Protein Structure Prediction: Methods and Algorithms, Edited by [Editor's Name].

ISBN 0-471-XXXXX-X

Chapter 17

Modeling and validation of transmembrane protein structures

Maya Schushan and Nir Ben-Tal

Abstract

Transmembrane (TM) proteins comprise some 15% to 30% of the proteome, and the number of reported 3D structures has grown rapidly over the past decade. Nevertheless, owing to technical difficulties, the vast majority of TM protein structures are yet to be determined. Computational modeling techniques can be used to provide the essential structural data needed to shed light on structure-function relationships in TM proteins.

In this chapter, we present some of the advanced modeling approaches that can help resolve the unique challenges encountered in predicting the three-dimensional (3D) structure of α-helical TM proteins. The usefulness of standard homology-modeling procedures is limited because the number of available TM-protein structures is small. In many cases, moreover, it is difficult to align the sequences of the query and the template proteins because of the weak sequence similarity between them. Additional ways to predict the location of TM helices in the polypeptide chain, by employing fold recognition, hydrophobicity scales, or other tools, may be helpful in improving the alignment accuracy. When a structural template is not available, low-resolution electron-density maps, obtained from cryo-electron microscopy (cryo-EM) or preliminary X-ray diffraction studies, can be used to restrict the search in conformational space. At the right resolution, the locations of TM helices can be roughly determined even when the amino acids are not visible. When these data are combined with physicochemical characteristics of amino acids (such as their hydrophobicity) and with evolutionary information, the location of the amino acids can be modeled.

After modeling, it is imperative to assess the quality of the structure and estimate the level of confidence of the prediction. To this end, it is often helpful to estimate the evolutionary conservation of the amino acids and project them onto the model-structure. The expectation is

that the protein core and functional regions (such as narrow channel pores and ligand-binding sites) will accommodate evolutionarily conserved amino acids while the periphery will be more variable. Deviations from this pattern might reflect inaccuracies in the model. Furthermore, because X-ray crystal structures of TM proteins are often determined on the basis of electron-density maps of limited resolution, it might be useful to examine their evolutionary profiles as an independent measure of their validity.

Contents

17.1 Introduction

TM proteins comprise an estimated 15% to 30% of bacterial and eukaryotic genomes [1-3]. As gateways to the cell, TM proteins participate in a variety of processes, such as energy production, transport of metabolites, and cell–cell communication. Structural information is needed in order to uncover the components that contribute to these diverse functions and to the structure-function relationship. In addition, TM-protein structure might provide an interpretation at the molecular level for mutations and enable structure-based drug design [4]. However, despite the substantially growing numbers of reported TM-protein structures, owing to technical difficulties only some have been experimentally solved to date [5]. As a result, TM-protein structures currently account for less than 2% of the Protein Data Bank (PDB) [6]. Moreover, most of the available structures are of bacterial origin, whereas only a small minority are of eukaryotic TM proteins [7].

Polytopic TM domains exhibit one of two possible folds: an α-helix bundle or a β-barrel [8-10]. The α-helical proteins are widespread, whereas distribution of the β-barrels is limited to mitochondria, chloroplasts, and the outer membranes of Gram-negative bacteria [9]. Because the two types display distinct characteristics, there is some variation in their 3D-modeling. In this chapter we deal only with the α-helical type. Owing to the uniqueness of the membrane environment, many features of TM proteins are quite distinct from those of soluble proteins (e.g. [11-14]). This has significant implications for the prediction of their structure.

17.1.1 Traits and topology of helical TM proteins

TM proteins display an amino-acid composition that is quite different from that of soluble proteins [11, 15, 16], for example, with regard to the proportion of hydrophobic residues [13, 17]. Strongly polar residues are less abundant in TM proteins than in soluble ones [18], as their transfer from the aqueous phase to the hydrocarbon region of the membrane is associated with a high energetic penalty [10, 19, 20]. As might be expected, the extra-membrane regions in the TM proteins, which interact with the aqueous phase, are much more hydrophilic than the membrane-embedded helices. There are two other noteworthy characteristics of TM proteins: von Heijne's "positive-inside" rule [21] (discussed in chapter 6) and the existence of "aromatic belts", i.e. an abundance of Trp and Tyr residues near both ends of the TM helices [15, 17]. Additionally, the structural context of proline residues in TM helices was explored in several studies [10, 22, 23], showing that in many cases proline residues induce distortions such as kinks or bends in central

regions of TM helices, where they contribute to function, conformational changes, and folding (e.g. [24-26]). In addition, Yohannan and co-workers showed that kinks often correspond to positions exhibiting an abundance of proline residues (>10%) in multiple sequence alignments (MSAs) [27]. Thus, even when a proline is not included in the sequence itself, identification of proline peaks in alignments might also offer some information about specific features of an α-helical TM protein.

The membrane topology of a query protein can often be identified on the basis of its amino-acid sequence, and is addressed in detail in chapter 6, with focus on TM prediction methods. Some 20 years ago, it was reported that algorithms for secondary structure prediction could not accurately predict the secondary structure of TM proteins [28]. A recent study, however, showed that modern algorithms, developed for soluble proteins, are almost as precise for TM proteins as they are for soluble proteins [29]. This suggested that TM and soluble α-helices are more similar than was previously assumed. Other studies have clearly demonstrated, however, that some of the secondary structure propensities of TM proteins are unique [14, 30].

A recent study showed that five kinds of specific interactions are abundant in TM structures, and can even be employed to correctly reassemble the native helix packing, starting from the backbone of the individual helices [31]. The contacts consist of hydrogen bonds, salt bridges, aromatic interactions, and packing of small and of aliphatic residues. These findings support the hypothesis that the interactions constrain the helix backbone, and facilitate folding and stability in TM proteins.

17.1.2 Fold space of helical TM structures

The contemporary view of the variety of folds presented by α-helical TM proteins differs significantly from the initial picture [7, 10, 16, 19]. In the past, α-helical TM structures were thought to be composed of strictly canonical helices that span the entire membrane in an approximately vertical direction, in correspondence with the first TM-protein structures to be solved (e.g. bacteriorhodopsin [32], Figure 1A). That view implied that the architecture of α-helical TM proteins was rather limited, and that their modeling might therefore be much simpler than that of water-soluble proteins, which manifest a variety of folds.

Once some additional structures were determined, however, it became clear that TM-protein structures can also possess non-canonical helices, half or discontinuous helices, and re-entrant loops. Furthermore, some helices are very short and do not span the entire membrane,

while others are extremely long and are tilted relative to the membrane plane. These observations are exemplified by the structures of the bacterial Na^+/H^+ exchanger NhaA [33] and the glycerol channel GlpF [34] (Figures 1B and 1C), as well as by numerous other TM structures. Overall, the fold space of the α-helical TM proteins is larger than initially estimated because their non-standard structural elements are broadly distributed. It is restricted, however, relative to the fold space of soluble proteins, owing to the membrane environment as well as the distinct secondary structural elements and composition [8-10, 16]. This implies that the development of specific computational modeling tools for α-helical membrane structures is likely to be more complicated than originally thought, but is nevertheless still attainable.

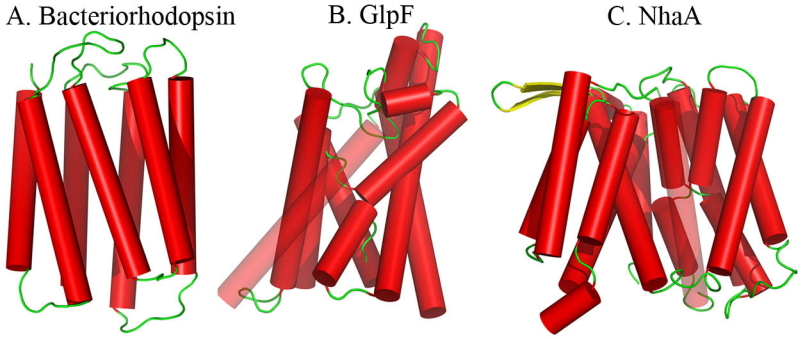

Figure 1. Simple vs. compound helical TM structures. The cytoplasmic side is downwards. TM helices are shown as red cylinders, beta-strands are yellow, and loops are green. In panels B and C, one of the TM helices is depicted as transparent for clarity. **A.** The structure of bacteriorhodopsin [32] shows a "classical" architecture, composed of almost straight helices spanning the membrane. **B.** In GlpF [34], many of the TM helices are tilted with respect to the membrane normal. The structure also features half-helices and intramembrane loops. **C.** The structure of NhaA [33] encompasses an assembly in which two of the segments are discontinued helices located opposite to one another. The structure also contains bent helices.

17.1.3 General computational approaches to the modeling of TM proteins

Given the scarcity of experimentally derived structures of TM proteins, especially those of human or other eukaryotic origin [5, 7, 35], computational modeling techniques provide an appealing alternative. Depending on the availability of data, there are in general three different modeling approaches: (a) comparative (or homology) modeling, (b) experimental data fitting, and (c) template-free prediction. For comparative modeling, the query protein should be related to a similar protein with a solved high-resolution structure. Methods known as fold recognition (or threading) are highly effective when the query protein has template structures that are difficult to detect on the basis of sequence similarity (addressed in chapters 9 and 10). Even in the absence of similarity to a known structure, experimental constraints combined with additional features such as the evolutionary profile might suffice to yield model-structures [7, 36]. In the next sections, we will discuss the principles and application of these two approaches.

Should these two alternatives fail, it is possible to resort to template-free prediction, an approach that has yet to become generally applicable [9, 37]. In principle, this approach utilizes only the rules of physics and chemistry to model the TM protein's structural features. However, "free modeling" techniques also include hybrid approaches. These methods incorporate the use of structural data in the form of libraries of the structures of short fragments, as well as "statistical potentials" that represent common proximities of amino acids or atoms in proteins [37, 38]. Actually, the most advanced approaches, such as Rosetta and TASSER, offer a unified modeling framework, combing the different modeling approaches to better address various modeling challenges [39].

