Chapter 9
Structure Prediction of Transmembrane Proteins

Gábor E. Tusnády and Dániel Kozma

9.1 Introduction

In this chapter we discuss the various structural aspects of transmembrane proteins (TMPs) and survey the tasks and methods needed for modeling their structure. The structure prediction of TMPs from the pure amino acid sequence translated from genome projects may go through the following steps: (i) remove annotated or predicted cleavable parts (transit sequences, signal peptides); (ii) determination of the protein type (TMP or not); (iii) localization of TM segments within the amino acid sequence (topography prediction, 2D prediction) and the soluble parts of the protein relative to the membrane (topology prediction, 2.5D prediction); (iv) modeling the tertiary structure (3D) of membrane embedded protein parts which, depending on the amino acid similarity to the available relatives whose structure are already solved, may be based on homology modeling; may use the advantage of threading or may be de novo predictions including the contact prediction of amino acids of TM segments; (v) prediction of oligomerization propensity; (vi) finding the orientation in the membrane. In the following sections we guide the reader through these consecutive steps (Fig. 9.1) on how to derive the biologically active form of an unknown TMP purely computationally.

G.E. Tusnády (✉) · D. Kozma
Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, P.O. Box 7, Budapest 1518, Hungary
e-mail: tusnady.gabor@ttk.mta.hu

G. Náray-Szabó (ed.), *Protein Modelling*, DOI 10.1007/978-3-319-09976-7_9

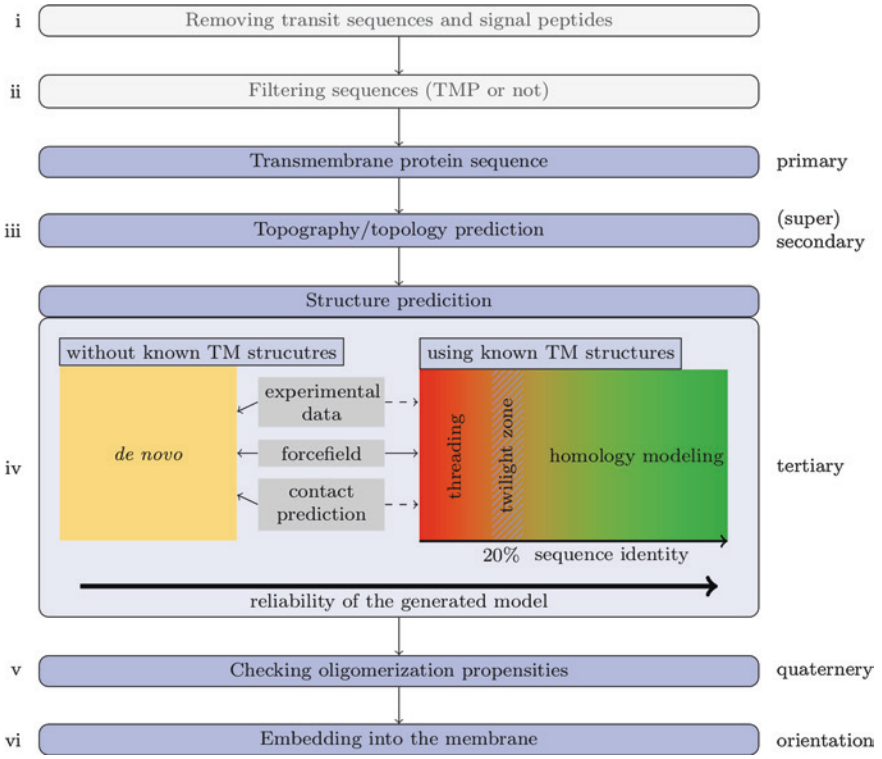

Fig. 9.1 Summary of the prediction pipeline of transmembrane proteins

9.2 Structural Aspects of Transmembrane Proteins

The lipid bilayer is an amphipathic slab with hydrophilic surfaces and a hydrophobic core region, from where the water molecules are excluded. Therefore the membrane segments of the polypeptide chains must adopt structures where all hydrogen donor and acceptor atoms are bound intramolecularly. This constraint leads to the formation of α-helical bundles and β-barrel secondary structures that are the most common secondary structures in the membrane spanning regions of TMPs. Therefore, based on the secondary structure of protein segments in the membrane regions, TMP can be classified into two main groups: α-helical and β-barrel. All plasma-membrane proteins are α-helical bundles with a large conformational variation, which is partly due to the water molecules penetrated into the membrane regions of the TMPs [161] forming water-filled cavities which makes the hydropathy-based topology predictions more difficult as well.

Very rarely, coil regions can be found in the membrane-embedded structure parts, mostly in re-entrant regions (that enter and exit on the same side of the membrane) or at kinks, where the translational symmetry breaks. Secondary

structures do not terminate necessarily at the membrane water interface; sometimes these penetrate to the hydrophilic water phase. Often, on the membrane—water barrier interfacial helices (α-helices laying close and approximately parallel to the membrane surface) can be found, which have various (but not fully understood) functional roles, e.g. gating regulation and co-factor shielding [104].

While α-helical TMPs exist in all super-kingdoms, β-barrel TMPs can be found only in bacterial porins and in the inner membrane of mitochondria of eukaryote cells. For a long time it was believed that β-barrel TMPs always have even number of strands and in the range between 8 and 22, but this is refuted by the recently solved structure of the voltage-dependent anion channel (VDAC) [9] and the translocation domain of bacterial usher proteins [113, 150] containing 19 and 24 transmembrane (TM) β-strands, respectively.

The number of TM segments in α-helical TMPs range from 1 up to 24 (sodium channel protein type 2, α-subunit), but regarding their number in autonomous protein domains the highest known number is 15. Genome-wide analyses showed that distribution of the number of TM segments in α-helical TMPs is not random, proteins with 6 and 12 transmembrane helices (TMHs), such as small-molecule transporters, sugar transporters and ABC transporters, are predominant in uni-cellular organisms [3, 28, 67, 73, 123, 157]. In contrast, proteins with 7 TMHs are frequent in worms and human due to the high abundance of G-protein coupled receptors (GPCRs) [100]. Partly due to this abundance, the seven-helix membrane protein family members are the most important current drug targets.

9.3 Estimated Size of the Structure Space of Transmembrane Proteins

For globular/soluble proteins the total number of distinct globular folds that exist in nature is predicted to be a rather limited number [23], probably no more than 10,000 [70, 162], regardless of the astronomical number of the possible combination of structural elements. In TMPs, due to the physical constraints imposed by the lipid bilayer the number of possible folds is much smaller. Most of the TMPs adhere to one principal topology, involving one or more α-helices arranged parallel to each other and oriented about perpendicular with respect to the membrane plane. For β-barrel TMPs, they have a smaller structural diversity than α-helical ones. The short loops between helices constrain the possible folds of TMHs, therefore conformation space can be sampled effectively for small numbers of helices, and there are only about 30 possible folds for a TMP with three transmembrane helices (TMHs) [15]. However, the number of combinatorially possible folds was shown to increase exponentially with the number of TMHs to 1.5 million folds for seven helices, studies have showed that increasing number of membrane regions does not mean the exponential expansion of the fold space. Moreover structures with 8 or more transmembrane helices have less different architectures which reuse elements of folds with 3 or 4 helices [100]. Therefore the size of the fold

space cannot be predicted based on combinatorial considerations. As an upper estimate of the number of different structures, the number of protein families can be used that are identifiable based on sequence similarity alone. Obviously this is a rough approximation, but provides a definite and reliable upper limit.

