

Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism

Yang Liu, Mark Gerstein, and Donald M. Engelman*

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520-8114

Contributed by Donald M. Engelman, December 29, 2003

Recombination of evolutionarily unrelated domains is a mechanism often used by evolution to produce variety in soluble proteins. By using a classification of polytopic transmembrane domains into families, we examined integral membrane proteins for evidence of this mechanism. Surprisingly, we found that domain recombination is not common for the transmembrane regions of membrane proteins, a majority of integral membrane proteins containing only a single transmembrane domain. We suggest that noncovalent oligomeric associations, which are common in membrane proteins, may provide an alternative source of evolutionary diversity.

Membrane proteins are known to evolve diversity of function by gene duplication followed by divergent evolution, and by recombination of homologous but different subunits. However, the strategy of fusing evolutionarily unrelated transmembrane domains has not been examined for membrane proteins, even though it is commonly observed in soluble proteins. Protein domains are considered to be structural units that can be mixed to facilitate evolution (1–4), either by recombination of homologous or of nonhomologous structures. Studies of domain recombination in proteomes from archae, prokarya, and eukarya by using the structure-based classification (5) shows that a large majority of domains ($\approx 65\%$ in prokarya and $\approx 80\%$ in eukarya) are combined with other domains (6). This result supports the idea that recombination of domains could generate new protein structures and functions in the course of evolution. However, the result applies principally to soluble proteins, as the structural database is strongly biased in favor of them. By using our classification of polytopic membrane proteins into ≈ 650 families (7), we have been able to study domain combinations in integral membrane proteins and find that they are much less abundant than in the soluble protein cases.

Methods and Results

Based on sequence similarities and the number of putative transmembrane helices (8), we classified families of polytopic membrane domains in 26 genomes (8 in archaea, 14 in prokarya, and 4 in eukarya). A family is required to have at least four members, each with at least two putative transmembrane helices. The classification was based in part on the Pfam assignments (9) and by clustering based on sequence similarities (7). Pfam is a large collection of families based on multiple sequence alignments and profile-hidden Markov models, leading to the classification of $>3,000$ families (Pfam release 6.1). Most (95%) polytopic membrane domains defined in the families have relatively short loops (<75 aa) between transmembrane helices. We identified >650 families, corresponding to $\approx 61\%$ of all predicted helical transmembrane proteins (7).

Because they are the best defined, we chose to examine the cases in the manually curated Pfam-A classified families (Fig. 1A), and we found that most integral membrane proteins ($\approx 78\%$ for archaea and prokarya and $\approx 67\%$ for eukarya) contain only a single classified membrane domain. It follows that the level of transmembrane domain recombination in membrane proteins is $<33\%$. Thus, membrane proteins do not exploit domain recombination to the large extent that soluble proteins do. The relative

paucity of domain combinations within integral membrane proteins might be understood as arising from the 2D 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 homooligomeric and heterooligomeric associations (10). Fig. 2 shows a cross section of cytochrome C oxidases (from bovine heart mitochondria) (11), photosynthetic reaction center (from *Rhodospseudomonas viridis*) (12), and cytochrome bc₁ complexes (in bovine heart mitochondria) (13) 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, in which polytopic membrane proteins expressed as fragments are observed to associate and function (see ref. 10 for review). The same kind of behavior has been documented *in vitro* by using fragments of proteins (14–17). That fragments can reassociate and function argues that the covalent linkage between them, although perhaps adding stability and/or control of expression, is not always essential.

Of the membrane proteins containing more than one domain, some seem to have resulted from domain duplication, containing two or more identical Pfam-A domains (Fig. 1B). Eukaryotes have a higher incidence ($\approx 16\%$ on average) of integral membrane proteins with two or more duplicated domains than do prokarya or archaea ($\approx 9\%$). Fig. 1B lists the most commonly duplicated domains in integral membrane proteins in the genomes. An interesting observation is that the seven-transmembrane chemoreceptors and seven-transmembrane rhodopsin families have high occurrences (48 and 46), and most of them occur in *Caenorhabditis elegans* (48 and 32). *C. elegans* has an exceptionally large number of seven-transmembrane receptors and rhodopsin-like membrane proteins (7, 17); therefore, it may be that the duplications imply possible functional relations between homologous seven-transmembrane domains. This observation is also supported by the idea that dimerization of G protein-coupled receptors may be important for their functions (19).