Within this category, two methodologies, namely Rosetta and TASSER, have featured novel membrane-specific adaptations. The Rosetta methodology has been used to successfully predict the structures of small soluble proteins [40], and was recently modified for helical TM proteins with promising performance in several test cases [41-43]. The TASSER methodology (discussed in chapter 12) has also been adapted for TM proteins, and was applied to predict the structures of hundreds of human G-protein-coupled receptors (GPCRs) [44]. Recently, the structure of a human GPCR protein, the β2-adrenergic receptor, was solved experimentally [45]. To assess the performance of the above two computational approaches, we compared their predicted β2-adrenergic receptor models to the native structure. Barth and Baker *(unpublished results*) predicted the structure of the β2-adrenergic receptor via the membrane-modified version of Rosetta. The starting point for the Rosetta model was a model obtained via homology modeling, with the structure of bovine rhodopsin serving as template. In the case of the TASSER algorithm this comparison was actually a blind test since the model was published, within the TASSER database of GPCRs [44], before the experimental structure came out. In both cases the TM region of the β2-adrenergic receptor model was reasonably accurate, with root mean square deviation (RMSD) values of 1.54 Å over 212 Cα atoms for the final Rosetta model (Figure 2A) and 1.7Å over 199 Cα atoms for the best TASSER model (Figure 2B).

As might be expected, the extra-membrane regions were more difficult to predict than the TM regions. On examining the helical structural elements, we could see that the predicted TM4 in the Rosetta model was longer than the native TM4 in the X-ray structure (Figure 2A), whereas in the TASSER-derived model a longer helical segment relative to the native structure was predicted for TM6 (Figure 2B). Notably, one of the unique features of the β2-adrenergic receptor structure is an extra helix in the second extracellular loop (ECL2) (Figure 2). Both the TASSER and the

Rosetta models failed to predict this helix (Figure 2). However, the Rosetta model did predict a short helical segment in the region preceding the ECL2 helix (Figure 2A). Overall, both produced reasonably good models.

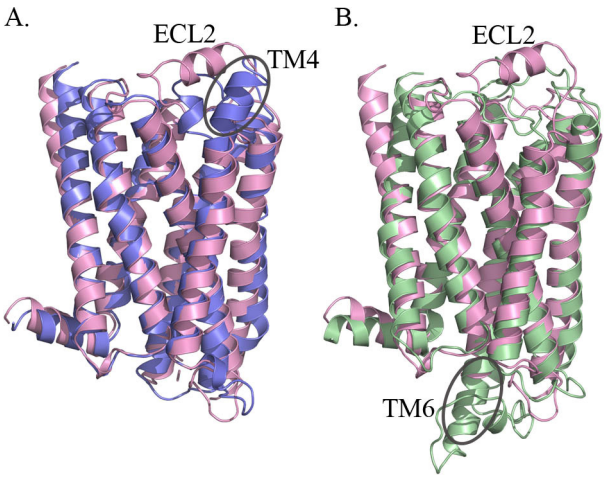

Figure 2. Performance of Rosetta and TASSER in predicting the structure of the human β2-adrenergic receptor. In both panels, the crystal structure of the β2-adrenergic receptor [45] is shown as pink ribbons; the cytoplasmic region is downward and the short helix in ECL2 is marked. The Rosetta model (Barth and Baker, *unpublished results*) (purple) and the best TASSER model [44] (green) are superimposed on the native structure in panels **A** and **B**, respectively. The prolonged segments in predicted TM4 and TM6 are marked, along with ECL2 of the native structure.

GPCRs, the largest family of TM signal-transduction proteins, include about 1000 human isoforms [46, 47]. Since they comprise approximately 50% of contemporary protein drug targets, there is particular interest in modeling their structures for purposes of drug discovery [48]. Up to now, efforts to model GPCRs have been based on a variety of computational approaches, ranging from homology modeling using the few GPCR structures available from experiments (reviewed, for example, in [49]) to specifically designed template-free methods (e.g. [44, 50, 51]), all of which are tailored for GPCRs. This is a research field of its own, and is beyond the scope of this chapter. The interested reader is referred to references [52-54].

An objective assessment of structure prediction methods is provided by the CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments. The object of these biennial experiments, which started in 1994, is to assess current abilities and inabilities in predicting protein structure. During the experiment, different groups submit blind modeling predictions of various proteins. These predictions are later compared to the native structures of the proteins, which have already been determined experimentally but are not yet known to the participating scientists during the experiment [55]. Unfortunately, TM proteins are not used as CASP targets; thus, there is currently no generally accepted way to assess, directly and without bias, the application of available modeling techniques for TM proteins.

17.2 Comparative modeling

In comparative (or homology) modeling, currently the leading computational approach for generating protein models, a high-resolution, experimentally solved structure is used to produce a

model-structure of a homologous protein for which structural data are not yet available [38, 56-58]. The technique has been successfully applied to numerous soluble proteins and is considered to be the most accurate approach to structural modeling available today [56-58]. Recently, Forrest and co-workers showed that comparative modeling can also be applied to TM proteins to produce models of similar accuracy to those of soluble proteins [29]. To this end, the structures of 11 TM-protein families were examined, covering a range of folds and sequence similarities. In line with previous observations for soluble proteins [58], the analysis of Forrest et al. [29] showed that the accuracy of TM model-structures produced via comparative modeling depends, as anticipated, on the similarity between the query and the template sequences. With decreasing similarity between the sequences the precision of the produced model structures also decreases, owing to two factors: alignment errors and inherent structural differences between the two proteins.

Using a previous division of TM proteins from 95 genomes into families [59], Granseth et al. examined the relationship between prokaryotic and eukaryotic TM proteins [35] with the object of assessing the extent to which comparative modeling could derive structural models of eukaryotic and even human TM proteins from available prokaryotic structures. Their analysis revealed that 13% of eukaryotic TM families include also members of the prokaryotic kingdom. Of these 256 families, solved structures exist for representatives of only 29 [35]. Although these data are not particularly encouraging, they nevertheless indicate that a significant number of eukaryotic TM models could be obtained by comparative modeling. In this respect it should also be noted that the sequence similarity between the eukaryotic query protein and the prokaryotic template is often low, further complicating the production of an accurate pairwise sequence alignment between them [29, 57, 58]. Since comparative modeling is largely dependent on pairwise alignment (discussed below), this presents a major obstacle in obtaining TM model-structures of high quality.

This difficulty is best illustrated by a description of some recent efforts to model TM human proteins based on their remote prokaryotic homologues. The serotonin transporter of the neurotransmitter:sodium symporter (NSS) family was modeled using the eubacterium *Aquifex aeolicus* leucine transporter (aaLeuT) as template [60]. An available alignment [61] was refined, because of low sequence identity between the prokaryotic and eukaryotic family members, by the use of various bioinformatics tools along with elaborate experimental data. Interestingly, the model-structure was utilized to identify a chloride ion-binding site in Cl^- dependent transporters, a prediction confirmed by experimental assays [60].

In another study, the extremely low sequence identity between the human Na^+/H^+ exchanger NHE1 and NhaA of *E. coli* (<15%) prompted a composite modeling approach in which various state-of-the-art computational modeling tools were integrated to achieve correct alignment of the two proteins [62]. Supported by elaborate mutagenesis, the model revealed common properties of the inhibitor-binding sites of NHE1 and NhaA, as well as a putative ion-transport mechanism for NHE1 [62].

In yet another example, although the intriguing cystic fibrosis TM conductance regulator (CFTR) is a chloride channel, the structure of Sav1866, a multi-drug transporter of the same superfamily, was used as template for its modeling [63]. To overcome the obstacle of low sequence identity, their divergent sequences were aligned through a technique of hydrophobic cluster analysis [64]. The model-structure, which demonstrated good correlation with experimental data, offers a molecular-level insight into the contacts that might be affected as a result of the deletion of Phe508, the most abundant cystic fibrosis-causing mutation [63].

17.2.1 Work scheme

The scheme for predicting a structure by means of homology modeling is generally the same for soluble and TM proteins. It can be divided into four major steps: (a) template search and selection, (b) pairwise sequence alignment of the query and the template sequences, (c) model building, and (d) evaluation and validation. It is noteworthy that depending on the outcome of the validation stage, it might be necessary to refine the model structure by repeating the previous stages. This cycle can be repeated until a model of the best possible quality is produced. Because of the paucity of experimentally solved TM structures on the one hand and the exclusive features of the proteins on the other, it is necessary to develop unique methods for each step. Accurate prediction of the membrane topology (addressed in chapter 6) is likely to be helpful in the first two stages, which are the keys to proper modeling. However, because the structural data are limited, the computational aspects of membrane-specific homology modeling are still underdeveloped [37]. In the following sections we provide a description of the fine points of these work steps for homology modeling of α-helical TM proteins. The last step, model evaluation and validation, will be presented in section 17.4, as it is identical for models generated via homology modeling (section 17.2) and via experimental data fitting (section 17.3).

17.2.2 Template search and selection

17.2.2.1 Simple and advanced search

When attempting to determine the structure of a TM protein, we first search for a potential template or templates. This is easy if the structure of a close homologue of the query protein has already been solved, but difficult if it has not. The number of available TM structures, however, is very small, which makes it hard to find suitable structural templates.

Because similar sequences adopt similar structures, the initial strategy in searching for a potential template is usually to employ sequence-search methods such as BLAST [65]. Many TM proteins possess water-soluble domains in addition to the TM region, which may appear at the N- or C-termini or between consecutive TM helices. By excluding large extra-membrane regions and using only the sequence of the TM domain to be matched with PDB-derived sequences, it may be possible to produce more accurate results since they will be focused only on the area of interest. It is worth noting that the "low-complexity filter" included among the common search tools might remove hydrophobic segments, and should therefore not be applied [66]. In addition, Hedman and co-workers showed that the search for homologues (tested on a benchmark of GPCR sequences) can be improved by utilizing predictions of the location of the TM segments in the sequence [66].

When the simple search for a template fails, advanced sequence-based search could be employed. It is also possible to apply fold recognition or threading algorithms to detect putative structural templates. These algorithms perform two modeling steps: template identification and alignment of the query and the template sequences [38]. Chapters 9 and 10 describe these approaches, which are currently employed in the same manner for both soluble and TM proteins.

