 | 2,589 | 28.9% | 81% | 33 | 1,923 | 76 | 625 | 1,222 | 62.5% |
| 6 | 4,628 | 180 | 1,411 | 3,037 | 34.4% | 85% | 27 | 2,464 | 108 | 1,024 | 1,332 | 63.2% |
| 7 | 2,076 | 82 | 515 | 1,479 | 28.8% | 75% | 25 | 1,075 | 44 | 330 | 701 | 62.5% |
| 8 | 1,965 | 82 | 572 | 1,311 | 33.3% | 73% | 26 | 1,037 | 52 | 398 | 587 | 63.2% |
| 9 | 2,015 | 100 | 704 | 1,211 | 39.9% | 74% | 30 | 1,033 | 67 | 501 | 465 | 63.0% |
| 10 | 2,061 | 89 | 525 | 1,447 | 29.8% | 74% | 31 | 1,090 | 42 | 293 | 755 | 66.4% |
| 11 | 2,045 | 94 | 625 | 1,326 | 35.2% | 75% | 23 | 1,087 | 62 | 400 | 625 | 65.7% |
| 12 | 2,541 | 132 | 794 | 1,615 | 36.4% | 82% | 22 | 1,286 | 80 | 499 | 707 | 64.3% |
| Total | 51,044 | 1,938 | 10,460 | 38,646 | 24.3% | - | 410 | 26,788 | 1,192 | 6,886 | 18,710 | 60.9% |

[1] "Known" and "putative" sequences are counted.

[2] Sequences in the representative clusters and "known" and "putative" sequences in other clusters (with 1-9 sequences) are included.

# References

[1] Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T. and Riley, M. 2001. A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* 2:RESEARCH0035.

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

[3] Sugiyama, Y., Polulyakh, N. and Shimizu, T. 2003. Identification of transmembrane protein functions by binary topology patterns. *Protein Eng.* 16:479-488.

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

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

[6] Lao, D.M. and Shimizu, T. 2001. Methods for detecting the signal peptide in transmembrane and globular proteins. In: Matsuda, H., Miyano, S., Takagi, T. and Wong, L., editors, *Proceedings of the Twelfth International Conference on Genome Informatics (GIW 2001)*, Universal Academy Press, Tokyo. pp. 340-342.