In contrast with the paucity of covalent combinations of transmembrane domains, covalent combinations between soluble domains and transmembrane domains are observed frequently. 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 ($\approx 34\%$ and 24% , respectively) of single-domain membrane proteins without flanking soluble domains than eukarya ($\approx 7\%$). Consistent with a study (6) of soluble proteins, 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

*To whom correspondence should be addressed. E-mail: donald.engelman@yale.edu.

© 2004 by The National Academy of Sciences of the USA

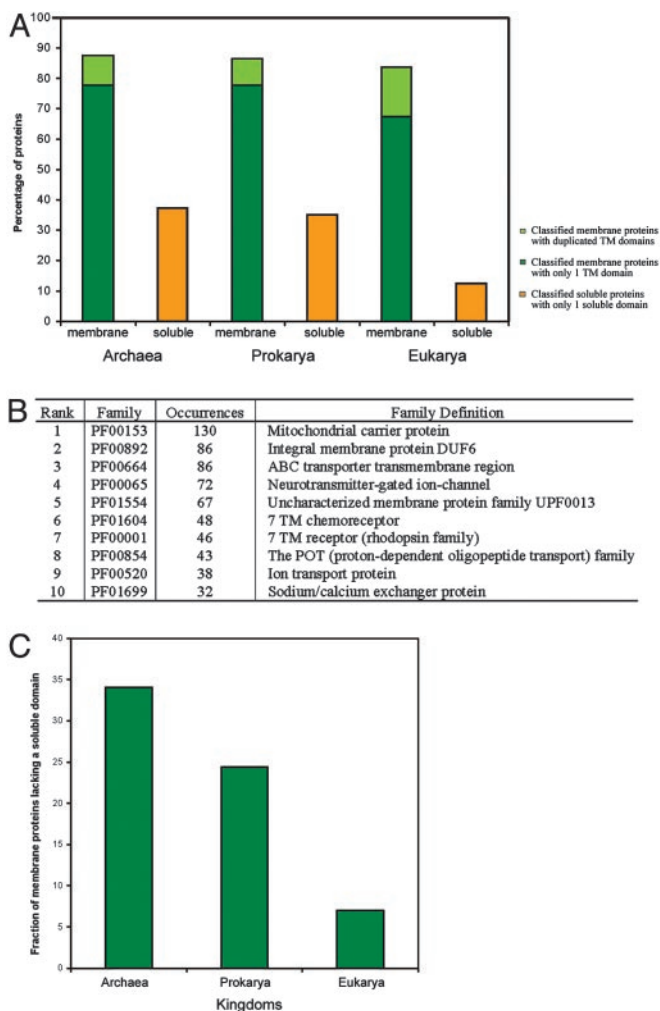


Fig. 1. Domain combination of polytopic membrane domains in genomes. (A) The green bars represent the percentage of classified membrane proteins by Pfam-A that consist of only one polytopic membrane domain, and the light green bars indicate the percentage of classified membrane proteins that consist 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*, *C. elegans*, and *Arabidopsis thaliana*. Notes on the assignment strategy: $\approx 65\%$ of the assigned membrane proteins had Pfam-A matches. Pfam-B and the clustered families were excluded because they are not classified as carefully 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 domain proteins have no more than 30 residue-flanking regions next to the membrane domains.

