Evolutionary Use of Domain Recombination: A Distinction Between Membrane and Soluble Proteins

Yang Liu, Mark Gerstein, Donald M. Engelman

Department of Molecular Biophysics and Biochemistry, Yale University,

P.O. Box 208114, New Haven, CT, 06520-8114, USA

To whom reprint requests and page proofs should be addressed:

Donald M. Engelman

Department of Molecular Biophysics and Biochemistry, Yale University,

P.O. Box 208114, New Haven, CT, 06520-8114, USA

Phone: *203 432 5601;* Fax: *203 432 6381*;

E-mail: *Donald.Engelman@yale.edu*

Abstract

Soluble proteins often contain multiple structural domains, which are shuffled or recombined to gain new functions in the course of evolution. We examined integral membrane proteins for evidence of this mechanism using a classification of polytopic transmembrane domains. Surprisingly, in contrast to the situation in aqueous solution, we found that recombination of structural domains is not common inside membranes: the majority of integral membrane proteins contain only a single polytopic membrane domain. We suggest that non-covalent oligomeric associations, which are common in membrane proteins, may provide an alternative source of evolutionary diversity in this class of proteins.

Introduction

Protein domains are often mixed to facilitate evolution, usually by recombination events that place them in single polypeptides (1-4). Proteomes from archae, prokarya, and eukarya were studied using a structure-based classification (SCOP) (5), and it was found that a large majority of domains (approximately 65% in prokarya and approximately 80% in eukarya) are combined with other domains (6). Thus, evolution appears to use recombination of domains to generate new protein structures and functions. However, the structural database is overwhelmingly biased in favor of soluble proteins, raising the question of whether the process of domain recombination is also used inside membranes. Using our classification of polytopic trans-membrane domains into ~650 families (7), we examined 26 proteomes, and found that mixed domain proteins are much less abundant inside membranes than in the aqueous regions of a

cell. We argue that the constraints of the membrane environment, which have been previously noted (8, 9), favor oligomerization, so that covalent links may not be required for domains to recombine to gain new functions during evolution.

Results

Using sequence data from 26 genomes (8 in archaea, 14 in prokarya, and 4 in eukarya), membrane domains with two or more putative transmembrane helices (10) were classified into families. Here, we use the word “domain” to designate a protein with more than one transmembrane helix, in distinction from the occasional use of the word to note the independent stability of single helices (9, 11). Popot (personal communication) has suggested that a useful distinction can be made between folding domains, which might be single helices, and functional domains, which would usually require multiple helices. Our classification was based in part on the Pfam assignments (12) and in part on clustering by sequence similarities (7). Most (95%) polytopic membrane domains defined in the families have relatively short loops (<80 residues) between transmembrane helices. To be counted as a polytopic domain family, at least four members must be present. Approximately 650 families were identified, corresponding to approximately 61% of all predicted integral membrane protein domains.

Because they are the best defined, we chose to examine the cases in the Pfam-A classified families (see Fig. 1A), and found that most integral membrane proteins (~78% for archaea and prokarya; ~67% for eukarya) contain only a single classified membrane domain. It follows that the level of transmembrane domain recombination in membrane proteins is less than 33%. Thus, membrane proteins do not exploit domain recombination to such a large extent as soluble proteins do. The relative paucity of domain combinations within integral membrane proteins might be understood as arising

from the two-dimensional structure of the phospholipid bilayer, which facilitates domain interaction without covalent linkages. Membranes restrict volume, translational freedom, and rotational freedom of proteins so that the entropic penalty for oligomerization is reduced. It is notable that the known membrane protein structures overwhelmingly consist of homo- and hetero-oligomeric associations (13). Figure 2 shows a cross-section of cytochrome C oxidases (from bovine heart mitochondria) (14), photosynthetic reaction center (from Rhodopseudomonas virdis) (15), and cytochrome bc1 complexes (in bovine heart mitochondria) (16) at the midplane of the bilayer, revealing that the identity of individual subunits cannot be seen in the structure: inspection of the gray representation does not lead to the identities color-coded in the other view.

