BIOINFORMATICS DISCOVERY NOTE

Vol. 29 no. 13 2013, pages 1589–1592
doi:10.1093/bioinformatics/btt249

Structural bioinformatics

Advance Access publication May 14, 2013

Relation between sequence and structure in membrane proteins

Mireia Olivella[1,*], Angel Gonzalez[2], Leonardo Pardo[2] and Xavier Deupi[2,†,*]

[1]Grup de Recerca en Bioinformàtica i Estadística Mèdica, Departament de Biologia de Sistemes, Escola Politècnica Superior, Universitat de Vic, 08500 Vic, Barcelona, Catalonia, Spain and [2]Laboratori de Medicina Computacional, Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Integral polytopic membrane proteins contain only two types of folds in their transmembrane domains: α-helix bundles and β-barrels. The increasing number of available crystal structures of these proteins permits an initial estimation of how sequence variability affects the structure conservation in their transmembrane domains. We, thus, aim to determine the pairwise sequence identity necessary to maintain the transmembrane molecular architectures compatible with the hydrophobic nature of the lipid bilayer.

Results: Root-mean-square deviation (rmsd) and sequence identity were calculated from the structural alignments of pairs of homologous polytopic membrane proteins sharing the same fold. Analysis of these data reveals that transmembrane segment pairs with sequence identity in the so-called 'twilight zone' (20–35%) display high-structural similarity (rmsd <1.5 Å). Moreover, a large group of β-barrel pairs with low-sequence identity (<20%) still maintain a close structural similarity (rmsd <2.5 Å). Thus, we conclude that fold preservation in transmembrane regions requires less sequence conservation than for globular proteins. These findings have direct implications in homology modeling of evolutionary-related membrane proteins.

Contact: Mireia.Olivella@uvic.cat or Xavier.Deupi@psi.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 27, 2013; revised on April 9, 2013; accepted on April 26, 2013

1 INTRODUCTION

Integral polytopic membrane proteins mediate the interaction of the cell with its surroundings, being involved in multiple cellular processes, as selective molecular transport, signaling, respiration and motility. Because of their relevance to cellular physiology and their accessibility from the extracellular environment, membrane proteins represent a significant portion of therapeutic drug targets (Arinaminpathy *et al.*, 2009; Hopkins and Groom, 2002). Particularly G protein-coupled receptors, transport proteins and ion channels are among the most prominent target families for the pharmaceutical industry.

Although membrane proteins represent ~20–30% of all proteins in sequenced genomes (Krogh *et al.*, 2001), only 2% of crystal structures deposited in the Protein Data Bank are membrane proteins (Tusnady *et al.*, 2005*a*), mainly because of the difficulty of their overexpression, purification and crystallization (Bill *et al.*, 2011).

Thus, because of the limited high-resolution structural information on membrane proteins, computational techniques to predict their three-dimensional (3D) structure from the amino acid sequence are a valuable tool (Pieper *et al.*, 2013). Recently, *de novo* techniques using evolutionary constraints have been applied to predict 3D structures for transmembrane (TM) proteins (Nugent and Jones, 2012). Homology models of proteins with unknown experimental structure can also be built from homologous proteins of known structure and similar sequence (templates). This method is based on the fact that in homologous proteins, structure is more conserved than sequence. In general, homologous proteins with a sequence identity >35% have a similar 3D structure. This similarity is less common in pairs of homologous proteins with sequence identity in the 'twilight zone' (threshold of 20–35%) (Chothia and Lesk, 1986; Krissinel and Henrick, 2004; Rost, 1999). Although these conclusions were achieved from the analysis of crystal structures of soluble proteins, homology modeling methods are also appropriate for membrane proteins, obtaining rmsd <2 Å relative to the native structure in the TM region for sequence identities of $\geq$30% (Forrest *et al.*, 2006). However, because of the relative scarcity of reference structures, these methods are frequently applied using templates with sequence identity below the 'twilight zone', with reasonable results (see, for instance, Callebaut *et al.*, 2006; Engel and Stahlberg, 2002; Patny *et al.*, 2006; Sansom *et al.*, 2002). This is particularly true for the core of the TM regions and has inspired new template-based coordinate generation protocols for membrane proteins (Kelm *et al.*, 2010).

This apparent difference between membrane and globular proteins probably arises because of their different environment. The lipid bilayer imposes a physical constraint that limits the number of folds that polypeptide chains can adopt when inserted in a membrane. These include α-helix bundles in bacterial, archaeal and eukaryotic cells and β-barrels in the outer membrane of bacteria, mitochondria and chloroplasts (Bowie, 2005; Wimley, 2003). These two folds contain secondary structure elements that maximize the hydrogen bond interactions among backbone atoms, whereas hydrophobic side chains are preferentially oriented toward the membrane lipids. As a consequence, many TM proteins share similar structural arrangements, even with marginal sequence identities (Gonzalez *et al.*, 2012; Sansom *et al.*, 2002), suggesting that relatively few conserved residues are sufficient to determine the molecular architecture of a particular TM fold.