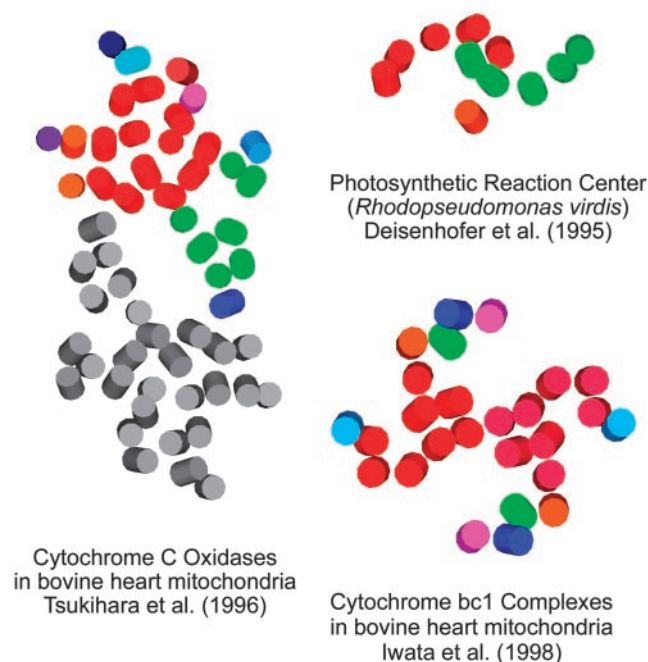


Fig. 2. Helix interactions in the membrane midplane. A five-residue section is defined at the apparent center of the membrane lipid bilayer (inferred from the hydrophobic exterior), and helix positions are indicated. The grayscale image emphasizes that the subunits shown in the colored image cannot be inferred from helix relationships.

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 (by using the Pfam-A classification) in the three kingdoms (Fig. 3). The 26 proteomes used in this study consist of 1,922 soluble domain families and 214 polytopic membrane domain families. The soluble proteins have ≈ 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

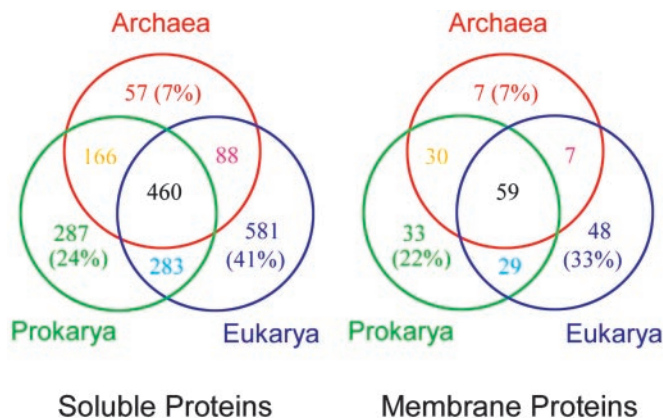


Fig. 3. Protein domain families shared among the archaea, prokarya, and eukarya kingdoms. The distributions of Pfam-A families in soluble and membrane proteins among the three kingdoms are shown. The common families shared by the three kingdoms represent 24% for soluble proteins and 28% for membrane proteins, whereas 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.

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 a substantial majority consists of only one membrane domain, indicating either that the functional diversity required is much less or that membrane proteins 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. Because the same constraint would not apply if one of the domains were a soluble domain, it is reasonable to find that covalent links are used between soluble protein domains and transmembrane domains. A challenge 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 in membrane proteins that is equivalent to domain swapping in soluble proteins.

We thank J.-L. Popot for comments. M.G. and D.M.E. thank the National Institutes of Health (Grant GM54160). Y.L. was supported by a postdoctoral fellowship (T15LMO7056).

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**, 83–86.
4. Enright, A. J., Iliopoulos, I., Kyriakides, 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.* **310**, 311–325.
7. Liu, Y., Engelman, D. M. & Gerstein, M. (2002) *Genome Biol.* **3**, research0054.
8. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001) *J. Mol. Biol.* **305**, 567–580.
9. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28**, 263–266.
10. Popot, J. L. & Engelman, D. M. (2000) *Annu. Rev. Biochem.* **69**, 881–922.
11. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996) *Science* **272**, 1136–1144.
12. Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. (1995) *J. Mol. Biol.* **246**, 429–457.
13. 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.
14. Liao, M. J., London, E. & Khorana, H. G. (1983) *J. Biol. Chem.* **258**, 9949–9955.
15. Kahn, T. W. & Engelman, D. M. (1992) *Biochemistry* **31**, 6144–6151.
16. Marti, T. (1998) *J. Biol. Chem.* **273**, 9312–9322.
17. Popot, J. L., Gerchman, S. E. & Engelman, D. M. (1987) *J. Mol. Biol.* **198**, 655–676.
18. Bargmann, C. (1998) *Science* **282**, 2028–2033.
19. Gomes, I., Jordan, B. A., Gupta, A., Rios, C., Trapaidze, N. & Devi, L. A. (2001) *J. Mol. Med.* **79**, 226–242.