Chapter 5
Advances in Computational Methods for Transmembrane Protein Structure Prediction

Tim Nugent, David Jones and Sikander Hayat

Abstract Transmembrane (TM) proteins fulfill many crucial cellular functions such as substrate transport, biogenesis and signalling, and make up a significant fraction of any given proteome. Estimates suggest that up to 30% of all human genes may encode α-helical TM proteins, while β-barrel TM proteins, which are found in the outer-membrane of gram-negative bacteria, mitochondria and chloroplast, are encoded by 2–3% of genes. However, relatively few high resolution TM protein structures are known, making it all the more important to extract as much structural information as possible from amino acid sequences. In this chapter, we review the existing methods for the identification, topology prediction and three-dimensional modelling of TM proteins, including a discussion of the recent advances in identifying residue-residue contacts from large multiple sequence alignments that have enabled impressive gains to be made in the field of TM protein structure prediction.

Keywords Transmembrane proteins · Structure prediction · 3D modelling

T. Nugent (✉)
Thomson Reuters, Corporate Research and Development, 30 South Colonnade, Canary Wharf, EC2A 4EG London, UK
e-mail: tim.nugent@thomsonreuters.com

D. Jones
Bioinformatics Group, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
e-mail: d.jones@cs.ucl.ac.uk

S. Hayat
Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York City, USA
e-mail: hayats@mskcc.org

D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_5

5.1 Introduction

Transmembrane (TM) proteins are involved in a wide range of essential biological processes including cell signalling, transport of membrane-impermeable molecules, cell-cell communication, cell recognition, cell adhesion and biogenesis of the bacterial outer membrane. Many are also prime drug targets, with approximately 60% of all drugs currently on the market targeting membrane proteins (Hopkins and Groom 2002). Despite recent progress in TM protein structure determination, the experimental difficulties associated with obtaining crystals that diffract to high resolution mean that TM protein are severely under-represented in structural databases, making up only 1% of known structures in the PDB (White 2004) of which only about 500 are unique. TM proteins, which have both hydrophobic and hydrophilic regions on their surfaces, are much more difficult to isolate than water-soluble proteins as the native membrane surrounding the protein must be disrupted and replaced with detergent molecules without causing any denaturation. Given the biological and pharmacological importance of TM proteins, an understanding of their structure and topology—the total number of TM helices, their boundaries and in/out orientation relative to the membrane—is essential for functional analysis and directing further experimental work. In the absence of vital structural data, bioinformatics strategies thus turn to sequence-based prediction methods.

5.2 Membrane Protein Structural Classes

TM proteins can be classified into two basic types: α-helical and β-barrel proteins. α-helical membrane proteins form the major category of TM proteins and are present in all type of biological membranes, including bacterial outer membranes. They consist of one or more α-helices, each of which contains a stretch of hydrophobic amino acids, embedded in the membrane and linked to subsequent helices by extra-membranous loop regions. It is thought such proteins may have up to 20 TM helices allowing a diverse range of differing topologies. Loop regions are known to contain substructures including re-entrant loops—short α-helices that enter and exit the membrane on the same side—as well as amphipathic helices that lie parallel to the membrane plane, and globular domains. β-barrel TM proteins (TMBs) mainly consist of transmembrane β-strands that form a closed barrel in the membrane. Analysis of solved β-barrel 3D structures show that these proteins can consist of 8–26 β-strands arranged in an anti-parallel manner in the bacterial outer-membrane. Some TMBs also have large plug-domains and outer loops that can interact with the barrel region to control substrate transport.

5.2.1 α-Helical Bundles

α-helical TM proteins can be further divided into a number of subtypes based on their topology. Type I and II membrane proteins consist of a single TM α helix, type III have multiple membrane-spanning helices while type IV membrane proteins have multiple domains which form an assembly that spans the membrane multiple times. Type I membrane proteins are attached to the membrane with an anchor sequence targeting their amino terminus to the endoplasmic reticulum lumen and the carboxy terminus exposed to the cytoplasmic side. These proteins are further divided into two subtypes. Type Ia—which constitutes most eukaryotic membrane proteins—contain cleavable signal sequences, while type Ib do not. Type II membrane proteins are similar to type I in that they span the membrane only once but their orientation is reversed; they have their amino terminus on the cytoplasmic side of the cell and the carboxy terminus on the exterior. Type III membrane proteins, which include G protein coupled receptors (e.g. PDB code 1gzm) consist of multiple TM helices and are also divided into two subtypes. Type IIIa have cleavable signal sequence while type IIIb do not, but do have their amino terminus exposed to the extracellular side of the membrane. Type IV membrane proteins have multiple domains which form an assembly that spans the membrane multiple times. Domains may reside on a single polypeptide chain but are often composed of more than one. Examples include Photosystem I, which comprises nine unique chains (1jb0).

5.2.2 Transmembrane β-Barrels

TMBs can be divided into two main categories depending on whether the barrel pore is formed from a single-chain, or via a homo-oligomeric complex, with each chain contributing 2–4 strands. All known bacterial transmembrane β-barrels consist of anti-parallel β-strands that traverse the outer-membrane in a regular manner (Fig. 5.1). Residues on a transmembrane β-strand follow a strict-dyad repeat such that alternate side-chain face the lipids and barrel pore, respectively. The lipid-facing residues are mostly hydrophobic, but the pore-facing residues can be a mixture of both polar and hydrophobic amino acids. Moreover, transmembrane β-strands generally have fewer residues than transmembrane α-helices and have a less prominent hydrophobic profile. Residues on adjacent β-strands are hydrogen bonded to each other such that alternate residues on strand S1 form a N–O and O–N bond with residues in-register on strand S2, where S1 and S2 are adjacent strands. Solved 3D structures of bacterial TMBs have 8 to 26 β-strands, while the only known Eukaryotic TMB structure - mitochondrial voltage dependent anion channel (VDAC) has 19 strands, where the first and the last strand are parallel to each other. TMBs have long extra-cellular loops that generally protrude away from the barrel pore region but can interact with the barrel domain and short inner loops.

Additionally, a few TMBs have plug domains (Fig. 5.1) that sit inside the barrel and participate in gating and signaling (Ferguson et al. 2002). It is generally estimated that TMBs account for 2–3% of the genes in bacteria, but there is scope for improvement in accurately determining the number of yet unknown TMB families.

Multi-chain TMBs mainly fall into one of four known superfamilies—(a) the pore-forming toxins (PDB codes 3w9t, 3o44, 4h56, 3b07, 7ahl) that are secreted by pathogenic bacteria such as *Staphylococcus aureus*, *Clostridium perfringens* and *Vibrio cholerae*, (b) outer membrane efflux proteins (PDB codes 4mt4, 4mt0, 2xmn, 3pik, 1wp1, 1yc9, 1ek9) that are used by bacteria to expel a wide range of molecules including antibacterial drugs thereby increasing multi-drug resistance, (c) mycobacterial porins (PDB code 1uun) in Mycobacteria that can be used to transport drugs through an otherwise low-permeability outer membrane environment that renders them resistant to many antibiotics, and (d) trimeric autotransporters (PDB codes 2lme, 2gr7) such as the Hia autotransporter of *Haemophilus*

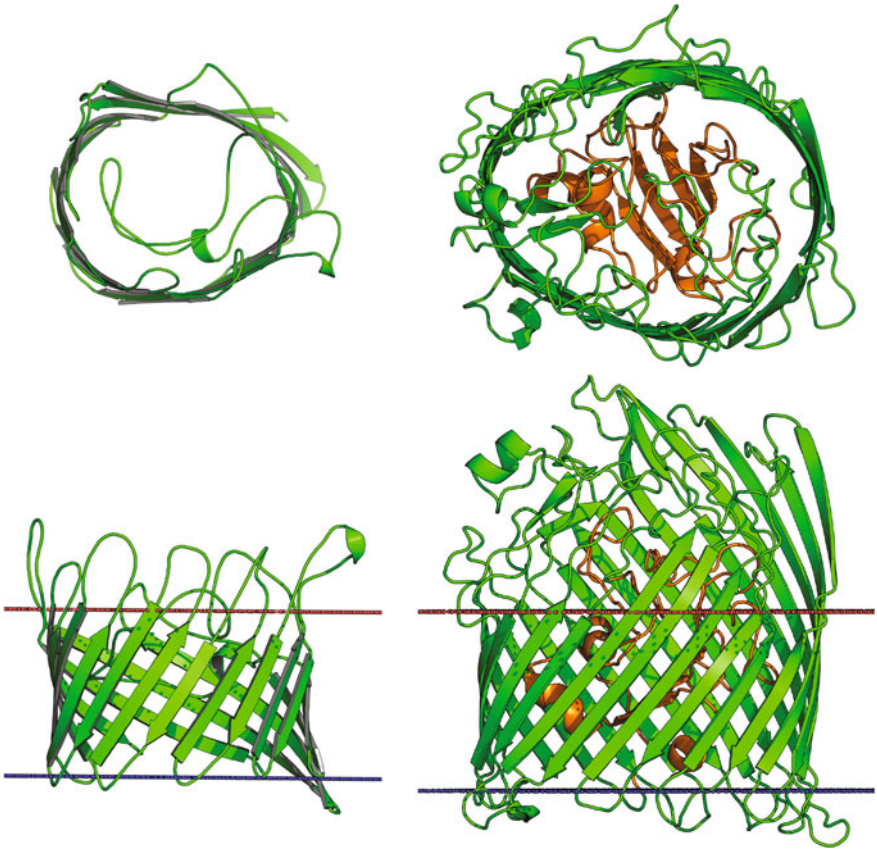

Fig. 5.1 Top and front views of a diffusion porin (PDB code 3prn) and outer membrane iron transporter FecA (PDB code 1kmp). Both proteins have long outer-loops. The large plug domain of FecA (*orange*) sits in the barrel and facilitates substrate transport and allosteric transitions

influenzae that belongs to the largest family of virulence proteins mediating bacterial adhesion, invasion and spread to host cells. Sequence-based analysis methods to identify protein sequences that belong to those families, and therefore estimate the number of multi-chain TMB families, are currently lacking. Additionally, better computational methods for their topology prediction and 3D assembly need to be developed to increase our understanding of their assembly mechanism and function.

