Internal duplications in α -helical membrane protein topologies are common but the nonduplicated forms are rare

Aron Hennerdal, Jenny Falk, Erik Lindahl,* and Arne Elofsson*

CBR,DBB,SBC, Stockholm University, SE-106 91 Stockholm, Sweden

Received 18 February 2010; Revised 19 August 2010; Accepted 9 September 2010

DOI: 10.1002/pro.510

Published online 29 September 2010 proteinscience.org

Abstract: Many α -helical membrane proteins contain internal symmetries, indicating that they might have evolved through a gene duplication and fusion event. Here, we have characterized internal duplications among membrane proteins of known structure and in three complete genomes. We found that the majority of large transmembrane (TM) proteins contain an internal duplication. The duplications found showed a large variability both in the number of TM-segments included and in their orientation. Surprisingly, an approximately equal number of antiparallel duplications and parallel duplications were found. However, of all 11 superfamilies with an internal duplication, only for one, the AcrB Multidrug Efflux Pump, the duplicated unit could be found in its nonduplicated form. An evolutionary analysis of the AcrB homologs indicates that several independent fusions have occurred, including the fusion of the SecD and SecF proteins into the 12-TM-protein SecDF in *Brucella* and *Staphylococcus aureus*. In one additional case, the Vitamin B₁₂ transporter-like ABC transporters, the protein had undergone an additional fusion to form protein with 20 TM-helices in several bacterial genomes. Finally, homologs to all human membrane proteins were used to detect the presence of duplicated and nonduplicated proteins. This confirmed that only in rare cases can homologs with different duplication status be found, although internal symmetry is frequent among these proteins. One possible explanation is that it is frequent that duplication and fusion events happen simultaneously and that there is almost always a strong selective advantage for the fused form.

Keywords: membrane proteins; gene duplication; protein evolution; gene fusion

Additional Supporting Information may be found in the online version of this article.

AE owns shares in a company that works with bioinformatics of membrane proteins. No direct use of this research.

Aron Hennerdal and Jenny Falk contributed equally to this work.

Grant sponsors: Swedish Research Council; SSF (the Foundation for Strategic Research); EU 6'th Framework Program (EMBRACE project); Grant number: LSHGCT- 2004-512092; Grant sponsor: EU 7'th Framework Program (EDICT project); Grant number: HEALTH-F4-2007-201924.

*Correspondence to: Erik Lindahl or Arne Elofsson, CBR,DBB,SBC, SERC, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: lindahl@cbr.su.se or arne@bioinfo.se

Introduction

The membrane environment imposes particular constraints onto the structure and evolution of transmembrane (TM) proteins that are not present for soluble proteins. Although it has recently been noticed that the structures of membrane proteins are perhaps as complex as soluble proteins,^{1–4} it is clear that in general α -helical TM-proteins can be described as mainly consisting of long α -helices that traverse the membrane, with non-TM regions that are either “inside” (cytosolic) or “outside” (extracellular or inside particular organelles). This enables the use of “topology” as a useful descriptor of a TM-protein. A topology describes which residues of a

Table I. Internal Duplications Found by Strucal. The SHRIMP-based Approach Using only Sequence Information is also noted, as are Results Using BLAST for Comparison

PDB-id	OPM family	No. TMHs	STRUCTAL						SHRIMP		BLAST	
			Dupl.	Z	RMSD	n_{aligned}	% Aligned	%ID	Dupl.	Z	Dupl.	Z
Parallel												
Duplication of 4 TM-helices												
1ehkA	1.1.05.01	12	1-4/5-8	6.4	2.76	123	88	11	1-4/9-12	3.5	2-4/5-7	6.4
1fftA	1.1.05.01	12	5-8/9-12	6.0	3.06	119	86	8	1-4/9-12	3.0	1-4/9-12	2.2
1m56A	1.1.05.01	12	1-4/9-12	7.7	3.27	133	73	11	1-3/9-11	2.8	1-3/7-9	4.8
1l7vA	1.1.09.01	11	2-5/6-9	4.9	2.53	90	82	13	1-4/6-9	2.4	1-3/5-7	2.0
Duplication of 6 TM-helices												
1pw4A	1.1.14.01	12	1-6/7-12	9.1	2.83	171	96	11	1-6/7-12	11.3	2-4/8-10	3.3
2cfpA	1.1.14.02	12	1-6/7-12	8.4	2.97	168	94	12	1-6/7-12	8.3	3-5/8-10	3.3
2gfpA	1.1.14.02	12	1-6/7-12	6.6	3.72	164	94	7	2-4/8-10	6.8	5-7/8-10	1.1
2gffB	1.1.15.01	12	1-6/7-12	12.0	6.44	399	83	9	1-6/7-12	45.0	2-6/8-12	6.6
Anti-parallel												
Duplication of 3 TM helices												
1ldfA	1.1.27.01	6	1-3/4-6	6.2	1.25	76	80	18	1-3/4-6	14.9	1-3/4-6	4.5
1rc2A	1.1.27.01	6	1-3/4-6	6.8	1.05	75	81	15	1-3/4-6	13.0	1-3/4-6	4.6
1z98A	1.1.27.01	6	1-3/4-6	7.4	1.20	81	82	22	1-3/4-6	12.0	1-3/4-6	7.0
2b6pA	1.1.27.01	6	1-3/4-6	7.1	0.96	75	79	31	1-3/4-6	8.3	1-3/4-6	6.6
2f2bA	1.1.27.01	6	1-3/4-6	6.5	0.84	68	61	34	1-3/4-6	9.5	1-3/4-6	9.7
Duplication of 5 TM helices												
1rh5A	1.1.12.01	10	1-5/6-10	4.9	3.26	132	84	13	3-5/8-10	4.6	2-4/5-7	4.8
1zcdA	1.1.17.01	12	1-5/8-12	3.1	1.88	79	48	17	2-4/10-12	3.3	3-5/8-10	1.3
2a65A	1.1.18.01	12	2-6/7-11	4.5	4.92	142	63	12	2-5/7-10	3.7	1-3/4-6	4.6
1u7gA	1.1.19.01	11	1-5/6-10	7.1	2.64	128	77	16	3-6/8-11	5.3	5-7/8-10	3.9
2b2fA	1.1.19.01	11	1-5/6-10	7.7	2.36	129	80	12	3-6/7-10	4.3	1-3/6-8	1.7
3b9wA	1.1.19.01	11	1-5/6-10	7.5	3.53	133	94	12	1-5/7-11	0.7	3-5/7-9	1.7
1kplB	1.1.28.01	11	2-6/7-11	8.3	1.97	123	63	18	1-4/6-9	7.6	3-5/9-11	3.3