Liu et al. [87] showed in a study of 26 proteomes that there are about 10 times more soluble protein families than membrane protein families. Oberai et al. [107] set up a numerical experiment to estimate the number of distinct TMP folds. They found that any given residue has an 80 % chance to fall into one of about 500 families and observed a significant decrease in the number of members between the first and the second 20 most populous families. These results indicate that there are only a few very large and many very small families of membrane proteins, similarly to soluble proteins. The largest families are populated by various signaling proteins (e.g. GPCRs) and channels (e.g. potassium channels) [24, 48, 129], different transporters (secondary transporters and the ABC transporter family [32, 128, 132]), and TMPs involved in energy production (cytochrome b and NADH ubiquinone oxidoreductases) [12, 39]. As a consequence of the rapid fall-of and the asymptotic tail of the family size distribution, Oberai et al. [107] concluded that 670 families will cover 80 % of the structured sequence space but 1,720 families are needed to cover 90 % of the structured sequence space for all extant polytopic membrane proteins. These numbers are still an upper limit, as in SCOP [2] hierarchy a family is a subset of a fold. Assuming that the distribution of folds over families is similar to the one of soluble proteins and applying a stretched exponential model [44], Oberai et al. [107] estimated that only 550 folds cover 90 % and 300 folds cover 80 % of membrane protein structured sequence space. Finally, taking into account the physical constraint that stem from the membrane bilayer environment, they expect this is still an overestimate. Currently about only a hundred distinct (good quality, X-ray) transmembrane folds are known from various organisms. Known TMP structures by now make possible to create model for 26 % of the human α-helical transmembrane proteome using homology modeling (see Sect. 9.4.2.1), this ratio could be increased up to 56 % with 100 more new evenly selected and determined structures [115].

Another interesting paper discusses the number of different helix-helix contact architectures as a function of the number of transmembrane segments [100]. They developed a method for predicting helical interaction graphs and found that membrane proteins with 8 and more helices have significantly fewer arrangements than proteins with up to 7 helices. The most striking cases are transporter proteins with either 8 or 11 transmembrane helices, which according to Neumann et al., all seem to share a common helix interaction pattern. It was observed that TMPs with 10, 12, 14 membrane segments have significantly more distinct interaction graph than TMPs with 11, 13 or 15. This implies a hypothesis, TMPs with more than 8 TM segments may originated from TMPs with 5, 6 and 7 membrane regions that themselves are distributed over many different helix interaction clusters. While odd number of regions cannot stem from gene duplication, this could be an acceptable explanation for the phenomena described above [100].

9.4 Predicting Different Levels of Structures

9.4.1 Topography and Topology Prediction

Starting structure modeling from the amino acid sequence the first task is to check the presence of signal peptides and to decide whether it codes a globular or transmembrane protein. Here we refer some recent reviews, where these problems are discussed [146].

As a second step, one has to locate membrane spanning segments within the sequence. The information refers the location of the membrane spanning regions within the sequence is called topography. While in the case of helical TMPs the transmembrane segments are formed by 15–20 hydrophobic amino acids, in case of the β-barrel TMPs the length of the TM segments are shorter and only every second amino acid has to be hydrophobic making their topography prediction harder. In this section we do not discuss topography prediction of β-barrel TMPs; instead we focus on helical transmembrane segment prediction.

Earlier topography prediction methods [35, 74] explored the fact that membrane spanning segments are more hydrophobic than other parts of the protein chain. These segments can be identified by averaging the hydrophobicity of the amino acids within a sliding window over the sequence investigated. Other statistical approaches, like the Dense Alignment Surface (DAS) algorithm [27] overcomes the difficulties caused by the different hydrophobicity scales by a special alignment procedure [26], where the unrelated TMPs recognize each other without applying any hydrophobicity scales. Later it was shown, that in the case of a properly chosen hydrophobicity scale, accuracy of topography prediction can be as high as of the best state-of-the-art prediction methods [11].

For topology prediction the next step is orienting the membrane spanning segments from outside to inside or vice versa. This is equivalent to localize the sequence segments between membrane spanning segments alternatively inside or outside. The difference between topography and topology is that topology refers the location of the non-membrane segments as well. However, there are only a few properties of TMPs that help this task. The first and most prevalent such feature of TMPs is that the positively charged amino acids are more abundant on the cytosolic part of polypeptide chain, than on the extra-cytosolic ones (positive-inside rule) [137, 151]. Most topology predictions apply this rule after the topography prediction to choose the more likely from the two possible models [126]. Some prediction methods, such as TOPPRED [137, 153], utilize this rule both for topography and for topology prediction, by generating several models with certain and possible transmembrane segments, and choosing the model where the differences of the number of lysines and arginines were the highest between the even and odd loops. The MEMSAT method [60] incorporates the positive-inside rule indirectly by maximizing the sum of log-likelihoods of amino acid preferences taken from various structural parts of membrane proteins in a model recognition approach.

By increasing the number of TMPs whose topology were experimentally proven, machine learning algorithms like hidden Markov model (HMM) [119],

support vector machine (SVM) [25] and artificial neural network (ANN) [94] can provide high prediction accuracies due to the fact, that the amino acid compositions of the various structural parts of TMPs are specific and machine learning algorithms are capable of learning these compositions during supervised learning [139]. Novel machine learning methods report higher and higher prediction accuracies due to the continuously growing and more reliable training sets and combining various techniques (e.g. using SVM or ANN for residue prediction and HMM for segment identification [149]). However, as these methods usually operate with parameter sets that are hard or near-impossible to integrate biochemically we cannot learn from these methods about the topology forming rules of TMPs. Moreover, to predict the topology of novel TMPs were never seen earlier by the machine learning methods, these methods may need to be retrained.

Replacing supervised learning technique by unsupervised one for HMMs, the training phase can be eliminated and the dependence on the training set can be avoided as well. Methods, such as HMMTOP, therefore do not need to be retrained from time to time. The success of unsupervised learning is based on the fact that a polypeptide chain of a TMP goes through various spaces of a cell with different physico-chemical properties (hydrophobic, polar, negatively charged, water-lipid interface etc.), therefore, the amino acid compositions of the TMP segments will be different in each type of regions. We do not need to know and as a consequence the constructed method does not need to learn these characteristic amino acid compositions to successfully predict the topology of TMPs. According to the law of maximal probability, these structural parts can be identified by segmenting in a way, that the amino acid compositions of the various structural parts show maximal divergence. This partitioning can be found by hidden Markov models.