5.3 Databases

There now exist a number of databases that serve as repositories for the sequences and structures of both α-helical and β-barrel TM proteins (Table 5.1). OPM (Lomize et al. 2006b, 2011), PDBTM (Tusnady et al. 2004, 2005a; Kozma et al. 2013), CGDB (Chetwynd et al. 2008) and the mpstruc database (http://blanco.biomol.uci.edu/mpstruc/) all contain TM proteins of known structure determined using X-ray and electron diffraction, nuclear magnetic resonance and cryo-electron microscopy. OPM, PDBTM and CGDB additionally contain orientation predictions of the protein relative to the membrane based on water-lipid transfer energy minimisation (Lomize et al. 2006a), hydrophobicity/structural feature analysis (Tusnady et al. 2005b) and coarse grained molecular dynamic simulations (Sansom et al. 2008), while MemProtMD (http://sbcb.bioch.ox.ac.uk/memprotmd/) contains orientations calculated using a knowledge-based statistical potential (Nugent and Jones 2013). TOPDB (Tusnady et al. 2008; Dobson et al. 2015a) and MPtopo (Jayasinghe et al. 2001) include topology data that has been experimentally validated using low-resolution techniques such as gene fusion, antibody and mutagenesis studies. Other TM protein databases tend to focus on specific families such as

Table 5.1 Transmembrane protein databases

Method	URL	Features
OPM	http://opm.phar.umich.edu/	Known structures
PDB_TM	http://pdbtm.enzim.hu/	Known structures
CGDB	http://sbcb.bioch.ox.ac.uk/cgdb/	Coarse grained simulations
MemProtMD	http://sbcb.bioch.ox.ac.uk/memprotmd/	Coarse grained simulations
TOPDB	http://topdb.enzim.hu/	Experimental validation
Mptopo	http://blanco.biomol.uci.edu/mptopo/	Experimental validation
VKCDB	http://vkcdb.biology.ualberta.ca/	Potassium channels
KDB	http://sbcb.bioch.ox.ac.uk/kdb/	Potassium channels
TCDB	http://www.tcdb.org/	Transporters
TMBB-DB	http://beta-barrel.tulane.edu/	Predicted TMBs
TMBETA-GENOME	http://tmbeta-genome.cbrc.jp/annotation	Predicted TMBs
OMPdb	http://bioinformatics.biol.uoa.gr/OMPdb	Predicted TMB families
HHomp	http://toolkit.tuebingen.mpg.de/hhomp	TMB remote homology detection

voltage-gated potassium channels, including VKCDB (Li and Gallin 2004; Gallin and Boutet 2011) and KDB (http://sbcb.bioch.ox.ac.uk/kdb/), while others such as the Transporter Classification Database (Saier et al. 2006, 2009, 2014) focus on particular structural or functional classes.

For TMBs, TMBB-DB (Freeman and Wimley 2012), TMBETA-GENOME (Gromiha et al. 2007) and OMPdb (Tsirigos et al. 2011) provide an exhaustive list of putative TMBs predicted using computational methods. In addition, HHomp (Remmert et al. 2009) provides a list of putative TMBs found by comprehensive, transitive homology search. As with all bioinformatics databases, care should be taken to ensure that a given resource is frequently updated. The rate at which new sequences and structures are deposited in GenBank and the PDB [and occasionally retracted e.g. (Chang et al. 2006)] results in significant manual annotation for database administrators, and much evidence suggests that this workload often exceeds the amount of time an administrator is willing to commit.

5.4 Multiple Sequence Alignments

As with globular proteins, multiple sequence alignments play an important role in TM protein structure prediction. Homologous sequences identified via database searches can be used to construct sequence profiles which can significantly enhance TM topology prediction accuracy (Henricson et al. 2005; Jones 2007), while recent co-evolution-based approaches (Jones et al. 2012, 2015) are dependent on high-quality alignments to infer residue-residue contacts which can be used for de novo modelling (Nugent and Jones 2012).

Conventional pair-wise alignment methods return possible matches based on a scoring function that relies on amino acid substitution matrices such as PAM (Dayhoff and Schwartz 1978) or BLOSUM (Henikoff and Henikoff 1992). Such matrices are derived from globular protein alignments, and as amino acid composition, hydrophobicity and conservation patterns differ between globular and TM proteins (Jones et al. 1994a), they are in principle unsuitable for TM protein alignment. A number of TM-specific substitution matrices have therefore been developed, which take into account such differences. For example, the JTT TM matrix (Jones et al. 1994b) was based on the observation that polar residues in TM proteins are highly conserved, while hydrophobic residues are more interchangeable. Other matrices such as SLIM (Muller et al. 2001), were reported to have the highest accuracy for detecting remote homologues in a manually curated GPCR dataset, while PHAT (Ng et al. 2000) has been shown to outperform JTT, especially on database searching.

More recently, a number of methods have been developed to improve actual TM protein alignment. HMAP (Tang et al. 2003) showed that alignment accuracy could be improved significantly using a profile-profile based approach incorporating structural information. STAM (Shafrir and Guy 2004) implemented higher penalties for insertion/deletions in TM segments compared to loop regions, with combinations of different substitution matrices to produce alignments resulting in more accurate

homology models. PRALINETM (Pirovano et al. 2008), which integrates state-of-the-art sequence prediction techniques with membrane-specific substitution matrices, was shown to outperform standard multiple alignment techniques such as ClustalW (Thompson et al. 1994) and MUSCLE (Edgar 2004) when tested on the TM alignment benchmark set within BAliBASE (Bahr et al. 2001). AlignMe (Stamm et al. 2014, 2013; Khafizov et al. 2010), which uses secondary structure matching combined with evolutionary information, also demonstrated high quality alignments when tested on BAliBASE, although it was noted that accuracy was generally lower when transmembrane topology predictions were also included, although the inclusion of this information may still be useful in cases of extremely distantly related proteins for which sequence information is less informative. PSI-Coffee—a modification of the T-Coffee method (Chang et al. 2012; Notredame et al. 2000)—employs a homology extension technique that can be used to reveal and use specific conservation patterns found within transmembrane proteins, such as amphiphilic α-helices, resulting in significant improvements to the accuracy of alignments. Hill and co-workers constructed substitution tables for different environments within membrane proteins, demonstrating that, in the 10–25% sequence identity range, alignments could be improved by an average of 28 correctly aligned residues compared with alignments made using default substitution tables, leading to improved structural models (Hill and Deane 2013; Hill et al. 2011).

For TMBs, Jimenez-Morales and Liang (2011) have estimated the evolutionary pattern of residue substitutions which can be useful for improved sequence alignment of TMBs, while Yan et al. (2011), have shown the utility of secondary structure element alignment for the identification of putative TMBs. Additionally, a structure based alignment method for TMBs that uses TMB-specific topology features has been shown to improve alignment (Wang et al. 2013).

5.5 Transmembrane Protein Topology Prediction

The under-representation of TM proteins in structural databases makes their study extremely difficult. As a result, tools to analyse TM proteins have historically focused on sequence-based topology prediction—identifying the total number of TM helices, their boundaries, and in/out orientation relative to the membrane. Experimental approaches for determining TM topology include glycosylation analysis, insertion tags, antibody studies and fusion protein constructs; however, such studies are time consuming, often conflicting (Mao et al. 2003; Kyttala et al. 2004; Ratajczak et al. 2014), and also risk upsetting the natural topology by altering the protein sequence. Theoretical prediction methods therefore provide an important strategy for furthering our understanding of these biological and pharmacological important proteins.

5.5.1 Early α-Helical Topology Prediction Approaches

Early topology predictions methods were based on physicochemical observations of TM proteins. Even before the arrival of the first crystal structures, stretches of hydrophobic residues long enough to span the lipid bilayer were identified as TM spanning α-helices. Prediction methods by Kyte and Doolittle (1982) and Engelman et al. (1986), and later by Wimley and White (1996), relied on experimentally determined hydropathy indices to create a hydropathy plot for a protein. This involved taking a sliding window of 19–21 residues and averaging the score with peaks in the plots (regions of high hydrophobicity) corresponding to the locations of TM helices. With more sequences came the discovery that aromatic Trp and Tyr residues tend to cluster near the ends of the transmembrane segments (Wallin et al. 1997), possibly acting as physical buffers to stabilise TM helices within the lipid bilayer. Later, studies identified the appearance of sequence motifs, such as the GxxxG motif (Senes et al. 2000), within TM helices and also periodic patterns implicated in helix-helix packing and 3D structure (Samatey et al. 1995). However, perhaps the most important realisation was that positively-charged residues tend to cluster on cytoplasmic loop—the 'positive-inside' rule of Gunnar von Heijne (von Heijne 1992). Combined with hydrophobicity-based prediction of TM helices, this led to early topology prediction methods such as TopPred (Claros and von Heijne 1994).