Predictions that differ from the structural analysis and Z-scores below 3.0 are shown in italics. "RMSD": Strucal RMSD values are averaged over aligned C α -atoms in the two fragments. "Z": Z-score. "Dupl.": duplicated unit, helix ordinal numbers. " n_{aligned} ": number of C α atoms aligned. "% ID": percentage of exact matches when compared with the number of amino acid residues with aligned C α atoms (n_{aligned}). "% Aligned": percentage of aligned amino acid residues when compared with the total number of residues in the shortest fragment.

particular protein that are within the membrane and what parts are on each side of it. For instance, the most frequent drug-target, G-protein Coupled Receptors (GPCRs), can be described as a 7 – TM – N_{out} protein. It consists of 7 TM-regions and has the N-terminal on the outside. Consequently, the C-terminal will be on the inside.

Two forces are the main determinants of the topology of membrane proteins. First, it is energetically unfavorable to put polar and charged groups in direct contact with the membrane, and therefore the long hydrophobic helices are formed. The few, often functionally important, polar and charged groups in the deep core membrane are only rarely in contact with the lipids but are instead either in contact with other polar/charged groups or with aqueous channels within the protein in question. Second, the orientation of the entire protein is governed by the "positive inside rule."^{5,6} This states that the orientation of the entire protein is decided from which side contains the most positively charged residues (Arg and Lys). For most proteins, there is a clear "positive inside"

signal, but it has recently been shown that a few protein most likely exist in a dual topology.^{7,8}

In addition to the structural constraints imposed by the membrane on TM-proteins, there are also imposed evolutionary constraints. It has been proposed that recombination events are uncommon for TM-regions and that the 2D structure of the phospholipid bilayer facilitates domain interaction without covalent linkages.⁹ Therefore, it should be more common for membrane proteins to form homo- or hetero-oligomeric complexes than to form multi-domain proteins. Arguing against this is that it has been noted that many α -helical TM-proteins contain internal symmetries that most likely origin from a duplication and fusion event.¹⁰⁻¹²

In particular, a few examples highlight how duplications have been an essential part in the evolution of TM-proteins. Five *Escherichia coli* protein families with a likely dual topology were found recently.⁸ Members of these protein families occur in genomes either as pairs or as singletons, where pairs encode two oppositely oriented proteins and

singletons encode dual-topology proteins.⁸ The functional complex is then constituted by either the two proteins or by two differently oriented dual-topology proteins. It has also been shown that only a few point mutations can change a dual-topology protein, the EmrE multidrug transporter, to a fixed orientation.¹³ Proteins in four of the *E. coli* families have an even number of TM-segments, while the fifth, a family of unknown function designated DUF606, has proteins with five TM-segments. Interestingly, representatives of DUF606 were found in a third form where the two 5-TM proteins had merged to a single 10-TM protein. A duplication including an odd number of TM-helices require that one of the duplicated units change its orientation and therefore initially after a duplication and fusion of a single protein there will not exist any “positive inside” bias for these proteins. Alternatively, if the symmetric protein that is created is a fusion of two homologous proteins, these need to be differently oriented before the fusion. Lolkema *et al.*¹⁴ studied the evolutionary history of DUF606 and revealed a total of nine independent duplication events in the DUF606 family. Five of these resulted in paired (two 5-TM proteins) genes, and four resulted in fused 10-TM genes.

A detailed analysis of internal symmetry in membrane proteins with known 3D structure showed that nearly half of them contain an internal symmetry.¹² It was also predicted that about one quarter of the multi-spanning membrane proteins in SWISSPROT¹⁵ contained an internal symmetry. As internal symmetry was more abundant among proteins involved in transport processes than other groups, it was suggested that it might be important for conformational changes needed for the transport process.

Although it is clear that symmetry is a common feature of membrane proteins, the evolutionary history of how these proteins became symmetric is not clear, with the exception of DUF606. To improve our understanding of the internal symmetries among membrane proteins, we have here examined their frequency and their evolutionary history.

Results and Discussion

Fifty-three proteins chains with more than six TM-segments were extracted from Orientations of Proteins in Membranes (OPM).¹⁶ Thereafter, structural alignments using Strucal¹⁷ were performed. Given the dominance of long hydrophobic helices in TM-proteins, the original scoring schemes used in Strucal might not provide correct statistics for TM-proteins.^{18,19} After some optimization, it was found that using Z-scores calculated from a background distribution of non-related fragments (different OPM¹⁶ superfamilies) with a fixed number of TM-regions was suitable for membrane proteins, see the Methods section for details. A Z-score of 3.0 or higher was

then used as threshold to detect internal duplications, see Table I.

