

Membrane protein structural biology - How far can the bugs take us? (Review)

Erik Granseth^{ab}, Susanna Seppälä^a, Mikaela Rapp^a, Daniel O. Daley^a, Gunnar von Heijne^{ab}

^aCenter for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

^bStockholm Bioinformatics Center, Stockholm, Sweden

(Received 13 February 2007; and in revised form 5 April 2007)

Correspondence: Prof. Gunnar von Heijne, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. Fax: +46 8 153 679. E-mail: gunnar@dbb.su.se

Abstract

Membrane proteins are core components of many essential cellular processes, and high-resolution structural data is therefore highly sought after. However, owing to the many bottlenecks associated with membrane protein crystallization, progress has been slow. One major problem is our inability to obtain sufficient quantities of membrane proteins for crystallization trials. Traditionally, membrane proteins have been isolated from natural sources, or for prokaryotic proteins, expressed by recombinant techniques. We are however a long way away from a streamlined overproduction of eukaryotic proteins. With this technical limitation in mind, we have probed the question as to how far prokaryotic homologues can take us towards a structural understanding of the eukaryotic/human membrane proteome(s).

Keywords: *Membrane protein, structural biology, crystallography, protein overexpression*

This is an electronic version of an article published in *Molecular Membrane Biology* (2007) 24, 329-332.

Molecular Membrane Biology is available online at:

<http://dx.doi.org/10.1080/09687680701413882>

Introduction

α -helical membrane proteins (MPs) are involved in a wide range of vital cellular processes, as well as in numerous medical conditions. Many scientists in both academia and industry therefore strive for a molecular understanding of how they function. Essential to this process is a high-resolution structure, so that structure-function analyses can be conducted at an atomic level. High-resolution structures are therefore much anticipated, but progress has been slow: as of 1 January 2007, MPs comprised less than 0.5% of the Protein Data Bank (Berman et al. 2002), even though they constitute 20-30% of a typical proteome (Krogh et al. 2001, Granseth et al. 2005, Wallin & von Heijne 1998).

The striking paucity of MP structures, as compared to soluble proteins, does not reflect the level

of investment or interest. Accordingly, MP structures are extremely well received by the scientific community. The majority of novel structures (92%) have found their way into high-impact journals like *Nature*, *Science* or *Cell* (Figure 1). These structures are also frequently rewarded with an editorial comment (55%) and/or an illustration on the cover (31%) of the journal. Most notably, 8% of the novel structures have helped win a Nobel Prize. So if the reward justifies the effort, what's the problem?

The poor progress of structural determination can be attributed to the fact that MPs are incompatible with the structural genomics approaches that have been so successful for their soluble counterparts. A major bottleneck is obtaining sufficient quantities of material for crystallization trials. Early crystallographic success was possible with proteins that were naturally enriched in biological membranes (Figure 2, Table 1), but this approach is obviously not applicable for the majority of MPs. A more viable approach is the production of recombinant proteins using overexpression-hosts such as *Escherichia*

coli. This approach has already had a significant impact, providing material for many structures in recent years (Figure 2). To date, however, recombinant overexpression has worked mainly for prokaryotic MPs, as eukaryotic MPs are still very difficult to overexpress (reviewed in Drew et al. 2003, Wagner et al. 2006). Despite considerable effort, only 5 eukaryotic MP structures have been obtained using material produced by recombinant techniques: three were overexpressed in *Pichia pastoris* and two were overexpressed in *E. coli*. Hopefully, this is the beginning of a positive trend.

Given the difficulties of overexpressing eukaryotic MPs, it seems as if the prokaryotic pathway is currently the most viable way forward in our aim for a structural understanding of MPs. The critical question then becomes: how far can prokaryotic MPs take us towards the ultimate goal of understanding eukaryotic/human MPs?

Eukaryotic/human MPs with prokaryotic homologs

To answer this question, we have searched for MP families that are shared between prokaryotes and eukaryotes in a recently compiled database (Oberai et al. 2006). This database contains 3961 MP families derived from the fully sequenced genomes of 13 eukaryotic and 82 bacterial/archaeal organisms (see Box 1). It is obvious at a first glance that prokaryotic membrane proteomes are poor models for eukaryotic organisms, as only 256 families are common between the two groups, representing a mere ~13% of the eukaryotic MP families (and ~14% of the human MP families) (Figure 3). On the positive side, a closer look reveals that while these so-called 'universal families' are few, they are on average larger than 'unique families'. Given their ubiquity in nature, these omnipresent families are of particular interest to the scientific community, and have therefore been preferentially targeted by the structural genomics community (Table 1).