5.5.2 Machine Learning Approaches for α-Helical Topology Prediction

Despite their early success, these methods based on hydrophobicity analysis combined with the 'positive-inside' rule have since been superseded by machine learning approaches which offer substantially higher prediction accuracy due to their probabilistic formulation (Table 5.2). Hidden Markov models (HMMs) were among the first supervised learning algorithms to be applied to TM topology prediction, with both TMHMM (Krogh et al. 2001) and HMMTOP (Tusnady and Simon 1998) proving highly successful. TMHMM implemented a cyclic model with seven states for a TM helix, while HMMTOP used HMMs to distinguish between five structural states [helix core, inside loop, outside loop, helix caps (C and N) and globular domains]. These states were connected by transition probabilities before dynamic programming was used to match a sequence against a model with the most probable topology. HMMTOP also allowed constrained predictions to be made, where specific residues could be fixed to a topological location based on experimental data, as did other methods such as HMM-TM (Bagos et al. 2006). Later HMM-based predictors include PRODIV-TMHMM and PolyPhobius, both of which made use of evolutionary information from homologs resulting in substantially increased performance (Viklund and Elofsson 2004; Kall et al. 2005).

Table 5.2 Topology prediction methods for α-helical transmembrane proteins

Method	Features	URL
TMHMM (Krogh et al. 2001)	HMM	http://www.cbs.dtu.dk/services/TMHMM/
HMMTOP (Tusnady and Simon 1998)	HMM	http://www.enzim.hu/hmmtop/
HMM-TM (Bagos et al. 2006)	HMM	http://bioinformatics.biol.uoa.gr/HMM-TM/
PRODIV-TMHMM (Viklund and Elofsson 2004)	HMM + Evolutionary information	https://www.pdc.kth.se/hakanv/prodiv-tmhmm
Phobius (Kall et al. 2005)	HMM + Evolutionary information + Signal peptide prediction	http://phobius.sbc.su.se/
OCTOPUS (Viklund and Elofsson 2008)	HMM + NN + Evolutionary information	http://octopus.cbr.su.se/
SPOCTOPUS (Viklund and Elofsson 2008)	HMM + NN + Evolutionary information + Signal peptide prediction	http://octopus.cbr.su.se/
PHDhtm (Rost et al. 1996)	NN	https://www.predictprotein.org/
MEMSAT3 (Jones 2007)	NN + Evolutionary information + Signal peptide prediction	http://bioinf.cs.ucl.ac.uk/psipred/
MEMSAT-SVM (Nugent and Jones 2009)	SVM + Evolutionary information + Signal peptide prediction	http://bioinf.cs.ucl.ac.uk/psipred/
Philius (Reynolds et al. 2008)	Dynamic Bayesian networks	http://noble.gs.washington.edu/proj/philius/
WRF-TMH (Hayat and Khan 2013)	Random forests	http://111.68.99.218/WRF-TMH/
TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009)	Consensus	http://topcons.cbr.su.se/
CCTOP (Dobson et al. 2015b)	Consensus	http://cctop.enzim.ttk.mta.hu/

Neural networks (NNs) were employed by early methods including PHDhtm (Rost et al. 1996) and MEMSAT3 (Jones 2007). PHDhtm used multiple sequence alignments to perform a consensus prediction of TM helices by combining two NNs. The first created a 'sequence-to-structure' network, which represented the structural propensity of the central residue in a window. A 'structure-to-structure' network then smoothed these propensities to predict TM helices, before the 'positive-inside' rule was applied to produce an overall topology. MEMSAT3 uses a neural network and dynamic programming in order to predict not only TM helices, but also to score the topology and to identify possible signal peptides.

Additional evolutionary information provided by multiple sequence alignments led to prediction accuracies increasing to as much as 80%. OCTOPUS (Viklund and Elofsson 2008) used a novel combination of hidden Markov models and artificial neural networks to further increase performance.

Later, Support Vector Machines (SVMs) gained in popularity and were successfully applied to TM protein topology prediction (Yuan et al. 2004; Lo et al. 2006, 2008). Particularly using non-linear kernel functions, SVMs are capable of learning complex relationships among the amino acids within a given window with which they are trained, particularly when provided with evolutionary information, and are also more resilient to the problem of over-training compared to other machine learning methods. MEMSAT-SVM (Nugent and Jones 2009), an extension of MEMSAT3, used multiple SVM models to classify sequence into one of four states [TM helix, inside or outside loop, re-entrant helix, or signal peptide] before calculating the most likely topologies using dynamic programming, while a further SVM was used to discriminate between globular and TM proteins. Although multiclass SVMs do exist, their performance is typically poorer than binary SVMs since in many cases no single mathematical function exists to separate all classes of data from one another.

More recently, other machine learning algorithms have been applied to TM helix and topology prediction including dynamic Bayesian networks (Reynolds et al. 2008), random forests (Hayat and Khan 2013), self-organizing maps (Deng 2006) and deep learning (Qi et al. 2012). A selection of machine learning-based predictors can be found in Table 5.2.

5.5.3 Signal Peptides and Re-entrant Helices

One significant challenge faced by topology predictors is the discrimination between TM helices and other highly hydrophobic structural features. These include targeting motifs such as signal peptides and signal anchors, amphipathic helices, and re-entrant helices, membrane penetrating helices that enter and exit the membrane on the same side, common in many ion channel families (Fig. 5.2). The similarity between such features and the hydrophobic profile of a TM helix frequently leads to crossover between the different types of predictions. Should these elements be predicted as TM helices, the ensuing topology prediction is likely to be severely disrupted. Some prediction methods, such as SignalP (Petersen et al. 2011; Bendtsen et al. 2004) and TargetP (Emanuelsson et al. 2007), are effective in identifying signal peptides in TM proteins, and may be used as a pre-filter prior to analysis using a TM topology predictor. Phobius (Kall et al. 2004) used a HMM to successfully address the problem of signal peptides in TM protein topology prediction, while PolyPhobius (Kall et al. 2005) further increased accuracy by including homology information. Other methods such as MEMSAT-SVM, OCTOPUS and SPOCTOPUS (Viklund et al. 2008) have also attempted to incorporate identification of re-entrant regions and signal peptides into TM

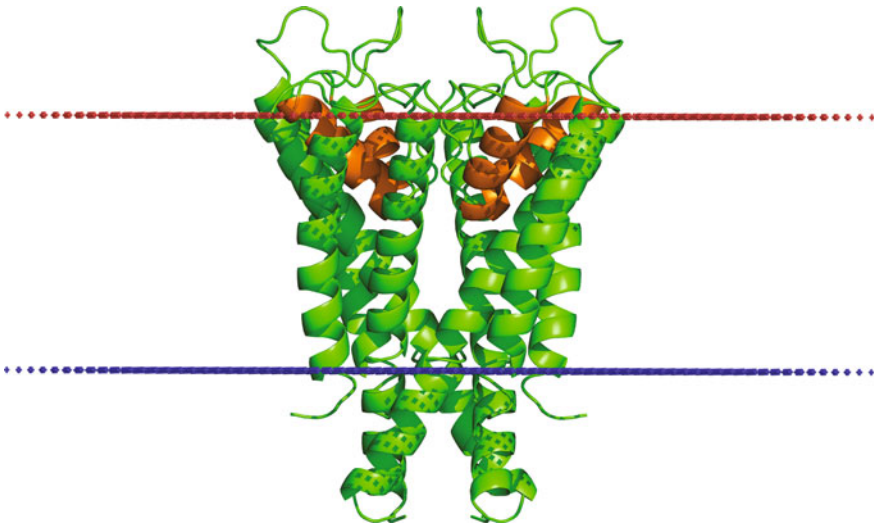

Fig. 5.2 Potassium channel KcsA (PDB code 1R3J). Each monomer of the homo-tetrameric complex consists of two TM helices and one re-entrant helix (*orange*), which surrounds the central pore and is involved in channel gating

topology prediction but there is significant room for improvement. The problem, particularly regarding re-entrant helices, is the lack of reliable data with which to train machine-learning based methods.

5.5.4 *Consensus Approaches for α-Helical Topology Prediction*

While a number of methods successfully combine multiple machine learning approaches, for example ENSEMBLE (Martelli et al. 2003) uses a NN and two HMMs while OCTOPUS uses two sets of four NNs and one HMM, perhaps the best overall methods are those which adopt a consensus approach by combining the results of several predictors to yield more reliable results. Early consensus predictors such as BPROMPT (Taylor et al. 2003) combined the outputs of five different predictors to produce an overall topology using a Bayesian belief network, while Nilsson et al. (2002) used a simple majority-vote approach to return the best topology from their five predictors. The PONGO server (Amico et al. 2006) returns the results of 5 high scoring methods in a graphical format for direct comparison. More recently, MetaTM (Klammer et al. 2009) is based on SVM models and combines the results of six TM topology predictors and two signal peptide predictors. TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009) combines a number of topology predictions into one consensus prediction, while also quantifying the reliability of the prediction based on

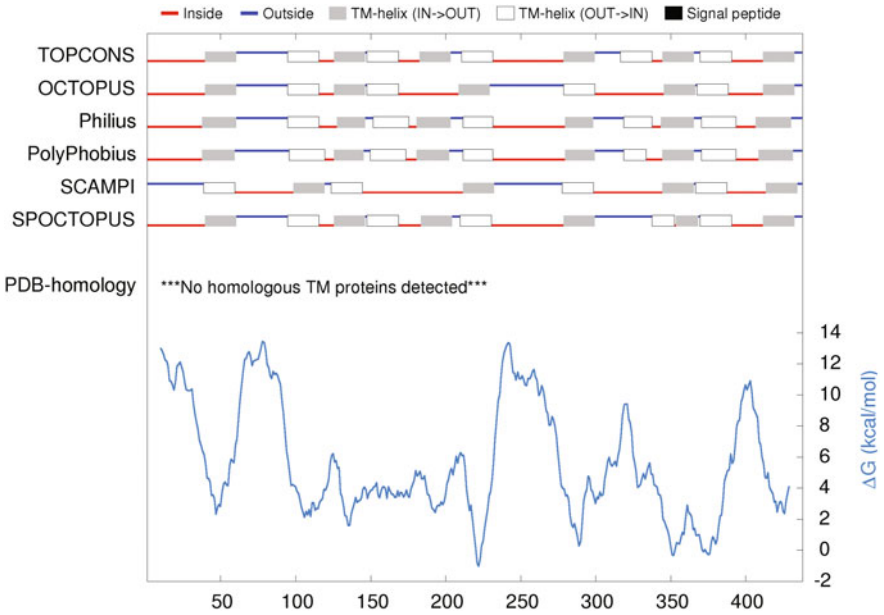

Fig. 5.3 Consensus topology prediction by TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009). The results from a number of individual predictors are combined to produce the TOPCONS prediction

the level of agreement between the underlying methods, both at the protein level and at the level of individual TM regions (Fig. 5.3). Results indicate an overall increase in performance by 4% compared to the currently available best-scoring methods. CCTOP (Dobson et al. 2015b) makes use of ten different topology prediction methods, while also incorporating topology information from existing experimental and computational resources such as the PDBTM, TOPDB and TOPDOM databases, using a HMM. In most cases, but particularly proteins whose topology is not straightforward, using a consensus-based method is highly advisable.