Twenty (38%) of the protein chains were classified as containing an internal duplication, see Table I. The identified fragments were then checked manually and superpositions of them are illustrated in Figure 1. In addition, RMSD values for the fragment pairs and the percentage of successfully aligned c_{α} -atoms were collected and are presented in Tables I and II. To give an idea of how sequence similarity influences the identification, the percentage of identical residues in these regions are also noted. In most cases, the sequence identity is quite low (10%), the RMSD is acceptable (2–4 Å) and the coverage is at least 75% of the length of the fragments, indicating that the structural alignment is good although the sequence identity is low.

Detecting internal duplications using sequence information

To understand how frequent internal duplications are among all TM-proteins, it is necessary to develop methods that can detect internal similarity between potentially duplicated units reliably using only sequence information. Therefore, two methods based on BLAST²⁰ and SHRIMP¹⁹ were tested for their ability to detect internal duplications in TM-proteins. It was found that SHRIMP could detect that duplications had occurred in 17 of the proteins, while BLAST could detect the duplication in 14 cases, Table I. Additionally, four putative internal duplications were found by SHRIMP, which all appeared to be real upon manual inspection, Table II. In the first case, a putative metal-chelate type ABC transporter (PDB-ID: 2nq2A), the internal duplication was found in another representative from the same ABC transporter superfamily by Strucal. Here, the corresponding Z-score from Strucal is indeed quite close to the threshold of 3.0. For a further two cases, photosystem II (2axtb) and a glutamate symport protein (2nw1A), one helix has undergone a lateral shift and therefore the duplications were not detected by the structural alignment. Superpositions of these proteins are included in Figure 1. The fourth case comprises the first chain of mitochondrial ADP/ATP carrier protein (1okcA), which turned out to be a triplicate of 2-TM-segments rather than a duplication of 3-TM-segments as predicted. Since the search space for the structural alignments only included fragments of between 3 and 6 TM-segments, this duplication was missed. The signal of the 3-TM duplication suggested by the sequence-based methods was discarded as too weak, see Table II. A visualization of this protein can be found in Supporting Information, Figure 3.

Despite the good performance of SHRIMP in identifying the presence of internal duplications, the “correct” duplication unit was only found in 10 of the

Figure 1. Visualisation of internal duplications detected by Strucal and SHRIMP, one example for each OPM superfamily. For all proteins, the duplicated regions are extracted and thereafter one of them is rotated and aligned to the other using Strucal. In 2nwlA and 2axtB, a duplication has most likely occurred that includes a lateral rotation of one helix.

Table II. Internal Duplications Found by SHRIMP but not by Strucal

PDB-id	OPM family	No. TMHs	STRUCTURAL						SHRIMP		BLAST	
			Dupl.	Z	RMSD	n_{align}	% Aligned	%ID	Dupl.	Z	Dupl.	Z
2axtb	1.1.02.03	6	1-3/4-6	2.1	2.03	53	54	13	1-3/4-6	5.9	No hits found	
2nq2A	1.1.09.01	10	1-3/4-6	2.1	2.63	53	58	9	2-4/5-7	3.6	2-4/7-9	1.5
2nwlA	1.1.16.01	13	1-3/4-6	2.3	2.15	58	62	9	1-3/4-6	4.5	No hits found	
1okcA	1.1.13.01	6	1-3/4-6	0.5	2.78	53	45	4	1-3/4-6	17.0	1-3/4-6	11.6

Predictions that differ from SHRIMP and Z-scores below 3.0 are shown in italics. "RMSD": Strucal RMSD values are averaged over aligned C α -atoms in the two fragments. "Z": Z-score. "Dupl.": duplicated unit, helix ordinal numbers. " n_{aligned} ": number of C α atoms aligned. "% ID": percentage of exact matches when compared with the number of amino acid residues with aligned C α atoms (n_{aligned}). "% aligned": percentage of aligned amino acid residues when compared with the total number of residues in the shortest fragment.

chains. In several cases, the duplicated unit detected by SHRIMP was too short, partly due to the fact that SHRIMP has to rely on predicted topologies. An attempt to promote longer fragments by biasing the scoring function did not yield promising results (data not shown).

Among the three proteins completely missed by SHRIMP, two have Z-scores close to the threshold of 3.0; 2.8 for cytochrome C oxidase (1m56A) and 2.4 for the ABC transporter BtuCD (117vA). The third, an Rh-like ammonia transporter (3b9wA), has a very weak signal of 0.7 but the internal duplication is found in another member of the same family (2b2fA).

In conclusion, in 24 out of 53 protein chains an internal duplication can be detected either using structural alignments or using SHRIMP. Most internal duplications that can be detected using Strucal can be detected using SHRIMP, indicating that it would have been possible to identify these even before the structures were solved. Therefore, internal duplications detected by SHRIMP should provide a good estimate of how frequent internal duplications are in membrane proteins. However, the exact extent of the duplication might be hard to identify using sequence alone, in particular as it is necessary to rely on predicted topologies.

Larger proteins are more likely to be internally duplicated

It can be noted that the frequency of internal duplications increases rapidly with the number of TM-regions in a protein. Of the nine OPM superfamilies that contain proteins with more than 10 TM-segments, all but one (89%) contains an internal duplication event, while for shorter chains (10 TM-segments or less) only two out of 12 (17%) contained a duplicated segment. It should also be noted that in many of the proteins the protein is not completely made up of the duplicated unit. Additional helices exist mainly at the N- and C-termini.