There are two additional possibilities to increase the prediction accuracy of topology predictions. The first one is the utilization of consensus prediction methods. In addition to getting better predictions, using the results of several prediction methods allows us to estimate the reliability of the predicted topology as well. The consensus approach was also applied to predict partial membrane topologies, i.e. the part of the sequence where the majority of the applied methods agree. The other technique to increase the prediction accuracy is the use of constrained prediction methods. These can be used if there is/are one or more experimental data about the topology and prediction method can handle these data as constraints and not only to filter results that agree with the given experimental data. Thus, given a constraint (e.g. the N-terminus is inside), a constrained prediction method gives a prediction that satisfies this criterion. In a HMM based method this is achieved by the modification of the Baum-Welch and Viterbi algorithms. The first such application was HMMTOP2 [145]. Later the two other HMM based methods, TMHMM and Phobius were also modified to include this feature [62, 139]. The mathematical details of the necessary modification can be found in Ref. [6]. The optimal placement of constraints was also investigated, and it was shown that the accuracy can be increased by 10 % if the N- or C-terminal of the polypeptide chain is constrained in the above mentioned way, and 20 % is the maximum obtainable increase if one of each loop or tail residue in turn is fixed to its experimentally annotated location

[120]. Constraints can be either experimental results or bioinformatical evidence. In the first molecular biology experiments transposons were used to create random chimera proteins [52], later more specific molecular biology techniques were applied to investigate the topology of TMPs of interest (for review of these techniques see [144, 148]). The continuous development of biotechnology allows scientist to analyse the topology of all TMPs in an organism. In the topology analyses of *E. coli* and *S. cerevisiae*, the results of C-terminal fusion proteins were applied as constraints [28, 31, 67, 68, 120]. Recently high through-put techniques became available, where the surface of a living cell is labeled by chemical agents and the labeled peptides are investigated by coupled analytical technique after purification and degradation [13, 46, 93, 101]. In TOPDB more than 4,500 experimental results were collected for ~1,500 TMPs, and these constraints were applied to make constrained topology predictions for the ~1,500 TMPs. Regarding bioinformatical approaches, locations of compartment specific domains and sequence motifs can also be used as constraints. Such domains and motifs were collected into the TOPDOM database [144] from various databases such as SMART [45, 84], Pfam [36] and Prosite [136] for the purpose of constrained prediction.

9.4.2 Tertiary Structure Prediction of Transmembrane Proteins

Despite of the theoretical and computational difficulties, during the last two decades scientists have developed valuable methods to approximately model the tertiary structure of TMPs. Predicting TMP structures, at first, seems to be a relatively easy problem compared to understanding soluble protein structures. The fact that many TMPs share similar folds even with marginal sequence identities [43, 129] proves that TMPs are more structurally conserved than globular proteins. This is due to the strict conformational constraints that come from the membrane lipid bilayer, which dramatically decreases the size of the conformational space. However, the presence of an additional environment may cause previously unforeseen difficulties.

There are three main strategies to solve the tertiary structure of unknown TMP sequences. Homology modeling can be used when there is a sequential homologue with sequence identity greater than 20 %. In the case when no sequential homologue is available but (ideally) all folds are known, one can use threading methods to select the best packing of the query sequence. When neither sequential relative nor all folds are captured solely, de novo methods are still usable. It is worth mentioning that the order of this enumeration reflects the reliability of the methods as well (Fig. 9.1). Therefore it's not surprising that—due to the fundamentally unfeasible sampling of the whole structure space while looking for native structures—de novo methods are at the end of this list.

In the following sections we go through these three main families of transmembrane tertiary structure prediction strategies, namely comparative modeling, threading and de novo methods.

9.4.2.1 Comparative Modeling Techniques

Comparative modeling (also known as homology modeling) is a structure building strategy for unknown protein structures, which can be used when at least one sequential homologue with known structure is available for a given query TMP. The 3D structure of the sequentially homologue protein are used as a template (or target). Once the template has been selected and an alignment is generated between the template and target sequences, the non-conserved residues are replaced and insertions (regions with no template structure) are modeled as loop regions using de novo methods [117]. It is important to note that as for globular proteins, the accuracy of a homology model is strongly dependent on the identity between the two sequences [38].

While this technique basically relies on sequence alignments, at first, we have to declare the sequence identity level from where two TMPs can be considered as structural homologue. It was shown for globular proteins [125] that proteins with 30 % sequence identity the probability of sharing the same fold is ~90 % (below 25 % identity this probability drops to 10 %), in alignments longer than 80 residues. Although the application of this well-known fact has become second nature for researchers in the case of globular proteins, shedding light on the twilight zone (where structural similarity starts to diverge rapidly as sequential identity decreases) of TMPs is only a recent improvement [38, 108]. This lagging is due to the difficulties in experimental structure determination methods [75] applied and its consequence, the relatively small number of known transmembrane structures.

In a recent study [108], sequence–structure relation was analyzed using TMP structures with resolution <4 Å. It was found for the membrane region of TMPs that at >35 % sequence identity the structure RMSDs (RMSD—Root Mean Square Deviation) were 0.89 ± 0.43 Å and 0.80 ± 0.32 Å for α-helical and β-barrel membrane proteins, respectively. In addition, at 20–30 % sequence identity RMSDs increased—as expected—to 1.59 ± 0.55 Å and 1.30 ± 0.35 Å. According to expectations, TMPs show lower RMSD values than globular proteins, as structure in the membrane region is more conserved or restricted than in the non-membrane regions. Consequently, in the case of membrane regions of TMPs it is possible to use structures even with low sequence identity (<20 %) for comparative modeling. Moreover, β-barrel architecture seems much more robust to sequence variations. They found that sequence–structure similarity is generally independent of the number of membrane regions. The authors [108] concluded that functional mechanisms are preserved by high structural conservation and their functional specificity is mainly determined by the variable solvent-exposed regions.

Although homology modeling of globular proteins is a tried-and-true technique to predict 3D structure of query sequences having a sequential homologue, but in the field of TMPs this approach is in its infancy. There are some examples for modeling GPCR receptors [4, 43], but there isn't any fully automated, membrane protein specific method. Other, non-specific methods [122] are used as well, but the constraints imposed by the membrane are not utilized in the modeling, and the applied scoring functions designed for globular proteins might lead to distorted models.

Neglecting the scoring function and other technical details, a typical template-based modeling protocol can be briefly described in the following steps. At first,

query protein searched against a related database containing TMP sequences with known structure and one or more homologue templates are selected based on their sequential identity. Next, query sequence is aligned to all template sequences. These steps are usually merged and performed together, while most methods for detecting templates rely on the production of sequence alignments. As known, the primary criteria of database search algorithms is speed, therefore alignments resulted in database searches may not be as accurate as alignment produced by non-searching techniques. However, these kinds of algorithms are widely used to detect, and to generate alignment for homologue templates from database, e.g. PSI-BLAST [1, 131] and HHsearch [138]. The alignment of the target to template sequence(s) is the most important step of the whole procedure. Aligning transmembrane sequences used to be a long-standing unsolved problem, but by now numerous TMP sequence aligner methods have been developed, e.g. AlignMe [140] and MP-T [53]. According to Forrest et al. [38], comparative modeling of TMPs has been estimated to obtain accuracy as high as that of soluble proteins if the alignment for TMPs achieves the accuracy of its soluble protein counterpart.

The last step is the coordinate generation based on the alignment. For predicting the conformation of loop regions one can use Loopy [163], which is one of the fastest or PLOP [57], which is one of the most accurate techniques. FREAD [22] uses environment specific scoring parameters to improve the sampling for their loop structure prediction algorithm. RAPPER [29] and FALCm4 [81] rely on fine-grained residue-specific φ/ψ propensity tables for conformational sampling. Recently a coarse-grained method for loop prediction [90] was also developed of which computational time scales better than others, while the accuracy was preserved.