5.5.5 *Transmembrane β-Barrel Topology Prediction*

Topology prediction of TMBs entails the estimation of the number and the location of TM β-strands. Traditional methods based on a sliding-window hydrophobicity profile are not sufficiently accurate, most likely due to the shorter size and less prominent hydrophobic nature of the TM β-strands. This problem is further complicated by the presence of other β-sheet rich regions in full protein sequences such as the pre-barrel region (seen, for example, in *EstA* Autotransporter protein; PDB code 3kvn) and large plug domains that reside inside the barrel (as seen in *FecA*

Table 5.3 Computational methods for identifying transmembrane β-barrels

Method	Features	URL
boctopus + PSORTb (Imai et al. 2013)	Predicted topology + Subcellular localization	http://boctopus.cbr.su.se/
BETAWARE (Savojardo et al. 2013a)	N-to-1 Extreme Learning Machine	http://betaware.biocomp.unibo.it/BetAware
SSEA-OMP (Yan et al. 2011)	Secondary structure element alignment	http://protein.cau.edu.cn/SSEA-OMP/index.html
TMB-Hunt (Garrow et al. 2005)	K-nearest neighbor	http://bioinformatics.leeds.ac.uk/betaBarrel/
TMBETA-NET (Gromiha et al. 2005)	Amino acid composition + NN	http://psfs.cbrc.jp/tmbeta-net/
BOMP (Berven et al. 2004)	C-terminal pattern + Integral b-score	http://services.cbu.uib.no/tools/bomp
F-W barrel analyzer (Freeman and Wimley 2010)	Empirical Score	http://www.tulane.edu/biochem/WW/apps.html

protein; PDB code 1fep). Additionally, the absence of long stretches of hydrophobic residues makes it harder to distinguish TM β-strands from β-sheets in globular proteins. One strategy to predict the topology of TMBs relies on first predicting if the query sequence is a TMB or not (Table 5.3) and then using dedicated computational methods to predict the topology of sequences that are predicted to be TMBs. This can potentially improve the accuracy of computational methods that are based on learning from data points available from known 3D structures. Boctopus in combination with PSORTb (Imai et al. 2013), which is a bacterial subcellular localization tool, can be used to identify putative TMBs. The idea here is that proteins for which topology predictor methods predict at least 8 strands with predicted subcellular localization as 'outer-membrane' can be potential TMBs. BETAWARE (Savojardo et al. 2013a) is a machine learning based tool that predicts if a protein is TMB using N-to-1 network encoding and then predicts the topology using a constrained grammar. Other methods employ a combination of secondary structure features, hydrophobicity, amino acid composition and empirical scores to identify putative TMBs. In general, TMB topology prediction methods can be classified as empirical, machine learning and consensus-based. A few of these methods are discussed below (Table 5.4).

5.5.6 *Empirical Approaches for β-Barrel Topology Prediction*

Traditionally, features based on knowledge gained from 3D structures, such as the hydrophobicity analyses over a sliding window, amino acid distribution, length of

Table 5.4 Topology prediction methods for transmembrane β-barrels

Method	Features	URL
BETAWARE (Savojardo et al. 2013a)	Conditional Random Fields	http://www.biocomp.unibo.it/
boctopus (Hayat and Elofsson 2012a)	HMM + SVM	http://boctopus.cbr.su.se/
tobmodel (Hayat and Elofsson 2012b)	HMM + SVM	http://tmbmodel.cbr.su.se/
TMBHMM (Singh et al. 2011)	HMM	http://www.zbi.uni-saarland.de/en
partiFold (Waldispühl et al. 2008)	http://partifold.csail.mit.edu/	Inter-strand residue interaction probabilities
PROFtmb (Bigelow and Rost 2006)	HMM	https://www.predictprotein.org/
transFold (Waldispühl et al. Waldispühl et al. 2006)	Multi-tape S-attribute grammars	http://bioinformatics.bc.edu/clotelab/transFold/
PRED-TMBB (Bagos et al. 2004)	HMM	http://bioinformatics.biol.uoa.gr/PRED-TMBB/
tbbpred (Natt et al. 2004)	SVM + NN	http://www.imtech.res.in/raghava/tbbpred/
TMBETAPRED-RBF (Ou et al. 2010)	SVM	http://rbf.bioinfo.tw/
TMBETA-NET (Gromiha et al. 2005)	NN	http://psfs.cbrc.jp/tmbeta-net/

TM β-strands and outer/inner loops, have been used for the topology prediction of TMBs (Schirmer and Cowan 1993; Gromiha et al. 1997; Gromiha and Ponnuswamy 1993; Diederichs et al. 1998). Wimley et al. (2002) combined features such as hydrophobicity profile, amino acid composition, known variation in the length of inner loops and the abundance of proteins facing the lipids of the barrel pore to formulate a computational score to predict TM stretches and also identify putative TMBs. The distribution of amino acids on a transmembrane β-strand along the membrane normal and the occurrence of the dyad-repeat pattern were employed by Jackups and Liang (2005) to improve the location of predicted strands and estimate the strand-registration such that the maximum number of hydrogen-bonds were satisfied between two adjacent β-strands.

5.5.7 Machine Learning Approaches for β-Barrel Topology Prediction

Machine learning-based methods for the topology prediction of TMBs are typically trained on a dataset of labeled data points extracted from known 3D structures. Rost and Sander (1993) showed early on that the use of information obtained from

multiple sequence alignments yields higher prediction accuracy as compared to using features from a single-sequence alone. SVMs, neural networks and hidden Markov models have all been used for TMB topology prediction (Table 5.4). The use of a sequence profile-based HMM for the identification and topology prediction of TMBs was first introduced by Martelli et al. (2002). PROFtmb (Bigelow and Rost 2006) and PRED-TMBB (Bagos et al. 2004) used a similar approach, where an HMM is used to predict strands, inner-loop and outer-loop states using a sequence profile. The HMM architecture employed in these methods was chosen such that it resembled a pair of strands (up and down), a self-loop representing long outer-loops that connect the two strands on the extracellular side and a self-loop of the inner-membrane side. The number of states representing the β-strand region was chosen to account for the variation in the length of these elements that form TMBs.

Recently, two-stage predictors such as BOCTOPUS (Hayat and Elofsson 2012a) and tobmodel (Hayat and Elofsson 2012b) have been implemented. These methods employ SVMs in the first stage to predict the local preference of each residue to form an outer-loop, inner-loop or membrane strand region. The output of this stage is then fed to an HMM that predicts the overall topology. Another approach called BETAWARE (Savojardo et al. 2013a) consists of two methods, first an N-to-1 Extreme Learning Machine algorithm is used for the identification of TMBs, followed by a Grammatical-Restrained Hidden Conditional Random Field approach to predict the topology. In contrast to other methods, transFold (Waldispühl et al. 2006) does not require a training set but uses a grammar to predict the β-strands and inter-strand residue contacts. Most of these topology prediction methods can also be used for distinguishing TMBs from non-TMBs.

5.5.8 *Consensus Approaches for β-Barrel Topology Prediction*

To our knowledge, conBBPRED (Bagos et al. 2005) is the only consensus method available for TMB topology prediction. conBBPRED assigns a per-residue score by averaging over contributions of each individual predictor followed by a dynamic programming step to obtain the overall topology. On a dataset of 20 proteins, conBBPRED increases the accuracy of predicted topologies by 15% (Bagos et al. 2005). With larger datasets and more topology predictors becoming available, it will be interesting to see if consensus topology prediction methods for TMBs show improved accuracy over single methods.

5.6 3D Structure Prediction

As with globular proteins, 3D structure prediction of TM proteins can be dealt with via two main approaches, homology modelling and de novo modelling, covered in Chaps. 1 and 4 of this book.

5.6.1 Homology Modelling of α-Helical Transmembrane Proteins

Homology modelling involves the use of a related template structure in order to build a 3D model of a target protein. The method is based on the observation that protein structure is conserved more highly than amino acid sequence, hence even proteins that have diverged significantly in sequence but still share detectable similarity may also share common structural properties, and in particular, the overall fold. When a suitable template is available, predicting TM protein structure by homology modelling can be highly effective, especially when tools specifically designed for modelling TM proteins are used. Compared to globular proteins, lower sequence conservation is required for fold preservation in transmembrane regions, so it may even be possible to generate useful 3D models with templates that share as little as 20% sequence identity to the target, although the paucity of high resolution membrane protein structures will still limit the number of families that such methods are applicable to (Olivella et al. 2013).