A few membrane proteins have been found to have a dual topology.^{8,14} In these proteins, the positive inside signal deciding which way the protein should be inserted into the membrane is so weak that both orientations have been observed. It has been suggested that at least some of these proteins are functional as homo-dimers, with the two chains oriented differently. Observations have been made where this kind of proteins have duplicated and merged as a functional unit, that is, with the two copies in opposite orientation in the membrane.

To try to find the frequency of anti-parallel duplications, the supposed orientation of the identified duplications was studied. To be able to adopt an opposite orientation of the duplicates, the number of duplicated TM-segments must be odd or an odd number of helices must have been inserted between

the duplicates. Surprisingly, parallel and anti-parallel duplications are found in an approximately equal amount. All even-number duplication units are parallel and all odd-number are antiparallel, see Table I. In total, six OPM superfamilies contain anti-parallel duplication units, but in three of these additional helices were also present at the N or C-termini, indicating that the evolutionary history of these proteins might not be as simple as for the previously described dual-topology protein family DUF606. We also examined the possibility to separate parallel and anti-parallel proteins from entire genomes, but due to SHRIMP's lack of ability to detect the exact duplication unit, no clear conclusion could be drawn.

Detection of duplicated proteins in genomes

Encouraged by the result of the analysis of the test-set, we decided to search for duplicated proteins in three genomes. For that purpose the complete genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo Sapiens* were downloaded and TM-segments were predicted for all peptide chains using PRODIV-TMHMM.²¹ Proteins with 6 or more predicted TM-segments were used for further analysis.

The fraction of membrane proteins with at least 6 TM-segments that are duplicated differs between the genomes. *E. coli* and yeast contain 47 and 43% internally duplicated membrane proteins respectively, while for only 22% of the membrane proteins in the human genome an internal duplication can be detected. This difference can to a large extent be attributed to the large family of GPCRs, which are non-duplicated 7 TM-helices proteins, see Figure 2. GPCRs represent almost 50% of the long human TM-proteins studied here. Anyhow, with the exception of 12 TM-proteins, it seems as if *E. coli* contains slightly more duplicated proteins given the same number of TM-helices than the other two genomes, and all genomes contain less internally duplicated TM-proteins than was predicted to have occurred in the structural data-set.

In all three genomes, the probability to contain an internal duplication is at least twice as high if a protein contains nine or more predicted TM-segments compared to if it contains less, Figure 2. For proteins with 6–8 TM-helices less than 20% contain an internal duplication, while for chains of medium length (9–11 TM-helices) the fraction ranges from 30 to 60% in the different genomes. In all genomes 12-TM proteins are most likely to contain an internal duplication with close to 70% of the chains duplicated. The fraction appears to be slightly lower for the few proteins with more than 12 TM-helices.

Despite the fact that identified duplicated units using SHRIMP are frequently shorter than the correct duplicated unit, it was found that in most cases the duplication unit is the longest possible one, that is, for proteins with eight or nine TM-helices it is

Figure 2. The length of the detected duplications in the three genomes and in the OPM set using predicted TM-segments. In all sets, longer chains seem to contain a higher fraction of duplications.

most frequent with an internal duplication unit of four, and so forth, Figure 2. In particular, duplications in 12-TM proteins consist very frequently of 6-TM units.

Search for the origin of the duplicated units

If membrane proteins frequently evolve by gene duplication followed by merging of the two gene copies, there might remain evidence of this in the form of nonmerged proteins in some genomes. In search of these, three iterations of PSI-BLAST²⁰ were made for all the proteins with a confirmed duplicated unit in a database containing close to 600 bacterial genomes, Figure 3. Surprisingly, for only one of the proteins, the 12 TM-helices AcrB Multidrug Efflux Pump (2gifB), a non-duplicated homolog was found. For another family of proteins, the Vitamin B₁₂ ABC transporter proteins (1l7vA and 2nq2A), an additional duplication of the protein was observed.

For all other proteins, most homologs found had the same number of TM-segments or possibly a few more or less. These discrepancies may be due to subsequent addition of helices at the ends of the proteins, or might simply be due to mispredictions of TM-segments or incorrect gene predictions. For the OPM superfamily of Cytochrome c oxidases (OPM: 1.1.05) three peaks can be seen; one at 12 TM-segments, one around 15 and one at 19, see 1m56A in Figure 3 for an example. The additional peak at 15 TM-helices was upon closer inspection not a result of an internal duplication. These proteins contained

an addition of one N-terminal and two C-terminal helices. The N-terminal region of the 19-TM protein aligned perfectly with our duplicated protein chain 1m56A, while we could not detect any similarity between the five C-terminal helices and 1m56A or the 19-TM protein itself.

Evolutionary history of AcrB related proteins

A total of 4180 homologs to the AcrB Multidrug Efflux Pump (PDB ID: 2gifB) were found after three rounds of PSI-BLAST.²⁰ A total of 76% (3197) of them were predicted to contain 11–13 TM-helices, and 19% (806) 5–7 TM-helices, Figure 3. The majority of the genomes show a trend of having exactly two or zero short homologs, Figure 4, while the number of long homologs varies more. On closer inspection of the short homologs, it was seen that most of them were annotated to be units of the Sec translational machinery, SecD or SecF. These are believed to be important for maintaining the proton motive force over the membrane and possibly for preventing backward sliding of the pre-protein chain. It is also known that in some bacterial genomes SecD and SecF has been fused to form the SecDF 12-TM protein.²²

A phylogenetic consensus tree was constructed using the PHYLIP package containing homologs from all 598 genomes. To obtain a more informative tree, a homology reduction of 50% was made, resulting in 693 long homologs and 197 short homologs. Proteins predicted to have 11, 12, or 13 TM-

Figure 3. The distribution of the number of TM-segments among the homologues found after three rounds of PSI-BLAST, for one representative per OPM superfamily of all duplicated proteins found.

segments were split into N- and C-terminal parts (1386 halves). This way all sequences in the tree were of roughly the same length. To further simplify visualization, only 10% of the homologs that remained were shown in the tree, which could be divided into five distinct clusters, Figure 5. A full tree without this selection of 10% can be found in the supplementary material, Figure 1.