*To whom correspondence should be addressed.
†Present address: Condensed Matter Theory Group and Laboratory of Biomolecular Research, Paul Scherrer Institut, Villigen PSI, Switzeland.

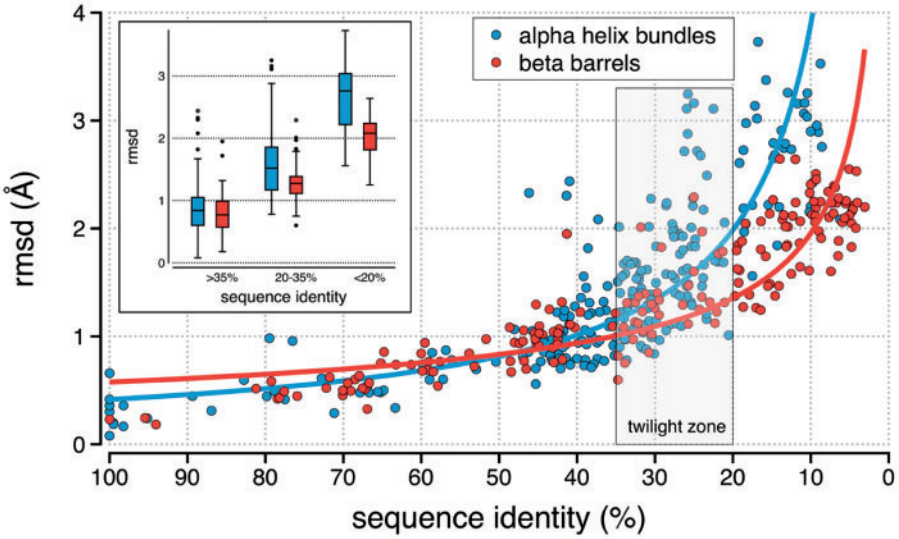

Fig. 1. Relation between sequence identity and rmsd calculated from the superimposition of the backbone α-carbon atoms in TM segments of homologous pairs of membrane proteins, containing both α-helix bundles (blue) and β-barrels (red). Power functions that fit sequence identity and rmsd for α-helix bundles (blue line; $y = 36.5x^{-0.97}$, $R^2 = 0.69$) and β-barrels (red line; $y = 6.6x^{-0.53}$, $R^2 = 0.72$) are also shown. At low-sequence identities, β-barrel domains seem to have a closer structural similarity than α-helix bundles

The present study aims to test the hypothesis that computational structural biologists have been implicitly using when building homology models of membrane proteins using templates of low-sequence identity: the fold in membrane proteins is less dependent on sequence variability than in globular proteins. By analyzing the relationship between structure and sequence in a database of membrane protein crystal structures, we show that fold preservation in the TM region of membrane proteins requires a lower degree of sequence conservation than in globular proteins.

2 METHODS

2.1 Membrane protein dataset

The coordinates of polytopic TM proteins with three or more homologous structures and resolution <4.0 Å were obtained from the Protein Data Bank (Berman *et al.*, 2000). Selected proteins were classified according to the SCOP (Murzin *et al.*, 1995) and OPM (Lomize *et al.*, 2006) databases and include receptors, energy transfer molecules, transporters and channels from different phyla. The native inactive state (i.e. without mutations or activating ligands) was selected for those proteins with more than one structure available. A total of 159 membrane proteins (111 α-helix bundles and 48 β-barrels) representing 25 different families were analyzed (Supplementary Table S1). This resulted in a comparison of 432 pairs (250 in α-helix bundles and 182 in β-barrels) of homologous TM protein subunits.

2.2 Determination of the transmembrane region

There exist several methods to annotate the membrane-spanning elements of TM proteins. We tested the TMDET (Tusnady *et al.*, 2005b) and OPM (Lomize *et al.*, 2006) algorithms, which yielded similar results when applied to our dataset. We observed that in the crystal structures of G protein-coupled receptors, some of the helices that form the TM bundle extend beyond the presumed boundaries of the membrane toward the cytoplasm. In such cases, we decided to extend the definition of TM segment obtained by TMDET to include these helical regions. To remove redundancy in our dataset, we selected one of the TM subunits as representative for homo-multimeric complexes.

2.3 Structural and sequence alignment

A pairwise structure alignment between members of each protein family was carried out using the Secondary Structure Matching (SSM) algorithm (Krissinel and Henrick, 2004). From the structural alignments, we obtained root-mean-square deviations (rmsd) of the backbone α-carbon atoms and sequence identity, as the fraction of identical residues in the total number of (structurally) aligned residues. Structural alignments were performed separately for either the entire protein (i.e. considering TM and less structured non-TM segments) or the TM domain.