A homology modelling protocol can be subdivided into a number of key steps which can each be performed iteratively to improve the quality of the final model: template selection, target-template alignment, model construction, and model quality assessment (Marti-Renom et al. 2000; Sanchez and Sali 1997). Aside from SWISS-MODEL (Peitsch 1996; Biasini et al. 2014) which has a 7TM/GPCR interface, few TM protein-specific homology modelling methods exist. MEDELLER (Kelm et al. 2010) is designed to approach the steps in structure prediction to take into account the differences between the physical environments of globular and TM proteins. The method is optimized to build a highly reliable core structure shared by the template and target proteins by first calculating membrane insertion using iMembrane (Kelm et al. 2009) which is used to guide target-template alignment by MP-T (Hill and Deane 2013). The core is gradually extended using a specialized membrane-specific substitution score, before loops are completed using the loop modelling protocols FREAD (Choi and Deane 2010) and Modeller (Marti-Renom et al. 2000). Results show that MEDELLER produces accurate core models and achieves a core model accuracy of 1.97 Å RMSD versus 2.57 Å for Modeller. The Memoir modeling pipeline now provides a fully automated web server that applies this protocol to both α-helical and β-barrel TM proteins (Ebejer et al. 2013).

Chen and co-workers developed a method specifically to deal with the issue of building homology models from very distantly related homologues exhibiting distinct loop and TM helix conformations (Chen et al. 2014). The approach is based on efficient sampling techniques of alternative TM helix structures, in order to reconstruct both TM core and loop regions from distant structural homologues, resulting in high quality models that were top-ranked when stringently validated in two blind predictions (Kufareva et al. 2011; Michino et al. 2009). Since the method requires only a single distant homolog, they estimate that around 60% of human membrane proteins can be reliably modeled using their approach, allowing the generation of 3D models for a large and diverse fraction of structurally uncharacterized TM proteins.

A number of tools also exist to model specific regions of TM proteins. These include TM loop regions, which have been shown to differ significantly from loop regions in globular proteins. Kelm and co-workers showed that it is possible to accurately predict the structure of TM loops using a database of small TM protein loop fragments (0.8–1.6 Å). Their findings show that while many globular protein fragments have similar shapes to their TM counterparts, their sequences are often very different, although they do not appear to differ in their substitution patterns. Their method is implemented in a modification to FREAD (Kelm et al. 2014). Modelling of TM kinks has also attracted a lot of attention, as they have been observed to provide important functional and structural roles in TM proteins (Yohannan et al. 2004). Tools to model TM kinks include the Monte Carlo method based algorithm, MC-HELAN, which determines helical axes alongside positions and angles of helical kinks (Langelaan et al. 2010), HELANAL-Plus (Kumar and Bansal 2012), a web server for analysis of helix geometry in TM protein structures, and TMKink, a neural network predictor which identifies over two-thirds of all bends with high sensitivity and specificity (Meruelo et al. 2011).

5.6.2 *Homology Modelling of Transmembrane β-Barrel Proteins*

For transmembrane β-barrel proteins, HHomp (Remmert et al. 2009) can be used to identify remote homologues with a known 3D structure that can act as template/s for 3D modelling of these proteins. Standard application of MEDELLER or MODELLER can then be used to generate all-atom homology models (Kelm et al. 2010; Marti-Renom et al. 2000). The TMBpro method (Randall et al. 2008) uses a combination of machine-learning to predict the location of β-strands and inter-strand contacts and then selects templates from TMBs with known 3D structure by matching the number of β-strands. However, as stated above, a key limitation of such an approach is that it is only limited to protein sequences for which a reliable template can be found. Additionally, for transmembrane β-barrels,

where identification of novel families is still an open issue, such an approach might miss reliable templates.

5.6.3 De Novo Modelling of α-Helical Transmembrane Proteins

De novo modelling, or ab initio modelling, involves the construction of a 3D model in the absence of any tertiary structural data relating to the target protein. As with homology modelling, most methods address globular proteins although recently a number of methods have emerged specifically to deal with TM proteins including FILM (Pellegrini-Calace et al. 2003), RosettaMembrane (Barth et al. 2007, 2009) and BCL::MP-fold (Weiner et al. 2013) (Table 5.5).

FILM (Folding In Lipid Membranes) is a modification of the globular protein structure prediction method FRAGFOLD (Jones and McGuffin 2003; Jones 1997). FRAGFOLD employs simulated annealing in order to perform a conformational search using high-resolution super-secondary structural fragments to assemble the tertiary fold, guided by a statistical function that includes pairwise, solvation, steric and hydrogen bonding energy terms. FILM added a knowledge-based membrane potential term to the FRAGFOLD energy function, derived from the statistical analysis of a data set of 640 transmembrane helices whose topologies had been determined experimentally. The relative frequencies of each amino acid at fixed distances from the membrane centre were assessed, allowing the membrane potential term to be calculated by transforming these values using the inverse Boltzmann equation. Results indicated that it was possible to predict both the topology and conformation of small proteins at a reasonable level of accuracy, although attaining the level of compactness observed in larger TM helix bundles was challenging, since TM helix bundles are usually not optimally compact despite neighboring helices being closely packed together. Further modification to FILM allowed progress to be made in the prediction of larger TM helix bundles by incorporating another term accounting for lipid exposure into the energy function. This allowed models of seven TM helix bacteriorhodopsin and rhodopsin to be

Table 5.5 3D modelling tools for α-helical transmembrane proteins

Method	Features	URL
RosettaMembrane (Barth et al. 2009, 2007)	Knowledge-based potential	https://www.rosettacommons.org/
Evfold_membrane (Hopf et al. 2012; Sheridan et al. 2015)	Evolutionary couplings	http://evfold.org/transmembrane
FILM3 (Nugent and Jones 2012)	Evolutionary couplings	http://bioinfadmin.cs.ucl.ac.uk/downloads/FILM3/
BCL::MP-fold (Weiner et al. 2013)	Knowledge-based potential	http://www.meilerlab.org/index.php/servers

generated to within 6–7 Å root mean square deviation (rmsd) of the native structure (Hurwitz et al. 2006).

RosettaMembrane is also a modification of a globular protein structure prediction method—Rosetta (Rohl et al. 2004; Simons et al. 1999), which, like FRAGFOLD, assembles folds using fragments of known structures using simulated annealing or parallel tempering—an effective algorithm to overcome the slow convergence in low-temperature protein simulation. RosettaMembrane added terms to the Rosetta energy function that described intra-protein and protein-solvent interactions in the anisotropic membrane environment, treating hydrogen bonds explicitly and membrane protein/lipid interactions implicitly. The method describes interactions between protein residues at atomic detail while applying continuum solvent models to the water, hydrophobic core, and lipid head group regions of the membrane. Results suggest that the model captures the essential physical properties that govern the solvation and stability of membrane proteins, allowing the structures of 12 small TM protein domain (<150 residues) to be predicted successfully to a resolution of <2.5 Å (129), comparing favourably with predictions obtained on small water-soluble protein domains. More recently, the method was extended to incorporate distance constraints into the predictions to direct helix-helix interactions, the constraints being derived from either experimental data or sequence-based predictions (Fuchs et al. 2009; Lo et al. 2009; Nugent et al. 2011; Nugent and Jones 2010). This allowed larger (90–300 residues) structures with more complicated topologies to be successfully modelled to within 4 Å rmsd in the best four cases, with results indicating that only a single constraint was sometimes sufficient to enrich the population of near-native models.

A recent method BCL::MP-fold (Weiner et al. 2013), a modification of BCL::Fold (Karakas et al. 2012), generates models within a static membrane object by evaluating conformations using a knowledge-based energy potential which takes into account the unique properties of the apolar membrane in the amino acid environment potential, as well as an increased radius of gyration along the membrane normal. Three additional terms are introduced first to describe the preferential orientation of secondary structure elements with respect to the membrane, secondly to penalise connection of two neighboring TM helices that would require passage through the membrane, and finally to assess the agreement of residue placement in TM regions with predictions from sequence. Additionally, a symmetry folding mode allows for the prediction of obligate homo-multimeric TM complexes. A benchmark test using 40 TM protein 3D structures demonstrated that the method is able to accurately predict the correct topology in 34 cases, suggesting the approach can successfully predict protein topology without the need for large multiple sequence alignments, homologous template structures, or experimental restraints.

5.6.4 De Novo Modelling of Transmembrane β-Barrels

The topological arrangement of β-strands in transmembrane β-barrels is regular and can be exploited to generate 3D models of TMBs based on an idealized geometry (Naveed et al. 2012; Hayat and Elofsson 2012b). Existing methods based on idealized geometry approximate the diameter of a TMB, calculated based on its number of strands. Additionally, 3D coordinates of Cα atoms along β-strands and their placement with respect to the in-register Cα atom can also be determined using a theoretical description (Chou et al. 1990; Murzin et al. 1994a, b). Tobmodel uses these regular structural features to generate idealized Cα atoms of TMBs (Hayat and Elofsson 2012b). Another method, 3d-SpoT, uses an empirical scoring function derived from frequencies of lipid-facing and pore-facing residues in known TMB structures to find the optimal strand-registration and then uses a geometric model of intertwined coils to generate 3D models (Naveed et al. 2012) (Table 5.6).