Two of the clusters—the two predominantly gray clusters at the bottom of Figure 5—contain the SecF and SecD proteins, respectively. Mixed in with the SecF proteins are C-terminal fragments of the 12-TM SecDF proteins and mixed with SecD proteins are the corresponding N-terminal parts found, marked by stars (“*”) in the figure. The separation of the N- and C-terminal parts shows that the SecDF proteins have

been formed by a fusion of two 6-TM proteins and not from a duplication of a single protein. Among the interspersed 12-TM proteins, we find two proteins from the *Brucella* genus of the *Proteobacteria* phylum, and one member of *Staphylococcus aureus* of the *Firmicutes*. The closest 6-TM homologs to these protein halves are unsurprisingly from related organisms from the same respective phyla. The only case where half of a 12-TM protein lacks its complimentary half within these two clusters is the C-terminal half of NP_224760.1, a protein from *Chlamydomonas reinhardtii*. The complimentary N-terminal half can be found close to the root of the cluster of other N-terminal halves and are described below.

A third cluster—in blue at the top of Figure 5—consists exclusively of C-terminal parts of acroflavin

Figure 4. Number of genomes with certain number of hits of different lengths. Homologs with 5,6, or 7 TM-segments generally come in pairs while the longer homologs with 11,12, or 13 TM-segments can exist in a varying number of copies in different genomes.

homologs, while just below it there is a fourth cluster mainly consisting of the corresponding N-terminal parts. Within the N-terminal cluster a few (four) 5–7 TM proteins are found.

The last cluster—in the middle of Figure 5—is comprised of a mix of N- and C-terminal parts of 12-TM proteins and can be divided into a set of sub-cluster pairs consisting of the N- and C-termini. Due to low bootstrap support values (< 43), it is not possible to confidently say whether all these proteins origin from the same ancient single 6-TM protein or not.

From the tree, a potential evolutionary history of the AcrB homologs could be described as follows. Most likely, a single 6-TM protein existed at first. This protein was then duplicated to form two proteins, SecD and SecF. In a few genomes, a fusion of these two into SecDF occurred later. It appears to have been several fusion events since both fused and non-fused forms exist in several different phyla. In addition, the 6-TM protein was early duplicated and fused into one or several 12-TM proteins that in turn diverged to form several sub-families.

Duplications in Vitamin B_{12} ABC transporters

ABC transporters can be divided into two subtypes on the basis of the direction of the transport reaction.²³ ABC importers are only found in prokaryotic genomes, while ABC exporters also exist in eukaryotic genomes. The functional units of ABC transporters are believed to consist of two integral membrane domains/proteins and two cytoplasmic domains/proteins. The TM domain can show a large variation and is not evolutionary conserved between different sub-groups of ABC transporters. In a study by Bornberg-Bauer and Weiner,²⁴ it has been suggested that ABC transporters have undergone a large number of

circular permutation events. It has been suggested that eukaryotic ABC exporters are expressed with all four domains in a single polypeptide chain, while in prokaryotic genomes they are separate genes.²³

The crystal structures of two ABC-transporters from the Vitamin B_{12} sub-family were included in the set of proteins that contained an internal duplication. The two proteins, 117vA and 2nq2A are homologs (35% sequence identity), but are annotated in OPM to contain a different number of TM-segments; 11 versus 10. However, the difference is due to that in 117vA the automatic annotation method in OPM assigns a TM-segment to a reentrant region, while in 2nq2A the corresponding region is not annotated as such. The internal duplication corresponds to four TM-helices to form helix 2–5 and 6–9 in the final structures.

Basically all homologs to 117vA and 2nq2A have between 8 and 10 predicted TM-segments. However, in a few genomes fused proteins with 18–20 predicted TM-segments were found, see Figure 3. From a phylogenetic analysis seems quite likely that this fusion occurred only once, see Supporting Information Figure 2.

Genome-wide search in *Homo sapiens* for duplication patterns

As an extension to the search for the origin of the duplicated units described, we conducted the same type of analysis on a genomic scale for all long TM-proteins in *Homo sapiens*. Here, SCAMPI²⁵ was used to find all membrane proteins with six or more predicted TM-segments and PSI-BLAST²⁰ was used to detect homologs in an extended database including both eukaryotic and prokaryotic proteins from completely sequenced genomes. For all homologs, the number of TM-segments were predicted, also using SCAMPI,²⁵ and for each query protein the distribution of the number of homologs with a given number of predicted segments were collected.

A methodology based on fitting gaussian functions to the data was used to single out patterns that could indicate the existence of homologs with difference in duplication status, see Methods. These proteins were thereafter clustered on the similarity in the fitted curves, so that protein families with a similar pattern of TM-segments were grouped together. This clustering matched very well with the domain content of the families, see Figure 7.