3 RESULTS AND DISCUSSION

3.1 Relation between sequence and structure in the transmembrane domains of membrane proteins

Figure 1 and Supplementary Figure S1 show the relation between sequence identity and rmsd calculated from the pairwise structure alignments of homologous pairs of polytopic membrane proteins (see Section 2). In the inset, the data are divided into three categories: pairs of high- (>35%), medium- (20–35%) and low- (<20%) sequence identity. Clearly, pairs of membrane proteins with high-sequence identity have highly similar structures, with rmsd values <1 Å (0.89 ± 0.43 Å and 0.80 ± 0.32 Å for α-helix bundles and β-barrels, respectively). Protein pairs with medium sequence identities in the 'twilight zone', as defined for globular proteins (Rost, 1999), also display high-structural similarity, with rmsd values <2.0 Å (1.59 ± 0.55 Å and 1.30 ± 0.35 Å). Obviously, as sequences diverge, structures become more dissimilar. However, a large group of protein pairs with low-sequence identity still maintains a close structural similarity, with rmsd values <3 Å (2.69 ± 0.51 Å and 2.00 ± 0.34 Å).

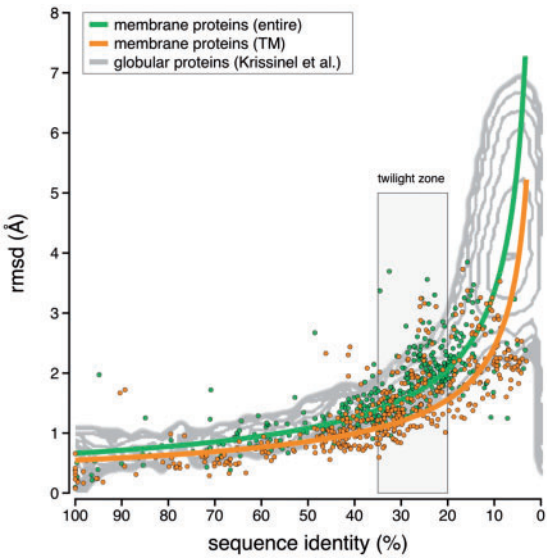

Fig. 2. Comparison of sequence identity and rmsd for membrane and globular proteins. The curves fit rmsd and sequence identity values for entire membrane proteins (in green; $y = 17.0x^{-0.70}$, $R^2 = 0.63$) and for TM segments (in orange; $y = 10.8x^{-0.65}$, $R^2 = 0.61$). Both curves are superimposed to data for globular proteins represented as contour maps of the reduced density of probability of obtaining 3D alignments with the corresponding rmsd (Krissinel and Henrick, 2004). Transmembrane segments present higher structural similarity (lower rmsd) than globular proteins at low values of sequence identity (<40%)

These rmsd values do not seem to depend on the number of transmembrane helices and strands (Supplementary Fig. S2). These results highlight the strong conservation of the TM fold even at low-sequence identities. This way, homology models of TM regions in evolutionary-related proteins sharing low-sequence identities (<20%) can result in acceptable 3D molecular models (rmsd < 3 Å).

3.2 Comparison of α-helix and β-barrel transmembrane domains

The power curves in Figure 1 fit the data calculated from the TM domains of α-helix bundles (blue line) and β-barrels (red line). The relation between sequence and structure in these two types of folds follows slightly different trends for sequence identities below the twilight zone. Clearly, β-barrel domains display on average a closer structural similarity at low-sequence identities (<30%). Therefore, β-barrel architectures seem to be more robust to sequence variations than α-helix bundles (Williams and Lovell, 2009).

3.3 Comparison of membrane and globular proteins

Figure 2 shows the relation between sequence identity and rmsd of membrane protein domains, both α-helix and β-barrels [considering only the TM domains, in orange or the entire (TM and non-TM segments) protein, in green] compared with a contour map of probability density calculated for globular proteins (Krissinel and Henrick, 2004). The plot shows that when considering only TM regions (orange), pairs of membrane proteins display lower rmsd values than pairs of globular proteins, particularly in the low-residue conservation range (<40%). Although in globular proteins, or in entire membrane protein domains (i.e. considering TM- and non-TM segments), a high-structural similarity (<2 Å) requires sequence identities of ~20% (Chothia and Lesk, 1986; Krissinel and Henrick, 2004), a similar degree of structural similarity within TM domains can be achieved with ~10% of sequence identity. This supports the idea that TM templates of relatively low-sequence identity can be used for homology modeling of the TM regions of membrane proteins.