To highly increase the accuracy of the final structure, a genetic algorithm developed by John et al. [58] can be used to iteratively build better alignment for distant homologues. This method builds target-template alignments and structure models, and after assessing generated models, the alignments of the best models are used for generating further alignments.

Here we sketched the basic principles of the homology modeling techniques, in the following we review some recent methods based on comparative modeling.

A web server for homology modeling of TMPs named Memoir [34] is a pipeline utilizing iMembrane [65], a membrane protein annotator using CGDB [21] coarse-grained database; MP-T [53] target-template aligner; Medeller [66], a coordinate generator and FREAD [22], a loop modeler. Memoir does not search for a homologue template, therefore it needs this as an input parameter and does not provide any information on the reliability of the resulting structures.

A novel method, GPCRM [79] is developed for GPCR membrane protein structure predictions with averaging of multiple template structures and profile-profile comparison. It also utilizes two distinct loop modeling techniques: Modeller [154] and Rosetta [124] and excluding models with lipid penetrated loops.

At the border of homology and de novo modeling, the SWISS-MODEL [4] 7TM interface is developed for the modeling of TMPs with 7 transmembrane helices. SWISS-MODEL 7TM performs homology modeling on experimental and theoretical templates; to use this server user needs to provide the location of TMHs in the query sequence and also a template.

9.4.2.2 Threading Algorithms

Threading becomes very useful in cases when a sequence does not have any sequential relative with a known structure. This is a common scenario in the case of TMPs, where only a highly restricted number of TMPs show significant sequence identity to any known structure. As a consequence, homology modeling techniques discussed above have serious limitations which can be bridged using threading. Nevertheless, for building an efficient and reliable pipeline first we need a representative structure set of the conformational space of TMPs.

As discussed in Sect. 9.3, only a very small ratio, about the one fifth of the TMP structure space is known. Therefore, an efficient threading algorithm must not only find the structure with the lowest energy, but it has to discriminate native and the 'most-stable' decoy structures as well. These structure assessing algorithms are discussed in Sect. 9.4.2.4.

Due to the significant physico-chemical differences between soluble and membrane proteins, threading methods developed for globular proteins cannot be used directly, however a few methods have been customized for TMPs.

TASSER [165] is a two-step method that threads the sequence onto parts of solved protein structures and then refines the resulting template. The method was validated on a set of 38 non-homologue TMP structures, a little fewer than half of which have the RMSDs less than 6.5 Å compared to the native structure, but in the other cases RMSDs are in excess of 10 Å. It was used systematically to predict human GPCRs and these seemed consistent with experimental data. However, when there was no significant sequential relative, it was uncertain if the results represent the native structure.

A recent method, TMFR [158] is a sequence based fold recognition algorithm and has the accuracy of 49.2 and 82.2 % for α-helical and β-barrel TMPs, respectively. It utilizes topological features which improve the fold recognition [49] and can accurately align the target sequence to the template structure and generate reliable alignment raw scores to evaluate the structural similarity between the target and template. This provides practically only a sequence alignment. Therefore, algorithm traces back structure prediction problem to something akin to homology modeling.

However, this type of approximation widen the horizon of TMP structure prediction significantly, unfortunately the lack of structural representatives limits the usability of threading henceforward. In the next subsection, de novo methods are discussed which try to get over these difficulties.

9.4.2.3 De Novo Methods

De novo modeling does not use homologue proteins of known structures to predict the structure of an unknown protein. For an effective de novo structure prediction method there are two crucial requirements: accurate energetic representation of a protein structure and an efficient sampling of conformational space [82]. While structural

space expands rapidly with the sequence length, these methods are mainly applicable for small soluble proteins [16], not for TMPs, which are often large structures [158]. Although methods do not have restrictions on the number of known structures as homology modeling or threading do. However, as combinatorial approaches these require large amounts of computing time which often cannot be run on a single desktop computer, hence reducing their availability for structural biologists.

Contact Aided Structure Prediction

Contact prediction methods originates from the article, written by Göbel et al. [42], that describes how one could infer spatial information from multiple sequence alignments. This concept is based on the observation that the structure is more conserved than the sequence. Therefore, if a residue fulfilling structurally important role in a protein mutates, than another spatially close residue has to change to preserve both the structure and the function. Later it turned out that this assumption is a poor approximation of real proteins and their evolutionary processes.

Contact prediction methods can be classified into two main categories, namely local and global methods. The first one contains the 'classical' correlated mutation algorithms (CMA), which could be subdivided into further subcategories. To extract spatially close residues from multiple sequence alignments simple covariance analysis with various substitution matrices [42, 109]), χ^2-test [64], information theoretic approaches [33], machine-learning [20, 103, 118], alignment perturbation (SCA [88], ELSC [30]), probabilistic and empirical matrix methods or formal language [155, 156] were used. Further on, consensus methods are developed [41], which did not succeed to significantly overcome the performance of the previous methods (Acc. ~10 %, see Ref. [55]). CASP10 [95] confirmed the need for the development of contact prediction methods. On a test set of newly identified structures the best algorithm performed at an accuracy of ~30 %. Machine learning methods (PROFcon [118], CMAPpro [20], MEMPACK [106], PhyCMAP [159]) generally outperform others based on statistical considerations. Despite of the better predictor abilities, machine learning based approaches make their results difficult to interpret biophysically. In addition, the lack of a physical model makes the limits of their usability ill-defined. A recent study [47] showed that using three representatively selected contact prediction methods, there is no such linear combination of selected local techniques which could reach a satisfiable performance level. In addition, when a consensus method was trained and tested on only two, ABC-B and ABC-C protein families, despite of a nearly over parameterized model, these techniques could not reach a satisfying performance limit.

The main problem is that the observable correlations among sites do not stem from spatial closure purely. Atchley et al. [5] formalized the sources of these correlations, that—apart from structural constraints—could came from phylogenetic noise, function and higher-order statistical non-independence of positions. In addition, random noise or uneven sampling could bias measured correlations as well. The orders of magnitude of these factors are investigated by Noivirt et al. [102];

they found that correlation from structural, functional and phylogenetic constraints are in the same order of magnitude. Therefore, using background co-evolution signal correction [33] proved to be a valuable tool to reduce phylogenetic noise and to increase precision of methods significantly albeit even these precisions remains low.

Even if we neglect the disturbingly high correlation from functional and phylogenetic sources, there still remains a significant problem, namely disentangling direct and indirect interactions [17]. Global methods can take it into account with estimating joint probabilities of multiple residues. For reliable statistics huge number of sequences is required, which is a limiting factor still could not be overcame yet. Burger and van Nimwegen [17] had developed a Bayesian-network based method, which can take into account that the probability for residues to be in contact depends on their primary sequence separation and that highly conserved residues tend to participate in a larger number of contacts [17]. With this or other methods using maximum-entropy model [77, 97], sparse inverse covariance estimating [59] approaches could break through the barriers set by indirect contacts and multiple correlations. This network-based conceptional change of view and the increased size of sequence families result in significant performance gain.