Using this automated detection of double-peaks, we found 48 proteins that could be grouped into 13 groups that seemed to exist in versions with different duplication states, see Figure 7. A few of these groups (A: Inositol phospholipid synthesis protein; B: a group of proteins without well-assigned Pfam domains; C: Peptidase family M48; G: membrane bound O-acyl transferases H: ABC transporters; L: Cytochrome C and Quinol oxidase polypeptide I; M:

6-TM proteins

N-terminal part of 12 TM protein

C-terminal part of 12 TM protein

Figure 5. Tree of homologous proteins from the 598 genomes. To simplify visualization, leaves and connecting branches to a sequence with 5, 6, or 7 predicted TM-segments are colored gray. Leaves and branches to sequences with 11, 12, or 13 predicted TM-segments that are split in half are colored red when being the N-terminal half and blue denoting the C-terminal half. Five clades can be seen; one containing all Sec-annotated proteins, one containing N-terminal parts of 11,12,13 TMH homologues and half of the six TMH homologues, and one containing C-terminal parts of long homologues and the other half of the short homologues. Bootstrap consensus tree of 2gifB and homologues taken from a database consisting of 10% randomly chosen sequences from 598 genomes.

Table III. Number of Homo sapiens TM Proteins that Contain an Internal Duplication (SHRIMP) and/or a Dual Peak of Homologous Proteins with a Different Number of TM-segments

		Internal duplication	
		+	–
2-peak pattern	+	31	17
	–	677	1616

Polycystin cation channel) contained 16 proteins where no internal duplication was found, see Figure 7. For most of these proteins families, the number of TM-segments seemed to vary and no very clear two peak pattern exist. However, in one protein from the Peptidase family M48, a clear pattern of 4 and 7 TM-segments could be observed indicating that proteins from this family mostly exist in two forms and perhaps have undergone an internal duplication that could not be detected by SHRIMP. Further, for a few members of the ABC-transporter family (Group H) a duplication from 6 to 12 TM-segments can be seen, once again showing that this group of proteins exist in forms with different number of TM-segments.

In four groups (D: The surfeit locus gene SURF4; F: membrane bound O-acyl transferases; I: hypothetical protein and J: the Patched family) of in total 16 proteins, shorter proteins matching the predicted non-duplicated form can be found. In the last two groups (E: Alanine dehydrogenases and K: Major Facilitator Superfamily), a duplication can be found, but the patterns does not match exactly the prediction duplication units. The Alanine dehydrogenases in the human protein have 14 TM-segments that is partly built up from a 3 or 4 TM-segments long duplication unit. However, most of the proteins in this family have 3, 5, or 8 TM-segments showing that this also is a family with a variable set of TM-segments.

In summary, out of the 2341 human long TM proteins 22% were predicted to contain an internal duplication, but for only 17 of these the non-duplicated form could be found in any genome, supporting the conclusions from the structural dataset, see Table III.

Conclusions

We have described the development of two methods for the detection of α -helical membrane protein duplications, based on structure and sequence information, respectively. It was found that the sequence-based method could detect most of the symmetries found by the structure-based method. The results are in some cases heavily dependent on the quality of membrane protein topology predictions, and it was difficult to reliably detect the exact extent of the

duplicated unit. Surprisingly, parallel and anti-parallel duplications were found at similar frequencies in the structural dataset. In contrast to earlier observations, it was found that in the majority of the superfamilies the protein is not completely made up of the duplicated unit. Additional TM-helices were found mainly at the termini, but helices were also inserted between the duplicated units in two superfamilies.

The sequence-based method was also used to identify internally duplicated TM-proteins in some completely sequenced organisms and also to detect homologs to all internally duplicated TM-proteins with known 3D-structure. It was found that the fraction of long membrane proteins containing an internal duplication was 47, 43, and 22% for *E. coli*, *S. cerevisiae*, and *H. Sapiens*, respectively. The lower fraction of duplicated proteins in *H. Sapiens* is due to the large fraction of non-duplicated 7-TM GPCRs. In general, it was also found that longer TM-protein are more likely to contain an internal duplication.

Surprisingly, only one protein was found where we could detect the existence of the duplicated units as a single protein, the Acrb Multidrug Efflux Pump. In addition, one example—the Vitamin B_{12} transporter-like ABC transporters—was found to have undergone an additional duplication in some genomes. All other proteins were only found in their duplicated version. This trend was then supported by a search for non-duplicated forms of all long TM-proteins in the human genome. This leaves the interesting question of whether this symmetry might be the result of convergent evolution, if a homo-multimer is very unfavorable, or if we just are unable to detect these proteins. Arguing against the latter is that we actually can detect the “homology” between the internally duplicated units.

Methods

Datasets

One hundred and fifty-two α -helical membrane proteins with known three-dimensional structure were downloaded from the OPM database¹⁶ (<http://opm.phar.umich.edu>). The structure files in PDB-format were used without modification, and the corresponding full protein sequences were downloaded from UniProt (<http://www.uniprot.org>). This dataset was homology reduced to 40% sequence identity using CD-Hit.²⁶ The set was further limited to sequences containing at least six TM-segments resulting in 53 protein chains, see Supporting Information. An extended dataset limited to sequences containing at least three TM-segments resulting in 81 chains, was used for the benchmark of scoring schemes.

Each of the 53 chains was split into fragments consisting of 3, 4, 5, and 6 TM-segments respectively, excluding N- and C-terminal loops. Structural

alignments for all combinations of fragments of equal number of TM-segments were constructed using Strucal¹⁷ and 20 proteins from 10 superfamilies were identified as having internal symmetries, see Table I. A representative protein from each of these 10 superfamilies were visualized and the fragments superposed using the PyMOL software (<http://www.pymol.org>), see Figure 1.