3.4 Comparison of transmembrane and non-transmembrane segments

Rmsd values calculated from the superimposition of the backbone α-carbon atoms of the entire protein domain (containing both water-exposed and membrane-embedded regions) are on average larger than the rmsd values calculated exclusively for the TM segments (Fig. 2). This observation reflects the fact that the structure of membrane-embedded regions is more conserved than in water-exposed domains. Most probably, the high-structural conservation of the TM core preserves a conserved functional mechanism, whereas the variable solvent-exposed regions are responsible for the specificity of a wide range of extracellular stimuli and comprise loops with non-conserved structure. Thus, because amino acid substitutions at the TM region have less influence in structure than those in solvent accessible regions, homology modeling of the TM domains can be far more accurate that the outer structural elements (Forrest *et al.*, 2006).

4 CONCLUSIONS

In this work, we have assessed the relation between sequence identity and structure conservation in a dataset of polytopic membrane proteins. Our analysis shows that the TM regions of membrane proteins present a high-structure similarity (rmsd = 1–2 Å) in the twilight zone (20–35% sequence identity). The degree of structure similarity differs between α-helix bundles and β-barrels (Fig. 1). Comparison of our results with similar studies in globular proteins (Chothia and Lesk, 1986; Krissinel and Henrick, 2004; Rost, 1999) shows that the TM region of membrane proteins presents a higher degree of structural similarity than globular proteins, particularly in the twilight zone (Fig. 2). Moreover, in contrast to globular proteins, a significant set of membrane proteins maintains a strong conservation of the TM structure even at low-sequence identity (<20%). This finding suggests that it is possible to obtain relatively accurate 3D models of the TM regions of membrane proteins by homology modeling techniques even at low-sequence identities. In this regard, there are several examples in the literature that support such applicability (Blattermann *et al.*, 2012; Callebaut *et al.*, 2006; Engel and Stahlberg, 2002; Patny *et al.*, 2006; Sansom *et al.*, 2002; Zeth, 2010).

Funding: Ministerio de Ciencia e Innovación [SAF2010- 22198-C02-02] to L.P.; Instituto de Salud Carlos III [RD07/0067/0008] to L.P.; Swiss National Science Foundation (SNSF) [31003A_132815] to X.D.; ETH Zürich within the framework of the National Center for Competence in Research in Structural

Biology Program to X.D.; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2009SGR-581] to M.O.

Conflict of Interest: none declared.

REFERENCES

Arinaminpathy,Y. *et al.* (2009) Computational analysis of membrane proteins: the largest class of drug targets. *Drug. Discov. Today*, **14**, 1130–1135.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bill,R.M. *et al.* (2011) Overcoming barriers to membrane protein structure determination. *Nat. Biotechnol.*, **29**, 335–340.

Blattermann,S. *et al.* (2012) A biased ligand for OXE-R uncouples Galpha and Gbetagamma signaling within a heterotrimer. *Nat. Chem. Biol.*, **8**, 631–638.

Bowie,J.U. (2005) Solving the membrane protein folding problem. *Nature*, **438**, 581–589.

Callebaut,I. *et al.* (2006) Hydrophobic cluster analysis and modeling of the human Rh protein three-dimensional structures. *Transfus. Clin. Biol.*, **13**, 70–84.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Engel,A. and Stahlberg,H. (2002) Aquaglyceroporins: channel proteins with a conserved core, multiple functions, and variable surfaces. *Int. Rev. Cytol.*, **215**, 75–104.

Forrest,L.R. *et al.* (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.*, **91**, 508–517.

Gonzalez,A. *et al.* (2012) Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors. *Chem. Bio. Chem.*, **13**, 1393–1399.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug. Discov.*, **1**, 727–730.

Kelm,S. *et al.* (2010) MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics*, **26**, 2833–2840.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lomize,M.A. *et al.* (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nugent,T. and Jones,D.T. (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA*, **109**, E1540–E1547.

Patny,A. *et al.* (2006) Homology modeling of G-protein-coupled receptors and implications in drug design. *Curr. Med. Chem.*, **13**, 1667–1691.

Pieper,U. *et al.* (2013) Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nat. Struct. Mol. Biol.*, **20**, 135–138.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sansom,M.S. *et al.* (2002) Potassium channels: structures, models, simulations. *Biochim. Biophys. Acta*, **1565**, 294–307.

Tusnady,G.E. *et al.* (2005a) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.

Tusnady,G.E. *et al.* (2005b) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.

Williams,S.G. and Lovell,S.C. (2009) The effect of sequence evolution on protein structural divergence. *Mol. Biol. Evol.*, **26**, 1055–1065.

Wimley,W.C. (2003) The versatile beta-barrel membrane protein. *Curr. Opin. Struct. Biol.*, **13**, 404–411.

Zeth,K. (2010) Structure and evolution of mitochondrial outer membrane proteins of beta-barrel topology. *Biochim. Biophys. Acta*, **1797**, 1292–1299.