17.2.2.2 Template selection

The next step is to select the most suitable template(s) from the detected hits. For TM proteins this is easy, simply because it is rare to find any templates at all. The challenge here is rather to estimate the suitability of a putative template whose sequence similarity to the query protein is often low. Because resemblance between the two sequences correlates with model accuracy [29], it is important -as in the case of soluble proteins- to assess their similarity and their evolutionary relationship [58]. The number of TM helices in the query protein and in the template proteins is likely to be the same. Thus, the known (or predicted) membrane topology of the query protein will probably aid in selection of the most suitable template. In view of the known relationship between structure and function, experimental evidence of functional similarity between the two

proteins might be taken as an indication of the template's suitability. This was done, for example, in the modeling of NHE1 based on the structure of NhaA [62].

17.2.3 Aligning the query and the template sequences

Aligning the query and the template sequences as accurately as possible is a crucial step in model building, as the alignment largely determines the 3D location of each of the residues of the query protein [29, 67]. Interestingly, a recent study by Gao and Stern [67] showed that the accuracy of TM models is significantly improved when sequence identity between the query protein and the template exceeds 30%, and is substantially reduced at weaker sequence identities. Exactly the same threshold values for correct modeling also apply in the case of soluble proteins [68]. We next address two cases: alignment with high and with low sequence similarities (Figure 3).

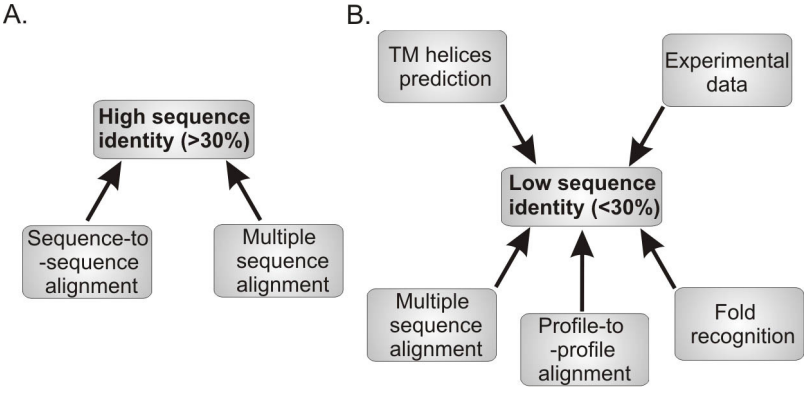

Figure 3. Computational approaches for the alignment of query and template proteins of high and low similarity. A. If the query and the template sequences are close enough (>30% identity) it is possible to use simple sequence-to-sequence alignment, but it is often more accurate to extract the pairwise alignment from an MSA. **B**. As sequence identity decreases, it becomes necessary to combine more sources of data in order to align the sequences correctly. These include fold recognition, profile-to-profile methods, TM prediction methods, MSAs, and available experimental data.

17.2.3.1 High similarity

A rather simple sequence-to-sequence alignment might suffice when the query and the template proteins exhibit sequence identity of more than 30%, covering all the TM segments of the sequence, whereas extraction of the pairwise alignment from an MSA might add essential evolutionary information and thus improve the alignment accuracy (Figure 3A) [29]. To ensure MSA integrity and avoid sequence fragments, it might be useful to include in the MSA only those sequences that share all of the query protein's TM helices. Forrest and co-workers, after examining the performance of different MSA algorithms in securing the pairwise alignments needed for TM-protein modeling [29], reported that advanced alignment methods, such as MUSCLE [69] or T-Coffee [70], are more effective than the traditional Clustal W [71].

Amino-acid substitution matrices are essential for generating both pairwise and multiple sequence alignments. The widely used substitution matrices, such as PAM [72] and BLOSUM [73], were derived from datasets of homologous soluble proteins. Additional substitution matrices, e.g. JTT TM [13], PHAT [74], and SLIM [75], were developed specifically for TM proteins. To the best of our knowledge, no study has yet compared all three matrices. However, when the abilities of both matrices in searching for homologues searches were compared using a dataset of GPCRs, SLIM was predicted to outperform PHAT [75]. When the same TM dataset was examined, both SLIM and PHAT were shown to be more accurate than the traditional BLOSUM. On the other hand, the PHAT matrix performed better than JTT TM [74]. In this case, the test set was composed of 100 sequences from 74 TM-protein families that were utilized as query proteins for database searching.

Theoretically, in bipartite alignments the membrane matrices should be utilized to align TM regions, while the extra-membrane regions should be aligned with the traditional matrices. Such an approach is implemented by STAM, which adds high gap penalties in predicted TM regions. However, the STAM method was evaluated only for GPCRs and was compared only to Clustal W [76]. Forrest et al., on examining the performance of bipartite alignments on a diverse dataset of TM proteins, found that performance was worse when they used a combination of the PHAT and BLOSUM matrices than BLOSUM alone [29]. Nevertheless, some improvements over the commonly used alignment methods were seen for the PRALINE^{TM} method, which also incorporates the PHAT matrix for TM alignment [77]. This was attributed to a more accurate prediction of the TM segments. A systematic evaluation of the substitution matrices and their performance in bipartite alignments has yet to be carried out.

17.2.3.2 Low similarity

When the query and the template proteins are distant homologues (sequence identity <30%), as is often the case when eukaryotic proteins are aligned to prokaryotic proteins (e.g. [61-63]), the straightforward approach described above might not suffice [29]. In such cases it is rather difficult to produce a fully continuous alignment. However, the TM helices are typically more conserved than the extra-membrane regions and it is often possible to align them properly. Indeed, such fragmented alignment can be used for model building of the TM domain. So the problem becomes a matter of detecting the TM helices of the query protein and their subsequent alignment to the known helices of the template. The TM helices are not only strongly

hydrophobic, but are also usually preserved within the protein family. Hence, in an MSA of the query protein and its homologues, these regions often appear as gap-less segments of a strongly hydrophobic nature; this observation simplifies their detection.

Fold recognition approaches, including profile-to-profile alignments, not only enable remote relationships between query and template proteins to be detected, but also produce a sensitive pairwise alignment. The HMAP method [78] produced more accurate alignments between query and template TM proteins than those produced via sequence-to-sequence and MSAs, especially in cases of low sequence similarity [29].

Overall, when attempting to properly align sequences of low identity, it is recommended to use a range of tools and all the available data (Figure 3B). These include MSAs, results of fold recognition approaches, TM-helix predictions, and the available biochemical and biophysical experimental data (for example, site-directed mutagenesis and accessibility measurements). In the optimal situation, data from all sources will overlap, thus consolidating the prediction. But even in less favorable situations, in which conflicting data might be obtained, often there is consensus at least with regard to the location of some of the TM helices. When data from various sources are in conflict concerning the location of a particular TM helix, and in the absence of a more compelling basis for resolution, decisions can be made based on the majority of (independent) data. Alternatively, 3D models can be built on the basis of more than one sequence alignment.

17.2.4 Building a 3D model-structure

The model-building process includes construction of the protein core based on an existing structural template, and modeling of the backbone and side chains of peripheral regions for which a template might not be available [57]. For α-helical TM proteins the core includes the TM helices, while the periphery contains the extra-membrane loops that tend to vary even between related TM proteins.

As in the case of soluble proteins, the building process is carried out via one of the many available applications, such as Modeller [79] and NEST [80]. The performance of model-building methods specifically for TM proteins was recently investigated in two studies. Reddy et al. [81] assessed the model-building performance of five methods (or combinations of methods): Modeller [58, 79], the MOE [82] homology module of InsightII [83], Swiss-PdbViewer [84, 85] and models produced via initial construction by InsightIIHomology followed by Modeller

refinement. Although this analysis did not include state-of-the-art programs such as NEST [80], PLOP [86-88], or Rosetta's template-based module [89], it was an initial attempt to evaluate the performance of some algorithms for building models of TM proteins. The results indicated that for this particular dataset of TM structures**,** Modeller generally outperformed the other methods.

Gao and Stern [67] compared Modeller [58, 79] to PLOP [86-88] while using a dataset of TM proteins that included both α-helix bundles and β-barrels. First they evaluated model-structures built via Modeller and PLOP on the basis of accurate pairwise alignments, constructed using structural alignments to eliminate alignment errors. PLOP outperformed Modeller in this case, probably as a result of its improved energy function. However, when those alignments were replaced with Clustal W-derived [71] pairwise alignments, which usually contain some alignment errors, Modeller and PLOP produced similar results, despite the fact that PLOP's energy function is considered to be superior to that of Modeller. The authors offered an explanation for this discrepancy by suggesting that the current sampling of conformational space by PLOP is not sufficient to detect the correct structural conformation. This suggestion was supported by their finding that in the refinement of the loop regions both methods performed poorly, again possibly owing to limited sampling of conformational space [67].

Since some traits of TM proteins differ from those of soluble proteins, future research should probably address specific adaptations of the abovementioned packages for TM proteins. These might include, for example, the use of rotamers constructed from TM structures, as well as novel scoring functions. To the best of our knowledge, only the template-based modeling application of Rosetta [89] has so far been modified to include a membrane-specific force field [42]. This modified method was utilized, for example, in a study of voltage-gated potassium channels, where models derived via Rosetta's homology*/de novo* membrane mode were used to provide a mechanism for voltage-dependent gating [90].

Although many energy (or scoring) functions do not currently encompass membrane-specific adaptations, Gou and Stern examined the ability of several high-quality energy functions to distinguish native loops in TM structures from decoys of loop conformations, generated via molecular dynamics. This analysis included, *inter alia,* the energy functions implanted in Modeller [58, 79], Rosetta [91, 92], PLOP [86-88] and the DFIRE potential [93]. The results indicated that some of the examined energy functions can reliably discriminate the native loop from the decoy conformations. Moreover, all but one energy function successfully ranked the energy of the entire TM structure lower than those of decoy models produced by homology

modeling. These findings raise hopes for the future implementation of energy functions in the refinement of TM models, possibly in parallel with the improvement of sampling in conformational space.

17.2.5 Useful tips for TM comparative modeling

- It is useful to evaluate the topology of the TM protein. This prediction might come in handy for template identification and/or query-template alignment.
- Identification of a common origin can indicate a shared fold between two TM proteins. When sequences diverge, however, this functional relationship is not always easy to detect.
- Both simple and advanced similarity searches can be helpful in the identification and alignment of possible templates, especially meta-servers that combine methods such as fold-recognition and membrane-topology predictions.
- When the similarity between query and template sequences is low, alignment accuracy might be improved by combining state-of-the-art bioinformatical tools with experimental data.
- It is sometimes useful to obtain and evaluate a number of models built using different alignments or templates until the best model or models are found.