Although they are hotly pursued, there are still quite a few (>200) universal families lacking a descriptive structure, and in these cases a prokaryotic MP may shed light on the structure and function of a eukaryotic/human homolog (see Box 2). These families include ABC transporter #1 (amino acid, phosphate, ferric, nitrate, nickel, taurine,

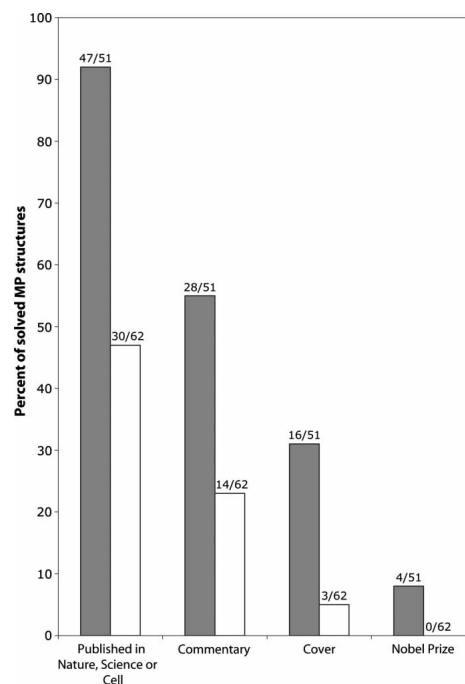


Figure 1: Membrane protein structures are hot. Analysis of high-resolution membrane protein structures collected by Stephen White (http://blanco.biomol.uci.edu/Membrane_proteins_xtal.html) indicates that novel structures (black bars) are often published in high-impact journals (i.e. Nature, Science or Cell) and accompanied by 'high-profile accessories' (i.e., commentaries and covers). Surprisingly, improved or follow-up structures (white bars) also maintain impact, albeit less than novel structures.

sugar), Transporter #1 (cationic acid, aromatic amino acid, choline K^+ uptake) and Transporter #2 (drug/metabolite, amino acid transport). With the ever-increasing genome-wide screenings of orthologous proteins for crystallization trials, structures for these families can be anticipated in the near future.

However, in our struggle to obtain a detailed structural understanding of the eukaryotic MPs, it is clear that prokaryotic MPs will only take us to 'first base' (see also Fleishman et al. 2006). Approximately 85% of the eukaryotic families do not have a prokaryotic homolog, so the structures of these proteins will need to be solved from eukaryotic sources (Figure 3). Further, prokaryotic MPs can ultimately only serve as models for eukaryotic/human proteins, and eventually there will be a desire to solve the structure for the actual protein of interest. This tells us that in the long term, we must solve the overexpression-bottleneck for eukaryotic MPs. Until then we might have to make the most of the natural sources.

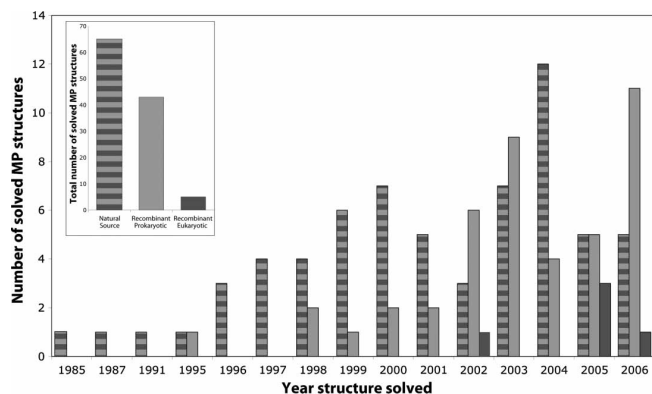


Figure 2: Eukaryotic membrane proteins produced by recombinant techniques are scarce in structural biology. Analysis of high resolution membrane protein structures collected by Stephen White (http://blanco.biomol.uci.edu/Membrane_proteins_xtal.html) indicates that most structures have been solved from proteins which have been purified from natural sources (striped bar), see inset. Of these, 44% were prokaryotic and 56% eukaryotic. Another successful approach is the production of recombinant prokaryotic proteins (light grey bar), see inset. In almost all of these cases the proteins have been produced in *E. coli* (not shown). Eukaryotic membrane proteins have not been successfully produced in the structural biology community to date (dark grey bar). Closer inspection indicates that most of the early structures were solved from proteins purified from natural sources, whereas recombinant technology arrived more recently. We are grateful to Niek Dekker (AstraZeneca, Sweden) for assistance with this Figure.