5.6.5 Covariation-Based Approaches

Up until recently, using knowledge-based potentials derived from the statistical analysis of known protein structures has been the standard approach for de novo structure prediction. Over the last five years, the field has seen dramatic progress as new methods have emerged that are capable of accurately inferring residue-residue contacts from large multiple sequence alignments (MSAs), allowing 3D structures to be computed directly from sequence data. Two key factors have led to this revolution; firstly, the rapid growth in the size of sequence databases, which has resulted in the number of sequences available for a typical protein family increasing by orders of magnitude (Sadowski and Taylor 2013), and secondly, the application of advanced statistical methods to this sequence data that allows the detection of true correlated mutations between sites in MSAs. The main idea behind correlated

Table 5.6 3D modelling tools for transmembrane β-barrels

Method	Features	URL
EVfold_bb (Hayat et al. 2015)	Evolutionary couplings + Strand-registration prediction	http://cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/
tobmodel (Hayat and Elofsson 2012b)	Topology + Strand-registration prediction	http://tmbmodel.cbr.su.se/
3D-SpoT (Naveed et al. 2012)	Inter-strand pairing + Idealized barrel	http://tanto.bioe.uic.edu/TMBB-Explorer/
TMBpro (Randall et al. 2008)	Machine learning + Templates	http://tmbpro.ics.uci.edu/

mutations is that residues that are proximal in 3D space are more likely to impose constraints on each other, which should lead to a correlation in their substitution patterns in the MSA. Mutation of either residue might disrupt the stability of the contact, which is likely to have an impact on the stability of the overall fold. Subsequent mutation of one or both residues to a more physicochemically complementary pairing may increase the likelihood of the contact being maintained; therefore residue pairs that form contacts are often seen to covary. It is this property that modern contact prediction methods seek to exploit.

A number of different methods have been developed for predicting contacts from sequence data based on the recognition of these residue covariation patterns. Up until now, they major obstacle in achieving performance useful for structure prediction has been in dealing with indirect coupling effects: should a direct contact exist at sites A–B and A–C, an apparent interaction may appear between B-C even though no direct contact exists. The approach of Lapedes et al. (1999) dealt with this so-called chaining problem by applying a maximum entropy approach, but at a high computational cost. The Direct Coupling Analysis (DCA) method reduced the problem to one of maximum entropy inference, applying a heuristic message passing approach to determine the solution of the contact weights (Weigt et al. 2009). This allowed the approach of Lapedes et al. to be put to practical use, with prediction accuracy achieving sufficient quality to be useful in structure prediction (Taylor and Sadowski 2011). PSICOV is based on sparse inverse covariance estimation (Jones et al. 2012). It applies the graphical lasso method (Friedman et al. 2008) to estimate the inverse of the covariance matrix, which is calculated from the MSA, whilst also constraining the solution to be sparse. The inverse covariance matrix, also known as the precision matrix, gives the correlation between any two sites in the MSA, conditional on observations at all other sites. This global statistical model was able to predict contacts with an accuracy approaching 80%, even for long-range contacts (those separated by >23 residues in the sequence), which is sufficient to identify to the native fold for medium sized (<200 residue) globular proteins, where sufficient numbers of aligned sequences are available. A more recent method, plmDCA (Ekeberg et al. 2013) uses a pseudo-likelihood approach applied to the Potts models. This has been shown to significantly outperform existing DCA-based approaches, while consensus approaches such as PconsC (Skwark et al. 2013) and MetaPSICOV further improve performance (Jones et al. 2015).

5.6.6 *Evolutionary Covariation-Based Methods for De Novo Modelling of α-Helical Membrane Proteins*

The performance of these methods has led to the development of a number of de novo structure prediction methods capable of generating accurate models for even large domains, guided primarily by predicted contacts. Evfold_membrane (Hopf et al. 2012) incorporates predicted transmembrane topology into the EVfold protocol

(Marks et al. 2011), which uses DCA in combination with the CNS molecular dynamics software suite to generate 3D models. A webserver to de novo fold proteins using EVfold protocol with DCA and plmDCA has also been implemented (Sheridan et al. 2015). It was shown to be capable of generating accurate models within the top-10 ranked structures for fifteen targets ranging in size from 50 to 260 residues to within 2.7–4.8 Å rmsd of their native structures over at least two-thirds of the protein length. The latest version of FILM, FILM3, replaces the statistical potential with a single scoring function based on predicted contacts and their estimated probabilities (Nugent and Jones 2012). Using contacts predicted by PSICOV, results indicate that models with TM-scores >0.5 could be generated for 25 out of 28 membrane protein targets with complex topologies and an average length over 300 residues (Fig. 5.4). In the most remarkable case, it was possible to build a model for all 514 residues of cytochrome c oxidase polypeptide I with a TM-score >0.75. As encouraging as these results are, data suggests that even with perfect distance constraints, folding methods are unable to generate models less than 2 Å rmsd of the native structure, suggesting that protein refinement protocols will play an increasingly important role in generating higher accuracy models.

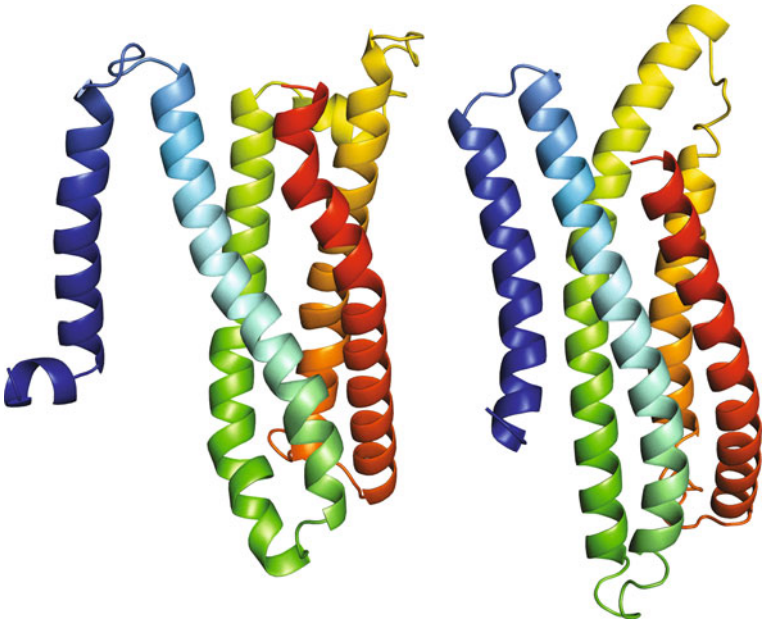

Fig. 5.4 Model of CASP 11 free modelling target T0836 (*right*)—a 5-helix TM protein. Predicted contacts were generated using MetaPSICOV (Jones et al. 2015) enabling a model to be produced using the FILM3 protocol (Nugent and Jones 2012) resulting in a TM-score of 0.60 (Kosciolek and Jones 2015). The native structure is on the *left*

5.6.7 Evolutionary Covariation-Based Methods for Transmembrane β-Barrel Structure Prediction

Transmembrane β-barrels have a uniform β-strand topological pattern, where alternate strands traverse from the inside to the outside and vice versa, and additionally, anti-parallel β-strands have a unique hydrogen-bonding pattern. These structural features can be exploited to enhance the accuracy of predicting residues pairs in contact between two adjacent β-strands. Further, these can also be used to estimate the registration (relative position of two strands with respect to each other) of two adjacent β-strands. This has been shown to be useful for 3D modelling of TMBs (Hayat and Elofsson 2012b; Naveed et al. 2012; Randall et al. 2008). Additionally, Hayat et al. (2015) have implemented a simple strand-shift algorithm, where adjacent strands are shifted up/down relative to each other to ascertain the position that gives the highest sum of evolutionary couplings (ECs) between paired residues to identify the correct registration of TM β-strands in TMBs. This hybrid algorithm that combines empirical knowledge about TM β-strands and evolutionary covariation analysis-based contact prediction improves the prediction accuracy of inter-strand residue contacts. These predicted inter-strands constraints can then used to identify the underlying hydrogen-bonding network and the resulting interactions are used as distance constraints to de novo fold large TMBs using a tool called EVfold_bb (Hayat et al. 2015). EVfold_bb method can correctly predict the 3D structure with an average TM-score of 0.54 for the top-ranking models. EVfold_bb can also identify the correct inter-strand registration with an accuracy of 44% (in generated models), which is an improvement over tobmodel (18%), which does not use ECs to guide optimal strand registration search. Moreover, the generated models are not restricted to idealized geometries and do not require a template. Most interestingly, EVfold_bb can also identify and model 3D interactions between the barrel and the large plug domain in FecA protein (TM-score 0.68). The plug domain sits in the TM barrel domain and participates in gating and signaling (Noinaj et al. 2012).

Furthermore, methods specifically meant for improving prediction of β-sheet contacts in both globular and membrane proteins have also been developed. These methods can be broadly divided into two groups based on the use of ECs. BetaPro (Cheng and Baldi 2005) and MLN-2S (Lippi and Frasconi 2009) use neural networks and Markov logic networks, respectively, to predict β-sheet contacts. Maximum entropy-based correlated mutation measures (CMM) (Burkoff et al. 2013), Bcov (Savojardo et al. 2013b), bbcontacts (Andreani and Söding 2015) and MetaPSICOV (Jones et al. 2015) all use evolutionary covariation. In addition, these methods employ an additional layer of machine-learning techniques such as deep learning or HMMs on predicted evolutionary couplings to increase the accuracy of predicted residue-residue contacts in β-sheets. In future, methods that combine the general principles of anti-parallel β-stands along with machine-learning based methods that employ predicted contacts should be able to improve the applicability of these techniques to TMBs.

5.7 Future Directions

Substantial progress has been made in the field of membrane protein structure prediction over recent years. Methods for the detection of remote homologues have drastically improved, making it possible to generate template-based models for a larger number of protein families. Advances in techniques for predicting pairwise residue contacts have made it possible to generate de novo 3D models of large membrane proteins. However, these techniques are only applicable to protein families with large multiple sequence alignments. It is anticipated that as more sequencing data becomes available, 3D models of yet unknown TM protein families will become model-able based on predicted contacts. Future challenges lie in further improving these contact prediction methods by optimizing multiple sequence alignments, generation of fragment libraries, statistical inference methods used and the tools employed to predict 3D models.