Another possibility for calculating structural constraints is the prediction of helical interaction only, instead of predicting directly residue—residue contacts. It is an easier task than identifying all individual residue contacts in an α-helical TMP. However, this does not give any information on the orientation of helices and all the helices are treated as perpendicular to the membrane plane [40, 100]. Using propensity estimation techniques, e.g. lipid exposure predictors, the precision of helix-helix interaction and orientation estimations can be improved [92, 103, 116]. It has been known for a while that the tilted orientation of transmembrane helices is a principal compensation mechanism for hydrophobic mismatch [111]. Nevertheless, spanning regions are not necessarily straight: kinked or bended helices exist as well [76, 152], which complicates helical contact prediction even further.

These methods are valuable tools in themselves, which help to get closer to the biological understanding of TMP structures, functions and their mutational processes, but unfortunately they are still not as trustworthy as e.g. topology prediction techniques.

It is worth to mention, contact predictions cannot be used to directly reconstruct the 3D structure of proteins [110] not even using perfect predictor, not even for TMPs. This is due to their contradictory results originates from the oligomerization or conformational changes of the studied proteins. In the case of oligomers we would need to distinguish between intra- and interchain contacts. Another problem arises from multiple conformations of proteins, as in the case of the open and closed conformations of the *E. coli* GlpT or human OCTN1 [96]. When neither conformation change nor oligomerization has an influence on the inspected protein structure, theoretically an essential set of structure determining residue contacts is enough to replicate the 3D structure [130].

If we approximate the given problem from a reverse way, we could use predicted contacts as constraints in simulated annealing simulations [54, 91] or to aid the separation of native like TMP folds from decoys [92, 103, 127]. Obviously,

one could use experimental techniques, such as e.g. NMR, to earn useful structural constraints to build TMP model structures, but in this section we discussed this problem from the computational point of view only.

Forcefield-Based Approaches

There are many forcefield-based methods for determining the 3D structure of proteins; here we review some of those that were developed for TMPs. Given the structural simplicity observed in the β-barrel conformational space, structure prediction methods focused on estimating structures of α-helical TMPs. This imbalance will be observable in this paragraph as well.

In the early studies, such as Fleishman and Ben-Tal [37] residue environment preferences were used to predict the likely arrangement of transmembrane helices, and they were able to predict the native structure of TMP glycophorin A. Ledesma et al. [80] suggesting a model for the uncoupling protein 1 (UCP1), utilizing a computational docking method. Chen and Chen [19] used a lattice model of membrane proteins with a composite energy function to study their folding dynamics and native structures in Monte Carlo simulations. This model successfully predicts the seven helix bundle structure of sensory rhodopsin I by employing a three-stage folding. FILM [112] was developed for predicting small TMP structures based on assembling super-secondary segments taken from a protein structure library. The native structure is searched by simulated annealing. The main limitation of FILM is that the potential function is not able to reproduce the compactness of transmembrane bundles.

RosettaMembrane [164] (a derivation of Rosetta [124]) uses an all-atom physical model to describe intra-protein and protein-solvent interactions in the membrane environment. The surrounding environment is divided into 5 layers: water-exposed, polar, interface, outer and inner hydrophobic in both directions of the membrane core. Here a log-likelihood pair and environment potential were used, which penalizes steric overlap but favours packing density like characteristics of membrane proteins and strands. The method was tested on 12 membrane proteins with known structure, The length of the query sequences were between 51 and 145, which was predicted with RMSD <4 Å. However, mainly due to the technically unfeasible sampling of the conformational space this method performs poorly for large and complex proteins, independent on being soluble or transmembrane. In a newer version [7] of RosettaMembrane, experimental and predicted constraints were used to aid structure prediction. A great advantage of this method is that it can take into account cofactors, which could significantly modify structures.

BCL::MP-Fold [160] uses a three layer (solution, transition, membrane) implicit membrane representation with transition regions and a knowledge-based potential derived using Bayes' theorem and the inverse-Boltzmann relation. The final score resulted as the linear combination of many energy terms with optimized weights. The search for the native structure starts from randomly placed helices oriented perpendicular to the membrane plane. As a next step, folding is performed with simulated annealing [63].

As GPCRs are the most common drug targets, specific methods for predicting GPCR structures, such as MembStruk [85, 147] and PREDICT [135] have been developed as well. Both methods provide full-atom models for GPCRs on the basis of physico-chemical principles. In the PREDICT algorithm a concept of structural decoys is employed to ensure that the algorithm identifies the correct structure and to avoid trapping in a local minimum.

3D-SPOT [99] is a template-free method utilizing a statistical mechanical model [98] and an empirical potential function; TMSIP [56] is to predict the 3D structure of a given β-barrel TMP. While this method is based on physical interactions and does not require template structures, it can be applied for predicting structures of novel folds. The method performs well; in a blind test it was able to generate accurate structures of the transmembrane regions with a median main chain RMSD of 3.9 Å, on a set of 23 proteins.

9.4.2.4 Validating Predicted 3D Structures

Several methods have been developed for judging the reliability of predicted structures and identifying erroneous regions. Methods like PROCHECK [78] can be used for TMPs as well, because it takes into account only fundamental properties, namely the ψ/φ backbone torsion angles. There are various attempts to develop a measure, like the normalized QMEAN score [10] for soluble proteins, describing absolute quality of each structure for membrane proteins too. Phatak et al. [114] described a method filtering near-native structures from decoys using low-complexity Support Vector Regression models for predicting relative lipid accessibility (RLA). The quality assessment is based on the consistency of the predicted and observed RLA profiles. ProQM [121] utilizes SVM and membrane protein specific features, tested on GPCR structures. As it turned out, this method is capable of disentangling correct models from incorrect ones and has the ability to identify poor quality regions. IQ (Interaction-based Quality assessment) [50] incorporates four types of inter-residue interactions and achieves high prediction power on the independently constructed dataset (GPCR Dock 2008 (206 models), GPCR Dock 2010 (284 models), and HOMEP (92 models)). However, further validation of this method is needed. Recent results suggest that among conformations very dissimilar to native structures, this scoring function cannot correctly identify the best one. This is largely understandable since this scoring function relies primarily on the number of hydrophobic interactions. Lots of incorrectly formed hydrophobic interactions in decoy conformations could bias the IQ value (Li Zhijun personal communication).

9.4.3 Quaternary Structures of Membrane Proteins

As it was discussed earlier, genome-wide analysis of domain combinations of helical membrane proteins revealed that α-helical TMPs exist mostly as single domains. Oligomerization within the membrane may could be the general

mechanism for membrane proteins to gain new biological functions [83, 86, 87]. Therefore, discovering principles of oligomer formation of TMPs is needed to understand theirs functions and to gain new therapeutic strategies.

For globular proteins there are various methods to predict oligomerization propensities. PQS [72] and PISA [51] can identify the biologically active oligomer from X-ray structures. However, these methods cannot be used for TMPs, therefore in the PDBTM [71, 142, 143] database a simple homology search was used for predicting oligomeric state of potentially existing novel transmembrane structures, independent on their type (α-helical or β-barrel).