17.3 Experimental data fitting

In contrast to the section on comparative modeling, here we describe a computational approach that does not rely on the existence of a high-resolution structure of a similar protein. Instead, other types of available data can be exploited as constraints in order to produce a model-structure [7, 36]. Essential data for this purpose might come from low-resolution structures (e.g. cryo-EM maps) and from mutagenesis studies. The former have been shown to produce more accurate models, and will be addressed here more thoroughly. It should be noted that only a few TM model-structures have been obtained by experimental data fitting (e.g. [94-98]). This is mainly because of a lack of the needed preliminary data, but might also result from a rather complicated modeling process, which requires manual intervention and specialized expertise. Nevertheless, the models obtained so far have raised considerable interest. As more experimental data emerge, especially cryo-EM maps of eukaryotic TM proteins, this approach is likely to become much more relevant and easier to use.

17.3.1 Starting from electron-density maps at intermediate resolution

Cryo-EM maps occasionally provide an intermediate-resolution image of the protein structure, with an in-plane resolution of 5Å–10Å but a much lower resolution along the membrane normal. Owing to the low resolution, such maps cannot reveal the exact features of TM-protein structures. In the case of α-helical TM proteins, they cannot even be used to assign the TM segments to the map helices, not to mention the coordinates of the amino acids of the TM helices in 3D space. Nevertheless, the maps provide important data concerning the number, tilt, and overall location of the TM helices in the structure.

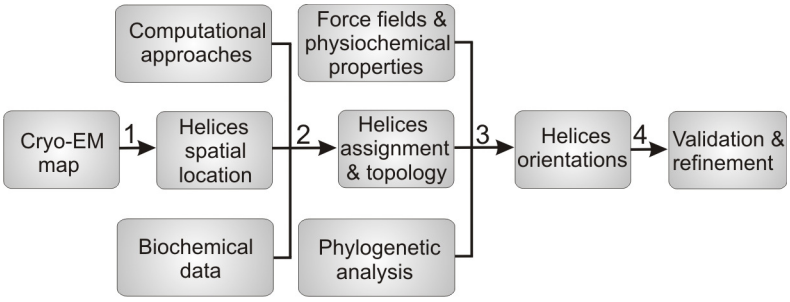

Figure 4. Predicting TM models from cryo-EM maps [7]. Step 1. Locations of TM helices, and specifically of their principal axes, are derived. Step 2. Using computational tools and biochemical data, the TM segments in the sequence are assigned to the helical rods in the density map. The topology of the TM protein in the map is also determined. Step 3. To correctly rotate each helix around its principal axis, additional data from phylogenetic analysis, physicochemical properties and/or force fields can be exploited. A Cα-trace model is generated. Side chains can be added to obtain a full-atom model. Step 4. Data that were not employed for model building can be used for validation. The model can then be refined by reviewing the preceding modeling steps. In addition, the model-structure can be used to design mutagenesis experiments and undergo subsequent refinement on the basis of the results.

For production of a molecular model from a cryo-EM map, additional data must be incorporated for the various modeling steps [94, 99, 100]. The overall modeling process is depicted by the flowchart in Figure 4 [7, 36]. First, spatial locations of the helices are obtained from an available intermediate-resolution cryo-EM map. Using the map, the principal axes of the helices can be detected and extracted. Next, TM segments in the protein sequence are assigned to the helical density rods, usually by employing both biochemical data and computational approaches. TM helices, corresponding to the TM-sequence segments, are then constructed using the principal axes. During this step, their register along the axes must be determined. Additional sources of data, such as evolutionary conservation and physicochemical properties of the protein sequence, are subsequently exploited to orient the helices around their principal axes. The result is a Cα-trace model-structure, i.e., the predicted location of the Cα atoms of the TM domain. The backbone atoms and the side chains of the residues can then be reconstructed in order to generate a full atom model of the TM protein. Finally, the model should be validated, typically via

experimental data that were not used to build the model (described in section 17.4). The results of the validation process can then be utilized for model refinement. Details of each step of this modeling process are addressed in the following sections.

17.3.1.1 Helix assignment and membrane topology

Before modeling is begun, the TM segments in the protein sequence must be identified and the topology of the protein in the membrane determined, as described above. These features are essential for TM-model building, but in most cases it is not easy to predict them with confidence, especially the precise helices boundaries. Thus, it is best to rely, as much as possible, on experimental data.

For n helices, the number of possibilities for assigning the TM segments detected in the protein's sequence to the helices in the map is $n!$. Adding the two possible membrane orientations (the cytoplasmic- and the extracellular-facing sides of the cryo-EM map) in relation to the protein's topology, the number of possible models is $2 \times n!$. This implies that the crucial step of helix assignment and selection of the membrane orientation is extremely complicated even for TM proteins of moderate size; a TM protein with 4 helices, for example, will have 48 combinations.

An attempt was made to develop a graph-theory approach for assigning TM helices and predicting topology based on the lengths of the loops connecting the helices [101]. The method worked well for short loops of up to 7 residues, but the accuracy of prediction depends strongly on exact determination of the boundaries of the TM helices, which is usually not available. Thus, there is no generally applicable way to determine the helix assignment and topology of a TM protein using a single automatic computational tool. The problem may occasionally be solved by combining manual analysis of biochemical data with use of the available computational tools. As in the modeling of the EmrE multidrug transporter [95], considerations from phylogenetic analysis, hydrophobicity, and experimental data might also be useful.

Generally speaking, the TM helices detected in the map can be divided into two groups: (a) core helices, surrounded by other TM helices in the bundle, and (b) peripheral helices, which are in contact with the core, but also have at least one lipid-exposed face. Owing to dissimilar evolutionary pressures, both the hydrophobicity and the evolutionary conservation patterns of the two types of helices are quite distinct. Relying on these differences of TM helices, Adamian and

Liang developed an automatic method to identify core TM helices, which are less accessible to the lipid membrane [102]. This method can help reduce the number of possibilities for helix assignment. Another useful observation for this modeling step is that interacting residues from neighboring TM helices might evolve cooperatively [96, 103]. Hence, detection of co-evolving positions by phylogenetic analysis (e.g. [103, 104]) can help guide the assignment of interacting TM-helix pairs. Experimental data such as distance constraints, site-directed mutagenesis, and accessibility assays can also be used.

The complexity of this step is best demonstrated by the example of the gap junction Cα-model, produced based on a cryo-EM map [96]. When a crystal structure (of a homologous protein) became available, it became clear that the helix assignment that was utilized for model-building was incorrect; only one of the four TM helices of each subunit in the homo-hexamer was assigned correctly [105]. The erroneous assignment was based on mutagenesis data that apparently was interpreted wrongly [106].

17.3.1.2 Helix building and rotation

An intermediate-resolution map for producing a model-structure of a TM protein was first employed by Baldwin and co-workers, who constructed a Cα-trace model of vertebrate rhodopsin [97]. The model was generated from a structure at 7Å resolution in the membrane plane using constraints derived from MSA and biochemical data. When the structure was later determined at high resolution by X-ray crystallography, the orientations of TM helices in the model-structure were found to be quite accurate (3.2Å RMSD). Most of the variation was attributed to difficulties encountered in the precise modeling of two kinked helices [7].

Expanding on this pioneering approach, Fleishman and co-workers developed an automatic method for TM-model prediction based essentially on evolutionary conservation analysis [100]. For predicting the orientation of each helix, the algorithm included a scoring function that favored the burial of conserved (and charged) residues in the protein core, as well as the exposure of variable amino acids to the lipid membrane. This method, in which only the Cα atoms of the TM domain were constructed, was later applied to predict the structure of the gap-junction [96] and the EmrE multidrug transporter [95]. The crystal structure of EmrE, determined a few months later, was markedly similar to the model structure, with an RMSD value of 1.4Å for the core region [107] (and of 3.52 Å for the entire TM domain, Figure 5).

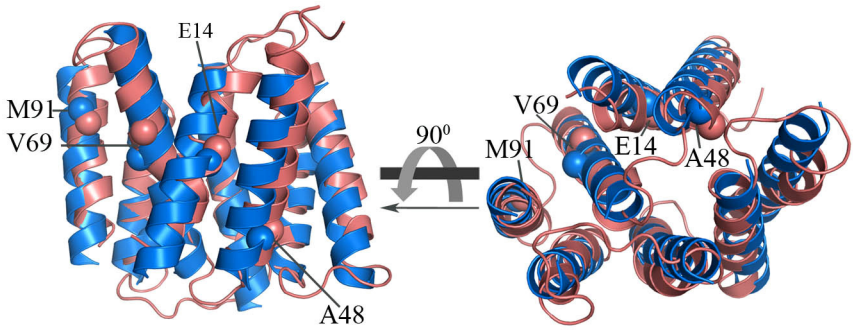

Figure 5. Model vs. crystal structure of EmrE. A model of the EmrE homodimer (blue) was derived using a cryo-EM map and the computational approach of Fleishman and co-workers [95]. The crystal structure of EmrE (red) was solved later [107]. The model and structure are aligned and viewed from the side (left) and top (right). The 3D location of specific Cα atoms (marked as spheres on one monomer of the model and structure) demonstrate the similarity between the model and native structure.

Beuming and Weinstein proposed a similar method [94], in which helix orientations are selected by employing both evolutionary conservation and a knowledge-based scale of the propensities of the 20 amino acids to be exposed to lipids. In addition to the Cα atoms, the backbones and side chains of residues are also constructed, and this is followed by structure minimization and some manual adjustment. The result is a full-atom model. This method was used to predict a molecular model of the bacterial oxalate transporter OxlT, using its electron density map of 6.5Å resolution in the membrane plane. The model was in agreement with cross-linking experiments and data concerning functional residues [94].

