

# Domain Architecture of Transmembrane Proteins: Focusing on Single-Spanning Proteins First

Masafumi Arai<sup>1,2,\*</sup>, Takafumi Fukushi<sup>1</sup>, Masanobu Satake<sup>2,3</sup>, Toshio Shimizu<sup>1</sup>

**Keywords:** single-spanning transmembrane protein, domain architecture, transmembrane topology, proteome-scale analysis, prokaryotic genome

## 1 Introduction.

Single-spanning transmembrane (TM) proteins could be regarded as soluble proteins rooted in the membrane lipid bilayer. The two tail loops, i.e., N-tail and C-tail loops, are mostly long enough, and each polypeptide chain, protruding to either side of the membrane, could behave freely with its open end. Naturally, these loops are expected to have structural, functional and evolutionary domains on them, many of which should be shared with soluble proteins, with some others specific to TM proteins only. The domain assignment in proteome-scale has shed a new light on structural, functional and evolutionary makeups of proteins. This is, however, limited only to soluble proteins, not for TM proteins including single-spanning proteins, unfortunately, although it is eagerly desired also for TM proteins.

In this study, we attempted to carry out a comprehensive analysis of domain architectures in the single-spanning TM proteins by assigning Pfam domains [1] to the prokaryotic (Bacterial and Archaeal) single-spanning sequences predicted from 87 sequenced genomes. We also analyzed three-dimensional structures of full-length multi-spanning TM proteins to elucidate the domain architecture of them.

## 2 Material and Methods.

A total 14,625 single-spanning TM proteins predicted from 87 sequenced prokaryotic (72 Bacterial and 15 Archaeal) proteomes were used in this study. These single-spanning sequences were subjected to the domain assignment by using Pfam-A families (in total 5,724 families) registered in Pfam 9.0 [1] and the hmmpfam program included in HMMER 2.3.2. The criterion for acceptable assignment was set to the E-value less than  $10^{-3}$ . The sequences of which TM segment (TMS) region is overlapped with assigned domains by over five residues were discarded not as targets for further analyses, and with assigned domains overlapped one another by even single residue were also thrown out. After this domain assignment procedure, a total of 10,966 single-spanning TM proteins which contain 7,850 (71.6%), 2,394 (21.8%) and 722 (6.6%) sequences with no, one and two assigned domains, respectively, were remained.

We also extracted 47 intact three-dimensional multi-spanning TM protein structures with a full-length sequence from PDB [2] (detailed not shown here), and used them for the domain architecture analysis.

---

<sup>1</sup> Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan

<sup>2</sup> Department of Developmental Biology and Neuroscience, Graduate School of Life Sciences, Tohoku University, Sendai 980-8577, Japan

<sup>3</sup> 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

### 3 Results and Discussion.

Depending on the distribution of assigned domains on the tail loops, the domain arrangement of single-spanning TM proteins can be classified into six “modes”: i.e., “NonE”, with no assigned domains; “DonN”, one domain on the N-tail loop; “DonC”, one domain on the C-tail loop; “DDonN”, two domains on the N-tail loop; “DDonNC”, one domain on the N-tail and C-tail loops each; “DDonC”, two domains on the C-tail loop.

Table 1 represents the distribution of the sequences in the NonE, DonN+DDonN, DDonNC and DonC+DDonC modes over four divisions according to N-tail and C-tail loop lengths with the threshold length of 60 residues (aa). The DonC+DDonC sequences with the N-tail and C-tail loop lengths of  $\geq 60$  and  $< 60$  aa, respectively, are predominant, i.e., 87.9% in the total DonC+DDonC ones. In contrast, those with the C-tail loop length of  $< 60$  aa are only 13 sequences (0.5%). We note that the NonE sequences with the N-tail and/or C-tail loop lengths of  $\geq 60$  aa occupy nearly three fourth of all the None ones. This is indicating we may have still a large number of functionally unknown domains on the tail loops of single-spanning TM proteins that haven’t defined yet in Pfam.

Domain arrangement mode	Sequences in individual divisions (%)				Total
	N-tail $\geq 60$ aa C-tail $\geq 60$ aa	N-tail $\geq 60$ aa C-tail $< 60$ aa	N-tail $< 60$ aa C-tail $\geq 60$ aa	N-tail $< 60$ aa C-tail $< 60$ aa	
NonE	691 (8.8)	1,152 (14.7)	3,966 (50.5)	2,041 (26.0)	7,850 (100)
DonN+DDonN	164 (31.2)	354 (67.4)	2 (0.4)	5 (1.0)	525 (100)
DDonNC	48 (98.0)	1 (2.0)	0 (0)	0 (0)	49 (100)
DonC+DDonC	295 (11.6)	1 (0)	2,234 (87.9)	12 (0.5)	2,542 (100)
Total	1,198	1,508	6,202	2,058	10,966

Table 1: Distribution of the sequences in the NonE, DonN+DDonN, DDonNC and DonC+DDonC modes over the four divisions (see in text).

Looking into the 47 multi-spanning TM proteins extracted from PDB, we can easily discover the characteristic domain architecture of them, which are constructed from a TM domain and soluble domains carried on long loops (both of tail and connecting) longer than 60 residues in most cases. The TM domain is formed from a recombination of a few bundles of consecutive TMSs connected by short loops (named “TM module”), basically less than 30 residues long. We defined 41 single-spanning and 67 multi-spanning TM modules from the 47 multi-spanning TM protein structures. Among the 67 multi-spanning TM modules, we have 58 modules (87%) in which all the TMSs are contacted with some other TMSs within 8 angstrom distance, i.e., 44 “fully contacted” and 14 “quasi-fully contacted”. The remaining 9 modules (13%) are not considered as single compacted structural units, being separated into two or more independent smaller sub-structures. These domain structures could be detected in multi-spanning TM proteins of unknown structure, provided accurately predicted TM topology data (the number and positions of TMSs, and N-tail location) are available.

### References

[1] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R. 2004. The Pfam protein families database. *Nucleic Acids Res.* 32:D138-D141.

[2] Westbrook, J., Feng, Z., Chen, L., Yang, H. and Berman, H.M. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31:489-491.