Genome-Wide Functional Classification/Identification of Prokaryotic Transmembrane Proteins Based on Transmembrane Topology Similarity

Masafumi Arai[1,2]
d01603@si.hirosaki-u.ac.jp

Kosuke Okumura[1]
gs03406@si.hirosaki-u.ac.jp

Masanobu Satake[2,3]
satake@idac.tohoku.ac.jp

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

[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

Keywords: transmembrane protein, transmembrane topology, functional classification/identification, prokaryotic genome, genome-wide analysis

1 Introduction

Functions of transmembrane (TM) proteins holding 20-30% [1] in individual proteomes are difficult to be inferred comprehensively based on only sequence similarity, which is partly because TM protein sequences of known function are lacking. 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 the N-tail location [5]. In this study, we propose a new method for the comprehensive classification/identification of TM protein functions by single-linkage clustering based on TM topology similarity. Applying this method to TM protein sequences predicted from 87 sequenced prokaryotic genomes, it was realized that TM topology information improved functional classification/identification greatly.

2 Materials and Methods

In total 239,359 open reading frames (ORFs) of 87 sequenced prokaryotic (72 bacterial and 15 archaean) genomes were downloaded from the GenBank database. Out of these ORFs, 51,044 sequences were extracted as TM protein and their TM topologies (1-12 TMSs) ($\sim$21%) were predicted, by using SOSUI [2] (TM protein sequence prediction, $\geq$98% accuracy), DetecSig (signal peptide prediction and removal, 88% accuracy) [4] and ConPred (TM topology prediction, 69.6% and 83.3% accuracies for the attributes on the number of plus positions of TMSs and the N-tail location, respectively) [3]. The procedures and the genome-wide analysis of TM topologies are described in detail in our previous paper [1].

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 the BLAST search and ALIGN alignment (details not shown here). These annotated sequences were clustered by the single-linkage method based on TM topology similarity, separately for the number of

TMSs, N_{in}/N_{out} and +/− signal peptide. TM topology similarity between sequences 1 and 2, $S_{1,2}$ is calculated as:

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

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 larger clusters (with ≥10 sequences) would occupy over 50% out of all the sequences.

3 Results and Discussion

More than 60% of TM protein functions were classified/identified by using TM topology similarity, while the rate based on sequence similarity is less than one forth (on average). This result indicates that TM topology information is remarkably more effective than sequence similarity in the classification/identification of TM protein functions.

Table 1: Results of the comprehensive functional classification/identification of TM proteins for 87 prokaryotic genomes based on sequence similarity and TM topology similarity.

TMSs	Total sequences	Based on sequence similarity				Based on TM topology similarity					
		"Known"	"Putative"	"Unknown"	Classified/ identified[1] (%)	TM topology similarity (%)	In the larger clusters (with =10 sequences)				Classified/ identified[2] (%)
							Total clusters	Total 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 included.
[2] "Known" and "putative" sequences in the smaller clusters with 1-9 sequences are also included (the figures are not shown here explicitly).

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 topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, *In Silico Biol.*, 2(1):19–33, 2002.

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

[5] Sugiyama, Y., Polulyakh, N., and Shimizu, T., Identification of transmembrane protein functions by binary topology patterns, *Protein Eng.*, 16(7):479–488, 2003.