The most recent work in this field was done by Kovacs et al., who presented a new method [99] in which helix orientations are determined by minimizing an energy function that takes into account van der Waals interactions, electrostatics, hydrogen bonding, and torsional and density correlation terms. Side chains are also predicted. The best conformations are then energy-minimized by a complex procedure in which atoms of the helical backbone are restrained to fit the observed cryo-EM map densities. This minimization step also relies on a solvent-accessibility grid map of the density rods. It should be noted that in constructing this accessibility map the membrane boundaries must be selected within the cryo-EM map. Correct prediction of these boundaries is not a simple task, and deviations from the real (unknown) boundaries can affect the model-structure. Another limitation of the approach is that because of the complicated energy calculations required for an all-atom representation, it is feasible only for TM proteins of moderate size (up to 4 helices per symmetric subunit). The approach worked well in three test-cases, but has not yet been used for *de novo* predictions.

Overall, several different methods are available for modeling the TM domains of α-helical proteins employing restraints derived from electron-density maps of sufficient resolution. So far the starting point has always been a cryo-EM map, but maps from X-ray crystallography experiments at intermediate resolution can be used as well. Each of the above methods was developed and tested on only a small number of cases. Until the performance of all methods is examined by means of a large-scale assessment of known structures, their efficacies cannot be determined. In particular, it would be interesting to know whether the addition of side chains increases or decreases the accuracy. It may well be that each method is suitable only for certain specific cases. When constructing new structural models, therefore, it might be advisable to examine several potential models each produced by various different methods. The models can then be inspected on the basis of, for example, reliable experimental data along with other evaluation procedures (described in section 17.4 below).

17.3.2 Modeling based on biochemical and biophysical data

Other computational approaches aimed at addressing cases for which there were no available structural data. These approaches have employed biochemical and biophysical data, obtained for example from site-directed mutagenesis and chemical cross-linking, as the only constraints on the protein structure. Because these data are difficult to interpret in an unequivocal manner, this approach is inherently less reliable than modeling based on intermediate resolution structure from cryo-EM maps, as described above. In particular, the results of mutagenesis often represent phenomena that are associated with more than one conformation of the TM protein, and the observed phenotypes of a mutation might be indicative of allosteric effects.

Sale and co-workers [108] developed an automatic method for TM-structure prediction based on distance restraints obtained from experimental assays such as chemical cross-linking, nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR) or fluorescence resonance energy transfer (FRET). A search of the conformational space for a TM model-structure that is compatible with the available distance restraints is followed by optimization using Monte Carlo simulated annealing. The optimization samples models that correspond well both with the experimental restraints and with knowledge-based structural parameters derived from a dataset of known TM structures. Although this approach produced an accurate model-structure of the 7 TM helices of bovine rhodoposin (RMSD of 3.2Å) [108], it has yet to be applied for the prediction of novel TM-protein structures.

The bacterial lactose permease (also referred to as lacY), a galactosidase transporter, is arguably the most extensively studied TM protein to date. This transporter was examined by means of various experimental approaches, including systematic site-directed mutagenesis, double-cysteine mutants, thiol cross-linking, engineered Mn(II) binding sites, *N*-ethylmaleimide (NEM) alkylation of single cysteine mutants, site-directed EPR, and discontinuous mAb epitope mapping [109, 110]. The accumulated experimental data related to each position in the 12 TM helices of lacY, comprising 417 residues. By utilizing helical backbone restraints and 99 long-range restraints derived from thiol cross-linking and engineered Mn(II) binding sites, Sorgen et al. obtained a single cluster of models for lacY with small deviations from one another [98]. They achieved this using an algorithm based on torsion-angle dynamics-simulated annealing, which was initially developed and utilized in NMR structure determination [111].

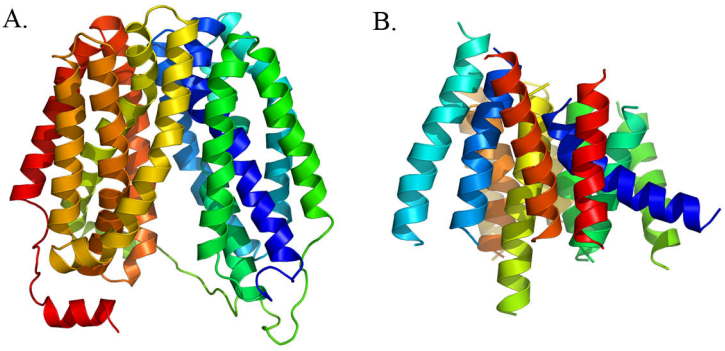

Figure 6. Comparison of the lacY model, produced via experimental constraints, and the solved crystal structure. In both panels the cytoplasmic side points downward. The lacY crystal structure (panel A) [112] and computational model (panel B) [98] are colored by rainbow. Although the overall fold and helix organization are quite distinct, there are regions of similarity, especially between the helices that contribute to the cytoplasmic-facing pore.

The crystal structure of lacY was later determined [112], making it possible to evaluate the effectiveness of this modeling approach. Comparison between the model and the native structure showed that various local arrangements of functional residues, such as sugar- and proton- binding residues, were fairly accurate. However, the overall architecture of the model and its structure were not superimposable (Figure 6) [7]. Given the crystal structure, it was possible to examine the experimentally measured distances that were used for modeling. While the distances on the periplasmic side of the crystal structure were in good agreement with the experimental data, many of the distances on the cytoplasmic side were underestimated. The distances obtained for the cytoplasmic side probably corresponded to the periplasmic-facing conformation or other conformational sub-states, and therefore did not agree with the crystal structure, which was solved in an inward-facing conformation [113]. The fact that the experimental data probably reflect different conformations might account for the discrepancies between the model and the structure. Overall, this case study demonstrated the difficulty of producing an accurate model

when the experimental data do not account for a single structural conformation. Obviously such "monochromatic" data are usually not available.

Prediction of the dimeric structure of the *E. coli* Na^+/H^+ antiporter NhaA is another example of constraint-based modeling of a TM protein. Although the protein in its physiological form is a dimer [114, 115], its crystal structure depicts only the monomer; the physiological dimeric contacts are not exhibited [33]. To obtain the dimeric structure, two NhaA monomeric structures were considered as rigid bodies. Nine long-range EPR distance measurements were then used as constraints to build a dimer by docking the two monomers [116]. The model-dimer showed good agreement with the interfacial domain observed in cryo-EM 2D crystals, which exhibited the electron density of the NhaA dimer. The suggested dimeric interface was further supported by chemical cross-linking [115] and deletion assays [117]. Although this is not a classical example of the use of experimental constraints to model a helical TM protein, it shows how the membrane plane and intrinsic symmetry reduces the degrees of freedom of the modeling process. Thus, even a small number of distance constraints was sufficient for inferring the oligomeric conformation.

17.3.3 Tips for modeling by experimental data fitting

- A cryo-EM map is a good starting point. When the map's resolution is high enough to detect the TM helices, at least Cα-trace models of TM proteins can be produced.
- Because helix assignment is a crucial and complicated step, several data sources are usually needed in order to correctly assign the TM sequence segments to the helix density contours in the map.
- Evolutionary conservation, physicochemical features and force fields are useful for rotating the TM helices around their principal axes.
- The use of empirical data to build models of TM proteins is complicated because: (a) the protein often undergoes conformational changes, and data relating to the effects of mutations might reflect a mixture of these conformations. (b) Mutagenesis data might reflect both direct interactions and remote (allosteric) effects.

17.4 Quality assessment

Computational methods of model evaluation are usually referred to as Model Quality Assessment Programs (MQAPs) [38, 118]. As reviewed in chapters 15 and 16, numerous computational

methods have already been developed for local and global assessments of model-structures, indicating the importance of this step in protein modeling (e.g., [119-124]). These are all general methods for model evaluation; to the best of our knowledge, specialized MQAPs for TM proteins are currently not available. Therefore, when these general evaluation tools are applied to TM model-structures, the results of the assessment should be viewed with caution. A good strategy for assessment of TM models predicted via homology modeling might be to apply the MQAPs to both model and template. The results of the template could be used thereafter as a reference point for the result of the model-structure. It is anticipated that in the future, existing approaches will be modified to better comply with the distinct traits of TM proteins. Alternatively, assessment of the performance of state-of-the-art MQAPs on TM structures might reveal that current methods are also adequate for evaluating this class of proteins.

The validity of the TM model-structure can be further assessed through an examination of its generic characteristics. These might comprise only external features, defined here as protein characteristics that were not accounted for during model building. A recent investigation of the local accuracy of TM models that were produced via homology modeling demonstrated that even when the membrane-embedded helices of the query and the template sequences are structurally similar, the extra-membrane regions that connect them might deviate in both sequence and length [29]. In another study it was demonstrated that refinement procedures cannot clearly improve the loop regions in TM proteins [67]. Thus, loops in TM model-structures should be considered *a priori* as regions of questionable accuracy, as in the modeling of soluble proteins.

17.4.1 Compatibility of the model with general characteristics of TM proteins

In section 17.1.1 we presented some of the general features that characterize α-helical TM proteins. These distinct traits were observed by analysis of TM structures that were solved experimentally. Those traits could therefore be utilized to assess the accuracy of TM model-structures, provided that they were not used in building the model. When model-structures are assessed in the future, it might be helpful to exploit the recent discovery that the structural determinants of TM helices appear to incorporate five specific types of interhelical interactions [31]. Accordingly, the expectation would be that trustworthy models will feature these interactions, and that wrong models will not. When erroneous X-ray crystal structures were retracted from the PDB, they were indeed found to include only very few interactions of that sort, which would not suffice to keep the fold intact.

17.4.1.1 The "positive-inside" rule and the "aromatic-belt"

The "positive-inside" rule [21] can be used to examine the overall architecture of a TM model-structure, as the distribution of lysines and arginines in the extra-membrane regions of the model-structure can serve as an indication of whether the TM segments and extra-membrane regions are correctly approximated. Moreover, this rule can point out cases where the template selection or the query-template alignments are entirely erroneous. Clearly however, this will not be of help in determining the exact TM boundaries, inter-helical structural arrangement, or packing. In addition, TM-protein structures frequently feature an "aromatic belt" near the borders of the hydrocarbon core region [17]. As with the "positive-inside" rule, this feature can also be evaluated to assess the overall topology of the TM model-structure, but will not help to validate its precise molecular details.