Bowie [69] and coworkers predicted the structure of α-helical TMP oligomers (glycophorin A and M2 proton channel) using knowledge of the oligomer symmetry. They used a simple softened van der Waals potential and Monte Carlo minimization to pack ideal α-helices. Bordner [14] have developed a method to predict biding sites of TMPs using a Random Forest classifier, trained on residue type distributions and evolutionary conservation for individual surface residues, followed by spatial averaging of the residue scores. Random Forest predictions were first made for individual surface residues and then the resulting scores of nearby residues were averaged in order to arrive at the final prediction score. Docking based approaches for predicting oligomerization has been developed as well [18]. In a recent review, the suitability of some widely-used docking algorithms for modeling complexes of α-helical TMPs was studied and the dependence of the docking performance on the protein features discussed as well [61].

Although α-helical TMPs pose a greater challenge, the oligomerization state of β-barrel membrane proteins can be accurately predicted computationally [98]. Based on the TMSIP [56] empirical potential function and the reduced conformational state model, extensive and contiguous weakly stable regions in many β-barrel membrane proteins seem to be an indicator of oligomerization propensities of β-barrel membrane proteins. Furthermore, as structural information is not essential for such predictions, the oligomerization state can also be predicted quite successful even when only sequence information is considered [98].

As it was discussed, there are various methods to model the quaternary structure of a TMP if it is a homomer or all the different subunits are known. Cases when the other subunits are unknown cannot be solved yet.

9.5 Orientation of Membrane Proteins in the Lipid Bilayer

Neither monomeric nor oligomeric TMPs do not exist alone without the amphiphilic membrane bilayer. By removing the hydrophobic environment the native structure breaks down. For experimental structure determination special handle with detergent is needed to extract TMP from the membrane and to preserve its native structure. Accordingly during experimental structure determination of TMPs, the information on the orientation disappears. While this information is essential for understanding the biological function and the mechanism of action of TMPs, experimental methods cannot recover it and thus has to be defined using

computational techniques. Orientation and burial of TMPs are very important e.g. for drug design to identify accessible parts of TMPs.

There are various attempts to predict orientation and burial of TMPs. One of the first methods was IMPALA [8], which uses amino acid propensities. TMDET [141] algorithm utilizes a geometrical algorithm to locate the most probable orientation of the given TMP in the membrane slab. OPM [89] applies a more sophisticated description of the problem, but does not outperform TMDET significantly. Senes et al. [134] have developed an empirical low-resolution potential called E_z, for protein insertion in the lipid membrane. A recent paper describes a method for predicting membrane protein orientation using a knowledge-based statistical potential [105, 133].

References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
2. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425
3. Arai M, Ikeda M, Shimizu T (2003) Comprehensive analysis of transmembrane topologies in prokaryotic genomes. Gene 304:77–86
4. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22(2):195–201
5. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17(1):164–178
6. Bagos PG, Liakopoulos TD, Hamodrakas SJ (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. BMC Bioinform 7:189
7. 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
8. Basyn F, Spies B, Bouffioux O, Thomas A, Brasseur R (2003) Insertion of X-ray structures of proteins in membranes. J Mol Graph Model 22(1):11–21
9. Bayrhuber M, Meins T, Habeck M, Becker S, Giller K, Villinger S, Vonrhein C, Griesinger C, Zweckstetter M, Zeth K (2008) Structure of the human voltage-dependent anion channel. Proc Natl Acad Sci USA 105(40):15370–15375
10. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27(3):343–350
11. Bernsel A, Viklund HK, Falk J, Lindahl E, von Heijne G, Elofsson A (2008) Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci USA 105(20):7177–7181
12. Berry EA, Guergova-Kuras M, Huang L-S, Crofts AR (2000) Structure and function of Cytochrome bc complexes. Annu Rev Biochem 69(1):1005–1075
13. Biller L, Matthiesen J, Kühne V, Lotter H, Handal G, Nozaki T, Saito-Nakano Y, Schümann M, Roeder T, Tannich E, Krause E, Bruchhaus I (2014) The cell surface proteome of Entamoeba histolytica. Mol Cell Proteomics MCP 13(1):132–144
14. Bordner AJ (2009) Predicting protein–protein binding sites in membrane proteins. BMC Bioinform 10:312
15. Bowie JU (1999) Helix-bundle membrane protein fold templates. Protein Sci 8(12):2711–2719

16. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871
17. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6(1):e1000633
18. Casciari D, Seeber M, Fanelli F (2006) Quaternary structure predictions of transmembrane proteins starting from the monomer: a docking-based approach. BMC Bioinform 7(1):340
19. Chen C-M, Chen C-C (2003) Computer simulations of membrane protein folding: structure and dynamics. Biophys J 84(3):1902–1908
20. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinform 8:113
21. Chetwynd AP, Scott KA, Mokrab Y, Sansom MSP (2008) CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. Mol Membr Biol 25(8):662–669
22. Choi Y, Deane CM (2010) FREAD revisited: accurate loop structure prediction using a database search algorithm. Proteins 78(6):1431–1440
23. Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357(6379):543–544
24. Chou K-C, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1(5):429–433
25. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
26. Cserzö M, Bernassau JM, Simon I, Maigret B (1994) New alignment strategy for transmembrane proteins. J Mol Biol 243(3):388–396
27. Cserzö M, Wallin E, Simon I, von Heijne G, Elofsson A (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 10(6):673–676
28. Daley DO, Rapp M, Granseth E, Melén K, Drew D, von Heijne G (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. Science 308(5726):1321–1323
29. de Bakker PIW, DePristo MA, Burke DF, Blundell TL (2003) Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. Proteins 51(1):21–40
30. Dekker JP, Fodor A, Aldrich RW, Yellen G (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 20(10):1565–1572
31. Drew D, Sjöstrand D, Nilsson J, Urbig T, Chin C-N, de Gier J-W, von Heijne G (2002) Rapid topology mapping of *Escherichia coli* inner-membrane proteins by prediction and PhoA/GFP fusion analysis. Proc Natl Acad Sci USA 99(5):2690–2695
32. Driessen AJ, Rosen BP, Konings WN (2000) Diversity of transport mechanisms: common structural principles. Trends Biochem Sci 25(8):397–401
33. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24(3):333–340
34. Ebejer J-P, 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–W383
35. Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179(1):125–142
36. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251
37. Fleishman SJ, Ben-Tal N (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices. J Mol Biol 321(2):363–378
38. Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91(2):508–517
39. Friedrich T, Böttcher B (2004) The gross structure of the respiratory complex I: a Lego system. Biochim Biophys Acta Bioenerg 1608(1):1–9