Detection of internal duplications

A benchmark was made comparing different normalization schemes using identical strategies for Strucal,¹⁷ SHRIMP,¹⁹ and BLAST²⁰ to adapt the scoring to membrane proteins. A background distribution of the score from each method was obtained using pairs of unrelated protein fragments, that is, fragment pairs that are guaranteed not to represent an internal duplication. “Unrelated” fragments were in this case defined as fragments of protein chains stemming from different OPM superfamilies.

Mean values and standard deviations of these background distributions were used to find particularly strong signals. Each possible internal nonoverlapping fragment pair within a protein chain was compared using each respective method. A Z-score was assigned based on aforementioned means and standard deviations, and it was found that a Z-score above 3.0 was suitable to define fragment pairs representing an internal duplication. If more than one pair within the same protein met this criterion, the pair with the highest Z-score was chosen, and the other pairs were discarded.

The same strategy was used to detect internal duplications in the three complete genomes with the difference that the topologies were predicted using PRODIV-TMHMM.²⁷ Two eukaryotic genomes were downloaded from Ensembl²⁸ using the versions “*Saccharomyces cerevisiae* SGD1.01.45” and “*Homo Sapiens* NCBI36.45”. In addition, the *E. coli* genome was downloaded from NCBI²⁹ using *Escherichia coli* strain K12, sub-strain MG1655. Comparisons using SHRIMP were made for each protein chain predicted to have 6 or more TM-segments. As in the test-set, a Z-score above 3.0 was seen as an indication of a duplication event. If more than one candidate was found within the same protein, the pair with the highest Z-score was considered to be the duplication unit.

Detection of homologs to duplicated proteins

The eukaryotic genomes were combined with 595 bacterial genomes downloaded on 2007-11-12 from the National Center for Biotechnology Information’s (NCBI) FTP-site (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>). For each of the 20 proteins where an internal duplication had been detected, PSIBLAST²⁰ (e-value cutoff: 10^{-3} , rounds: 3, maximum number of hits: 10^6) was used to detect homologs. Thereafter, the to-

Figure 6. Example of results from curve fitting of mono- and bi-topic Gaussian functions to homolog TM-segment distribution data. The protein exemplified is “ENSP00000284476”, a 12-TM protein from the “Patched” family of proteins involved in hedgehog signalling.

pology for each identified homolog was predicted using PRODIV-TMHMM.²⁷

Phylogenetic tree of acroflavin resistance protein homologs

The tree in Figure 5 was constructed starting from the sequence of the acroflavin resistance protein (UniProt ID: P31224, PDB ID: 2gifB) downloaded from Uniprot (<http://www.uniprot.org>). Homologs were found in the 598 genomes using PSI-BLAST²⁰ as described above. Sequences of homologs predicted to contain 5, 6, 7, 11, 12, or 13 TM-segments, consistent with the two sets of peaks seen for 2gifB in Figure 3, were then extracted.

To obtain a quality measure for the different clusters in the tree and assess the credibility of the results seen, 100 bootstrap sets were created from the alignment data. These bootstrap sets were then used independently to create 100 trees, which were in turn merged into one consensus tree, Figure 5. Branch lengths represent the number of times the particular branch was seen in the bootstrap trees.

The set of homologs was then homology reduced to 50% using CD-Hit.²⁶ Thereafter, the longer sequences of 11, 12, or 13 TM-segments were split into N-terminal and C-terminal halves and a multiple sequence alignment of all these sequences and sequence fragments was constructed using kalign.³⁰

The PHYLIP³¹ software package was subsequently used to construct the phylogenetic tree. Protein distances were calculated using PROTDIST, the tree was then created by neighbor-joining using NEIGHBOR, and drawn using DRAWGRAM. A more visually appealing but for all other purposes equivalent tree was made by out-grouping the node separating the Sec F/D homologs from the other proteins using RETREE. The resulting tree was finally modified to annotate N-terminal and C-terminal

Figure 7. Distribution of the number of TM-segments in homologs to human proteins that implicates a gene duplication and merge in the past. The included query proteins (y-axis) are from *Homo sapiens* and have six or more predicted transmembrane segments. The heat-map shows the number of homologs to the query sequence in 650 genomes and their predicted number of TM-segments. “Q” denotes the number of TM-segments of the query sequence, “D” denotes the duplication size suggested by our SHRIMP-based duplication detection method. If no duplication was detected no “D” is present. At the right side of each row, the Pfam domains present in each protein is listed. All 48 proteins are clustered according to the similarity of the fitted curves. The proteins can roughly be divided into 13 groups, labeled A to M.

halves and proteins of 5,6,7 TM-segments in red, blue, and gray colors, respectively.

Homolog TM-segment pattern heat-map

Starting out with the downloaded human genome from Ensemble²⁸ described previously, the single sequence version of SCAMPI²⁵ was used to identify membrane proteins with six or more predicted TM regions. SCAMPI was chosen for its relatively good performance and comparably low computational requirements.

The resulting 2341 proteins sequences were aligned to the sequences of 649 eukaryotic and prokaryotic genomes using PSI-BLAST²⁰ (e-value cutoff: 10^{-4} , rounds: 3, maximum number of hits: 10^6). The genomes comprise the 595 bacterial genomes described previously with an addition of 54 eukaryotic genomes downloaded from the Ensemble²⁸ FTP site. The number of TM segments for each found homolog was predicted using SCAMPI²⁵. For each human query sequence, the frequency of homologs with each number of TM-segments between 1 and 30 were recorded, yielding 2341 curves as exemplified by Figure 6. This distribution of number of TM-segments in homologs was then normalized so that the frequencies for each query protein summed to 1.

$$f_{\text{monotopic}}(x) = A \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (1)$$

$$f_{\text{bitopic}}(x) = A_1 \cdot \exp\left(-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right) + A_2 \cdot \exp\left(-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right) \quad (2)$$