It should be mentioned that both of the above features are implemented in many of the advanced methods for predicting the membrane topology from the sequence. If such methods were used for building the model, it is fairly obvious that the model will inevitably be compatible with both thumb-rules.

17.4.1.2 Hydrophobicity of lipid-facing residues

Based on knowledge derived from available TM structures, most of the lipid-exposed residues of the TM model-structure are expected to be hydrophobic [17]. In some of the methods for predicting TM structure by the use of experimental constraints, this trait is exploited to produce the 3D model (e.g. [94, 100]). In comparative modeling, this trait can be addressed indirectly when integrating the results of TM-helix prediction for correct alignment of the query-template sequences.

Such examination is likely to be useful for validation provided that the nature of the lipid-exposed residues was not taken into consideration during model building. To ensure that this requirement is met, the lipid-exposed positions in the TM model-structure should be reviewed using a hydrophobicity scale (e.g. [20]). Mapping of the scale on the residues of the TM model-structure reveals its degree of correspondence with the structure's expected hydrophobicity pattern. It should be noted that this evaluation process is useful for peripheral helices, in which residues are exposed to the membrane, but not for helices that are buried in the TM core. If, on

examining a model generated by comparative modeling, a peripheral helix is found in which polar residues face the membrane while hydrophobic residues face the protein core, this might indicate that the pairwise alignment of the query-template in this region needs to be refined. During this evaluation process, moreover, the physiological oligomeric state should also be taken into account. Regions that appear to be lipid-exposed might actually participate in inter-protein interfaces within the oligomeric structure of the TM protein. These might include polar residues.

17.4.1.3 Prolines and kinks

Cordes and co-workers have demonstrated that proline residues are more abundant at the ends of TM helices [22]. Proline residues interrupt helical segments and are also commonly found in irregular regions of TM helices [10, 22, 23]. Furthermore, inspection of hinge regions of TM proteins revealed that 60% of the prolines comprise the hinge itself or are located up to 4 residues (approximately one turn) before it [22]. Another study showed that positions in which the MSA exhibits a high content of prolines are likely to correspond to proline-induced kinked or otherwise disrupted regions [27]. Taking all of the above into consideration, it is interesting to examine the predicted location of proline residues and the positions in which the MSA shows an abundance of proline. Many of these positions, especially the conserved ones, would be expected to cluster at the ends of the helix or in the vicinity of helix irregularities in the model-structure. This can provide a rough validation for the TM model, especially with respect to the assignment of irregular TM helices.

17.4.2 Evolutionary conservation profile

Proteins are usually subjected to evolutionary pressure in areas of structural or functional importance. A number of studies have shown that α-helical TM proteins exhibit a distinct conservation pattern in which the protein core is conserved while the loops and lipid-exposed residues are rather variable [12, 36, 94, 100, 102, 125-127]. Thus, peripheral helices frequently present distinct variable and conserved helical faces, with the variable faces exposed to the membrane. Because the core region contributes to structural stability and function, it is typically under stronger evolutionary pressure, and would accordingly generally exhibit a high level of conservation. This evolutionary conservation pattern has been demonstrated for various membrane proteins (for example, bacteriorhodopsin [7] and the sodium/proton transporter NhaA of *E. coli* [62, 128]). By contrast, mapping of evolutionary conservation analysis on erroneous

TM structures, such as two of the structures of the EmrE transporter [129, 130], does not fit this paradigmatic pattern [36]. The empirical principle can also be demonstrated by a comparison of the conservation pattern of another retracted structure, the crystal structure of the ATP-binding cassette multidrug transporter MsbA [131, 132], to that of the correct structure of homologous sav1866 [133]. The conserved residues of sav1866 are evidently buried in the protein core while variable residues face the lipids, as anticipated (Figure 7B). However, the evolutionary profile of the retracted structure of MSbA shows a different pattern: some conserved residues face the lipids and some variable residues are buried in the core (Figure 7A). In both proteins, however, the cytoplasmic ends are highly conserved and form contacts with the cytoplasmic domains.

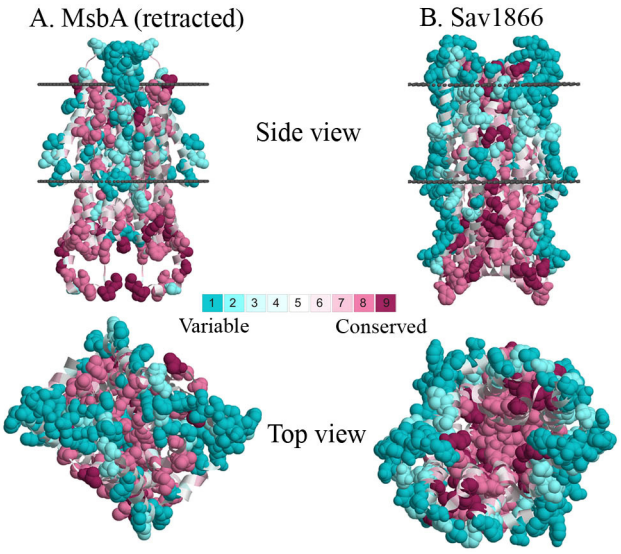

Figure 7. Conservation analysis of erroneous and correct structures of ABC transporters. The retracted structure of MsbA [131] (panel A) and the structure of sav1866 [133] (panel B) are colored according to conservation, using the ConSurf color scale [134]. Highly conserved residues, receiving grades of 8 or 9, along with the outermost variable (grades of 1 or 2) are shown as spheres. The two upper panels show a side view of the two proteins with their cytoplasmic sides facing down. Approximated membrane boundaries are shown in grey. The nucleotide-binding cytoplasmic domains of both MsbA and sav1866 were omitted for clarity. The two lower panels show a top (and closer) view of the same proteins.

Incorporating this notion, mapping of evolutionary conservation analysis on the model is highly effective in assessing putative structural models. The examination is applicable only if conservation was not taken into account during generation of the model. The approach was recently utilized, for example, to validate models produced for the SERT transporter [60] and the NHE1 Na^+/H^+ exchanger [62]. In both studies, conservation scores were calculated via the ConSurf webserver (http://consurf.tau.ac.il [134]). This evaluation procedure can still only be performed manually. Future developments are likely to include its automatization, assigning a score that can then be utilized to compare the quality of different models.

In the case of comparative modeling, examination of the evolutionary conservation analysis mapped on the model-structure can indicate if the pairwise alignment or the template selection procedures should be revised. For example, a common error such as a single shift in the alignment of a TM helix might result in the placement of conserved residues towards the lipid

while variable residues are buried. Such inaccuracy is easily visible from the conservation analysis, but might be difficult to decipher using other evaluation tools. Overall, mapping of the evolutionary conservation analysis on TM model-structures can be considered a highly effective method of evaluation. Close examination of this analysis can allow large or small errors to be detected in the model-structure. This will help not only in the assessment of the TM model's local and global quality, but also in the refinement of problematic regions.

It is noteworthy that water-soluble proteins also exhibit similar evolutionary profiles. That is, their interior is more conserved than their exterior. Indeed, this property has been used to evaluate the quality of structural models [135-137].

17.4.3 Correspondence with experimental and clinical data

As already mentioned, some types of experimental data provide constraints that are useful in predicting TM structures, and if not used for model building, these data can be utilized for validation. Generally speaking, residues in which mutations disrupt a protein's function would normally be found in the TM-protein core, and would typically be of structural or functional importance. They might, for example, contribute to stabilization through inter-helical interfaces or to a function such as direct binding of a substrate. By contrast, most of the positions that are less sensitive to mutations are typically exposed to the lipid, owing to the fact that membrane-facing positions are in general not directly involved in structural stabilization or in function. This paradigm was well exemplified by mapping of elaborate mutagenesis data on the model-structure of the Na^+/H^+ exchanger NHE1 [62].

The above general logic can be applied on examination of the model-predicted locations of polymorphisms and disease-causing mutations. The former are predicted to reside on peripheral regions of the TM protein, whereas the latter typically comprise the core. Besides site-directed mutagenesis and clinical data, other types of empirical data are also helpful in validating the structure of TM proteins. These include, for example, accessibility assays (employed, for instance, to evaluate the EmrE model-structure [95]) and distance assessments using chemical cross-linking (used, for example, in assessing the model of CFTR [63]) or other measurements.

Nevertheless, it is worth re-emphasizing that the experimental data should be treated with caution, especially with regard to intrinsic conformational changes. This was well illustrated in a recent study by Forrest and co-workers of the *Aquifex aeolicus* leucine transporter (aaLeuT),

whose structure had been previously solved in its extra-cellular-facing conformation [138]. By exploiting the pseudo-symmetry observed in the crystal structure, they produced a model of the cytoplasmic-facing conformation of the aaLeuT transporter [139]. To assess their cytoplasmic-facing model-structure they used two inhibitors of the homologous SERT transporter, each of which stabilizes a distinct structural conformation (inward or outward). Accessibility measurements obtained for the inhibitor-stabilized inward state of the SERT transporter provided experimental support for the cytoplasmic-facing model of the aaLeuT transporter. These results demonstrated that when the available data correspond to a single conformational state of a TM protein, it is possible to obtain accurate validation of its model-structure.

17.4.4 Tips for evaluation

- MQAPs should be used with caution as their performance on TM proteins has yet to be examined.
- It is helpful to inspect the predicted location of specific amino-acid types that exhibit special traits in TM structures.
- Membrane-exposed residues usually exhibit marked hydrophobicity. Thus, the presence of too many polar residues in lipid-facing regions, especially if the residues are charged, might be indicative of an inadequate model-structure.
- Evolutionary conservation analysis is a useful tool for TM-model assessment and refinement. It is important to bear in mind that such analysis is profoundly affected by the quality of the input MSA.
- TM-model validation via experimental data is extremely helpful. Data that are reliable and easy to interpret offer the best available external assessment of TM model-structures.

Acknowledgments

This work was supported by grant 611/07 from the Israel Science Foundation to N.B-T. M.S. was supported by the Edmond J. Safra Bioinformatics program at Tel-Aviv University.

References

1. Liu, J. and B. Rost, *Comparing function and structure between entire proteomes.* Protein Sci, 2001. **10**(10): p. 1970-9.
2. Kahsay, R.Y., G. Gao, and L. Liao, *An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes.* Bioinformatics, 2005. **21**(9): p. 1853-8.