Box 1. Membrane proteins: what is a family and what is a fold? According to the Structural Classification of Proteins (SCOP), two proteins belong to the same *fold* if they are similarly ordered in three dimensions, i.e., if their secondary structures match and they have a similar topological arrangement (Murzin et al. 1995). Although two proteins share the same fold, they do not have to be evolutionary related. A deeper level of structural similarity is achieved on the *family* level: two proteins belong to the same family if they are evolutionary related, in which case they typically have a pairwise sequence identity of $\geq 30\%$. Currently, SCOP (release 1.71) contains 941 different folds and 3004 families. These numbers are based on proteins with known three-dimensional structures. Some 34 folds and 44 families contain α -helical MPs. Notably, for both soluble and membrane-integrated proteins, the majority of families can be attributed to a few folds, while there is large number of folds that comprise

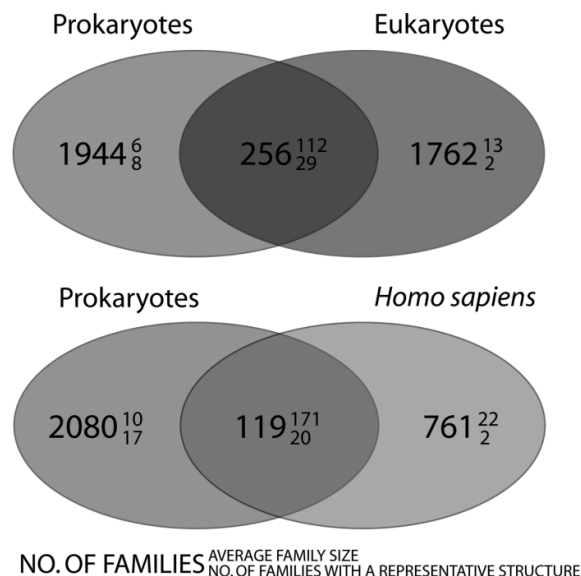


Figure 3: Venn diagrams, showing the distribution of membrane protein families. The upper panel shows the distribution of families between prokaryotes and eukaryotes and the lower panel shows the distribution of families between prokaryotes and *Homo sapiens*. The large number denotes the number of families that are common to that category. The number in superscript denotes the average family size and subscript, how many families in that category contain at least one member with known structure. It is evident that few are common to both prokaryotes and eukaryotes/*Homo sapiens*. These families are however very large (i.e., conserved across many species).

only a handful of families (Govindarajan et al. 1999, Ubarretxena-Belandia & Engelman 2001, Oberai et al. 2006).

Since many protein classification resources (such as SCOP) only classify MPs with known three-dimensional structures, and owing to the scarcity of structures, several groups have attempted to make projections about the size of the structural space of MPs, based on predicted transmembrane regions and sequence similarity (Jones 1998, Martin-Galiano & Frishman 2006, Oberai et al. 2006). Oberai and coworkers conclude that in order to have 80% of all MPs assigned to a fold or family, ~300 folds and ~700 families are required. They estimate that if the number of MP structures increases exponentially, as predicted by Stephen White (White 2004), this goal should be reached somewhere between the years 2020 and 2034. In a similar study by Martin-Galiano and Frishman (2006), 24 of 266 MP sequence clusters (corresponding to folds) already contain at least one structure, and these clusters cover approximately 70% of all MPs. However, this study was based on prokaryotic MPs only.

Box 2. 3D structure modelling of membrane proteins How far can the structure of a prokaryotic protein take us towards an understanding of a eukaryotic/human homolog? In a recent study (Forrest et al. 2006), the authors assessed how successful homology modelling is for 8 different MP families. It was possible to create acceptable models (the C_{α} Root Mean Square Deviation (RMSD) $< 2 \text{ \AA}$ compared to the native structure) if the sequence identity was above 30%. One of the limitations when assessing the quality of homology modeling of membrane proteins is that there are not very many families that contain more than two structures and hence that can be used as test cases.

But one should not despair: if there is no suitable target structure to begin the modeling from, recent progress of *de novo* structure prediction methods show some promise (Yarov-Yarovoy et al. 2006). Twelve multipass membrane proteins, both fragments and full structures, were modeled by the Rosetta-Membrane method and the corresponding predictions often had significant regions with an RMSD within 4 \AA from the native struc-

ture. The results were comparable to the accuracy of low-resolution predictions made for watersoluble proteins of the same length. However, Rosetta-Membrane is not yet able to create the full-atom models of the proteins needed for docking studies.