Competing Interests The authors declare that they have no competing interests.

References

Amico M, Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G, Jones D, Krogh A, Fariselli P, Luigi Martelli P, Casadio R (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. Nucleic Acids Res 34 (Web Server issue):W169–172

Andreani J, Söding J (2015) bbcontacts: prediction of β-strand pairing from direct coupling patterns. Bioinformatics:btv041

Bagos PG, Liakopoulos TD, Hamodrakas SJ (2005) Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics 6(1):7

Bagos PG, Liakopoulos TD, Hamodrakas SJ (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. BMC Bioinformatics 7:189

Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004) PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. Nucleic Acids Res 32(suppl 2):W400–W404

Bahr A, Thompson JD, Thierry JC, Poch O (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 29(1):323–326

Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci USA 104(40):15682–15687

Barth P, Wallner B, Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. Proc Natl Acad Sci USA 106(5):1409–1414

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783–795

Bernsel A, Viklund H, Hennerdal A, Elofsson A (2009) TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res 37(Web Server issue):W465–468

Berven FS, Flikka K, Jensen HB, Eidhammer I (2004) BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 32(suppl 2):W394–W399

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(Web Server issue): W252–258

Bigelow H, Rost B (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res 34(suppl 2):W186–W188

Burkoff NS, Várnai C, Wild DL (2013) Predicting protein β-sheet contacts using a maximum entropy based correlated mutation measure. Bioinformatics:btt005

Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ, Chen AP (2006) Retraction. Science 314 (5807):1875

Chang JM, Di Tommaso P, Taly JF, Notredame C (2012) Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13(Suppl 4):S1

Chen KY, Sun J, Salvo JS, Baker D, Barth P (2014) High-resolution modeling of transmembrane helical protein structures from distant homologues. PLoS Comput Biol 10(5):e1003636

Cheng J, Baldi P (2005) Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics 21(suppl 1):i75–i84

Chetwynd AP, Scott KA, Mokrab Y, Sansom MS (2008) CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. Mol Membr Biol 25(8):662–669

Choi Y, Deane CM (2010) FREAD revisited: Accurate loop structure prediction using a database search algorithm. Proteins 78(6):1431–1440

Chou KC, Carlacci L, Maggiora GM (1990) Conformational and geometrical properties of idealized beta-barrels in proteins. J Mol Biol 213(2):315–326

Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10(6):685–686

Dayhoff MO, Schwartz RM (1978) Chapter 22: A model of evolutionary change in proteins. In: in Atlas of Protein Sequence and Structure. Silver Spring

Deng Y (2006) TSFSOM: transmembrane segments prediction by fuzzy self-organizing map. In: Advances in neural networks-ISNN 2006. Springer, pp 728–733

Diederichs K, Freigang J, Umhau S, Zeth K, Breed J (1998) Prediction by a neural network of outer membrane β-strand protein topology. Protein Sci 7(11):2413–2420

Dobson L, Lango T, Remenyi I, Tusnady GE (2015a) Expediting topology data gathering for the TOPDB database. Nucleic Acids Res 43(Database issue):D283–289

Dobson L, Remenyi I, Tusnady GE (2015b) CCTOP: a Consensus constrained TOPology prediction web server. Nucleic Acids Res

Ebejer JP, Hill JR, Kelm S, Shi J, Deane CM (2013) Memoir: template-based structure prediction for membrane proteins. Nucleic Acids Res 41(Web Server issue):W379–383

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E: Stat, Nonlin, Soft Matter Phys 87(1):012707

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971

Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353

Ferguson AD, Chakraborty R, Smith BS, Esser L, van der Helm D, Deisenhofer J (2002) Structural basis of gating by the outer membrane transporter FecA. Science 295(5560):1715–1719

Freeman TC Jr, Wimley WC (2010) A highly accurate statistical approach for the prediction of transmembrane beta-barrels. Bioinformatics 26(16):1965–1974. doi:10.1093/bioinformatics/btq308

Freeman TC, Wimley WC (2012) TMBB-DB: a transmembrane β-barrel proteome database. Bioinformatics 28(19):2425–2430

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
Fuchs A, Kirschner A, Frishman D (2009) Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. Proteins 74(4):857–871
Gallin WJ, Boutet PA (2011) VKCDB: voltage-gated K+ channel database updated and upgraded. Nucleic Acids Res 39(Database issue):D362–366
Garrow AG, Agnew A, Westhead DR (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane β-barrel proteins. Nucleic Acids Res 33(suppl 2):W188–W192
Gromiha MM, Ahmad S, Suwa M (2005) TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins. Nucleic Acids Res 33(suppl 2): W164–W167
Gromiha MM, Majumdar R, Ponnuswamy P (1997) Identification of membrane spanning beta strands in bacterial porins. Protein Eng 10(5):497–500
Gromiha MM, Ponnuswamy P (1993) Prediction of transmembrane β-strands from hydrophobic characteristics of proteins. Int J Pept Protein Res 42(5):420–431
Gromiha MM, Yabuki Y, Kundu S, Suharnan S, Suwa M (2007) TMBETA-GENOME: database for annotated β-barrel membrane proteins in genomic sequences. Nucleic Acids Res 35(suppl 1):D314–D316
Hayat M, Khan A (2013) WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids. Amino Acids 44(5):1317–1328
Hayat S, Elofsson A (2012a) BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Bioinformatics 28(4):516–522
Hayat S, Elofsson A (2012b) Ranking models of transmembrane β-barrel proteins using Z-coordinate predictions. Bioinformatics 28(12):i90–i96
Hayat S, Sander C, Marks DS, Elofsson A (2015) All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. Proc Natl Acad Sci 112(17):5413–5418
Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
Henricson A, Kall L, Sonnhammer EL (2005) A novel transmembrane topology of presenilin based on reconciling experimental and computational evidence. FEBS J 272(11):2727–2733
Hill JR, Deane CM (2013) MP-T: improving membrane protein alignment for structure prediction. Bioinformatics 29(1):54–61. doi:10.1093/Bioinformatics/Bts640
Hill JR, Kelm S, Shi J, Deane CM (2011) Environment specific substitution tables improve membrane protein alignment. Bioinformatics 27(13):15–23
Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149(7):1607–1621. doi:10.1016/j.cell.2012.04.012
Hopkins AL, Groom CR (2002) The druggable genome. Nat Rev Drug Discov 1(9):727–730
Hurwitz N, Pellegrini-Calace M, Jones DT (2006) Towards genome-scale structure prediction for transmembrane proteins. Philos Trans R Soc Lond B Biol Sci 361(1467):465–475
Imai K, Hayat S, Sakiyama N, Fujita N, Tomii K, Elofsson A, Horton P (2013) Localization prediction and structure-based in Silico analysis of bacterial proteins: with emphasis on outer membrane proteins. In: Data mining for systems biology. Springer, Berlin, pp 115–140
Jackups R, Liang J (2005) Interstrand pairing patterns in β-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol 354 (4):979–993
Jayasinghe S, Hristova K, White SH (2001) MPtopo: a database of membrane protein topology. Protein Sci 10(2):455–458
Jimenez-Morales D, Liang J (2011) Pattern of amino acid substitutions in transmembrane domains of β-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria. PLoS ONE 6(11):e26400
Jones DT (1997) Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins Suppl 29(1):185–191

Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23(5):538–544
Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190. doi:10.1093/bioinformatics/btr638
Jones DT, McGuffin LJ (2003) Assembling novel protein folds from super-secondary structural fragments. Proteins 53(suppl 6):480–485
Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31(7):999–1006
Jones DT, Taylor WR, Thornton JM (1994a) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33(10):3038–3049
Jones DT, Taylor WR, Thornton JM (1994b) A mutation data matrix for transmembrane proteins. FEBS Lett 339(3):269–275
Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338(5):1027–1036
Kall L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21(Suppl 1):i251–i257
Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J (2012) BCL:Fold-de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS ONE 7(11):e49240. doi:10.1371/journal.pone.0049240
Kelm S, Shi J, Deane CM (2009) iMembrane: homology-based membrane-insertion of proteins. Bioinformatics 25(8):1086–1088
Kelm S, Shi J, Deane CM (2010) MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 26(22):2833–2840
Kelm S, Vangone A, Choi Y, Ebejer JP, Shi J, Deane CM (2014) Fragment-based modeling of membrane protein loops: successes, failures, and prospects for the future. Proteins 82(2):175–186. doi:10.1002/prot.24299
Khafizov K, Staritzbichler R, Stamm M, Forrest LR (2010) A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe. Biochemistry 49 (50):10702–10713
Klammer M, Messina DN, Schmitt T, Sonnhammer EL (2009) MetaTM—a consensus method for transmembrane protein topology prediction. BMC Bioinformatics 10:314
Kosciolek T, Jones DT (2015) Accurate contact predictions using coevolution techniques and machine learning. Proteins. doi:10.1002/prot.24863
Kozma D, Simon I, Tusnady GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41(Database issue):D524–529
Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580
Kufareva I, Rueda M, Katritch V, Stevens RC, Abagyan R (2011) Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. Structure 19(8):1108–1126
Kumar P, Bansal M (2012) HELANAL-Plus: a web server for analysis of helix geometry in protein structures. J Biomol Struct Dyn 30(6):773–783
Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
Kyttala A, Ihrke G, Vesa J, Schell MJ, Luzio JP (2004) Two motifs target Batten disease protein CLN3 to lysosomes in transfected nonneuronal and neuronal cells. Mol Biol Cell 15(3):1313–1323
Langelaan DN, Wieczorek M, Blouin C, Rainey JK (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. J Chem Inf Model 50(12):2213–2220