3. Mitaku, S., et al., *Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system.* Biophys Chem, 1999. **82**(2-3): p. 165-71.
4. Lundstrom, K., *Structural genomics and drug discovery.* J Cell Mol Med, 2007. **11**(2): p. 224-38.
5. White, S.H., *The progress of membrane protein structure determination.* Protein Sci, 2004. **13**(7): p. 1948-9.
6. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
7. Fleishman, S.J., V.M. Unger, and N. Ben-Tal, *Transmembrane protein structures without X-rays.* Trends Biochem Sci, 2006. **31**(2): p. 106-13.
8. Hurwitz, N., M. Pellegrini-Calace, and D.T. Jones, *Towards genome-scale structure prediction for transmembrane proteins.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1467): p. 465-75.
9. Punta, M., et al., *Membrane protein prediction methods.* Methods, 2007. **41**(4): p. 460-74.
10. Bowie, J.U., *Solving the membrane protein folding problem.* Nature, 2005. **438**(7068): p. 581-9.
11. Liu, Y., D.M. Engelman, and M. Gerstein, *Genomic analysis of membrane protein families: abundance and conserved motifs.* Genome Biol, 2002. **3**(10): p. research0054.0051-00.54.0012.
12. Donnelly, D., et al., *Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues.* Protein Sci, 1993. **2**(1): p. 55-70.
13. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A mutation data matrix for transmembrane proteins.* FEBS Lett, 1994. **339**(3): p. 269-75.
14. Blondelle, S.E., et al., *Secondary structure induction in aqueous vs membrane-like environments.* Biopolymers, 1997. **42**(4): p. 489-98.
15. Wallin, E., et al., *Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria.* Protein Sci, 1997. **6**(4): p. 808-15.
16. Ubarretxena-Belandia, I. and D.M. Engelman, *Helical membrane proteins: diversity of functions in the context of simple architecture.* Curr Opin Struct Biol, 2001. **11**(3): p. 370-6.
17. Ulmschneider, M.B., M.S. Sansom, and A. Di Nola, *Properties of integral membrane protein structures: derivation of an implicit membrane potential.* Proteins, 2005. **59**(2): p. 252-65.
18. Tourasse, N.J. and W.H. Li, *Selective constraints, amino acid composition, and the rate of protein evolution.* Mol Biol Evol, 2000. **17**(4): p. 656-64.
19. von Heijne, G., *Membrane-protein topology.* Nat Rev Mol Cell Biol, 2006. **7**(12): p. 909-18.
20. Kessel, A. and N. Ben-Tal, *Free energy determinants of peptide association with lipid bilayers.* Current Topics in Membranes, 2002. **52**: p. 205-253.
21. Heijne, G.V., *The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology.* EMBO J, 1986. **5**(11): p. 3021-27.
22. Cordes, F.S., J.N. Bright, and M.S. Sansom, *Proline-induced distortions of transmembrane helices.* J Mol Biol, 2002. **323**(5): p. 951-60.
23. Barlow, D.J. and J.M. Thornton, *Helix geometry in proteins.* J Mol Biol, 1988. **201**(3): p. 601-19.
24. Tieleman, D.P., et al., *Proline-induced hinges in transmembrane helices: possible roles in ion channel gating.* Proteins, 2001. **44**(2): p. 63-72.
25. Lu, H., T. Marti, and P.J. Booth, *Proline residues in transmembrane alpha helices affect the folding of bacteriorhodopsin.* J Mol Biol, 2001. **308**(2): p. 437-46.
26. Brandl, C.J. and C.M. Deber, *Hypothesis about the function of membrane-buried proline residues in transport proteins.* Proc Natl Acad Sci U S A, 1986. **83**(4): p. 917-21.
27. Yohannan, S., et al., *The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors.* Proc Natl Acad Sci U S A, 2004. **101**(4): p. 959-63.
28. Wallace, B.A., M. Cascio, and D.L. Mielke, *Evaluation of methods for the prediction of membrane protein secondary structures.* Proc Natl Acad Sci U S A, 1986. **83**(24): p. 9423-7.
29. Forrest, L.R., C.L. Tang, and B. Honig, *On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins.* Biophys J, 2006. **91**(2): p. 508-17.
30. Li, S.C. and C.M. Deber, *A measure of helical propensity for amino acids in membrane environments.* Nat Struct Biol, 1994. **1**(6): p. 368-73.
31. Harrington, S.E. and N. Ben-Tal, *Structural determinants of transmembrane helical proteins.* Structure, 2009. **17**(8): p. 1092-1103.
32. Grigorieff, N., et al., *Electron-crystallographic refinement of the structure of bacteriorhodopsin.* J Mol Biol, 1996. **259**(3): p. 393-421.

33. Hunte, C., et al., *Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH.* Nature, 2005. **435**(7046): p. 1197-202.
34. Fu, D., et al., *Structure of a glycerol-conducting channel and the basis for its selectivity.* Science, 2000. **290**(5491): p. 481-6.
35. Granseth, E., et al., *Membrane protein structural biology--how far can the bugs take us?* Mol Membr Biol, 2007. **24**(5-6): p. 329-32.
36. Fleishman, S.J. and N. Ben-Tal, *Progress in structure prediction of alpha-helical membrane proteins.* Curr Opin Struct Biol, 2006. **16**(4): p. 496-504.
37. Elofsson, A. and G. von Heijne, *Membrane protein structure: prediction versus reality.* Annu Rev Biochem, 2007. **76**: p. 125-40.
38. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
39. Das, R. and D. Baker, *Macromolecular Modeling with Rosetta.* Annual Review of Biochemistry, 2008. **77**(1): p. 363-382.
40. Bradley, P., K.M. Misura, and D. Baker, *Toward high-resolution de novo structure prediction for small proteins.* Science, 2005. **309**(5742): p. 1868-71.
41. Yarov-Yarovoy, V., J. Schonbrun, and D. Baker, *Multipass membrane protein structure prediction using Rosetta.* Proteins, 2006. **62**(4): p. 1010-25.
42. Barth, P., J. Schonbrun, and D. Baker, *Toward high-resolution prediction and design of transmembrane helical protein structures.* Proceedings of the National Academy of Sciences, 2007. **104**(40): p. 15682-15687.
43. Barth, P., B. Wallner, and D. Baker, *Prediction of membrane protein structures with complex topologies using limited constraints.* Proceedings of the National Academy of Sciences, 2009. **106**(5): p. 1409-1414.
44. Zhang, Y., M.E. Devries, and J. Skolnick, *Structure modeling of all identified G protein-coupled receptors in the human genome.* PLoS Comput Biol, 2006. **2**(2): p. e13.
45. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-65.
46. Takeda, S., et al., *Identification of G protein-coupled receptor genes from the human genome sequence.* FEBS Lett, 2002. **520**(1-3): p. 97-101.
47. Pierce, K.L., R.T. Premont, and R.J. Lefkowitz, *Seven-transmembrane receptors.* Nat Rev Mol Cell Biol, 2002. **3**(9): p. 639-50.
48. Lundstrom, K., *Latest development in drug discovery on G protein-coupled receptors.* Curr Protein Pept Sci, 2006. **7**(5): p. 465-70.
49. Patny, A., P.V. Desai, and M.A. Avery, *Homology modeling of G-protein-coupled receptors and implications in drug design.* Curr Med Chem, 2006. **13**(14): p. 1667-91.
50. Shacham, S., et al., *PREDICT modeling and in-silico screening for G-protein coupled receptors.* Proteins, 2004. **57**(1): p. 51-86.
51. Trabanino, R.J., et al., *First principles predictions of the structure and function of g-protein-coupled receptors: validation for bovine rhodopsin.* Biophys J, 2004. **86**(4): p. 1904-21.
52. Fanelli, F. and P.G. De Benedetti, *Computational modeling approaches to structure-function analysis of G protein-coupled receptors.* Chem Rev, 2005. **105**(9): p. 3297-351.
53. Oliveira, L., et al., *Heavier-than-air flying machines are impossible.* FEBS Lett, 2004. **564**(3): p. 269-73.
54. Becker, O.M., et al., *Modeling the 3D structure of GPCRs: advances and application to drug discovery.* Curr Opin Drug Discov Devel, 2003. **6**(3): p. 353-61.
55. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Curr Opin Struct Biol, 2005. **15**(3): p. 285-9.
56. Ginalski, K., *Comparative modeling for protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(2): p. 172-7.
57. Petrey, D. and B. Honig, *Protein structure prediction: inroads to biology.* Mol Cell, 2005. **20**(6): p. 811-9.
58. Fiser, A. and A. Sali, *Modeller: generation and refinement of homology-based protein structure models.* Methods Enzymol, 2003. **374**: p. 461-91.