To identify patterns with two distinct peaks in the TM-segment distributions, a monotopic (1) and a bitopic Gaussian (2) were fitted to each normalized distribution using the leastSquaresFit function of the python package Scientific.Functions.LeastSquares (<http://dirac.cnrs-orleans.fr/ScientificPython/ScientificPythonManual/Scientific.Functions.LeastSquares-module.html>). This function gave fitted parameter values for the different A , μ and σ parameters in the above equations. It also provides a measure of the quality of the fit in the form of a χ^2 -value. A few intuitive rules allowed the filtering out of 48 proteins showing a two-peak pattern. The rules derived after a limited amount of optimization are in the form of conditions on the fitted parameters. A distribution is considered to have a double peak if the following applies: 1) the bitopic Gaussian fits better than the monotopic Gaussian (lower χ^2 -value), 2) the μ -values of the bitopic Gaussian differ more than 2.5, 3) the σ -values of the bitopic Gaussian are lower than 0.7 and 4) the TM-segment frequency

corresponding to the peaks of the bitopic Gaussian are greater than 0.05 (Fig. 6).

Figure 7 was constructed using the python package Matplotlib (<http://matplotlib.sourceforge.net>).

Acknowledgment

We also thank Dr. Erik Granseth for inspiration.

References

1. Granseth E, von Heijne G, Elofsson A (2005) A study of the membrane-water interface region of membrane proteins. *J Mol Biol* 346:377–385.
2. Viklund H, Granseth E, Elofsson A (2006) Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J Mol Biol* 361:591–603.
3. Kauko A, Illergard K, Elofsson A (2008) Coils in the membrane core are conserved and functionally important. *J Mol Biol* 380:170–180.
4. Elofsson A, von Heijne G (2007) Membrane protein structure: prediction versus reality. *Annu Rev Biochem* 76:125–140.
5. von Heijne G, Gavel Y (1988) Topogenic signals in integral membrane proteins. *Eur J Biochem* 174:671–678.
6. Nilsson J, Persson B, von Heijne G (2005) Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins* 60:606–616.
7. Saaf A, Johansson M, Wallin E, von Heijne G (1999) Divergent evolution of membrane protein topology: the Escherichia coli RnfA and RnfE homologues. *Proc Natl Acad Sci USA* 96:8540–8544.
8. Rapp M, Granseth E, Seppala S, von Heijne G (2006) Identification and evolution of dual-topology membrane proteins. *Nat Struct Mol Biol* 13:112–116.
9. Liu Y, Gerstein M, Engelman D (2004) Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *Proc Natl Acad Sci USA* 101:3495–3497.
10. Shimizu T, Mitsuke H, Noto K, Arai M (2004) Internal gene duplication in the evolution of prokaryotic transmembrane proteins. *J Mol Biol* 339:1–15.
11. Dutzler R, Campbell E, Cadene M, Chait B, MacKinnon R (2002) X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* 415:287–294.
12. Choi S, Jeon J, Yang J, Kim S (2008) Common occurrence of internal repeat symmetry in membrane proteins. *Proteins* 71:68–80.
13. Rapp M, Seppala S, Granseth E, von Heijne G (2007) Emulating membrane protein evolution by rational design. *Science* 315:1282–1284.
14. Lolkema J, Dobrowolski A, Slotboom D (2008) Evolution of antiparallel two-domain membrane proteins: tracing multiple gene duplication events in the DUF606 family. *J Mol Biol* 378:596–606.
15. Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24:21–25.
16. Lomize M, Lomize A, Pogozheva I, Mosberg H (2006) OPM: orientations of proteins in membranes database. *Bioinformatics* 22:623–625.
17. Gerstein M, Levitt M (1996) Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc Int Conf Intell Syst Mol Biol* 4:59–67.

18. Hedman M, Deloof H, Von Heijne G, Elofsson A (2002) Improved detection of homologous membrane proteins by inclusion of information from topology predictions. *Protein Sci* 11:652–658.
19. Bernsel A, Viklund H, Elofsson A (2008) Remote homology detection of integral membrane proteins using conserved sequence features. *Proteins* 71:1387–1399.
20. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
21. Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* 13:1908–1917.
22. Bolhuis A, Broekhuizen C, Sorokin A, van Roosmalen M, Venema G, Bron S, Quax W, van Dijk J (1998) SecDF of *Bacillus subtilis*, a molecular Siamese twin required for the efficient secretion of proteins. *J Biol Chem* 273:21217–21224.
23. Hollenstein K, Dawson R, Locher K (2007) Structure and mechanism of ABC transporter proteins. *Curr Opin Struct Biol* 17:412–418.
24. Weiner J, III, Bornberg-Bauer E (2006) Evolution of circular permutations in multidomain proteins. *Mol Biol Evol* 23:734–743.
25. Bernsel A, Viklund H, Falk J, Lindahl E, Heijne GV, Elofsson A (2008) Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci USA* 105:7177–7181.
26. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17:282–283.
27. Viklund H, Elofsson A (2004) Best-helical transmembrane protein topology predictions are achieved using hidden Markov models. *Protein Sci* 13:1908–1917.
28. Flicek P, Aken B, Beal K, Ballester B, Caccamo M (2008) Ensembl 2008. *Nucleic Acids Res* 36:D707–D714.
29. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J (2007) GenBank. *Nucleic Acids Res* 35:D21–D25.
30. Lassmann T, Sonnhammer E (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6:298.
31. Retief J (2000) Phylogenetic analysis using PHYLIP. *Methods Mol Biol* 132:243–258.