Lapedes AS, Giraud B, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: phylogenetic and structural effects. In: Seillier-Moiseiwitsch F (ed) Statistics in molecular biology and genetics, vol 33. Lecture Notes-Monograph Series. Institute of Mathematical Statistics, Hayward, CA, pp 236–256

Li B, Gallin WJ (2004) VKCDB: voltage-gated potassium channel database. BMC Bioinformatics 5:3

Lippi M, Frasconi P (2009) Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics 25(18):2326–2333

Lo A, Chiu HS, Sung TY, Hsu WL (2006) Transmembrane helix and topology prediction using hierarchical SVM classifiers and an alternating geometric scoring function. Comput Syst Bioinformatics Conf, 31–42

Lo A, Chiu HS, Sung TY, Lyu PC, Hsu WL (2008) Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. J Proteome Res 7(2):487–496

Lo A, Chiu YY, R?dland EA, Lyu PC, Sung TY, Hsu WL (2009) Predicting helix-helix interactions from residue contacts in membrane proteins. Bioinformatics 25(8):996–1003

Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI (2006a) Positioning of proteins in membranes: a computational approach. Protein Sci 15(6):1318–1333

Lomize AL, Pogozheva ID, Mosberg HI (2011) Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. J Chem Inf Model 51(4):930–946

Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006b) OPM: orientations of proteins in membranes database. Bioinformatics 22(5):623–625. doi:10.1093/bioinformatics/btk023

Mao Q, Foster BJ, Xia H, Davidson BL (2003) Membrane topology of CLN3, the protein underlying Batten disease. FEBS Lett 541(1–3):40–46

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766. doi:10.1371/journal.pone.0028766

Martelli PL, Fariselli P, Casadio R (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. Bioinformatics 19(Suppl 1):i205–i211

Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. Bioinformatics 18(suppl 1):S46–S53

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Meruelo AD, Samish I, Bowie JU (2011) TMKink: a method to predict transmembrane helix kinks. Protein Sci 20(7):1256–1264

Michino M, Abola E, Brooks CL, Dixon JS, Moult J, Stevens RC (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. Nat Rev Drug Discov 8(6):455–463

Muller T, Rahmann S, Rehmsmeier M (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics 17(Suppl 1):S182–S189

Murzin AG, Lesk AM, Chothia C (1994a) Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 236(5):1369–1381

Murzin AG, Lesk AM, Chothia C (1994b) Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. J Mol Biol 236(5):1382–1400

Natt NK, Kaur H, Raghava G (2004) Prediction of transmembrane regions of β-barrel proteins using ANN-and SVM-based methods. Proteins: Struct Funct Bioinf 56(1):11–18

Naveed H, Xu Y, Jackups R Jr, Liang J (2012) Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. J Am Chem Soc 134(3):1775–1781

Ng PC, Henikoff JG, Henikoff S (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16(9):760–766

Nilsson J, Persson B, Von Heijne G (2002) Prediction of partial membrane protein topologies using a consensus approach. Protein Sci 11(12):2974–2980

Noinaj N, Easley NC, Oke M, Mizuno N, Gumbart J, Boura E, Steere AN, Zak O, Aisen P, Tajkhorshid E, others (2012) Structural basis for iron piracy by pathogenic Neisseria. Nature 483(7387):53–58

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205–217

Nugent T, Jones DT (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 10:159

Nugent T, Jones DT (2010) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol 6(3):e1000714

Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proc Natl Acad Sci USA 109(24):E1540–E1547

Nugent T, Jones DT (2013) Membrane protein orientation and refinement using a knowledge-based statistical potential. BMC Bioinformatics 14:276

Nugent T, Ward S, Jones DT (2011) The MEMPACK alpha-helical transmembrane protein structure prediction server. Bioinformatics 27(10):1438–1439

Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29(13):1589–1592

Y-y Ou, S-a Chen, Gromiha MM (2010) Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy. J Comput Chem 31(1):217–223

Peitsch MC (1996) ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. Biochem Soc Trans 24(1):274–279

Pellegrini-Calace M, Carotti A, Jones DT (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. Proteins 50(4):537–545

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786

Pirovano W, Feenstra KA, Heringa J (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 24(4):492–497

Qi Y, Oja M, Weston J, Noble WS (2012) A unified multitask architecture for predicting local protein properties. PLoS ONE 7(3):e32235

Randall A, Cheng J, Sweredoski M, Baldi P (2008) TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins. Bioinformatics 24(4):513–520

Ratajczak E, Petcherski A, Ramos-Moreno J, Ruonala MO (2014) FRET-assisted determination of CLN3 membrane topology. PLoS ONE 9(7):e102593

Remmert M, Linke D, Lupas AN, Söding J (2009) HHomp?prediction and classification of outer membrane proteins. Nucleic Acids Res 37(suppl 2):W446–W451

Reynolds SM, Kall L, Riffle ME, Bilmes JA, Noble WS (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 4(11): e1000213

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Meth Enzymol 383:66–93

Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 5(8):1704–1718

Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci 90(16):7558–7562

Sadowski MI, Taylor WR (2013) Prediction of protein contacts from correlated sequence substitutions. Sci Prog 96(Pt 1):33–42

Saier MH, Reddy VS, Tamang DG, Vastermark A (2014) The transporter classification database. Nucleic Acids Res 42(Database issue):D251–258

Saier MH, Tran CV, Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34(Database issue): D181–186

Saier MH, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucleic Acids Res 37(Database issue):D274–278

Samatey FA, Xu C, Popot JL (1995) On the distribution of amino acid residues in transmembrane alpha-helix bundles. Proc Natl Acad Sci USA 92(10):4577–4581

Sanchez R, Sali A (1997) Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7(2):206–214

Sansom MS, Scott KA, Bond PJ (2008) Coarse-grained simulation: a high-throughput computational approach to membrane proteins. Biochem Soc Trans 36(Pt 1):27–32

Savojardo C, Fariselli P, Casadio R (2013a) BETAWARE: a machine-learning tool to detect and predict transmembrane beta barrel proteins in Prokaryotes. Bioinformatics:bts728

Savojardo C, Fariselli P, Martelli PL, Casadio R (2013b) BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. Bioinformatics: btt555

Schirmer T, Cowan SW (1993) Prediction of membrane-spanning β-strands and its application to maltoporin. Protein Sci 2(8):1361–1363

Senes A, Gerstein M, Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol 296(3):921–936

Shafrir Y, Guy HR (2004) STAM: simple transmembrane alignment method. Bioinformatics 20(5):758–769

Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, Hopf T, Marks DS, Sander C (2015) EVfold. org: Evolutionary Couplings and Protein 3D Structure Prediction. bioRxiv:021022

Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176

Singh NK, Goodman A, Walter P, Helms V, Hayat S (2011) TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1814(5):664–670

Skwark MJ, Abdel-Rehim A, Elofsson A (2013) PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 29(14):1815–1816. doi:10.1093/bioinformatics/btt259

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2013) Alignment of helical membrane protein sequences using AlignMe. PLoS ONE 8(3):e57731

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2014) AlignMe–a membrane protein sequence alignment web server. Nucleic Acids Res 42(Web Server issue):W246–251

Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 334(5):1043–1062

Taylor PD, Attwood TK, Flower DR (2003) BPROMPT: a consensus server for membrane protein prediction. Nucleic Acids Res 31(13):3698–3700

Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. PLoS ONE 6(12):e28265

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Tsirigos KD, Bagos PG, Hamodrakas SJ (2011) OMPdb: a database of β-barrel outer membrane proteins from Gram-negative bacteria. Nucleic Acids Res 39(suppl 1):D324–D331

Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res

Tusnady GE, Dosztanyi Z, Simon I (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 20(17):2964–2972

Tusnady GE, Dosztanyi Z, Simon I (2005a) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 33(Database issue):D275–278

Tusnady GE, Dosztanyi Z, Simon I (2005b) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 21(7):1276–1277

Tusnady GE, Kalmar L, Simon I (2008) TOPDB: topology data bank of transmembrane proteins. Nucleic Acids Res 36(Database issue):D234–239

Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283(2):489–506

Viklund H, Bernsel A, Skwark M, Elofsson A (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 24(24):2928–2929. doi:10.1093/bioinformatics/btn550

Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13(7):1908–1917

Viklund H, Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 24(15):1662–1668. doi:10.1093/bioinformatics/btn221

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225(2):487–494

Waldispühl J, Berger B, Clote P, Steyaert J-M (2006) transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res 34(suppl 2): W189–W193

Waldispühl J, O'Donnell CW, Devadas S, Clote P, Berger B (2008) Modeling ensembles of transmembrane β-barrel proteins. Proteins: Structure, Function, Bioinform 71(3):1097–1112

Wallin E, Tsukihara T, Yoshikawa S, von Heijne G, Elofsson A (1997) Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. Protein Sci 6(4):808–815

Wang H, Liu B, Sun P, Ma Z (2013) A topology structure based outer membrane proteins segment alignment method. Mathematical Problems in Engineering 2013

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106(1):67–72

Weiner BE, Woetzel N, Karakas M, Alexander N, Meiler J (2013) BCL:MP-fold: folding membrane proteins through assembly of transmembrane helices. Structure 21(7):1107–1117. doi:10.1016/j.str.2013.04.022

White SH (2004) The progress of membrane protein structure determination. Protein Sci 13(7):1948–1949

Wimley WC (2002) Toward genomic identification of β-barrel membrane proteins: Composition and architecture of known structures. Protein Sci 11(2):301–312

Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3(10):842–848

Yan R-X, Chen Z, Zhang Z (2011) Outer membrane proteins can be simply identified using secondary structure element alignment. BMC Bioinformatics 12(1):76

Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 101(4):959–963

Yuan Z, Mattick JS, Teasdale RD (2004) SVMtm: support vector machines to predict transmembrane segments. J Comput Chem 25(5):632–636