40. 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
41. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D (2007) Co-evolving residues in membrane proteins. Bioinformatics 23(24):3312–3319
42. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317
43. Gonzalez A, Cordom A, Caltabiano G, Pardo L (2012) Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors. ChemBioChem 13(10):1393–1399
44. Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. Proteins 35(4):408–414
45. Granseth E, Daley DO, Rapp M, Melén K, von Heijne G (2005) Experimentally constrained topology models for 51,208 bacterial inner membrane proteins. J Mol Biol 352(3):489–494
46. Gu B, Zhang J, Wang W, Mo L, Zhou Y, Chen L, Liu Y, Zhang M (2010) Global expression of cell surface proteins in embryonic stem cells. PLoS One 5(12):e15795
47. Gulyás-Kovács A (2012) Integrated analysis of residue coevolution and protein structure in ABC transporters. PLoS One 7(5):e36546
48. Harte R, Ouzounis CA (2002) Genome-wide detection and family clustering of ion channels. FEBS Lett 514(2–3):129–134
49. Hedman M, Deloof H, Von Heijne G, Elofsson A (2002) Improved detection of homologous membrane proteins by inclusion of information from topology predictions. Protein Sci 11(3):652–658
50. Heim AJ, Li Z (2012) Developing a high-quality scoring function for membrane protein structures based on specific inter-residue interactions. J Comput Aided Mol Des 26(3):301–309
51. Henrick K (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23(9):358–361
52. Herrero M, de Lorenzo V, Neilands JB (1988) Nucleotide sequence of the iucD gene of the pColV-K30 aerobactin operon and topology of its product studied with phoA and lacZ gene fusions. J Bacteriol 170(1):56–64
53. Hill JR, Deane CM (2013) MP-T: improving membrane protein alignment for structure prediction. Bioinformatics 29(1):54–61
54. 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
55. Horner DS, Pirovano W, Pesole G (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 9(1):46–56
56. Jackups R, Liang J (2005) Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol 354(4):979–993
57. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55(2):351–367
58. John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 31(14):3982–3992
59. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2011) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190
60. Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33(10):3038–3049
61. Kaczor AA, Selent J, Sanz F, Pastor M (2013) Modeling complexes of transmembrane proteins: systematic analysis of protein–protein docking tools. Mol Inform 32(8):717–733
62. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338(5):1027–1036

63. Karakaş 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
64. Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 48(4):611–617
65. Kelm S, Shi J, Deane CM (2009) iMembrane: homology-based membrane-insertion of proteins. Bioinformatics 25(8):1086–1088
66. Kelm S, Shi J, Deane CM (2010) MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 26(22):2833–2840
67. Kim H, Melén K, Osterberg M, von Heijne G (2006) A global topology map of the *Saccharomyces cerevisiae* membrane proteome. Proc Natl Acad Sci USA 103(30):11142–11147
68. Kim H, Melén K, von Heijne G (2003) Topology models for 37 *Saccharomyces cerevisiae* membrane proteins based on C-terminal reporter fusions and predictions. J Biol Chem 278(12):10208–10213
69. Kim S, Chamberlain AK, Bowie JU (2003) A simple method for modeling transmembrane helix oligomers. J Mol Biol 329(4):831–840
70. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420(6912):218–223
71. Kozma D, Simon I, Tusnády GE (2012) PDBTM: protein data bank of transmembrane proteins after 8 years. Nucleic Acids Res 41(D1):D524–D529
72. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372(3):774–797
73. 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
74. Kyte J, Doolittle R (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
75. Lacapère J-J, Pebay-Peyroula E, Neumann J-M, Etchebest C (2007) Determining membrane protein structures: still a challenge! Trends Biochem Sci 32(6):259–270
76. 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
77. Lapedes A, Giraud B, Jarzynski C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv:1207.2484
78. Laskowski R, Macarthur M, Moss D, Thornton J (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291
79. Latek D, Pasznik P, Carlomagno T, Filipek S (2013) Towards improved quality of GPCR models by usage of multiple templates and profile-profile comparison. PLoS One 8(2):e56742
80. Ledesma A, de Lacoba MG, Arechaga I, Rial E (2002) Modeling the transmembrane arrangement of the uncoupling protein UCP1 and topological considerations of the nucleotide-binding site. J Bioenerg Biomembr 34(6):473–486
81. Lee J, Lee D, Park H, Coutsias EA, Seok C (2010) Protein loop modeling by using fragment assembly and analytical loop closure. Proteins 78(16):3428–3436
82. Lee J, Lee J, Sasaki TN, Sasai M, Seok C, Lee J (2011) De novo protein structure prediction by dynamic fragment assembly and conformational space annealing. Proteins 79(8):2403–2417
83. Lehnert U, Xia Y, Royce TE, Goh C-S, Liu Y, Senes A, Yu H, Zhang ZL, Engelman DM, Gerstein M (2004) Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. Q Rev Biophys 37(2):121–146
84. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res 32:D142–D144
85. Li Y, Goddard WA (2008) Prediction of structure of G-protein coupled receptors and of bound ligands, with applications for drug design. Pac Symp Biocomput 344–353

86. Liang J, Naveed H, Jimenez-Morales D, Adamian L, Lin M (2012) Computational studies of membrane proteins: models and predictions for biological understanding. Biochim Biophys Acta 1818(4):927–941
87. Liu Y, Gerstein M, Engelman DM (2004) Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. Proc Natl Acad Sci USA 101(10):3495–3497
88. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286(5438):295–299
89. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. Bioinformatics 22(5):623–625
90. Macdonald JT, Kelley LA, Freemont PS (2013) Validating a coarse-grained potential energy function through protein loop modelling. PLoS One 8(6):e65770
91. 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
92. Miller CS, Eisenberg D (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models. Bioinformatics 24(14):1575–1582
93. Mindaye ST, Ra M, Lo Surdo J, Bauer SR, Alterman MA (2013) Improved proteomic profiling of the cell surface of culture-expanded human bone marrow multipotent stromal cells. J Proteomics 78:1–14
94. Minsky M, Seymour P (1969) Perceptrons. MIT Press, Oxford
95. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2014) Evaluation of residue-residue contact prediction in CASP10. Proteins 82(Suppl 2):138–153
96. Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci USA 110(51):20533–20538
97. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108(49):E1293–E1301
98. Naveed H, Jackups R, Liang J (2009) Predicting weakly stable regions, oligomerization state, and protein–protein interfaces in transmembrane domains of outer membrane proteins. Proc Natl Acad Sci USA 106(31):12735–12740
99. Naveed H, Xu Y, Jackups R, Liang J (2012) Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. J Am Chem Soc 134(3):1775–1781
100. Neumann S, Fuchs A, Hummel B, Frishman D (2013) Classification of α-helical membrane proteins using predicted helix architectures. PLoS One 8(10):e77491
101. Niehage C, Steenblock C, Pursche T, Bornhäuser M, Corbeil D, Hoflack B (2011) The cell surface proteome of human mesenchymal stromal cells. PLoS One 6(5):e20399
102. Noivirt O, Eisenstein M, Horovitz A (2005) Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel 18(5):247–253
103. Nugent T, Jones DT (2010) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol 6(3):e1000714
104. Nugent T, Jones DT (2011) Membrane protein structural bioinformatics. J Struct Biol 179(3):327–337
105. Nugent T, Jones DT (2013) Membrane protein orientation and refinement using a knowledge-based statistical potential. BMC Bioinform 14(1):276
106. Nugent T, Ward S, Jones DT (2011) The MEMPACK alpha-helical transmembrane protein structure prediction server. Bioinformatics 27(10):1438–1439
107. Oberai A, Ihm Y, Kim S, Bowie JU (2006) A limited universe of membrane protein families and folds. Protein Sci 15(7):1723–1734
108. Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29(13):1589–1592
109. Olmea O, Rost B, Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 293(5):1221–1239
110. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Ab initio folding of proteins using restraints derived from evolutionary information. Proteins Struct Funct Genet 37(S3):177–185