Acknowledgements

We thank Drs Amit Oberai and James U. Bowie for kindly providing data. This work was supported by grants from the Swedish Research Council, the Marianne and Marcus Wallenberg Foundation, and the Swedish Foundation for Strategic Research to GvH and DOD.

References

- Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. 2002. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**:899-907.
- Drew D, Froderberg L, Baars L, de Gier JW. 2003. Assembly and overexpression of membrane proteins in *Escherichia coli*. *Biochim Biophys Acta* **1610**:3-10.
- Fleishman SJ, Unger VM, Ben-Tal N. 2006. Transmembrane protein structures without X-rays. *Trends Biochem Sci* **31**:106-113.
- Forrest LR, Tang CL, Honig B. 2006. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J* **91**:508-517.
- Govindarajan S, Recabarren R, Goldstein RA. 1999. Estimating the total number of protein folds. *Proteins* **35**:408-414.
- Granseth E, Daley DO, Rapp M, Melen K, von Heijne G. 2005. Experimentally constrained topology models for 51,208 bacterial inner membrane proteins. *J Mol Biol* **352**:489-494.
- Jones DT. 1998. Do transmembrane protein superfolds exist? *FEBS Lett* **423**:281-285.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**:567-580.
- Martin-Galiano AJ, Frishman D. 2006. Defining the fold space of membrane proteins: the CAMPS database. *Proteins* **64**:906-922.
- Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**:536-540.

Oberai A, Ihm Y, Kim S, Bowie JU. 2006. A limited universe of membrane protein families and folds. *Protein Sci* **15**:1723-1734.

Ubarretxena-Belandia I, Engelman DM. 2001. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr Opin Struct Biol* **11**:370-376.

Wagner S, Bader ML, Drew D, de Gier JW. 2006. Rationalizing membrane protein overexpression. *Trends Biotechnol* **24**:364-371.

Wallin E, von Heijne G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Prot Sci* **7**(4):1029-1038.

White SH. 2004. The progress of membrane protein structure determination. *Protein Sci* **13**:1948-1949.

Yarov-Yarovoy V, Schonbrun J, Baker D. 2006. Multipass membrane protein structure prediction using Rosetta. *Proteins* **62**:1010-1025.

Table 1: Structural status of the 25 largest MP families. Proteins collected by Stephen White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html).

Family number	Present in <i>Homo sapiens</i>	Present in prokarya	Structure available	Family name
1	*		*	GPCR #1 (rhodopsin like)
2	*	*	*	Major facilitator
3		*		ABC transporter #1 (amino acid, phosphate, ferric, nitrate, nickel, taurine, sugar)
4				GPCR #2 (serpentine)
5	*	*	*	ABC transporter #2 (multidrug resistance)
6	*	*	*	Potassium channel
7	-	-	-	^a Receptor protein kinase
8	*	*		Transporter #1 (cationic acid, aromatic amino acid, choline, K ⁺ uptake)
9	*	*	*	Cytochrome <i>b</i> #1 (N-terminal part)
10	*	*	*	Cytochrome <i>b</i> #2 (C-terminal part)
11		*	*	ABC transporter #3 (branched amino acid, sugar, iron, zinc, manganese)
12	*	*	*	ATPase transporter
13	*	*		Transporter #2 (drug/metabolite, amino acid transport)
14	*	*	*	Transporter #3 (SecF protein export, cation, multidrug efflux)
15	*	*		NADH ubiquinone oxidoreductase
16	-	-	-	^a Cytochrome P450
17	*	*		Transporter #4 (arsenic efflux, Na ⁺ /H ⁺ antiporter, gluconate, mannitol, Na ⁺ /sulfate symporter, Mg ²⁺ /citrate transporter)
18	*	*	*	Cytochrome <i>c</i> oxidase
19		*		ABC transporter #4 (heme, O-Ag, multidrug resistance, antibiotic resistance, polysaccharide export)
20	*	*		Transporter #5 (DNA damage, inducible protein, virulence factor, polysaccharide export protein, Na ⁺ driven multidrug efflux)
21	*		*	^b Acetylcholine receptor
22	*	*		Adenylate cyclase
23	*	*		Permease (purines, pyrimidines, sulfate)
24		*		Acetyl transferase
25	*			Glutamate receptor

^aMight be monotopic membrane families, according to Oberai et al. (2006).

^bThe available structure from this protein family is of prokaryotic origin; The protein family is however not present in any of the prokaryotic organisms used in this study.