Relation between sequence and structure in membrane proteins

Mireia Olivella¹,*, Angel Gonzalez², Leonardo Pardo² and Xavier Deupi²,†,*

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

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 < 1.5Å). Moreover, a large group of β-barrel pairs with low-sequence identity (<20%) still maintain a close structural similarity (rmsd < 2.5Å). 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 probably arise 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, archaegal and eukaryotic cells and β-barrels in the outer membrane of bacteria, mitochondria and chloroplasts (Bowie, 2005; Winley, 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.

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 < 2Å relative to the native structure in the TM region for sequence identities of >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; Pattny 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, archaegal and eukaryotic cells and β-barrels in the outer membrane of bacteria, mitochondria and chloroplasts (Bowie, 2005; Winley, 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.
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 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.
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 than 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 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.
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