59. Oberai, A., et al., *A limited universe of membrane protein families and folds.* Protein Sci, 2006. **15**(7): p. 1723-34.
60. Forrest, L.R., et al., *Identification of a chloride ion binding site in Na+/Cl -dependent transporters.* Proc Natl Acad Sci U S A, 2007. **104**(31): p. 12761-6.
61. Beuming, T., et al., *A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na+ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function.* Mol Pharmacol, 2006. **70**(5): p. 1630-42.
62. Landau, M., et al., *Model structure of the Na+/H+ exchanger 1 (NHE1): functional and clinical implications.* J Biol Chem, 2007. **282**(52): p. 37854-63.
63. Mornon, J.P., P. Lehn, and I. Callebaut, *Atomic model of human cystic fibrosis transmembrane conductance regulator: membrane-spanning domains and coupling interfaces.* Cell Mol Life Sci, 2008. **65**(16): p. 2594-612.
64. Callebaut, I., et al., *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives.* Cell Mol Life Sci, 1997. **53**(8): p. 621-45.
65. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
66. Hedman, M., et al., *Improved detection of homologous membrane proteins by inclusion of information from topology predictions.* Protein Sci, 2002. **11**(3): p. 652-8.
67. Gao, C. and H.A. Stern, *Scoring function accuracy for membrane protein structure prediction.* Proteins, 2007. **68**(1): p. 67-75.
68. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
69. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
70. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
71. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
72. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary change in proteins.* Atlas of Protein Sequence and Structure., 1978. **5**: p. 345–352.
73. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
74. Ng, P.C., J.G. Henikoff, and S. Henikoff, *PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane.* Bioinformatics, 2000. **16**(9): p. 760-6.
75. Muller, T., S. Rahmann, and M. Rehmsmeier, *Non-symmetric score matrices and the detection of homologous transmembrane proteins.* Bioinformatics, 2001. **17**(Suppl 1): p. S182-9.
76. Shafrir, Y. and H.R. Guy, *STAM: simple transmembrane alignment method.* Bioinformatics, 2004. **20**(5): p. 758-69.
77. Pirovano, W., K.A. Feenstra, and J. Heringa, *PRALINETM: a strategy for improved multiple alignment of transmembrane proteins.* Bioinformatics, 2008. **24**(4): p. 492-7.
78. Tang, C.L., et al., *On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles.* J Mol Biol, 2003. **334**(5): p. 1043-62.
79. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
80. Petrey, D., et al., *Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling.* Proteins, 2003. **53**(Suppl 6): p. 430-5.
81. Reddy, C.S., et al., *Homology modeling of membrane proteins: a critical assessment.* Comput Biol Chem, 2006. **30**(2): p. 120-6.
82. Kelly, K., *3D bioinformatics and comparative protein modeling in MOE.* J Chem Comp Group, 1999. **autumn ed**.
83. Dayringer, H.E., A. Tramontano, and R.J. Fletterick, *Interactive program for visualization and modelling of proteins, nucleic acids and small molecules.* J Mol Graph, 1986(4): p. 82-87.

84. Schwede, T., et al., *SWISS-MODEL: An automated protein homology-modeling server.* Nucleic Acids Res, 2003. **31**(13): p. 3381-5.
85. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-23.
86. Jacobson, M. and S. A., *Comparative protein structure modeling and its applications to drug discovery.* Annual Reports in Medicinal Chemistry, 2004. **39**: p. pp 259–67.
87. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins, 2004. **55**(2): p. 351-67.
88. Jacobson, M.P., et al., *Force Field Validation Using Protein Side Chain Prediction , Force Field Validation Using Protein Side Chain Prediction.* Journal of Physical and Colloid Chemistry, 2002. **106**(44,): p. 11673-80.
89. Rohl, C.A., et al., *Modeling structurally variable regions in homologous proteins with rosetta.* Proteins, 2004. **55**(3): p. 656-77.
90. Yarov-Yarovoy, V., D. Baker, and W.A. Catterall, *Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels.* Proc Natl Acad Sci U S A, 2006. **103**(19): p. 7292-7.
91. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.
92. Simons, K.T., et al., *Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.* Proteins, 1999. **34**(1): p. 82-95.
93. Zhou, H. and Y. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein Sci, 2002. **11**(11): p. 2714-26.
94. Beuming, T. and H. Weinstein, *Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT.* Protein Eng Des Sel, 2005. **18**(3): p. 119-25.
95. Fleishman, S.J., et al., *Quasi-symmetry in the cryo-EM structure of EmrE provides the key to modeling its transmembrane domain.* J Mol Biol, 2006. **364**(1): p. 54-67.
96. Fleishman, S.J., et al., *A Calpha model for the transmembrane alpha helices of gap junction intercellular channels.* Mol Cell, 2004. **15**(6): p. 879-88.
97. Baldwin, J.M., G.F. Schertler, and V.M. Unger, *An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors.* J Mol Biol, 1997. **272**(1): p. 144-64.
98. Sorgen, P.L., et al., *An approach to membrane protein structure without crystals.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14037-40.
99. Kovacs, J.A., M. Yeager, and R. Abagyan, *Computational prediction of atomic structures of helical membrane proteins aided by EM maps.* Biophys J, 2007. **93**(6): p. 1950-9.
100. Fleishman, S.J., et al., *An automatic method for predicting transmembrane protein structures using cryo-EM and evolutionary data.* Biophys J, 2004. **87**(5): p. 3448-59.
101. Enosh, A., et al., *Assigning transmembrane segments to helices in intermediate-resolution structures.* Bioinformatics, 2004. **20**(Suppl 1): p. i122-9.
102. Adamian, L. and J. Liang, *Prediction of transmembrane helix orientation in polytopic membrane proteins.* BMC Struct Biol, 2006. **6**: p. 13.
103. Fuchs, A., et al., *Co-evolving residues in membrane proteins.* Bioinformatics, 2007. **23**(24): p. 3312-9.
104. Fleishman, S.J., O. Yifrach, and N. Ben-Tal, *An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels.* J Mol Biol, 2004. **340**(2): p. 307-18.
105. Maeda, S., et al., *Structure of the connexin 26 gap junction channel at 3.5[thinsp]A resolution.* Nature, 2009. **458**(7238): p. 597-602.
106. Skerrett, I.M., et al., *Identification of amino acid residues lining the pore of a gap junction channel.* J. Cell Biol., 2002. **159**(2): p. 349-360.
107. Chen, Y.J., et al., *X-ray structure of EmrE supports dual topology model.* Proc Natl Acad Sci U S A, 2007. **104**(48): p. 18999-9004.

108. Sale, K., et al., *Optimal bundling of transmembrane helices using sparse distance constraints.* Protein Sci, 2004. **13**(10): p. 2613-27.
109. Kaback, H.R., M. Sahin-Toth, and A.B. Weinglass, *The kamikaze approach to membrane transport.* Nat Rev Mol Cell Biol, 2001. **2**(8): p. 610-20.
110. Kaback, H.R. and J. Wu, *From membrane to molecule to the third amino acid from the left with a membrane transport protein.* Q Rev Biophys, 1997. **30**(4): p. 333-64.
111. Stein, E.G., L.M. Rice, and A.T. Brunger, *Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation.* J Magn Reson, 1997. **124**(1): p. 154-64.
112. Abramson, J., et al., *Structure and mechanism of the lactose permease of Escherichia coli.* Science, 2003. **301**(5633): p. 610-5.
113. Abramson, J., et al., *The lactose permease of Escherichia coli: overall structure, the sugar-binding site and the alternating access model for transport.* FEBS Lett, 2003. **555**(1): p. 96-101.
114. Williams, K.A., et al., *Projection structure of NhaA, a secondary transporter from Escherichia coli, at 4.0 A resolution.* EMBO J, 1999. **18**(13): p. 3558-63.
115. Gerchman, Y., et al., *Oligomerization of NhaA, the Na+/H+ antiporter of Escherichia coli in the membrane and its functional and structural consequences.* Biochemistry, 2001. **40**(11): p. 3403-12.
116. Hilger, D., et al., *High-resolution structure of a Na+/H+ antiporter dimer obtained by pulsed electron paramagnetic resonance distance measurements.* Biophys J, 2007. **93**(10): p. 3675-83.
117. Rimon, A., T. Tzubery, and E. Padan, *Monomers of the NhaA Na+/H+ antiporter of Escherichia coli are fully functional yet dimers are beneficial under extreme stress conditions at alkaline pH in the presence of Na+ or Li+.* J Biol Chem, 2007. **282**(37): p. 26810-21.
118. Fischer, D., *Servers for protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(2): p. 178-82.
119. Fasnacht, M., J. Zhu, and B. Honig, *Local quality assessment in homology models using statistical potentials and support vector machines.* Protein Sci, 2007. **16**(8): p. 1557-68.
120. Eisenberg, D., R. Luthy, and J.U. Bowie, *VERIFY3D: assessment of protein models with three-dimensional profiles.* Methods Enzymol, 1997. **277**: p. 396-404.
121. Sippl, M.J., *Recognition of errors in three-dimensional structures of proteins.* Proteins, 1993. **17**(4): p. 355-62.
122. Wallner, B. and A. Elofsson, *Identification of correct regions in protein models using structural, alignment, and consensus information.* Protein Sci, 2006. **15**(4): p. 900-13.
123. Wallner, B. and A. Elofsson, *Can correct protein models be identified?* Protein Sci, 2003. **12**(5): p. 1073-86.
124. Tosatto, S.C., *The victor/FRST function for model quality estimation.* J Comput Biol, 2005. **12**(10): p. 1316-27.
125. Briggs, J.A., J. Torres, and I.T. Arkin, *A new method to model membrane protein structure based on silent amino acid substitutions.* Proteins, 2001. **44**(3): p. 370-5.
126. Stevens, T.J. and I.T. Arkin, *Substitution rates in alpha-helical transmembrane proteins.* Protein Sci, 2001. **10**(12): p. 2507-17.
127. Jones, D.T., *Improving the accuracy of transmembrane protein topology prediction using evolutionary information.* Bioinformatics, 2007. **23**(5): p. 538-44.
128. Kozachkov, L., K. Herz, and E. Padan, *Functional and structural interactions of the transmembrane domain X of NhaA, Na+/H+ antiporter of Escherichia coli, at physiological pH.* Biochemistry, 2007. **46**(9): p. 2419-30.
129. Pornillos, O., et al., *X-ray structure of the EmrE multidrug transporter in complex with a substrate.* Science, 2005. **310**(5756): p. 1950-3.
130. Ma, C. and G. Chang, *Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2852-7.
131. Chang, G., et al., *Retraction.* Science, 2006. **314**(5807): p. 1875.
132. Reyes, C.L. and G. Chang, *Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide.* Science, 2005. **308**(5724): p. 1028-31.
133. Dawson, R.J. and K.P. Locher, *Structure of a bacterial multidrug ABC transporter.* Nature, 2006. **443**(7108): p. 180-5.

134. Landau, M., et al., *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W299-302.
135. Olmea, O., B. Rost, and A. Valencia, *Effective use of sequence correlation and conservation in fold recognition.* J Mol Biol., 1999. **293**(5): p. 1221-39.
136. Muppirala , U.K. and Z. Li, *A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues.* Protein Eng Des Sel., 2006. **19**(6): p. 265-75. .
137. Mihalek, I., et al., *Combining inference from evolution and geometric probability in protein structure evaluation.* J Mol Biol., 2003. **331**(1): p. 263-79.
138. Yamashita, A., et al., *Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters.* Nature, 2005. **437**(7056): p. 215-23.
139. Forrest, L.R., et al., *Mechanism for alternating access in neurotransmitter transporters.* Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10338-43.