Further support for the idea that oligomers emerge as a consequence of the membrane environment can be found in “split protein” experiments, where polytopic membrane proteins expressed as fragments are observed to associate and function (see, e.g., (17); for review see (13)). The same kind of behavior has been documented *in vitro*, using fragments of proteins (18, 19) (20, 21). Separate evolution of subunits that associate has also been observed (22). That fragments can re-associate and function argues that the covalent linkage between them, while perhaps adding stability and/or control of expression, is not essential.

Of the membrane proteins containing more than one domain, many appear to have resulted from domain duplication, containing two or more identical Pfam-A domains (Fig. 1B) (see, e.g., (23)). Eukaryotes have a higher incidence (~16% on average) of

integral membrane proteins with two or more duplicated domains than do prokarya or archaea (~9%). Figure 1B lists the most commonly duplicated domains in integral membrane proteins in the genomes. An interesting observation is that the 7-TM chemoreceptors and 7-TM rhodopsin families have high occurrences (48 and 46) and most of them occur in *C. elegans* (48 and 32). Knowing that *C. elegans* has an exceptionally large number of 7-TM receptors and rhodopsin-like membrane proteins (7, 24), it may be that the duplications imply possible functional relations between homologous 7-TM domains. This observation is also supported by the idea that dimerization of G-protein-coupled receptors may be important for their functions (25).

By contrast with the paucity of covalent combinations of transmembrane domains, combinations between soluble domains and membrane domains are frequently observed. We analyzed the membrane proteins having only one membrane domain to see how many had flanking soluble domains (Fig. 1C). We found that archaea and prokarya have a much larger proportion (~34% and 24%, respectively) of single-domain membrane proteins without flanking soluble domains than eukarya (~7%). Consistent with a previous study of soluble proteins (6), this observation indicates that genetic recombination can happen for membrane protein genes in a similar fashion to soluble ones. That the membrane portions do not show such recombination with each other must then reflect different constraints.

Another similarity shared by membrane and soluble proteins is the distribution pattern of protein domain families in the three kingdoms (Figure 3). Based on previous analysis

(7), using just the Pfam-A families, we found that the 26 proteomes used in this study consist of 1922 soluble domain families and 214 polytopic membrane domain families. The soluble proteins have almost 10 times more families than integral membrane proteins, suggesting a higher diversity of structure for proteins when the membrane constraints are absent. On the other hand, the proportions of the common and unique families in the three kingdoms are similar between membrane and soluble proteins, implying that a similar evolutionary process is shared by these two kinds of proteins.

Discussion

Our survey of domain combinations in the helical, transmembrane parts of membrane proteins reveals that most have only one membrane domain. Either the required functional diversity is much less for membrane proteins or they may exploit a different strategy to attain diversity in evolution. The latter possibility is supported by the observation of a widespread occurrence of membrane protein oligomers, by studies of split membrane proteins, and by the argument that oligomer formation is facilitated by the constraints of the membrane bilayer. Since the same constraint would not apply if one of the domains were a soluble domain, it is reasonable to find that covalent links are frequently used between soluble protein domains and membrane domains. A challenge for future work will be to document the extent to which alternative oligomerization (the degree to which a given domain may participate in different oligomeric complexes) may provide an evolutionary mechanism.

Acknowledgement

MG thanks an NIH grant (GMS4160-07) for financial support. YL was supported by an NLM postdoctoral fellowship (T15 LM07056).

References

1. Chothia, C. (1992) *Nature* **357,** 543-544.
2. Doolittle, R. F. (1995) *Annu. Rev. Biochem.* **64,** 287-314.
3. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999) *Nature* **402,** 83-6.
4. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999) *Nature* **402,** 86-90.
5. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536-540.
6. Apic, G., Gough, J. & Teichmann, S. A. (2001) *J. Mol. Biol.***,** 311-325.
7. Liu, Y., Engelman, D. M. & Gerstein, M. (2002) *Genome Biol* **3,** research0054.
8. Engelman, D. M. & Steitz, T. A. (1981) *Cell* **23,** 411-22.
9. Popot, J. L., Engelman, D. M., Zaccai, G. & de Vitry, C. (1990) *Prog Clin Biol Res* **343,** 237-62.
10. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001) *J. Mol. Biol.* **305,** 567-580.
11. Popot, J. L. & Engelman, D. M. (1990) *Biochemistry* **29,** 4031-7.
12. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28,** 263-266.
13. Popot, J. L. & Engelman, D. M. (2000) *Annu. Rev. Biochem.* **69,** 881-922.

14. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996) *Science* **272,** 1136-44.

15. Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. (1995) *J Mol Biol* **246,** 429-57.

16. Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S. & Jap, B. K. (1998) *Science* **281,** 64-71.

17. Zen, K. H., McKenna, E., Bibi, E., Hardy, D. & Kaback, H. R. (1994) *Biochemistry* **33,** 8198-206.

18. Popot, J. L., Gerchman, S. E. & Engelman, D. M. (1987) *J Mol Biol* **198,** 655-76.

19. Liao, M. J., London, E. & Khorana, H. G. (1983) *J Biol Chem* **258,** 9949-55.

20. Kahn, T. W. & Engelman, D. M. (1992) *Biochemistry* **31,** 6144-51.

21. Marti, T. (1998) *J Biol Chem* **273,** 9312-22.

22. Claros, M. G., Perea, J., Shu, Y., Samatey, F. A., Popot, J. L. & Jacq, C. (1995) *Eur J Biochem* **228,** 762-71.

23. Pao, S. S., Paulsen, I. T. & Saier, M. H., Jr. (1998) *Microbiol Mol Biol Rev* **62,** 1-34.

24. Bargmann, C. (1998) *Science* **282,** 2028-2033.

25. Gomes, I., Jordan, B. A., Gupta, A., Rios, C., Trapaidze, N. & Devi, L. A. (2001) *J. Mol. Med.* **79,** 226-242.

Figures

Figure 1. Domain combination of polytopic membrane domains in genomes

(A) The green bars represent the percentage of classified membrane proteins by Pfam-A that consist of only one polytopic membrane domain, and the light green bars indicate the percentage of classified membrane proteins that consist of duplicated polytopic membrane domains. The archaea group includes *Archaeoglobus fulgidus*, *Aeropyrum pernix K1*, *Halobacterium sp.*, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Pyrococcus abyssi*, *Pyrococcus horikoshii*, and *Thermoplasma acidophilum*; the prokarya group includes *Aquifex aeolicus*, *Borrelia burgdorferi*, *Bacillus subtilis*, *Chlamydia pneumoniae* strain AR39, *Chlamydia trachomatis*, *Escherichia coli* strain K12, *Haemophilus influenzae*, *Helicobacter pylori* strain 26695, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Rickettsia prowazekii*, *Synechocystis sp*, and *Treponema pallidum*; and the eukarya group includes *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*. Notes on the assignment strategy: ~65% of the assigned membrane proteins had Pfam-A matches. Pfam-B and the clustered families were excluded, as they are not as carefully classified as Pfam-A families. Integral membrane proteins that contain only one classified membrane domain with no more than one extra TM-helix were considered to be single membrane domain proteins; otherwise, they were considered to be multiple membrane domain proteins. (The Pfam classification does not always consider TM-helix regions). The orange bars indicate the percentage of single domain soluble proteins based on the classification of Pfam-A, which can have up to 30 residues next to their Pfam-A domains.

(B) The table shows the Pfam-A membrane-protein families that occur most often in tandem duplicated fashion. It ranks the families by the number of sequences where they are found more than once in a given gene.

(C) The plot shows the percentage of classified single domain membrane proteins lacking a soluble domain. The single domains proteins have no more than 30 residues flanking regions next to the membrane domains.

Figure 2. Helix interactions in the membrane midplane

A 5-residue section is defined at the apparent center of the membrane lipid bilayer (inferred from the hydrophobic exterior) and helix positions are indicated. The grayscale image emphasizes that the subunits shown in the colored image cannot be inferred from helix relationships.

Figure 3. Protein domain families shared between the archaea, prokarya, and eukarya kingdoms

This figure shows the distributions of Pfam-A families in soluble and membrane proteins among the three kingdoms. The common families shared by the three kingdoms represent 24% for soluble proteins and 28% for membrane proteins; while the unique families represent 7%, 24%, and 41% for soluble proteins and 7%, 22% and 33% for membrane proteins in archaea, prokarya, and eukarya respectively.