111. Park SH, Opella SJ (2005) Tilt angle of a trans-membrane helix is determined by hydrophobic mismatch. J Mol Biol 350(2):310–318
112. 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
113. Phan G, Remaut H, Wang T, Allen WJ, Pirker KF, Lebedev A, Henderson NS, Geibel S, Volkan E, Yan J, Kunze MBA, Pinkner JS, Ford B, Kay CWM, Li H, Hultgren SJ, Thanassi DG, Waksman G (2011) Crystal structure of the FimD usher bound to its cognate FimC-FimH substrate. Nature 474(7349):49–53
114. Phatak M, Adamczak R, Cao B, Wagner M, Meller J (2011) Solvent and lipid accessibility prediction as a basis for model quality assessment in soluble and membrane proteins. Curr Protein Pept Sci 12(6):563–573
115. Pieper U, Schlessinger A, Kloppmann E, Chang GA, Chou JJ, Dumont ME, Fox BG, Fromme P, Hendrickson WA, Malkowski MG, Rees DC, Stokes DL, Stowell MHB, Wiener MC, Rost B, Stroud RM, Stevens RC, Sali A (2013) Coordinating the impact of structural genomics on the human α-helical transmembrane proteome. Nat Struct Mol Biol 20(2):135–138
116. Pilpel Y, Ben-Tal N, Lancet D (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. J Mol Biol 294(4):921–935
117. Punta M, Forrest LR, Bigelow H, Kernytsky A, Liu J, Rost B (2007) Membrane protein prediction methods. Methods 41(4):460–474
118. Punta M, Rost B (2005) PROFcon: novel prediction of long-range contacts. Bioinformatics 21(13):2960–2968
119. Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
120. Rapp M, Drew D, Daley DO, Nilsson J, Carvalho T, Melén K, De Gier J-W, Von Heijne G (2004) Experimentally based topology models for *E. coli* inner membrane proteins. Protein Sci 13(4):937–945
121. Ray A, Lindahl E, Wallner B (2010) Model quality assessment for membrane proteins. Bioinformatics 26(24):3067–3074
122. Reddy CS, Vijayasarathy K, Srinivas E, Sastry GM, Sastry GN (2006) Homology modeling of membrane proteins: a critical assessment. Comput Biol Chem 30(2):120–126
123. Remm M, Sonnhammer E (2000) Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs. Genome Res 10(11):1679–1689
124. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
125. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng Des Sel 12(2):85–94
126. Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86 % accuracy. Protein Sci 5(8):1704–1718
127. Sadowski MI, Maksimiak K, Taylor WR (2011) Direct correlation analysis improves fold recognition. Comput Biol Chem 35(5):323–332
128. Saier MJ, Beatty J, Goffeau A, Harley K, Heijne W, Huang S, Jack D, Jähn P, Lew K, Liu J, Pao S, Paulsen I, Tseng T, Virk P (1999) The major facilitator superfamily. J Mol Microbiol Biotechnol 1(2):257–279
129. Sansom MS, Shrivastava IH, Bright JN, Tate J, Capener CE, Biggin PC (2002) Potassium channels: structures, models, simulations. Biochim Biophys Acta Biomembr 1565(2):294–307
130. Sathyapriya R, Duarte JM, Stehr H, Filippis I, Lappe M (2009) Defining an essence of structure determining residue contacts in proteins. PLoS Comput Biol 5(12):e1000584
131. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29(14):2994–3005
132. Schmitt L (2002) Structure and mechanism of ABC transporters. Curr Opin Struct Biol 12(6):754–760

133. Schramm CA, Hannigan BT, Donald JE, Keasar C, Saven JG, Degrado WF, Samish I (2012) Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. Structure 20(5):924–935
134. Senes A, Chadi DC, Law PB, Walters RFS, Nanda V, Degrado WF (2007) E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. J Mol Biol 366(2):436–448
135. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M, Naor Z, Noiman S, Becker OM (2004) PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins 57(1):51–86
136. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347
137. Sipos L, von Heijne G (1993) Predicting the topology of eukaryotic membrane proteins. Eur J Biochem/FEBS 213(3):1333–1340
138. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960
139. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182
140. Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2013) Alignment of helical membrane protein sequences using AlignMe. PLoS One 8(3):e57731
141. Tusnády GE, Dosztányi Z, Simon I (2004) Transmembrane proteins in the protein data bank: identification and classification. Bioinformatics 20(17):2964–2972
142. Tusnády GE, Dosztányi Z, Simon I (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 33:D275–D278
143. Tusnády GE, Dosztányi Z, Simon I (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 21(7):1276–1277
144. Tusnády GE, Kalmár L, Hegyi H, Tompa P, Simon I (2008) TOPDOM: database of domains and motifs with conservative location in transmembrane proteins. Bioinformatics 24(12):1469–1470
145. Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17(9):849–850
146. Tusnády GE, Simon I (2010) Topology prediction of helical transmembrane proteins: how far have we reached? Curr Protein Pept Sci 11(7):550–561
147. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA (2002) Prediction of structure and function of G protein-coupled receptors. Proc Natl Acad Sci USA 99(20):12622–12627
148. van Geest M, Lolkema JS (2000) Membrane topology and insertion of membrane proteins: search for topogenic signals. Microbiol Mol Biol Rev 64(1):13–33
149. Viklund HK, Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 24(15):1662–1668
150. Volkan E, Kalas V, Pinkner JS, Dodson KW, Henderson NS, Pham T, Waksman G, Delcour AH, Thanassi DG, Hultgren SJ (2013) Molecular basis of usher pore gating in *Escherichia coli* pilus biogenesis. Proc Natl Acad Sci USA 110(51):20741–20746
151. von Heijne G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5(11):3021–3027
152. von Heijne G (1991) Proline kinks in transmembrane alpha-helices. J Mol Biol 218(3):499–503
153. von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225(2):487–494
154. Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
155. Waldispühl J, Berger B, Clote P, Steyaert J-M (2006) Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. Proteins 65(1):61–74

156. Waldispühl J, Steyaert J-M (2005) Modeling and predicting all-α transmembrane proteins including helix–helix pairing. Theor Comput Sci 335(1):67–92
157. Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 7(4):1029–1038
158. Wang H, He Z, Zhang C, Zhang L, Xu D (2013) Transmembrane protein alignment and fold recognition based on predicted topology. PLoS One 8(7):e69744
159. Wang Z, Xu J (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics 29(13):i266–i273
160. Weiner BE, Woetzel N, Karakaş M, Alexander N, Meiler J (2013) BCL:MP-fold: folding membrane proteins through assembly of transmembrane helices. Structure 21(7):1107–1117
161. White SH (2009) Biophysical dissection of membrane proteins. Nature 459(7245):344–346
162. Wolf YI, Grishin NV, Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. J Mol Biol 299(4):897–905
163. Xiang Z, Soto CS, Honig B (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 99(11):7432–7437
164. Yarov-Yarovoy V, Schonbrun J, Baker D (2006) Multipass membrane protein structure prediction using Rosetta. Proteins 62(4):1010–1025
165. Zhang Y, Devries ME, Skolnick J (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol 2(3):e13